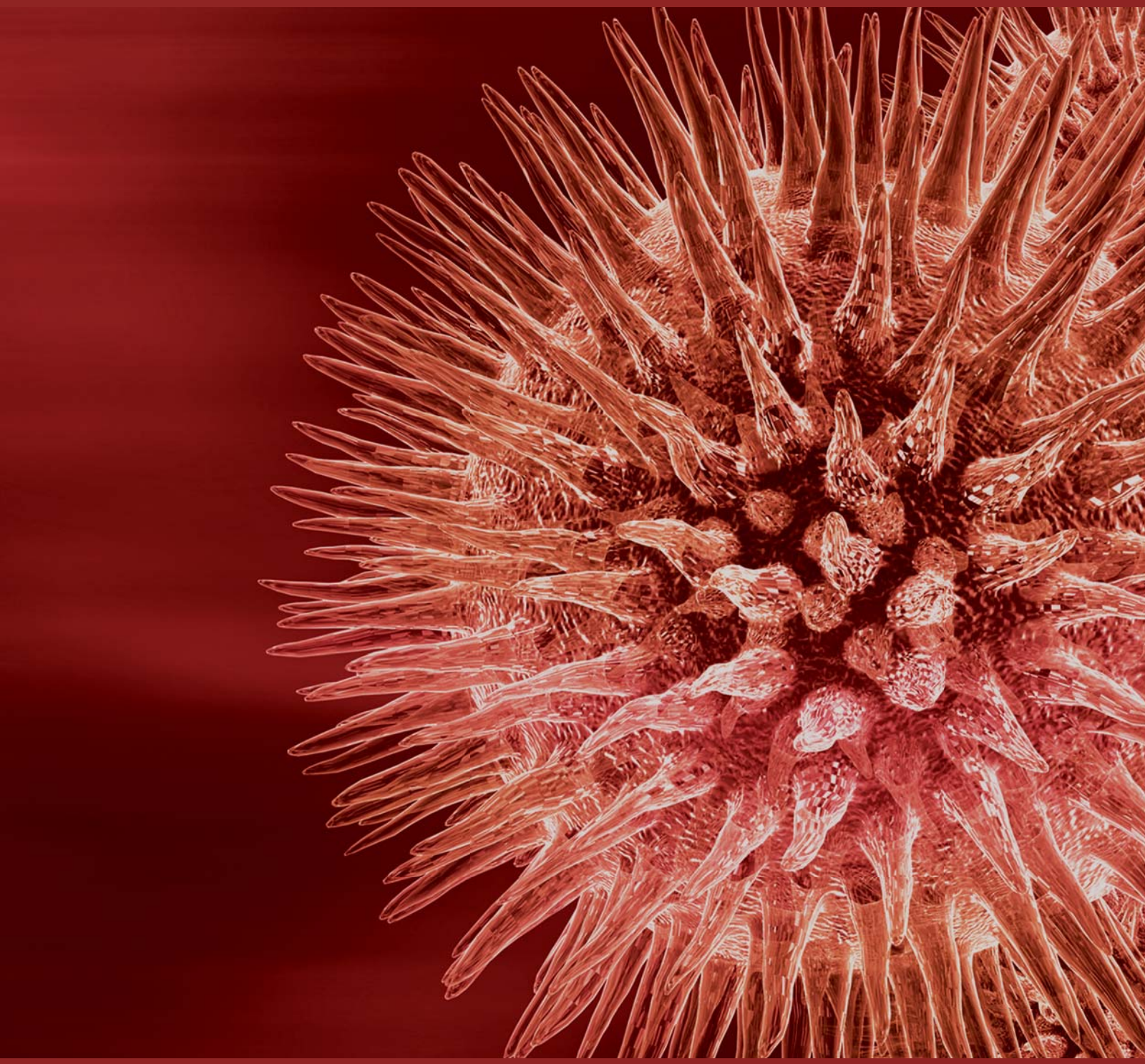
Journal of Biomedicine and Biotechnology

Vaccine Informatics

Guest Editors: Yongqun He, Rino Rappuoli, Anne S. De Groot, and Robert T. Chen



Vaccine Informatics

Journal of Biomedicine and Biotechnology

Vaccine Informatics

Guest Editors: Yongqun He, Rino Rappuoli, Anne S. De Groot, and Robert T. Chen

Editorial Board

The editorial board of the journal is organized into sections that correspond to the subject areas covered by the journal.

Agricultural Biotechnology

Guihua H. Bai, USA
Christopher P. Chanway, Canada
Ravindra N. Chibbar, Canada
Ian Godwin, Australia
Hari B. Krishnan, USA
Carol A. Mallory-Smith, USA
Dennis P. Murr, Canada
Rodomiro Ortiz, Mexico
B. C. Saha, USA
Mariam B. Sticklen, USA
Chiu-Chung Young, Taiwan

Animal Biotechnology

E. S. Chang, USA
Bhanu P. Chowdhary, USA
Noelle E. Cockett, USA
Peter Dovc, Slovenia
Scott C. Fahrenkrug, USA
Dorian J. Garrick, USA
Thomas A. Hoagland, USA
Tosso Leeb, Switzerland
James D. Murray, USA
Anita M. Oberbauer, USA
Jorge A. Piedrahita, USA
Daniel Pomp, USA
Kent M. Reed, USA
Lawrence Reynolds, USA
Lawrence B. Schook, USA
Mari A. Smits, The Netherlands
Leon Spicer, USA
J. Verstegen, USA
Matthew B. Wheeler, USA
Kenneth L. White, USA

Biochemistry

Robert Blumenthal, USA
David Ronald Brown, UK
Saulius Butenas, USA
Vittorio Calabrese, Italy
F. Castellino, USA
Roberta Chiaraluce, Italy
D. M. Clarke, Canada
Francesca Cutruzzolà, Italy
Paul W. Doetsch, USA
Hicham Fenniri, Canada
Nick V. Grishin, USA
J. Guy Guillemette, Canada
Paul W. Huber, USA
Chen-Hsiung Hung, Taiwan
Michael Kalafatis, USA
B. E. Kemp, Australia
Phillip E. Klebba, USA
Wen-Hwa Lee, USA
Richard D. Ludescher, USA
George Makhatadze, USA
Leonid Medved, USA
Susan A. Rotenberg, USA
Jason Shearer, USA
Andrei Surguchov, USA
John B. Vincent, USA
Yujun George Zheng, USA

Bioinformatics

T. Akutsu, Japan
Miguel A. Andrade, Germany
Mark Y. Borodovsky, USA
Rita Casadio, Italy
Artem Cherkasov, Canada
David Corne, UK
Sorin Draghici, USA
Stavros J. Hamodrakas, Greece
Paul Harrison, USA
George Karypis, USA
Jack A. Leunissen, The Netherlands
Guohui Lin, Canada
Satoru Miyano, Japan
Zoran Obradovic, USA
Florencio Pazos, Spain
Zhirong Sun, China
Ying Xu, USA
A. Zelikovsky, USA
Albert Zomaya, Australia

Biophysics

Miguel Castanho, Portugal
P. Bryant Chase, USA
Kuo-Chen Chou, USA
Rizwan Khan, India
Ali A. Khraibi, Saudi Arabia
Rumiana Koynova, USA
Serdar Kuyucak, Australia
Jianjie Ma, USA
S. B. Petersen, Denmark
Peter Schuck, USA
Claudio M. Soares, Portugal

Cell Biology

Omar Benzakour, France
Sanford I. Bernstein, USA
Phillip I. Bird, Australia
Eric Bouhassira, USA
Mohamed Boutjdir, USA
Chung-Liang Chien, Taiwan
Richard Gomer, USA
Paul J. Higgins, USA
Pavel Hozak, Czech Republic
Xudong Huang, USA
Anton M. Jetten, USA
Seamus J. Martin, Ireland
Manuela Martins-Green, USA
Shoichiro Ono, USA
George Perry, USA
M. Piacentini, Italy
George E. Plopper, USA
Lawrence Rothblum, USA
Michael Sheetz, USA
James L. Sherley, USA
G. S. Stein, USA
Richard Tucker, USA
Thomas van Groen, USA
Andre Van Wijnen, USA
Steve Winder, UK
Chuanyue Wu, USA
Bin-Xian Zhang, USA

Genetics

Adewale Adeyinka, USA
Claude Bagnis, France
J. Birchler, USA
Susan Blanton, USA
Barry J. Byrne, USA
R. Chakraborty, USA
Domenico Coviello, Italy
Sarah H. Elsea, USA
Celina Janion, Poland
J. Spencer Johnston, USA
M. Ilyas Kamboh, USA
Feige Kaplan, Canada
Manfred Kayser, The Netherlands
Brynn Levy, USA
Xiao Jiang Li, USA
Thomas Liehr, Germany
James M. Mason, USA
Mohammed Rachidi, France
Raj S. Ramesar, South Africa
Elliot D. Rosen, USA
Dharambir K. Sanghera, USA
Michael Schmid, Germany
Markus Schuelke, Germany
Wolfgang Arthur Schulz, Germany
Jorge Sequeiros, Portugal
Mouldy Sioud, Norway
Rongjia Zhou, China

Genomics

Vladimir Bajic, Saudi Arabia
Margit Burmeister, USA
Settara Chandrasekharappa, USA
Yataro Daigo, Japan
J. Spencer Johnston, USA
Vladimir Larionov, USA
Thomas Lufkin, Singapore
Joakim Lundeberg, Sweden
John L. McGregor, France
John V. Moran, USA
Yasushi Okazaki, Japan
Gopi K. Podila, USA
Momiao Xiong, USA

Immunology

Hassan Alizadeh, USA
Peter Bretscher, Canada
Robert E. Cone, USA
Terry L. Delovitch, Canada
Anthony L. DeVico, USA
Nick Di Girolamo, Australia
Don Mark Estes, USA
Soldano Ferrone, USA
Jeffrey A. Frelinger, USA
John Robert Gordon, Canada
James D. Gorham, USA
Silvia Gregori, Italy
Thomas Griffith, USA
Young S. Hahn, USA
Dorothy E. Lewis, USA
Bradley W. McIntyre, USA
R. Mosley, USA
Marija Mostarica-Stojković, Serbia
Hans Konrad Muller, Australia
Ali Ouaissi, France
Kanury V. S. Rao, India
Yair Reisner, Israel
Harry W. Schroeder, USA
Wilhelm Schwaeble, UK
Nilabh Shastri, USA
Yufang Shi, China
Piet Stinissen, Belgium
Hannes Stockinger, Austria
J. W. Tervaert, The Netherlands
Graham R. Wallace, UK

Microbial Biotechnology

Jozef Anné, Belgium
Yoav Bashan, Mexico
Marco Bazzicalupo, Italy
Nico Boon, Belgium
Luca Simone Cocolin, Italy
Peter Coloe, Australia
Daniele Daffonchio, Italy
Han de Winde, The Netherlands
Yanhe Ma, China
Bernd H. A. Rehm, New Zealand
Angela Sessitsch, Austria
Effie Tsakalidou, Greece
J. Wiegel, USA

Microbiology

D. Beighton, UK
Steven R. Blanke, USA
Stanley Brul, The Netherlands
Isaac K. O. Cann, USA
Peter Dimroth, Switzerland
Stephen K. Farrand, USA
Alain Filloux, UK
Gad Frankel, UK
Roy Gross, Germany
Hans-Peter Klenk, Germany
Tanya Parish, UK
Gopi K. Podila, USA
Frederick D. Quinn, USA
Didier A. Raoult, France
Isabel Sá-Correia, Portugal
P. L. C. Small, USA
Lori Snyder, UK
Michael Thomm, Germany
H. C. van der Mei, The Netherlands
Schwan William, USA

Molecular Biology

Rudi Beyaert, Belgium
Michael Bustin, USA
Douglas Cyr, USA
K. Iatrou, Greece
Lokesh Joshi, Ireland
David W. Litchfield, Canada
Wuyuan Lu, USA
Patrick Matthias, Switzerland
John L. McGregor, France
S. L. Mowbray, Sweden
Elena Orlova, UK
Yeon-Kyun Shin, USA
William S. Trimble, Canada
Lisa Wiesmuller, Germany
Masamitsu Yamaguchi, Japan

Oncology

Colin Cooper, UK
F. M. J. Debruyne, The Netherlands
Nathan Ames Ellis, USA
Dominic Fan, USA
Gary E. Gallick, USA
Daila S. Gridley, USA
Xin-yuan Guan, Hong Kong
Anne Hamburger, USA
Manoor Prakash Hande, Singapore
Beric Henderson, Australia
Steve B. Jiang, USA
Daehee Kang, Republic of Korea
Abdul R. Khokhar, USA
Rakesh Kumar, USA
Macus Tien Kuo, USA
Eric W. Lam, UK
Sue-Hwa Lin, USA
Kapil Mehta, USA
Orhan Nalcioglu, USA
Vincent C. O. Njar, USA
P. J. Oefner, Germany
Allal Ouhtit, USA
Frank Pajonk, USA
Waldemar Priebe, USA
F. C. Schmitt, Portugal
Sonshin Takao, Japan
Ana Maria Tari, USA
Henk G. Van Der Poel, The Netherlands
Haodong Xu, USA
David J. Yang, USA

Pharmacology

Abdel A. Abdel-Rahman, USA
M. Badr, USA
Stelvio M. Bandiera, Canada
Ronald E. Baynes, USA
R. Keith Campbell, USA
Hak-Kim Chan, Australia
Michael D. Coleman, UK
J. Descotes, France
Dobromir Dobrev, Germany
Ayman El-Kadi, Canada
Jeffrey Hughes, USA
Kazim Husain, USA
Farhad Kamali, UK
Michael Kassiou, Australia
Joseph J. McArdle, USA
Mark J. McKeage, New Zealand
Daniel T. Monaghan, USA
T. Narahashi, USA
Kennerly S. Patrick, USA
Vickram Ramkumar, USA
Michael J. Spinella, USA
Quadiri Timour, France
Todd W. Vanderah, USA
Val J. Watts, USA
David J. Waxman, USA

Plant Biotechnology

P. L. Bhalla, Australia
J. R. Botella, Australia
Elvira Gonzalez De Mejia, USA
H. M. Häggman, Finland
Liwen Jiang, Hong Kong
Pulugurtha Bharadwaja Kirti, India
Yong Pyo Lim, Republic of Korea
Gopi K. Podila, USA
Ralf Reski, Germany
Sudhir Kumar Sopory, India

Toxicology

Michael Aschner, USA
Michael L. Cunningham, USA
Laurence D. Fechter, USA
Hartmut Jaeschke, USA
Youmin James Kang, USA
M. Firoze Khan, USA
Pascal Kintz, France
R. S. Tjeerdema, USA
Kenneth Turteltaub, USA
Brad Upham, USA

Virology

Contents

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 765762, 2 pages
doi:10.1155/2010/765762

Editorial

Vaccine Informatics

Yongqun He,[1] Rino Rappuoli,[2] Anne S. De Groot,[3,4] and Robert T. Chen[5]

[1] *Unit for Laboratory Animal Medicine, Department of Microbiology and Immunology, Center for Computational Medicine and Bioinformatics and Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, MI 48109, USA*
[2] *Novartis Vaccines and Diagnostics, 53100 Siena, Italy*
[3] *EpiVax, Inc., Providence, RI 02903, USA*
[4] *Institute for Immunology and Informatics, University of Rhode Island, Providence, RI 02903, USA*
[5] *HIV Vaccine and Special Studies Team, Centers for Disease Control and Prevention (CDC/DHAP/EB), Atlanta, GA 30333, USA*

Correspondence should be addressed to Yongqun He, yongqunh@umich.edu

Received 31 December 2010; Accepted 31 December 2010

Global public health has dramatically increased due to the successful and effective implementation of immunization programs that utilize major infectious disease vaccines. Among the most famous vaccines are Jenner's smallpox vaccine, Pasteur's rabies and anthrax vaccines, and the Bacillus Calmette-Guérin (or Bacille Calmette-Guérin, BCG) vaccine for tuberculosis. The positive effect of vaccination on public health is highlighted by the success of the global smallpox eradication campaign, and the scope of vaccine development for the prevention and treatment of various diseases in both humans and animals continues to grow on a daily basis.

Vaccine informatics is an emerging field that develops and applies computational, statistical, and bioinformatics methods to study vaccine and vaccination-related issues in different stages of research, development, clinical trial, post-licensure clinical uses, and surveillance. The field began centuries ago with widespread vaccination monitoring to determine vaccine safety and clinical applicability. Yet it was not until the growth of the immunoinformatics field in the 1980s that vaccine informatics emerged as an important area of study, when many methods were developed and implemented to predict T-cell and B-cell immune epitopes. These T-cell and B-cell immune epitopes were identified as potential vaccine targets and were recognized as crucial to understanding basic protective immunity. In the 1990s, bioinformatics became an emerging science after more and more DNA, RNA, and protein sequences were identified and studied. In the postgenomics and information era, the application of bioinformatics tools within the vaccinology field has fostered dramatic progression for both literature and data on vaccine informatics, creating the potential to revolutionize every aspect of the pre- and postlicensure vaccine enterprise.

Due to the differing data types available, separate bioinformatics technologies have been developed and applied to address specific vaccine-related issues.

(1) Immune epitope data: an immune epitope is a part of an antigen recognized by antibodies, B cells, or T cells within the immune system. The discovery and documentation of epitope data constructs a foundation which new epitope prediction and vaccine development algorithms are built upon.

(2) Gene and genome DNA sequence data: microbial gene DNA sequences have been used for vaccine development through gene cloning and engineering, protein isolation and purification, and vaccine construction. The availability of whole genome sequences for microbial pathogens makes it possible to predict and select protective antigens as vaccine candidates through *in silico* analysis within the reverse vaccinology strategy.

(3) RNA and protein gene expression data: advanced computational approaches have been developed to analyze high throughput DNA microarray and proteomics data. Detection of RNA and protein gene expression profiles in hosts or pathogens facilitates vaccine design and analysis of hosts' vaccine-induced immune responses.

(4) Clinical data: statistics methods have routinely been used to analyze data from clinical vaccine trials. Tracking immunization history in computerized registries, modeling the impact of alternative immunization strategies, timely detection of vaccine preventable disease outbreaks, and analyses of vaccine safety concerns are all possible due to analysis of current clinical data. In addition, informatics is also changing postlicensure immunization policy and programs.

(5) Literature data: as the vaccinology field expands and the volume of available data increases, the ability to determine and retrieve useful data is crucial in improving and facilitating further vaccine development and protective immune mechanism research. High throughput literature mining programs are being developed toward these goals.

In addition, mathematical modeling approaches are able to combine both synthetic and real vaccine data to engage specific issues. It is also important to integrate various vaccine-associated data types for more advanced vaccine research and development. To support better vaccine data integration and analysis, manually curated vaccine data have been stored in various vaccine databases (e.g., VIOLIN: http://www.violinet.org/) or represented using a community-based vaccine ontology (VO).

This special issue aspires to become an international forum for researchers to provide testimony regarding new and exciting developments in the field. The first special issue on vaccine informatics includes 20 articles that cover a variety of vaccine research areas. This special issue contains 7 research papers, 5 methodology papers, and 7 review papers. The topics in this issue include vaccine informatics reviews, immunoinformatics analysis of immune epitopes, vaccine design based on immunoinformatics and reverse vaccinology, transcriptomics and proteomics analyses, postlicensure vaccine informatics, vaccine literature mining, and immune database management with potential for vaccine research and development. In addition to the many uses of informatics highlighted by the included articles, we should note that systems biology approaches to vaccines are rapidly emerging. These approaches rely entirely on informatics to analyze large datasets deriving from complex and multidimensional systems.

The field of vaccine informatics has attracted some of the most innovative scientists from the vaccinology, microbiology, immunology, epidemiology, and bioinformatics branches.

Acknowledgments

As Editors, we are truly privileged to have had many of these individuals contribute to this special issue. It is our hope that the following literature inspires the scientific community to devote their time and efforts into this exciting area. It is through the hard work and dedication of scientists that we are able to develop and utilize cutting edge vaccine informatics methods to advance vaccine R&D and improve immunization programs. We would like to express our sincere gratitude to the contributing authors as well as to the Journal of Biomedicine and Biotechnology.

Yongqun He
Rino Rappuoli
Anne S. De Groot
Robert T. Chen

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 218590, 26 pages
doi:10.1155/2010/218590

Review Article

Emerging Vaccine Informatics

Yongqun He,[1] Rino Rappuoli,[2] Anne S. De Groot,[3,4] and Robert T. Chen[5]

[1] *Department of Microbiology and Immunology, Unit for Laboratory Animal Medicine, Center for Computational Medicine and Bioinformatics, and Comprehensive Cancer Center, University of Michigan Medical School, Ann Arbor, MI 48109, USA*
[2] *Novartis Vaccines and Diagnostics, 53100 Siena, Italy*
[3] *EpiVax, Inc., Providence, RI 02903, USA*
[4] *Institute for Immunology and Informatics, University of Rhode Island, Providence, RI 02903, USA*
[5] *HIV Vaccine and Special Studies Team, Centers for Disease Control and Prevention (CDC/DHAP/EB), Atlanta, GA 30333, USA*

Correspondence should be addressed to Yongqun He, yongqunh@med.umich.edu

Received 8 December 2010; Accepted 31 December 2010

Academic Editor: Rodomiro Ortiz

Vaccine informatics is an emerging research area that focuses on development and applications of bioinformatics methods that can be used to facilitate every aspect of the preclinical, clinical, and postlicensure vaccine enterprises. Many immunoinformatics algorithms and resources have been developed to predict T- and B-cell immune epitopes for epitope vaccine development and protective immunity analysis. Vaccine protein candidates are predictable *in silico* from genome sequences using reverse vaccinology. Systematic transcriptomics and proteomics gene expression analyses facilitate rational vaccine design and identification of gene responses that are correlates of protection *in vivo*. Mathematical simulations have been used to model host-pathogen interactions and improve vaccine production and vaccination protocols. Computational methods have also been used for development of immunization registries or immunization information systems, assessment of vaccine safety and efficacy, and immunization modeling. Computational literature mining and databases effectively process, mine, and store large amounts of vaccine literature and data. Vaccine Ontology (VO) has been initiated to integrate various vaccine data and support automated reasoning.

1. Introduction

While the history of vaccines is relatively short, vaccines have contributed to dramatic improvements in public health worldwide. Jenner's description of smallpox prevention in 1796 [1] is the most commonly recognized "start" of vaccine research in European historical documents, although variolation had been practiced in Asia centuries earlier. Critical advances in vaccine science took place in the late 19th and early 20th centuries, by scientists such as Pasteur, Koch, von Behring, Calmette, Guérin, and Ehrlich [2]. Discoveries by these early vaccine researchers contributed to the development of antiserums, antitoxins, and live, attenuated bacterial vaccines. The discovery of tissue culture methods for viral and bacterial propagation *in vitro* during the period from 1930 to 1950 was a technical advance that enabled the development of vaccines against many viruses including measles and polio. Further advances in cell culture techniques, carbohydrate chemistry, molecular biology, and immunology have led to the modern era of "subunit" vaccine development. The recombinant hepatitis B vaccine, one of the first subunit vaccines, was licensed in 1986 [2]. This marked the beginning of the molecular biology phase of vaccine development. At present, human vaccines are used in the prevention of more than thirty infectious diseases. Due to the success of the smallpox eradication campaign in 1960s and 1970s, the powerful impact of vaccines on human health is universally recognized [3]. In addition, there exist a large number of animal vaccines [4].

With the advent of computers and informatics, new approaches have been devised that facilitate vaccine research and development. Immunoinformatics targets the use of mathematical and computational approaches to address immunological questions. Since the 1980s, many immunoinformatics methods have been developed and used to predict T-cell and B-cell immune epitopes [5]. Indeed, many predicted T- and B-cell immune epitopes are possible epitope vaccine targets. Experimentally verified immune epitopes are

now stored in web-based databases which are freely available for further analysis [6]. Immune epitope studies are crucial to uncover basic protective immune mechanisms.

A new era of vaccine research began in 1995, when the complete genome of *Haemophilus influenzae* (a pathogenic bacterium) was published [7]. In parallel with advances in molecular biology and sequencing technology, bioinformatics analysis of microbial genome data has allowed *in silico* selection of vaccine targets. Further advances in the field of immunoinformatics have led to the development of hundreds of new vaccine design algorithms. This novel approach for developing vaccines has been named reverse vaccinology [8] or immunome-derived vaccine design [9]. Reverse vaccinology was first applied to the development of vaccines against serogroup B *Neisseria meningitides* (MenB) [10]. With the availability of multiple genomes sequenced for pathogens, it is now possible to run comparative genomics analyses to find vaccine targets shared by many pathogenic organisms.

In the postgenomics era, high throughput-omics technologies-genomics, transcriptomics, proteomics, and large-scale immunology assays enable the testing and screening of millions of possible vaccine targets in real time. Bioinformatics approaches play a critical role in analyzing large amounts of high throughput data at differing levels, ranging from data normalization, significant gene expression detection, function enrichment, to pathway analysis.

Mathematical simulation methods have also been developed to model various vaccine-associated areas, ranging from analysis of host-pathogen interactions and host-vaccine interactions to cost cost-effectiveness analyses and simulation of vaccination protocols. The mathematical modeling approaches have contributed dramatically to the understanding of fundamental protective immunity and optimization of vaccination procedures and vaccine distribution.

Informatics is also changing postlicensure immunization policies and programs. Computerized immunization registries or immunization information systems (IIS) are effective approaches to track vaccination history. Bioinformatics has widely been used to improve surveillance of (1) vaccine safety using systems such as the Vaccine Adverse Event Reporting System (VAERS, http://vaers.hhs.gov/) [11] and the Vaccine Safety Datalink (VSD) [12] project and (2) vaccine effectiveness for each of the target vaccine preventable diseases via their respective public health surveillance systems. Computational methods have also been applied to model the impact of alternative immunization strategies and to detect outbreaks of vaccine preventable diseases and safety concerns related to vaccinations as well.

With the large amounts of vaccine literature and data becoming available, it is not only challenging but crucial to perform vaccine literature mining, generate well-annotated and comprehensive vaccine databases, and integrate various vaccine data to enhance vaccine research. Computational vaccine literature mining will allow us to efficiently find vaccine information. To effectively organize and analyze the huge amounts of vaccine data produced and published in the postgenomics and information era, many vaccine-related databases, such as the VIOLIN vaccine database and analysis system (http://www.violinet.org/) [13] and AIDS vaccine trials database (http://www.iavireport.org/trials-db/), have been developed and are available on the web. However, relational databases are not ideal for data sharing since different databases may use different schemas and formats. A biomedical ontology is a consensus-based controlled vocabulary of terms and relations, with associated definitions that are logically formulated in such a way as to promote automated reasoning. Ontologies are able to structure complex biomedical domains and relate the myriads of data accumulated in such a fashion as to permit shared understanding of vaccines among different resources. The Vaccine Ontology (VO; http://www.violinet.org/vaccineontology/) is a novel open-access ontology in the domain of vaccine [14]. Recent studies show that VO can be used to support vaccine data integration and improve vaccine literature mining [15, 16].

In summary, vaccine informatics is an emerging field of research that focuses on the development and applications of computational approaches to advance vaccine research and development (R&D) and improve immunization programs. Vaccine informatics plays an important role in every aspect of pre- and postlicensure vaccine enterprises (Figure 1). This paper summarizes the history of vaccine informatics developments in advancing vaccine research and development and immunization programs.

2. Immunoinformatics and Vaccine Design

This section describes immunoinformatics and how it is used for vaccine design and to study protective immune responses to vaccines.

2.1. Brief History of Immunoinformatics Approaches for Vaccine Design. The first immunoinformatics tools for vaccine design were developed in the 1980s by DeLisi and Berzofsky and others [17]. Chief among vaccine design informatics tools are epitope-mapping algorithms. Since the T-cell epitopes are bound in a linear form to the human leukocyte antigen (HLA), the interface between ligands and T-cells can now be modeled with accuracy. A large number of T-cell epitope-mapping algorithms have consequently been developed [18, 19]. These tools now make it possible to start with the entire proteome of a pathogen and rapidly identify putative T-cell epitopes. Such information is immensely valuable for the development of new vaccines, diagnostic purposes, and for studying the pathology of infectious diseases [5, 20–27].

Several different routes for vaccine development have been pursued. One method, which has been used by De Groot and Martin [24, 28], is to synthesize the putative T-cell epitopes and screen peripheral blood mononuclear cells (PBMC) isolated from human subjects infected with the target pathogen (or have a target cancer) for immune response to the epitopes. A T-cell *in vitro* response to a specific peptide epitope (typically measured by ELISA or ELISpot assay) served as an indicator that the protein from which the peptide was derived was expressed, processed, and presented to the immune system in the course of a "natural" immune response. This approach, often considered a means of

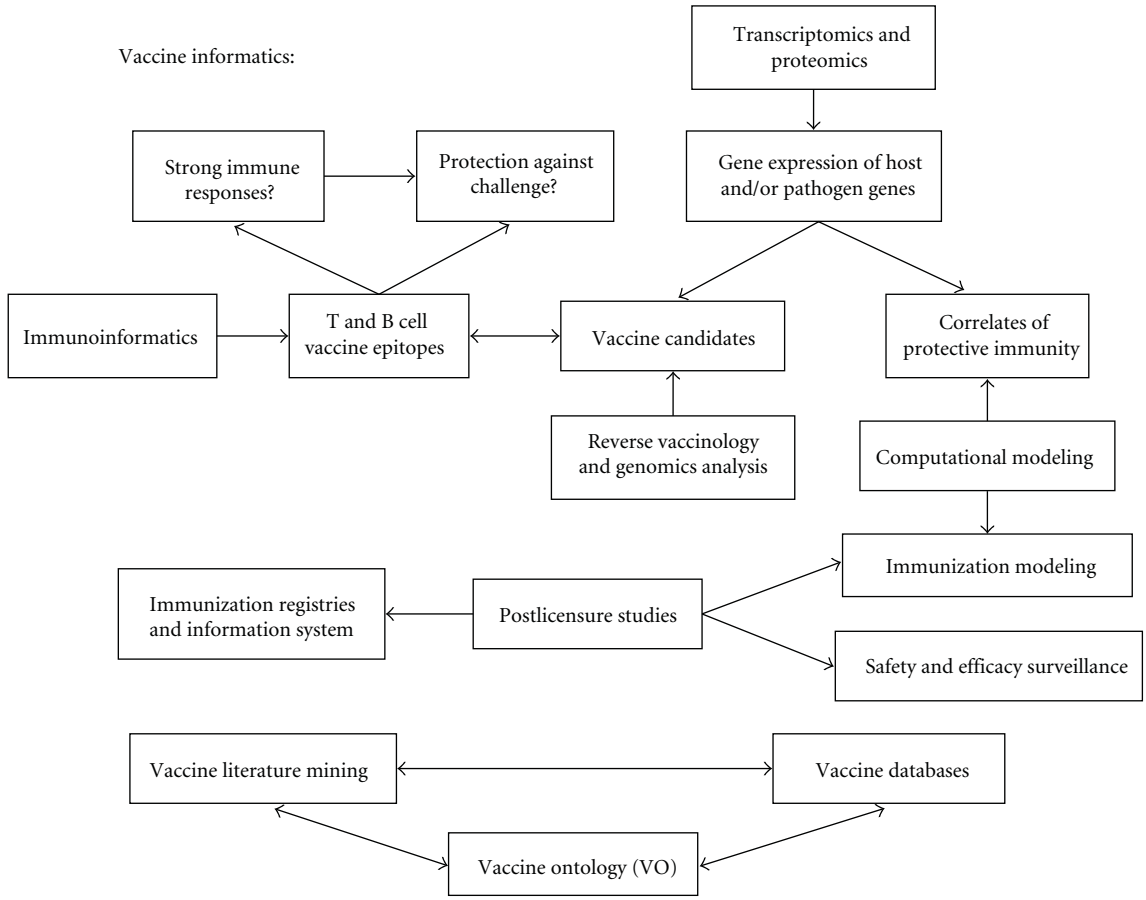


Figure 1: Overview of current vaccine informatics topics.

making epitope-based vaccines (see below), can also be used to identify proteins for use in vaccine development. This approach, described by De Groot and Martin's group as "fishing for antigens using epitopes as bait", has been used to discover new vaccine antigens for *F. tularensis* (a bioterror agent) [28], tuberculosis [24], smallpox [29], and *H. pylori* [30].

The proteome of *M. tuberculosis* (*Mtb*), the etiologic agent of TB, contains almost 4,000 proteins. Evaluating each one using the straightforward but expensive and laborious approach of synthesizing and testing overlapping peptides could take decades. Using epitope mapping tools, it is now possible to screen a whole proteome *in silico*, followed by a finer focus on the resulting sets of peptides [5].

The ability to accurately predict T-cell epitopes from raw genomic data is fundamental to the development of novel vaccines, and serves as the starting point for a number of research projects. Freeing the researchers from the constraints of predetermined sets of "virulence genes" has resulted in some remarkable discoveries. McMurry and De Groot [24] found extraordinary diversity of human immune responses to proteins in the *Mtb* genome that have yet to be ascribed a function, suggesting that human immune response is omnivorous and is not focused on recognition of a single "immunodominant" protein. In addition, these investigators have found a remarkable similarity between *Francisella tularensis* (the etiologic agent for Tularemia) and (human) self, at the epitope level [28]. Thus informatics, starting at the genome, may reveal potential antigenic relationships between human proteins and pathogens, or even commensal organisms, which might predetermine individual immune response, that is, prior exposure to a given pathogen may tune immune response to a second pathogen [31].

An alternative approach, "Reverse Vaccinology," a term coined by Rappuoli, starts with predicting putative vaccine candidates by *in silico* genomics analysis based on yet different criteria. The predicted vaccine candidates (e.g., bacterial surface proteins) are thought to stimulate protective immunity. Candidate proteins can be evaluated experimentally by demonstrating an immune response that correlates with *in vivo* protection [10]. The Reserve Vaccinology approach is well discussed later on (see below).

2.2. MHC Polymorphism, Epitope Variations, and Vaccine Design. The success rate of vaccine development decreases with the increasing variability of the surface antigens of pathogens and the decreasing ability of antibodies to confer protective immunity [32]. Fortunately, vaccine informatics

tools are being developed that increase the accuracy of vaccine target prediction for variable pathogens and help vaccinologists triage antigens.

T-cells are activated by direct interaction with antigen presenting cells (APCs). On the molecular level, the initial interaction occurs between the T-cell receptor and peptides derived from endogenous and exogenous proteins that are bound in the cleft of MHC class I or class II molecules. In general, MHC class I molecules present peptides 8–10 amino acids in length and are predominantly recognized by $CD8^+$ cytotoxic T lymphocytes (CTLs). Class I peptides usually contain an MHC I-allele-specific motif composed of two conserved anchor residues [33–35]. Peptides presented by class II molecules are longer, more variable in size, and have more complex anchor motifs than those presented by class I molecules [36–38]. MHC class II molecules bind peptides consisting of 11–25 amino acids and are recognized by $CD4^+$ T helper (Th) cells.

MHC class I molecules present peptides obtained from proteolytic digestion of endogenously synthesized proteins. Host- or pathogen-derived intracellular proteins are cleaved by a complex of proteases in the proteasome. Small peptide fragments are then typically transported by ATP-dependent transporters associated with antigen processing (TAPs) and also by TAP-independent means into the endoplasmic reticulum (ER), where they form complexes with nascent MHC class I heavy chains and beta-2-microglobulin. The peptide-MHC class I complexes are transported to the cell surface for presentation to the receptors of $CD8^+$ T-cells [39–41].

MHC class II molecules generally bind peptides derived from the cell membrane or from extracellular proteins that have been internalized by APCs. The proteins are initially processed in the MHC class II compartment (MIIC). Inside the MIIC, MHC is initially bound to class II-associated invariant chain peptide (CLIP) which protects the MHC from binding to endogenous peptides. Peptides generated by proteolytic processing within endosomes replace CLIP in a reaction catalyzed by the protein HLA-DM [42, 43]. The class II molecules bound to peptide fragments are transported to the surface of APCs for presentation

To complicate matters further, HLA molecules bind different peptides due to the configuration of their HLA binding pockets. This is the source of genetic diversity of immune responses [34]. Fortunately, there is some conservation between HLA pockets, and both DeLisi and Sette have addressed the issue of HLA coverage for epitope predictions by demonstrating that epitope-based vaccines containing epitopes restricted by selected "supertype" Class I and Class II HLA can provide the broadest possible coverage of the human population [44, 45]. De Groot and Martin have constructed an algorithm, Aggregatrix, which uses the "set cover" method to identify the best set of peptides from a pathogen that would yield the broadest coverage of HLA if included in a vaccine. The Aggregatrix algorithm selects optimized epitope sets which, in terms of immunogenicity and genetic conservation, collectively "cover" a wide variety of known circulating strain variants of a given pathogen and a majority of the common human HLA types [46]. The Conservatrix algorithm is used to identify highly conserved peptide segments contained within multiple isolates of variable pathogens such as retroviruses [47]. The amino acid sequences of protein isolates are parsed into 9 mer frames overlapping by eight amino acids. The resulting peptide set yields a list of unique segments and appearance frequencies. Highly conserved sequences are thought to be important in the evolutionary "fitness" of pathogens and thus are unlikely to change in an attempt to evade the immune system. Conserved sequences can be analyzed using epitope prediction software.

2.3. T-Cell Epitope Mapping. Although textbooks teach that protective immune response is attributed to the development of protective antibodies, the immune response to attenuated intact viruses and subunit vaccines is to a very large degree dependent on T-cell recognition of peptide epitopes bound to MHC. Thus targeting antigens that contain many $CD4^+$ T helper epitopes may lead to the selection of good B-cell antigens as well as immunogens for effective CD8 responses—this is because $CD4^+$ T helper cells are critically important to the development of memory B-cell (antibody) and memory CTL (cytotoxic T-cell) responses, in addition to being active against pathogens on their own. T helper cells have been called the "conductors of the immune system orchestra" [20]. CTLs generally play a role in the containment of viral and bacterial infection [48], and the prevalence of CTLs usually correlates with the rate of pathogen clearance. Regulatory T-cells are also represented among $CD4^+$ T-cells, although some $CD8^+$ Tregs have been described.

T-cell epitope algorithms now achieve a high degree of prediction accuracy (in the range of 90 to 95% Positive Predictive Value). For example, epitope mapping tools can now be compared to other available tools, using the Immune Epitope Database "gold standard" as described by Wang et al. [49]. A list of epitope mapping tools, ancillary algorithms, and their comparative features is provided in Table 1. A number of the epitope mapping tools are available to researchers via the web. These include the tool available at the SYFPEITHI website [50] and an HLA binding prediction tool available on at the National Institutes of Health (BIMAS) [51]. A recently developed set of tools has now been made available through the Immunome Epitope Database. Each of these tools has been described and validated [49]. One such proprietary algorithm, EpiMatrix, is in active use in the pharmaceutical industry [52]. While none of these sites yield exactly the same predictions, all predictions are quite accurate, especially when compared to results obtained with early epitope mapping tools (e.g., SYFPEITHI and BIMAS) [49, 52]. In general, the newer and more actively maintained algorithms tend to outperform the older more static predictive methods.

With many machine learning techniques developed since early 1990s for T-cell epitope predictions, it is possible to comparatively evaluate them through prediction performance assessments [49, 53–55]. Lin et al. compared 30 servers developed by 19 groups that can predict HLA-I binding peptides [53]. Their benchmarking study showed that predictions of six out of seven of HLA-I binding peptides achieved excellent classification accuracy. In general, nonlinear predictors outperform matrix-based predictors,

Table 1: List of online tools for T-cell epitope prediction.

Tool	Website	Type	Class	Refs
ANNPRED	http://www.imtech.res.in/raghava/nhlapred/neural.html	ANN	I	[234]
Bimas	http://www-bimas.cit.nih.gov/molbio/hla_bind/	QM	I	[51]
EpiJen	http://www.ddg-pharmfac.net/epijen/EpiJen/EpiJen.htm	Multi-step algorithm	I	[235]
EPIMHC	http://imed.med.ucm.es/epimhc/	user made Profiles	I, II	[236]
EpiMatrix	http://www.epivax.com/	Matrix-based and pocket profile	I, II	[237]
HLABIND	http://atom.research.microsoft.com/hlabinding/hlabinding.aspx	Adaptive Double Threading	I	[238]
IEDB	http://tools.immuneepitope.org/analyze/html/mhc_binding.html	ARB-QM, SMM-QM, ANN-regression	I	[239]
KISS	http://cbio.ensmp.fr/kiss/	SVM	I	[240]
MHC2PRED	http://www.imtech.res.in/raghava/mhc2pred/	SVM	II	[241]
MHC Pred	http://www.jenner.ac.uk/MHCPred	Partial least-squares-based multivariate statistical method	I, II	[242]
MULTIPRED	http://antigen.i2r.a-star.edu.sg/multipred/	ANN, pHMM	I, II	[243]
MOTIF_SCAN	http://www.hiv.lanl.gov/content/immunology/motif_scan/motif_scan	Sequence Motifs	I, II	—
NetCTL	http://www.cbs.dtu.dk/services/NetCTL/	Multi-step algorithm	I, CTL	[60]
NetCTLSpan	http://www.cbs.dtu.dk/services/NetCTLpan/	Multi-step algorithm	I, CTL	[62]
NetMHC	http://www.cbs.dtu.dk/services/NetMHC/	ANN	I	[244]
netMHCII	http://www.cbs.dtu.dk/services/NetMHCII/	SMM-QM	II	[245]
netMHCpan	http://www.cbs.dtu.dk/services/NetMHCpan/	ANN-regression	I	[246]
netMHCIIpan	http://www.cbs.dtu.dk/services/NetMHCIIpan/	ANN-regression	II	[247]
PEPVAC	http://imed.med.ucm.es/PEPVAC/	Profiles or PSSM	I	[248]
PREDEP	http://margalit.huji.ac.il/Teppred/mhc-bind/index.html	Threading	I	[249]
POPI	http://iclab.life.nctu.edu.tw/POPI/	SVM	I, II	[250]
PROPREDI	http://www.imtech.res.in/raghava/propred1/	QM	I	[251]
PROPRED	http://www.imtech.res.in/raghava/propred/	QM	II	[252]
RANKPEP	http://imed.med.ucm.es/Tools/rankpep.html	Profiles or PSSM	I, II	[253]
SVMHC	http://abi.inf.uni-tuebingen.de/SVMHC	SVM	I, II	[254]
SVRMHC	http://svrmhc.biolead.org/	SVM-regression	I, II	[255]
SYFPEITHI	http://www.syfpeithi.de/	Motif matrices	I, II	[50]
TEPITOPE	http://www.vaccinome.com/	QM	II	[256]
Vaxign	http://www.violinet.org/vaxign/	PSSM	I, II	[110]

Abbreviations: ANN: artificial neural networks; PSSM: position-specific scoring matrix; QM: quantitative matrices; SMM: stabilized matrix method; SVM: support vector machine; multistep algorithm: integrating predictions of proteasomal cleavage, TAP transport efficiency, and MHC class I affinity.

and most predictors can be improved by non-linear transformations of their raw prediction scores [53]. While good performance has been achieved for MHC class I predictions, there is still limited success for prediction of epitopes for HLA class II [54, 55]. The low prediction accuracy of HLA-II binding peptides is due to several factors: (a) insufficient or low-quality training data, (b) difficulty in identifying 9-mer binding cores within longer peptides used for training and lack of consideration of the influence of flanking residues, and (c) relative permissiveness of the binding groove of HLA-II molecules for peptide binding, which limits the stringency of binding [54].

Adequate predictors are lacking for predicting epitopes for HLA-C, HLA-DQ, and HLA-DP. However, Wang et al. have made a significant effort in peptide binding predictions for HLA DR, DP, and DQ molecules [56]. Their research with a large-scale datasets of over 17,000 HLA-peptide binding affinities for 11 HLA DP and DQ alleles found that prediction methodologies developed for HLA DR molecules perform equally well for DP and DQ molecules.

The generation of an MHC class-I epitope starts with the degradation of endogenous proteins into oligomeric fragments by cytosolic proteases, mainly the proteasome. These oligomeric fragments may escape from the attack of amino peptidases by entering the endoplasmic reticulum (ER) by the transporter associated with antigen presentation (TAP) [57]. The prediction algorithms for TAP binding and proteasomal cleavage have been developed [58, 59]. For

example, Peters et al. used a stabilized matrix method to predict TAP affinity of peptides [58]. This scoring method took advantage of the fact that binding of peptides to TAP is mainly determined by the C terminus and three N-terminal residues of a peptide. Predictions of the MHC class I pathway can be improved by predictions of proteasomal cleavage, TAP transport efficiency, and MHC class I binding affinity [58, 60–62].

While many successes have been made in the area of T-cell epitope prediction, the limitations of all these predictors should be noted. Our goal is to identify good vaccine targets that will induce productive immune responses. However, our ability to measure is usually done indirectly: peptide-binding assays, induction and measurement of immune responses *ex vivo*, use of animal models, and so forth. Only a small number of HLA-binding peptides are good targets. In clinical vaccine trials, wrong peptides were often selected using indirect methods and tested [63]. For example, the virulence and tumor maintenance capacity of high-risk Human Papillomavirus 16 (HPV-16) is mediated by two viral oncoproteins, E6 and E7. Of 21 E6 and E7 peptides computed to bind HLA-A*0201, 10 were confirmed through TAP-deficient T2 cell HLA stabilization assay. By testing their physical presence among peptides eluted from HPV-16-transformed epithelial tumor HLA-A*0201 immunoprecipitates, only one epitope (E7(11–19)) highly conserved among HPV-16 strains was detected. This 9-mer serves to direct cytolysis by T-cell lines. However, a related 10-mer (E7(11–20)), previously used as a vaccine candidate, was not detected by immune-precipitation or cytolysis assays. These data underscore the importance of precisely defining CTL epitopes on tumor cells and offer a paradigm for T-cell-based vaccine design [63].

2.4. B-Cell Epitope Mapping. It is important to clarify that limited immunoinformatics tools are currently available to identify B-cell antigens (recognized by antibodies). While humoral, or antibody-based, response represents the first line of defense against most viral and bacterial pathogens, the protein target of this arm of defense is usually too complex to model *in silico*. Antibodies that recognize B-cell epitopes, composed of either linear peptide sequences or conformational determinants, are present only in the three-dimensional form of the antigens. Several B-cell epitope prediction tools, including 3DEX, CEP, and Pepito, are at various stages in development and are in the process of being refined [64–67]. IEDB has collected a list of web prediction tools for B-cell epitope prediction (http://toolsl.immuneepitope.org/main/html/bcell_tools.htm). Unfortunately, the computational resources and modeling complexity required to predict B-cell epitopes are enormous. This complexity is due, in part, to the inherent flexibility in the complementarity-determining regions (CDR) of the antibody and, in part, attributable to posttranslational modifications such as glycosylation, all of which can result in modification of B-cell epitopes.

B-cell epitopes include linear and discontinuous epitopes. Linear epitopes comprise a single continuous stretch of amino acids within a protein sequence. An epitope whose residues are distantly separated in the sequence but have physical proximity through protein folding is named a discontinuous epitope. Although most epitopes are discontinuous [68], experimental epitope detection is primarily for linear epitopes. Tools for prediction of linear B-cell epitopes exist but in general are not predictive [69, 70]. The benchmarking B-cell epitope prediction by Blythe and Flower [69] found that with the best set of scales and parameters, amino acid propensity profiles can predict linear B-cell epitopes only marginally better than random. Such a conclusion has been confirmed by another study where the dismal performance of five predictors was tested against a set of reported linear B-cell epitopes [70].

Although devising accurate B-cell epitope mapping tools remains difficult, the selection of potent B-cell antigens can be accelerated using T-cell epitope mapping tools. When considering B-cell antigens as potential subunit vaccines, it also may be important to also consider their T-cell epitope content since the quality and kinetics of the antibody response is dependent upon the presence of T help. B-cell antigens that contain significant T help may outperform B-cell antigens lacking cognate help. In some cases, an identified T-cell epitope may also contain a B-cell epitope. Different epitopes activate T and B-cells. Despite this observation, it has been widely reported that B-cell epitopes may colocalize near, or overlap, Class II (Th, $CD4^+$) epitopes [71, 72].

2.5. Immunoinformatics-Based Vaccine Design Strategies. Different epitope-based vaccine design strategies exist, for example, mosaic vaccines [73], consensus [74, 75], centralized or ancestor immunogen [76, 77], or COT^+ [78]. Mosaic vaccines are comprised of "mosaic" proteins that are assembled from fragments of natural sequences via a computational optimization method [73]. Many immunogens, such as HIV envelope proteins, have high amino acid sequence divergences. To minimize the genetic differences between vaccine strains and contemporary isolates, immunogenic consensus sequences can be detected and used in vaccine design [74, 75]. Computer programs can also be developed to generate "centralized" vaccine that consists of consensus, ancestor, or center of the tree, modeled from phylogenetic trees. These "centralized" sequences can decrease the genetic distances between the "centralized" and wild-type gene immunogens [76, 77]. In an effort to develop antigens that capture both consensus and mutation sequences among strains, Nickle et al. reconstructed COT^+ antigens by including the ancestral state sequence at the center of phylogenic tree (COT) and extending the COT immunogen through addition of a composite sequence that includes high-frequency variable sites preserved in their native contexts [78]. These epitope-based vaccine designs have proven effective and provided vaccine researchers with different options in rational vaccine design. It is promising to combine various epitope methods to improve target discovery [56, 60, 79].

Integrated systems and workflows for computational vaccinology are likely to be key for automation of vaccine target discovery [80–82]. For example, Sollner et al. introduced the

pBone/pView computational workflow that supports design and execution of immunoinformatics workflow modules, results visualization, and knowledge sharing and reuse [80]. Pappalardo et al. developed ImmunoGrid, an integrative environment for large-scale simulation of the immune system for vaccine discovery, design, and optimization [81]. Feldhahn et al. developed FRED, an extendable, open source software framework for T-cell eptiope detection that integrates many prediction methods and supports implementation of custom-tailored prediction pipelines [82]. The effectiveness of these systems has been demonstrated with different applications.

The EpiVax vaccine design tools (EpiMatrix, ClustiMer, VaccineCAD, EpiAssembler, BlastiMer) are available to researchers through a portal at the Institute for Immunology and Informatics (the iVAX toolkit) [9]. The team of De Groot, Moise, and Martin have implemented the iVAX toolkit to develop four vaccines, a multiepitope TB vaccine [24], a cross-clade HIV vaccine [74], a prototype *H. pylori* vaccine [83], and a tularemia vaccine [84]. In collaboration with the TRIAD (Translational Immunology Research and Accelerated [vaccine] Development) program at the University of Rhode Island, iVAX is now being used to design additional vaccines including a multipathogen biodefense vaccine against Tularemia and *Burkholderia* spp, an epitope-based vaccine for HCV, and a vaccine derived from the deer tick saliva to prevent the acquisition of Tick-borne-pathogens. In addition, iVAX has recently been used to scan the entire genome of *Salmonella typhi* for vaccine candidates. This program is also accessible to researchers working on Neglected Tropical Diseases through the immunome website http://immunome.org/.

3. Reverse Vaccinology

3.1. Basic Principles of In Silico Antigen Prediction. Initially, when Reverse Vaccinology (RV) was developed, prediction of putative vaccine candidates was based solely on *in silico* analysis of the genome of a single strain. Now that selection criteria have been implemented, however, *in silico* analysis remains the central step in an RV project (see Figure 2).

The first step in the process of genome interpretation, usually referred as gene finding, consists in the prediction and localization of genes onto the chromosome. This is accomplished using prediction programs, which scan the sequence in search of regions that are likely to encode proteins. In prokaryotic systems, the identification of potential coding regions or open reading frames requires implementation of a few basic rules. In the simplest formulation, open reading frames (ORFs) are identified as segments of the same frame comprised between one of the three standard start codons (ATG, TTG, GTG) and one of the three standard stop codons (TAA, TAG, TGA). It is generally accepted that there is approximately one gene for every 1000 DNA base pairs. This suggests that significantly long start-to-stop segments are likely to encode for proteins.

Genome annotation procedures can be automated to different extents. Automated methods for prokaryotic gene finding such as GLIMMER [85], ORPHEUS [86], and GeneMark [87] have been used in genome sequencing projects [88–91]. GLIMMER uses interpolated Markov models, GeneMark uses hidden Markov models, and ORPHEUS is mainly based on codon usage and ribosome binding site statistics derived from annotated genes.

An exhaustive summary of software tools and websites that can be used to obtain bacterial genome annotations was presented by Stothard and Wishart [92].

The annotation procedure allows the translation of the bacterial genome sequence into a list of all the proteins that a bacterium virtually expresses at any time in its life cycle. Each of these amino acid sequences is then compared to the content of public databases of proteins or DNA sequences in an attempt to identify related sequences. When there exist obvious sequence similarities, it is reasonable to transfer this information on the filed sequence to the query. The functional annotation of a protein is sometimes sufficient for the selection of the protein as vaccine candidate, especially when the prediction of protein subcellular localization is uncertain, for example, a protein annotated as fibronectin binding protein may be a good vaccine candidate even when localization algorithms classify it as cytoplasmic. A critical aspect is represented by sequences that lack homologues or contain only remote homologues filed in the databases. ORFs having 20% or less of amino acid identity to any amino acid sequence found in the databases are generally considered to have unreliable homologues. These could represent novel uncharacterized proteins or random open reading frames misidentified as genes. Although homology searches can identify to a limited extent ORFs that are likely to encode functional proteins, experimental authentication by proteomic techniques is usually a more powerful approach for distinguishing genes from random ORFs.

The ensemble of hypothetical proteins can be processed with software programs dedicated to deduce their possible cellular localization. One of the basic assumptions utilized for candidate searches is that a good antigen will be located on the cell surface of a bacterium, where it is readily available for antibody recognition. Several algorithms have been developed that predict the subcellular localization of proteins based solely on the amino acid sequence and composition (see Table 2). The basic assumption made is that the N-terminal sequence of the protein predicts its cellular destination. The presence of a "leader sequence" provides evidence that the proteins will be exported to extra-cytoplasmic compartments. Additional signatures may also be exploited such as the presence of a cleavage site immediately after the leader peptide. Such sites imply that the protein is released into the extra-cellular environment of Gram-positive bacteria or into the periplasmic space of Gram-negatives. Similarly, proteins that contain an LXXC motif, where X is any amino acid, positioned at the end of the leader peptide are often lipoproteins. Anchoring of proteins to the Gram-positive bacterium cell wall often requires a specific carboxy-terminal sorting sequence. This sequence is identified by an LPXTG motif followed by approximately 20 hydrophobic amino acids and a charged tail. In Gram-negative bacteria, additional secretion pathways exist that promote the passage of extracellular proteins across the outer

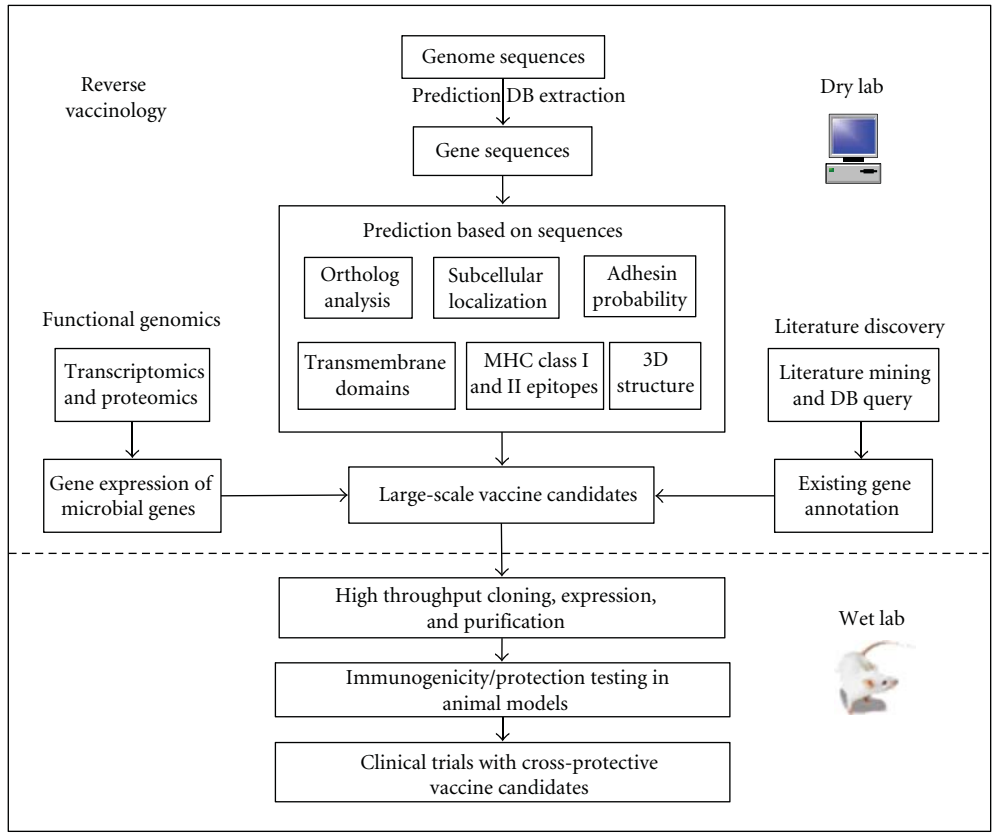


Figure 2: A schematic demonstration of integrative reverse vaccinology strategy towards vaccine development.

membrane. At least six distinct extracellular protein secretion systems have been reported in Gram-negative and Gram-positive bacterium (type I–VI, T1SS–T6SS) that can deliver proteins through the multilayered bacterial cell membrane and in some instances pass directly into the target host cell [93]. The six secretion systems exist in Gram-negative bacteria and the common Gram positive bacteria. Gram positive bacteria contain an additional specific secretion system (type VII) [94]. This increases the variety and complexity of secretion signals, making the identification of outer membrane and secreted proteins yet more challenging [95, 96].

Several computational methods have been generated to predict extracellular proteins in Gram-negative microorganisms [97].

PSORTb is the most widely used tool for predicting subcellular multiple localizations of organelles in Gram-negative bacteria. This program uses biological knowledge to elaborate "if-then" rules, combining information on amino acid composition, similarity to proteins of known subcellular localization, presence of signal peptides, transmembrane helices, and motifs diagnostics of specific subcellular localization. Recently, two predictive methods CELLO [98] and Proteome Analyst [99] have been proposed for Gram-negative bacteria. These programs are providing comparable performances in terms of accuracy and recall with respect to PSORTb [97].

Despite the recent progress, identification of secretion systems components *in silico* and their effectors still mainly relies on the detection of amino acid sequence [94] and the structural [100] similarities of selected proteins. Caution is necessary in applying these predictions, as sequence similarities can be very weak and do not necessarily imply any functional analogy.

In conclusion, by knowing the genome sequence it becomes possible to select using bioinformatics tools to generate a list of potential antigens without cultivating the microorganism. This methodology has a huge advantage over conventional vaccinology approaches for two major reasons. First of all, *in silico* analysis is very fast and cheap, and secondly, proteins not expressed *in vitro* are also identified. However, this approach only provides a prediction of a protein's subcellular localization and it cannot reveal if a protein is expressed and under what conditions. Therefore, use of a bioinformatics approach may need to be complemented with other techniques, for example, a Mass Spectrometry-based approach to aid vaccine candidate prediction. The first RV project employed a single genome. Indeed, at that time there was only one genome available for *N. meningitidis*. Nowadays in most cases, there are more than five genomes available for

Table 2: Tools used for reverse vaccinology.

Tool name	Website URL	Comment	Refs
ORF prediction and genome annotation			
GLIMMER	http://www.cbcb.umd.edu/software/glimmer/	Interpolated Markov models	[85]
ORPHEUS	http://pedant.gsf.de/orpheus/ (not working now)	Codon usage and ribosome binding statistics	[86]
GeneMark	http://exon.gatech.edu/GeneMark/	HMM	[87]
Bacterial protein localization prediction			
PSORTb	http://www.psort.org/psortb/	Multicomponent	[97]
Proteome Analyst	http://webdocs.cs.ualberta.ca/~bioinfo/PA/Sub/	Annotation keywords	[99]
SubLoc	http://www.bioinfo.tsinghua.edu.cn/SubLoc/	SVM	[257]
CELLO	http://cello.life.nctu.edu.tw/	SVM	[98]
PSLpred	http://www.imtech.res.in/raghava/pslpred/	SVM	[258]
LOCtree	http://cubic.bioc.columbia.edu/cgi/var/nair/loctree/query	SVM	[259]
SignalP	http://www.cbs.dtu.dk/services/SignalP/	NN, HMM	[260]
Sequence conservation			
BLAST	http://blast.ncbi.nlm.nih.gov/	Best reciprocal BLAST hit	[261]
OrthoMCL	http://www.orthomcl.org/cgi-bin/OrthoMclWeb.cgi	SVM	[262]
Transmembrane domain prediction			
HMMTOP	http://www.enzim.hu/hmmtop/	HMM	[263]
PRED-TMBB	http://biophysics.biol.uoa.gr/PRED-TMBB/	HMM (β-barrel)	[264]
TBBpred	http://www.imtech.res.in/raghava/tbbpred/	NN, SVM (β-barrel)	[265]
PROFtmb	http://cubic.bioc.columbia.edu/services/proftmb/	HMM	[266]
Bacterial adhesin prediction			
SPAAN	ftp://203.195.151.45	NN	[112]
Reverse vaccinology software			
NERVE	http://www.bio.unipd.it/molbinfo/ReverseVaccinology-NERVE.html	Multicomponent	[113]
Vaxign	http://www.violinet.org/vaxign/	Web-based, multicomponent	[110]

Abbreviations: HMM: hidden Markov model; NN: neuron network; SVM: support vector machine.

any human pathogen. Therefore, the *in silico* analysis can take advantage of comparative genomics.

3.2. Comparative Genomics and the Pangenome Concept. Today, the number of fully sequenced microbial genomes exceeds 1000 (http://www.ncbi.nlm.nih.gov/bioproject/) (Many are not from pathogens). It is clear that microbial diversity has been vastly underestimated, and a single genome does not exhaust the genomic diversity of any bacterial species [101, 102]. In many cases, an extensive genomic plasticity exists. For example, completion of the genome sequence of *E. coli* O157:H7 revealed that it contains >1,300 strain-specific genes compared to *E. coli* K12, which encode proteins that are involved in virulence and metabolic capabilities [103, 104]. Additional reports have revealed the occurrence of an extensive amount of genomic diversity among the strains of a single species [105–107].

These early findings were formalized with the definition of the bacterial pangenome, as the sum of the genes present in each individual species. This concept was originally introduced during study of the genome variability in eight isolates of *Streptococcus agalactiae* (also known as Group B *streptococcus* or GBS). It was found that each new genome had an average of 30 genes that were not present in any of the previously sequenced genomes. Not every bacterial species has the same level of complexity as GBS. For instance, the pangenome for *Bacillus anthracis* can be adequately described by four genome sequences. Hence, scientists refer to certain species as having an "open" and others a "closed" pangenome. In species with an open pangenome, there

are an unlimited number of new genes found for every genome. In closed pangenomes, there are only a limited number of strain-specific genes. The differences in the nature of the pangenome reflect several factors: differing lifestyles of two organisms, the number of closely related species in the same environment and physiological state, the ability of each species to acquire and stably incorporate foreign DNA (an advantage in niche adaptation from the acquisition of laterally transferred DNA), and the recent evolutionary history of each species. It should be noted that the imperfection of our definition of a bacterial species, for example, *B. anthracis* and *B. cereus,* can be considered the same species by some criteria, may render pangenome analysis more complicated. As the definition of a pangenome improves, the coverage of strains included in a bacterial species will change and thus alter the analysis results.

A pangenome can be divided into three elements: (1) a core genome that is shared by all strains (2) a set of dispensable genes that are shared by some but not all isolates, and (3) a set of strain-specific genes that are unique to each isolate. For *S. agalactiae*, the core genome encodes the basic aspects of *S. agalactiae* biology and was as such predicted to rapidly converge to 80% of the genome in each isolate. Conversely, dispensable and strain-specific genes, which are largely composed of hypothetical, phage-related and transposon-related genes [108], contribute to its genetic diversity. The concept of the pangenome and comparative genomics has practical applications in vaccine research. In fact, while obviously the ideal vaccine candidate is a conserved protein encoded by a gene present in every isolate of the species, in the case of GBS it was shown that the design of a universal protein-based vaccine against GBS was possible using dispensable genes [109]. Of note, capsular-specificity genes and other pathogenicity traits are often identified in an accessory genome. Moving forward, bacterial taxonomy and epidemiology must take into consideration whole genome sequences and not just a few genetic loci, as has been the case so far with methods such as ribosomal RNA sequences, capsular typing, and multilocus sequence typing (MLST). Comparison of the whole genome sequences of GBS strains has shown that the genomic diversity does not necessarily correlate with serotypes or MLST sequence-types. The application of additional whole genome sequence analysis will require that epidemiology studies have a reliable, systematic correlation between strains and disease and permit a standardization of the classification for clinical isolates. These observations are instrumental for developing a protective vaccine that covers a broad range of pathogenic strains.

Comparative genomics is also important for the identification of pathogenic factors since they potentially represent good vaccine candidates. The level of distinction and the function played by carrier versus virulent strains of *streptococci* and *neisseriae*, for example, has been the matter of discussion for a long time and still lacks an answer. There is, as yet, no clear and strict correlation between the presence of apparent virulence factors and the diseases caused by these organisms. The epidemiological evidence is vague and does not provide definitive clues. There may be multiple reasons for this apparent lack of correlation. It is likely that in species that only rarely result in disease, there exist multiple virulence factors and toxins that are uniquely associated with infection. Therefore, comparative genomics can be used to identify the "pathogenicity signature" associated with the most virulent bacterial strains or the strains that are successful in colonization. Comparative genomics can also be used to compare various strains that exhibit different virulence levels, for example, commensal nonpathogenic strains versus virulent ones, to find specific vaccine candidates [110]. The advantages include making a vaccine against commensal strains and narrowing down the pool of vaccine candidates. It is anticipated that most virulence factors will be found in accessory genomes, at least the ones that determine increased pathogenicity. However, presently comparative genomics is not able to identify expression variability that contributes to the different manifestations of pathogenicity of bacterial strains. Hence, functional studies are still critically needed to shed light on the relevance of specific virulence factors.

Another potential application could be studies of certain species of bacterial symbionts such as *Mycoplasma*, *Rickettsiae,* and *Chlamydiae*. These species, instead of acquiring genes during evolution, have actually lost significant levels of their genetic information [111]. Primarily biosynthetic pathway genes have been lost because intracellular bacteria have a relatively constant environment with access to much of what they require for survival. By applying the concept of pangenomics to these species, we would obtain a "microgenome" representative of the set of genes necessary to live in the intracellular niche. Comparing this "microgenome" with the pangenome for free-living species will likely simplify the identification of genes necessary for the microorganism to survive in varying and unfavorable environments. Housekeeping genes specific to the pathogen are considered vaccine candidates.

In comparison to a half decade ago, comparative genomics studies have become incredibly easy to perform. For the most important human pathogens, the average number of genomes for the different available strains is above five. Therefore, for new RV studies, the conservation level of selected antigens can be determined. Antigen conservation level is important since conserved antigens can be used to develop a broad strain protective vaccine [32].

In addition to the basic mechanisms of the RV strategy described above, additional criteria can be added. For example, since outer membrane proteins containing more than one transmembrane helix are difficult to clone and purify [10], the number of transmembrane domains of a candidate protein is often used as an additional filtering criterion. Bacterial adhesins play critical roles in adherence, colonization, and invasion of microbial pathogens to host cells [112]. Therefore, adhesins are essential for bacterial survival and are possible targets for vaccine development. Two RV software programs, NERVE [113] and Vaxign [110], utilize these criteria. Since RV focuses on predicting antigens using protein sequences, immune epitope prediction based on amino acid sequences can also be considered as a criterion for RV vaccine design [110].

Vaxign (http://www.violinet.org/vaxign/) is the first web-based vaccine design program based on genome sequences

utilizing the RV strategy. Predicted features in the Vaxign pipeline include protein subcellular location, transmembrane helices, adhesin probability, conservation to human and/or mouse proteins, sequence exclusion from genome(s) of nonpathogenic strain(s), and epitope binding to MHC class I and class II. Vaxign has been demonstrated to successfully predict vaccine targets for *Brucella* spp. [114, 115] and uropathogenic *Escherichia coli* [110]. Currently, more than 100 genomes have been precomputed using the Vaxign pipeline and available for query in the Vaxign website. Vaxign also performs dynamic vaccine target prediction based on input sequences.

The availability of three-dimensional structure may facilitate epitope prediction and antigen discovery [116, 117]. It would be ideal to also consider inclusion of analysis of high throughput transcriptomics and proteomics data to aide in complementary identification of vaccine candidates.

4. Transcriptomics and Proteomics Data Analysis for Vaccine R&D

Beside genomics methods in vaccine studies (described above), high-throughput transcriptomics and proteomics technologies (i.e., microarray) have been used for vaccine target design and analysis of vaccine-induced host immune responses. These assay systems are able to measure the expression pattern of thousands of genes in parallel, permitting the generation of large amounts of gene expression data. Bioinformatics techniques will play a critical role in analyzing such data and in making novel discoveries. In general, bioinformatics analysis of transcriptomics and proteomics data includes the following: (1) data preprocessing such as data quality controls and normalization, (2) statistical analysis of significantly regulated genes, (3) gene grouping and pattern discovery analyses, and (4) inference of biological pathways and networks [118, 119]. Depending on the specific research goals of any given project, different informatics tools may be applied individually or in combination.

Data processing is important in minimizing the effects of experimental artifacts and random noise. Companies that market microarrays usually provide their own methods for raw data processing and data quality control. For example, the GeneChip Operating Software (GCOS) expression analysis software provided by Affymetrix (Santa Clara, CA) can be used to process image data and the signals from the Affymetrix DNA microarrays [120]. The probe sets of Affymetrix microarray data are labeled present (P), absent (A), or marginal (M) based on the default P values set up in the GCOS system. Such labeling provides a useful approach for gene filtering. Commonly used microarray normalization methods include the Affymetrix MicroArray Suite MAS 5.0 (implemented in GCOS), the Robust Multichip Analysis (RMA) method [121], and the method of Li and Wong [122]. The software programs implementing these methods can be downloaded from the BioConductor (http://www.bioconductor.org/), a repository for open source and open development software programs developed specifically for the analysis and comprehension of omics data [123].

A common task in analyzing microarray data is to identify up- or down-regulated gene lists [124]. Fold changes of gene expression values between treatment group and nontreated controls were first used by biologists. However, this method may miss biologically important genes that exhibit small fold changes but have statistical significance. It also overemphasizes those genes with large fold changes but have little or no statistical significance [119]. Frequently used statistical methods for the determination of significantly changed genes include analysis of variance (ANOVA) [125], significance analysis of microarrays (SAM) [126], and the BioConductor package Linear Models for Microarray Data (LIMMA) [127]. ANOVA is a highly flexible analytical approach and is used in various commercial and open-source software packages [125]. SAM identifies genes with statistically significant expression changes by assimilating a set of gene-specific t-tests [126]. LIMMA uses linear models and empirical Bayesian methods to assess differential expression in microarray experiments [127].

Once the lists of up- or down-regulated genes are determined, they can be grouped into expression classes to identify patterns of gene expression and to provide greater insight into their biological functions and relevance. "Unsupervised and supervised" computational methods can be used for gene clustering analysis [128]. "Unsupervised" methods arrange genes and samples in groups or clusters based solely on the similarities in gene expression. Examples of unsupervised clustering methods include hierarchical clustering [129], self-organizing maps [12], and model-based clustering (e.g., CRCView [130]). "Supervised" methods, for example, EASE [131] and gene set enrichment analysis (GSEA) [132], use sample classifiers and gene expression to identify hypothesis-driven correlations. The Gene Ontology program (GO) is frequently used for gene enrichment analysis by many software programs, such as DAVID [133] and GOStat [134]. Additional GO-based microarray data analysis approaches can be found at http://www.geneontology.org/GO.tools.microarray.shtml.

The next level of DNA and protein array data analysis is the inference of biological pathways and networks [135, 136]. Several methods have been explored to model gene expression data including simple correlation [137], differential equations [138], neural networks [139], and Bayesian networks [140, 141]. These methods have different advantages and disadvantages [135, 136]. Simple correlation assumes linear and typically pairwise relationships. These limitations render it difficult for the investigator to identify multidimensional relationships between variables [142]. While methods utilizing differential equations are accurate, they are often "hand created" and as such are limited to the use of a small number of variables [142]. In contrast, neural networks make accurate predictions by mapping the data onto a high-dimensional polynomial. This allows the variables to influence each other in complex ways [139]. However, the use of neural networks assumes that everything is affected by the changing variable. This renders it difficult to identify such mechanisms. Bayesian networks (BN) represent a powerful method for identifying causal or apparently causal patterns in gene expression data. A key advantage

of Bayesian networks is that they are relatively agnostic to the complexity of the relationships predicted and can model linear, nonlinear, combinatorial, stochastic, and other types of relationships among variables across multiple levels of biological organizations [143]. However, current Bayesian network approaches are also subject to limitations. For example, the expression levels must be discretized, leading to varying degrees of loss of information [135].

The combined application of transcriptomics and proteomics experiments in conjugation with specialized informatics analyses has many applications in the field of vaccine research and development. First, these "omics" methods can be used to discover vaccine targets for many microorganism-induced diseases as well as cancers [144, 145]. For example, the sexual stages of malarial parasites are essential for transmission of the disease by the mosquito and as such are the targets for malaria vaccine development. To better understand how genes participate in the sexual development process, Young et al. utilized microarrays to profile the transcriptomes of high-purity stage I-V *Plasmodium falciparum* gametocytes [146]. An ontology-based pattern identification algorithm was applied to identify a 246 gene sexual development cluster. Some of the genes have the potential of being used for vaccine development. Sturniolo et al. [147] developed a matrix-based computational algorithm when applied to DNA microarray experiments all data was used successfully to predict human leukocyte antigen (HLA) class II ligands and differentially expressed colon cancer genes. A list of peptides uniquely associated with colon cancer was identified. These are potentially immunogenic. These peptides provide a basis for rational vaccine development against colon cancer.

One practical problem in vaccine investigation is that for most diseases, no immune response correlates well with protection. To solve this issue, systems biology (Omics and bioinformatics) approaches have also been used to detect gene signatures induced in vaccinated hosts (e.g., humans) that correlate and even predict protective immunity. For example, two recently published studies examined early gene signatures induced in humans vaccinated with the attenuated yellow fever vaccine YF17D [148, 149]. Each study analyzed total peripheral-blood mononuclear cells from different cohorts of human volunteers at various time points following vaccination with YF17D. Early effects (3 and 7 days postvaccination) on gene expression were determined using microarrays and were analyzed using bioinformatics approaches. Many genes involved in innate immune response (e.g., Toll-like receptor signaling and inflammasome) were discovered. Gaucher et al. [149] identified a group of transcription factors, including interferon-regulatory factor 7 (IRF7), signal transducer and activator of transcription 1 (STAT2), and ETS2, as key regulators of the early immune response to the YF17D vaccine [149]. YF17D was found to trigger the proliferation of several leukocyte subtypes including macrophages, dendritic cells, natural killer cells, and lymphocytes [149]. Definition of this "baseline" innate immunity response subsequently allowed detection of defective hyperresponse (excessive CCR5 activation) in a YF17D vaccinee who had developed a serious viscerotopic adverse event [150]. In another study, Querec et al. [148] discovered gene signatures that correlate with the magnitude of antigen-specific $CD8^+$ T-cell responses and antibody titers [148]. EIF2AK4, a key gene in the integrated stress response, was found among most of the predictive signatures. The actual predictive capacity of a gene signature was verified using the signatures for $CD8^+$ T-cell responses from the first trial to predict the outcome of the second trial and vice versa. Another distinct early gene signature that included TNFRSF17 (a receptor for B-cell-activating factor) was found to predict the neutralizing antibody titers as late as 90 days following vaccination [148].

Microarray-based methods have also been used to investigate vaccine safety [151]. For example, McKinney et al. used protein microarrays to compare 108 serum cytokines and chemokines in vaccine recipients before and one week after smallpox vaccination [151]. Among 74 individuals studied, 22 experienced systemic adverse events. Machine-learning and statistical analyses identified six cytokines that accurately discriminate between individuals on the basis of their adverse event status. A DNA microarray-based system has also been developed to evaluate the genetic signatures of the toxicity of many vaccines including pertussis vaccine [152] and influenza vaccines [153].

5. Mathematical Simulations for Vaccine R&D

Integrative research, development, and uses of vaccines follow a cyclical fashion where mathematical and computational simulations are connected with experimentation leading to improved accuracy and reduced cost in vaccine R&D [154]. Many mathematic simulations have been developed to support different areas of vaccine research and development (R&D). These studies support various vaccine-associated aspects including vaccine discovery and development, vaccine production and stockpiling, vaccination protocol optimization, vaccine distribution, and vaccine regulation. Here we introduce some striking examples.

Mathematical models have been developed to study the dynamics of host-pathogen and host-vaccine interactions [155]. For example, Kirschner et al. integrate information over relevant biological and temporal scales to generate a model for major histocompatibility complex class II-mediated antigen presentation [156]. This multiscale mathematical model simulates molecular, cellular, tissue, and organ/organism, and the interactions between different levels. This model has been used to answer questions about mechanisms of infection and new strategies for treatment and vaccines. The same group has developed a multifaceted approach to modeling tuberculosis-induced granuloma, a self-organizing structure of immune cells forming in the lung and lymph nodes in response to bacterial invasion [157–159]. Many mathematical models have been developed to understand the mechanisms and limitations of HIV control by humoral and cell-mediated immunity [160]. These studies suggest that $CD8^+$ T-cells do "too little too late" to prevent the establishment of HIV infection. However, passively administered antibody acts very early to reduce the initial viral count and slow HIV growth [160]. Cell culture-based

influenza vaccine manufacturing is of growing importance. Influenza virus is able to replicate and induce apoptosis in host cells. Combined with experiments, Schulze-Horsel et al. have formulated a mathematical model to describe changes in the concentration of uninfected and influenza A virus-infected adherent cells, dynamics of virus particle release, and the time course of the percentage composition of the cell population [161]. This model can be used to characterize and maximize viral titer yield in the bioreactors meant to produce virus for use in influenza vaccines.

Cost-effectiveness analyses (CEA) of vaccination programs can be performed using mathematical modeling [162]. For effective evaluation of cost effectiveness, a model is generally required which considers the relevant biological, clinical, epidemiological, and economic factors of a vaccination program. CEA modeling methods have been categorized based on three main attributes: static/dynamic, stochastic/deterministic, and aggregate/individual based. The modeling methods for CEAs of vaccination programs can be improved in the areas of model choice, construction, assessment, and validation [162]. CFA has been applied to study different vaccination programs such as human papillomavirus (HPV) vaccination [163], influenza vaccination [164], and vacation with pneumococcal conjugate vaccine [165].

Mathematical modeling can be used to simulate and optimize vaccination protocols. The combination of *in silico* and *in vivo* studies has the ability to reduce the time, effort, and cost of vaccine studies by orders of magnitude [166, 167]. For example, Pappalardo et al. designed and implemented SimTriplex, an agent-based model specifically tailored to simulate the effects of tumor-preventive cell vaccines in HER-2/neu transgenic mice prone to mammary carcinoma development [168]. The SimTriplex mathematical model combined with genetic algorithm has been used to search for new vaccination schedules to prevent tumors in HER-2/neu transgenic mice [166, 169, 170]. It has been found that the computational model can be used for simulation of immune responses ("*in silico*" experiments), leading to optimization of vaccine protocols. Pennisi et al. also developed MetastaSim, a hybrid Agent Based-ODE model for the simulation of the Triplex cell vaccine-elicited immune system response against lung metastases in mice [167]. MetastaSim simulates the main features of the immune system. Both innate and adaptive immune responses are covered. This model includes different cell types and molecules, such as dendritic cells, macrophages, cytotoxic lymphocytes, antibodies, antigens, IL-12, and IFN-γ. Their study with MetastaSim demonstrated that it is possible to obtain *in silico* a 45% reduction in the number of vaccinations [167].

Mathematical modeling plays an important role in postlicensure vaccine informatics and in assessing the impact of immunizations against target diseases. For example, Blower et al. developed a mathematic model to predict the tradeoff between efficacy and safety of live attenuated HIV vaccines [171]. More details in this topic are introduced in the following section.

6. Postlicensure Vaccine Informatics

Successful vaccine immunization induces protective immunity in the individual. Equally important for most infectious diseases, when a sufficiently high threshold of a group of individuals is immunized, a "herd effect" is observed at the population level where the incidence of the disease in the remaining unimmunized members of the group is lower than it would be otherwise [172]. Due to the large societal benefits of immunizations, almost all governments (generally at the state/provincial or national levels) organize formal targeted immunization programs to maximize vaccine coverage. The impact of the immunization programs is to reduce the incidence of the targeted disease. For some infectious diseases where the characteristics permit [173] (e.g., smallpox, polio, measles, neonatal tetanus), regional or global initiatives to eliminate or eradicate the targeted disease may be organized. Routine or special immunization programs are incredibly complex to initiate. Ongoing endeavors not uncommonly require careful orchestration and planning for sustained and repeated immunizations of millions of persons annually in most jurisdictions. Vaccine informatics is critical to providing accurate data and facilitates the smooth planning, organization, implementation, and monitoring of almost every aspect of such complex immunization programs. The introduction of each new recommended vaccine into an already crowded pediatric immunization schedule adds to this complexity [174]. We describe next some of the better known postlicensure vaccine informatic systems: tracking immunization history in computerized immunization information systems (IIS) or registries, informatics methods for improving surveillance of vaccine safety and efficacy, and modeling impact of alternative immunization strategies against target diseases.

6.1. Computerized Immunization Information Systems (or Immunization Registries). Accurate tracking of vaccination history is essential to ensure proper completion of the primary immunization schedule and subsequent booster doses. This seemingly straightforward task is nontrivial system-wide when compounded by an increasingly mobile population, immunization schedules of increasing complexity, multiple vaccine manufacturers of the same vaccine, multiple health care providers and/or health insurance for the same individual (a problem in the U.S.), multiple individual with same name, and so forth. Add in small vaccine vials with hard to read small fonts in a busy pediatric clinic serving many crying babies simultaneously, the opportunities for inaccurate or nonrecording of an administered vaccination is substantial in developed and developing countries.

Computerized immunization information systems (IIS) provide an obvious potential solution to these challenges. In the U.S., the first large IISs were organized in Delaware in the early 1970s [175]. This action was followed by several health maintenance organization (HMOs) with the dual purpose of linking the IIS to medical visits for rigorous studies of vaccine safety [176]. The Robert Wood Johnson Foundation funded the All Kids Count I and II programs in the 1990s in multiple communities. This provided an important

impetus to the field [175]. The Centers for Disease Control and Prevention (CDC) now provide some financial and technical assistance for public sector IIS in almost every state (http://www.cdc.gov/vaccines/programs/iis/default.htm). This work is aided by partners such as the American Immunization Registry Association (http://immregistries.org/) and the Public Health Informatics Institute (http://phii.org/). Internationally, Australia [177], Canada [178], and Norway [179] are some of the other countries with active IIS.

IISs also have the potential to provide a foundation of child health registries [175] and electronic health records [180]. While initially focused on routine pediatric immunizations, registries in many locations have been expanded to meet other needs, including adolescent and adult immunizations [181], disasters [182], targeting of at risk populations [182, 183], study vaccine refusal [184], and facilitating accurate and timely reporting of vaccine adverse events [185]. Substantial progress has also been attained in the protection of privacy and confidentiality; in ensuring participation of all immunization providers and recipients, to ensure appropriate functioning of registries, and to ensure sustainable funding for registries [186]. However, challenges remain in exchanging information among different IISs, and across state lines. The National Vaccine Advisory Committee has issued recommendations on how to overcome these challenges (http://www.hhs.gov/nvpo/nvac/IISRecolmmendationsSep08.htm).

6.2. Informatics Methods for Improving Surveillance of Vaccine Safety and Efficacy. Before a vaccine is licensed, it undergoes rigorous testing in preclinical (laboratory and animal) and phased human clinical trials for safety and efficacy [187]. Due mainly to cost (intensive monitoring perprotocol is expensive) and ethical (once a vaccine is determined to be safe and effective, it is no longer ethical to withhold it from others in need) considerations, however, the sample size and duration of followup in prelicensure trials are usually limited. This means surveillance for both vaccine safety and effectiveness [188] in larger immunized population postlicensure and postmarketing is needed. This is challenging because trial conditions (e.g., double-blinding, randomization) that permit straightforward comparison between vaccinated and unvaccinated groups no longer hold. Substantial data collection and adjustments on possible confounders, when possible, are needed to fully analyze and interpret such observational studies.

Post-licensure monitoring for vaccine safety can generally be divided into hypothesis generating and hypothesis testing. Since vaccine coverage for many vaccines can be close to universal, by definition, anyone with a medical adverse event will have previously been vaccinated. Spontaneous reporting or passive surveillance systems like the U.S. Vaccine Adverse Event Reporting System (VAERS, http://vaers.hhs.gov/) in the U.S. [189] and elsewhere [190], where medical problems suspected to be caused by the vaccination can be reported to health authorities, provide the bulk of new vaccine safety hypotheses. Due to the large number of reports (>20,000 annually to VAERS), data mining techniques are beginning to be applied to triage reports worthy of further attention [191].

Once a vaccine safety concern is provisionally identified, based on our understanding of likely pathophysiology and nonrandom clustering of cases in onset time after vaccination, a formal study is usually needed to (1) confirm whether the etiologic link with vaccination is real and not coincidental, and (2) identify the magnitude of the risk (to assist in risk-benefit determination for the immunization). Since these safety concerns are likely to be rare (otherwise they would have been detected pre-licensure), confirmatory pharmacoepidemiologic studies of large vaccinated populations are usually needed to "test the hypothesis". Large national (e.g., Denmark) or population (e.g., Managed Care Organization (MCO)) health care systems, where members have unique personal identifiers and most of the care for both vaccinations (exposure) and medical visits (outcome) are automated, provide an efficient platform for piggy-backing vaccine safety pharmacoepidemiologic studies [192]. The Vaccine Safety Datalink (VSD) project in the US, a consortium of 8 MCO's representing ~3% of the population, has been used as a prototype of how such large linked databases can be used for rigorous vaccine safety studies [176, 193]. Examples include rotavirus vaccine and intussusception [194], thimerosal and neurologic adverse events [195], and vaccinations and central demyelination [196]. Similar large-linked vaccine safety databases have been created in England [197] and Vietnam [198].

Safety issues cannot be assessed directly and can only be inferred from the relative absence of multiple adverse events. Therefore, standardizing the case definitions used to assess adverse events is needed to allow for meaningful comparison of vaccine safety data in various settings. Recognizing this need, the Brighton Collaboration (https://brightoncollaboration.org/public) was formed in 1999 as a voluntary global collaboration to facilitate the development, evaluation, and dissemination of high quality information about the safety of human vaccines in both pre- and post-licensure settings. To date, 28 guidelines and case definitions have been developed and are freely available to users. The case definitions are tiered by the level of evidence available and will differ based on whether the data are gathered in prospective clinical trials or passive postmarketing surveillance and on the level of resource availability (e.g., developed versus developing countries). Since its inception, the Collaboration has helped to form a critical mass of experts interested in vaccine safety that can potentially be convened or accessed as new vaccine safety issues arise. The Brighton Collaboration Viral Vector Vaccine Safety Working Group is exploring using the "wiki" model of mass collaboration for completing and maintaining standard templates on characteristics of various viral vectors [199].

Post-licensure monitoring for vaccine effectiveness is usually done by examining the impact on targeted diseases. For reasonable sensitivity and specificity for monitoring trends of the disease, this usually requires the establishment of some type of public health surveillance system. For example, the recent reintroduction of rotavirus vaccine in the US has resulted in delayed onset and diminished magnitude of

rotavirus activity [200]. A decline in invasive pneumococcal disease was observed after the introduction of conjugate pneumococcal vaccine [201]. Similar data was obtained in developing countries, as was done with introduction of conjugate *Haemophilus influenza* type b vaccine in Mali [202]. When disease remains high or an outbreak occurs despite high vaccine coverage, a special epidemiologic study to assess vaccine effectiveness may be needed. This type of action was undertaken after a posthoneymoon period measles outbreak in Burundi [203], the resurgence of diphtheria in the former Soviet Union [204], and the introduction of a monovalent oral type 1 poliovirus vaccine in India [205].

6.3. Modeling of Impact of Immunizations against Target Diseases. The cyclical nature of epidemics of many infectious diseases such as plague and smallpox in humans (and other animals) was noted by ancient historians prior to the introduction of immunization in modern times [206]. This periodicity was described as a mathematical relationship between susceptible and immune individuals in a population over time that interacted with an external infectious force by Ronald Ross and Anderson Gray McKendrick at the beginning of the 20th Century [207]. It was not until the early 1980's, however, that Anderson and May systematically organized and effectively organized disparate works in population biology, ecology, and epidemiology into mathematical models of infectious diseases that linked the theory with practical translation into public health policy (e.g., vaccinations) [208]. Their 1991 textbook "Infectious Diseases of Humans: Dynamics and Control" [209] has helped to create a cohort of mathematical modelers that have furthered our understanding of transmission of infectious agents within human communities and design programs for their control. As Geographic Information Systems (GISs) that integrate and analyze spatial information become increasingly available for linkage with public health databases [210], this should aid continued refinements in various assumptions used in mathematical models of infectious diseases.

Irrespective of the model or the target disease, a key concept in any mathematical model is the basic reproductive rate or R_0 of a microorganism—the average number of secondary infections produced when one infected individual is introduced into a totally susceptible population. The goal of any control program (e.g., immunizations) is to reduce the R_0 as much as possible. For disease elimination or eradication programs, it must be <1 [211]. Another key concept is "herd immunity", the indirect effect of some vaccines on reduction of disease transmission beyond the protection in actual vaccine recipients [172]. Most mathematical models attempt to describe as accurately as possible the flow of a human population from susceptibility (usually at birth or with the waning of maternally derived immunity), infected (by wild disease or vaccination), and immune states (adjusting for various variables such as mixing) transmission coefficient, vaccine effectiveness, and duration of protection. Each of these variables in turn can be further modeled (e.g., mixing can differ with age-classes or other subpopulations).

Historically, one of the more successful integrations of mathematical modeling of vaccine-preventable diseases and immunization program policies has been for measles [203, 211, 212] and rubella [213]. Mathematical models have also been critical for understanding how best to (a) introduce newly licensed vaccines like human papillomavirus vaccine [214], (b) control new emerging public health problems, such as pandemic influenza [215], (c) how best to optimize control of a vaccine-preventable disease (such as impact of pneumococcal conjugate vaccine on emergence of penicillin-resistant strains) [216], or (d) how spatio-temporal variation in birth rates may explain the observed patterns of rotavirus disease after the introduction of new rotavirus vaccine [200].

7. Vaccine Literature Mining, Databases, and Data Integration

Vaccine informatics is dedicated to the acquisition, processing, storage, distribution, analysis, and interpretation of vaccine-associated data by means of computing methods and tools. Advanced DNA sequencing, molecular, cellular, and immunological methods have provided a huge amount of vaccine-related data. These data have been processed and analyzed by exponentially expanded computational power and new algorithms. To facilitate advanced vaccine research and development, the large amounts of vaccine literature data need to be processed and mined. Different types of vaccine databases are also needed to store various vaccine data. Eventually, all these data need to be integrated within the vaccine domain and with other biomedical data for computational reasoning and discovery of new knowledge.

7.1. Vaccine Literature Mining. The papers and authors related to vaccine/vaccination have increased exponentially. Only six vaccine-related papers were published and recorded in PubMed before 1900. In the first half of the 20th century, 1,210 vaccine-related papers were published. This number has increased almost 100-fold in the second half of the last century. In addition, the numbers of vaccine publications have increased exponentially (Figure 3). For example, 6,399 vaccine-related papers were published during the period of 1951–1960, and 96,938 in 2001–2010. Therefore, the number of papers published annually in PubMed has increased more than 15-fold during the past 50 years.

It has become increasingly challenging to retrieve useful vaccine data for research purposes from the huge amount of vaccine literature. Literature mining has been used to facilitate the discovery and analysis of potential vaccine targets. For example, cross-matching and analysis of the literature and *in silico*-derived data allowed the selection of 189 putative vaccine candidates from the entire *Mycobacterium tuberculosis* genome [217]. In this study, the first step towards the selection of vaccine candidates was to accumulate published experimental data from a literature scan of documented studies with a focus on global analyses. The literature sources were then grouped based on different categories (e.g., macrophage). This literature mining approach detected 189 potential vaccine candidates. These were studied further through *in silico* functional analysis and immunoinformatics epitope prediction. A qualitative score

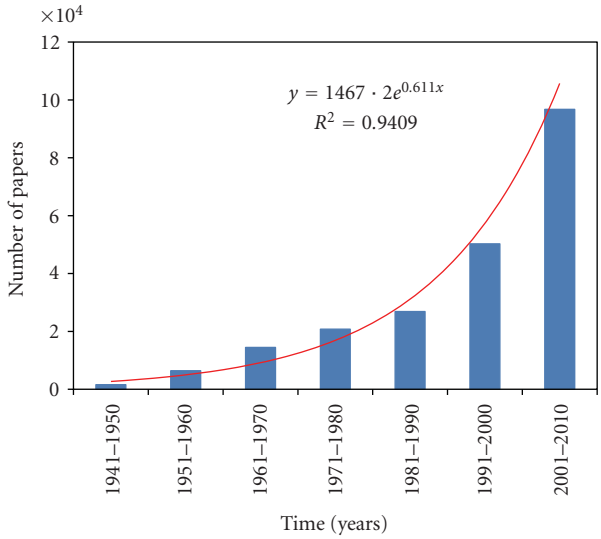


Figure 3: Exponential growth of papers in the area of vaccine and vaccination. The data was obtained by analysis of available papers in PubMed.

was designed based on a total of 14 criteria and used to rank and prioritize the gene list [217].

Literature mining can also be used to analyze vaccine-associated host immune response networks. For example, Ozgur et al. recently applied a literature mining and network centrality analysis [218] to analyze the IFN-γ and vaccine-associated gene networks [219]. Among approximately 1,000 genes found to interact with IFN-γ, 102 genes were predicted to be vaccine-associated and 52 of them were verified by manual curation. The production of IFN-γ is crucial for successful immune response induced by vaccines against various viruses and intracellular bacteria. For example, these include HIV [220], *M. tuberculosis* [221], *Leishmania* spp. [222], and *Brucella* spp. [223]. The discovery of the IFN-γ and vaccine-mediated gene network provides a comprehensive view of the vaccine-induced protective immune network and generates new hypotheses for further experimental testing.

Two literature mining programs presented in the Vaccine Investigation and Online Information Network (VIOLIN; see next section) were developed for general vaccine literature searching and analysis [13]. Vaxpresso (http://www.violinet.org/textpresso/cgi-bin/home) is a vaccine literature mining program using natural language processing (NLP) and ontology-based literature searching [224]. For a list of selected pathogens, Vaxpresso contains all possible vaccine-related papers extracted from PubMed (http://www.ncbi.nlm.nih.gov/pubmed). Vaxpresso is able to retrieve and sort article sentences that match specific keywords and ontology-based categories. Vaxmesh (http://www.violinet.org/litesearch/meshtree/meshtree.php) is a vaccine literature browser based on the Medical Subject Headings (MeSH). MeSH is a controlled vocabulary of medical and scientific terms that is used for indexing PubMed articles in a consistent way supporting PubMed literature mining. Vaxmesh enables users to locate articles using MeSH terms in a hierarchical MeSH tree structure.

7.2. Web-Based Vaccine Databases and Online Resources. Many publicly available vaccine databases and online resources exist (Table 3). For example, the USA CDC Vaccine Information Statements (VISs) system (http://www.cdc.gov/vaccines/pubs/vis/) provides information sheets that explain to vaccine recipients, their parents, or their legal representatives both the benefits and risks of a vaccine. Federal law in the US requires that VISs be handed out for all vaccines before their use. The licensed vaccine information is provided by the U.S. FDA (http://www.fda.gov/BiologicsBloodVaccines/Vaccines/default.htm). The Vaccine Resource Library (VRL, http://www.path.org/vaccineresources/) offers various high quality, scientifically accurate documents and links to specific diseases and topics in immunization.

These databases focus primarily on the clinical uses and regulations of existing vaccines for vaccine users. To store and analyze research data concerning commercial vaccines and vaccines under clinical trials, or in early stages of development, the Vaccine Investigation and Online Information Network (VIOLIN, http://www.violinet.org/) was developed. VIOLIN is a web-based vaccine database and analysis system primarily targeted for vaccine researchers [13]. The VIOLIN vaccine database currently contains more than 2,700 vaccines, or vaccine candidates, for more than 160 pathogens through manual curation from >1500 peer-reviewed papers or other reliable sources. The stored vaccine data includes vaccine preparation, pathogen genes used and gene engineering, vaccine adjuvants and vectors, vaccine-induced host immune responses, and vaccine efficacy in host after virulent challenge. VIOLIN curates more than 500 protective antigens (http://www.violinet.org/protegen/) [225]. Vaccine-related pathogen and host genes are also annotated and available for searching through customized BLAST programs. VIOLIN also stores and processes all possible vaccine literature through different text mining programs [13]. Vaxign, a web-based vaccine design program based on reverse vaccinology strategy [110], is also a program in VIOLIN.

Besides the above databases which focused on vaccine awareness and vaccine research, many other databases are available that are useful for vaccine research and development. For example, more than 65,000 antibody and T-cell epitopes have been deposited in the Immune Epitope Database and Analysis Resource (http://www.immuneepitope.org/) since the database was established in 2004 [6]. These immune epitopes cover a broad range of species including humans, nonhuman primates, rodents, and other animal species as related to all infectious diseases [6]. AntigenDB is an immunoinformatics database of pathogen antigens and store sequences, structures, origins, and epitopes [226].

7.3. Development of a Community-Based Vaccine Ontology (VO). Although public vaccine databases provide help with different aspects of vaccine knowledge and research, it remains a challenge to integrate this disparate body of information on vaccines. Data integration is hampered since the data are often collected using incompatible or poorly described methods for data capture, storage, and

Table 3: Vaccine web resources.

Resource name	Website URL	Comment
Vaccines and Immunization		
WHO Immunization/vaccines	http://www.who.int/immunization/en/	WHO vaccine site
Immunization Action Coalition	http://www.immunize.org/	Vaccination Information for Healthcare Professionals
USA CDC Vaccine Information Statements	http://www.cdc.gov/vaccines/pubs/vis/	Benefits and risks of vaccines
USA CDC listed vaccines	http://www.cdc.gov/vaccines/vpd-vac/vaccines-list.htm	Vaccines used in USA
US FDA licensed vaccine information	http://www.fda.gov/BiologicsBlood-Vaccines/Vaccines/default.htm	Licensed vaccines used in USA
Canada licensed vaccine information	http://www.phac-aspc.gc.ca/dpg-eng.php#vaccines	Canada
DH Immunization at UK	http://www.dh.gov.uk/en/Publichealth/Immunisation/index.htm	Official UK vaccination site
PATH Vaccine Resource Library	http://www.path.org/vaccineresources/	Collection of vaccine resources
NNii: National Network for Immunization Information	http://www.immunizationinfo.org/vaccines	Scientific valid information
GAVI Alliance (Global Alliance for Vaccines and Immunisation)	http://www.gavialliance.org/	Goal: save children's lives
Vaccine Clinical Trials		
Nonhuman primate HIV/SIV vaccine trials database	http://www.hiv.lanl.gov/content/vaccine/home.html	Vaccine studies of HIV/SIV using nonhuman primates
AIDS vaccine trials database	http://www.iavireport.org/trials-db/Pages/default.aspx	AIDS vaccine clinical trials
Clinical trials database (USA NIH)	http://clinicaltrials.gov/	Vaccine or other trials
Vaccine Safety		
WHO Immunization Safety	http://www.who.int/immunization_safety/en/	WHO vaccination safety site
VAERS: Vaccine Adverse Event Reporting System (USA FDA & CDC)	http://vaers.hhs.gov/index	USA vaccine adverse event reporting website
USA CDC-Vaccine Safety	http://www.cdc.gov/vaccinesafety/index.html	USA official vaccine safety site
The Brighton Collaboration	https://brightoncollaboration.org/public	Setting standards in vaccine safety
Vaccine Research Database		
VIOLIN	http://www.violinet.org/	Comprehensive vaccine data
Vaccine Manufacturers		
EVM: European Vaccine Manufacturers	http://www.efpia.org/Content/Default.asp	Collection of European vaccine companies
List of vaccine manufactures	http://www.ontobee.org/browser/rdf.php?o=VO&iri=http://purl.obolibrary.org/obo/VO_0000299	Collected in Vaccine Ontology

dissemination. Integration is also complicated as investigators use independently derived local terminologies and data schemas. These problems can be alleviated through the use of a common ontology, that is, a consensus-based controlled vocabulary of terms and relations, with associated definitions that are logically formulated in such a way as to promote automated reasoning. Ontologies are able to structure complex biomedical domains and relate a myriad of data to allow for a shared understanding of vaccines.

The collaborative, community-based Vaccine Ontology (VO; http://www.violinet.org/vaccineontology/) was recently initiated to promote vaccine data standardization, integration, and computer-assisted reasoning. VO can be used for different applications, including vaccine data integration and

literature mining. Currently, VO contains more than 3,000 terms, including more than 700 vaccines and vaccine candidates that are represented in an appropriately structured ontological hierarchy. These vaccines or vaccine candidates are targeted to 70 pathogens and have been studied in more than 20 animal species (e.g., human, mouse, cattle, and fish). VO also stores terms related to different vaccine components (e.g., protective antigens, vaccine adjuvants and vectors), vaccine-induced immune responses, vaccine adverse events, and protection efficacy. The known relations between these terms are also listed. These representations are readable by computer programs and support computer-assisted reasoning. This knowledge is also exchangeable across multiple scientific domains to facilitate hypothesis generation and validation. This approach will undoubtedly lead to new scientific discoveries.

VO has been used for several different applications. For example, VO, in combination with other ontologies, has been used to model and study vaccine protection investigation [15]. Reported vaccine protection data from different reports can be systematically analyzed [227]. VO can also be used to improve vaccine literature mining. For example, a direct PubMed search for "live attenuated *Brucella* vaccine" returned 69 papers (as of August 2010). VO includes 13 live attenuated *Brucella* vaccines that are defined as "live" and "attenuated". When specific "live, attenuated" *Brucella* vaccine terms are included in a PubMed search, the number of papers found in PubMed increased by more than 10-fold [228, 229]. The application of VO has also enhanced the discovery of IFN-γ and vaccine-associated gene networks [16].

8. Discussion

In summary, vaccine informatics has been widely implemented in the areas of basic vaccine research, translational vaccine development, prolicensure vaccine immunization registry and surveillance, and vaccine data mining and integration.

Vaccine informatics is an emerging interdisciplinary research, with close relationships to several similar research fields. Vaccine informatics overlaps with immunological bioinformatics (or immunoinformatics). The latter field applies informatics technologies to investigate the immune system at a systems biology level [5, 230]. Vaccine informatics emphasizes understanding of vaccine-induced immunity. Vaccine informatics uses information of OMICS (genomics, transcriptomics, proteomics, and metabolomics). This is in contrast to reverse vaccinology that primarily uses genomics, that is, informatics analysis of genome sequences. Other OMICS technologies may also have the potential to aid in rational vaccine design. Recently Poland et al. defined a new area of vaccinomics that will focus on the development of personalized vaccines based on our increasing understanding of genotype information [231]. Vaccine informatics is also closely associated with clinical immunology in the areas of post-licensure vaccine assessment and surveillance. Mathematical modeling also plays an important role in vaccine informatics by modeling various aspects of pre- and post-licensure vaccine research and clinical investigations.

Vaccine informatics still faces many challenges. Many infectious diseases, including HIV/AIDS, tuberculosis, and malaria, still lack effective and safe vaccines. Although extensive progress has been made towards the genetic structure and pathogenesis of HIV and other infectious pathogens, significant gaps in our understanding of host-pathogen interactions still remain [232, 233]. These gaps are attributable to imperfect and nonstandardized animal models, the absence of precise immunological correlates of protection, and the prohibitive cost of confirmatory clinical trials. The development of vaccines against many noninfectious diseases including cancer, autoimmune diseases, and allergy remains a challenge. While many vaccine adverse events are likely genetically determined (and thus predictable), it remains challenging to predict possible vaccine adverse events with available genotype data and possibly design personalized vaccine. These challenges will undoubtedly be met with improved rational vaccine design and a better understanding of fundamental protective immunity mechanisms obtained with improving vaccine informatics technologies.

New bioinformatics technologies are constantly being devised and applied to address various vaccine-related questions using high throughput sequencing, gene expression data, and experimental results from experimental and clinical studies. Efforts during the 21st century vaccinology will witness more successes of application of vaccine informatics in vaccine research.

Acknowledgments

This work has been supported by grant R01AI081062 from the National Institute of Allergy and Infectious Diseases USA. The authors wish to acknowledge the assistance of John Glasser and Gary Urquhart as reviewers for the sections on mathematical modeling and immunization information systems, respectively. Editorial review by Dr. George W. Jourdian is also appreciated. The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of the Centers for Disease Control and Prevention.

References

[1] E. Jenner, *An Inquiry into the Causes and Effects of the Variolae Vaccinae*, Low, London, UK, 1798.

[2] M. R. Hilleman, "Vaccines in historic evolution and perspective: a narrative of vaccine discoveries," *Vaccine*, vol. 18, no. 15, pp. 1436–1447, 2000.

[3] J. Parrino and B. S. Graham, "Smallpox vaccines: past, present, and future," *Journal of Allergy and Clinical Immunology*, vol. 118, no. 6, pp. 1320–1326, 2006.

[4] E. N. T. Meeusen, J. Walker, A. Peters, P. P. Pastoret, and G. Jungersen, "Current status of veterinary vaccines," *Clinical Microbiology Reviews*, vol. 20, no. 3, pp. 489–510, 2007.

[5] A. S. De Groot, H. Sbai, C. S. Aubin, J. McMurry, and W. Martin, "Immuno-informatics: mining genomes for vaccine components," *Immunology and Cell Biology*, vol. 80, no. 3, pp. 255–269, 2002.

[6] B. Peters, J. Sidney, P. Bourne et al., "The immune epitope database and analysis resource: from vision to blueprint," *PLoS Biology*, vol. 3, no. 3, p. e91, 2005.

[7] R. D. Fleischmann, M. D. Adams, O. White et al., "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd," *Science*, vol. 269, no. 5223, pp. 496–521, 1995.

[8] R. Rappuoli, "Reverse vaccinology," *Current Opinion in Microbiology*, vol. 3, no. 5, pp. 445–450, 2000.

[9] A.S. De Groot, J. McMurry, L. Moise, and B. Martin, "Epitope-based Immunome-derived vaccines: a strategy for improved design and safety," in *Applications of Immunomics*, A. Falus, Ed., Springer Immunomics Series, Springer, New York, NY, USA, 2009.

[10] M. Pizza, V. Scarlato, V. Masignani et al., "Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing," *Science*, vol. 287, pp. 1816–1820, 2000.

[11] W. Zhou, V. Pool, J. K. Iskander et al., "Surveillance for safety after immunization: Vaccine Adverse Event Reporting System (VAERS)—United States, 1991–2001," *MMWR. Surveillance Summaries*, vol. 52, no. 1, pp. 1–24, 2003.

[12] P. Tamayo, D. Slonim, J. Mesirov et al., "Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 96, no. 6, pp. 2907–2912, 1999.

[13] Z. Xiang, T. Todd, K. P. Ku et al., "VIOLIN: vaccine investigation and online information network," *Nucleic Acids Research*, vol. 36, no. 1, pp. D923–D928, 2008.

[14] Y. He, L. Cowell, A. D. Diehl et al., "VO: vaccine ontology," in *Proceedings of the 1st International Conference on Biomedical Ontology (ICBO '09)*, Buffalo, NY, USA, August 2009.

[15] R. R. Brinkman, M. Courtot, D. Derom et al., "Modeling biomedical experimental processes with OBI," *Journal of Biomedical Semantics*, vol. 1, supplement 1, p. S7, 2010.

[16] A. Ozgur, Z. Xiang, D. Radev, and Y. He, "Mining of vaccine associated IFN-γ gene interaction networks using the Vaccine Ontology," *Journal of Biomedical Semantics*. 2(Suppl 2):S8, 2011, http://www.jbiomedsem.com/qc/content/2/S2/S8.

[17] C. DeLisi and J. A. Berzofsky, "T-cell antigenic sites tend to be amphipathic structures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 82, no. 20, pp. 7048–7052, 1985.

[18] A. S. De Groot and J. A. Berzofsky, "From genome to vaccine—new immunoinformatics tools for vaccine design," *Methods*, vol. 34, no. 4, pp. 425–428, 2004.

[19] A. S. De Groot and L. Moise, "New tools, new approaches and new ideas for vaccine development," *Expert Review of Vaccines*, vol. 6, no. 2, pp. 125–127, 2007.

[20] J. D. Ahlers, I. M. Belyakov, E. K. Thomas, and J. A. Berzofsky, "High-affinity T helper epitope induces complementary helper and APC polarization, increased CTL, and protection against viral infection," *The Journal of Clinical Investigation*, vol. 108, no. 11, pp. 1677–1685, 2001.

[21] H. Inaba, W. Martin, A. S. De Groot, S. Qin, and L. J. De Groot, "Thyrotropin receptor epitopes and their relation to histocompatibility leukocyte antigen-DR molecules in graves' disease," *Journal of Clinical Endocrinology and Metabolism*, vol. 91, no. 6, pp. 2286–2294, 2006.

[22] A. S. De Groot, B. M. Jesdale, E. Szu, J. R. Schafer, R. M. Chicz, and G. Deocampo, "An interactive web site providing major histocompatibility ligand predictions: application to HIV research," *AIDS Research and Human Retroviruses*, vol. 13, no. 7, pp. 529–531, 1997.

[23] K. B. Bond, B. Sriwanthana, T. W. Hodge et al., "An HLA-directed molecular and bioinformatics approach identifies new HLA-A11 HIV-1 subtype E cytotoxic T lymphocyte epitopes in HIV-1-infected Thais," *AIDS Research and Human Retroviruses*, vol. 17, no. 8, pp. 703–717, 2001.

[24] J. McMurry, H. Sbai, M. L. Gennaro, E. J. Carter, W. Martin, and A. S. De Groot, "Analyzing Mycobacterium tuberculosis proteomes for candidate vaccine epitopes," *Tuberculosis*, vol. 85, no. 1-2, pp. 95–105, 2005.

[25] Y. Dong, S. Demaria, X. Sun et al., "HLA-A2-restricted CD8-cytotoxic-T-Cell responses to novel epitopes in mycobacterium tuberculosis superoxide dismutase, alanine dehydrogenase, and glutamine synthetase," *Infection and Immunity*, vol. 72, no. 4, pp. 2412–2415, 2004.

[26] O. A. Koita, D. Dabitao, I. Mahamadou et al., "Confirmation of immunogenic consensus sequence HIV-1 T-cell epitopes in Bamako, Mali and Providence, Rhode Island," *Human Vaccines*, vol. 2, no. 3, pp. 119–128, 2006.

[27] A. S. De Groot, "Immunomics: discovering new targets for vaccines and therapeutics," *Drug Discovery Today*, vol. 11, no. 5-6, pp. 203–209, 2006.

[28] J. A. McMurry, S. H. Gregory, L. Moise, D. Rivera, S. Buus, and A. S. De Groot, "Diversity of Francisella tularensis Schu4 antigens recognized by T lymphocytes after natural infections in humans: identification of candidate epitopes for inclusion in a rationally designed tularemia vaccine," *Vaccine*, vol. 25, no. 16, pp. 3179–3191, 2007.

[29] L. Moise, J. A. McMurry, S. Buus, S. Frey, W. D. Martin, and A. S. De Groot, "In silico-accelerated identification of conserved and immunogenic variola/vaccinia T-cell epitopes," *Vaccine*, vol. 27, no. 46, pp. 6471–6479, 2009.

[30] L. Moise, J. A. McMurry, J. Pappo et al., "Identification of genome-derived vaccine candidates conserved between human and mouse-adapted strains of H. pylori," *Human Vaccines*, vol. 4, no. 3, pp. 219–223, 2008.

[31] S. K. Kim, M. Cornberg, X. Z. Wang, H. D. Chen, L. K. Selin, and R. M. Welsh, "Private specificities of CD8 T cell responses control patterns of heterologous immunity," *Journal of Experimental Medicine*, vol. 201, no. 4, pp. 523–533, 2005.

[32] R. Rappuoli, "Bridging the knowledge gaps in vaccine design," *Nature Biotechnology*, vol. 25, no. 12, pp. 1361–1366, 2007.

[33] T. Elliott, V. Cerundolo, J. Elvin, and A. Townsend, "Peptide-induced conformational change of the class I heavy chain," *Nature*, vol. 351, no. 6325, pp. 402–406, 1991.

[34] K. Falk, O. Rotzschke, S. Stevanovic, G. Jung, and H. G. Rammensee, "Allele-specific motifs revealed by sequencing of self-peptides eluted from MHC molecules," *Nature*, vol. 351, no. 6324, pp. 290–296, 1991.

[35] O. Rotzschke, K. Falk, S. Stevanovic, G. Jung, P. Walden, and H. G. Rammensee, "Exact prediction of a natural T cell epitope," *European Journal of Immunology*, vol. 21, no. 11, pp. 2891–2894, 1991.

[36] E. R. Unanue, "Cellular studies on antigen presentation by class II MHC molecules," *Current Opinion in Immunology*, vol. 4, no. 1, pp. 63–69, 1992.

[37] J. H. Brown, T. S. Jardetzky, J. C. Gorga et al., "Three-dimensional structure of the human class II histocompatibility antigen HLA-DR1," *Nature*, vol. 364, no. 6432, pp. 33–39, 1993.

[38] R. M. Chicz, R. G. Urban, J. C. Gorga, D. A. A. Vignali, W. S. Lane, and J. L. Strominger, "Specificity and promiscuity among naturally processed peptides bound to HLA-DR alleles," *Journal of Experimental Medicine*, vol. 178, no. 1, pp. 27–47, 1993.

[39] R. N. Germain and D. H. Margulies, "The biochemistry and cell biology and antigen processing and presentation," *Annual Review of Immunology*, vol. 11, pp. 403–450, 1993.

[40] P. Cresswell and A. Lanzavecchia, "Antigen processing and recognition," *Current Opinion in Immunology*, vol. 13, no. 1, pp. 11–12, 2001.

[41] E. S. Trombetta and I. Mellman, "Cell biology of antigen processing in vitro and in vivo," *Annual Review of Immunology*, vol. 23, pp. 975–1028, 2005.

[42] E. Appella, E. A. Padlan, and D. F. Hunt, "Analysis of the structure of naturally processed peptides bound by class I and class II major histocompatibility complex molecules," *EXS*, vol. 73, pp. 105–119, 1995.

[43] R. N. Germain, F. Castellino, R. Han et al., "Processing and presentation of endocytically acquired protein antigens by MHC class II and class I molecules," *Immunological Reviews*, no. 151, pp. 5–30, 1996.

[44] A. Sette and J. Sidney, "Nine major HLA class I supertypes account for the vast preponderance of HLA-A and -B polymorphism," *Immunogenetics*, vol. 50, no. 3-4, pp. 201–212, 1999.

[45] A. S. De Groot, J. McMurry, and L. Moise, "Prediction of immunogenicity: in silico paradigms, ex vivo and in vivo correlates," *Current Opinion in Pharmacology*, vol. 8, no. 5, pp. 620–626, 2008.

[46] A. S. De Groot, D. S. Rivera, J. A. McMurry, S. Buus, and W. Martin, "Identification of immunogenic HLA-B7 "Achilles' heel" epitopes within highly conserved regions of HIV," *Vaccine*, vol. 26, no. 24, pp. 3059–3071, 2008.

[47] A. S. De Groot, E. A. Bishop, B. Khan et al., "Engineering immunogenic consensus T helper epitopes for a cross-clade HIV vaccine," *Methods*, vol. 34, no. 4, pp. 476–487, 2004.

[48] H. Plotnicky, D. Cyblat-Chanal, J. P. Aubry et al., "The immunodominant influenza matrix t cell epitope recognized in human induces influenza protection in HLA-A2/K transgenic mice," *Virology*, vol. 309, no. 2, pp. 320–329, 2003.

[49] P. Wang, J. Sidney, C. Dow, B. Mothé, A. Sette, and B. Peters, "A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach," *PLoS Computational Biology*, vol. 4, no. 4, Article ID e1000048, 2008.

[50] H. G. Rammensee, J. Bachmann, N. P. N. Emmerich, O. A. Bachor, and S. Stevanović, "SYFPEITHI: database for MHC ligands and peptide motifs," *Immunogenetics*, vol. 50, no. 3-4, pp. 213–219, 1999.

[51] K. C. Parker, M. A. Bednarek, and J. E. Coligan, "Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains," *Journal of Immunology*, vol. 152, no. 1, pp. 163–175, 1994.

[52] A. S. De Groot and W. Martin, "Reducing risk, improving outcomes: bioengineering less immunogenic protein therapeutics," *Clinical Immunology*, vol. 131, no. 2, pp. 189–201, 2009.

[53] H. H. Lin, S. Ray, S. Tongchusak, E. L. Reinherz, and V. Brusic, "Evaluation of MHC class I peptide binding prediction servers: applications for vaccine research," *BMC Immunology*, vol. 9, article 8, 2008.

[54] H. H. Lin, G. L. Zhang, S. Tongchusak, E. L. Reinherz, and V. Brusic, "Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research," *BMC Bioinformatics*, vol. 9, no. 12, article S22, 2008.

[55] U. Gowthaman and J. N. Agrewala, "In silico tools for predicting peptides binding to HLA-class II molecules: more confusion than conclusion," *Journal of Proteome Research*, vol. 7, no. 1, pp. 154–163, 2008.

[56] P. Wang, J. Sidney, Y. Kim et al., "Peptide binding predictions for HLA DR, DP and DQ molecules," *BMC Bioinformatics*, vol. 11, 2010.

[57] S. Bulik, B. Peters, C. Ebeling, and H. Holzhütter, "Cytosolic processing of proteasomal cleavage products can enhance the presentation efficiency of MHC-1 epitopes," *Genome Informatics*, vol. 15, no. 1, pp. 24–34, 2004.

[58] B. Peters, S. Bulik, R. Tampe, P. M. Van Endert, and H. G. Holzhütter, "Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors," *Journal of Immunology*, vol. 171, no. 4, pp. 1741–1749, 2003.

[59] M. Nielsen, C. Lundegaard, O. Lund, and C. Keşmir, "The role of the proteasome in generating cytotoxic T-cell epitopes: insights obtained from improved predictions of proteasomal cleavage," *Immunogenetics*, vol. 57, no. 1-2, pp. 33–41, 2005.

[60] M. V. Larsen, C. Lundegaard, K. Lamberth, S. Buus, O. Lund, and M. Nielsen, "Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction," *BMC Bioinformatics*, vol. 8, article 424, 2007.

[61] S. Tenzer, B. Peters, S. Bulik et al., "Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding," *Cellular and Molecular Life Sciences*, vol. 62, no. 9, pp. 1025–1037, 2005.

[62] T. Stranzl, M. V. Larsen, C. Lundegaard, and M. Nielsen, "NetCTLpan: pan-specific MHC class I pathway epitope predictions," *Immunogenetics*, vol. 62, no. 6, pp. 357–368, 2010.

[63] A. B. Riemer, D. B. Keskin, G. Zhang et al., "A conserved E7-derived cytotoxic T lymphocyte epitope expressed on human papillomavirus 16-transformed HLA-A2+ epithelial cancers," *The Journal of Biological Chemistry*, vol. 285, pp. 29608–29622, 2010.

[64] D. Enshell-Seijffers, D. Denisov, B. Groisman et al., "The mapping and reconstitution of a conformational discontinuous B-cell epitope of HIV-1," *Journal of Molecular Biology*, vol. 334, no. 1, pp. 87–101, 2003.

[65] A. Schreiber, M. Humbert, A. Benz, and U. Dietrich, "3D-Epitope-Explorer (3DEX): localization of conformational epitopes within three-dimensional structures of proteins," *Journal of Computational Chemistry*, vol. 26, no. 9, pp. 879–887, 2005.

[66] U. Kulkarni-Kale, S. Bhosle, and A. S. Kolaskar, "CEP: a conformational epitope prediction server," *Nucleic Acids Research*, vol. 33, no. 2, pp. W168–W171, 2005.

[67] M. J. Sweredoski and P. Baldi, "PEPITO: improved discontinuous B-cell epitope prediction using multiple distance thresholds and half sphere exposure," *Bioinformatics*, vol. 24, no. 12, pp. 1459–1460, 2008.

[68] D. J. Barlow, M. S. Edwards, and J. M. Thornton, "Continuous and discontinuous protein antigenic determinants," *Nature*, vol. 322, no. 6081, pp. 747–748, 1986.

[69] M. J. Blythe and D. R. Flower, "Benchmarking B cell epitope prediction: underperformance of existing methods," *Protein Science*, vol. 14, no. 1, pp. 246–248, 2005.

[70] U. Reimer, "Prediction of linear B-cell epitopes," *Methods in Molecular Biology*, vol. 524, pp. 335–344, 2009.

[71] E. Rajnavolgyi, N. Nagy, B. Thuresson et al., "A repetitive sequence of Epstein-Barr virus nuclear antigen 6 comprises overlapping T cell epitopes which induce HLA-DR-restricted

CD4(+) T lymphocytes," *International Immunology*, vol. 12, no. 3, pp. 281–293, 2000.

[72] C. M. Graham, B. C. Barnett, I. Hartlmayr et al., "The structural requirements for class II (I-A(d))-restricted T cell recognition of influenza hemagglutinin: b cell epitopes define T cell epitopes," *European Journal of Immunology*, vol. 19, no. 3, pp. 523–528, 1989.

[73] W. Fischer, S. Perkins, J. Theiler et al., "Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants," *Nature Medicine*, vol. 13, no. 1, pp. 100–106, 2007.

[74] A. S. De Groot, L. Marcon, E. A. Bishop et al., "HIV vaccine development by computer assisted design: the GAIA vaccine," *Vaccine*, vol. 23, no. 17-18, pp. 2136–2148, 2005.

[75] B. Gaschen, J. Taylor, K. Yusim et al., "Diversity considerations in HIV-1 vaccine selection," *Science*, vol. 296, no. 5577, pp. 2354–2360, 2002.

[76] F. Gao, B. T. Korber, E. A. Weaver, H. X. Liao, B. H. Hahn, and B. F. Haynes, "Centralized immunogens as a vaccine strategy to overcome HIV-1 diversity," *Expert Review of Vaccines*, vol. 3, no. 4, pp. S161–S168, 2004.

[77] F. Gao, E. A. Weaver, Z. Lu et al., "Antigenicity and immunogenicity of a synthetic human immunodeficiency virus type 1 group M consensus envelope glycoprotein," *Journal of Virology*, vol. 79, no. 2, pp. 1154–1163, 2005.

[78] D. C. Nickle, M. Rolland, M. A. Jensen et al., "Coping with viral diversity in HIV vaccine design," *PLoS Computational Biology*, vol. 3, no. 4, article e75, pp. 754–762, 2007.

[79] L. A. McNamara, Y. He, and Z. Yang, "Using epitope predictions to evaluate efficacy and population coverage of the Mtb72f vaccine for tuberculosis," *BMC Immunology*, vol. 11, article 18, 2010.

[80] J. Söllner, A. Heinzel, G. Summer et al., "Concept and application of a computational vaccinology workflow," *Immunome Research*, vol. 6, supplement 2, 2010.

[81] F. Pappalardo, M. D. Halling-Brown, N. Rapin et al., "ImmunoGrid, an integrative environment for large-scale simulation of the immune system for vaccine discovery, design and optimization," *Briefings in Bioinformatics*, vol. 10, no. 3, pp. 330–340, 2009.

[82] M. Feldhahn, P. Dönnes, P. Thiel, and O. Kohlbacher, "FRED—a framework for T-cell epitope detection," *Bioinformatics*, vol. 25, no. 20, pp. 2758–2759, 2009.

[83] L. Moise, J. A. McMurry, J. Pappo et al., "Identification of genome-derived vaccine candidates conserved between human and mouse-adapted strains of H. pylori," *Human Vaccines*, vol. 4, no. 3, pp. 219–223, 2008.

[84] S. H. Gregory, S. Mott, J. Phung et al., "Epitope-based vaccination against pneumonic tularemia," *Vaccine*, vol. 27, no. 39, pp. 5299–5306, 2009.

[85] A. L. Delcher, D. Harmon, S. Kasif, O. White, and S. L. Salzberg, "Improved microbial gene identification with GLIMMER," *Nucleic Acids Research*, vol. 27, no. 23, pp. 4636–4641, 1999.

[86] D. Frishman, A. Mironov, H. W. Mewes, and M. Gelfand, "Combining diverse evidence for gene recognition in completely sequenced bacterial genomes," *Nucleic Acids Research*, vol. 26, no. 12, pp. 2941–2947, 1998.

[87] J. Besemer, A. Lomsadze, and M. Borodovsky, "GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions," *Nucleic Acids Research*, vol. 29, no. 12, pp. 2607–2618, 2001.

[88] S. T. Fitz-Gibbon, H. Ladner, U. J. Kim, K. O. Stetter, M. I. Simon, and J. H. Miller, "Genome sequence of the hyperthermophilic crenarchaeon Pyrobaculum aerophilum," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 2, pp. 984–989, 2002.

[89] A. M. Cerdeño-Tárraga, A. Efstratiou, L. G. Dover et al., "The complete genome sequence and analysis of Corynebacterium diphtheriae NCTC13129," *Nucleic Acids Research*, vol. 31, no. 22, pp. 6516–6523, 2003.

[90] J. Wei, M. B. Goldberg, V. Burland et al., "Complete genome sequence and comparative genomics of Shigella flexneri serotype 2a strain 2457T," *Infection and Immunity*, vol. 71, no. 5, pp. 2775–2786, 2003.

[91] M. P. McLeod, X. Qin, S. E. Karpathy et al., "Complete genome sequence of Rickettsia typhi and comparison with sequences of other rickettsiae," *Journal of Bacteriology*, vol. 186, no. 17, pp. 5842–5855, 2004.

[92] P. Stothard and D. S. Wishart, "Automated bacterial genome analysis and annotation," *Current Opinion in Microbiology*, vol. 9, no. 5, pp. 505–510, 2006.

[93] A. Economou, P. J. Christie, R. C. Fernandez, T. Palmer, G. V. Plano, and A. P. Pugsley, "Secretion by numbers: protein traffic in prokaryotes," *Molecular Microbiology*, vol. 62, no. 2, pp. 308–319, 2006.

[94] T. T. Tseng, B. M. Tyler, and J. C. Setubal, "Protein secretion systems in bacterial-host associations, and their description in the Gene Ontology," *BMC Microbiology*, vol. 9, no. 1, article S2, 2009.

[95] H. Tjalsma, A. Bolhuis, J. D. H. Jongbloed, S. Bron, and J. M. Van Dijl, "Signal peptide-dependent protein transport in Bacillus subtilis: a genome-based survey of the secretome," *Microbiology and Molecular Biology Reviews*, vol. 64, no. 3, pp. 515–547, 2000.

[96] H. Tjalsma, H. Antelmann, J. D. H. Jongbloed et al., "Proteomics of protein secretion by Bacillus subtilis: separating the "secrets" of the secretome," *Microbiology and Molecular Biology Reviews*, vol. 68, no. 2, pp. 207–233, 2004.

[97] J. L. Gardy and F. S. L. Brinkman, "Methods for predicting bacterial protein subcellular localization," *Nature Reviews Microbiology*, vol. 4, no. 10, pp. 741–751, 2006.

[98] C. S. Yu, C. J. Lin, and J. K. Hwang, "Predicting subcellular localization of proteins for Gram-negative bacteria by support vector machines based on n-peptide compositions," *Protein Science*, vol. 13, no. 5, pp. 1402–1406, 2004.

[99] Z. Lu, D. Szafron, R. Greiner et al., "Predicting subcellular localization of proteins using machine-learned classifiers," *Bioinformatics*, vol. 20, no. 4, pp. 547–556, 2004.

[100] A. P. Tampakaki, V. E. Fadouloglou, A. D. Gazi, N. J. Panopoulos, and M. Kokkinidis, "Conserved features of type III secretion," *Cellular Microbiology*, vol. 6, no. 9, pp. 805–816, 2004.

[101] P. D. Schloss and JO. Handelsman, "Status of the microbial census," *Microbiology and Molecular Biology Reviews*, vol. 68, no. 4, pp. 686–691, 2004.

[102] F. M. Cohan and E. B. Perry, "A Systematics for Discovering the Fundamental Units of Bacterial Diversity," *Current Biology*, vol. 17, no. 10, pp. R373–R386, 2007.

[103] N. T. Perna, G. Plunkett, V. Burland et al., "Genome sequence of enterohaemorrhagic Escherichia coli O157:H7," *Nature*, vol. 409, no. 6819, pp. 529–533, 2001.

[104] Y. Zhang, C. Laing, M. Steele et al., "Genome evolution in major Escherichia coli O157:H7 lineages," *BMC Genomics*, vol. 8, 2007.

[105] M. Brochet, E. Couvé, M. Zouine et al., "Genomic diversity and evolution within the species Streptococcus agalactiae," *Microbes and Infection*, vol. 8, no. 5, pp. 1227–1243, 2006.

[106] H. Tettelin, V. Masignani, M. J. Cieslewicz et al., "Genome analysis of multiple pathogenic isolates of Streptococcus agalactiae: implications for the microbial "pan-genome"," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 39, pp. 13950–13955, 2005.

[107] E. Brzuszkiewicz, H. Brüggemann, H. Liesegang et al., "How to become a uropathogen: comparative genomic analysis of extraintestinal pathogenic Escherichia coli strains," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 34, pp. 12879–12884, 2006.

[108] H. Tettelin, D. Medini, C. Donati, and V. Masignani, "Towards a universal group B Streptococcus vaccine using multistrain genome analysis," *Expert Review of Vaccines*, vol. 5, no. 5, pp. 687–694, 2006.

[109] D. Maione, I. Margarit, C. D. Rinaudo et al., "Immunology: identification of a universal group B Streptococcus vaccine by multiple genome screen," *Science*, vol. 309, no. 5731, pp. 148–150, 2005.

[110] Y. He, Z. Xiang, and H. L.T. Mobley, "Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development," *Journal of Biomedicine and Biotechnology*, vol. 2010, Article ID 297505, 2010 pages, 2010.

[111] H. Ochman and N. A. Moran, "Genes lost and genes found: evolution of bacterial pathogenesis and symbiosis," *Science*, vol. 292, no. 5519, pp. 1096–1098, 2001.

[112] G. Sachdeva, K. Kumar, P. Jain, and S. Ramachandran, "SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks," *Bioinformatics*, vol. 21, no. 4, pp. 483–491, 2005.

[113] S. Vivona, F. Bernante, and F. Filippini, "NERVE: new enhanced reverse vaccinology environment," *BMC Biotechnology*, vol. 6, article 35, 2006.

[114] Y. He and Z. Xiang, "Bioinformatics analysis of Brucella vaccines and vaccine targets using VIOLIN," *Immunome Research*, vol. 6, supplement 1, p. S5, 2010.

[115] Z. Xiang and Y. He, "Procedia in vaccinology," in *Proceedings of the 2nd Global Congress on Vaccines*, vol. 1, pp. 23–29, Boston, Mass, USA, 2009.

[116] D. Serruto and R. Rappuoli, "Post-genomic vaccine development," *The FEBS Letters*, vol. 580, no. 12, pp. 2985–2992, 2006.

[117] M. Mora, C. Donati, D. Medini, A. Covacci, and R. Rappuoli, "Microbial genomes and vaccine design: refinements to the classical reverse vaccinology approach," *Current Opinion in Microbiology*, vol. 9, no. 5, pp. 532–536, 2006.

[118] E. L. Hendrickson, R. J. Lamont, and M. Hackett, "Tools for interpreting large-scale protein profiling in microbiology," *Journal of Dental Research*, vol. 87, no. 11, pp. 1004–1015, 2008.

[119] Y. Liang and A. Kelemen, "Associating phenotypes with molecular events: recent statistical advances and challenges underpinning microarray experiments," *Functional and Integrative Genomics*, vol. 6, no. 1, pp. 1–13, 2006.

[120] Affymetrix, *GeneChip Expression Data Analysis Fundamentals*, Affymetrix, Inc., Santa Clara, Calif, USA, 2004.

[121] R. A. Irizarry, B. Hobbs, F. Collin et al., "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.

[122] C. Li and W. H. Wong, "Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 1, pp. 31–36, 2001.

[123] R. C. Gentleman, V. J. Carey, D. M. Bates et al., "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, no. 10, p. R80, 2004.

[124] W. Pan, "A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments," *Bioinformatics*, vol. 18, no. 4, pp. 546–554, 2002.

[125] J. F. Ayroles and G. Gibson, "Analysis of variance of microarray data," *Methods in Enzymology*, vol. 411, pp. 214–233, 2006.

[126] V. G. Tusher, R. Tibshirani, and G. Chu, "Significance analysis of microarrays applied to the ionizing radiation response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 5116–5121, 2001.

[127] G. K. Smyth, "Linear models and empirical bayes methods for assessing differential expression in microarray experiments," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, article 3, 2004.

[128] S. Raychaudhuri, P. D. Sutphin, J. T. Chang, and R. B. Altman, "Basic microarray analysis: grouping and feature reduction," *Trends in Biotechnology*, vol. 19, no. 5, pp. 189–193, 2001.

[129] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, "Cluster analysis and display of genome-wide expression patterns," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 25, pp. 14863–14868, 1998.

[130] Z. Xiang, Z. S. Qin, and Y. He, "CRCView: a web server for analyzing and visualizing microarray gene expression data using model-based clustering," *Bioinformatics*, vol. 23, no. 14, pp. 1843–1845, 2007.

[131] D. A. Hosack, G. Dennis Jr., B. T. Sherman, H. C. Lane, and R. A. Lempicki, "Identifying biological themes within lists of genes with EASE," *Genome Biology*, vol. 4, no. 10, p. R70, 2003.

[132] A. Bild and P. G. Febbo, "Application of a priori established gene sets to discover biologically important differential expression in microarray data," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 43, pp. 15278–15279, 2005.

[133] D. A. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2009.

[134] T. Beißbarth and T. P. Speed, "GOstat: find statistically overrepresented Gene Ontologies with a group of genes," *Bioinformatics*, vol. 20, no. 9, pp. 1464–1465, 2004.

[135] Y. U. Xia, H. Yu, R. Jansen et al., "Analyzing cellular biochemistry in terms of molecular networks," *Annual Review of Biochemistry*, vol. 73, pp. 1051–1087, 2004.

[136] J. Goutsias and N. H. Lee, "Computational and experimental approaches for modeling gene regulatory networks," *Current Pharmaceutical Design*, vol. 13, no. 14, pp. 1415–1436, 2007.

[137] J. Hardin, A. Mitani, L. Hicks, and B. VanKoten, "A robust measure of correlation between two genes on a microarray," *BMC Bioinformatics*, vol. 8, Article 220, 2007.

[138] P. A. Morel, S. Ta'asan, B. F. Morel, D. E. Kirschner, and J. L. Flynn, "New insights into mathematical modeling of the

immune system," *Immunologic Research*, vol. 36, no. 1–3, pp. 157–165, 2006.

[139] G. Dreyfus, *Neural Networks : Methodology and Applications*, Springer, New York, NY, USA, 2005.

[140] N. Friedman, M. Linial, I. Nachman, and D. Pe'er, "Using Bayesian networks to analyze expression data," *Journal of Computational Biology*, vol. 7, no. 3-4, pp. 601–620, 2000.

[141] Z. Xiang, R. M. Minter, X. Bi, P. J. Woolf, and Y. He, "miniTUBA: medical inference by network integration of temporal data using Bayesian analysis," *Bioinformatics*, vol. 23, no. 18, pp. 2423–2432, 2007.

[142] A. Stuart, J. K. Ord, and S. F. Arnold, *Kendall's Advanced Theory of Statistics*, Oxford University Press, New York, NY, USA, 2004.

[143] J. Yu, V. A. Smith, P. P. Wang, A. J. Hartemink, and E. D. Jarvis, "Advances to Bayesian network inference for generating causal networks from observational biological data," *Bioinformatics*, vol. 20, no. 18, pp. 3594–3603, 2004.

[144] N. Dhiman, R. Bonilla, D. O'Kane J, and G. A. Poland, "Gene expression microarrays: a 21st century tool for directed vaccine design," *Vaccine*, vol. 20, no. 1-2, pp. 22–30, 2001.

[145] T. Chen, "DNA microarrays—an armory for combating infectious diseases in the new century," *Infectious Disorders—Drug Targets*, vol. 6, no. 3, pp. 263–279, 2006.

[146] J. A. Young, Q. L. Fivelman, P. L. Blair et al., "The Plasmodium falciparum sexual development transcriptome: a microarray analysis using ontology-based pattern identification," *Molecular and Biochemical Parasitology*, vol. 143, no. 1, pp. 67–79, 2005.

[147] T. Sturniolo, E. Bono, J. Ding et al., "Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices," *Nature Biotechnology*, vol. 17, no. 6, pp. 555–561, 1999.

[148] T. D. Querec, R. S. Akondy, E. K. Lee et al., "Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans," *Nature Immunology*, vol. 10, no. 1, pp. 116–125, 2009.

[149] D. Gaucher, R. Therrien, N. Kettaf et al., "Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses," *Journal of Experimental Medicine*, vol. 205, no. 13, pp. 3119–3131, 2008.

[150] B. Pulendran, J. Miller, T. D. Querec et al., "Case of yellow fever vaccine-associated viscerotropic disease with prolonged viremia, robust adaptive immune responses, and polymorphisms in CCR5 and RANTES genes," *Journal of Infectious Diseases*, vol. 198, no. 4, pp. 500–507, 2008.

[151] B. A. McKinney, D. M. Reif, M. T. Rock et al., "Cytokine expression patterns associated with systemic adverse events following smallpox immunization," *Journal of Infectious Diseases*, vol. 194, no. 4, pp. 444–453, 2006.

[152] I. Hamaguchi, J.-i. Imai, H. Momose et al., "Two vaccine toxicity-related genes Agp and Hpx could prove useful for pertussis vaccine safety control," *Vaccine*, vol. 25, no. 17, pp. 3355–3364, 2007.

[153] T. Mizukami, J. I. Imai, I. Hamaguchi et al., "Application of DNA microarray technology to influenza A/Vietnam/1194/2004 (H5N1) vaccine safety evaluation," *Vaccine*, vol. 26, no. 18, pp. 2270–2283, 2008.

[154] V. Brusic and N. Petrovsky, "Immunoinformatics and its relevance to understanding human immune disease," *Expert Review of Clinical Immunology*, vol. 1, pp. 145–157, 2005.

[155] C. A.A. Beauchemin and A. Handel, "A review of mathematical models of influenza A infections within a host or cell culture: lessons learned and challenges ahead," *BMC Public Health*, vol. 11, supplement 1, p. S7, 2011.

[156] D. E. Kirschner, S. T. Chang, T. W. Riggs, N. Perry, and J. J. Linderman, "Toward a multiscale model of antigen presentation in immunity," *Immunological Reviews*, vol. 216, no. 1, pp. 93–118, 2007.

[157] M. A. Fishman and A. S. Perelson, "Th1/Th2 cross regulation," *Journal of Theoretical Biology*, vol. 170, no. 1, pp. 25–56, 1994.

[158] B. F. Keele, E. E. Giorgi, J. F. Salazar-Gonzalez et al., "Identification and characterization of transmitted and early founder virus envelopes in primary HIV-1 infection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 21, pp. 7552–7557, 2008.

[159] S. Marino, J. J. Linderman, and D. E. Kirschner, "A multifaceted approach to modeling the immune response in tuberculosis," *Wiley Interdisciplinary Reviews. Systems Biology and Medicine*, In Press.

[160] M. P. Davenport, R. M. Ribeiro, L. Zhang, D. P. Wilson, and A. S. Perelson, "Understanding the mechanisms and limitations of immune control of HIV," *Immunological Reviews*, vol. 216, no. 1, pp. 164–175, 2007.

[161] J. Schulze-Horsel, M. Schulze, G. Agalaridis, Y. Genzel, and U. Reichl, "Infection dynamics and virus-induced apoptosis in cell culture-based influenza vaccine production-Flow cytometry and mathematical modeling," *Vaccine*, vol. 27, no. 20, pp. 2712–2722, 2009.

[162] S. Y. Kim and S. J. Goldie, "Cost-effectiveness analyses of vaccination programmes: a focused review of modelling approaches," *PharmacoEconomics*, vol. 26, no. 3, pp. 191–215, 2008.

[163] S. J. Goldie, J. J. Kim, K. Kobus et al., "Cost-effectiveness of HPV 16, 18 vaccination in Brazil," *Vaccine*, vol. 25, no. 33, pp. 6257–6270, 2007.

[164] S. Aballéa, J. Chancellor, M. Martin et al., "The cost-effectiveness of influenza vaccination for people aged 50 to 64 years: an international model," *Value in Health*, vol. 10, no. 2, pp. 98–116, 2007.

[165] D. J. Isaacman, D. R. Strutton, E. A. Kalpas et al., "The impact of indirect (herd) protection on the cost-effectiveness of pneumococcal conjugate vaccine," *Clinical Therapeutics*, vol. 30, no. 2, pp. 341–357, 2008.

[166] A. Palladini, G. Nicoletti, F. Pappalardo et al., "In silico modeling and in vivo efficacy of cancer-preventive vaccinations," *Cancer Research*, vol. 70, pp. 7775–7763, 2010.

[167] M. Pennisi, F. Pappalardo, A. Palladini et al., "Modeling the competition between lung metastases and the immune system using agents," *BMC Bioinformatics*, vol. 11, supplement 7, article S13, 2010.

[168] F. Pappalardo, P. L. Lollini, F. Castiglione, and S. Motta, "Modeling and simulation of cancer immunoprevention vaccine," *Bioinformatics*, vol. 21, no. 12, pp. 2891–2897, 2005.

[169] P.-L. Lollini, S. Motta, and F. Pappalardo, "Discovery of cancer vaccination protocols with a genetic algorithm driving an agent based simulator," *BMC Bioinformatics*, vol. 7, article 352, 2006.

[170] F. Pappalardo, M. Pennisi, F. Castiglione, and S. Motta, "Vaccine protocols optimization: in silico experiences," *Biotechnology Advances*, vol. 28, no. 1, pp. 82–93, 2010.

[171] S. M. Blower, K. Koelle, D. E. Kirschner, and J. Mills, "Live attenuated HIV vaccines: predicting the tradeoff between efficacy and safety," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 6, pp. 3618–3623, 2001.

[172] T. J. John and R. Samuel, "Herd immunity and herd effect: new insights and definitions," *European Journal of Epidemiology*, vol. 16, no. 7, pp. 601–606, 2000.

[173] "Recommendations of the International Task Force for Disease Eradication," *MMWR—Recommendations and Reports*, vol. 42, no. 16, pp. 1–38, 1993.

[174] National Center for Immunization and Respiratory Diseases, "General recommendations on immunization—recommendations of the Advisory Committee on Immunization Practices (ACIP)," *MMWR—Recommendations and Reports*, vol. 60, no. 2, pp. 1–64, 2011.

[175] D. Wood, K. N. Saarlas, M. Inkelas, and B. T. Matyas, "Immunization registries in the United States: implications for the practice of public health in a changing health care system," *Annual Review of Public Health*, vol. 20, pp. 231–255, 1999.

[176] R. T. Chen, J. W. Glasser, P. H. Rhodes et al., "Vaccine safety datalink project: a new tool for improving vaccine safety monitoring in the United States," *Pediatrics*, vol. 99, no. 6, pp. 765–773, 1997.

[177] B. P. Hull, S. L. Deeks, and P. B. McIntyre, "The Australian Childhood Immunisation Register-A model for universal immunisation registers?" *Vaccine*, vol. 27, no. 37, pp. 5054–5060, 2009.

[178] "National standards for immunization coverage assessment: recommendations from the Canadian Immunization Registry Network," *Canada Communicable Disease Report*, vol. 31, pp. 93–97, 2005.

[179] C. Trewin, H. B. Strand, and E. K. Grøholt, "Norhealth: norwegian health information system," *Scandinavian Journal of Public Health*, vol. 36, no. 7, pp. 685–689, 2008.

[180] M. L. Popovich, J. J. Aramini, and M. Garcia, "Immunizations: the first step in a personal health record to empower patients," *Stud Health Technol Inform*, vol. 137, pp. 286–295, 2008.

[181] D. B. Fishbein, B. C. Willis, W. M. Cassidy et al., "Determining indications for adult vaccination: patient self-assessment, medical record, or both?" *Vaccine*, vol. 24, no. 6, pp. 803–818, 2006.

[182] J. A. Boom, A. C. Dragsbaek, and C. S. Nelson, "The success of an immunization information system in the wake of Hurricane Katrina," *Pediatrics*, vol. 119, no. 6, pp. 1213–1217, 2007.

[183] K. A. Feemster, C. V. Spain, M. Eberhart, S. Pati, and B. Watson, "Identifying infants at increased risk for late initiation of immunizations: maternal and provider characteristics," *Public Health Reports*, vol. 124, no. 1, pp. 42–53, 2009.

[184] F. Wei, J. P. Mullooly, M. Goodman et al., "Identification and characteristics of vaccine refusers," *BMC Pediatrics*, vol. 9, no. 1, article 18, 2009.

[185] V. L. Hinrichsen, B. Kruskal, M. A. O'Brien, T. A. Lieu, and R. Platt, "Using electronic medical records to enhance detection and reporting of vaccine adverse events," *Journal of the American Medical Informatics Association*, vol. 14, no. 6, pp. 731–735, 2007.

[186] A. R. Hinman, G. A. Urquhart, R. A. Strikas et al., "Immunization information systems: national vaccine advisory committee progress report, 2007," *Journal of Public Health Management and Practice*, vol. 13, no. 6, pp. 553–558, 2007.

[187] C. P. Farrington and E. Miller, "Vaccine trials," *Molecular Biotechnology*, vol. 17, no. 1, pp. 43–58, 2001.

[188] D. S. Fedson, "Measuring protection: efficacy versus effectiveness," *Developments in biological standardization*, vol. 95, pp. 195–201, 1998.

[189] R. T. Chen, S. C. Rastogi, J. R. Mullen et al., "The vaccine adverse event reporting system (VAERS)," *Vaccine*, vol. 12, no. 6, pp. 542–550, 1994.

[190] K. S. Lankinen, S. Pastila, T. Kilpi, H. Nohynek, P. H. Mäkelä, and P. Olin, "Vaccinovigilance in Europe—need for timeliness, standardization and resource," *Bulletin of the World Health Organization*, vol. 82, no. 11, pp. 828–835, 2004.

[191] D. Banks, E. J. Woo, D. R. Burwen, P. Perucci, M. M. Braun, and R. Ball, "Comparing data mining methods on the VAERS database," *Pharmacoepidemiology and Drug Safety*, vol. 14, no. 9, pp. 601–609, 2005.

[192] T. Verstraeten, F. DeStefano, R. T. Chen, and E. Miller, "Vaccine safety surveillance using large linked databases: opportunities, hazards and proposed guidelines," *Expert Review of Vaccines*, vol. 2, no. 1, pp. 21–29, 2003.

[193] J. Baggs, J. Gee, E. Lewis et al., "The Vaccine Safety Datalink: a model for monitoring immunization safety," *Pediatrics*, vol. 127, supplement 1, pp. S45–S53, 2011.

[194] P. Kramarz, E. K. France, F. Destefano et al., "Population-based study of rotavirus vaccination and intussusception," *Pediatric Infectious Disease Journal*, vol. 20, no. 4, pp. 410–416, 2001.

[195] W. W. Thompson, C. Price, B. Goodson et al., "Early thimerosal exposure and neuropsychological outcomes at 7 to 10 years," *The New England Journal of Medicine*, vol. 357, no. 13, pp. 1281–1292, 2007.

[196] F. DeStefano, T. Verstraeten, L. A. Jackson et al., "Vaccinations and risk of central nervous system demyelinating diseases in adults," *Archives of Neurology*, vol. 60, pp. 504–509, 2003.

[197] P. Farrington, S. Pugh, A. Colville et al., "A new method for active surveillance of adverse events from diphtheria/tetanus/pertussis and measles/mumps/rubella vaccines," *The Lancet*, vol. 345, no. 8949, pp. 567–569, 1995.

[198] M. Ali, C. G. Do, J. D. Clemens et al., "The use of a computerized database to monitor vaccine safety in Viet Nam," *Bulletin of the World Health Organization*, vol. 83, no. 8, pp. 604–610, 2005.

[199] D. Tapscott, *Wikinomics*, Penguin, New York, NY, USA, 2008.

[200] V. E. Pitzer, C. Viboud, L. Simonsen et al., "Demographic variability, vaccination, and the spatiotemporal dynamics of rotavirus epidemics," *Science*, vol. 325, no. 5938, pp. 290–294, 2009.

[201] C. G. Whitney, M. M. Farley, J. Hadler et al., "Decline in invasive pneumococcal disease after the introduction of protein-polysaccharide conjugate vaccine," *The New England Journal of Medicine*, vol. 348, no. 18, pp. 1737–1746, 2003.

[202] S. O. Sow, M. D. Tapia, S. Diallo et al., "Haemophilus influenzae type b conjugate vaccine introduction in Mali: impact on disease burden and serologic correlate of protection," *American Journal of Tropical Medicine and Hygiene*, vol. 80, no. 6, pp. 1033–1038, 2009.

[203] R. T. Chen, R. Weierbach, Z. Bisoffi et al., "A 'post-honeymoon period' measles outbreak in Muyinga sector, Burundi," *International Journal of Epidemiology*, vol. 23, no. 1, pp. 185–193, 1994.

[204] R. T. Chen, I. R. Hardy, P. H. Rhodes, D. K. Tyshchenko, A.V. Moiseeva, and V. F. Marievsky, "Ukraine, 1992: first assessment of diphtheria vaccine effectiveness during the recent resurgence of diphtheria in the Former Soviet Union," *Journal of Infectious Diseases*, vol. 181, supplement 1, pp. S178–S183, 2000.

[205] N. C. Grassly, J. Wenger, S. Durrani et al., "Protective efficacy of a monovalent oral type 1 poliovirus vaccine: a case-control study," *The Lancet*, vol. 369, no. 9570, pp. 1356–1362, 2007.

[206] W. H. McNeill, *Plagues and Peoples*, Blackwell Scientific Publications, Oxford, UK, 1976.

[207] M. A. Kermack, "A contribution to the mathematical theory of epidemics," *Proceedings of the Royal Society of London. Series A*, vol. 115, pp. 700–721, 1927.

[208] R. M. Anderson and R. M. May, "Directly transmitted infectious diseases: control by vaccination," *Science*, vol. 215, no. 4536, pp. 1053–1060, 1982.

[209] R. M. Anderson, R. M. May, and B. Anderson, *Infectious Diseases of Humans: Dynamics and Control*, Oxford University Press, Oxford, UK, 1991.

[210] K. Ellen and S. M. Cromley, *GIS and Public Health*, Guilford Press, New York, NY, USA, 2002.

[211] N. J. Gay, "The theory of measles elimination: implications for the design of elimination strategies," *Journal of Infectious Diseases*, vol. 189, no. 1, pp. S27–S35, 2004.

[212] H. R. Babad, D. J. Nokes, N. J. Gay, E. Miller, P. Morgan-Capner, and R. M. Anderson, "Predicting the impact of measles vaccination in England and Wales: model validation and analysis of policy options," *Epidemiology and Infection*, vol. 114, no. 2, pp. 319–344, 1995.

[213] J. M. Best, C. Castillo-Solorzano, J. S. Spika et al., "Reducing the global burden of congenital rubella syndrome: report of the World Health Organization Steering Committee on Research Related to Measles and Rubella Vaccines and Vaccination, June 2004," *Journal of Infectious Diseases*, vol. 192, no. 11, pp. 1890–1897, 2005.

[214] J. J. Kim, "Mathematical model of HPV provides insight into impacts of risk factors and vaccine," *PLoS Medicine*, vol. 3, no. 5, article e164, pp. 587–588, 2006.

[215] N. M. Ferguson, D. A. T. Cummings, C. Fraser, J. C. Cajka, P. C. Cooley, and D. S. Burke, "Strategies for mitigating an influenza pandemic," *Nature*, vol. 442, no. 7101, pp. 448–452, 2006.

[216] L. Temime, D. Guillemot, and P. Y. Boëlle, "Short- and long-term effects of pneumococcal conjugate vaccination of children on penicillin resistance," *Antimicrobial Agents and Chemotherapy*, vol. 48, no. 6, pp. 2206–2213, 2004.

[217] A. Zvi, N. Ariel, J. Fulkerson, J. C. Sadoff, and A. Shafferman, "Whole genome identification of Mycobacterium tuberculosis vaccine candidates by comprehensive data mining and bioinformatic analyses," *BMC Medical Genomics*, vol. 1, p. 18, 2008.

[218] A. Özgür, T. Vu, G. Erkan, and D. R. Radev, "Identifying gene-disease associations using centrality on a literature mined gene-interaction network," *Bioinformatics*, vol. 24, no. 13, pp. i277–i285, 2008.

[219] A. Ozgur, D. R. Radev, Z. Xiang, and Y. He, "Literature-based discovery of IFN-γ and vaccine-mediated gene interaction networks," *Journal of Biomedicine and Biotechnology*, vol. 2010, Article ID 426479, p. 13, 2010.

[220] H. Streeck, N. Frahm, and B. D. Walker, "The role of IFN-γ Elispot assay in HIV vaccine research," *Nature Protocols*, vol. 4, no. 4, pp. 461–469, 2009.

[221] H. A. Fletcher, "Correlates of immune protection from tuberculosis," *Current Molecular Medicine*, vol. 7, no. 3, pp. 319–325, 2007.

[222] P. Mansueto, G. Vitale, G. Di Lorenzo, G. B. Rini, S. Mansueto, and E. Cillari, "Immunopathology of leishmaniasis: an update," *International Journal of Immunopathology and Pharmacology*, vol. 20, no. 3, pp. 435–445, 2007.

[223] Y. He, R. Vemulapalli, A. Zeytun, and G. G. Schurig, "Induction of specific cytotoxic lymphocytes in mice vaccinated with Brucella abortus RB51," *Infection and Immunity*, vol. 69, no. 9, pp. 5502–5508, 2001.

[224] H. M. Müller, E. E. Kenny, and P. W. Sternberg, "Textpresso: an ontology-based information retrieval and extraction system for biological literature," *PLoS Biology*, vol. 2, no. 11, article e309, 2004.

[225] B. Yang, S. Sayers, Z. Xiang, and Y. He, "Protegen: a web-based protective antigen database and analysis system," *Nucleic Acids Research*, vol. 39, supplement 1, pp. D1073–D1078, 2011.

[226] H. R. Ansari, D. R. Flower, and G. P. S. Raghava, "AntigenDB: an immunoinformatics database of pathogen antigens," *Nucleic Acids Research*, vol. 38, no. 1, Article ID gkp830, pp. D847–D853, 2009.

[227] Y. He, Z. Xiang, T. Todd et al., "Bio-Ontologies 2010: semantic applications in life sciences," in *Proceedings of the 18th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB '10)*, p. 4, Boston, Mass, USA, July 2010.

[228] Z. Xiang and Y. He, "Improvement of pubmed literature searching using biomedical ontology," in *Proceedings of the 1st International Conference on Biomedical Ontology (ICBO '09)*, Buffalo, NY, USA, July 2009.

[229] J. Hur, Z. Xiang, E. Feldman, and Y. He, "Bio-Ontologies 2010: semantic applications in life sciences," in *Proceedings of the 18th Annual International Conference on Intelligent Systems for Molecular Biology (ISMB '10)*, Boston, Mass, USA, July 2010.

[230] O. Lund, M. Nielsen, C. Lundergaard, C. Kesmir, and S. Brunak, *Immunological Bioinformatics*, MIT Press, Cambridge, Mass, USA, 2005.

[231] G. A. Poland, I. G. Ovsyannikova, and R. M. Jacobson, "Personalized vaccines: the emerging field of vaccinomics," *Expert Opinion on Biological Therapy*, vol. 8, no. 11, pp. 1659–1667, 2008.

[232] M. Sullivan, "Moving candidate vaccines into development from research: lessons from HIV," *Immunology and Cell Biology*, vol. 87, no. 5, pp. 366–370, 2009.

[233] S. K. Pierce and L. H. Miller, "World Malaria Day 2009: what malaria knows about the immune system that immunologists still do not," *Journal of Immunology*, vol. 182, no. 9, pp. 5171–5177, 2009.

[234] M. Bhasin and G. P. S. Raghava, "A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes," *Journal of Biosciences*, vol. 32, no. 1, pp. 31–42, 2007.

[235] I. A. Doytchinova, P. Guan, and D. R. Flower, "EpiJen: a server for multistep T cell epitope prediction," *BMC Bioinformatics*, vol. 7, article 131, 2006.

[236] P. A. Reche, H. Zhang, J. P. Glutting, and E. L. Reinherz, "EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology," *Bioinformatics*, vol. 21, no. 9, pp. 2140–2141, 2005.

[237] A. S. De Groot, M. Ardito, E. M. McClaine, L. Moise, and W. D. Martin, "Immunoinformatic comparison of T-cell epitopes contained in novel swine-origin influenza A (H1N1) virus with epitopes in 2008-2009 conventional influenza vaccine," *Vaccine*, vol. 27, no. 42, pp. 5740–5747, 2009.

[238] N. Jojic, M. Reyes-Gomez, D. Heckerman, C. Kadie, and O. Schueler-Furman, "Learning MHC I—peptide binding," *Bioinformatics*, vol. 22, no. 14, pp. e227–e235, 2006.

[239] R. Vita, L. Zarebski, J. A. Greenbaum et al., "The immune epitope database 2.0," *Nucleic Acids Research*, vol. 38, no. 1, Article ID gkp1004, pp. D854–D862, 2009.

[240] L. Jacob and J. P. Vert, "Efficient peptide-MHC-I binding prediction for alleles with few known binders," *Bioinformatics*, vol. 24, no. 3, pp. 358–366, 2008.

[241] M. Bhasin and G. P. S. Raghava, "SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence," *Bioinformatics*, vol. 20, no. 3, pp. 421–423, 2004.

[242] P. Guan, I. A. Doytchinova, C. Zygouri, and D. R. Flower, "MHCPred: bringing a quantitative dimension to the online prediction of MHC binding," *Applied Bioinformatics*, vol. 2, no. 1, pp. 63–66, 2003.

[243] G. L. Zhang, A. M. Khan, K. N. Srinivasan, J. T. August, and V. Brusic, "MULTIPRED: a computational system for prediction of promiscuous HLA binding peptides," *Nucleic Acids Research*, vol. 33, no. 2, pp. W172–W179, 2005.

[244] S. Buus, S. L. Lauemøller, P. Worning et al., "Sensitive quantitative predictions of peptide-MHC binding by a 'Query by Committee' artificial neural network approach," *Tissue Antigens*, vol. 62, no. 5, pp. 378–384, 2003.

[245] M. Nielsen and O. Lund, "NN-align. An artificial neural network-based alignment algorithm for MHC class II peptide binding prediction," *BMC Bioinformatics*, vol. 10, article 1471, p. 296, 2009.

[246] I. Hoof, B. Peters, J. Sidney et al., "NetMHCpan, a method for MHC class i binding prediction beyond humans," *Immunogenetics*, vol. 61, no. 1, pp. 1–13, 2009.

[247] M. Nielsen, C. Lundegaard, T. Blicher et al., "Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan," *PLoS Computational Biology*, vol. 4, no. 7, Article ID e1000107, 2008.

[248] P. A. Reche and E. L. Reinherz, "PEPVAC: a web server for multi-epitope vaccine development based on the prediction of supertypic MHC ligands," *Nucleic Acids Research*, vol. 33, no. 2, pp. W138–W142, 2005.

[249] O. Schueler-Furman, Y. Altuvia, A. Sette, and H. Margalit, "Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles," *Protein Science*, vol. 9, no. 9, pp. 1838–1846, 2000.

[250] C. W. Tung and S. Y. Ho, "POPI: predicting immunogenicity of MHC class I binding peptides by mining informative physicochemical properties," *Bioinformatics*, vol. 23, no. 8, pp. 942–949, 2007.

[251] H. Singh and G. P. S. Raghava, "ProPred1: prediction of promiscuous MHC class-I binding sites," *Bioinformatics*, vol. 19, no. 8, pp. 1009–1014, 2003.

[252] H. Singh and G. P. S. Raghava, "ProPred: prediction of HLA-DR binding sites," *Bioinformatics*, vol. 17, no. 12, pp. 1236–1237, 2002.

[253] P. A. Reche, J. P. Glutting, H. Zhang, and E. L. Reinherz, "Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles," *Immunogenetics*, vol. 56, no. 6, pp. 405–419, 2004.

[254] P. Dönnes and A. Elofsson, "Prediction of MHC class I binding peptides, using SVMHC," *BMC Bioinformatics*, vol. 3, article 25, 2002.

[255] W. Liu, J. Wan, X. Meng, D. R. Flower, and T. Li, "In silico prediction of peptide-MHC binding affinity using SVRMHC," *Methods in Molecular Biology*, vol. 409, pp. 283–291, 2007.

[256] H. Bian and J. Hammer, "Discovery of promiscuous HLA-II-restricted T cell epitopes with TEPITOPE," *Methods*, vol. 34, no. 4, pp. 468–475, 2004.

[257] H. U. Chen, NI. Huang, and Z. Sun, "SubLoc: a server/client suite for protein subcellular location based on SOAP," *Bioinformatics*, vol. 22, no. 3, pp. 376–377, 2006.

[258] M. Bhasin, A. Garg, and G. P. S. Raghava, "PSLpred: prediction of subcellular localization of bacterial proteins," *Bioinformatics*, vol. 21, no. 10, pp. 2522–2524, 2005.

[259] R. Nair and B. Rost, "Mimicking cellular sorting improves prediction of subcellular localization," *Journal of Molecular Biology*, vol. 348, no. 1, pp. 85–100, 2005.

[260] O. Emanuelsson, S. Brunak, G. von Heijne, and H. Nielsen, "Locating proteins in the cell using TargetP, SignalP and related tools," *Nature Protocols*, vol. 2, no. 4, pp. 953–971, 2007.

[261] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman, "Basic local alignment search tool," *Journal of Molecular Biology*, vol. 215, no. 3, pp. 403–410, 1990.

[262] F. Chen, A. J. Mackey, C. J. Stoeckert, and D. S. Roos, "OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups," *Nucleic Acids Research*, vol. 34, pp. D363–368, 2006.

[263] G. E. Tusnády and I. Simon, "The HMMTOP transmembrane topology prediction server," *Bioinformatics*, vol. 17, no. 9, pp. 849–850, 2001.

[264] P. G. Bagos, T. D. Liakopoulos, I. C. Spyropoulos, and S. J. Hamodrakas, "PRED-TMBB: a web server for predicting the topology of β-barrel outer membrane proteins," *Nucleic Acids Research*, vol. 32, pp. W400–W404, 2004.

[265] N. K. Natt, H. Kaur, and G. P. S. Raghava, "Prediction of transmembrane regions of β-barrel proteins using ANN- and SVM-based methods," *Proteins*, vol. 56, no. 1, pp. 11–18, 2004.

[266] H. R. Bigelow, D. S. Petrey, J. Liu, D. Przybylski, and B. Rost, "Predicting transmembrane beta-barrels in proteomes," *Nucleic Acids Research*, vol. 32, no. 8, pp. 2566–2577, 2004.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 864029, 17 pages
doi:10.1155/2010/864029

Review Article

Bioinformatics in New Generation Flavivirus Vaccines

Penelope Koraka, Byron E. E. Martina, and Albert D. M. E. Osterhaus

Department of Virology, Erasmus Medical Centre, 3000 CA Rotterdam, The Netherlands

Correspondence should be addressed to Albert D. M. E. Osterhaus, a.osterhaus@erasmusmc.nl

Received 2 October 2009; Revised 21 December 2009; Accepted 2 March 2010

Academic Editor: Yongqun Oliver He

Flavivirus infections are the most prevalent arthropod-borne infections world wide, often causing severe disease especially among children, the elderly, and the immunocompromised. In the absence of effective antiviral treatment, prevention through vaccination would greatly reduce morbidity and mortality associated with flavivirus infections. Despite the success of the empirically developed vaccines against yellow fever virus, Japanese encephalitis virus and tick-borne encephalitis virus, there is an increasing need for a more rational design and development of safe and effective vaccines. Several bioinformatic tools are available to support such rational vaccine design. In doing so, several parameters have to be taken into account, such as safety for the target population, overall immunogenicity of the candidate vaccine, and efficacy and longevity of the immune responses triggered. Examples of how bio-informatics is applied to assist in the rational design and improvements of vaccines, particularly flavivirus vaccines, are presented and discussed.

1. Introduction

Flavivirus infections are among the most important viral pathogens of humans and animals, causing significant morbidity and mortality. They belong to genus Flavivirus of the family *Flaviviridae*. Flaviviruses are transmitted via arthropods and are classified into mosquito-borne and tick-borne flaviviruses. These groups of viruses are distinct not only by their modes of transmission, but also phylogenetically and by the clinical manifestations of the infections they cause [1]. Tick-borne flaviviruses are mainly associated with encephalitis and comprise a group of closely related viruses, which collectively belong to the tick-borne encephalitis (TBE) complex. The TBE complex comprises the Powassan virus, Louping ill virus, Kyasanur Forest disease virus, Omsk hemorrhagic fever virus, Langat virus and the Tick-Borne encephalitis viruses (TBEV). The TBEV species includes three subtypes which are all known to cause encephalitis in humans: Western European (previously known as Central European Encephalitis or CEE), Far Eastern (previously known as Russian Spring and Summer Encephalitis or RSSE), and Siberian (previously west-Siberian). Mosquito-borne flaviviruses are further divided based on phylogenetic and antigenic differences. The Japanese encephalitis serocomplex includes Japanese encephalitis virus (JEV), West Nile virus (WNV) and Saint Louis encephalitis virus (SLEV). The Dengue virus (DENV) group comprises an independent serogroup of four closely related but antigenically distinct serotypes. Finally, yellow fever virus (YFV), the prototype of flaviviruses, constitutes an independent serogroup [2].

The majority of flavivirus infections are manifested by mild acute febrile syndromes. A low percentage of infected individuals may develop severe neurological, hepatic and/or hemorrhagic disease with high mortality rates. JEV causes frequent outbreaks of meningo-encephalitis in Asia, affecting mainly children [3]. WNV may also cause severe outbreaks of meningo-encephalitis, with outbreaks confined to West Africa, Middle-East, and since 1999 also in North America. WNV is now spreading all over the Americas, posing a risk to millions of people [4]. DENV is endemic throughout tropical and subtropical areas of the world where more than 2.5 billion people are at risk of infection. Infection with DENV may result in development of hemorrhagic manifestations and/or shock in untreated patients [5]. YFV causes serious infections manifested by fulminant hepatitis and severe hemorrhagic disease. YFV still kills a considerable

number of people annually, despite the availability of an effective vaccine [6].

Flaviviruses are relatively small (approx. 11 kb) positive single-stranded RNA viruses coding for three structural (Capsid, C; Precursor membrane, prM; and envelope, E) and seven nonstructural proteins (Figure 1(a)). The single open reading frame (ORF) is flanked by 5′ and 3′ untranslated regions (UTR), the structures of which are important in viral replication [7]. The E protein is the major component of the virus bearing sites for attachment and fusion with the host cells and induction of immune responses. Several B- and T cell epitopes have been recognized on the E protein [8, 9]. The E protein of flaviviruses forms a head-to-tail dimer both in solution and on the viral membrane surface, with each monomer divided into three domains (D I, II, and III, Figure 1(b)). DI folds into an eight-stranded antiparallel β-barrel, containing about 120 residues and divided in three segments. The two long loops between these three segments form the dimerization DII, which contains the fusion peptide. DIII contains the carboxy-terminal 100 amino acids that form seven antiparallel β-sheets. In contrast to DI and DII, this domain represents an independently folding domain that can be expressed as a recombinant protein. Studies on the B cell repertoire upon flavivirus infection suggest that the human antibody response is predominantly directed to epitopes located in DII [10, 11]. However, antibodies specific to epitopes in DII have been shown to be weakly neutralizing, highly cross-reactive with other flaviviruses, and nonprotective in animal models. On the other hand, the potent, type-specific, neutralizing epitopes are located in the upper lateral surface of DIII [12–14]. Mutations in the DIII region have been associated with attenuated virulence or the ability of virus to escape specific neutralization, suggesting a role of DIII in receptor recognition [15]. Flavivirus infection triggers both innate and adaptive immunity in naïve individuals. In animal models, adaptive immune responses have been shown to play an important role in controlling primary WNV infection as exemplified by the high viral loads and high mortality observed in IgM deficient mice. It has been shown that the level of WNV-specific IgM at day four after infection has a prognostic value [16]. The role of IgM in controlling infection with other flaviviruses is however still unclear. Although T-helper (CD4+) and T-cytotoxic (CD8+) cells have been shown to play a role in controlling infection of mice with either WNV [17, 18] or DENV [19], the presence of antibodies is generally considered more relevant in terms of vaccine-induced protection.

For vector-borne viruses it is reasonable to assume that effective vector control would greatly reduce the morbidity of infection with such viruses. In the case of flaviviruses, mosquito control has at least in the long run, proven to be largely ineffective [20–22]. For example, efforts to eradicate the mosquito vectors of DENV during the 1970's were successful and resulted in the disappearance of the virus from the region. However, as soon as these programs were discontinued, *Ae. aegypti* (the main vector for urban transmission of DENV) re-infested the region, which coincided with the re-emergence of DENV. Therefore, vaccination against these pathogens may be the most efficient and effective way to control disease. In fact, one of the most effective vaccines ever developed is the live-attenuated vaccine against YFV [23]. This vaccine was developed in the late 1930's by Theiler and co-workers and although the correlates of protection even today are not completely clear, it has been used to immunize and protect more than 400 million people against yellow fever. Also licensed vaccines against JEV and TBEV infections in humans exist. Routine childhood immunization against JEV has eliminated the disease from many Asian countries, whereas the disease continues to cause devastating outbreaks in countries where the vaccine is not used. The mouse brain-derived inactivated vaccine is relatively expensive to produce and the lack of long-term immunity and risk of allergic reactions makes large-scale vaccination with this vaccine, especially in Asia where it is most needed, not feasible. Cell culture-derived inactivated and live-attenuated JEV vaccines have been extensively used in China, but purity of these vaccine preparations represents the main hurdle for approval of these vaccines in Western countries. More recently a cell-derived, alum-adjuvanted JEV vaccine was licensed for use in adults in USA and some EU member states. However, the longevity of the immune response to this vaccine is not yet established. For other important flavivirus such as WNV and DENV, no vaccine is available yet.

This review aims to give a brief overview of flavivirus vaccine candidates and discuss how bio-informatics can assist in the rational design and improvements of vaccines. In doing so we will discuss examples of how bioinformatics has been or could be used in the development of novel generations flavivirus vaccines. It is beyond the scope of this review to summarize the recent advances in immunoinformatics for the prediction of T and B cell epitopes and the interaction of proteins with the immune system (the reader is referred to the excellent review of Brusic et al. [24] and the references therein).

2. Rational Design of Flavivirus Vaccines

We have entered a new era of vaccine development, where rational design of candidate vaccines has replaced the trial-and-error approach. Undoubtedly, the trial-and-error approach has been successful in the past as exemplified by the success of the YFV vaccine, one of the most effective vaccines ever developed. Rational design of vaccines is a step-wise approach as depicted in Figure 2: (1) identify target antigen(s), (2) determine the vaccine platform, (3) optimize factors for gene expression, (4) produce and characterize vaccine, (5) test safety, immunogenicity and efficacy in animal models. The recent advances in molecular biology and bio-information technology have contributed significantly to our understanding of viral structure, viral replication, attenuation, and determinants of pathogenicity. Understanding these aspects are crucial for identifying target proteins. Several bio-informatic tools are available that can help in the process of antigen identification. The majority of bioinformatic tools used in computational vaccinology however, focus on antigen presentation and processing, with the ultimate goal to map T and B cell epitopes (step 5). Although

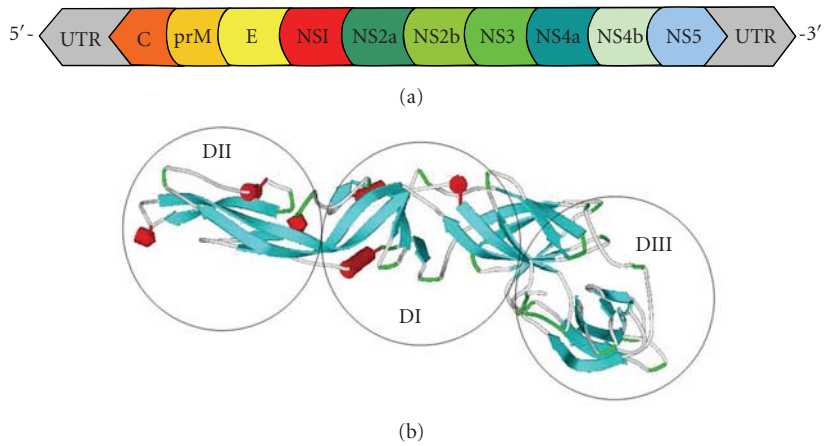


Figure 1: (a) Flavivirus genomic RNA encoding one long ORF cleaved co- and post-translationally into three structural and seven nonstructural proteins. The viral genome is flanked by 5′and 3′UTRs which play an important role in virus transcription and replication. (b) Three dimensional structure of the monomeric form of the E protein of DENV-2. Each monomer is divided into three discernable domains (DI, II and III). Serocomplex and group specific cross reactive epitopes are mainly located on DI and II. DIII is the receptor binding domain and contains mainly type specific epitopes. The potent neutralizing epitopes are located at the lateral site of DIII.

the mapping and identification of highly immunogenic epitopes on target antigens is important, several additional steps need to be considered when embarking on rational design of new generation vaccines against these viruses (step 2 and 3). Common problems encountered during development of vaccines include low yields of proteins and consequently low immunogenicity.

3. Inactivated and Subunit Candidate Vaccines

The great impact that flavivirus infections have on public health and the fact that prevention is difficult to achieve through vector control have made the development of flavivirus vaccines essential for long-term control and elimination of these infections, and reduction of the associated morbidity and mortality. The currently available vaccines against YFV, JEV and TBEV have proven to be quite successful. Efforts are ongoing to develop safe and effective vaccines against the other medical important flaviviruses. In addition efforts aiming to improve the existing ones are ongoing. Almost all vaccine platforms have been used in the development of candidate flavivirus vaccines, ranging from whole inactivated, live-attenuated, subunit, DNA and vectored candidate vaccines (summarized in Table 1). The currently licensed vaccines against JEV and TBEV contain formalin-inactivated whole virus, purified from mouse brains [25] (JEV) or cultured cells (JEV, TBEV). Studies to compare the safety and immunogenicity of candidate vaccines based on cell-derived inactivated JEV show promising results [26, 27]. This formalin-inactivated JEV vaccine is now licensed in the USA and EU for use in adults. However, the duration of immunity and the need for booster vaccination needs to be further investigated [28]. There are several concerns with the use of inactivated vaccines. They are expensive, require two or three doses to achieve protective efficacy, and perhaps more importantly, require further booster doses to maintain immunity [29]. Therefore, large scale use of inactivated flavivirus vaccines for humans may prove problematic due to the need of an adjuvant to potentiate the immune response and to provide long-lasting immunity. Furthermore, formalin-inactivation may pose other safety concerns as has been shown in case of respiratory syncytia virus and human metapneumo virus [30, 31], although extensive use of formalin-inactivated JEV and TBEV vaccines does not corroborate this concern.

Subunit flavivirus vaccines, expressing recombinant E protein, combination of prM-E or DIII alone, have been shown to be effective and immunogenic in animal models [32, 33]. The minimalistic approach of using DIII alone instead of full length E protein is based on the observations that DIII contains epitopes responsible for type specific neutralization, whereas DI and DII contain epitopes responsible for cross-neutralization and/or sensitivity to disease enhancement [10, 11]. Because DIII is poorly immunogenic, the use of an adjuvant is necessary for efficient induction of immune responses.

A common problem of inactivated and subunit vaccines is the poor immunogenicity when compared to live-attenuated vaccines. Given the safety issues often associated with the use of live-attenuated vaccines, modern vaccinology also focuses on the improvement of the immunogenicity of inactivated and subunit vaccine candidates. Until recently aluminium hydroxide (Alum) was the only adjuvant registered for human use. Alum triggers antibody responses and T-helper 2 responses, but no CD8+ T cell immune responses. Increasing understanding of the mechanisms that leads to protective versus detrimental immune responses will be instructive in the rational choice and design of novel adjuvants. For example, the activation of several Toll like receptors (TLR) has been shown to stimulate innate and orchestrate adaptive immune responses via stimulation of antigen presenting cells. Several studies have shown the potential of TLRs agonists as effective adjuvants [34–36]. CpG oligodeoxynucleotides were shown to be potent

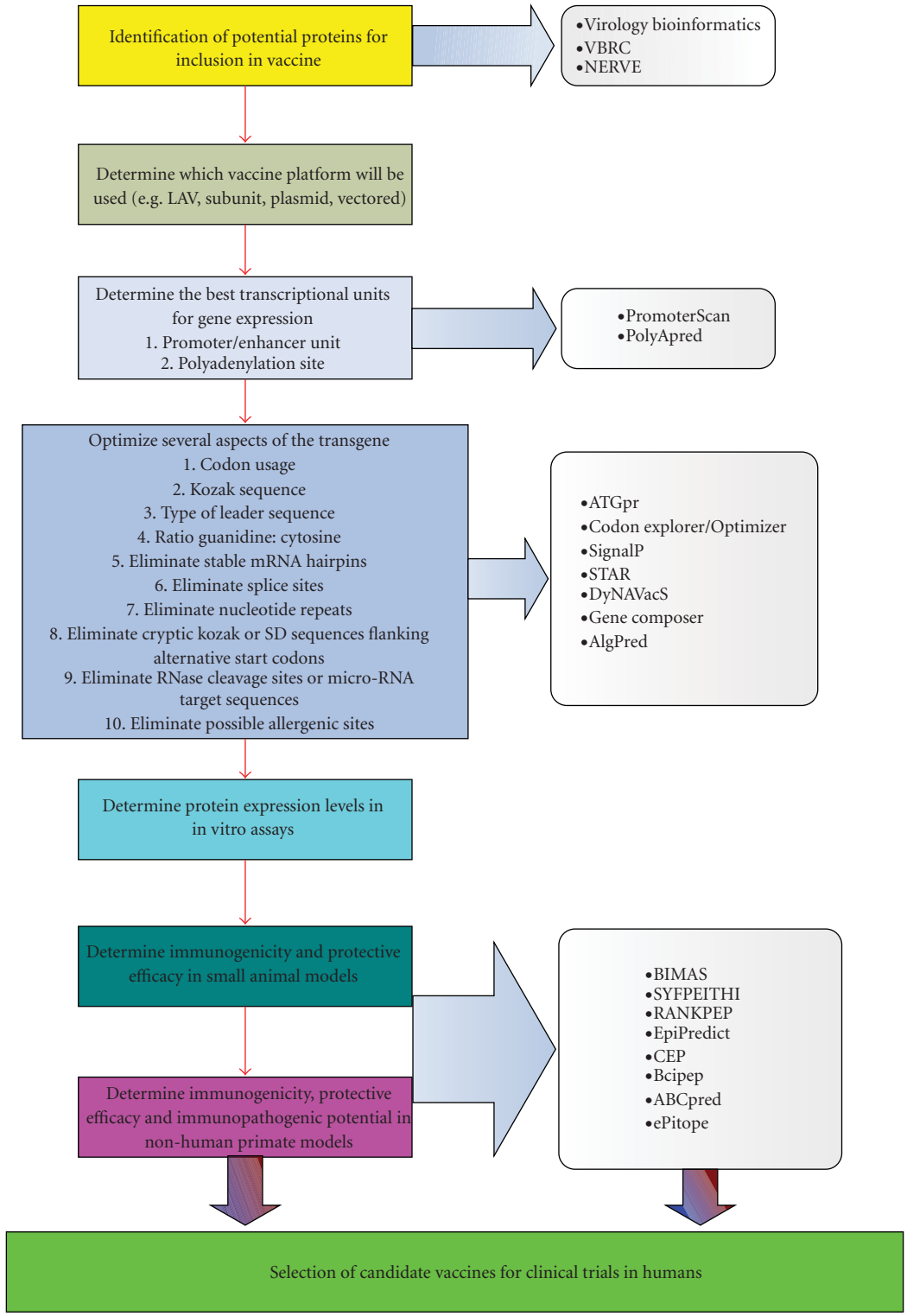


Figure 2: Steps involved in the rational design of vaccines. A number of parameters should be taken into account from choosing the right immunogen until the selection of candidate vaccines for clinical trials. Several bioinformatics tools can be applied in this process to assist in the improvement of antigen selection, maximizing expression, determination of immunogenicity and selection of candidate vaccines.

Table 1: Summary of flavivirus candidate vaccines based on different vaccine platforms.

Type of vaccine	Virus	Antigen	Comments
Inactivated/killed	JEV	Whole virus	Licenced for human use
	DENV-2	Whole virus	Pre clinical studies in mice and non-human primates
	WNV	E, prM, whole virus	Low immunogenicity, use of adjuvant necessary
Subunit	DENV-2	D III, NS1, E protein	Immunogenicity tested only in mice
	DENV-4	C-M-E-NS1	Immunogenicity tested only in mice
	WNV	D III, E protein, NS1	Low immunogenicity, only tested in mice
	JEV	prM-E	Immunogenicity tested only in mice
DNA	DENV	prM-E, NS1	Low immunogenicity
	JEV	E	Low immunogenicity
	WNV	E, prM, C	Low immunogenicity
Live attenuated	YFV		Licenced for human use
	Mono- and tetravalent DENV		Phase I, and II clinical trials ongoing
	JEV		Phase I and II clinical trials conducted
	WNV		
Vectored	WNV	E, prM	Vectors used: MVA, MV, AdV, pox-, alphaviruses

stimulators of TLR-9 and able to significantly enhance both the humoral and cellular immune responses to several viruses in mice [37, 38]. Furthermore, the use of CpG resulted in substantial reduction of the antigen dose needed in several models. Several programs (e.g., DyNAVacS, Table 2) allow optimization of CpG content in the gene of interest, which may enhance vaccine immunogenicity. An interesting development in this field is the availability of a database (TollML) to retrieve and deposit agonists for the different TLR (Table 2). In this database, the known structural motifs of TLRs are deposited allowing for structural modelling of other TLRs. Furthermore the ligands of TLRs can be easily retrieved. Elucidation of the crystal structure of TLRs and improvement of powerful bioinformatics tools for protein homology and *ab intitio* modelling will allow structure-based design of safe and effective TLR agonists that could be subsequently used as adjuvants for vaccines [39–41]. The potential of such adjuvants was recently demonstrated in a study that reported how the poorly immunogenic DIII of WNV was remarkably more immunogenic in mice when fused with the TLR-5 agonist flagellin [42].

4. Live-Attenuated Vaccine Candidates

In light of the safety issues that may be related with the use of current experimental adjuvants, the concept of using live-attenuated or vectored vaccines is more and more preferred over that of inactivated or subunit vaccines. The current vaccine against YFV is a live-attenuated whole virus vaccine, which has been shown to be safe and highly effective. The advantage of live-attenuated vaccines is that both antibody and T cell responses are induced, although, there may be risks related to the use of live-attenuated vaccines [43]. For instance, the efforts for the development of a successful vaccine against DENV have mainly focused on the attenuation of the four DENV serotypes through several passages in cell culture and use of these viruses in monovalent or multivalent formulations. The leading hypothesis to explain pathogenesis of DENV infections is the antibody-dependent enhancement of infection (ADE), which suggests that cross reactive subneutralizing antibodies may enhance the infection of target cells, rather than protect them against infection [5]. Since the borderline between protective immunity and disease enhancing immunity is not well defined, the most effective vaccine against DENV would be one that confer solid immunity against all four serotypes. A live-attenuated tetravalent DENV vaccine candidate has been shown to be immunogenic in flavivirus naïve and immune nonhuman primates [44, 45], whereas phase I and II clinical trials in humans have shown promising results [46–48]. The molecular basis for attenuation of wild type virus through passage on cell cultures or in animals is usually not understood. A more rational way of generating live-attenuated vaccines is by understanding correlates of virulence. Several algorithms have been designed and are available either free online or as commercial software programs to aid in the prediction of secondary structures of RNA (Table 2). The secondary structures of the 5′ and 3′-UTRs of several flaviviruses, including YFV, DENV, and TBEV have been predicted and conserved structural elements that are important in viral replication were identified [49–52]. In a series of studies with DENV-4, the secondary structure of wild type viruses were predicted and subsequent studies were designed to create deletion mutants lacking structural elements that could influence viral replication. It was indeed shown that when a 30-nucleotide region was deleted, the mutant virus was attenuated and had lost its ability to be transmitted by *Ae. aegypti*. Subsequently it was shown that these mutant viruses were able to induce strong immune responses in rhesus macaques and in humans, similar to what is seen with wild type viruses [53, 54]. The success of the DENV-4 Δ30 was followed by similar results using a DENV-1 Δ30 mutant virus in vaccination trials of humans and nonhuman primates [55]. In the same line of

Table 2: Selection of servers and programs that could assist in the rational design of vaccines.

Server	Web address	Server description	Availability
ORF-FINDER	http://www.ncbi.nlm.nih.gov/projects/gorf/	Program identifies open reading frames in a genome.	Free online
GeneMark	http://exon.biology.gatech.edu/	A set of gene prediction servers for prokaryotes, eukaryotes and viruses	Free online
NERVE	http://www.bio.unipd.it/molbinfo/	In silico prediction of vaccine candidates from complete proteomes of bacterial pathogens	Free for download
Gene Composer	http://www.genecomposer.net/	Gene Composer is a software program for rational design or optimization of genes	Free demo for 1 year
PromoterScan	http://www-bimas.cit.nih.gov/molbio/proscan/	Predicts Promoter regions based on scoring homologies with putative eukaryotic Pol II promoter sequences	Free online
ATGpr	http://flj.hinv.jp/ATGpr/atgpr/index.html	A program for identifying the initiation codons in cDNA sequences	Free online
PolyApred	http://www.imtech.res.in/raghava/polyapred/	PolyApred is a SVM based method for the prediction of polyadenylation signal in human DNA sequence.	Free online
NetGene2	http://genome.cbs.dtu.dk/services/NetGene2/	Predictions of splice sites in human	Free online
SignalP	http://www.cbs.dtu.dk/services/SignalP/	Signal peptide and cleavage sites in prokaryotic and eukaryotic sequences	Free online
Optimizer	http://genomes.urv.es/OPTIMIZER/	A web server for optimization of codon usage of a DNA sequence to increase its expression level.	Free online
CodonExplorer	http://bmf2.colorado.edu/codonexplorer/index.psp	Online tool for analyzing GC content and codon usage frequency	Free online, registration required
STAR	http://biology.leidenuniv.nl/~batenburg/STROrder.html	STAR simulates the folding pathway of RNA and is able to predict formation of pseudoknots formation	Commercially available
VIENNA package	http://rna.tbi.univie.ac.at/cgi-bin/RNAfold.cgi	RNAfold predicts secondary structure of single stranded RNA or DNA based on minimum free energy predictions	Commercially available
mFOLD	http://mfold.bioinfo.rpi.edu/	Prediction of RNA and DNA secondary structure based on thermodynamic methods	Free online
DyNAVacS	http://miracle.igib.res.in/dynavac/	An integrative tool for optimized DNA vaccine design	Free onliune
AlgPred	http://www.imtech.res.in/raghava/algpred/	Algpred allows prediction of allergens based on similarity of known epitope with any region of protein	Free online

Table 2: Continued.

Server	Web address	Server description	Availability
BIMAS	http://www-bimas.cit.nih.gov/molbio/hla_bind/	Server for prediction of MHC-I epitopes	Free online
EpiPredict	http://www.imtech.res.in/raghava/mhc2pred/	Software to predict HLA-class II restricted T cell epitopes and ligands	Free online
SYFPEITHI	http://www.syfpeithi.de/	Server for prediction of MHC-I and MHC-II epitopes	Free online
RANKPEP	http://bio.dfci.harvard.edu/RANKPEP/	Server for prediction of MHC-I and MHC-II epitopes	Free online
CEP	http://202.41.70.74:8080/cgi-bin/cep.pl	Server for prediction conformational B cell epitopes	Free online
Bcipep	http://www.imtech.res.in/raghava/bcipep/pep_src.html	Bcipep is a database with a collection of immunogenic B cell peptides	Free online
ABCpred	http://www.imtech.res.in/raghava/abcpred/	Server for prediction conformational B cell epitopes	Free online
ePitope	http://www.epitope-informatics.com/	Server for prediction linear B cell epitopes	Online, commercial
TollML	http://tollml.lrz.de/	A user-friendly database to retrieve sequence and structural data as well as ligands for TLRs.	Free online
Virology Bioinformatics	http://www.bioinfo.de/isb/2008/08/0008/	A collection of bioinformatic tools for virology research	Free online
VBRC	http://athena.bioc.uvic.ca/	VBRC provides access to viral genomes and a variety of bioinformatic tools for comparative genomic analyses	Free online

research, DENV-2 and DENV-3 Δ30 mutants were found to be under-attenuated [56, 57]. Using prediction algorithms for secondary structure of RNA a larger region in the 3′UTR of DENV-3 was defined as possibly associated with attenuation and the new mutant virus that was engineered based on these predictions showed a much more promising phenotype in preclinical studies [58]. Similarly, prediction of RNA secondary structure of YFV strains revealed differences between attenuated vaccine strains and wild type viruses. In particular, the conserved long stable hairpin (LSH, a crucial structural element for virus replication) was shorter in vaccine strains than in wild type viruses [50]. As depicted in Figure 3, a 30-nt deletion in the 3′UTR of DENV-3 results in altered secondary structure of the RNA and possibly loss of structural elements necessary for replication. These examples demonstrate that *in silico* simulation of RNA secondary structure of viral genomes could be a first step in rational design of candidate vaccines. Similarly, it has been shown that the nucleotide sequence of the attenuated YFV vaccine strain differs only 68 nucleotides (translated in 32 amino acid differences) from the parental wild type virulent strain, suggesting that some of these mutations might be associated with virulence and attenuation [59]. The advances in reverse genetics technology allow construction of recombinant viruses with the desired mutations. Such mutant viruses are attractive attenuated vaccine candidates, which could also be used as vectors to express genes of interest, or serve as backbones for chimeric vaccines.

5. DNA and Vectored Vaccine Candidates

A common problem encountered during the development of DNA and some vectored vaccines is the low immunogenicity of such vaccine candidates. Variation in transgene stability and level of protein expression are typically attributed to properties of the promoter elements and recombinant gene.

5.1. Transcriptional Regulator

5.1.1. Promoter. Molecular biology has made a significant contribution to the development of new generation vaccines. Among the first steps in development of novel generation vaccines is the cloning and expression of genes that encode

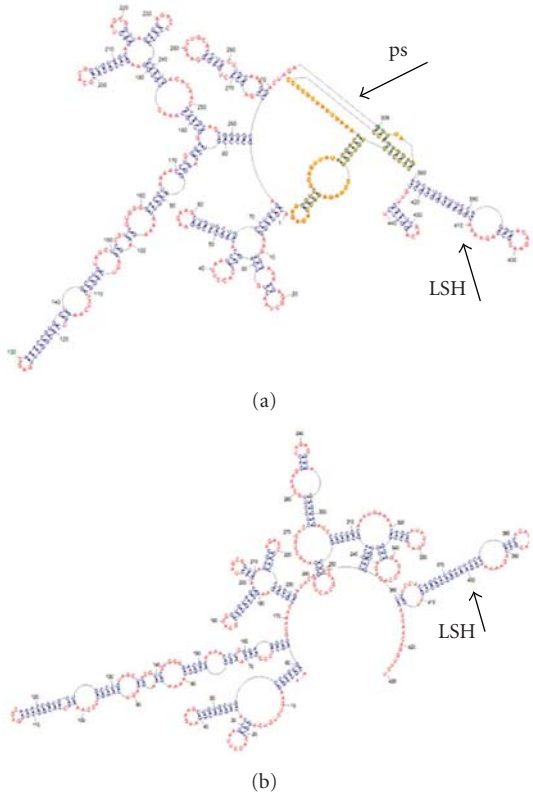


Figure 3: Prediction of RNA secondary structure of the 3′UTR of DENV-3 strain H87 (DENV-2 Jamaica strain) using the GA algorithm included in the STAR software package. In the top panel is depicted the predicted structure of the naturally occurring strain (shown with arrows the predicted pseudoknot and the conserved LSH structures), bottom panel: predicted structure after *in silico* introduction of a 29-nt deletion in the proximal part of the UTR. The LSH is preserved but the pseudoknot is lost after deletion. LSH: long stable hairpin, ps: pseudoknot.

highly immunogenic proteins of the virus. Several plasmids have been designed for cloning and expression of both prokaryotic and eukaryotic genes (Figure 4(a)). Obviously, the choice of plasmid for cloning of the genes of interest depends on the type of vaccine platform which is envisaged: subunit, DNA, or vectored. For example, in case of subunit vaccines, it is imperative for the antigen used in the vaccine to resemble its native form as much as possible. Therefore, it is important to realize that posttranslational modifications, which differ between prokaryotes and eukaryotes may affect the formation of B cell epitopes. For DNA and vectored vaccines it is well appreciated that antigen expression levels, determined at both the transcriptional and translational level, affect immunogenicity and efficacy of vaccines. Promoter and enhancer elements affect the levels of messenger RNA (mRNA) available, which have a significant impact on the protein expression levels. The human cytomegalovirus (CMV) immediate-early and the simian virus 40 (SV40) early promoters are the most commonly used transcriptional regulators. Several studies have shown that virus-derived promoters, including the CMV, SV40, and Rous sarcoma virus (RSV), are stronger than other eukaryotic promoters [60] in driving gene expression. In particular the CMV promoter has been shown to be a potent regulator of transcription compared to most other viral promoters [60]. The sensitivity of the CMV promoter to inactivation by cytokines and methylation, especially in muscle tissues [61–63], in addition to the regulatory problems that may be associated with the use of transcriptional elements derived from pathogenic viruses, have prompted the search and use of other potent promoters. Promoter regions can be predicted *in silico* in a given sequence and/or vector (Table 2). To this end, mammalian promoters including β-actin [64], human Muscle Creatine-Kinase (MCK) [65], human elongation factor 1α (EF-1α) [66], human phosphoglycerate kinase-1 (PGK) [67], major histocompatibility class II [68], human telomerase reverse transcriptase (TRT) [69–71], and tissue plasminogen activator (tPA) [72] promoters have been tested. It is important to realize that promoters may regulate expression of genes in all cell types (ubiquitous) or they may express tissue-specific activity. Therefore, careful consideration must be given to the selection of the promoter driving transgene expression. For instance, while CMV promoter drives transient, but high-level expression of proteins, RSV promotes high-levels of gene expression for a longer time. Therefore, the eukaryotic promoter EF-1α, which provides long-term and high-level gene expression [73] may represent an alternative to viral promoters like RSV. Their activity may be increased by the addition of introns upstream the gene sequence as exemplified by CMV driven vectors [74–77]. This increase in protein expression has been attributed to the presence of enhancer elements in the intron sequence and increased rate of polyadenylation and RNA splicing [77]. However, inclusion of an intron sequence must be applied with caution as it may also lead to aberrant splicing [78]. The use of programs that predict splicing (Table 2) in the context of promoter and intron could help in the rational design or choice of transcriptional units that do not result in aberrant gene splicing. The use of synthetic promoter/enhancer sequences from different sources has also proven to be promising. For example the CAG promoter, a chimera of a CMV enhancer, a chicken β-actin promoter and a rabbit β-globulin splicing site was shown to induce strong expression of genes in several tissues [79–81]. Although different promoters have been used in the design of flavivirus candidate DNA vaccines, the effect of different promoters on immunogenicity and efficacy has not been investigated [82]. However, the choice of promoters has been shown to affect immunogenicity of DNA vaccines against human immunodeficiency virus type 1, human hepatitis B virus, and herpes simplex virus type 1 [83–85], and therefore should be taken into account when designing candidate vaccines.

5.1.2. Polyadenylation. Polyadenylation (addition of a polyA tail) of eukaryotic mRNA has several functions: (i) it aids the export of the mRNA from the nucleus, (ii) it protects the molecule from enzymatic degradation in the cytoplasm, (iii) it directs termination of transcription, and (iv) enhances translation, although poly(A) is not necessary for translation of all mRNAs. In principle, polyadenylation occurs in three

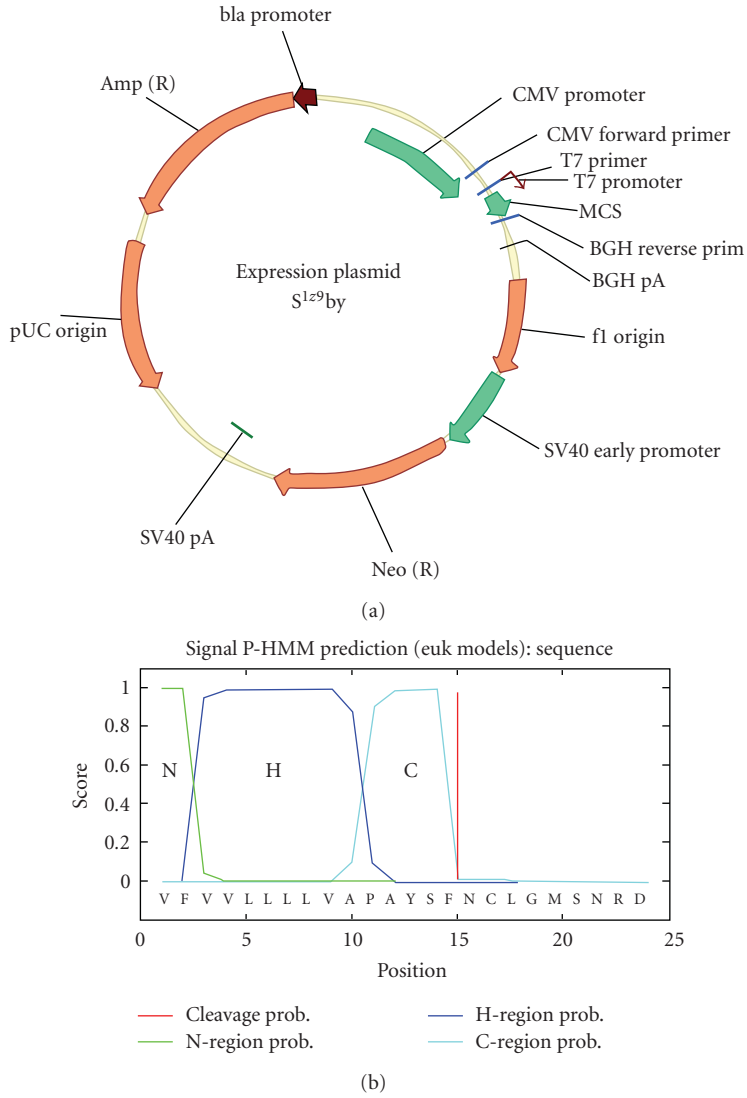


Figure 4: (a) A basic expression plasmid showing the CMV promoter, multiple cloning sites (MCS) and a BGH polyadenylation site, (b) Predicted signal peptide sequence for the WNV E protein. The signal peptide probability is determined by the positive polarity of the n-region, the hydrophobicity of the h-region and the uncharged amino acids occupying positions −1 and −3 in the c-region. The cleavage site is located in the c-region and is determined by the physicochemical characteristics and length of the n- and h-regions.

stages: polyadenylation site choice, cleavage of the pre-mRNA, and addition of the poly(A) tail to the newly formed 3′-end [86]. The first step, polyadenylation site choice, can be defined as the preparation of the pre-mRNA to allow efficient and accurate cleavage [87]. Any mutation of the pre-mRNA sequence elements involved in polyadenylation site choice may result in the inefficient polyadenylation of the pre-mRNA, limited nuclear export and decreased translation of the protein [86, 88, 89]. Therefore, polyadenylation site choice is an important first step in polyadenylation and is essential for optimal gene expression. The poly(A) site of SV40 is highly efficient and frequently used as a late polyadenylation signal in many DNA plasmids. It contains efficiency elements both upstream and downstream of the AAUAAA region, and the downstream region contains three defined elements, two U-rich elements and one G-rich element, instead of the single U- or GU-rich element found in most polyadenylation signals [90]. The bovine growth hormone polyadenylation signal (BGH) is another highly efficient and frequently used poly(A) signal in DNA vaccines. It is known that the BGH poly(A) pre-mRNA forms an extensive hairpin loop secondary structure at the 3′-UTR [91], possibly explaining the more efficient polyadenylation sequence. However, although both SV40 and BGH poly(A) signals have been shown to be highly efficient, depending on the background of the constructs used, they may negatively influence plasmid stability. The issue of promoter and poly(A) signal becomes particularly important when designing DNA vaccines. Inefficient nuclear delivery of plasmid DNA remains a major bottleneck in DNA vaccination [92],

which may be increased by addition of a nuclear localization signal [93]. However, nuclease degradation of plasmid DNA after administration and during trafficking to the cell nucleus represents one of the main reasons of this inefficiency [94]. In this respect it has been shown that there is a correlation between the number of purine-rich regions in the poly(A) site and the susceptibility of the plasmid to nuclease degradation [95, 96]. This indicates the importance of this region in conferring plasmid stability in addition to determining efficiency of post-transcriptional modification [95]. BGH has been shown to have the highest frequency of purine-rich regions compared to the SV40 poly(A) sequence and hence renders plasmids more susceptibility to degradation. Consistently, modification of the poly(A) site was shown to increase plasmid stability, although it negatively affected protein expression levels [95]. Taken together, plasmid resistance should be taken into account during DNA vaccine design and more efforts should be deployed to understand how rational modification of the poly(A) signal sequence may increase plasmid stability while maintaining the same levels of protein expression.

5.2. Translational Regulator

5.2.1. Codon Usage. It is generally appreciated that not all codons within a synonymous codon family are used at the same frequency, a phenomenon called "codon usage bias". Optimal codons are determined by high concentrations of particular transfer RNAs (tRNAs) and strong codon-anticodon interactions [97, 98]. In general, closely related organisms use similar codons compared to taxonomically distant organisms [99–101]. In addition to the codon usage bias between different organisms, there are substantial differences between genes of the same organism [102, 103]. Certain codons are preferentially used in highly expressed genes [98], while use of rare codons may represent a suppressive mechanism of gene expression under inappropriate conditions. Codon usage influences translation initiation [104, 105], protein folding [106], and consequently protein expression levels. Therefore, codon usage adaptation of the target gene to those of the expression host is one way to enhance translational efficiency. Methods for optimising genes for high expression in prokaryotes, yeast, plants, and mammalian cells are becoming increasingly sophisticated and well-established in the field of vaccinology (Table 2). One measure of codon quality is the Codon Adaptation Index (CAI), a measure for the relative adaptiveness of the codon usage of a gene towards the codon usage of highly expressed genes. The index uses a reference set of highly expressed genes from a species to assess the relative merits of each codon, and a score for a gene is calculated from the frequency of use of all codons in that gene [107]. Optimizing codon usage has been shown to increase immunogenicity of candidate vaccines against both viruses and bacteria [108–114]. The increased immunogenicity is the result of improved protein transcription and translation rate, which leads to stimulation of stronger antibody and T cell responses. In addition, the optimized GC content contributes to a better induction of T cell responses through the TLR-9 pathway. Several programs are now available to analyse and adapt codon usage of a transgene to that of the host (Table 2).

5.2.2. Kozak and Leader Sequences. Sequences surrounding the start codon (AUG) within the mRNA, the so-called kozak sequences, influence the quality and quantity of the synthesized protein. The optimal context for initiation of translation in mammals is GCCRCCAUGG [115]. Optimal kozak sequences contain a high frequency of A at the -3 position and G at positions -9, -6, and $+4$, and A or C are predominantly found at positions -5, -4, -2, and -1 [82, 116]. Although it is recommended to provide viral genes with a kozak sequence, most DNA and vectored vaccines have not included this sequence in their cloning strategy. Prokaryotic genes possess an analogue of the kozak sequences, the Shine-Dalgarno (SD). Analysis of several E-coli genes indicates that SD sequences are present in virtually all genes [115]. Therefore, expression of genes in prokaryotic and eukaryotic systems may benefit from the insertion of respectively an SD and kozak sequence around the translation initiation codon.

Leader sequences encode signal peptides that play an important role in targeting secretory and membrane proteins in both prokaryotes and eukaryotes to the right compartment. Leader sequences are usually found at the N-terminal side of proteins, although they may also be located within a protein or at its C-terminal end [117]. Signal peptides are divided into three different regions: N-terminal (n), hydrophobic (h), and cleavage (c) (Figure 4(b)). The h-region, which is the most essential part responsible for targeting and membrane insertion, comprises 6–15 amino acid residues. The c-region consists of small uncharged residues at positions -3 and -1, which determines the site and probability of signal cleavage. Cleavage is more likely to occur when an amino acid with a short side-chain is present at the -1 position and no charged amino acids are present at the -3 position (reviewed in [118]). In addition, the amino acid composition of the n-region as well as the length of the h-region may affect cleavage probability [118]. The choice of the leader sequence may also influence the behaviour of proteins as type I or type II as well as the rate of protein folding within the cell. In this regard, it is interesting to note that the codon usage of signal peptides plays an important role in correct folding of proteins [119, 120]. The use of signal sequences have been shown to affect vaccine immunogenicity and is therefore an important parameter in rational design of vaccines [121–124]. In this respect, signal peptide can have an "early" or "delayed" effect on protein maturation and secretion in that it can either cleave the signal peptide soon after translocation in the ER or delay its cleavage. It is thus of paramount importance to use custom designed vectors that carry the gene of interest for candidate DNA or subunit vaccines and with the use of bioinformatic tools determine the presence and potency of kozak and leader sequences.

6. Flavivirus DNA and Vectored Vaccines

Plasmids have been studied as potential candidate vaccines against several pathogens. Immunogenicity of inactivated, subunit, or vectored vaccines may be compromised in areas

where several flaviviruses cocirculate due to interference of pre-existing crossreactive antibodies with the vaccine. DNA vaccines may therefore represent an attractive alternative for use in flavivirus endemic areas, since these are not sensitive to neutralization by flavivirus cross-reacting antibodies. However, most DNA vaccine candidates have not proven sufficiently immunogenic, although many have provided protection against disease and death in animal models [125, 126]. During virus replication, prM is essential for protection of the E protein against denaturation due to low pH and some believe it is crucial for correct folding of the E protein. Consequently, most candidate DNA vaccines encode the complete prM and E genes of flaviviruses. It is difficult to explain why certain vaccines proved superior to others, since different vaccines were constructed using different plasmids. Comparison of different JEV DNA vaccines revealed differences in kozak sequences surrounding the start codon [82], and differences between signal peptides of the prM. These differences have been proposed to affect the immunogenicity of the different candidate vaccines [82]. The major difference between the different signal peptide sequences are the length and composition of the n-region. Signal peptides with a short n-region and with no positively charged amino acids may promote a type I orientation, which may affect processing efficiency and topology of the expressed prM and E proteins, and consequently immunogenicity [127, 128]. The prediction of optimal signal peptides and use of codon optimization programs have been exploited in the development of WNV candidate vaccines, which proved to be effective in animal models and in humans [72, 129, 130]. This is the first flavivirus DNA vaccine to reach phase I clinical trials, showing promising safety and immunogenicity results [130]. The ability of such vaccines to induce neutralizing antibodies can be explained by the formation of virus-like particles with morphology similar to virus particles. The choice of complete versus truncated E proteins may also determine whether proteins are secreted or will remain membrane-anchored and thus determine vaccine immunogenicity [131, 132]. Although DNA vaccines have not been shown to be very immunogenic, proof-of-principle for DNA vaccines has been validated with a number of flavivirus vaccine candidates in a variety of animal models. The concept of DNA vaccines as a generic flavivirus vaccine platform, especially in endemic areas, still remains viable and attractive. Many strategies are being explored to enhance the immunogenicity of DNA vaccines. The easiest and most straightforward approach that can be quickly transitioned to a clinical trial setting is vaccine delivery by a needle-free jet injector [133]. This approach has shown much potential and is the first and most forward way to enhance immunogenicity of DNA vaccines. Other approaches include the co-expression of cytokines [134, 135] or inclusion of sequences that enhance MHC class I and II antigen presentation [136–138]. Another promising DNA vaccine was recently described based on a single-round infectious particle system [139], which resulted in enhanced immunogenicity and efficacy against WNV.

Several viral vectors have been developed and evaluated as candidate vaccines against flaviviruses, including poxviruses [140–143], adenoviruses [144–146], measles virus [147], alphavirus [148–150], and vesicular stomatitis virus [151]. Modified Vaccina virus Ankara (MVA) is an orthopox virus vaccine vector of particular interest. MVA has been shown to be immunogenic and safe, even in severely immunocompromised animals [152]. The safety profile of MVA can be explained by the fact that the virus establishes only one round of replication, a property that is stably maintained over several passages. Furthermore, MVA and other viral vectors like the complex adenovirus (CAdVax), can harbour multiple genes, rendering them more suitable candidates for developing pan-flavivirus vaccines [153]. A variety of MVA promoters such as PmH5, P7.5, P11K, Psyn, PsynII, and Pk1L have been used for regulation of gene expression. Optimization of target DNA sequences and use of strong but not overexpressing promoter systems have been applied in the development of several MVA vaccine candidates. Viral vectors expressing prM and E, E alone or DIII have all been shown to be effective in animal models [141, 144–146, 151, 154, 155]. Most vectored vaccines developed against flaviviruses were not optimized for polyadenylation, codon and promoter usage, providing a partial explanation for the often moderate immunogenicity observed in many cases. Similar to the immunogenicity of DNA vaccines, that of vectored vaccines can be improved by the careful choice and reconfiguration of promoters [62, 66, 156], optimizing gene codon usage, inclusion of a consensus Kozak sequence and this all by exploiting the power of bioinformatics (Table 2).

Another promising approach for vectored vaccines is the chimeric vaccines developed using the YFV-17D backbone. Chimeric WNV candidate vaccines based on a related flavivirus vector, the YFV vaccine strain 17D, are among the most promising vectored WNV vaccine candidates to date [157]. A vaccine based on this technology is now licensed for use in horses [158] and a similar one has undergone phase I and phase II clinical trials in humans. These chimeric vaccines exploit the safety record of the yellow fever virus vaccine strain 17D in healthy individuals, and the immunogenicity of this vector that has already been shown in experimental animals and horses as well as for a vectored vaccine against other flaviviruses, such as JEV and DENV [159–161].

7. Concluding Remarks

Despite numerous and continuous efforts to develop safe and effective vaccines against flaviviruses causing disease of major medical importance, there are still areas of vaccine research that need to be explored in this respect. The traditional trial-and-error approach that has dominated the field of flavivirus vaccine development until today may benefit from recent advances in biotechnology and bioinformatics allowing for a more rational approach in vaccine design. The steps that should be taken to optimally exploit bioinformatics in the rational design of vaccines are shown in Figure 2.

Minimalistic approaches using certain genes or even parts thereof associated with high immunogenicity might replace whole virus-based candidate vaccines. There are several advantages in using subunit over whole virus

vaccines. A notorious example from the flavivirus field is the immune enhancement theory that is frequently stated to explain the pathogenesis of severe DENV infections (ADE, see above). In light of this hypothesis, an ideal DENV vaccine should elicit strong antibody responses to type specific and not to cross-reactive epitopes. In this regard, the DNA and vectored vaccines tested till now did not induce better immune responses than tetravalent live-attenuated vaccines. Taking advantage of, for instance, the optimization of gene and protein expression and development of safe and potent adjuvants, seems to be imperative in the rational design of DENV vaccines. Finally, bioinformatics should play a prominent role in the process of gene optimization as well as the downstream processes of testing and selection.

Considering the advantages and disadvantages of inactivated, live-attenuated, subunit, and DNA vaccines, vectored vaccines should be considered among the most promising ones. The advantage of vectored vaccines over the others is that long-lived B and T cell responses may be induced. Furthermore, vectored vaccines based on YFV, MVA, or Adenovirus have a well-established safety record. Still, the immunogenicity and efficacy of vectored vaccines should be improved by using the knowledge and tools acquired from the DNA vaccine field. To this end, bioinformatic tools routinely used for design of DNA vaccines should also be used for careful design of synthetic genes for optimal protein expression taking the genetic background of the vector and host into account. However, bioinformatics is not systematically used to guide the rational development of safer vaccines. Several tools are available that can be used separately to eliminate for instance allergenic, immunosuppressive, oncogenic, or DNA binding sequences from the target protein. However, there is a need for integration of these tools into a software program that will allow a more rational design of safe and effective vaccines.

References

[1] M. W. Gaunt, A. A. Sall, X. de Lamballerie, A. K. I. Falconar, T. I. Dzhivanian, and E. A. Gould, "Phylogenetic relationships of flaviviruses correlate with their epidemiology, disease association and biogeography," *Journal of General Virology*, vol. 82, part 8, pp. 1867–1876, 2001.

[2] G. Kuno, G.-J. J. Chang, K. R. Tsuchiya, N. Karabatsos, and C. B. Cropp, "Phylogeny of the genus Flavivirus," *Journal of Virology*, vol. 72, no. 1, pp. 73–83, 1998.

[3] A. Oya and I. Kurane, "Japanese encephalitis for a reference to international travelers," *Journal of Travel Medicine*, vol. 14, no. 4, pp. 259–268, 2007.

[4] B. J. Blitvich, "Transmission dynamics and changing epidemiology of West Nile virus," *Animal Health Research Reviews*, vol. 9, no. 1, pp. 71–86, 2008.

[5] S. B. Halstead, "Dengue," *The Lancet*, vol. 370, no. 9599, pp. 1644–1652, 2007.

[6] E. Gould and T. Solomon, "Pathogenic flaviviruses," *The Lancet*, vol. 371, no. 9611, pp. 500–509, 2008.

[7] B. D. Lindenbach and C. M. Rice, "Molecular biology of flaviviruses," *Advances in Virus Research*, vol. 59, pp. 23–61, 2003.

[8] W. D. Crill and G.-J. J. Chang, "Localization and characterization of flavivirus envelope glycoprotein cross-reactive epitopes," *Journal of Virology*, vol. 78, no. 24, pp. 13975–13986, 2004.

[9] J. T. Roehrig, R. A. Bolin, and R. G. Kelly, "Monoclonal antibody mapping of the envelope glycoprotein of the dengue 2 virus, Jamaica," *Virology*, vol. 246, no. 2, pp. 317–328, 1998.

[10] M. S. Diamond, T. C. Pierson, and D. H. Fremont, "The structural immunology of antibody protection against West Nile virus," *Immunological Reviews*, vol. 225, no. 1, pp. 212–225, 2008.

[11] T. Oliphant, G. E. Nybakken, M. Engle, et al., "Antibody recognition and neutralization determinants on domains I and II of West Nile virus envelope protein," *Journal of Virology*, vol. 80, no. 24, pp. 12149–12159, 2006.

[12] D. E. Volk, D. W. C. Beasley, D. A. Kallick, M. R. Holbrook, A. D. T. Barrett, and D. G. Gorenstein, "Solution structure and antibody binding studies of the envelope protein domain III from the New York strain of West Nile virus," *The Journal of Biological Chemistry*, vol. 279, no. 37, pp. 38755–38761, 2004.

[13] D. W. C. Beasley and A. D. T. Barrett, "Identification of neutralizing epitopes within structural domain III of the West Nile virus envelope protein," *Journal of Virology*, vol. 76, no. 24, pp. 13097–13100, 2002.

[14] T. C. Pierson, Q. Xu, S. Nelson, et al., "The stoichiometry of antibody-mediated neutralization and enhancement of West Nile virus infection," *Cell Host and Microbe*, vol. 1, no. 2, pp. 135–145, 2007.

[15] T. Oliphant, M. Engle, G. E. Nybakken, et al., "Development of a humanized monoclonal antibody with therapeutic potential against West Nile virus," *Nature Medicine*, vol. 11, no. 5, pp. 522–530, 2005.

[16] M. S. Diamond, E. M. Sitati, L. D. Friend, S. Higgs, B. Shrestha, and M. Engle, "A critical role for induced IgM in the protection against West Nile virus infection," *Journal of Experimental Medicine*, vol. 198, no. 12, pp. 1853–1862, 2003.

[17] B. Shrestha and M. S. Diamond, "Role of $CD8^+$ T cells in control of West Nile virus infection," *Journal of Virology*, vol. 78, no. 15, pp. 8312–8321, 2004.

[18] J. D. Brien, J. L. Uhrlaub, and J. Nikolich-Zugich, "West Nile virus-specific CD4 T cells exhibit direct antiviral cytokine secretion and cytotoxicity and are sufficient for antiviral protection," *Journal of Immunology*, vol. 181, no. 12, pp. 8568–8575, 2008.

[19] L. E. Yauch, R. M. Zellweger, M. F. Kotturi, et al., "A protective role for dengue virus-specific $CD8^+$ T cells," *Journal of Immunology*, vol. 182, no. 8, pp. 4865–4873, 2009.

[20] D. J. Gubler, "Dengue and dengue hemorrhagic fever," *Clinical Microbiology Reviews*, vol. 11, no. 3, pp. 480–496, 1998.

[21] G. P. Kouri, M. G. Guzman, J. R. Bravo, and C. Triana, "Dengue haemorrhagic fever/dengue shock syndrome: lessons from the Cuban epidemic, 1981," *Bulletin of the World Health Organization*, vol. 67, no. 4, pp. 375–380, 1989.

[22] E. E. Ooi, K. T. Goh, and D. J. Gubler, "Dengue prevention and 35 years of vector control in Singapore," *Emerging Infectious Diseases*, vol. 12, no. 6, pp. 887–893, 2006.

[23] M. Theiler and H. H. Smith, "The use of yellow fever virus modified by in vitro cultivation for human immunization," *The Journal of Experimental Medicine*, vol. 65, no. 6, pp. 787–800, 1937.

[24] V. Brusic, J. T. August, and N. Petrovsky, "Information technologies for vaccine research," *Expert Review of Vaccines*, vol. 4, no. 3, pp. 407–417, 2005.

[25] K. H. Eckels and R. Putnak, "Formalin-inactivated whole virus and recombinant subunit flavivirus vaccines," *Advances in Virus Research*, vol. 61, pp. 395–418, 2003.

[26] M. B. Appaiahgari and S. Vrati, "Immunogenicity and protective efficacy in mice of a formaldehyde-inactivated Indian strain of Japanese encephalitis virus grown in Vero cells," *Vaccine*, vol. 22, no. 27-28, pp. 3669–3675, 2004.

[27] K. Sugawara, K. Nishiyama, Y. Ishikawa, et al., "Development of vero cell-derived inactivated Japanese encephalitis vaccine," *Biologicals*, vol. 30, no. 4, pp. 303–314, 2002.

[28] A. Wilder-Smith and D. O. Freedman, "Japanese encephalitis: is there a need for a novel vaccine?" *Expert Review of Vaccines*, vol. 8, no. 8, pp. 969–972, 2009.

[29] A. M. Plesner, "Allergic reactions to Japanese encephalitis vaccine," *Immunology and Allergy Clinics of North America*, vol. 23, no. 4, pp. 665–697, 2003.

[30] R. L. de Swart, B. G. van den Hoogen, T. Kuiken, et al., "Immunization of macaques with formalin-inactivated human metapneumovirus induces hypersensitivity to hMPV infection," *Vaccine*, vol. 25, no. 51, pp. 8518–8528, 2007.

[31] R. L. De Swart, T. Kuiken, H. H. Timmerman, et al., "Immunization of macaques with formalin-inactivated respiratory syncytial virus (RSV) induces interleukin-13-associated hypersensitivity to subsequent RSV infection," *Journal of Virology*, vol. 76, no. 22, pp. 11561–11569, 2002.

[32] B. E. Martina, P. Koraka, P. van den Doel, G. van Amerongen, G. F. Rimmelzwaan, and A. D. M. E. Osterhaus, "Immunization with West Nile virus envelope domain III protects mice against lethal infection with homologous and heterologous virus," *Vaccine*, vol. 26, no. 2, pp. 153–157, 2008.

[33] J.-H. J. Chu, C.-C. S. Chiang, and M.-L. Ng, "Immunization of flavivirus West Nile recombinant envelope domain III protein induced specific immune response and protection against West Nile virus infection," *Journal of Immunology*, vol. 178, no. 5, pp. 2699–2705, 2007.

[34] W. Zhang, X. Du, G. Zhao, et al., "Levamisole is a potential facilitator for the activation of Th1 responses of the subunit HBV vaccination," *Vaccine*, vol. 27, no. 36, pp. 4938–4946, 2009.

[35] C. Stahl-Hennig, M. Eisenblätter, E. Jasny, et al., "Synthetic double-stranded RNAs are adjuvants for the induction of t helper 1 and humoral immune responses to human papillomavirus in rhesus macaques," *PLoS Pathogens*, vol. 5, no. 4, Article ID e1000373, 2009.

[36] Y. F. Lau, L.-H. Tang, and E.-E. Ooi, "A TLR3 ligand that exhibits potent inhibition of influenza virus replication and has strong adjuvant activity has the potential for dual applications in an influenza pandemic," *Vaccine*, vol. 27, no. 9, pp. 1354–1364, 2009.

[37] K. Gupta and C. Cooper, "A review of the role of CpG oligodeoxynucleotides as toll-like receptor 9 agonists in prophylactic and therapeutic vaccine development in infectious diseases," *Drugs in R and D*, vol. 9, no. 3, pp. 137–145, 2008.

[38] D. Higgins, J. D. Marshall, P. Traquina, G. Van Nest, and B. D. Livingston, "Immunostimulatory DNA as a vaccine adjuvant," *Expert Review of Vaccines*, vol. 6, no. 5, pp. 747–759, 2007.

[39] B. S. Park, D. H. Song, H. M. Kim, B. S. Choi, H. Lee, and J. O. Lee, "The structural basis of lipopolysaccharide recognition by the TLR4-MD-2 complex," *Nature*, vol. 458, no. 7242, pp. 1191–1195, 2009.

[40] J. N. Leonard, J. K. Bell, and D. M. Segal, "Predicting Toll-like receptor structures and characterizing ligand binding," *Methods in Molecular Biology*, vol. 517, pp. 55–67, 2009.

[41] T. P. Monie, N. J. Gay, and M. Gangloff, "Bioinformatic analysis of Toll-like receptor sequences and structures," *Methods in Molecular Biology*, vol. 517, pp. 69–79, 2009.

[42] W. F. McDonald, J. W. Huleatt, H. G. Foellmer, et al., "A West Nile virus recombinant protein vaccine that coactivates innate and adaptive immunity," *Journal of Infectious Diseases*, vol. 195, no. 11, pp. 1607–1617, 2007.

[43] S. J. Seligman and E. A. Gould, "Live flavivirus vaccines: reasons for caution," *The Lancet*, vol. 363, no. 9426, pp. 2073–2075, 2004.

[44] P. Koraka, S. Benton, G. van Amerongen, K. J. Stittelaar, and A. D. M. E. Osterhaus, "Efficacy of a live attenuated tetravalent candidate dengue vaccine in naive and previously infected cynomolgus macaques," *Vaccine*, vol. 25, no. 29, pp. 5409–5416, 2007.

[45] W. Sun, A. Nisalak, M. Gettayacamin, et al., "Protection of rhesus monkeys against dengue virus challenge after tetravalent live attenuated dengue virus vaccination," *Journal of Infectious Diseases*, vol. 193, no. 12, pp. 1658–1665, 2006.

[46] W. Sun, D. Cunningham, S. S. Wasserman, et al., "Phase 2 clinical trial of three formulations of tetravalent live-attenuated dengue vaccine in flavivirus-naive adults," *Human Vaccines*, vol. 5, no. 1, pp. 33–40, 2009.

[47] S. Simasathien, S. J. Thomas, V. Watanaveeradej, et al., "Safety and immunogenicity of a tetravalent live-attenuated dengue vaccine in flavivirus naive children," *American Journal of Tropical Medicine and Hygiene*, vol. 78, no. 3, pp. 426–433, 2008.

[48] S. Kitchener, M. Nissen, P. Nasveld, et al., "Immunogenicity and safety of two live-attenuated tetravalent dengue vaccine formulations in healthy Australian adults," *Vaccine*, vol. 24, no. 9, pp. 1238–1241, 2006.

[49] D. E. Alvarez, A. L. De Lella Ezcurra, S. Fucito, and A. V. Gamarnik, "Role of RNA structures present at the 3′UTR of dengue virus on translation, RNA synthesis, and viral replication," *Virology*, vol. 339, no. 2, pp. 200–212, 2005.

[50] V. Proutski, M. W. Gaunt, E. A. Gould, and E. C. Holmes, "Secondary structure of the 3′-untranslated region of yellow fever virus: implications for virulence, attenuation and vaccine development," *Journal of General Virology*, vol. 78, part 7, pp. 1543–1549, 1997.

[51] V. Proutski, E. A. Gould, and E. C. Holmes, "Secondary structure of the 3′ untranslated region of flaviviruses: similarities and differences," *Nucleic Acids Research*, vol. 25, no. 6, pp. 1194–1202, 1997.

[52] V. Proutski, T. S. Gritsun, E. A. Gould, and E. C. Holmes, "Biological consequences of deletions within the 3′-untranslated region of flaviviruses may be due to rearrangements of RNA secondary structure," *Virus Research*, vol. 64, no. 2, pp. 107–123, 1999.

[53] J. E. Blaney Jr., D. H. Johnson, G. G. Manipon, et al., "Genetic basis of attenuation of dengue virus type 4 small plaque mutants with restricted replication in suckling mice and in SCID mice transplanted with human liver cells," *Virology*, vol. 300, no. 1, pp. 125–139, 2002.

[54] A. P. Durbin, R. A. Karron, W. Sun, et al., "Attenuation and immunogenicity in humans of a live dengue virus type-4 vaccine candidate with a 30 nucleotide deletion in its 3′-untranslated region," *American Journal of Tropical Medicine and Hygiene*, vol. 65, no. 5, pp. 405–413, 2001.

[55] S. S. Whitehead, B. Falgout, K. A. Hanley, J. E. Blaney Jr., L. Markoff, and B. R. Murphy, "A live, attenuated dengue virus

type 1 vaccine candidate with a 30-nucleotide deletion in the 3′ untranslated region is highly attenuated and immunogenic in monkeys," *Journal of Virology*, vol. 77, no. 2, pp. 1653–1657, 2003.

[56] J. E. Blaney Jr., C. T. Hanson, C. Y. Firestone, K. A. Hanley, B. R. Murphy, and S. S. Whitehead, "Genetically modified, live attenuated dengue virus type 3 vaccine candidates," *American Journal of Tropical Medicine and Hygiene*, vol. 71, no. 6, pp. 811–821, 2004.

[57] J. E. Blaney Jr., C. T. Hanson, K. A. Hanley, B. R. Murphy, and S. S. Whitehead, "Vaccine candidates derived from a novel infectious cDNA clone of an American genotype dengue virus type 2," *BMC Infectious Diseases*, vol. 4, no. 4, article 39, 2004.

[58] J. E. Blaney Jr., N. S. Sathe, L. Goddard, et al., "Dengue virus type 3 vaccine candidates generated by introduction of deletions in the 3′ untranslated region (3′-UTR) or by exchange of the DENV-3 3′-UTR with that of DENV-4," *Vaccine*, vol. 26, no. 6, pp. 817–828, 2008.

[59] C. S. Hahn, J. M. Dalrymple, J. H. Strauss, and C. M. Rice, "Comparison of the virulent Asibi strain of yellow fever virus with the 17D vaccine strain derived from it," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 84, no. 7, pp. 2019–2023, 1987.

[60] Z.-L. Xu, H. Mizuguchi, A. Ishii-Watabe, E. Uchida, T. Mayumi, and T. Hayakawa, "Optimization of transcriptional regulatory elements for constructing plasmid vectors," *Gene*, vol. 272, no. 1-2, pp. 149–156, 2001.

[61] C. Beck, H. Uramoto, J. Boren, and L. M. Akyurek, "Tissue-specific targeting for cardiovascular gene transfer. Potential vectors and future challenges," *Current Gene Therapy*, vol. 4, no. 4, pp. 457–467, 2004.

[62] A. R. Brooks, R. N. Harkins, P. Wang, H. S. Qian, P. Liu, and G. M. Rubanyi, "Transcriptional silencing is associated with extensive methylation of the CMV promoter following adenoviral gene delivery to muscle," *Journal of Gene Medicine*, vol. 6, no. 4, pp. 395–404, 2004.

[63] L. Qin, Y. Ding, D. R. Pahud, E. Chang, M. J. Imperiale, and J. S. Bromberg, "Promoter attenuation in gene therapy: interferon-γ and tumor necrosis factor-α inhibit transgene expression," *Human Gene Therapy*, vol. 8, no. 17, pp. 2019–2029, 1997.

[64] M. Al-Dosari, G. Zhang, J. E. Knapp, and D. Liu, "Evaluation of viral and mammalian promoters for driving transgene expression in mouse liver," *Biochemical and Biophysical Research Communications*, vol. 339, no. 2, pp. 673–678, 2006.

[65] F. Takeshita, K. Takase, M. Tozuka, et al., "Muscle creatine kinase/SV40 hybrid promoter for muscle-targeted long-term transgene expression," *International Journal of Molecular Medicine*, vol. 19, no. 2, pp. 309–315, 2007.

[66] M. Mori-Uchino, T. Takeuchi, I. Murakami, et al., "Enhanced transgene expression in the mouse skeletal muscle infected by the adeno-associated viral vector with the human elongation factor 1α promoter and a human chromatin insulator," *Journal of Gene Medicine*, vol. 11, no. 7, pp. 598–604, 2009.

[67] S. Hong, D. Y. Hwang, S. Yoon, et al., "Functional analysis of various promoters in lentiviral vectors at different stages of in vitro differentiation of mouse embryonic stem cells," *Molecular Therapy*, vol. 15, no. 9, pp. 1630–1639, 2007.

[68] T. Vanniasinkam, S. T. Reddy, and H. C. J. Ertl, "DNA immunization using a non-viral promoter," *Virology*, vol. 344, no. 2, pp. 412–420, 2006.

[69] J.-S. Song, "Activity of the human telomerase catalytic subunit (hTERT) gene promoter could be increased by the SV40 enhancer," *Bioscience, Biotechnology and Biochemistry*, vol. 68, no. 8, pp. 1634–1639, 2004.

[70] M. Yago, R. Ohki, S. Hatakeyama, T. Fujita, and F. Ishikawa, "Variant forms of upstream stimulatory factors (USFs) control the promoter activity of hTERT, the human gene encoding the catalytic subunit of telomerase," *FEBS Letters*, vol. 520, no. 1–3, pp. 40–46, 2002.

[71] I. Horikawa, P. L. Cable, C. Afshari, and J. C. Barrett, "Cloning and characterization of the promoter region of human telomerase reverse transcriptase gene," *Cancer Research*, vol. 59, no. 4, pp. 826–830, 1999.

[72] M. S. Ashok and P. N. Rangarajan, "Protective efficacy of a plasmid DNA encoding Japanese encephalitis virus envelope protein fused to tissue plasminogen activator signal sequences: studies in a murine intracerebral virus challenge model," *Vaccine*, vol. 20, no. 11-12, pp. 1563–1570, 2002.

[73] Z. S. Guo, Q. Li, D. L. Bartlett, J. Y. Yang, and B. Fang, "Gene transfer: the challenge of regulated gene expression," *Trends in Molecular Medicine*, vol. 14, no. 9, pp. 410–418, 2008.

[74] A. H. Lee, Y. S. Suh, J. H. Sung, S. H. Yang, and Y. C. Sung, "Comparison of various expression plasmids for the induction of immune response by DNA immunization," *Molecules and Cells*, vol. 7, no. 4, pp. 495–501, 1997.

[75] B. S. Chapman, R. M. Thayer, K. A. Vincent, and N. L. Haigwood, "Effect of intron A from human cytomegalovirus (Towne) immediate-early gene on heterologous expression in mammalian cells," *Nucleic Acids Research*, vol. 19, no. 14, pp. 3979–3986, 1991.

[76] J. Chinsangaram, C. Beard, P. W. Mason, M. K. Zellner, G. Ward, and M. J. Grubman, "Antibody response in mice inoculated with DNA expressing foot-and- mouth disease virus capsid proteins," *Journal of Virology*, vol. 72, no. 5, pp. 4454–4457, 1998.

[77] M. T. F. Huang and C. M. Gorman, "Intervening sequences increase efficiency of RNA 3′ processing and accumulation of cytoplasmic RNA," *Nucleic Acids Research*, vol. 18, no. 4, pp. 937–947, 1990.

[78] M. T. F. Huang and C. M. Gorman, "The simian virus 40 small-t intron, present in many common expression vectors, leads to aberrant splicing," *Molecular and Cellular Biology*, vol. 10, no. 4, pp. 1805–1810, 1990.

[79] L.-M. You, J. Luo, A.-P. Wang, et al., "A hybrid promoter-containing vector for direct cloning and enhanced expression of PCR-amplified ORFs in mammalian cells," *Molecular Biology Reports*. In press.

[80] S. Garg, A. E. Oran, H. Hon, and J. Jacob, "The hybrid cytomegalovirus enhancer/chicken β-actin promoter along with woodchuck hepatitis virus posttranscriptional regulatory element enhances the protective efficacy of DNA vaccines," *Journal of Immunology*, vol. 173, no. 1, pp. 550–558, 2004.

[81] A. N. Alexopoulou, J. R. Couchman, and J. R. Whiteford, "The CMV early enhancer/chicken β actin (CAG) promoter can be used to drive transgene expression during the differentiation of murine embryonic stem cells into vascular progenitors," *BMC Cell Biology*, vol. 9, article 2, 2008.

[82] G.-J. J. Chang, B. S. Davis, A. R. Hunt, D. A. Holmes, and G. Kuno, "Flavivirus DNA vaccines: current status and potential," *Annals of the New York Academy of Sciences*, vol. 951, pp. 272–285, 2001.

[83] A. P. Jathoul, J. L. Holley, and H. S. Garmory, "Efficacy of DNA vaccines expressing the type F botulinum toxin Hc fragment using different promoters," *Vaccine*, vol. 22, no. 29-30, pp. 3942–3946, 2004.

[84] K. Santos, C. M. P. Duke, S. M. Rodriguez-Colon, et al., "Effect of promoter strength on protein expression and immunogenicity of an HSV-1 amplicon vector encoding HIV-1 Gag," *Vaccine*, vol. 25, no. 9, pp. 1634–1646, 2007.

[85] S. Wang, D. J. Farfan-Arribas, S. Shen, et al., "Relative contributions of codon usage, promoter efficiency and leader sequence to the antigen expression and immunogenicity of HIV-1 Env DNA vaccine," *Vaccine*, vol. 24, no. 21, pp. 4531–4540, 2006.

[86] J. Zhao, L. Hyman, and C. Moore, "Formation of mRNA 3′ ends in eukaryotes: mechanism, regulation, and interrelationships with other steps in mRNA synthesis," *Microbiology and Molecular Biology Reviews*, vol. 63, no. 2, pp. 405–445, 1999.

[87] G. M. Gilmartin and J. R. Nevins, "An ordered pathway of assembly of components required for polyadenylation site recognition and processing," *Genes and Development*, vol. 3, no. 12B, pp. 2180–2190, 1989.

[88] C. Montell, E. F. Fisher, M. H. Caruthers, and A. J. Berk, "Inhibition of RNA cleavage but not polyadenylation by a point mutation in mRNA 3′ consensus sequence AAUAAA," *Nature*, vol. 305, no. 5935, pp. 600–605, 1983.

[89] M. D. Sheets, S. C. Ogg, and M. P. Wickens, "Point mutations of AAUAAA and the poly (A) addition site: effects on the accuracy and efficiency of cleavage and polyadenylation in vitro," *Nucleic Acids Research*, vol. 18, no. 19, pp. 5799–5805, 1990.

[90] H. Hans and J. C. Alwine, "Functionally significant secondary structure of the simian virus 40 late polyadenylation signal," *Molecular and Cellular Biology*, vol. 20, no. 8, pp. 2926–2932, 2000.

[91] E. R. Gimmi, M. E. Reff, and I. C. Deckman, "Alterations in the pre-mRNA topology of the bovine growth hormone polyadenylation region decrease poly(A) site efficiency," *Nucleic Acids Research*, vol. 17, no. 17, pp. 6983–6998, 1989.

[92] C. M. Roth and S. Sundaram, "Engineering synthetic vectors for improved DNA delivery: insights from intracellular pathways," *Annual Review of Biomedical Engineering*, vol. 6, pp. 397–426, 2004.

[93] R. Schirmbeck, S. A. Konig-Merediz, P. Riedl, et al., "Priming of immune responses to hepatitis B surface antigen with minimal DNA expression constructs modified with a nuclear localization signal peptide," *Journal of Molecular Medicine*, vol. 79, no. 5-6, pp. 343–350, 2001.

[94] H. Pollard, G. Toumaniantz, J. L. Amos, et al., "Ca^{2+}-sensitive cytosolic nucleases prevent efficient delivery to the nucleus of injected plasmids," *Journal of Gene Medicine*, vol. 3, no. 2, pp. 153–164, 2001.

[95] A. R. Azzoni, S. C. Ribeiro, G. A. Monteiro, and D. M. F. Prazeres, "The impact of polyadenylation signals on plasmid nuclease-resistance and transgene expression," *Journal of Gene Medicine*, vol. 9, no. 5, pp. 392–402, 2007.

[96] S. C. Ribiero, G. A. Monteiro, and D. M. Prazeres, "The role of polyadenylation signal secondary structures on the resistance of plasmid vectors to nucleases," *Journal of Gene Medicine*, vol. 6, no. 5, pp. 565–573, 2004.

[97] T. Ikemura, "Correlation between the abundance of *Escherichia coli* transfer RNAs and the occurrence of the respective codons in its protein genes," *Journal of Molecular Biology*, vol. 146, no. 1, pp. 1–21, 1981.

[98] J. Zhou, W. J. Liu, S. W. Peng, X. Y. Sun, and I. Frazer, "Papillomavirus capsid protein expression level depends on the match between codon usage and tRNA availability," *Journal of Virology*, vol. 73, no. 6, pp. 4972–4982, 1999.

[99] T. Ikemura, "Codon usage and tRNA content in unicellular and multicellular organisms," *Molecular Biology and Evolution*, vol. 2, no. 1, pp. 13–34, 1985.

[100] T. Ikemura, "Correlation between the abundance of yeast transfer RNAs and the occurrence of the respective codons in protein genes. Differences in synonymous codon choice patterns of yeast and *Escherichia coli* with reference to the abundance of isoaccepting transfer RNAs," *Journal of Molecular Biology*, vol. 158, no. 4, pp. 573–597, 1982.

[101] R. Grantham, C. Gautier, M. Gouy, M. Jacobzone, and R. Mercier, "Codon catalog usage is a genome strategy modulated for gene expressivity," *Nucleic Acids Research*, vol. 9, no. 1, pp. r43–r74, 1981.

[102] P. M. Sharp, T. M. Tuohy, and K. R. Mosurski, "Codon usage in yeast: cluster analysis clearly differentiates highly and lowly expressed genes," *Nucleic Acids Research*, vol. 14, no. 13, pp. 5125–5143, 1986.

[103] A. Wang, L. Ren, G. Abenes, and R. Hai, "Genome sequence divergences and functional variations in human cytomegalovirus strains," *FEMS Immunology and Medical Microbiology*, vol. 55, no. 1, pp. 23–33, 2009.

[104] C. M. Stenstrom and L. A. Isaksson, "Influences on translation initiation and early elongation by the messenger RNA region flanking the initiation codon at the 3′ side," *Gene*, vol. 288, no. 1-2, pp. 1–8, 2002.

[105] A. C. Looman, J. Bodlaender, L. J. Comstock, et al., "Influence of the codon following the AUG initiation codon on the expression of a modified lacZ gene in Escherichia coli," *EMBO Journal*, vol. 6, no. 8, pp. 2489–2492, 1987.

[106] A. A. Komar, T. Lesnik, and C. Reiss, "Synonymous codon substitutions affect ribosome traffic and protein folding during in vitro translation," *FEBS Letters*, vol. 462, no. 3, pp. 387–391, 1999.

[107] P. M. Sharp and W. H. Li, "The codon adaptation index—a measure of directional synonymous codon usage bias, and its potential applications," *Nucleic Acids Research*, vol. 15, no. 3, pp. 1281–1295, 1987.

[108] L. Deml, A. Bojak, S. Steck, et al., "Multiple effects of codon usage optimization on expression and immunogenicity of DNA candidate vaccines encoding the human immunodeficiency virus type 1 Gag protein," *Journal of Virology*, vol. 75, no. 22, pp. 10991–11001, 2001.

[109] M. Uchijima, A. Yoshida, T. Nagata, and Y. Koide, "Optimization of codon usage of plasmid DNA vaccine is required for the effective MHC class I-restricted T cell responses against an intracellular bacterium," *Journal of Immunology*, vol. 161, no. 10, pp. 5594–5599, 1998.

[110] T. Nagata, M. Uchijima, A. Yoshida, M. Kawashima, and Y. Koide, "Codon optimization effect on translational efficiency of DNA vaccine in mammalian cells: analysis of plasmid DNA encoding a CTL epitope derived from microorganisms," *Biochemical and Biophysical Research Communications*, vol. 261, no. 2, pp. 445–451, 1999.

[111] Y.-K. Cheung, S. C.-S. Cheng, F. W.-Y. Sin, and Y. Xie, "Plasmid encoding papillomavirus type 16 (HPV16) DNA constructed with codon optimization improved the immunogenicity against HPV infection," *Vaccine*, vol. 23, no. 5, pp. 629–638, 2004.

[112] L. Frelin, G. Ahlen, M. Alheim, et al., "Codon optimization and mRNA amplification effectively enhances the immunogenecity of the hepatitis C virus nonstructural 3/4A gene," *Gene Therapy*, vol. 11, no. 6, pp. 522–533, 2004.

[113] F. Gao, Y. Li, J. M. Decker, et al., "Codon usage optimization of HIV type 1 subtype C gag, pol, env, and nef genes: in vitro

expression and immune responses in DNA-vaccinated mice," *AIDS Research and Human Retroviruses*, vol. 19, no. 9, pp. 817–823, 2003.

[114] H. J. Ko, S. Y. Ko, Y. J. Kim, E. G. Lee, S.-N. Cho, and C.-Y. Kang, "Optimization of codon usage enhances the immunogenicity of a DNA vaccine encoding mycobacterial antigen Ag85B," *Infection and Immunity*, vol. 73, no. 9, pp. 5666–5674, 2005.

[115] M. Kozak, "Regulation of translation via mRNA structure in prokaryotes and eukaryotes," *Gene*, vol. 361, no. 1-2, pp. 13–37, 2005.

[116] D. R. Cavener and S. C. Ray, "Eukaryotic start and stop translation sites," *Nucleic Acids Research*, vol. 19, no. 12, pp. 3185–3192, 1991.

[117] U. Kutay, G. Ahnert-Hilger, E. Hartmann, B. Wiedenmann, and T. A. Rapoport, "Transport route for synaptobrevin via a novel pathway of insertion into the endoplasmic reticulum membrane," *EMBO Journal*, vol. 14, no. 2, pp. 217–223, 1995.

[118] B. Martoglio and B. Dobberstein, "Signal sequences: more than just greasy peptides," *Trends in Cell Biology*, vol. 8, no. 10, pp. 410–415, 1998.

[119] Y. M. Zalucki, I. R. Beacham, and M. P. Jennings, "Biased codon usage in signal peptides: a role in protein export," *Trends in Microbiology*, vol. 17, no. 4, pp. 146–150, 2009.

[120] M. Sakaguchi, R. Tomiyoshi, T. Kuroiwa, K. Mihara, and T. Omura, "Functions of signal and signal-anchor sequences are determined by the balance between the hydrophobic segment and the N-terminal charge," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 1, pp. 16–19, 1992.

[121] I. Sominskaya, E. Alekseeva, D. Skrastina, et al., "Signal sequences modulate the immunogenic performance of human hepatitis C virus E2 gene," *Molecular Immunology*, vol. 43, no. 12, pp. 1941–1952, 2006.

[122] Z. Li, A. Howard, C. Kelley, G. Delogu, F. Collins, and S. Morris, "Immunogenicity of DNA vaccines expressing tuberculosis proteins fused to tissue plasminogen activator signal sequences," *Infection and Immunity*, vol. 67, no. 9, pp. 4780–4786, 1999.

[123] M. Luo, P. Tao, J. Li, S. Zhou, D. Guo, and Z. Pan, "Immunization with plasmid DNA encoding influenza A virus nucleoprotein fused to a tissue plasminogen activator signal sequence elicits strong immune responses and protection against H5N1 challenge in mice," *Journal of Virological Methods*, vol. 154, no. 1-2, pp. 121–127, 2008.

[124] C. Jiang, D. M. Magee, F. D. Ivey, and R. A. Cox, "Role of signal sequence in vaccine-induced protection against experimental coccidioidomycosis," *Infection and Immunity*, vol. 70, no. 7, pp. 3539–3545, 2002.

[125] R. Putnak, K. Porter, and C. Schmaljohn, "DNA vaccines for flaviviruses," *Advances in Virus Research*, vol. 61, pp. 445–468, 2003.

[126] C. J. Wu, T. L. Li, H. W. Huang, M. H. Tao, and Y. L. Chan, "Development of an effective Japanese encephalitis virus-specific DNA vaccine," *Microbes and Infection*, vol. 8, no. 11, pp. 2578–2586, 2006.

[127] G.-J. J. Chang, A. R. Hunt, and B. Davis, "A single intramuscular injection of recombinant plasmid DNA induces protective immunity and prevents Japanese encephalitis in mice," *Journal of Virology*, vol. 74, no. 9, pp. 4244–4252, 2000.

[128] E. Konishi, M. Yamaoka, Khin-Sane-Win, I. Kurane, and P. W. Mason, "Induction of protective immunity against japanese encephalitis in mice by immunization with a plasmid encoding Japanese encephalitis virus premembrane and envelope genes," *Journal of Virology*, vol. 72, no. 6, pp. 4925–4930, 1998.

[129] B. S. Davis, G.-J. J. Chang, B. Cropp, et al., "West Nile virus recombinant DNA vaccine protects mouse and horse from virus challenge and expresses in vitro a noninfectious recombinant antigen that can be used in enzyme-linked immunosorbent assays," *Journal of Virology*, vol. 75, no. 9, pp. 4040–4047, 2001.

[130] J. E. Martin, T. C. Pierson, S. Hubka, et al., "A West Nile virus DNA vaccine induces neutralizing antibody in healthy adults during a phase 1 clinical trial," *Journal of Infectious Diseases*, vol. 196, no. 12, pp. 1732–1740, 2007.

[131] R. Kaur, G. Sachdeva, and S. Vrati, "Plasmid DNA immunization against Japanese encephalitis virus: immunogenicity of membrane-anchored and secretory envelope protein," *Journal of Infectious Diseases*, vol. 185, no. 1, pp. 1–12, 2002.

[132] K. Raviprakash, T. J. Kochel, D. Ewing, et al., "Immunogenicity of dengue virus type 1 DNA vaccines expressing truncated and full length envelope protein," *Vaccine*, vol. 18, no. 22, pp. 2426–2434, 2000.

[133] K. Raviprakash and K. R. Porter, "Needle-free injection of DNA vaccines: a brief overview and methodology," *Methods in Molecular Medicine*, vol. 127, pp. 83–89, 2006.

[134] K. Raviprakash, E. Marques, D. Ewing, et al., "Synergistic neutralizing antibody response to a dengue virus type 2 DNA vaccine by incorporation of lysosome-associated membrane protein sequences and use of plasmid expressing GM-CSF," *Virology*, vol. 290, no. 1, pp. 74–82, 2001.

[135] K. Bharati, M. B. Appaiahgari, and S. Vrati, "Effect of cytokine-encoding plasmid delivery on immune response to Japanese encephalitis virus DNA vaccine in mice," *Microbiology and Immunology*, vol. 49, no. 4, pp. 349–353, 2005.

[136] Y. Lu, K. Raviprakash, I. C. Leao, et al., "Dengue 2 PreM-E/LAMP chimera targeted to the MHC class II compartment elicits long-lasting neutralizing antibodies," *Vaccine*, vol. 21, no. 17-18, pp. 2187–2198, 2003.

[137] K. Isaji, A. Kawase, M. Matono, X. Guan, M. Nishikawa, and Y. Takakura, "Enhanced CTL response by controlled intracellular trafficking of antigen in dendritic cells following DNA vaccination," *Journal of Controlled Release*, vol. 135, no. 3, pp. 227–233, 2009.

[138] S. J. Kim, C. Lee, S. Y. Lee, et al., "Enhanced immunogenicity of human papillomavirus 16 L1 genetic vaccines fused to an ER-targeting secretory signal peptide and RANTES," *Gene Therapy*, vol. 10, no. 15, pp. 1268–1273, 2003.

[139] D. C. Chang, W. J. Liu, I. Anraku, et al., "Single-round infectious particles enhance immunogenicity of a DNA vaccine against West Nile virus," *Nature Biotechnology*, vol. 26, no. 5, pp. 571–577, 2008.

[140] J. H. Nam, H. S. Bang, H. W. Cho, and Y. H. Chung, "Different contribution of co-stimulatory molecules B7.1 and B7.2 to the immune response to recombinant modified vaccinia virus ankara vaccine expressing prM/E proteins of Japanese encephalitis virus and two hepatitis B virus vaccines," *Acta Virologica*, vol. 51, no. 2, pp. 125–130, 2007.

[141] R. Men, L. Wyatt, I. Tokimatsu, et al., "Immunization of rhesus monkeys with a recombinant of modified vaccinia virus Ankara expressing a truncated envelope glycoprotein of dengue type 2 virus induced resistance to dengue type 2 virus challenge," *Vaccine*, vol. 18, no. 27, pp. 3113–3122, 2000.

[142] E. Konishi, I. Kurane, P. W. Mason, R. E. Shope, and F. A. Ennis, "Poxvirus-based Japanese encephalitis vaccine candidates induce JE virus-specific $CD8^+$ cytotoxic T lymphocytes in mice," *Virology*, vol. 227, no. 2, pp. 353–360, 1997.

[143] K. K. Seino, M. T. Long, E. P. J. Gibbs, et al., "Comparative efficacies of three commercially available vaccines against West Nile virus (WNV) in a short-duration challenge trial involving an equine WNV encephalitis model," *Clinical and Vaccine Immunology*, vol. 14, no. 11, pp. 1465–1471, 2007.

[144] K. Raviprakash, D. Wang, D. Ewing, et al., "A tetravalent dengue vaccine based on a complex adenovirus vector provides significant protection in rhesus monkeys against all four serotypes of dengue virus," *Journal of Virology*, vol. 82, no. 14, pp. 6927–6934, 2008.

[145] S. Khanam, N. Khanna, and S. Swaminathan, "Induction of neutralizing antibodies and T cell responses by dengue virus type 2 envelope domain III encoded by plasmid and adenoviral vectors," *Vaccine*, vol. 24, no. 42-43, pp. 6513–6525, 2006.

[146] M. B. Appaiahgari, M. Saini, M. Rauthan, Jyoti, and S. Vrati, "Immunization with recombinant adenovirus synthesizing the secretory form of Japanese encephalitis virus envelope protein protects adenovirus-exposed mice against lethal encephalitis," *Microbes and Infection*, vol. 8, no. 1, pp. 92–104, 2006.

[147] P. Despres, C. Combredet, M.-P. Frenkiel, C. Lorin, M. Brahic, and F. Tangy, "Live measles vaccine expressing the secreted form of the West Nile virus envelope glycoprotein protects against West Nile virus encephalitis," *Journal of Infectious Diseases*, vol. 191, no. 2, pp. 207–214, 2005.

[148] L. J. White, M. M. Parsons, A. C. Whitmore, B. M. Williams, A. De Silva, and R. E. Johnston, "An immunogenic and protective alphavirus replicon particle-based dengue vaccine overcomes maternal antibody interference in weanling mice," *Journal of Virology*, vol. 81, no. 19, pp. 10329–10339, 2007.

[149] M. N. Fleeton, P. Liljeström, B. J. Sheahan, and G. J. Atkins, "Recombinant Semliki Forest virus particles expressing louping ill virus antigens induce a better protective response than plasmid-based DNA vaccines or an inactivated whole particle vaccine," *Journal of General Virology*, vol. 81, part 3, pp. 749–758, 2000.

[150] G. Colombage, R. Hall, M. Pavy, and M. Lobigs, "DNA-based and alphavirus-vectored immunisation with PrM and E proteins elicits long-lived and protective immunity against the flavivirus, Murray Valley encephalitis virus," *Virology*, vol. 250, no. 1, pp. 151–163, 1998.

[151] A. V. Iyer, B. Pahar, M. J. Boudreaux, et al., "Recombinant vesicular stomatitis virus-based West Nile vaccine elicits strong humoral and cellular immune responses and protects mice against lethal challenge with the virulent West Nile virus strain LSU-AR01," *Vaccine*, vol. 27, no. 6, pp. 893–903, 2009.

[152] K. J. Stittelaar, T. Kuiken, R. L. de Swart, et al., "Safety of modified vaccinia virus Ankara (MVA) in immune-suppressed macaques," *Vaccine*, vol. 19, no. 27, pp. 3700–3709, 2001.

[153] D. L. Swenson, D. Wang, M. Luo, et al., "Vaccine to confer to nonhuman primates complete protection against multistrain Ebola and Marburg virus infections," *Clinical and Vaccine Immunology*, vol. 15, no. 3, pp. 460–467, 2008.

[154] J. Schepp-Berglind, M. Luo, D. Wang, et al., "Complex adenovirus-mediated expression of West Nile Virus C, preM, E, and NS1 proteins induces both humoral and cellular immune responses," *Clinical and Vaccine Immunology*, vol. 14, no. 9, pp. 1117–1126, 2007.

[155] J. H. Nam, S. L. Chae, and H. W. Cho, "Immunogenicity of a recombinant MVA and a DNA vaccine for Japanese encephalitis virus in swine," *Microbiology and Immunology*, vol. 46, no. 1, pp. 23–28, 2002.

[156] B. Wang, J. Li, F. H. Fu, et al., "Construction and analysis of compact muscle-specific promoters for AAV vectors," *Gene Therapy*, vol. 15, no. 22, pp. 1489–1499, 2008.

[157] R. A. Hall and A. A. Khromykh, "ChimeriVax-West Nile vaccine," *Current Opinion in Molecular Therapeutics*, vol. 9, no. 5, pp. 498–504, 2007.

[158] G. Dauphin and S. Zientara, "West Nile virus: recent trends in diagnosis and vaccine development," *Vaccine*, vol. 25, no. 30, pp. 5563–5576, 2007.

[159] D. W. C. Beasley, L. Li, M. T. Suderman, et al., "Protection against Japanese encephalitis virus strains representing four genotypes by passive transfer of sera raised against ChimeriVax-JE experimental vaccine," *Vaccine*, vol. 22, no. 27-28, pp. 3722–3726, 2004.

[160] C. E. McGee, M. G. Lewis, M. S. Claire, et al., "Recombinant chimeric virus with wild-type dengue 4 virus premembrane and envelope and virulent yellow fever virus Asibi backbone sequences is dramatically attenuated in nonhuman primates," *Journal of Infectious Diseases*, vol. 197, no. 5, pp. 693–697, 2008.

[161] F. Guirakhoo, K. Pugachev, J. Arroyo, et al., "Viremia and immunogenicity in nonhuman primates of a tetravalent yellow fever-dengue chimeric vaccine: genetic reconstructions, dose adjustment, and antibody responses against wild-type dengue virus isolates," *Virology*, vol. 298, no. 1, pp. 146–159, 2002.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 236528, 11 pages
doi:10.1155/2010/236528

Review Article

Comparative Pathogenesis and Systems Biology for Biodefense Virus Vaccine Development

Gavin C. Bowick and Alan D. T. Barrett

Departments of Pathology and Microbiology & Immunology, Center for Biodefense and Emerging Infectious Diseases, Sealy Center for Vaccine Development, and Institute of Human Infections & Immunity, University of Texas Medical Branch, Galveston, TX 77555-0609, USA

Correspondence should be addressed to Gavin C. Bowick, gabowick@utmb.edu

Received 30 September 2009; Revised 21 January 2010; Accepted 8 March 2010

Academic Editor: Yongqun Oliver He

Developing vaccines to biothreat agents presents a number of challenges for discovery, preclinical development, and licensure. The need for high containment to work with live agents limits the amount and types of research that can be done using complete pathogens, and small markets reduce potential returns for industry. However, a number of tools, from comparative pathogenesis of viral strains at the molecular level to novel computational approaches, are being used to understand the basis of viral attenuation and characterize protective immune responses. As the amount of basic molecular knowledge grows, we will be able to take advantage of these tools not only to rationally attenuate virus strains for candidate vaccines, but also to assess immunogenicity and safety in silico. This review discusses how a basic understanding of pathogenesis, allied with systems biology and machine learning methods, can impact biodefense vaccinology.

1. Vaccines for Biodefense

We have traditionally associated the term "Biodefense" with military applications. However, since October 2001 when anthrax spores were sent in envelopes through the US Postal Service, our understanding of biodefense has shifted significantly. We now see biodefense as the process to protect both civilian and military populations. It has become clear that many highly pathogenic microorganisms can be considered as either agents of biological warfare or naturally occurring emerging disease threats. In addition, we consider bioterrorism from varying points of view, including public health threats, veterinary threats, and agricultural threats. Taken together, they can be considered to be "biothreats". The biodefense field represents a unique challenge for vaccine development as the traditional economic models for vaccine development are based on large populations purchasing a vaccine to protect against common infectious diseases in order that a vaccine pharma can make a profit. The situation in biodefense is very different where the goal is to stockpile vaccines with the hope that they will never be used. Relatively small markets, combined with the difficulties of working with many of these agents—many of which require biosafety level 3 or 4 containment—means that there are currently no vaccines licensed for general use in the United States and most other countries for nearly all of the biothreat agents. There are major hurdles to the development of biodefense vaccines. In addition to the traditional issues of identifying protective immunogens and platforms to deliver the vaccines, there are major difficulties in undertaking efficacy studies as most of the diseases are rare, occur sporadically, or are not found naturally. Accordingly, emphasis is being placed on appropriate animal models to demonstrate efficacy in support of licensure. However, at this time, no vaccine has been approved based on animal efficacy trials only. Nonetheless, a significant amount of data are being accumulated describing the molecular basis of disease, which provides a strong foundation for continuing product development. The current state of biodefense vaccines in clinical development is summarized in Table 1.

Table 1: Current status of vaccines in clinical trials against priority pathogens and potential biothreat agents. Data for the table come from http://www.clinicaltrials.gov/ and other sources as indicated. Many vaccines are in preclinical development, including a number of candidate vaccines for Lassa virus [1–6] and Sin Nombre [7, 8] virus, both being important hemorrhagic fever viruses.

Virus	Disease	Candidates in clinical development
Dengue	Dengue fever, dengue hemorrhagic fever, and dengue shock syndrome	Live attenuated Live chimeric (yellow fever 17D backbone) Recombinant subunit (envelope protein) DNA plasmid Molecular attenuated based on infectious clone-derived virus
Ebola and Marburg viruses	Ebola hemorrhagic fever/Marburg hemorrhagic fever	Recombinant adenoviral vector Recombinant Vesicular stomatitis virus vector [9] DNA plasmid
Junin virus	Argentine hemorrhagic fever	Live-attenuated candid #1 currently used in Argentina. Candid #1 has had investigational new drug status in USA [10].
Rift Valley fever virus	Rift Valley fever	Inactivated Rift valley fever investigational new drug status in USA Live-attenuated MP-12
SARS-CoV	Severe acute respiratory syndrome	Inactivated SARS Coronavirus
Variola virus	Smallpox	Vaccinia virus vaccines were successfully used to eradicate smallpox. Continuing efforts to develop and refine vaccines to reduce incidence of adverse effects, Modified Vaccinia Ankara (MVA), LC16m8, and so forth.
Venezuelan equine encephalitis virus	Viral encephalitis	Live-attenuated TC-83, investigational new drug status in USA Formalin-inactivated C-84, investigational new drug status in USA Molecular-attenuated infectious clone-derived V3526

The advent of the 21st century has seen our expertise in molecular biology increase exponentially. With the increasing affordability of high-throughput sequencing, concepts such as reverse vaccinology—using pathogen genome sequences to produce peptides and nucleic acids for vaccine testing—have become more established [11, 12]. As our understanding of the molecular basis of pathogenesis continues to grow, coupled with large numbers of genome sequences for various pathogens becoming available and novel bioinformatics tools to facilitate their analysis, we are heading towards a position where vaccines can be rationally designed based on our molecular knowledge. This includes selection of an optimum immunogen and delivery system to maximize the host protective immune response rather than empirical approaches that were used for much of the 20th century. The continuing generation of high-throughput datasets characterizing high-priority biodefense and public health pathogens, and their host responses in the post-genomic era, will require novel bioinformatic techniques. These computational approaches will allow a more complete and thorough understanding of infection and pathogenesis, and greatly facilitate rational vaccine design. This review describes some of these methods and how they can be used in the development of novel vaccines against biothreat viruses.

2. Global Host-Pathogen Interactions of Biothreat Agents

Techniques such as microarray expression profiling, 2D gels, automated spot picking and peptide sequencing, and other high-throughput methods have revolutionized the study of virus interactions with the host cell. In particular, microarray analysis has proven a very useful tool in the emerging virus and biodefense field as the RNA extraction process facilitates removal of samples from the biosafety level 3 (BSL-3) or BSL-4 laboratory where only inactivated samples are allowed to leave biocontainment.

A successful vaccine must be both safe and immunogenic. Many pathogenic viruses interfere with the development of a protective immune response by preventing

the activation of key aspects of the immune response, from inhibition of IRF-3 phosphorylation to block type-I interferon synthesis to sequestering MHC-I in the cell to prevent display of virus-derived peptides. A central feature of many biothreat pathogens is the infection of macrophages and dendritic cells as the primary target for replication. Infection of these cell types enables viruses to prevent immune response development at an early, central stage. For example, dendritic cells infected with *Lymphocytic choriomeningitis virus* (LCMV), an Old World arenavirus used to infect macaques as a model for Lassa fever, lead to downregulation of MHC-II on the dendritic cell surface, inhibiting the ability of these cells to present antigen and develop the adaptive immune response [13] while *Marburg virus* infection of macaques leads to downregulation of MHC-II on dendritic cells in the spleen [14]. Effects such as these are common following infection with these viruses resulting in significantly impairment of the host to develop a protective immune response at an early stage of infection. The downregulation of MHC-II has also been observed in a microarray study of PBMCs from LCMV-infected macaques with pathogenic virus inducing a greater downregulation than infection with attenuated virus [15]. Interestingly, MHC-II was upregulated in liver tissues [16], highlighting the importance of animal models and role of the host as having different functional systems, which may respond differently to a similar challenge at the molecular level.

The contributions that host-pathogen interaction profiling can make to vaccine development are exemplified by a study which determined that substitution of a single amino acid in the VP35 protein of *Ebola virus* is sufficient to disrupt the viral inhibition of innate immune signaling, while maintaining the ability to replicate to wild-type levels in cell culture [17]. Similarly, deletion of an entire multigene family, with a hypothesized role in modulation of the interferon system, still results in viral replication [18]. If the viral proteins responsible for mediating inhibition of the host immune response can be determined, then targeted mutations can be made to disrupt these effects, counteracting the inhibition of the host immune response and allowing the host to induce either an innate and/or a protective immune response. In this way, live-attenuated viruses may be developed for further characterization as novel vaccine candidates.

Large DNA viruses, such as *Variola virus*, the causative agent of smallpox, contain large genomes of almost 200 kb encoding several genes that modulate the host immune response. In these situations, individual proteins may have a single function only, allowing entire genes to be knocked out to attenuate the virus, as is seen with vaccinia virus. In some cases, sufficient redundancy may exist requiring the deletion of multiple genes or whole genome sections to sufficiently attenuate the virus. For example, *African swine fever virus*, from another family of large DNA viruses, causes a hemorrhagic disease in domestic pigs and encodes a number of genes that modulate the host immune response. One such gene, *A238L*, encodes a protein which inhibits the transcription factor NF-κB, a central regulator of inflammation and the innate immune response [19]. However, deleting this gene does not cause any difference in porcine disease [20]. Similarly, deletion of an entire multigene family with a hypothesized role in modulation of the interferon system still results in viral replication [21, 22].

However, all of the current biothreat viruses, with the exception of *Variola virus*, are RNA viruses, with limited coding capacity and proteins that have multiple functions. In these cases, the deletion of entire genes may result in a non-viable virus. This illustrates the importance of identifying the individual residues/regions responsible for specific virus-host interactions. If specific residues are identified that are important in determining virulence or in modulating the host immune response, but do not interfere in the critical functions of viral replication, mutation of these sites may lead to novel vaccine candidates. For example, recent mapping of the *Nipah virus* P gene has identified a region that is required for inhibition of STAT-1 signaling, but without disrupting its function as a cofactor for the viral polymerase [23].

The use of global approaches, particularly microarray gene expression technology, has provided a wealth of information detailing the host responses to infection. While these data are often limited to transcription rather than expression at the protein level, they have provided significant insight into which classes of host genes are up- and downregulated in response to virus infections. A novel application of this technology is to use the differentially expressed genes as a marker of upstream transcription factor activation using bioinformatics tools such as CARRIE [24, 25]. In this way, host signaling pathways activated or repressed by virus infection can be inferred and studied further, in particular those inhibited by virus-induced proteins that are critical in the pathogenesis of the virus (see below).

3. Comparative Molecular Pathogenesis

The study of virus-host interactions has helped to identify host proteins required for particular stages in the virus life cycle. These studies have also identified innate immune mechanisms that are suppressed during infection. Identifying such cellular proteins may allow the development of novel therapeutics, for example, by inhibiting a cellular kinase required for viral entry or replication [26–28]. However, it is likely that these studies may not provide the key information required to aid the development of novel vaccines as elucidating correlates of protection is not possible if the study compares candidate vaccine-infected versus mock-infected groups without including wild-type virulent virus-infected groups or vice versa. Currently, there are few high-throughput systems-level studies employing these types of comparisons in vivo. Analyses of human transcriptional responses to the yellow fever 17D vaccine have provided significant insights into the host response to this attenuated virus [29], but such studies, not surprisingly, lack inclusion of the human response to Asibi virus (the wild-type parent to vaccine strain 17D) and limit our knowledge of the molecular basis of attenuation of 17D virus. However, studies using virulent/attenuated pairs, such

as LCMV WE/Armstrong infection in macaques [15, 16], allow one set of responses to be "subtracted" from the other, filtering the dataset down to a more manageable size for further analysis and potential correlation of transcriptional profiles with pathogenesis or protection.

As higher-throughput sequencing becomes ever more affordable, whole genome sequencing of many virus strains and species is becoming more commonplace. Many of these types of study have been undertaken from an epidemiological perspective in order to trace the natural history of virus strains and determine and predict their spread. However, these approaches have potential important applications in the field of vaccinology. Genomic sequence analysis of large numbers of naturally occurring field strains of viruses will be important in the process to identify specific mutations that correlate with differences in pathogenicity of these viruses, and subsequently the process of attenuation.

Comparing the cellular responses to infection by virulent and attenuated viruses has provided a large amount of information on the molecular basis of pathogenesis that may have application to developing correlates of protection. A significant trend that seems to be emerging is that infection with attenuated viruses promotes a strong immune response, while these events are suppressed in virulent infection. A number of virus pairs have been used for these types of analyses. A microarray study of mouse brain tissue following infection with wild-type or a laboratory-adapted strain of *Rabies virus* revealed extensive activation of the inflammatory response and the type-I interferon system in mice infected with the attenuated virus [30].

Studies using *Pichindé* virus, a guinea pig model for Lassa fever, have taken advantage of two passage variants of the virus which cause either a severe hemorrhagic fever or a mild, self-limiting infection from which animals recover [31, 32]. Proteomic and kinomic level studies using this system have been analyzed using pathway analysis and have shown that infection with the attenuated virus induced significantly more cellular signaling events and immune response activation than infection with the virulent virus, which more closely resembles patterns of protein expression and kinase activity seen following mock infection [33–35]. These systems-level analyses can be used to generate hypotheses and tested using pathway-focused assays. Analysis of NF-κB family activity showed increased expression and DNA binding of a transcriptionally repressive NF-κB protein following virulent infection [36] consistent with observations that pathogenic arenavirus infection fails to activate cells.

The genomes of the virulent and attenuated *Pichindé* viruses have been sequenced and the amino acid differences between them identified [37]. The striking differences in phenotype are caused by a small number of mutations, with three amino acid differences in the envelope glycoprotein precursor, one in the nucleoprotein, and five in the polymerase. Interestingly, while the nucleoprotein has been shown to play an important role in the inhibition of interferon induction [38, 39], the mutation observed in the *Pichindé virus* nucleoprotein does not inhibit this function [37]. Taken together with the observation described above that a single amino acid in the VP35 protein of *Ebola virus* is sufficient to disrupt the viral inhibition of innate immune signaling, it suggests that modulation of immune function is not the sole determinant of pathogenicity for this virus. The recent development of an infectious clone system [40] will allow further characterization of the roles of these mutations and shed light on the relation of individual amino acid changes to arenavirus pathogenesis.

Similar findings have been reported using non-high-throughout assays. A number of studies have compared *Lassa virus* with the related naturally nonvirulent arenavirus *Mopeia virus*. Infection of monocytes, targets of arenavirus infection in vivo and central mediators of the development of a protective immune response, with *Lassa virus* does not result in cell activation or production of TNF-α or IL-8, and consequently the host immune response to virus infection is inhibited. In contrast, IL-8 production is not suppressed and interferon signaling is activated in *Mopeia virus*-infected cells, allowing the host immune response to be induced [41, 42]. Dendritic cells, which are professional antigen-presenting cells, are also targets of hemorrhagic fever virus infection. *Lassa virus* infection of dendritic cells inhibits upregulation of the correct cell-surface expression of proteins involved in antigen presentation, adhesion, or activation, again inhibiting the development of the adaptive immune response [43]. Observations of Lassa fever patients demonstrated higher IL-8 and interferon-inducible protein 10 levels in patients who recovered from infection compared to those with fatal outcome [44], which corroborates in vitro studies suggesting that functional impairment of immune responses contributes to pathogenesis. However, given the ability of runaway immune responses to cause an immunopathology or a "cytokine storm"-like disease following infection by some viruses, immune mechanisms need to be appropriate and controlled to allow the development of the correct responses and lead to clearance of the virus. In the case of SARS coronavirus, a microarray study demonstrated that early induction of host responses and proinflammatory signaling was correlated with a fatal outcome of infection in aged mice [45]. Two attenuated variants of *Rift Valley fever virus* (RVFV), a Bunyavirus and category A priority pathogen, have also provided insights into how differences in viral genome, host interactions, and pathogenesis can identify important viral virulence determinants. The NSs protein mediates the inhibition of the type-I interferon response, a critical mediator of the innate response to viral infection important in establishing appropriate secondary immune developments. The clone 13 strain of RVFV has a large deletion in the NSs coding region, removing its ability to inhibit the type-I interferon response, which leads to a highly attenuated phenotype in mice, but retains immunogenicity, and is a potential candidate vaccine [46–49].

Integrating the virus infection data using pathway analysis and functional systems biology tools can provide an overview of the similarities and differences between infections by either a virulent or an attenuated variant of a virus. For example, inferring the upstream transcription factors that are differentially activated between attenuated

and virulent infections from microarray data may allow the responsible signaling pathways to be identified. If activation of a specific signaling pathway is associated with the expression of a subset of genes associated with the development of a protective immune response, this pathway could be targeted for stimulation as a novel adjuvant strategy, or its activity assayed as a proxy for potential efficacy if screening several vaccine candidates.

While the continued discovery of the types of virulence determinants discussed above by the study of virus strains that differ in pathogenesis illustrates the power and simplicity of this approach, the above examples demonstrate the important point that characterization of the appropriate immunological events associated with protection differs from pathogen to pathogen and from model to model. As datasets from global host-pathogen interaction studies continue to become available, novel meta-analyses at the systems level may allow particular patterns associated with pathogen, host, and disease pathology to be observed, which can feed back into models for vaccine development. At present, most studies are concentrating on the innate immune response, in particular the interferon signaling pathway, rather than the adaptive immune response. The major difficulty of all these studies is undertaking studies on wild-type virulent virus in human cells due to biocontainment requirements for studies in cell culture and animal models, while studies in humans are very hard to undertake as patient samples do not usually become available until late in the disease course or from deceased individuals.

4. Applications of Systems Biology and Machine Learning in Virus-Host Interactions

High-throughput "omics" experiments provide an almost overwhelming amount of data for validation and further studies, which makes interpretation difficult. As such, analysis is often limited to a heatmap following hierarchical clustering analysis to obtain a broad overview of the dataset and the relationships between the treatment groups, with the transcripts showing the highest fold-changes selected for confirmation by RT-PCR and further experiments. With the majority of recent work focusing on either the underlying molecular biology of virus infection and cellular responses or disease pathology at the histological level, emerging bioinformatics tools, and the continued use of experimental methods that characterize cellular regulation at a number of levels, from mRNA expression to metabolomics, will allow these two extremes to be linked, providing a holistic "systems-level" view of pathogenesis [50]. Importantly, these levels of increased understanding will be beneficial to many areas of research, including vaccinology, as we will ultimately be able to screen vaccine candidates based on the effects on particular host genes and/or proteins following infection (e.g., lack of inhibition of interferon signaling pathway) of a particular cell type, and later in appropriate animal models, with the vaccine candidates.

Techniques such as *k*-means clustering can be of use in "filtering" large numbers of genes down to manageable sizes for further analysis. In this approach, the differentially expressed genes are portioned into a number of clusters determined arbitrarily. An iterative algorithm then refines the clustering until each transcript is in the cluster with the closest mean average. In this way, transcripts with similar expression characteristics will collect in the same cluster. If gene expression analysis is used to compare the cellular responses following infection with mock, wild-type and attenuated or candidate vaccine strain viruses, groups of genes that may be specifically up or downregulated by a particular virus variant, will become apparent. These genes may be selected for further investigation using other methods. Figure 1 shows a generic example of the type of output produced following *k*-means clustering. As can be seen, the three clusters show groups of genes that are upregulated following wild-type but not attenuated virus infection, genes upregulated as a consequence of both infections, and a cluster of genes which are upregulated by attenuated but not wild-type infection. This latter group of genes may represent those that are required to initiate a protective immune response and indicate potential biomarkers of protection.

An approach that is becoming more widely used in this field is that of pathway analysis. Using knowledgebases of known interactions between proteins, DNA, RNA, and small molecules, large datasets can be placed into a more "functional" context on the basis of their known interactions. Pathway analysis tools such as Cytoscape [51, 52] and the Ingenuity Pathway Analysis software [53] are increasingly used to understand protein and gene interactions and construct so-called "signaling networks", often following characterization of gene and protein expression by microarray or proteomics analysis. Figure 2 shows an example network constructed from microarray analysis of viral infection.

These types of approaches have also been used to analyze interactions and network relationships between viral proteins as well as to characterize the interaction of viral proteins with cellular proteins [54–56]. However, to date, these types of mapping analyses have required high-throughput yeast two-hybrid type assays to generate the input data. Consequently, this type of approach may not yet be suited to the high-throughput identification of protein interaction networks for large numbers of virus strains, including potentially attenuating mutations. Furthermore, the yeast 2-hybrid system does not always generate practical meaningful interactions and a significant number of experiments are required to substantiate protein interactions. However, machine learning approaches have been applied to the determination of protein-protein interactions. In particular, support vector machines have been used to predict protein-protein interactions on the basis of sequence information alone.

Support vector machines (SVMs) are a machine learning approach which allows binary classification following training of an algorithm on a "test-set" of data. The SVM is trained on a dataset which includes a number of parameters where the classification is known. The SVM develops an algorithm that, when given a set of parameters (n) of unknown classification, can be used to predict the class on the basis of the training set. The SVM transforms the data,

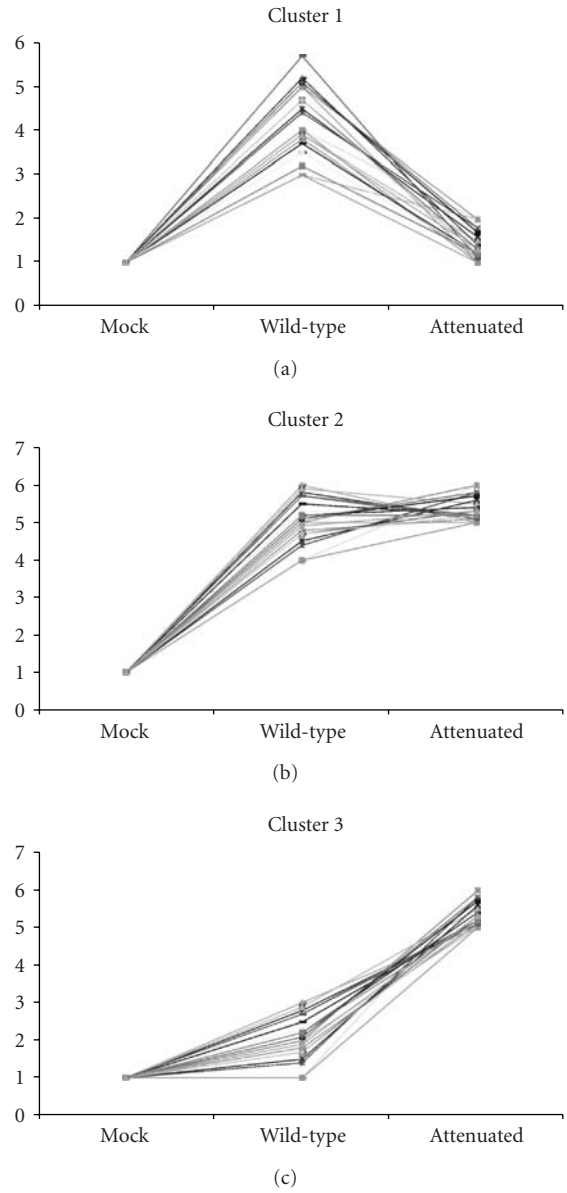


Figure 1: A generic example of analysis using *k*-means clustering. Data, such as from an mRNA expression microarray following mock, virulent, and attenuated infection of cell cultures or animals, are placed into clusters, on the basis of a similar pattern of expression to other genes in that cluster; lines represent an individual transcript. As can be seen, this type of analysis can quickly determine groups of transcripts which may associate with pathogenesis or protection.

places it into *n*-dimensional space, and attempts to separate the data by class by defining a hyperplane through this "feature space" (Figure 3). The advantage of these types of methods is that the number of parameters can be tailored to strike a balance between what is experimentally feasible and predictive accuracy. As the number of parameters increases, classification accuracy may improve, but collecting the data may become less feasible experimentally. With training sets constructed from known protein-protein interactions defined experimentally, SVM-based approaches have shown increasing accuracy for predicting these interactions [57–59]. Using these approaches, it may be possible to screen large numbers of protein sequences in silico to elucidate amino acid mutations to disrupt protein-protein interactions that may attenuate a virus, such as the single amino acid substitution that significantly attenuated *Ebola virus* [17].

Computational approaches have also demonstrated their power in de novo attenuation of viruses. Coleman et al. coined the term "synthetic attenuated virus engineering", SAVE, to describe their approach for attenuating virulent viruses [60]. Their approach takes advantage of the redundancy of the triplet genetic code and the fact that there is often a codon usage bias, where particular codons are used more or less often to code for a particular amino acid than would be predicted by chance. They developed a computer algorithm that recoded a given amino acid sequence at the genetic level, changing the codon pairs to those which were under-or overrepresented in the wild-type sequence. Their algorithm also included other features such as RNA folding-free energy, a factor shown to be implicated in hostadaptation of influenza virus [61]. By altering codon pair usage, they were able to construct a poliovirus that was attenuated in mice and provided protective immunity [60]. Of particular interest with this type of strategy is that the protein sequence remains unchanged from the wild-type virus. This reduces the likelihood of reduced protection by the attenuating mutations disrupting important epitopes.

As discussed above, the majority of high-throughput data to emerge from the biodefense field has been microarray mRNA expression data, rather than from high-throughput protein interaction data that directly lend itself to network analysis. However, genomics- and proteomics-level analyses have produced interaction networks that are beginning to yield some consistent and overlapping signaling pathways across different model systems, and novel computational approaches are providing alternative strategies for virus comparison and attenuation.

5. Towards Targeted Vaccine Design

Traditionally, live-attenuated viruses have been developed by serial passage of wild-type viruses in culture. The live-attenuated yellow fever 17D and the polio Sabin vaccines were developed in this way. However, the attenuated poliovirus vaccine can revert to virulence and has been associated with vaccine-associated poliomyelitis [62]. In recent years, the yellow fever 17D vaccine has caused a number of serious adverse effects with a significant case fatality rate [63–65]. Some of these outcomes may be associated with polymorphisms in the CCR5 chemokine receptor and RANTES chemokine genes in some vaccinees [66]. This finding complicates the vaccine safety field. If the vaccine-associated disease is associated with hostpolymorphisms rather than with virus reversion, then determining vaccine safety becomes increasingly complicated and suggests that in the 21st century each vaccine may become restricted to individuals with a particular host genetic background.

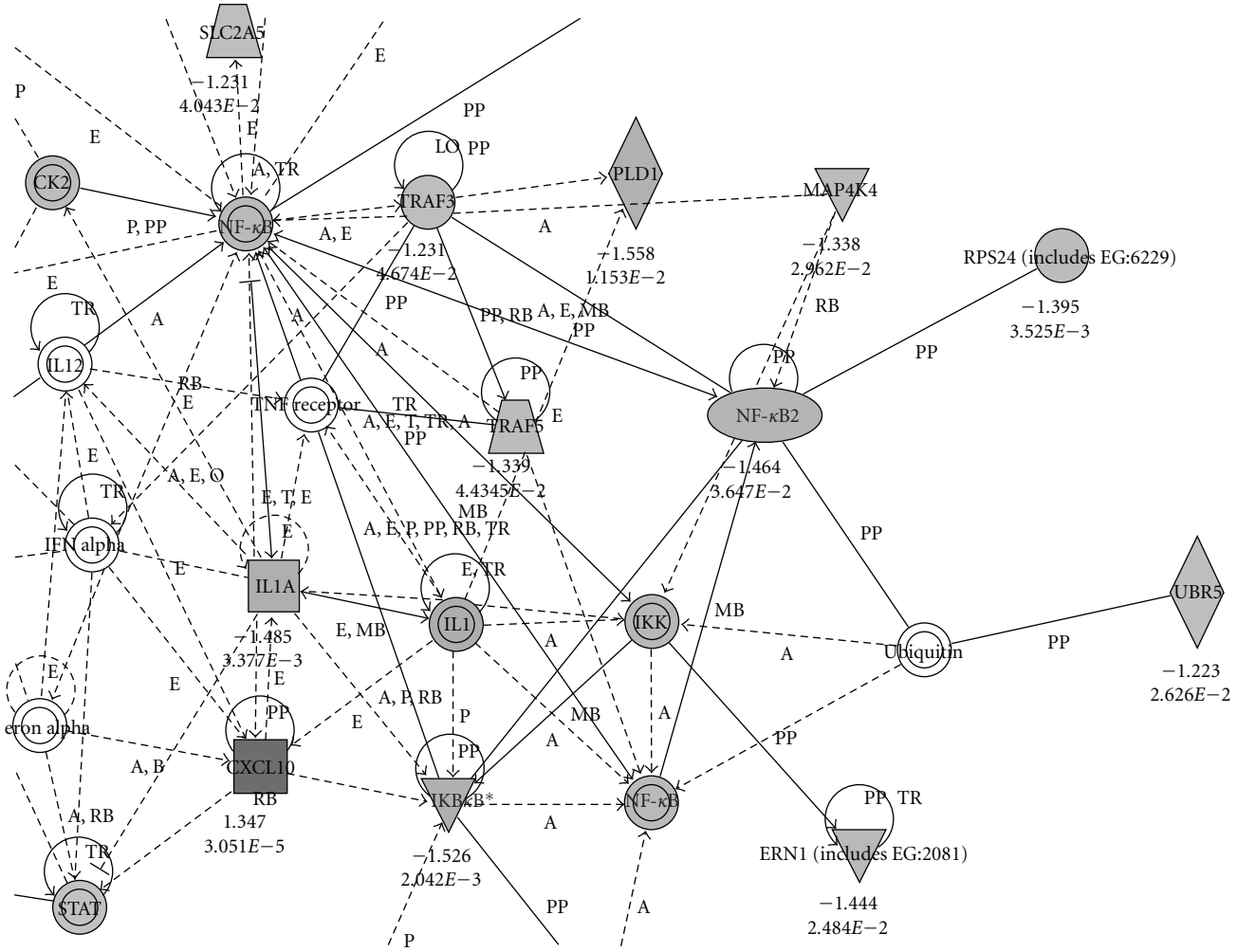


Figure 2: Placing datasets into a functional context using pathway analysis. Example transcriptional expression data were uploaded to the Ingenuity Pathway Analysis application (http://www.ingenuity.com/), and its knowledgebase was used to construct signaling networks on the basis of known interactions from the literature. These approaches can allow the visualization of networks associated with viral infection and identify signaling "hubs" which may act as master switches of the host response.

As high-throughput methods become more affordable, personalized vaccinology may be a strategy that becomes more feasible. Personalized vaccinology refers to the ability to tailor particular vaccines to specific individuals [67, 68]. If, for example, adverse responses to the yellow fever vaccine can be attributed to a specific genetic polymorphism, then the vaccinee can be screened for this polymorphism, and decisions whether or not to vaccinate can be made on that basis. If alternate vaccines are available, which perhaps offer a shorter-lived immunity but without reported adverse effects, appropriate risk assessments can be made and a course of action determined, in the best interests of the individual vaccinee.

The two basic requirements for vaccine design are the induction of a protective immune response and safety. Bioinformatics can assist with both of these. Studies of human responses to the yellow fever vaccine using mRNA microarrays have revealed the immune pathways responsible for protection [29, 69, 70]. Unfortunately, as these studies were performed in human volunteers, there is no directly comparable dataset following exposure to wild-type yellow fever virus infection. As discussed above, this raises one of the problems of utilizing system biology in vaccine development; namely, studies with candidate and licensed vaccines in humans are comparatively easy while obtaining samples from wild-type infections and serious adverse events are very difficult for a variety of reasons. However, these studies do reveal significant insights at the molecular level into how a historically very successful vaccine induces protection. These studies showed activation of several arms of the immune response but identified key transcription factors which acted as "master switches" in immune response development. Pathway analysis revealed these to be central "hubs": highly connected "nodes" in a signaling network. There are a number of biodefense vaccines in Investigation New Drug (IND) Status, due to limited studies on safety and efficacy, which are administered to individuals who are at risk from the disease. Bioinformatic studies on vaccinees who receive these IND vaccines may help generate important data that have application to safety and efficacy.

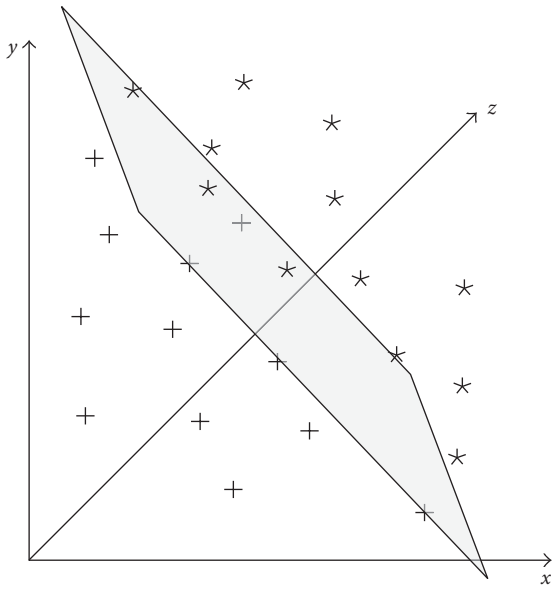


Figure 3: Separation of datasets using support vector machines. In this representation, three variables are observed for each data point, leading to a three-dimensional feature space. Transformation of the data into the feature space allows the two classes of observation to be split by the hyperplane (grey).

These types of analysis may also assist in the development of novel adjuvants. Microarray studies characterizing the global host responses to the adjuvants alum, CpG, MF59, and LTK63 have identified transcript signatures common between and specific to these adjuvants [71, 72]. By understanding how these transcriptional responses correlate with the development of specific immune mechanisms, appropriate pathways could be stimulated with specific vaccines to induce the appropriate immune effector functions for that pathogen: as efforts to design vaccines to be as safe as possible continue, a potential side effect could be over-attenuation leading to increased safety at the expense of immunogenicity and lasting protection. A possible solution could be the use of novel adjuvants to stimulate the signaling pathways associated with a strong immunogenic response. As an example, inflammasome activation was found to be a central regulator of the immune response to yellow fever vaccine [69]. Interestingly, a novel nanoparticle-based vaccine delivery method has recently been reported which includes lipopolysaccharide to activate the inflammasome alongside delivery of a protein subunit vaccine [73]. Combining findings such as these will be central to future themes of vaccinology. As our understanding of the correlates of protection for currently licensed vaccines increases, we may be able to target specific cellular pathways to tailor immune responses to a new generation of highly attenuated, safe vaccines.

Recently, a machine learning algorithm, based on a support vector machine, was used to identify potential correlates of immune protection, with potential transcript-based indicators of cell-mediated and antibody responses identified [29]. While these findings require further validation and comparison with potential indicators for responses to other vaccines, they suggest that bioinformatics could provide tools to predict immunization outcome soon after vaccination. Additionally, pathway analysis can be used to screen for transcriptional responses which are associated with other diseases or adverse effects, potentially providing a means to assay vaccine safety in vivo early in the development process.

Currently, a limiting factor in the rational attenuation of biodefense viruses for vaccine design is in obtaining datasets from equivalent platforms and model systems, facilitating meta-analyses to determine potential correlates of attenuation. However, while the virus-host interaction field may not yet be addressable in this way, efforts are being made to compile easily accessible databases of viral sequences for query. One such database has collected over 300 protein sequences from pathogenic arenaviruses to aid in epitope discovery studies [74]. Combining these types of databases with patient data such as antibody titer or disease severity may allow consistent peptide sequences which correlate with immune protection to be determined. This type of approach, combined with computational MHC epitope prediction, will significantly further our ability to rationally develop immunogenic vaccine candidates.

6. Summary and Conclusions

The increased availability and affordability of high-throughput technologies, such as whole genome sequencing and "omics" analyses, will allow the continuing development of large, multilevel datasets to characterize host responses to pathogens of differing virulence, and the changes in viral sequences responsible. As bioinformatics continues to evolve, and as microbiologists, immunologists, and vaccinologists become more familiar with the potential of these technologies, the stage will be set for novel approaches to vaccine design for biodefense, combining rational attenuation with immunogenicity and safety testing. By integrating these approaches early in the process, and using a combination of in vitro and in silico techniques, vaccine candidates with the most promise can be selected for further testing with a greater confidence, and funding can be better appropriated.

The use of comparative studies between virulent and attenuated strains could provide a wealth of data to assist in vaccine development. While many datasets have been published, comparison across different experimental systems remains an issue. Ideally, cell lines, timepoints, and other experimental criteria need to be standardized to allow informed comparison of the results. The ability to perform meta-analyses across different studies could allow a better understanding of what host responses are required to lead to a long-lasting protective immune response. If a dataset which characterized the host responses to yellow fever vaccine could be compared with similar datasets showing responses to other live vaccines, for example, smallpox vaccine and Junin vaccine, along with response to the appropriate wild-type agents, we will undoubtedly find several "groups" of similarities and differences which associate with protection,

virulence, or type of virus or disease pathology. These approaches to vaccine design illustrate how vaccinology can use the data from basic science, and particularly studies of molecular pathogenesis, to develop novel vaccine candidates. By identifying common cellular responses to attenuated virus vaccines, across different viral families and different wild-type disease pathologies, we may characterize a central "core" of responses which are always activated by a historically successful vaccine, regardless of virus type. These responses can guide us in the development of novel vaccine candidates or adjuvants.

As new viruses continue to emerge and the risks from emerging and remerging pathogens remain, the need for vaccines against these pathogens remains significant. The introduction and spread of *West Nile virus* to North America illustrates how quickly viruses can become established in a new area and present novel threats to public health [75]. The continuing investment in biodefense and emerging pathogens will result in even greater amounts of basic data to input into novel bioinformatics tools. As our understanding of the molecular determinants of protection and pathogenesis continues to grow, the development of safe, effective vaccines to these important pathogens will continue to accelerate.

Acknowledgments

The first author is supported by the Sealy Center for Vaccine Development, the Departments of Pathology and Microbiology & Immunology by the Department of Defense, through a National Biocontainment Training Center Fellowship. The authors thank Dr. Heidi Spratt for helpful discussions.

References

[1] R. Carrion Jr., J. L. Patterson, C. Johnson, et al., "A ML29 reassortant virus protects guinea pigs against a distantly related Nigerian strain of Lassa virus and can provide sterilizing immunity," *Vaccine*, vol. 25, no. 20, pp. 4093–4102, 2007.

[2] I. S. Lukashevich, J. Patterson, R. Carrion, et al., "A live attenuated vaccine for Lassa fever made by reassortment of Lassa and Mopeia viruses," *Journal of Virology*, vol. 79, no. 22, pp. 13934–13942, 2005.

[3] P. J. Bredenbeek, R. Molenkamp, W. J. M. Spaan, et al., "A recombinant Yellow Fever 17D vaccine expressing Lassa virus glycoproteins," *Virology*, vol. 345, no. 2, pp. 299–304, 2006.

[4] S. P. Fisher-Hoch, L. Hutwagner, B. Brown, and J. B. McCormick, "Effective vaccine for lassa fever," *Journal of Virology*, vol. 74, no. 15, pp. 6777–6783, 2000.

[5] T. W. Geisbert, S. Jones, E. A. Fritz, et al., "Development of a new vaccine for the prevention of Lassa fever," *PLoS Medicine*, vol. 2, no. 6, article e183, 2005.

[6] S. P. Fisher-Hoch, J. B. McCormick, D. Auperin, et al., "Protection of rhesus monkeys from fatal Lassa fever by vaccination with a recombinant vaccinia virus containing the Lassa virus glycoprotein gene," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 1, pp. 317–321, 1989.

[7] M. Bharadwaj, C. R. Lyons, I. A. Wortman, and B. Hjelle, "Intramuscular inoculation of Sin Nombre hantavirus cDNAs induces cellular and humoral immune responses in BALB/c mice," *Vaccine*, vol. 17, no. 22, pp. 2836–2843, 1999.

[8] M. Bharadwaj, K. Mirowsky, C. Ye, et al., "Genetic vaccines protect against Sin Nombre hantavirus challenge in the deer mouse (Peromyscus maniculatus)," *Journal of General Virology*, vol. 83, no. 7, pp. 1745–1751, 2002.

[9] A. Tuffs, "Experimental vaccine may have saved Hamburg scientist from Ebola fever," *British Medical Journal*, vol. 338, article b1223, 2009.

[10] J. I. Maiztegui, K. T. McKee Jr., J. G. B. Oro, et al., "Protective efficacy of a live attenuated vaccine against argentine hemorrhagic fever," *Journal of Infectious Diseases*, vol. 177, no. 2, pp. 277–283, 1998.

[11] S. Bambini and R. Rappuoli, "The use of genomics in microbial vaccine development," *Drug Discovery Today*, vol. 14, no. 5-6, pp. 252–260, 2009.

[12] R. Rappuoli and A. Covacci, "Reverse vaccinology and genomics," *Science*, vol. 302, no. 5645, p. 602, 2003.

[13] N. Sevilla, D. B. McGavern, C. Teng, S. Kunz, and M. B. A. Oldstone, "Viral targeting of hematopoietic progenitors and inhibition of DC maturation as a dual strategy for immune subversion," *Journal of Clinical Investigation*, vol. 113, no. 5, pp. 737–745, 2004.

[14] E. A. Fritz, J. B. Geisbert, T. W. Geisbert, L. E. Hensley, and D. S. Reed, "Cellular immune response to marburg virus infection in cynomolgus macaques," *Viral Immunology*, vol. 21, no. 3, pp. 355–363, 2008.

[15] M. M. Djavani, O. R. Crasta, J. C. Zapata, et al., "Early blood profiles of virus infection in a monkey model for Lassa fever," *Journal of Virology*, vol. 81, no. 15, pp. 7960–7973, 2007.

[16] M. Djavani, O. R. Crasta, Y. Zhang, et al., "Gene expression in primate liver during viral hemorrhagic fever," *Virology Journal*, vol. 6, p. 20, 2009.

[17] A. L. Hartman, L. Ling, S. T. Nichol, and M. L. Hibberd, "Whole-genome expression profiling reveals that inhibition of host innate immune response pathways by Ebola virus can be reversed by a single amino acid change in the VP35 protein," *Journal of Virology*, vol. 82, no. 11, pp. 5348–5358, 2008.

[18] A. L. Hartman, B. H. Bird, J. S. Towner, Z.-A. Antoniadou, S. R. Zaki, and S. T. Nichol, "Inhibition of IRF-3 activation by VP35 is critical for the high level of virulence of ebola virus," *Journal of Virology*, vol. 82, no. 6, pp. 2699–2704, 2008.

[19] P. P. Powell, L. K. Dixon, and R. M. E. Parkhouse, "An IκB homolog encoded by African swine fever virus provides a novel mechanism for downregulation of proinflammatory cytokine responses in host macrophages," *Journal of Virology*, vol. 70, no. 12, pp. 8527–8533, 1996.

[20] J. G. Neilan, Z. Lu, G. F. Kutish, L. Zsak, T. L. Lewis, and D. L. Rock, "A conserved African swine fever virus IκB homolog, 5EL, is nonessential for growth in vitro and virulence in domestic swine," *Virology*, vol. 235, no. 2, pp. 377–385, 1997.

[21] C. L. Afonso, M. E. Piccone, K. M. Zaffuto, et al., "African swine fever virus multigene family 360 and 530 genes affect host interferon response," *Journal of Virology*, vol. 78, no. 4, pp. 1858–1864, 2004.

[22] L. Zsak, Z. Lu, T. G. Burrage, et al., "African swine fever virus multigene family 360 and 530 genes are novel macrophage host range determinants," *Journal of Virology*, vol. 75, no. 7, pp. 3066–3076, 2001.

[23] M. J. Ciancanelli, V. A. Volchkova, M. L. Shaw, V. E. Volchkov, and C. F. Basler, "Nipah virus sequesters inactive STAT1 in the nucleus via a P gene-encoded mechanism," *Journal of Virology*, vol. 83, no. 16, pp. 7828–7841, 2009.

[24] P. M. Haverty, M. C. Frith, and Z. Weng, "CARRIE web service: automated transcriptional regulatory network inference and interactive analysis," *Nucleic Acids Research*, vol. 32, pp. W213–W216, 2004.

[25] P. M. Haverty, U. Hansen, and Z. Weng, "Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification," *Nucleic Acids Research*, vol. 32, no. 1, pp. 179–188, 2004.

[26] E. M. Vela, G. C. Bowick, N. K. Herzog, and J. F. Aronson, "Genistein treatment of cells inhibits arenavirus infection," *Antiviral Research*, vol. 77, no. 2, pp. 153–156, 2008.

[27] E. M. Vela, G. C. Bowick, N. K. Herzog, and J. F. Aronson, "Exploring kinase inhibitors as therapies for human arenavirus infections," *Future Virology*, vol. 3, no. 3, pp. 243–251, 2008.

[28] A. A. Kolokoltsov, M. F. Saeed, A. N. Freiberg, M. R. Holbrook, and R. A. Davey, "Identification of novel cellular targets for therapeutic intervention against ebola virus infection by siRNA screening," *Drug Development Research*, vol. 70, no. 4, pp. 255–265, 2009.

[29] T. D. Querec, R. S. Akondy, E. K. Lee, et al., "Systems biology approach predicts immunogenicity of the yellow fever vaccine in humans," *Nature Immunology*, vol. 10, no. 1, pp. 116–125, 2009.

[30] Z. W. Wang, L. Sarmento, Y. Wang, et al., "Attenuated rabies virus activates, while pathogenic rabies virus evades, the host innate immune responses in the central nervous system," *Journal of Virology*, vol. 79, no. 19, pp. 12554–12565, 2005.

[31] P. B. Jahrling, R. A. Hesse, and J. B. Rhoderick, "Pathogenesis of a pichinde virus strain adapted to produce lethal infections in guinea pigs," *Infection and Immunity*, vol. 32, no. 2, pp. 872–880, 1981.

[32] J. F. Aronson, N. K. Herzog, and T. R. Jerrells, "Pathological and virological features of arenavirus disease in guinea pigs: comparison of two Pichinde virus strains," *American Journal of Pathology*, vol. 145, no. 1, pp. 228–235, 1994.

[33] G. C. Bowick, S. M. Fennewald, B. L. Elsom, et al., "Differential signaling networks induced by mild and lethal hemorrhagic fever virus infections," *Journal of Virology*, vol. 80, no. 20, pp. 10248–10252, 2006.

[34] G. C. Bowick, S. M. Fennewald, E. P. Scott, et al., "Identification of differentially activated cell-signaling networks associated with Pichinde virus pathogenesis by using systems kinomics," *Journal of Virology*, vol. 81, no. 4, pp. 1923–1933, 2007.

[35] G. C. Bowick, H. M. Spratt, A. E. Hogg, et al., "Analysis of the differential host cell nuclear proteome induced by attenuated and virulent hemorrhagic arenavirus infection," *Journal of Virology*, vol. 83, no. 2, pp. 687–700, 2009.

[36] G. C. Bowick, S. M. Fennewald, L. Zhang, et al., "Attenuated and lethal variants of pichinde virus induce differential patterns of NF-κB activation suggesting a potential target for novel therapeutics," *Viral Immunology*, vol. 22, no. 6, pp. 457–462, 2009.

[37] S. Lan, L. McLay, J. Aronson, H. Ly, and Y. Liang, "Genome comparison of virulent and avirulent strains of the Pichinde arenavirus," *Archives of Virology*, vol. 153, no. 7, pp. 1241–1250, 2008.

[38] L. Martinez-Sobrido, E. I. Zuniga, D. Rosario, A. Garcia-Sastre, and J. C. de la Torre, "Inhibition of the type I interferon response by the nucleoprotein of the prototypic arenavires lymphocytic choriomeningitis virus," *Journal of Virology*, vol. 80, no. 18, pp. 9192–9199, 2006.

[39] L. Martinez-Sobrido, P. Giannakas, B. Cubitt, A. Garcia-Sastre, and J. C. de la Torre, "Differential inhibition of type I interferon induction by arenavirus nucleoproteins," *Journal of Virology*, vol. 81, no. 22, pp. 12696–12703, 2007.

[40] S. Lan, L. M. Schelde, J. Wang, N. Kumar, H. Ly, and Y. Liang, "Development of infectious clones for virulent and avirulent pichinde viruses: a model virus to study arenavirus-induced hemorrhagic fevers," *Journal of Virology*, vol. 83, no. 13, pp. 6357–6362, 2009.

[41] I. S. Lukashevich, R. Maryankova, A. S. Vladyko, et al., "Lassa and Mopeia virus replication in human monocytes/macrophages and in endothelial cells: different effects on IL-8 and TNF-α gene expression," *Journal of Medical Virology*, vol. 59, no. 4, pp. 552–560, 1999.

[42] D. Pannetier, C. Faure, M.-C. Georges-Courbot, V. Deubel, and S. Baize, "Human macrophages, but not dendritic cells, are activated and produce alpha/beta interferons in response to Mopeia virus infection," *Journal of Virology*, vol. 78, no. 19, pp. 10516–10524, 2004.

[43] S. Baize, J. Kaplon, C. Faure, D. Pannetier, M.-C. Georges-Courbot, and V. Deubel, "Lassa virus infection of human dendritic cells and macrophages is productive but fails to activate cells," *Journal of Immunology*, vol. 172, no. 5, pp. 2861–2869, 2004.

[44] S. Mahanty, D. G. Bausch, R. L. Thomas, et al., "Low levels of interleukin-8 and interferon-inducible protein-10 in serum are associated with fatal infections in acute Lassa fever," *Journal of Infectious Diseases*, vol. 183, no. 12, pp. 1713–1721, 2001.

[45] B. Rockx, T. Baas, G. A. Zornetzer, et al., "Early upregulation of acute respiratory distress syndrome-associated cytokines promotes lethal disease in an aged-mouse model of severe acute respiratory syndrome coronavirus infection," *Journal of Virology*, vol. 83, no. 14, pp. 7062–7074, 2009.

[46] R. Muller, J.-F. Saluzzo, N. Lopez, et al., "Characterization of clone 13, a naturally attenuated avirulent isolate of Rift Valley fever virus, which is altered in the small segment," *American Journal of Tropical Medicine and Hygiene*, vol. 53, no. 4, pp. 405–411, 1995.

[47] A. Billecocq, M. Spiegel, P. Vialat, et al., "NSs protein of Rift Valley fever virus blocks interferon production by inhibiting host gene transcription," *Journal of Virology*, vol. 78, no. 18, pp. 9798–9806, 2004.

[48] M. Bouloy, C. Janzen, P. Vialat, et al., "Genetic evidence for an interferon-antagonistic function of Rift Valley fever virus nonstructural protein NSs," *Journal of Virology*, vol. 75, no. 3, pp. 1371–1377, 2001.

[49] P. Vialat, A. Billecocq, A. Kohl, and M. Bouloy, "The S segment of Rift Valley fever phlebovirus (Bunyaviridae) carries determinants for attenuation and virulence in mice," *Journal of Virology*, vol. 74, no. 3, pp. 1538–1543, 2000.

[50] E. E. Schadt, "Molecular networks as sensors and drivers of common human diseases," *Nature*, vol. 461, no. 7261, pp. 218–223, 2009.

[51] P. Shannon, A. Markiel, O. Ozier, et al., "Cytoscape: a software environment for integrated models of biomolecular interaction networks," *Genome Research*, vol. 13, no. 11, pp. 2498–2504, 2003.

[52] S. Killcoyne, G. W. Carter, J. Smith, and J. Boyle, "Cytoscape: a community-based framework for network modeling," *Methods in Molecular Biology*, vol. 563, pp. 219–239, 2009.

[53] S. E. Calvano, W. Xiao, D. R. Richards, et al., "A network-based analysis of systemic inflammation in humans," *Nature*, vol. 437, no. 7061, pp. 1032–1037, 2005.

[54] E. Fossum, C. C. Friedel, S. V. Rajagopala, et al., "Evolutionarily conserved herpesviral protein interaction networks," *PLoS Pathogens*, vol. 5, no. 9, Article ID e1000570, 2009.

[55] P. Uetz, Y.-A. Dong, C. Zeretzke, et al., "Herpesviral protein networks and their interaction with the human proteome," *Science*, vol. 311, no. 5758, pp. 239–242, 2006.

[56] M. A. Calderwood, K. Venkatesan, L. Xing, et al., "Epstein-Barr virus and virus human protein interaction maps," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 18, pp. 7606–7611, 2007.

[57] S. Dohkan, A. Koike, and T. Takagi, "Improving the performance of an SVM-based method for predicting protein-protein interactions," *In Silico Biology*, vol. 6, no. 6, pp. 515–529, 2006.

[58] A. Koike and T. Takagi, "Prediction of protein-protein interaction sites using support vector machines," *Protein Engineering, Design and Selection*, vol. 17, no. 2, pp. 165–173, 2004.

[59] J. Shen, J. Zhang, X. Luo, et al., "Predicting protein-protein interactions based only on sequences information," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 11, pp. 4337–4341, 2007.

[60] J. R. Coleman, D. Papamichail, S. Skiena, B. Futcher, E. Wimmer, and S. Mueller, "Virus attenuation by genome-scale changes in codon pair bias," *Science*, vol. 320, no. 5884, pp. 1784–1787, 2008.

[61] R. Brower-Sinning, D. M. Carter, C. J. Crevar, E. Ghedin, T. M. Ross, and P. V. Benos, "The role of RNA folding free energy in the evolution of the polymerase genes of the influenza A virus," *Genome Biology*, vol. 10, no. 2, article R18, 2009.

[62] A. J. Cann, G. Stanway, P. J. Hughes, et al., "Reversion to neurovirulence of the live-attenuated Sabin type 3 oral poliovirus vaccine," *Nucleic Acids Research*, vol. 12, no. 20, pp. 7787–7792, 1984.

[63] H.-G. Bae, C. Domingo, A. Tenorio, et al., "Immune response during adverse events after 17D-derived yellow fever vaccination in Europe," *Journal of Infectious Diseases*, vol. 197, no. 11, pp. 1577–1584, 2008.

[64] C. Domingo and M. Niedrig, "Safety of 17D derived yellow fever vaccines," *Expert Opinion on Drug Safety*, vol. 8, no. 2, pp. 211–221, 2009.

[65] A. D. Barrett and D. E. Teuwen, "Yellow fever vaccine—how does it work and why do rare cases of serious adverse events take place?" *Current Opinion in Immunology*, vol. 21, no. 3, pp. 308–313, 2009.

[66] B. Pulendran, J. Miller, T. D. Querec, et al., "Case of yellow fever vaccine-associated viscerotropic disease with prolonged viremia, robust adaptive immune responses, and polymorphisms in CCR5 and RANTES genes," *Journal of Infectious Diseases*, vol. 198, no. 4, pp. 500–507, 2008.

[67] G. A. Poland, I. G. Ovsyannikova, and R. M. Jacobson, "Personalized vaccines: the emerging field of vaccinomics," *Expert Opinion on Biological Therapy*, vol. 8, no. 11, pp. 1659–1667, 2008.

[68] G. A. Poland, I. G. Ovsyannikova, and R. M. Jacobson, "Application of pharmacogenomics to vaccines," *Pharmacogenomics*, vol. 10, no. 5, pp. 837–852, 2009.

[69] D. Gaucher, R. Therrien, N. Kettaf, et al., "Yellow fever vaccine induces integrated multilineage and polyfunctional immune responses," *Journal of Experimental Medicine*, vol. 205, no. 13, pp. 3119–3131, 2008.

[70] P. G. Thomas and P. C. Doherty, "Rules to 'prime' by," *Nature Immunology*, vol. 10, no. 1, pp. 14–16, 2009.

[71] F. Mosca, E. Tritto, A. Muzzi, et al., "Molecular and cellular signatures of human vaccine adjuvants," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 105, no. 30, pp. 10501–10506, 2008.

[72] E. Tritto, A. Muzzi, I. Pesce, et al., "The acquired immune response to the mucosal adjuvant LTK63 imprints the mouse lung with a protective signature," *Journal of Immunology*, vol. 179, no. 8, pp. 5346–5357, 2007.

[73] S. L. Demento, S. C. Eisenbarth, H. G. Foellmer, et al., "Inflammasome-activating nanoparticles as modular systems for optimizing vaccine efficacy," *Vaccine*, vol. 27, no. 23, pp. 3013–3021, 2009.

[74] H.-H. Bui, J. Botten, N. Fusseder, et al., "Protein sequence database for pathogenic arenaviruses," *Immunome Research*, vol. 3, no. 1, 2007.

[75] F. A. Murphy, "Emerging zoonoses: the challenge for public health and biodefense," *Preventive Veterinary Medicine*, vol. 86, no. 3-4, pp. 216–223, 2008.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 459798, 16 pages
doi:10.1155/2010/459798

Review Article

Inflammatory and Autoimmune Reactions in Atherosclerosis and Vaccine Design Informatics

Michael Jan,[1,2,3] Shu Meng,[1,2] Natalie C. Chen,[1,2] Jietang Mai,[1,2] Hong Wang,[1,2,3] and Xiao-Feng Yang[1,2]

[1] *Department of Pharmacology, Temple University School of Medicine, Philadelphia, PA 19140, USA*
[2] *Cardiovascular Research Center, Temple University School of Medicine, Philadelphia, PA 19140, USA*
[3] *Sol Sherry Thrombosis Research Center, Temple University School of Medicine, Philadelphia, PA 19140, USA*

Correspondence should be addressed to Xiao-Feng Yang, xfyang@temple.edu

Received 29 October 2009; Revised 15 January 2010; Accepted 28 January 2010

Academic Editor: Yongqun Oliver He

Atherosclerosis is the leading pathological contributor to cardiovascular morbidity and mortality worldwide. As its complex pathogenesis has been gradually unwoven, the regime of treatments and therapies has increased with still much ground to cover. Active research in the past decade has attempted to develop antiatherosclerosis vaccines with some positive results. Nevertheless, it remains to develop a vaccine against atherosclerosis with high affinity, specificity, efficiency, and minimal undesirable pathology. In this review, we explore vaccine development against atherosclerosis by interpolating a number of novel findings in the fields of vascular biology, immunology, and bioinformatics. With recent technological breakthroughs, vaccine development affords precision in specifying the nature of the desired immune response—useful when addressing a disease as complex as atherosclerosis with a manifold of inflammatory and autoimmune components. Moreover, our exploration of available bioinformatic tools for epitope-based vaccine design provides a method to avoid expenditure of excess time or resources.

1. Introduction

Atherosclerosis and its pursuant complications are the leading causes of death worldwide. Its complex inflammatory and autoimmune pathogenesis has recently provided much insight for treatment and therapy. The suspicion of a role for the immune system in atherosclerosis began with the discovery that atherosclerotic plaques contain cellular and molecular mediators of innate and adaptive immunity [1–4]. Studies with cytokine gene knockout mice further confirmed the role of inflammation in mediating the innate response involved in atherosclerosis [5–8]. Moreover, experiments identifying a role for T-cells in atherogenesis indicate adaptive immune and potential autoimmune mechanisms [9–12]. Despite a myriad of available treatments for an ever-increasing list of targets and risk factors, it remains to find effective treatments with the potential for not only therapeutic but also prophylactic anti-atherosclerotic benefits. Success of vaccination campaigns against infectious diseases has demonstrated the proof-of-principle that identification of antigens and pathologically associated immune responses makes vaccination an attractive therapeutic option. While the idea of a vaccine against atherosclerosis has a seemingly mythic origin, the demonstrated inflammatory and autoimmune components of atherosclerosis provide a convincing rationale to investigate such a vaccine. Vaccines have the potential to induce and/or enhance protective immune responses generated by antigens without pathogenic consequences. In fact, active investigation during the past decade has led to clinical trials for anti-atherosclerosis vaccines with some positive results. Nevertheless, it remains to develop a vaccine against atherosclerosis with high affinity, specificity, efficiency, and minimal undesirable pathology. This effort can potentially consume excessive time and resources. However, a great boon to this field is the wealth of bioinformatic tools available to design such vaccines. In this brief review, our aim is to summarize the most recent developments in designing vaccines against atherosclerosis with a focus on how to apply a variety of useful bioinformatic tools and approaches.

2. Risk Factors for Atherosclerosis

Atherosclerosis is an inflammatory disease characterized by endothelial activation and dysfunction, lipid accumulation, monocyte infiltration and differentiation, T-cell infiltration and activation, foam cell formation, and fibrosis in the lesion area. Two of its common final results, coronary and cerebrovascular disease, are the major causes of morbidity worldwide. Nowadays, atherosclerosis is increasingly considered an immune-mediated inflammatory process of the vasculature in which intense immunological activity is occurring. Hypertension, hyperlipidemia, hyperglycemia, diabetes, obesity, cigarette smoking, age, and sex are well-established risk factors for atherosclerosis [13, 14]. Recent research from our team and others' laboratories has also identified hyperhomocysteinemia as an independent risk factor for atherosclerosis [15–17]. However, these well-defined risk factors do not account for all incidences of the disease. Infectious disease, such as *Chlamydia pneumoniae* might also contribute to atherosclerotic plaque formation [18]. This latter suggestion opens the door to the further complexity of atherogenesis—a process involving inflammatory and immune responses to a number of targets derived from both foreign and self antigens.

3. Inflammatory Mechanisms: Roles of Endothelial Cells, Monocytes, and Macrophages

A mature atherosclerotic plaque contains macrophages, vascular smooth muscle cells laden with lipid, endothelial cells, and extracellular matrix. The initial pathophysiological step in atherosclerosis is endothelial injury and monocyte infiltration. Endothelial injury manifests itself as increased adhesion molecule expression on the cell surface and cytokine or chemokine expression and secretion. In atherosclerotic animal models, vascular cell-adhesion molecule 1 (VCAM-1) appears in arterial endothelial cells during the initial vascular response to cholesterol accumulation in the intima. VCAM-1 expressed in endothelial cells and very late antigen 4 (VLA-4) expressed in leukocytes are important interacting mediators for leukocytes to infiltrate into the subendothelial compartment of the vessel. In addition, the patchy distribution of adhesion molecules progresses into fatty streaks. Low-density lipoprotein receptor deficient ($Ldlr^{-/-}$) mice that express a truncated nonfunctional VCAM-1 develop less atherosclerotic plaques [19]. Intercellular adhesion molecule 1 (ICAM-1) is constitutively expressed in endothelial cells, and its expression increases in the plaque. However, there are conflicting results of ICAM-1 having proatherosclerotic effects [19, 20]. In addition to integrins, selectins are also involved in the initial steps of the process. Apolipoprotein E deficient ($ApoE^{-/-}$) mice lacking both endothelial cell selectin (E-selectin) and platelet selectin (P-selectin) have less severe atherosclerosis [21]. Monocyte chemoattractant protein 1/Chemokine (C-C motif) ligand 2 (MCP-1/CCL2) and its receptor, CC-chemokine receptor 2 (CCR2), are noted to play an important role in the initiation of atherosclerosis, probably due to the increased monocyte and T-cell attraction and infiltration into the plaque. Several studies have demonstrated that oxidized low-density lipoprotein (oxLDL) is chemoattractant and that its oxidized phospholipid components can induce endothelial activation as judged by upregulated expression of MCP-1 [22, 23]. Other chemokines are also detected in atherosclerotic plaques, such as the cell-surface anchored CX3-chemokine ligand 1 (CX3CL1), which is expressed on vascular smooth muscle cells. MCP-1 is crucial for recruiting monocytes to atherosclerotic lesions. Crossing $ApoE^{-/-}$ or $Ldlr^{-/-}$ mice with mice lacking MCP-1 or CCR2 leads to significant lesion decrease [24–26]. Many therapeutic drugs that have anti-atherosclerotic effects may work via their anti-inflammatory effects that specifically target leukocyte adhesion and/or chemotaxis. CX3-chemokine receptor 1 (CX3CR1) is expressed on monocytes and macrophages. The results from the compound deficient mice made by crossing $ApoE^{-/-}$ mice with CX3CR1 deficient mice suggest that CX3CR1 may be involved in monocyte recruitment to the vessel wall, thereby promoting atherosclerotic plaque formation [27]. These studies all suggest that chemokines and chemokine receptors are strongly involved in atherogenesis because they increase monocyte attraction and infiltration into the lesion areas in vessels. Our recent study found that hyperhomocysteinemia accelerates atherosclerosis by promoting inflammatory monocyte differentiation/macrophage accumulation in lesions in a new hyperhomocysteinemia mouse model with an inducible human cystathionine β-synthase (CBS) transgene in the background of $ApoE^{-/-}$ and $CBS^{-/-}$ compound knock-out (Tg-hCBS/$ApoE^{-/-}$/$CBS^{-/-}$) mice fed a high fat diet [17]. Thus it is seen that the risk factors for atherosclerosis serve to promote any of a number of inflammatory processes involved in the formation or progression of atherosclerotic plaques.

Vascular cells and infiltrated cells have to be activated in response to stimulation by risk factors for atherosclerosis. In the last ten years, identification of pathogen-associated molecular patterns (PAMPs) and their receptors (PRRs) has provided a bridge to link risk factors to initiation of inflammation and secretion of proinflammatory cytokines. Toll-like receptors (TLRs) are a group of PRRs that can sense a broad range of PAMPs. In addition to TLRs, PRRs also include Nod-like receptors (NLRs) [28]. The PRRs may have implications in the development of an anti-atherosclerotic vaccine for two reasons: (1) they initiate innate immune responses and inflammation by sensing exogenous and endogenous PAMPs; and (2) they regulate adaptive immune responses. There may be many TLRs present in atherosclerotic plaques, mainly on macrophages and endothelial cells [29]. Crosses of $TLR4^{-/-}$ and $ApoE^{-/-}$ mice show reduced atherosclerosis and altered plaque phenotype [30]. Other studies also show that oxLDL and endogenous heat shock protein 60 (HSP60) can bind to the TLR4-CD14 complex and trigger inflammatory reactions with the features of pro-inflammatory cytokine secretion [31, 32], suggesting that PRRs not only recognize exogenous PAMPs but also may be activated by endogenous metabolic stress. To determine the expression of components

in the TLR/NLR/inflammasome/caspase-1/interleukin (IL)-1β pathway, our team examined the expression profiles of those genes. Among 11 tissues examined, vascular tissues and heart express fewer types of TLRs and NLRs than immune and defense tissues including blood, lymph nodes, thymus, and trachea. Based on the expression data of three characterized inflammasomes (NALP1, NALP3, and IPAF), the examined tissues can be classified into three tiers: the first tier tissues include brain, placenta, blood, and thymus express inflammasome(s) in constitutive status; the second tier tissues have inflammasome(s) in nearly ready expression status (with the requirement of upregulation of one component); the third tier tissues, like heart and bone marrow, require upregulation of at least two components in order to assemble functional inflammasomes. This original model of three-tier expression of inflammasomes would suggest a new concept of tissues' inflammation privilege and provides an insight into the differences among tissues in initiating acute inflammation in response to stimuli. This model also suggests the possibility that atherogenic risk factors induce the upregulation of PRRs and inflammasome components to trigger the chronic process of inflammatory atherogenesis [33]. Innate immune system activation is a critical step in the initiation of an effective adaptive immune response; therefore, activation of a class of receptors for PAMPs is a central feature of many adjuvant systems in vaccine preparations. One member of an intracellular PRR, the NALP3 inflammasome, is activated by a number of classical adjuvants including aluminum hydroxide and saponins [34, 35]. Inflammasome activation in vitro requires signaling of both the TLR and NALP3 in antigen-presenting cells (APCs) [36].

After infiltration, monocytes undergo differentiation into macrophages, which then become foam cells that contain a lipid-laden cytosol. Although it is unclear how these monocytes differentiate and according to what signal they differentiate, scavenger receptors, a group of proteins that mediate internalization and lysosomal degradation of lipid, lipopolysaccharide, and apoptotic bodies [37], are considered to be crucial for lipid accumulation in the macrophage to finally form the foam cell. These scavenger receptors include CD36, CD68, CXCL16, scavenger receptor A (SR-A), SR-B1, and lectin-type oxLDL receptor 1 (LOX1). The internalization and digestion of self and foreign proteins will facilitate major histocompatibility complex (MHC) class II antigen presentation [38]. Their MHC II is directly involved in antigen presentation to $CD4^+$ T-cells and $CD4^+$ T-cell activation. MHC II is widely expressed or upregulated in the lesion area on macrophages, endothelial cells, and smooth muscle cells [39]. This basic function of macrophages is antigen presentation, which triggers adaptive immunity and potentially contributes to the autoimmune character of atherosclerosis. Thus, macrophages are considered to be professional APCs. Among the internalized materials, oxLDL is one of the most important contributors to atherosclerosis and pathologically foam cells laden with cholesterol and fatty acids are composed of degraded oxLDL cholesterol ester content. Studies suggest that SR-A [40], CD36, and CXCL [41] have important functions in mediating uptake of oxLDL and promotion of atherosclerotic development.

In addition, MHC class I molecules, which present antigenic epitopes to $CD8^+$ T-cells, are constitutively expressed in macrophages. MHC molecule/antigen epitope complexes bind T-cell antigen receptors to constitute the first signal in stimulating T-cells, whereas activated macrophages in lesions also upregulate T-cell costimulation molecules on the cell surface to form the second signal for T-cell activation. Presumptive dendritic cells (DCs) bearing the CD11c integrin and other markers have previously been identified in normal mouse and human aorta. DCs are proved to be particularly abundant in the cardiac valves and aortic sinus. In all aortic locations, the $CD11c^+$ cells are localized to the subintimal space with occasional processes probing the vascular lumen. Aortic DCs express little CD40 but generate low levels of CD1d, CD80, and CD86. Aortic DCs can cross-present two different protein antigens on MHC class I to $CD8^+$ TCR transgenic T-cells. In addition, after intravenous injection, aortic DCs can capture anti-CD11c antibody and cross-present ovalbumin to T-cells. These results indicate that *bona fide* DCs are a constituent of the normal aorta and cardiac valves [42]. Studies from $ApoE^{-/-}$ mice also show the involvement of T-cell costimulation through the B7s (CD80 and CD86)/CD28 pathway in atherosclerotic plaques. The other costimulatory factor binding pair, CD40 and CD40 ligand (CD40L), is also widely expressed in the lesion cells [43]. Inhibition of CD40 signaling, including genetic disruption of CD40L in $ApoE^{-/-}$ mice or treating $Ldlr^{-/-}$ mice with CD40L antibody, reduces atherosclerotic lesion formation [44]. Treatment with CD40L antibody also inhibits the progression of formed plaques and maintains their stable phenotype [45]. Thus, macrophages are involved in both innate and adaptive immune responses during atherosclerosis.

4. Autoimmune Mechanisms: New T-Helper Cell Subsets, Tregs, and Autoantigens

The discovery that the innate immune system had an active role in the inflammatory process of atherosclerosis was critical in reinventing the established paradigm [7]. When it was discovered that the adaptive immune system also had a significantly more complicated role in atherosclerosis—namely an autoimmune component—the paradigm was revolutionized yet again [1]. It is now seen that some immune responses protect against atherosclerosis whereas others promote it.

Classically, the adaptive immune response follows from a scenario in which a pathogen escapes elimination by the innate immune system. The lymphocyte effectors of the adaptive immune system are activated by interactions between MHC/antigen complexes, T-/B-cell antigen receptors, and costimulatory molecules on the surface of innate immune system cells [46]. Professional APCs—dendritic cells, B-cells, and macrophages—take up, process, and present antigen epitopes to T-cells, thereby activating them (Figure 1). B-cell activation requires additional involvement

of $CD4^+$ T-helper cells (Figure 1). Further, the cytokines produced during the innate immune response determine the nature of the adaptive immune response—whether cellular (T-cell mediated) or humoral (B-cell mediated) [46]. Most $CD4^+$ T-cells are T-helper (Th) cells, which may be further classified on the basis of their induction and the cytokines that they secrete—an indication of the Th cells' roles and effects. For example, interferon-γ (IFN-γ) and IL-12 will induce activated $CD4^+$ T-cells to become type 1 Th (Th1) cells that secrete IFN-γ and IL-2 in promotion of cell-mediated immunity. Similarly, IL-4 will induce activated $CD4^+$ T-cells to become type 2 Th (Th2) cells that secrete IL-4, IL-5, IL-10, and IL-13 and promote humoral immunity via B-cell activation [1, 47]. These two Th subsets further cross-regulate each other. Subsequent characterization of T-cell differentiation has additionally identified IL-9-producing and IL-17-producing Th9 cells and Th17 cells subsets, respectively [48, 49]. Th9 have been observed to enhance T-cell proliferation while Th17 mediate defenses against bacteria as well as various autoimmune responses [48, 49]. Most recently, T follicular helper cells (Tfh) have been identified as having the unique ability to home to B-cell follicles where they can induce B-cell antibody production [50]. In contradistinction to the effectors of the innate immune response, the effectors of the adaptive immune response operate with a great deal of specificity to eliminate foreign pathogens and generate immunological memory. T-/B-cell receptors recognize epitopes with great specificity and affinity. It has been estimated that there are as many as 10^{18} B-cell and 10^{14} T-cell receptors [47]. While the purpose of the adaptive immune response is to eliminate agents containing antigens perceived as nonself, the scenario in which an adaptive immune response is mounted against self antigens does sometimes occur. It is now believed that atherosclerosis proceeds from such an autoimmune mechanism [12].

Atherosclerotic plaques are seen to contain elements of both innate and adaptive immunity—mostly macrophages and T-cells with few B-cells [2]. To confirm a role for adaptive immunity in atherogenesis, atherosclerotic plaque formation was observed in T- and B-cell deficient hypercholesterolemic mice and seen to be attenuated [51–54]. Further, it was observed that $CD4^+$ T-cell transfer to these deficient mice restored lesion formation to that of the control [55]. It was also noted that atherosclerotic plaques contain numerous cells producing IFN-γ, IL-12, IL-15, IL-18, and tumor necrosis factor (TNF) with few in contrast producing IL-4 [56–58]. This suggests a role for Th1 cells and not Th2 cells in promoting atherosclerosis. Further support is provided by attenuation of atherosclerotic plaque formation in animal studies targeting mediators of T-cell differentiation into Th1 cells—for example, IFN-γ, IL-12, IL-18, and TNF [10, 59–63]. Administration of IFN-γ in order to induce Th1 differentiation was seen to increase atherosclerosis in mice whereas pharmacological inhibition of Th1 cells was seen to decrease atherosclerosis in these mice [11, 64]. In contrast, Th2 cells seem to mediate anti-atherosclerotic effects in animal studies. Genetically modified mice prone to Th2 immune responses (as opposed to Th1 immune responses) developed reduced fatty streak formation, which was reversed by preventing T-cell differentiation into Th2 cells [65, 66]. Moreover, overexpression of Th2 cytokine IL-10 in hypercholesterolemic mice reduced lesion size by 50% [67]. Similarly, Th2 cytokine IL-5 deficiency has been shown to increase atherosclerosis [9]. Nevertheless, studies with Th2-inducing cytokine IL-4 overexpression show inconsistent results [45, 68]. Thus, Th1 cells have a clear proatherogenic role, whereas Th2 cells have a more complex role in atherosclerosis that is yet to be fully elucidated.

Th1 cytokines are proinflammatory and largely promote atherosclerosis by perpetuating the inflammatory mechanisms discussed earlier (Figure 1). IL-12, IL-18, IFN-γ, and TNF, found in atherosclerotic plaques, not only induce Th1 differentiation to produce still more cytokines but also activate macrophages and vascular cells to accelerate atherosclerosis. IFN-γ activates macrophages and inhibits endothelial cell and smooth muscle cell proliferation [69, 70]. This produces pro-inflammatory cytokines, prothrombotic factors, and vasoactive mediators. Meanwhile, TNF activates the NF-κB pathway in vascular cells to trigger still more inflammation and the generation of reactive oxygen species, proteolytic enzymes, and prothrombotic factors [71–73]. These pathways all implicate Th1 adaptive immune responses in the process of atherosclerosis. On the other hand, potentially pro-atherosclerotic pathways, associated with Th2, Th9, Th17, Tfh, and $CD8^+$ T-cells, are not as well defined [12, 74]. Nevertheless, these cells have obvious implications in modulating pro- and anti-atherosclerotic pathways since they differentiate from a common precursor, which means that the nuances governing their differentiation ultimately determine the course of pro- or anti-atherosclerotic responses. Tfh cells modulate B-cell response and antibody production (Figure 1); Th2 cells modulate Th1 cell differentiation and activity; Th17 cells regulate and are mutually exclusive of immunosuppressive regulatory T-cells [49, 74]. Further, IL-17 production by Th17 may also play a role in modulating autoimmune responses in atherosclerosis (Figure 1) [75–77].

$CD4^+CD25^{high}$ regulatory T-cells (Tregs) have an important role in suppressing both innate and adaptive immune responses (Figure 1) [28]. Our lab among others has shown that loss of Treg function results in autoimmune responses [78, 79]. We have characterized a Treg-specific, IL-2-dependent apoptotic pathway and demonstrated that Treg apoptotic/survival pathways are therapeutic targets for Treg-based immunotherapy. Moreover, we and others have shown that Treg inhibition accelerates vascular inflammation, with the obvious link to exacerbate atherosclerosis [80–82]. Like effector T-cells, Tregs require T-cell receptor activation and costimulation for activation. However, it remains to distinguish a pathway that exclusively regulates Tregs as opposed to effector T-cells [83]. Though it was seen that superagonistic CD28 antibody preferentially activated Tregs in animal models, phase I clinical trials showed a dangerous lack of specificity that triggered a cytokine storm [84–86]. Nevertheless, direct Treg transplant has proven more successful [87]. As atherosclerotic plaques contain a depleted number of Tregs (1%–5% of all T-cells versus normally 25% of all T-cells), this may prove to be a useful strategy for

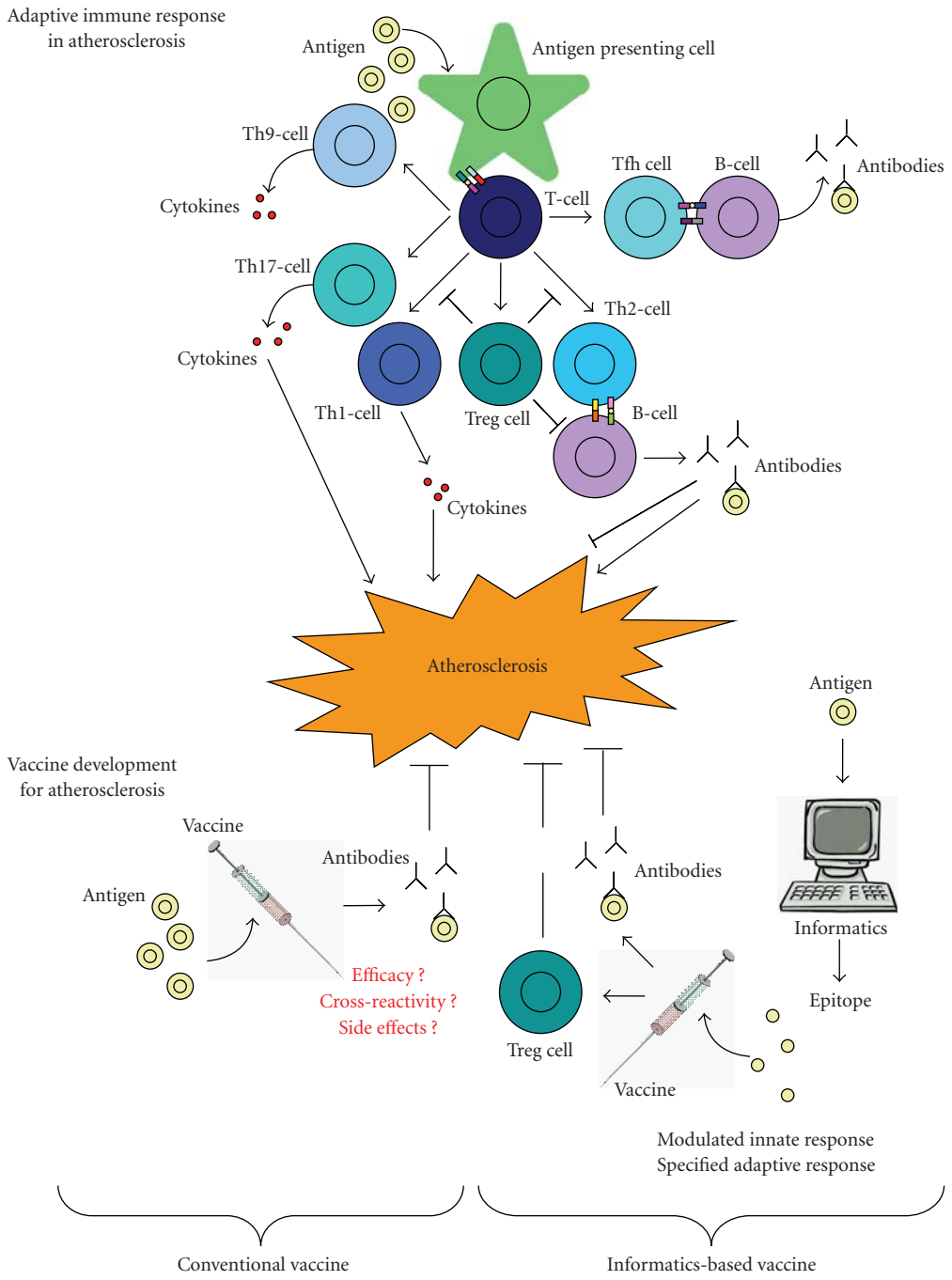


Figure 1: Approaches for vaccine development. This schematic diagram illustrates the adaptive immune response in atherosclerosis (top) as well as the conventional approach (bottom left) and our proposed bioinformatic approach (bottom right) to vaccine development for atherosclerosis.

regulation of T-cell dysfunction [88]. It should be cautioned that excessive T-cell suppression can compromise cellular defense systems. Ideally, Treg therapy should be targeted to specific autoantigens involved in atherogenesis and localized in atherosclerotic plaques.

In addition to T-cell mediated cellular responses, the adaptive immune response also includes humoral responses mediated by B-cells. Studies with T- and B-cell deficient mice confirmed a net pro-atherogenic role for these mediators of adaptive immunity [51–54]. However, other studies demonstrated that splenectomy and consequent loss of some T- and B-cells led to increased atherosclerosis that could be reversed by adoptive transfer of B-cells [89–91]. The primary role of B-cells in atherosclerosis is thought to consist of antibody production, which may be either pro- or anti-atherogenic [1, 74]. B-cells can also secrete cytokines

and serve as APCs, thereby modulating the T-cell adaptive immune response, though such a role for B-cells has not been sufficiently investigated in atherosclerosis. It has been suggested that a major limitation in appreciating the cellular responses mediated by B-cells is the failure to appreciate the existence of B-cell subsets that can mediate specific responses [92, 93]. While specific B-cell subsets have been identified in mice, it still remains to identify them in humans [93]. It is thought, however, that cells from a B-cell subset termed "B1-cells" produce "natural antibodies" of the IgM subtype in a T-cell and antigen independent manner [94]. Primarily located in the peritoneal cavity, they provide a first line of defense against foreign pathogens [92, 94]. Intriguingly, IgM natural antibodies specific for oxLDL are found in the circulation of humans without significant levels of atherosclerosis [95].

$CD4^+$ T-cells from atherosclerotic plaques have been shown to have specificity for oxLDL and HSP60 among other potential exogenous and endogenous antigens [57, 96–98]. oxLDL is one of the earliest and best characterized autoantigens involved in atherogenesis [99]. As an initiating step in atherogenesis, LDL particles that are trapped within arterial intima become oxidized and begin to accumulate as a fatty streak [100–102]. APCs internalize oxLDL via the scavenger-receptor pathway and, after proteolytic processing, bind MHC class II molecules to present peptide fragments to T-cells [1]. As T-cells do not react with native LDL components, it is thought that LDL oxidation leads to a loss of immunogenic tolerance. It has also been suggested that hypercholesterolemia impairs DC mobility with a consequent loss of immunogenic tolerance as they are not able to perform their tolerogenic clearance functions [12, 103–105]. Moreover, the existence of antibodies specific for oxLDL particles further indicates that an adaptive immune response to oxLDL occurs during atherosclerosis [97]. It has been suggested that the nature of this adaptive immune response is a combination of pathological autoimmunity and protective immunity [12].

HSPs' expressions are widespread amongst different organisms, and HSPs largely function as chaperones. They have also been implicated in autoimmune atherosclerosis [106]. Interest in HSPs as targets, developed from an observation that *Chlamydia pneumoniae* have HSPs that are found in atherosclerotic plaques [18, 107]. Epidemiologically, elevated HSPs have been found in patients with systemic hypertension, coronary artery disease, carotid atherosclerosis, myocardial infarction (MI), and ischemia [107]. Further, high levels of antibodies against mycobacterial HSP65 independently predict MI, stroke, and cardiovascular death [108]. The argument for HSPs as autoantigens relates to their cross-reactivity with other bacterial HSPs (e.g., *Mycobacteria*, *Chlamydia*) and potential recognition by antibodies against infection by those bacteria [107]. Such autoimmune recognition has been considered a consequence of a molecular mechanism termed "molecular mimicry" [109]. Other infectious agents like herpes simplex and cytomegalovirus have also been found in atherosclerotic lesions [47]. Though infection is not a necessary contributor to atherosclerosis, it may certainly play a role in activating PRRs that initiate innate immune responses leading to subsequent adaptive immune responses. For example, it has been demonstrated that TLR9 activation on DCs in atherosclerotic plaques leads to an autoimmune T-cell attack on vascular smooth muscle cells [110]. This study demonstrates the mechanism of how a pathogen can generate a pro-atherogenic autoimmune response.

5. Vaccine Development for Atherosclerosis

Historically, vaccines have been proved to be a safe and efficient tool for protection against infectious diseases [111]. Following decades of conceptual and technological breakthroughs, the concept of vaccination has been extended to a range of diseases not strictly limited to those caused by infection. Not surprisingly, vaccine design for the treatment of atherosclerosis has been an actively investigated field for a number of years. Several antigen targets have been proposed as well as several approaches to vaccine design, implementation, and efficacy. In some cases this has already led to funded clinical trials. Vaccines for atherosclerosis are somewhat different from traditional vaccines for infectious diseases. Ideal vaccines for atherosclerosis should provide recipients with (1) protective immunity against infection-derived pro-atherogenic antigens and (2) immune tolerance for autoimmunogenic self-antigens.

In the previous section, we discussed the immune responses to pro-atherogenic autoantigens. In this section, we focus on the research progress using these autoantigens as the targets of vaccine development. As LDL is a major mediator of atherosclerosis, it is not surprising that this was the earliest target for anti-atherosclerotic vaccine therapy [112–116]. These early studies provide conflicting, yet encouraging results. Subsequent studies focusing on knock-down of oxLDL showed a decrease in atherosclerotic lesion size [117, 118]. The complexity of LDL has, however, made identification of antigenic epitopes difficult. To this end, it has been necessary to develop several different experimental models. Early experiments targeted modified LDL with some success [119–121]. Still other experiments targeted components involved in LDL metabolism [122–124]. Targeting oxidized phospholipid components of oxLDL such as phosphatidyl choline (PC) generated enthusiasm in the field since both oxLDL and apoptotic cells express PC which promotes targeted removal by receptor-scavenger and IgM pathways [125]. Binder et al. were able to demonstrate that immunization with PC-containing *Streptococcus pneumoniae* vaccine generated oxLDL-specific antibodies that also correlated with reduced atherosclerosis [126]. Caligiuri et al. further showed that immunization with PC linked to a carrier protein was sufficient to induce specific antibodies and significantly reduce atherosclerosis [127]. Collectively, these studies confirm that antibodies against oxLDL epitopes have atheroprotective effects. A major consideration remains, however, as to what potential cross-reactivity will occur with non-oxLDL endogenous products containing oxidized phospholipids. For greater oxLDL specificity, more recent studies have targeted ApoB-100 peptide fragments following aldehyde modification and proteolytic degradation, which

are specific for LDL [12]. Immunization of mice with several of these ApoB-100 peptides was seen to reduce atherosclerosis [128–131].

Studies in animal models have confirmed the potential antigenicity of HSPs. Xu et al. have demonstrated that mycobacterial HSP65 elicits a pro-atherogenic immune response [132, 133]. Based on similar modes of delivery, other studies corroborate that the immune response against HSP65 is a pro-atherogenic most likely because of the innate immune response and Th1 activation [134, 135]. A different strategy using mucosal delivery in order to elicit a Th2 response showed decreased atherosclerosis [136, 137]. Immunization of patients with HSP70 also showed decreased atherosclerosis [138]. Following the idea that molecular mimicry is the mechanism responsible for human cross-reactivity with foreign HSPs, vaccination against infectious agents like influenza virus has been examined in the context of atherosclerosis. Though some results are encouraging, the benefits of the clinical trials have not yet proven convincing [107, 139]. Similarly, vaccination against hepatitis A was not seen to change atherosclerotic development in animal models [140]. Vaccines against still more risk factors for atherosclerosis have been examined like nicotine [141, 142], angiotensin I [143], ghrelin [144], and periodontitis [145] with some interesting results. While all avenues of thought have serious implications for preventing atherosclerosis, it is unrealistic to believe that a vaccine constructed from a single underlying molecular target of a clearly multifaceted disease process will effectively prevent it. In light of recent advances in knowledge and technology, it is useful to reevaluate the delivery and mode of action of vaccines in addition to simply choosing its targets.

Adjuvants are agents administered along with vaccines in order to generate an increased immune response to the vaccine. They often enhance the potency and longevity of a specific immune response without themselves inducing toxicity or initiating long-lasting immune effects [146, 147]. Aluminum salts—aluminum hydroxide, aluminum phosphate, or simply "Alum"—have been used as adjuvants since 1926 [148]. In fact, Alum is still used in a number of US FDA-approved vaccines for diphtheria-pertussis-tetanus, *Haemophilus influenzae* type b, hepatitis B, hepatitis A, inactivated poliovirus, *Streptococcus pneumonia*, and human papilloma virus [146, 149]. Despite its widespread use, Alum's mechanism for enhancing immune response has only recently begun to be understood. Numerous suggestions abound the following: (1) depot formation facilitating continuous antigen release, (2) promotion of phagocytosis by APCs, (3) induction of chemokine secretion by monocytes and macrophages, (4) inflammatory monocyte recruitment, (5) Th2 persistence and cytokine secretion to facilitate humoral response, (6) NALP3-dependent caspase-1 activation with IL-1β secretion [34, 150–154]. Each mechanism offers a conceivable route by which Alum may enhance different specific immune responses, though it is not certain under which circumstances which pathways are activated. A recent study of Alum in hypercholesterolemic mice showed that conditions of hypercholesterolemia affect Alum-induced immune responses. Wigren et al. demonstrated that Alum induced Treg expansion and inhibition of T-cell proliferation in hypercholesterolemic mice, but not in normal control littermates [155]. Aside from Treg induction being an important finding in its own right, this result points to a very important consideration in vaccine design—proper adjuvant formulation is necessary for efficacy. For more details, Reed et al. provide an excellent review of adjuvant design and development [146].

In addition to adjuvants, the route of vaccine delivery is also important to consider in order to elicit a specific immune response. Mucosal (oral or intranasal) immunization is often used to induce tolerogenic responses [156]. It is thought that this mode of delivery induces a Th2 mediated response, resulting in Th1 suppression, decreased inflammation, and reduced development of delayed-type hypersensitivity. Van Puijvelde et al. have shown that orally administered oxLDL can suppress atherosclerosis via Treg induction [157]. As an additional consideration, the time of delivery can also affect the efficacy of the vaccine. The neonatal immune system is distinctly different from the adult immune system with a Th2 bias and an abundance of Tregs [158]. Consequently, it has been shown that immunization of neonatal mice with oxLDL results in suppressed T-cell responses to oxLDL and inhibition of atherosclerosis [159].

Thus, while vaccine development for atherosclerosis is already underway, it's still important to explore a myriad of possibilities in order to anticipate future obstacles and in order to avoid a slow stepwise progression from one innovation to the next. Identification of a single antigen target is not a sufficient end. It would be better to identify a process that determines how many other targets there may be and how and when to deliver vaccines against them. Immunomodulation involves working with the entirety of the immune system—both innate and adaptive. Determination of target antigens should be made with the types of induced immune responses in mind—immunosuppressive versus immune-provocative, cellular immune responses versus humoral immune responses, and so forth.

The great availability of databases and bioinformatic tools characterizing B- and T-cell epitopes allows for evaluation of sequences among known autoantigens to identify suitable epitopes for inducing humoral and/or cellular immune responses. We previously found that the transcripts of all the autoantigens involved in various autoimmune diseases including atherosclerosis are highly modulated by alternative splicing. In contrast, the RNA transcripts of only 42% $\pm$ 5% randomly selected human genes are modulated by alternative splicing. However, housekeeping splicing of RNA transcripts could not generate autoantigen epitopes encoded by alternatively spliced exon(s). The only splicing events that respond to inflammatory/pathological stimulation include extra-exon sequences in mRNA transcripts, which then generate antigen epitopes previously untolerized in thymic T-cell development. In support of our observation via database mining, our experimental reports demonstrated that (1) the expression of an essential alternative splicing factor ASF/SF2 is modulated in the autoimmune/inflammation affected tissue in patients with autoimmune disease [160], (2) unmutated tumor antigens are actually a special group

of autoantigens that elicit antitumor immune responses in patients with tumors [161], and (3) the alternatively spliced long isoform of unmutated self-tumor antigen CML66 is immunogenic but not the alternatively spliced short form of CML66 in patients who developed anti-CML66 immune responses [162]. Based on these analyses, we proposed a new "stimulation-responsive splicing" model for generating untolerized autoantigenic epitopes (see our invited review for the details) [163]. We propose a similar analysis of known pro-atherosclerotic antigens in order to determine ideal autoantigenic epitopes rather than mere guesswork as to what moieties and modifications are potentially good targets. Moreover, by generating a systematic approach to autoantigen epitope identification, we can also lend additional support in identifying candidate targets hypothesized by present methods.

6. Bioinformatic Approach for Vaccine Development

We would like to suggest a bioinformatic approach for vaccine design that utilizes the vast array of epitope data available from prediction algorithms and databases. The idea is to use sequences of known autoantigens to characterize autoantigenic epitopes recognized by either/both B- and T-cells. As these recognized autoantigenic epitopes are generated from experimentally verified autoantigenic sequences, it follows that their homology may be characterized. We can then make statistical comparisons with candidate sequences to assess their antigenicity. This is a novel approach to vaccine design as it uses well-characterized information in order to generate a systematic approach of characterizing new antigenic candidates. Moreover, this approach has a structural basis as epitope recognition largely proceeds from sequence and structure rather than purported pathology. Presently, most available vaccines utilize weakened or inactivated forms of pathogens to either raise neutralizing antibodies or stimulate secondary immune responses specific to the antigen of interest (Figure 1) [111]. This method of vaccinology aims to induce immunity similar to that elicited by natural infections, without the pathogenic autoimmune consequences. This approach, however, has not been able to yield effective vaccines for diseases including most types of cancers, human immunodeficiency virus (HIV), tuberculosis, and atherosclerosis [166]. To improve the specificity, efficacy and safety of traditional vaccines, the concept of "epitope-based vaccines" was developed, the execution of which relies on the availability of genomic sequence databases as well as modern bioinformatic technologies (Figure 1). Since protein antigens mutate periodically (especially in viruses implicated in diseases—e.g., influenza virus, HIV), it is advantageous to target common epitope sequences that are conserved across subtypes. The identification of conserved genetic variants can be performed using public databases such as the NCBI Entrez protein database (http://www.ncbi.nlm.nih.gov/entrez) [163]. Consequently, a polypeptide that contains genetic information of multiple epitopes could be engineered to protect against a broad spectrum of microbial antigenic strains [167]. Utilizing bioinformatic approaches, combined with biological tools such as microarray, epitope-based studies of vaccines for HIV, malaria, and *B. meningococcus* have produced positive results [168–170]. Long-term studies in determining the relationship between antigen structure and antigenicity have generated numerous antigenicity prediction algorithms. Once the antigenic sequences of interest are identified, the prediction and mapping of relevant T-cell and B-cell epitopes is the next critical step in the creation of vaccines for atherosclerosis [107].

B-cells mediate a humoral immune response through the secretion of antibodies that neutralize invading microbes and pathogenic antigens. B-cells are stimulated when their antigen receptors recognize an antigenic epitope, which could be linear continuous amino acid sequences with about 15 amino acids or discontinuous amino acids connected via tertiary structures. The conformational aspect of discontinuous epitopes complicates the accurate prediction of B-cell epitopes [164, 171]. The algorithms developed for the prediction of B-cell epitopes have not been as effective or accurate [164]. B-cell epitope prediction tools fall into three categories: prediction based on primary amino acid sequences, based on conformational structure, and based on phage-display data. For continuous B-cell epitopes, prediction algorithms analyze physicochemical properties of primary sequences including amino acid hydrophilicity, flexibility, polarity, and turns. Sequence-based tools for the analysis of continuous B-cell epitopes include ABCprep [172], Bepipred, BEPITOPE [173], and IEDB B-cell epitope tools [164] (Table 1). For discontinuous B-cell epitopes, conformational structures play a more important role in epitope prediction compared to continuous epitopes, and structure-based tools have been developed such as Epitopia [174] and Ellipro [175] for the prediction of continuous epitopes (Table 1). Programs such as Epitope Mapping Tool (EMT) and EPIMAP utilize phage libraries to screen epitopes specific for known antigens. By analyzing many experimentally identified antigenic epitopes, we have generated statistical confidence intervals of antigenicity scores in order to ascertain the antigenicity of predicted B-cell epitopes [176]. This may be used to characterize potential vaccine targets with the aim of inducing a B-cell immune response in two ways. First, it can be applied to determine if a candidate antigen can be recognized by B-cells. Also, this method may be used to generate candidate epitopes within candidate antigens.

Compared to B-cell epitopes, T-cell epitopes are simple, short linear 9-15 amino acid sequences. As a result, the prediction tools for T-cell epitopes are better defined and give consistent high-quality results. T-cell epitopes are presented by MHCs for recognition by specific T-cell antigen receptors (TCRs). The interactions between epitopes and MHC as well as ones between epitope and TCR are critical in antigen recognition [177]. The binding of epitopes with MHC is a necessary step in dominant epitope generation and has been used as a predictor for epitope identification. Epitopes bind to special grooves in MHC molecules, which consist of specific sets of amino acids called anchor motifs

Table 1: B-cell epitope prediction tools. This table, adapted from Table 1 [164] summarizes available websites for B-cell epitope prediction. The prediction mechanisms employed, URLs, or authors' email addresses are included. Detailed descriptions and references are included in the paper.

Name	Description	URL
ABCpred	Based on sequence using recurrent neural network module	http://www.imtech.res.in/raghava/abcpred/
BEPITOPE	Based on sequence to predict continuous epitopes	jlpellequer@cea.fr
Bepro	Based on structure of the antigen to predict discontinuous B cell epitopes	http://pepito.proteomics.ics.uci.edu/
CEP	Based on structure for the prediction of continuous and discontinuous epitopes	http://bioinfo.ernet.in/cep.htm
COBEpro	Based on primary sequence of continuous B cell epitopes. Secondary structure and solvent accessibility information can be incorporated to improve prediction	http://scratch.proteomics.ics.uci.edu/
DiscoTope	Based on sequence and structure for the prediction of discontinuous epitopes	http://www.cbs.dtu.dk/services/DiscoTope/
Ellipro	Based on solvent accessibility and protein flexibility	http://tools.immuneepitope.org/tools/ElliPro/iedb_input
EMT	Based on Phage-display for the prediction of continuous and discontinuous epitopes	elro@novozymes.com/
EPIMAP	Based on Phage-display for the prediction of continuous and discontinuous epitopes	mumey@cs.montana.edu
Epitopia	Based on either protein 3-D structure or linear sequence using training data	http://epitopia.tau.ac.il
IEDB B-cell epitope tools	Predict continuous B cell epitope based on amino acids scales and discontinuous epitopes based on 3D structures	http://tools.immuneepitope.org/main/html/bcell_tools.html

Table 2: T-cell epitope prediction tools. This table, adapted from Table 1 [165], summarizes available websites for T-cell epitope prediction. The prediction mechanisms employed, URLs, or authors' email addresses are included. Detailed descriptions and references are included in the paper.

Name	Description	URL
BIMAS	Half time dissociation to HLA class I molecules	http://thr.cit.nih.gov/molbio/hla_bind
EpiMatrix	Binding efficiency with MHC Class I and II	http://www.epivax.com/
FRAGPREDICT	Binding score of Proteasome Cleavage Sites	http://www.mpiib-berlin.mpg.de/MAPPP/cleavage.html
Immune Epitope Database and Analysis Resource (IEDB)	Analyze proteasomal processing, TAP transport, and MHC I&II binding to produce an overall score for a peptide's potential for of being a T cell epitope	http://www.immuneepitope.org/
MHCPred	Based on MHC/peptide or TAP/peptide IC50 binding values	http://www.jenner.ac.uk/
MMBPred	Mutated high affinity MHC binding peptides	http://www.imtech.res.in/raghava/mmbpred/
NetChop	Proteasome or immunoproteasome cleavage sites	http://www.cbs.dtu.dk/services/NetChop/
NetCTL	Combined scores for MHC subtype binding, TAP transport and NetChop proteasome scores	http://www.cbs.dtu.dk/services/NetCTL/
NetMHC	MHC binding propensity of peptides	http:/www.cbs.dtu.dk/services/NetMHC
ProPred-1	Efficiency of MHC I peptide binding, optional proteasome/immunoproteasome cleavage filter	http://www.imtech.res.in/raghava/propred1
SYFPEITHI	Binding motifs to MHC Class I and II	http://www.syfpeithi.com/
TAPPred	Binding affinity of TAP proteins	http://www.imtech.res.in/raghava/tappred/
TEPITOPE	Promiscuous MHC II epitopes	http://www.vaccinome.com/

[165]. We found that in addition to the primary anchor amino acid residues E2 (the second in the epitope) and E9, some secondary residues including E3, E4, E6, E7, and E8 as well as some residues in the N-terminal and C-terminal flanking regions also contribute to binding to MHC molecule and epitope cleavage [178]. Predictions of epitopes based on analysis of anchor motif sequences are the most widely used including SYFPEITHI, BIMAS, and EpiMatrix (Table 2). The high affinity binding grooves accommodate peptides in a highly promiscuous manner such that each

MHC molecule could bind approximately between 1,000 and 10,000 peptide sequences with different affinities [179]. Moreover, MHC anchor motif sequences are prone to mutation and result in polymorphisms. To accommodate variations, programs such as TEPITOPE (Table 2) have been created. Furthermore, an assumption underlying prediction methods analyzing anchor motifs is that each amino acid residue contributes independently to the overall binding efficiency. In reality, interactions among different sites could influence the binding property of individual sites [180]. To address this discrepancy, machine learning methods such as artificial neural networks (ANNs) have been devised to address the complex relationships in the data sets [181].

Epitope-MHC binding is a prerequisite for T-cell epitope presentation. Generally, the higher the affinity binding is between the antigenic epitope and MHC molecule, the better the immunogenicity of the epitope. However, epitope-MHC binding is not the only definitive epitope indicator. Before epitope-MHC binding, the peptide needs to be processed in order to be presented by MHC molecules for T-cell recognition. This requires two steps: (1) proper protease cleavage in cytosolic ubiquitin-proteasome-peptidase systems and (2) translocation from cytosol into endoplasmic reticulum lumen by the transporter associated with antigen processing (TAP). Prediction programs such as NetCTL and IEDB (Table 2) have been designed to account for epitope proteasome cleavage and TAP binding efficiencies.

As with B-cell epitopes, we have generated statistical confidence intervals of antigenicity scores, proteasome processing scores, as well as TAP binding scores for various tumor antigen epitopes by analyzing many experimentally identified antigenic epitopes. This allows us to ascertain the antigenicity of predicted T-cell epitopes. These statistical confidence intervals suggest that over 95% of experimentally identified antigenic epitopes have the prediction scores within the intervals [163]. Similar measures could be made for the design of atherosclerosis vaccines. Proceeding from experimentally verified pro-atherosclerotic autoantigens and mediators will be possible to characterize relevant B- and T-cell epitopes. Further, candidate antigens may be analyzed using this result to determine potential antigenicity and mediation of the desired immune response. The effective utilization of B-cell and T-cell epitope prediction tools will greatly save time and effort and enhance future endeavors to design effective vaccines for atherosclerosis.

7. Discussion

In this review, we have interpolated a number of novel findings in the fields of vascular biology, immunology, and bioinformatics in order to discuss the development of a vaccine against atherosclerosis. We began with a brief discussion of traditional and novel risk factors for atherosclerosis that led to the determination of the inflammatory components of atherosclerosis. Now it is seen that the inflammatory pathways of the innate immune response set a course for an adaptive immune response—specifically an autoimmune response—that mediates the progression of atherosclerosis. Thinking of atherosclerosis in this way makes the disease a candidate for treatment by vaccine in the traditional sense of modulating immune responses. However, with recent technological breakthroughs, vaccine development affords precision in specifying the nature of the desired immune response—a useful tool when addressing a disease as complex as atherosclerosis with a manifold of inflammatory and autoimmune components. Moreover, our exploration of available bioinformatic tools for epitope-based vaccine design affords a thorough consideration of vaccine design without expenditure of excess time or resources.

Acknowledgment

M. Jan, S. Meng, and N. C. Chen made equal contributions to this work.

References

[1] G. K. Hansson and P. Libby, "The immune response in atherosclerosis: a double-edged sword," *Nature Reviews Immunology*, vol. 6, no. 7, pp. 508–519, 2006.

[2] L. Jonasson, J. Holm, O. Skalli, et al., "Regional accumulations of T cells, macrophages, and smooth muscle cells in the human atherosclerotic plaque," *Arteriosclerosis*, vol. 6, no. 2, pp. 131–138, 1986.

[3] L. Jonasson, J. Holm, O. Skalli, et al., "Expression of class II transplantation antigen on vascular smooth muscle cells in human atherosclerosis," *Journal of Clinical Investigation*, vol. 76, no. 1, pp. 125–131, 1985.

[4] P. T. Kovanen, M. Kaartinen, and T. Paavonen, "Infiltrates of activated mast cells at the site of coronary atheromatous erosion or rupture in myocardial infarction," *Circulation*, vol. 92, no. 5, pp. 1084–1088, 1995.

[5] G. Caligiuri, M. Rudling, V. Ollivier, et al., "Interleukin-10 deficiency increases atherosclerosis, thrombosis, and low-density lipoproteins in apolipoprotein E knockout mice," *Molecular Medicine*, vol. 9, no. 1-2, pp. 10–17, 2003.

[6] Z. Mallat, A. Gojova, C. Marchiol-Fournigault, et al., "Inhibition of transforming growth factor-β signaling accelerates atherosclerosis and induces an unstable plaque phenotype in mice," *Circulation Research*, vol. 89, no. 10, pp. 930–934, 2001.

[7] V. Z. Rocha and P. Libby, "Obesity, inflammation, and atherosclerosis," *Nature Reviews*, vol. 6, no. 6, pp. 399–409, 2009.

[8] A. Tedgui and Z. Mallat, "Cytokines in atherosclerosis: pathogenic and regulatory pathways," *Physiological Reviews*, vol. 86, no. 2, pp. 515–581, 2006.

[9] C. J. Binder, K. Hartvigsen, M.-K. Chang, et al., "IL-5 links adaptive and natural immunity specific for epitopes of oxidized LDL and protects from atherosclerosis," *Journal of Clinical Investigation*, vol. 114, no. 3, pp. 427–437, 2004.

[10] C. Buono, C. J. Binder, G. Stavrakis, J. L. Witztum, L. H. Glimcher, and A. H. Lichtman, "T-bet deficiency reduces atherosclerosis and alters plaque antigen-specific immune responses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 5, pp. 1596–1601, 2005.

[11] E. Laurat, B. Poirier, E. Tupin, et al., "In vivo downregulation of T helper cell 1 immune responses reduces atherogenesis in

apolipoprotein E-knockout mice," *Circulation*, vol. 104, no. 2, pp. 197–202, 2001.

[12] J. Nilsson and G. K. Hansson, "Autoimmunity in atherosclerosis: a protective response losing control?" *Journal of Internal Medicine*, vol. 263, no. 5, pp. 464–478, 2008.

[13] G. S. Berenson, S. R. Srinivasan, W. Bao, W. P. Newman III, R. E. Tracy, and W. A. Wattigney, "Association between multiple cardiovascular risk factors and atherosclerosis in children and young adults," *The New England Journal of Medicine*, vol. 338, no. 23, pp. 1650–1656, 1998.

[14] T. J. Tegos, E. Kalodiki, M. M. Sabetai, and A. N. Nicolaides, "The genesis of atherosclerosis and risk factors: a review," *Angiology*, vol. 52, no. 2, pp. 89–98, 2001.

[15] B. A. Maron and J. Loscalzo, "The treatment of hyperhomocysteinemia," *Annual Review of Medicine*, vol. 60, pp. 39–54, 2009.

[16] H. Wang, X. Jiang, F. Yang, et al., "Hyperhomocysteinemia accelerates atherosclerosis in cystathionine β-synthase and apolipoprotein E double knock-out mice with and without dietary perturbation," *Blood*, vol. 101, no. 10, pp. 3901–3907, 2003.

[17] D. Zhang, X. Jiang, P. Fang, et al., "Hyperhomocysteinemia promotes inflammatory monocyte generation and accelerates atherosclerosis in transgenic cystathionine β-synthase-deficient mice," *Circulation*, vol. 120, no. 19, pp. 1893–1902, 2009.

[18] L. A. Campbell and C.-C. Kuo, "Chlamydia pneumoniae—an infectious risk factor for atherosclerosis?" *Nature Reviews Microbiology*, vol. 2, no. 1, pp. 23–32, 2004.

[19] M. I. Cybulsky, K. Iiyama, H. Li, et al., "A major role for VCAM-1, but not ICAM-1, in early atherosclerosis," *Journal of Clinical Investigation*, vol. 107, no. 10, pp. 1255–1262, 2001.

[20] R. G. Collins, R. Velji, N. V. Guevara, M. J. Hicks, L. Chan, and A. L. Beaudet, "P-selectin or intercellular adhesion molecule (ICAM)-1 deficiency substantially protects against atherosclerosis in apolipoprotein E-deficient mice," *Journal of Experimental Medicine*, vol. 191, no. 1, pp. 189–194, 2000.

[21] Z. M. Dong, S. M. Chapman, A. A. Brown, P. S. Frenette, R. O. Hynes, and D. D. Wagner, "The combined role of P- and E-selectins in atherosclerosis," *Journal of Clinical Investigation*, vol. 102, no. 1, pp. 145–152, 1998.

[22] S. D. Cushing, J. A. Berliner, A. J. Valente, et al., "Minimally modified low density lipoprotein induces monocyte chemotactic protein 1 in human endothelial cells and smooth muscle cells," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 13, pp. 5134–5138, 1990.

[23] G. Subbanagounder, J. W. Wong, H. Lee, et al., "Epoxyisoprostane and epoxycyclopentenone phospholipids regulate monocyte chemotactic protein-1 and interleukin-8 synthesis. Formation of these oxidized phospholipids in response to interleukin-1β," *The Journal of Biological Chemistry*, vol. 277, no. 9, pp. 7271–7281, 2002.

[24] L. Boring, J. Gosling, M. Cleary, and I. F. Charo, "Decreased lesion formation in CCR2$^{-/-}$ mice reveals a role for chemokines in the initiation of atherosclerosis," *Nature*, vol. 394, no. 6696, pp. 894–897, 1998.

[25] J. Gosling, S. Slaymaker, L. Gu, et al., "MCP-1 deficiency reduces susceptibility to atherosclerosis in mice that overexpress human apolipoprotein B," *Journal of Clinical Investigation*, vol. 103, no. 6, pp. 773–778, 1999.

[26] L. Gu, Y. Okada, S. K. Clinton, et al., "Absence of monocyte chemoattractant protein-1 reduces atherosclerosis in low density lipoprotein receptor-deficient mice," *Molecular Cell*, vol. 2, no. 2, pp. 275–281, 1998.

[27] C. Combadiere, S. Potteaux, J.-L. Gao, et al., "Decreased atherosclerotic lesion formation in CX3CR1/apolipoprotein E double knockout mice," *Circulation*, vol. 107, no. 7, pp. 1009–1016, 2003.

[28] X.-F. Yang, Y. Yin, and H. Wang, "Vascular inflammation and atherogenesis are activated via receptors for PAMPs and suppressed by regulatory T cells," *Drug Discovery Today: Therapeutic Strategies*, vol. 5, no. 2, pp. 125–142, 2008.

[29] K. Edfeldt, J. Swedenborg, G. K. Hansson, and Z.-Q. Yan, "Expression of toll-like receptors in human atherosclerotic lesions: a possible pathway for plaque activation," *Circulation*, vol. 105, no. 10, pp. 1158–1161, 2002.

[30] K. S. Michelsen, M. H. Wong, P. K. Shah, et al., "Lack of toll-like receptor 4 or myeloid differentiation factor 88 reduces atherosclerosis and alters plaque phenotype in mice deficient in apolipoprotein E," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 29, pp. 10679–10684, 2004.

[31] A. Kol, A. H. Lichtman, R. W. Finberg, P. Libby, and E. A. Kurt-Jones, "Cutting edge: heat shock protein (HSP) 60 activates the innate immune response: CD14 is an essential receptor for HSP60 activation of mononuclear cells," *Journal of Immunology*, vol. 164, no. 1, pp. 13–17, 2000.

[32] X. H. Xu, P. K. Shah, E. Faure, et al., "Toll-like receptor-4 is expressed by macrophages in murine and human lipid-rich atherosclerotic plaques and upregulated by oxidized LDL," *Circulation*, vol. 104, no. 25, pp. 3103–3108, 2001.

[33] Y. Yin, Y. Yan, X. Jiang, et al., "Inflammasomes are differentially expressed in cardiovascular and other tissues," *International Journal of Immunopathology and Pharmacology*, vol. 22, no. 2, pp. 311–322, 2009.

[34] S. C. Eisenbarth, O. R. Colegio, W. O'Connor Jr., F. S. Sutterwala, and R. A. Flavell, "Crucial role for the Nalp3 inflammasome in the immunostimulatory properties of aluminium adjuvants," *Nature*, vol. 453, no. 7198, pp. 1122–1126, 2008.

[35] H. Li, S. B. Willingham, J. P.-Y. Ting, and F. Re, "Cutting edge: inflammasome activation by alum and alum's adjuvant effect are mediated by NLRP3," *Journal of Immunology*, vol. 181, no. 1, pp. 17–21, 2008.

[36] S. L. Demento, S. C. Eisenbarth, H. G. Foellmer, et al., "Inflammasome-activating nanoparticles as modular systems for optimizing vaccine efficacy," *Vaccine*, vol. 27, no. 23, pp. 3013–3021, 2009.

[37] L. Peiser, S. Mukhopadhyay, and S. Gordon, "Scavenger receptors in innate immunity," *Current Opinion in Immunology*, vol. 14, no. 1, pp. 123–128, 2002.

[38] A. Nicoletti, G. Caligiuri, I. Tornberg, T. Kodama, S. Stemme, and G. K. Hansson, "The macrophage scavenger receptor type A directs modified proteins to antigen presentation," *European Journal of Immunology*, vol. 29, no. 2, pp. 512–521, 1999.

[39] C. K. Glass and J. L. Witztum, "Atherosclerosis: the road ahead," *Cell*, vol. 104, no. 4, pp. 503–516, 2001.

[40] M. Krieger, "Scavenger receptor class B type I is a multiligand HDL receptor that influences diverse physiologic systems," *Journal of Clinical Investigation*, vol. 108, no. 6, pp. 793–797, 2001.

[41] M. Febbraio, D. P. Hajjar, and R. L. Silverstein, "CD36: a class B scavenger receptor involved in angiogenesis, atherosclerosis, inflammation, and lipid metabolism," *Journal of Clinical Investigation*, vol. 108, no. 6, pp. 785–791, 2001.

[42] J.-H. Choi, Y. Do, C. Cheong, et al., "Identification of antigen-presenting dendritic cells in mouse aorta and cardiac valves," *Journal of Experimental Medicine*, vol. 206, no. 3, pp. 497–505, 2009.

[43] F. Mach, U. Schonbeck, G. K. Sukhova, et al., "Functional CD40 ligand is expressed on human vascular endothelial cells, smooth muscle cells, and macrophages: implications for CD40-CD40 ligand signaling in atherosclerosis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 5, pp. 1931–1936, 1997.

[44] F. Mach, U. Schonbeck, G. K. Sukhova, E. Atkinson, and P. Libby, "Reduction of atherosclerosis in mice by inhibition of CD40 signalling," *Nature*, vol. 394, no. 6689, pp. 200–203, 1998.

[45] E. Lutgens, K. B. J. M. Cleutjens, S. Heeneman, V. E. Koteliansky, L. C. Burkly, and M. J. A. P. Daemen, "Both early and delayed anti-CD40L antibody treatment induces a stable plaque phenotype," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 13, pp. 7464–7469, 2000.

[46] C. A. T. Janeway, M. Walport, and M. J. Shlomchik, *Immunobiology*, Garland Science Publishing, New York, NY, USA, 6th edition, 2005.

[47] C. J. Binder, M.-K. Chang, P. X. Shaw, et al., "Innate and acquired immunity in atherogenesis," *Nature Medicine*, vol. 8, no. 11, pp. 1218–1226, 2002.

[48] M. Veldhoen, C. Uyttenhove, J. van Snick, et al., "Transforming growth factor-β 'reprograms' the differentiation of T helper 2 cells and promotes an interleukin 9-producing subset," *Nature Immunology*, vol. 9, no. 12, pp. 1341–1346, 2008.

[49] C. T. Weaver and R. D. Hatton, "Interplay between the TH 17 and TReg cell lineages: a (co-)evolutionary perspective," *Nature Reviews Immunology*, vol. 9, no. 12, pp. 883–889, 2009.

[50] C. King, "New insights into the differentiation and function of T follicular helper cells," *Nature Reviews Immunology*, vol. 9, no. 11, pp. 757–766, 2009.

[51] H. M. Dansky, S. A. Charlton, M. M. Harper, and J. D. Smith, "T and B lymphocytes play a minor role in atherosclerotic plaque formation in the apolipoprotein E-deficient mouse," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 9, pp. 4642–4646, 1997.

[52] A. Daugherty, E. Pure, D. Delfel-Butteiger, et al., "The effects of total lymphocyte deficiency on the extent of atherosclerosis in apolipoprotein $E^{-/-}$ mice," *Journal of Clinical Investigation*, vol. 100, no. 6, pp. 1575–1580, 1997.

[53] C. A. Reardon, L. Blachowicz, T. White, et al., "Effect of immune deficiency on lipoproteins and atherosclerosis in male apolipoprotein E-deficient mice," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 21, no. 6, pp. 1011–1016, 2001.

[54] L. Song, C. Leung, and C. Schindler, "Lymphocytes are important in early atherosclerosis," *Journal of Clinical Investigation*, vol. 108, no. 2, pp. 251–259, 2001.

[55] X. Zhou, A. Nicoletti, R. Elhage, and G. K. Hansson, "Transfer of $CD4^+$ T cells aggravates atherosclerosis in immunodeficient apolipoprotein E knockout mice," *Circulation*, vol. 102, no. 24, pp. 2919–2922, 2000.

[56] J. Frostegard, A.-K. Ulfgren, P. Nyberg, et al., "Cytokine expression in advanced human atherosclerotic plaques: dominance of pro-inflammatory (Th1) and macrophage-stimulating cytokines," *Atherosclerosis*, vol. 145, no. 1, pp. 33–43, 1999.

[57] S. Stemme, B. Faber, J. Holm, O. Wiklund, J. L. Witztum, and G. K. Hansson, "T lymphocytes from human atherosclerotic plaques recognize oxidized low density lipoprotein," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 9, pp. 3893–3897, 1995.

[58] K. Uyemura, L. L. Demer, S. C. Castle, et al., "Cross-regulatory roles of interleukin (IL)-12 and IL-10 in atherosclerosis," *Journal of Clinical Investigation*, vol. 97, no. 9, pp. 2130–2138, 1996.

[59] L. Branen, L. Hovgaard, M. Nitulescu, E. Bengtsson, J. Nilsson, and S. Jovinge, "Inhibition of tumor necrosis factor-α reduces atherosclerosis in apolipoprotein E knockout mice," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 24, no. 11, pp. 2137–2142, 2004.

[60] C. Buono, C. E. Come, G. Stavrakis, G. F. Maguire, P. W. Connelly, and A. H. Lichtman, "Influence of interferon-γ on the extent and phenotype of diet-induced atherosclerosis in the LDLR-deficient mouse," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 23, no. 3, pp. 454–460, 2003.

[61] P. Davenport and P. G. Tipping, "The role of interleukin-4 and interleukin-12 in the progression of atherosclerosis in apolipoprotein E-deficient mice," *American Journal of Pathology*, vol. 163, no. 3, pp. 1117–1125, 2003.

[62] R. Elhage, J. Jawien, M. Rudling, et al., "Reduced atherosclerosis in interleukin-18 deficient apolipoprotein E-knockout mice," *Cardiovascular Research*, vol. 59, no. 1, pp. 234–240, 2003.

[63] S. Gupta, A. M. Pablo, X.-C. Jiang, N. Wang, A. R. Tall, and C. Schindler, "IFN-γ, potentiates atherosclerosis in ApoE knock-out mice," *Journal of Clinical Investigation*, vol. 99, no. 11, pp. 2752–2761, 1997.

[64] S. C. Whitman, P. Ravisankar, H. Elam, and A. Daugherty, "Exogenous interferon-γ enhances atherosclerosis in apolipoprotein $E^{-/-}$ mice," *American Journal of Pathology*, vol. 157, no. 6, pp. 1819–1824, 2000.

[65] S. A. Huber, P. Sakkinen, C. David, M. K. Newell, and R. P. Tracy, "T helper-cell phenotype regulates atherosclerosis in mice under conditions of mild hypercholesterolemia," *Circulation*, vol. 103, no. 21, pp. 2610–2616, 2001.

[66] B. Paigen, A. Morrow, C. Brandon, et al., "Variation in susceptibility to atherosclerosis among inbred strains of mice," *Atherosclerosis*, vol. 57, no. 1, pp. 65–73, 1985.

[67] L. J. Pinderski, M. P. Fischbein, G. Subbanagounder, et al., "Overexpression of interleukin-10 by activated T lymphocytes inhibits atherosclerosis in LDL receptor-deficient mice by altering lymphocyte and macrophage phenotypes," *Circulation Research*, vol. 90, no. 10, pp. 1064–1071, 2002.

[68] V. L. King, S. J. Szilvassy, and A. Daugherty, "Interleukin-4 deficiency decreases atherosclerotic lesion formation in a site-specific manner in female LDL $receptor^{-/-}$ mice," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 22, no. 3, pp. 456–461, 2002.

[69] R. Friesel, A. Komoriya, and T. Maciag, "Inhibition of endothelial cell proliferation by γ-interferon," *Journal of Cell Biology*, vol. 104, no. 3, pp. 689–696, 1987.

[70] G. K. Hansson, M. Hellstrand, L. Rymo, L. Rubbia, and G. Gabbiani, "Interferon γ inhibits both proliferation and expression of differentiation-specific α-smooth muscle actin in arterial smooth muscle cells," *Journal of Experimental Medicine*, vol. 170, no. 5, pp. 1595–1608, 1989.

[71] E. Lee, D. E. Vaughan, S. H. Parikh, et al., "Regulation of matrix metalloproteinases and plasminogen activator inhibitor-1 synthesis by plasminogen in cultured human

vascular smooth muscle cells," *Circulation Research*, vol. 78, no. 1, pp. 44–49, 1996.

[72] P. Sarén, H. G. Welgus, and P. T. Kovanen, "TNF-α and IL-1β selectively induce expression of 92-kDa gelatinase by human macrophages," *Journal of Immunology*, vol. 157, no. 9, pp. 4159–4165, 1996.

[73] V. W. M. van Hinsbergh, E. A. van den Berg, W. Fiers, and G. Dooijewaard, "Tumor necrosis factor induces the production of urokinase-type plasminogen activator by human endothelial cells," *Blood*, vol. 75, no. 10, pp. 1991–1998, 1990.

[74] J. Andersson, P. Libby, and G. K. Hansson, "Adaptive immunity and atherosclerosis," *Clinical Immunology*, vol. 134, no. 1, pp. 33–46, 2010.

[75] R. E. Eid, D. A. Rao, J. Zhou, et al., "Interleukin-17 and interferon-γ are produced concomitantly by human coronary artery-infiltrating T cells and act synergistically on vascular smooth muscle cells," *Circulation*, vol. 119, no. 10, pp. 1424–1432, 2009.

[76] S. Taleb, M. Romain, B. Ramkhelawon, et al., "Loss of SOCS3 expression in T cells reveals a regulatory role for interleukin-17 in atherosclerosis," *Journal of Experimental Medicine*, vol. 206, no. 10, pp. 2067–2077, 2009.

[77] J.-J. Xie, J. Wang, T.-T. Tang, et al., "The Th17/Treg functional imbalance during atherogenesis in ApoE$^{-/-}$ mice," *Cytokine*, vol. 49, no. 2, pp. 185–193, 2010.

[78] M. G. von Herrath and L. C. Harrison, "Antigen-induced regulatory T cells in autoimmunity," *Nature Reviews Immunology*, vol. 3, no. 3, pp. 223–232, 2003.

[79] Y. Yan, Y. Chen, F. Yang, et al., "HLA-A2.1-restricted T cells react to SEREX-defined tumor antigen CML66L and are suppressed by CD4+CD25+ regulatory T cells," *International Journal of Immunopathology and Pharmacology*, vol. 20, no. 1, pp. 75–89, 2007.

[80] Z. Mallat, H. Ait-Oufella, and A. Tedgui, "Regulatory T-cell immunity in atherosclerosis," *Trends in Cardiovascular Medicine*, vol. 17, no. 4, pp. 113–118, 2007.

[81] Z. Xiong, J. Song, Y. Yan, et al., "Higher expression of Bax in regulatory T cells increases vascular inflammation," *Frontiers in Bioscience*, vol. 13, pp. 7143–7155, 2008.

[82] Z. Xiong, Y. Yan, J. Song, et al., "Expression of TCTP antisense in CD25 high regulatory T cells aggravates cuff-injured vascular inflammation," *Atherosclerosis*, vol. 203, no. 2, pp. 401–408, 2009.

[83] J. L. Riley and C. H. June, "The CD28 family: a T-cell rheostat for therapeutic control of T-cell activation," *Blood*, vol. 105, no. 1, pp. 13–21, 2005.

[84] N. Beyersdorf, S. Gaupp, K. Balbach, et al., "Selective targeting of regulatory T cells with CD28 superagonists allows effective therapy of experimental autoimmune encephalomyelitis," *Journal of Experimental Medicine*, vol. 202, no. 3, pp. 445–455, 2005.

[85] T. Lin, J. Hu, D. Wang, and D. M. Stocco, "Interferon-γ inhibits the steroidogenic acute regulatory protein messenger ribonucleic acid expression and protein levels in primary cultures of rat Leydig cells," *Endocrinology*, vol. 139, no. 5, pp. 2217–2222, 1998.

[86] G. Suntharalingam, M. R. Perry, S. Ward, et al., "Cytokine storm in a phase 1 trial of the anti-CD28 monoclonal antibody TGN1412," *The New England Journal of Medicine*, vol. 355, no. 10, pp. 1018–1028, 2006.

[87] J. L. Riley, C. H. June, and B. R. Blazar, "Human T regulatory cell therapy: take a billion or so and call me in the morning," *Immunity*, vol. 30, no. 5, pp. 656–665, 2009.

[88] O. J. de Boer, J. J. van der Meer, P. Teeling, C. M. van der Loos, and A. C. van der Wal, "Low numbers of FOXP3 positive regulatory T cells are present in all developmental stages of human atherosclerotic lesions," *PLoS ONE*, vol. 2, no. 8, article e779, 2007.

[89] G. Caligiuri, A. Nicoletti, B. Poirierand, and G. K. Hansson, "Protective immunity against atherosclerosis carried by B cells of hypercholesterolemic mice," *Journal of Clinical Investigation*, vol. 109, no. 6, pp. 745–753, 2002.

[90] A. S. Major, S. Fazio, and M. F. Linton, "B-lymphocyte deficiency increases atherosclerosis in LDL receptor-null mice," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 22, no. 11, pp. 1892–1898, 2002.

[91] J. L. Witztum, "Splenic immunity and atherosclerosis: a glimpse into a novel paradigm?" *Journal of Clinical Investigation*, vol. 109, no. 6, pp. 721–724, 2002.

[92] M. van Leeuwen, J. Damoiseaux, A. Duijvestijn, and J. W. C. Tervaert, "The therapeutic potential of targeting B cells and anti-oxLDL antibodies in atherosclerosis," *Autoimmunity Reviews*, vol. 9, no. 1, pp. 53–57, 2009.

[93] K. Yanaba, J.-D. Bouaziz, T. Matsushita, C. M. Magro, E. W. St.Clair, and T. F. Tedder, "B-lymphocyte contributions to human autoimmune disease," *Immunological Reviews*, vol. 223, no. 1, pp. 284–299, 2008.

[94] C. J. Binder and G. J. Silverman, "Natural antibodies and the autoimmunity of atherosclerosis," *Springer Seminars in Immunopathology*, vol. 26, no. 4, pp. 385–404, 2005.

[95] P. X. Shaw, S. Horkko, S. Tsimikas, et al., "Human-derived anti-oxidized LDL autoantibody blocks uptake of oxidized LDL by macrophages and localizes to atherosclerotic lesions in vivo," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 21, no. 8, pp. 1333–1339, 2001.

[96] O. J. de Boer, A. C. van der Wal, M. A. Houtkamp, J. M. Ossewaarde, P. Teeling, and A. E. Becker, "Unstable atherosclerotic plaques contain T-cells that respond to Chlamydia pneumoniae," *Cardiovascular Research*, vol. 48, no. 3, pp. 402–408, 2000.

[97] W. Palinski, M. E. Rosenfeld, S. Yla-Herttuala, et al., "Low density lipoprotein undergoes oxidative modification in vivo," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 4, pp. 1372–1376, 1989.

[98] G. Paulsson, X. Zhou, E. Tornquist, and G. K. Hansson, "Oligoclonal T cell expansions in atherosclerotic lesions of apolipoprotein E-deficient mice," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 20, no. 1, pp. 10–17, 2000.

[99] J. L. Witztum and D. Steinberg, "Role of oxidized low density lipoprotein in atherogenesis," *Journal of Clinical Investigation*, vol. 88, no. 6, pp. 1785–1792, 1991.

[100] D. J. Rader and A. Daugherty, "Translating molecular discoveries into new therapies for atherosclerosis," *Nature*, vol. 451, no. 7181, pp. 904–913, 2008.

[101] K. Skalen, M. Gustafsson, E. Knutsen Rydberg, et al., "Subendothelial retention of atherogenic lipoproteins in early atherosclerosis," *Nature*, vol. 417, no. 6890, pp. 750–754, 2002.

[102] I. Tabas, K. J. Williams, and J. Boren, "Subendothelial lipoprotein retention as the initiating process in atherosclerosis: update and therapeutic implications," *Circulation*, vol. 116, no. 16, pp. 1832–1844, 2007.

[103] H. Ait-Oufella, K. Kinugawa, J. Zoll, et al., "Lactadherin deficiency leads to apoptotic cell accumulation and accelerated atherosclerosis in mice," *Circulation*, vol. 115, no. 16, pp. 2168–2177, 2007.

[104] V. Angeli, J. Llodra, J. X. Rong, et al., "Dyslipidemia associated with atherosclerotic disease systemically alters dendritic cell mobilization," *Immunity*, vol. 21, no. 4, pp. 561–574, 2004.

[105] C. M. Weyand and J. J. Goronzy, "Medium- and large-vessel vasculitis," *The New England Journal of Medicine*, vol. 349, no. 2, pp. 160–169, 2003.

[106] M. Gupta, C. Vavasis, and W. H. Frishman, "Heat shock proteins in cardiovascular disease: a new therapeutic target," *Cardiology in Review*, vol. 12, no. 1, pp. 26–30, 2004.

[107] E. Riley, V. Dasari, W. H. Frishman, and K. Sperber, "Vaccines in development to prevent and treat atherosclerotic disease," *Cardiology in Review*, vol. 16, no. 6, pp. 288–300, 2008.

[108] A. Veres, G. Fust, M. Smieja, et al., "Relationship of anti-60 kDa heat shock protein and anti-cholesterol antibodies to cardiovascular events," *Circulation*, vol. 106, no. 22, pp. 2775–2780, 2002.

[109] K. Bachmaier and J. M. Penninger, "Chlamydia and antigenic mimicry," *Current Topics in Microbiology and Immunology*, vol. 296, pp. 153–163, 2005.

[110] A. Niessner, K. Sato, E. L. Chaikof, I. Colmegna, J. J. Goronzy, and C. M. Weyand, "Pathogen-sensing plasmacytoid dendritic cells stimulate cytotoxic T-cell function in the atherosclerotic plaque through interferon-α," *Circulation*, vol. 114, no. 23, pp. 2482–2489, 2006.

[111] R. Rappuoli, H. I. Miller, and S. Falkow, "Medicine: the intangible value of vaccination," *Science*, vol. 297, no. 5583, pp. 937–939, 2002.

[112] J. M. Bailey and R. Tomar, "lipoprotein immunization and induced atherosclerosis in rabbits. 1. Cockerel β-lipoproteins as antigens," *The Lancet*, vol. 5, no. 2, pp. 203–214, 1965.

[113] J. F. de Carvalho, R. M. Pereira, and Y. Shoenfeld, "Vaccination for atherosclerosis," *Clinical Reviews in Allergy and Immunology*, vol. 38, no. 2-3, pp. 135–140, 2010.

[114] S. Gerö, "Some data on the influence of cholesterol atherosclerosis by immunological means," *Revue de l'Atherosclerose et des Arteriopathies Peripheriques*, vol. 9, no. 1, supplement 1, pp. 194–198, 1967.

[115] S. Gerö, J. Gergely, L. Jakab, et al., "Inhibition of cholesterol atherosclerosis by immunisation with β-lipoprotein," *The Lancet*, vol. 274, no. 7088, pp. 6–7, 1959.

[116] P. E. Siegler, "Some immunological aspects of atherosclerosis," *Revue de l'Atherosclerose et des Arteriopathies Peripheriques*, vol. 9, no. 1, supplement 1, p. 199, 1967.

[117] S. Ameli, A. Hultgardh-Nilsson, J. Regnström, et al., "Effect of immunization with homologous LDL and oxidized LDL on early atherosclerosis in hypercholesterolemic rabbits," *Arterioscleros, Thrombosis, and Vascular Biology*, vol. 16, no. 8, pp. 1074–1079, 1996.

[118] W. Palinski, E. Miller, and J. L. Witztum, "Immunization of low density lipoprotein (LDL) receptor-deficient rabbits with homologous malondialdehyde-modified LDL reduces atherogenesis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 3, pp. 821–825, 1995.

[119] S. Freigang, S. Horkko, E. Miller, J. L. Witztum, and W. Palinski, "Immunization of LDL receptor-deficient mice with homologous malondialdehyde-modified and native LDL reduces progression of atherosclerosis by mechanisms other than induction of high titers of antibodies to oxidative neoepitopes," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 18, no. 12, pp. 1972–1982, 1998.

[120] J. George, A. Afek, B. Gilburd, et al., "Hyperimmunization of apo-E-deficient mice with homologous malondialdehyde low-density lipoprotein suppresses early atherogenesis," *Atherosclerosis*, vol. 138, no. 1, pp. 147–152, 1998.

[121] X. Zhou, G. Caligiuri, A. Hamsten, A. K. Lefvert, and G. K. Hansson, "LDL immunization induces T-cell-dependent antibody formation and protection against atherosclerosis," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 21, no. 1, pp. 108–114, 2001.

[122] R. S. Barkowski and W. H. Frishman, "HDL metabolism and CETP inhibition," *Cardiology in Review*, vol. 16, no. 3, pp. 154–162, 2008.

[123] M. H. Davidson, K. Maki, D. Umporowicz, A. Wheeler, C. Rittershaus, and U. Ryan, "The safety and immunogenicity of a CETP vaccine in healthy adults," *Atherosclerosis*, vol. 169, no. 1, pp. 113–120, 2003.

[124] W. H. Frishman and R. S. Barkowski, "Cholesteryl ester transfer protein inhibition for coronary heart disease prevention: real hope or despair?" *American Journal of Medicine*, vol. 121, no. 8, pp. 644–646, 2008.

[125] M.-Y. Chou, K. Hartvigsen, L. F. Hansen, et al., "Oxidation-specific epitopes are important targets of innate immunity," *Journal of Internal Medicine*, vol. 263, no. 5, pp. 479–488, 2008.

[126] C. J. Binder, S. Horkko, A. Dewan, et al., "Pneumococcal vaccination decreases atherosclerotic lesion formation: molecular mimicry between *Streptococcus pneumoniae* and oxidized LDL," *Nature Medicine*, vol. 9, no. 6, pp. 736–743, 2003.

[127] G. Caligiuri, J. Khallou-Laschet, M. Vandaele, et al., "Phosphorylcholine-targeting immunization reduces atherosclerosis," *Journal of the American College of Cardiology*, vol. 50, no. 6, pp. 540–546, 2007.

[128] K.-Y. Chyu, X. Zhao, O. S. Reyes, et al., "Immunization using an Apo B-100 related epitope reduces atherosclerosis and plaque inflammation in hypercholesterolemic apo $E^{-/-}$ mice," *Biochemical and Biophysical Research Communications*, vol. 338, no. 4, pp. 1982–1989, 2005.

[129] G. N. Frederikson, L. Andersson, I. Soderberg, et al., "Atheroprotective immunization with MDA-modified apo B-100 peptide sequences is associated with activation of Th2 specific antibody expression," *Autoimmunity*, vol. 38, no. 2, pp. 171–179, 2005.

[130] G. N. Fredrikson, I. Soderberg, M. Lindholm, et al., "Inhibition of atherosclerosis in apoE-null mice by immunization with apoB-100 peptide sequences," *Arteriosclerosis, Thrombosis, and Vascular Biology*, vol. 23, no. 5, pp. 879–884, 2003.

[131] A. Schiopu, J. Bengtsson, I. Soderberg, et al., "Recombinant human antibodies against aldehyde-modified apolipoprotein B-100 peptide sequences inhibit atherosclerosis," *Circulation*, vol. 110, no. 14, pp. 2047–2052, 2004.

[132] G. Wick, G. Schett, A. Amberger, R. Kleindienst, and Q. Xu, "Is atherosclerosis an immunologically mediated disease?" *Immunology Today*, vol. 16, no. 1, pp. 27–33, 1995.

[133] Q. Xu, H. Dietrich, H. J. Steiner, et al., "Induction of arteriosclerosis in normocholesterolemic rabbits by immunization with heat shock protein 65," *Arteriosclerosis and Thrombosis*, vol. 12, no. 7, pp. 789–799, 1992.

[134] G. K. Hansson and J. Holm, "Interferon-γ inhibits arterial stenosis after injury," *Circulation*, vol. 84, no. 3, pp. 1266–1272, 1991.

[135] D. J. Lamb and G. A. A. Ferns, "The magnitude of the immune response to heat shock protein-65 following BCG immunisation is associated with the extent of experimental

atherosclerosis," *Atherosclerosis*, vol. 165, no. 2, pp. 231–240, 2002.

[136] D. Harats, N. Yacov, B. Gilburd, Y. Shoenfeld, and J. George, "Oral tolerance with heat shock protein 65 attenuates Mycobacterium tuberculosis-induced and high-fat-diet-driven atherosclerotic lesions," *Journal of the American College of Cardiology*, vol. 40, no. 7, pp. 1333–1338, 2002.

[137] R. Maron, G. Sukhova, A.-M. Faria, et al., "Mucosal administration of heat shock protein-65 decreases atherosclerosis and inflammation in aortic arch of low-density lipoprotein receptor-deficient mice," *Circulation*, vol. 106, no. 13, pp. 1708–1715, 2002.

[138] A. G. Pockley, A. Georgiades, T. Thulin, U. De Faire, and J. Frostegard, "Serum heat shock protein 70 levels predict the development of atherosclerosis in subjects with established hypertension," *Hypertension*, vol. 42, no. 3, pp. 235–238, 2003.

[139] J. Danesh, "Antibiotics in the prevention of heart attacks," *The Lancet*, vol. 365, no. 9457, pp. 365–367, 2005.

[140] M. S. Burnett, J. Zhu, J. M. Miller, and S. E. Epstein, "Effects of hepatitis A vaccination on atherogenesis in a murine model," *Journal of Viral Hepatitis*, vol. 10, no. 6, pp. 433–436, 2003.

[141] J. Foulds, M. Burke, M. Steinberg, J. M. Williams, and D. M. Ziedonis, "Advances in pharmacotherapy for tobacco dependence," *Expert Opinion on Emerging Drugs*, vol. 9, no. 1, pp. 39–53, 2004.

[142] M. Nides, "Update on pharmacologic options for smoking cessation treatment," *American Journal of Medicine*, vol. 121, no. 4, supplement, pp. S20–S31, 2008.

[143] R. Reade, J. B. Michel, C. Carelli, H. Huang, T. Baussant, and P. Corvol, "Immunization against angiotensin I in the spontaneously hypertensive rat," *Archives des Maladies du Coeur et des Vaisseaux*, vol. 82, no. 7, pp. 1323–1328, 1989.

[144] E. P. Zorrilla, S. Iwasaki, J. A. Moss, et al., "Vaccination against weight gain," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 35, pp. 13226–13231, 2006.

[145] Y. Koizumi, T. Kurita-Ochiai, S. Oguchi, and M. Yamamoto, "Nasal immunization with Porphyromonas gingivalis outer membrane protein decreases P. gingivalis-induced atherosclerosis and inflammation in spontaneously hyperlipidemic mice," *Infection and Immunity*, vol. 76, no. 7, pp. 2958–2965, 2008.

[146] S. G. Reed, S. Bertholet, R. N. Coler, and M. Friede, "New horizons in adjuvants for vaccine development," *Trends in Immunology*, vol. 30, no. 1, pp. 23–32, 2009.

[147] A. Wack and R. Rappuoli, "Vaccinology at the beginning of the 21st century," *Current Opinion in Immunology*, vol. 17, no. 4, pp. 411–418, 2005.

[148] N. Petrovsky, "Novel human polysaccharide adjuvants with dual Th1 and Th2 potentiating activity," *Vaccine*, vol. 24, supplement 2, pp. S2/26–S2/29, 2006.

[149] C. J. Clements and E. Griffiths, "The global impact of vaccines containing aluminium adjuvants," *Vaccine*, vol. 20, supplement 3, pp. S24–S33, 2002.

[150] J. M. Brewer, M. Conacher, C. A. Hunter, M. Mohrs, F. Brombacher, and J. Alexander, "Aluminium hydroxide adjuvant initiates strong antigen-specific Th2 responses in the absence of IL-4- or IL-13-mediated signaling," *Journal of Immunology*, vol. 163, no. 12, pp. 6448–6454, 1999.

[151] M. Kool, V. Pétrilli, T. de Smedt, et al., "Cutting edge: alum adjuvant stimulates inflammatory dendritic cells through activation of the NALP3 inflammasome," *Journal of Immunology*, vol. 181, no. 6, pp. 3755–3759, 2008.

[152] M. Kool, T. Soullié, M. van Nimwegen, et al., "Alum adjuvant boosts adaptive immunity by inducing uric acid and activating inflammatory dendritic cells," *Journal of Experimental Medicine*, vol. 205, no. 4, pp. 869–882, 2008.

[153] A. Seubert, E. Monaci, M. Pizza, D. T. O'hagan, and A. Wack, "The adjuvants aluminum hydroxide and MF59 induce monocyte and granulocyte chemoattractants and enhance monocyte differentiation toward dendritic cells," *Journal of Immunology*, vol. 180, no. 8, pp. 5402–5412, 2008.

[154] M. Ulanova, A. Tarkowski, M. Hahn-Zoric, and L. A. Hanson, "The common vaccine adjuvant aluminum hydroxide up-regulates accessory properties of human monocytes via an interleukin-4-dependent mechanism," *Infection and Immunity*, vol. 69, no. 2, pp. 1151–1159, 2001.

[155] M. Wigren, D. Bengtsson, P. Duner, et al., "Atheroprotective effects of alum are associated with capture of oxidized LDL antigens and activation of regulatory T cells," *Circulation Research*, vol. 104, no. 12, pp. e62–e70, 2009.

[156] J. Holmgren, C. Czerkinsky, K. Eriksson, and A. Mharandi, "Mucosal immunisation and adjuvants: a brief overview of recent advances and challenges," *Vaccine*, vol. 21, supplement 2, pp. S89–S95, 2003.

[157] G. H. M. van Puijvelde, A. D. Hauer, P. De Vos, et al., "Induction of oral tolerance to oxidized low-density lipoprotein ameliorates atherosclerosis," *Circulation*, vol. 114, no. 18, pp. 1968–1976, 2006.

[158] V. J. Philbin and O. Levy, "Developmental biology of the innate immune response: implications for neonatal and infant vaccine development," *Pediatric Research*, vol. 65, supplement 5, pp. 98R–105R, 2009.

[159] A. Nicoletti, G. Paulsson, G. Caligiuri, X. Zhou, and G. K. Hansson, "Induction of neonatal tolerance to oxidized lipoprotein reduces atherosclerosis in ApoE knockout mice," *Molecular Medicine*, vol. 6, no. 4, pp. 283–290, 2000.

[160] Z. Xiong, A. Shaibani, Y.-P. Li, et al., "Alternative splicing factor ASF/SF2 is down regulated in inflamed muscle," *Journal of Clinical Pathology*, vol. 59, no. 8, pp. 855–861, 2006.

[161] F. Yang and X. F. Yang, "New concepts in tumor antigens: their significance in future immunotherapies for tumors," *Cellular & Molecular Immunology*, vol. 2, no. 5, pp. 331–341, 2005.

[162] Y. Yan, L. Phan, F. Yang, et al., "A novel mechanism of alternative promoter and splicing regulates the epitope generation of tumor antigen CML66-L1," *Journal of Immunology*, vol. 172, no. 1, pp. 651–660, 2004.

[163] F. Yang, I. H. Chen, Z. Xiong, Y. Yan, H. Wang, and X.-F. Yang, "Model of stimulation-responsive splicing and strategies in identification of immunogenic isoforms of tumor antigens and autoantigens," *Clinical Immunology*, vol. 121, no. 2, pp. 121–133, 2006.

[164] J. A. Greenbaum, P. H. Andersen, M. Blythe, et al., "Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools," *Journal of Molecular Recognition*, vol. 20, no. 2, pp. 75–82, 2007.

[165] B. Korber, M. LaBute, and K. Yusim, "Immunoinformatics comes of age," *PLoS Computational Biology*, vol. 2, no. 6, article e71, 2006.

[166] A. Sette and J. Fikes, "Epitope-based vaccines: an update on epitope identification, vaccine design and delivery," *Current Opinion in Immunology*, vol. 15, no. 4, pp. 461–470, 2003.

[167] A. M. Khan, O. Miotto, A. T. Heiny, et al., "A systematic bioinformatics approach for selection of epitope-based vaccine targets," *Cellular Immunology*, vol. 244, no. 2, pp. 141–147, 2006.

[168] M. M. Giuliani, J. Adu-Bobie, M. Comanducci, et al., "A universal vaccine for serogroup B meningococcus," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 29, pp. 10834–10839, 2006.

[169] H. L. Robinson and R. R. Amara, "T cell vaccines for microbial infections," *Nature Medicine*, vol. 11, no. 4, supplement, pp. S25–S32, 2005.

[170] C. C. Wilson, D. McKinney, M. Anders, et al., "Development of a DNA vaccine designed to induce cytotoxic T lymphocyte responses to multiple conserved epitopes in HIV-1," *Journal of Immunology*, vol. 171, no. 10, pp. 5611–5623, 2003.

[171] B. A. Jameson and H. Wolf, "The antigenic index: a novel algorithm for predicting antigenic determinants," *Computer Applications in the Biosciences*, vol. 4, no. 1, pp. 181–186, 1988.

[172] S. Saha and G. P. S. Raghava, "Prediction of continuous B-cell epitopes in an antigen using recurrent neural network," *Proteins: Structure, Function and Genetics*, vol. 65, no. 1, pp. 40–48, 2006.

[173] M. Odorico and J.-L. Pellequer, "BEPITOPE: predicting the location of continuous epitopes and patterns in proteins," *Journal of Molecular Recognition*, vol. 16, no. 1, pp. 20–22, 2003.

[174] N. D. Rubinstein, I. Mayrose, E. Martz, and T. Pupko, "Epitopia: a web-server for predicting B-cell epitopes," *BMC Bioinformatics*, vol. 10, p. 287, 2009.

[175] J. Ponomarenko, H.-H. Bui, W. Li, et al., "ElliPro: a new structure-based tool for the prediction of antibody epitopes," *BMC Bioinformatics*, vol. 9, p. 514, 2008.

[176] B. Ng, F. Yang, D. P. Huston, et al., "Increased noncanonical splicing of autoantigen transcripts provides the structural basis for expression of untolerized epitopes," *Journal of Allergy and Clinical Immunology*, vol. 114, no. 6, pp. 1463–1470, 2004.

[177] V. Brusic, V. B. Bajic, and N. Petrovsky, "Computational methods for prediction of T-cell epitopes—a framework for modelling, testing, and applications," *Methods*, vol. 34, no. 4, pp. 436–443, 2004.

[178] X.-F. Yang, D. Mirkovic, S. Zhang, et al., "Processing sites are different in the generation of HLA-A2.1-restricted, T cell reactive tumor antigen epitopes and viral epitopes," *International Journal of Immunopathology and Pharmacology*, vol. 19, no. 4, pp. 853–870, 2006.

[179] V. Brusic and J. Zeleznikow, "Computational binding assays of antigenic peptides," *Letters in Peptide Science*, vol. 6, no. 5-6, pp. 313–324, 1999.

[180] S. Buus, "Description and prediction of peptide-MHC binding: the 'human MHC project'," *Current Opinion in Immunology*, vol. 11, no. 2, pp. 209–213, 1999.

[181] M. Bhasin and G. P. S. Raghava, "Prediction of CTL epitopes using QM, SVM and ANN techniques," *Vaccine*, vol. 22, no. 23-24, pp. 3195–3204, 2004.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 325720, 12 pages
doi:10.1155/2010/325720

Review Article

High Throughput T Epitope Mapping and Vaccine Development

Giuseppina Li Pira,[1] Federico Ivaldi,[2] Paolo Moretti,[2] and Fabrizio Manca[2]

[1] *Laboratory of Cellular Immunology, Advanced Biotechnology Center, Largo Benzi 10, 16132 Genoa, Italy*
[2] *Laboratory of Clinical and Experimental Immunology, G. Gaslini Institute, Largo G. Gaslini 5, 16148 Genoa, Italy*

Correspondence should be addressed to Giuseppina Li Pira, lipira@email.it

Received 12 October 2009; Revised 18 February 2010; Accepted 20 April 2010

Academic Editor: Yongqun Oliver He

Mapping of antigenic peptide sequences from proteins of relevant pathogens recognized by T helper (Th) and by cytolytic T lymphocytes (CTL) is crucial for vaccine development. In fact, mapping of T-cell epitopes provides useful information for the design of peptide-based vaccines and of peptide libraries to monitor specific cellular immunity in protected individuals, patients and vaccinees. Nevertheless, epitope mapping is a challenging task. In fact, large panels of overlapping peptides need to be tested with lymphocytes to identify the sequences that induce a T-cell response. Since numerous peptide panels from antigenic proteins are to be screened, lymphocytes available from human subjects are a limiting factor. To overcome this limitation, high throughput (HTP) approaches based on miniaturization and automation of T-cell assays are needed. Here we consider the most recent applications of the HTP approach to T epitope mapping. The alternative or complementary use of in silico prediction and experimental epitope definition is discussed in the context of the recent literature. The currently used methods are described with special reference to the possibility of applying the HTP concept to make epitope mapping an easier procedure in terms of time, workload, reagents, cells and overall cost.

1. Introduction

Mapping of T epitopes on protein antigens derived from relevant pathogens implies the identification of amino acid sequences that are recognized by CD4 or CD8 T cells. The term "epitope" is frequently used interchangeably for "immunodominant peptide". For the sake of precision, though, epitope should refer to the shortest sequence that maintains stimulatory capacity for T cells, while the immunodominant peptide can be of any length and its deletions allow identification of the corresponding epitope.

Epitope mapping is important in vaccine development [1, 2] for several reasons: (1) epitopes are presented in the context of one or few MHC alleles [3–7] and in some instances they are promiscuous (i.e. presented by more alleles) [8–11] or universal (i.e. presented by most of the alleles) [12–16]; (2) epitopes represent the antigenic portion of a protein and thus the nonantigenic regions can be deleted; (3) epitopes restricted to different alleles can be collected to obtain selected peptide libraries that are recognized by the majority of the immune population [17, 18], thereby providing a valuable diagnostic tool [19–27]; (4) epitopes can be selected to construct peptide-based vaccines.

Peptides derived from HIV and restricted to human DR alleles were used as pools to prime mice to induce vigorous CD4 responses [28]. Peptides as strings of beads were produced as a recombinant protein to express and immunize against P. falciparum epitopes [29]. As additional examples, peptides as strings of beads have also been used as immunogens specific for HIV [30–32], for mycobacyteria [33], for tumor-associated viral antigens [34], for LCMV [35] and for immunodominant epitopes of five different viruses [36]. In addition, selected peptide fragments can be assembled as mosaic proteins to produce polyvalent vaccines for coverage of potential T-cell epitopes in HIV variants. [37].

In fact, epitope variability in pathogens like HIV [38] is instrumental for development of escape mutants [39] and represents the greatest challenge for the immune system that must deal with these mutations.

The converse situation, defined as deimmunization, permits selective removal by protein re-engineering of immunogenic epitopes identified on proteins to be used as therapeutic agents. In this case, in fact, potential antigenicity must be avoided to improve clinical effectiveness. This innovative possibility has been described in detail [40, 41].

In the present review we mention the predictive and experimental systems that can be used for T-cell epitope mapping. Since the experimental work requires functional T lymphocytes, the number of available cells is a limiting factor with current methods [42]. Furthermore, the time to perform the assays, the cost of reagents and the workload are additional limiting factors in epitope mapping. Peptide synthesis in particular has a remarkable impact when epitope mapping is budgeted. Therefore, in addition to in silico epitope prediction, as discussed later, any miniaturization that reduces the amount of required peptides is helpful. The possibility of using a high throughput (HTP) approach based on assay miniaturization and automation to test extended peptide panels will be dealt with in more detail. The HTP approach, in fact, permits reduction of the number of cells to be tested, of the amount of tested peptides, of the cost of reagents and of the workload referred to the information that can be obtained. The HTP concept can be applied to different techniques that can be automated and miniaturized, as described later.

2. Predictive Models of T-Cell Responses

Algorithms have been developed to test the potential capacity of a given peptide to bind a predetermined MHC allele. From this information it can be assumed that a T-cell with such a specificity can be present in the T-cell repertoire. Obviously, specific T-cell are expected to be at very low frequency in the specific naïve repertoire, while their frequency should be much higher in a primed repertoire. As pointed out later, this is probably the major difference between in silico and in vitro epitope mapping: in the former case an assumption is made while in the latter case an experimental approach is taken that must take into account the naïve or memory state of the specific repertoire that is being tested. Another important point is that in silico screening is faster and cheaper and certainly permits deletion of certain peptides that have no MHC binding potential and thus can be excluded from the list of potential candidates. Therefore, a preliminary prediction may result in remarkable reduction in the number of peptides to be synthesized and screened against responding lymphocytes.

Predictive methods have been developed for human and murine MHC alleles, including class I and class II molecules [43–50]. A large scale validation study for CTL epitope prediction has been recently published by comparing five different web-based tools [51]. Prediction of binding affinity, in addition to specificity, has been recently discussed in [43]. An updated description of predictive methods can be found in the Los Alamos Database website [38]. This site has links to tools for T epitope prediction based on MHC binding and also on cleavage sites to predict antigenic peptide fragments produced upon processing by the human proteasome [52, 53]. In fact, while MHC-peptide binding predictions consider virtual peptides derived from arbitrary cleavage of the protein sequence, anticipation of peptide fragments, that can be actually produced by antigen processing and presenting cells, can be of remarkable relevance also to reduce the number of peptides to be examined as possible candidates for immunogenicity. The combined predictions for TAP binding peptides [54–56] and protein cleavage have been modeled and considered for their potential applications, in combination with MHC binding predictions. [57, 58]. The data generated by using prediction tools to define MHC or TAP binding peptides have been collected in a dedicated database [59]. This comprehensive database includes a large collection of peptide sequences whose binding affinities for MHC or TAP molecules have been assayed experimentally.

The in silico methods based on MHC-peptide binding prediction carry the obvious advantage that T epitopes can be defined independently of the primed or naïve status of a subject. Several recent reviews have reported on this subject and its integration with in vitro data [45, 60, 61]. Peptide binding to MHC class I for CD8 T-cell recognition is certainly predictable with high confidence, while MHC class II binding algorithms have been recently criticized with respect to reliability [62].

It should be stressed here that a higher integration between in silico prediction and in vitro experiments [60] should be considered, each one of the two approaches with its pros and cons.

These aspects are considered in more detail in the different chapters.

MHC binding is a prerequisite for epitope presentation to specific T cells and thus prediction of MHC binding screens for sequences that are potentially antigenic. Nevertheless, MHC binding, even if experimentally determined to appreciate actual peptide binding to solid phase MHC [63], does not grant in itself that a T-cell with such a specificity is present indeed in the T-cell repertoire. Therefore, experimental data generated by in vitro work in which T cells are challenged with peptides, are necessary to confirm the prediction.

3. Experimental Approaches for T-Cell Epitope Mapping and Data Collection

Several assays measure various features of T-cell activation that follows antigen recognition, as described below in more detail. Using different methods, immunodominant peptides have been identified on proteins of various pathogens. The best example of an available database for experimentally identified epitopes recognized by CD4 and CD8 T-cell s is represented by the Los Alamos National Laboratory database [38] that provides updated information on HIV T epitopes. Another exhaustive database is represented by the Immune Epitope Data Base (IEDB, http://tools.immuneepitope.org) [64–66], in which an accurate description of the individual contributions is given. This database is fed with experimental data obtained by literature searches or by voluntary submissions, with an accurate description of methods used for epitope identification.

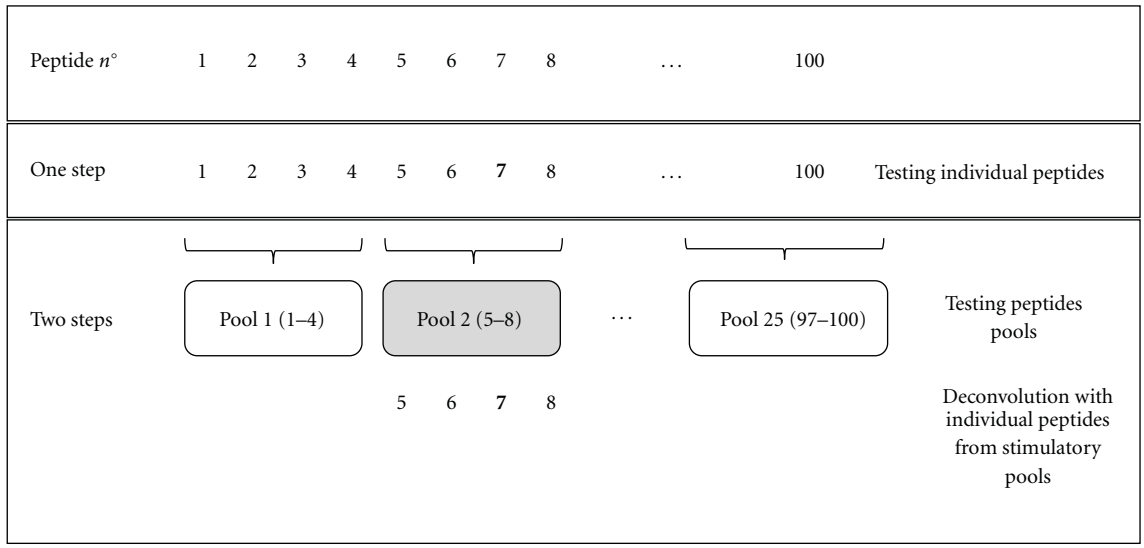


Figure 1: Conventional layout of peptide panels for T epitope mapping. In this example of a panel of 100 peptides, peptides can be tested individually in one step or as pools in two steps. Peptide 7 in bold represents the immunodominant peptide. Response to a positive pool can be further deconvoluted by testing the individual peptides encompassed by the stimulatory pool. The one step approach is simpler, but it requires more cells. The two-step approach requires fewer cells, but cells are needed at two different times. Thus, the cases in which a small fraction of pools are positive can take advantage of this approach, while in the worst case, in which the majority of the pools are positive and need to be individually deconvoluted, the two-step approach provides no significant advantage.

4. Peptide Layout for High Throughput T Epitope Mapping

One aspect of the T epitope mapping methodologies is represented by the use of synthetic peptides. The straightforward approach is to produce a complete panel of peptides, with a given length and a predetermined overlapping sequence, that encompass the candidate protein. This comprehensive procedure, though, is limited by the cost of peptide synthesis and thus peptide panels can be restricted by deleting those peptides that have been predicted as non binders using in silico screening. A detailed discussion on synthetic peptides to be used for epitope mapping has been published [67].

An epitope with a sequence spanning two sequential peptides would be missed if it is not encompassed by the overlapping sequence. The ideal peptide scan should be with one amino acid progression (sequential peptides). Apart from increased cost, this would be possible when mapping epitopes recognized by antibodies, that are available, in practice, at unlimited amount from diluted serum and that can be tested on microarray formats. For cellular assays that detect T-cell immunity, available PBMC are an incumbent limitation. In order to induce CTL and Th responses, 15–18 mer peptides overlapping by 11 amino acid residues are commonly used. This peptide size is selected so to have potential binding capacity to MHC class I and class II. In order to avoid the need for a CD4 peptide panel and a CD8 peptide panel, a compromise can be reached as described in an extensive study [68] using peptides derived from the immunodominant pp65 protein of cytomegalovirus. Even if it was confirmed that 9 mers are optimal stimulators for CD8 responses, it was clearly shown that 15 mers can also be used to detect CTL activation. This paper also presents the conundrum of which should be the optimal overlap between contiguous peptides in order not to miss a stimulatory epitope. It is concluded that in the case of 15 mers overlapping by 11 residues, each possible 9 mer is present in two 15 mers, except for the one with a central location, that is, present only once in the middle of the 15 mer. To obviate this limit and to have all 9 mers, including the central ones, represented in at least two 15 mers, peptides should overlap by 3 residues. This implies a 25% increase in the number of peptides. Therefore, also in this case a compromise must be accepted between number/cost of synthesized peptides and confidence that no epitopes are present in suboptimal configurations (i.e. at the extreme N- or C-terminus of longer peptides). Another point to be considered is the calculated risk of missing some epitopes when overlapping instead of sequential peptides are used. It is reasonable to say that it is acceptable to miss some epitopes when dealing with large, highly antigenic proteins on which numerous epitopes are expected, as in the case of proteins derived from pathogens. On the contrary, in the case of small proteins or antigens on which a few antigenic epitopes are predicted, as in the case of tumor antigens, the search should be more thorough.

If sequential peptides from panels are individually tested (Figure 1), this results in the use of large numbers of lymphocytes for screening, independently of the method in use (see following paragraphs). Alternatively, discrete pools of peptides can be used for preliminary screening, followed by deconvolution based on testing individual peptides present in the stimulatory pools (Figure 1).

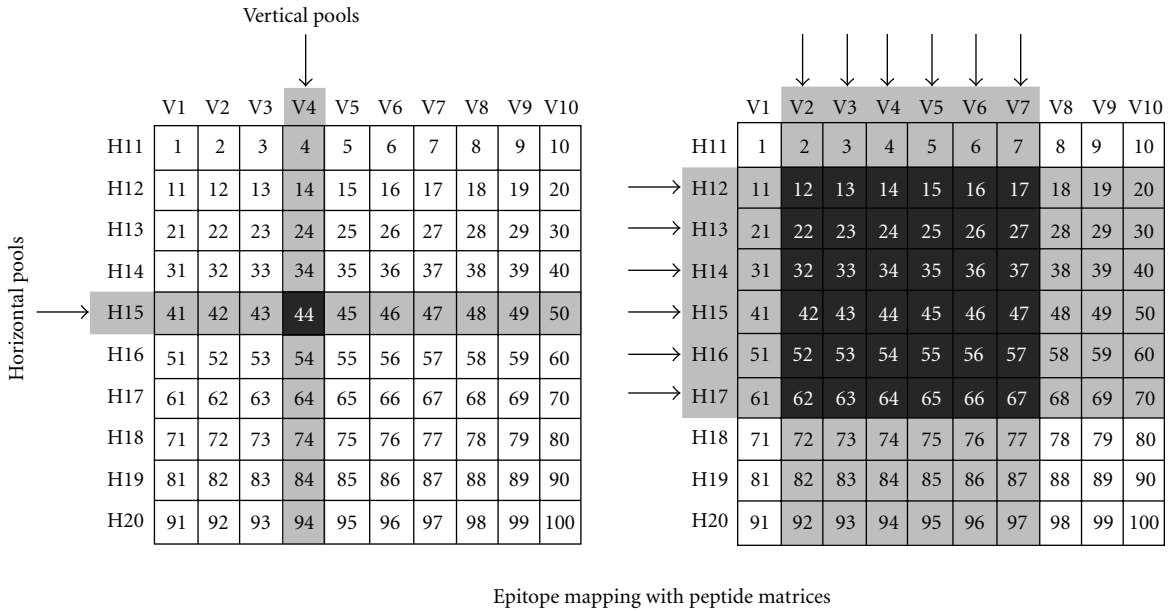


Figure 2: Epitope mapping with peptide matrices. As an example, a panel of 100 peptides is considered. Instead of being tested individually, peptides can be arranged in a matrix of 10 vertical pools (V1-V10) and 10 horizontal pools (H11-H20). The left panel depicts an unambiguous identification. In fact, if peptide 44 only is antigenic, pools V4 and H15 are stimulatory (arrows), thereby pointing at the relevant peptide at the intersection. The right panel depicts an ambiguous identification. In fact, if V2-V7 and H12-H7 pools are stimulatory (arrows), all peptides in the shaded area can be antigenic, but the same result would be obtained if only peptides 12,23,34,45,56,67 were antigenic. Therefore, the intersections of the stimulatory pools could indicate from a minimum of 6 to a maximum of 36 antigenic peptides. This ambiguity would require an additional deconvolution step in which individual peptides identified by the intersection area are individually tested.

In order to save on cells, peptide matrices have been introduced [69, 70]. In this case, horizontal and vertical peptide pools are assembled as shown in Figure 2. In this example, 20 pools are screened instead of testing 100 individual peptides. The intersection resulting from response to one vertical and one horizontal pool reveal the antigenic peptide. This is a convenient approach in principle. In fact, unambiguous definition of the antigenic peptide results at the intersection (left panel). Nevertheless, in case more peptides are antigenic, the intersections can be ambiguous (right panel) and a second deconvolution step with individual candidate peptides present in the stimulatory pools is needed as detailed in the figure legend. In the layout of this type of experiments, two types of peptide containing plates should be designed: the first one containing the peptides as pools or in a matrix format, the second one containing the full peptide panel. The second deconvolution step can be run by seeding lymphocytes only in the wells that contain peptides of stimulatory pools, rather than preparing a new plate each time with the ambiguous peptides contained in the stimulatory pools. This can reduce errors and labour. Handling of large peptide panels can benefit from automated dispensers that use individual 96 or 384 channels for distribution from masterplates into secondary replica plates.

Conventional peptide synthesis on mg scale provides more material than actually needed. In fact, peptides are currently used at 0.1–10 ug/mL in culture. Miniaturized peptide synthesis on pins [71–73] or on paper spots [74, 75] permits a less expensive peptide synthesis. In addition, microsynthesis in 96 well plate formats permits an easier transfer of peptides to 96 well or to 384 well plates, thereby abating the risk of positional errors or cross-contaminations.

Overlapping peptides have also been applied on microarray slides for antibody epitope mapping [76], but this approach is not applicable at the present time to T epitope mapping with functional assays. In fact, there is no evidence that solid phase peptides can be captured by membrane MHC and pulse APC for presentation to specific T cells. Furthermore, at least one specific T-cell should be present in the microarea of one peptide spot and this depends on the frequency of specific T cells, as discussed in a later paragraph. Biotechnological developments in this field are desirable, even though it should be noted that the microarray slide approach has already been used with MHC-peptide complexes to retain and stimulate specific T cells, as described later.

When the immunodominant protein of a given pathogen is not known, conventional overlapping peptides panels cannot be designed. Therefore, combinatorial peptide libraries consisting of 10-mer mixtures containing every possible combination of residues with only one fixed position could reveal the unknown epitopes [77–79].

5. Current Methods for T Epitope Mapping

Once peptide panels in different formats have been produced, T cells can be tested with different methods to determine the stimulatory immunodominant peptides. Methods to detect T-cell responses can be defined as dynamic or static [42], depending on whether a T-cell function is detected or not. Most of the assays reveal a functional T-cell response as upregulation of activation markers, cytokine synthesis, proliferation, cytolytic, and helper function. Only MHC-peptide multimers (MM) can be used independently of T-cell functions. These fluorescent reagents, in fact, bind and label the specific T-cell receptors by simulating an antigen presenting cell that displays an MHC-peptide complex.

All functional assays require also an intact APC function. Therefore, optimized culture conditions are needed, in particular to provide all specific T cells a chance to interact with an APC. This means that cell density (cells versus surface) is more important than cell concentration (cells versus volume). This is crucial in order not to underestimate responses, particularly if the method reads the actual frequency of specific T cells (e.g. ELIspot, ICS).

5.1. MHC-Peptide Multimers. Multimers constructed with specific MHC class I alleles and carrying a well-defined MHC binding peptide and a fluorescent tag [80] are simple to use, require no culture steps and are revealed by flow cytometric analysis. Additional phenotypic and functional markers can be simultaneous identified on the multimer (MM) binding cells [81]. Cells are not sacrificed and thus can be further cultured or even used for in vivo reinfusion for adoptive immunoreconstitution [82]. Besides these advantages, major limitations are represented by the fact that purified MHC class I molecules are available for a limited number of alleles and by the fact that the antigenic peptide(s) need to be known in advance, as a result of preliminary epitope mapping. In order to apply MM to epitope mapping, a novel approach has been proposed [83] based on reversible binding of MHC and peptide. Here, the first peptide contains a UV cleavable site that results in its dissociation from MHC once exposed to UV light. If a second peptide with the same MHC binding capacity is present, it can replace the cleavable peptide and generate a new MM with a different specificity (i.e. recognized by another T-cell receptor differing for fine specificity but not for MHC restriction). If this process is applied in microplates in which different second peptides are distributed in wells, a whole panel of MM can be constructed for T-cell screening. This approach is appealing, but it is still limited to a few MHC class I alleles and has not been proven to function with MHC class II molecules.

Multimers with MHC class II alleles [84] are not as broadly available as MHC class I multimers and technological improvements are still under development [85].

5.2. Solid Phase MHC-Peptide Complexes. Spotting MHC class I molecules on microarray slides, followed by binding of immunodominant peptides, has been used to create solid phase MHC-peptide complexes that can be recognized by specific T cells. Lymphocytes bind and can be selectively retained [86]. In addition, if the slide is also coated with an anticytokine antibody, the cytokine secreted by specifically activated T cells is captured as a spot and can be revealed by immunofluorescence with an appropriate scanner [87]. This method could be applied for epitope mapping by deposition of different peptides on the MHC spots, but the complexity of the manipulations and of the instrumentation required make this approach not broadly applicable at the present time.

5.3. Lymphoproliferation. The 3H-thymidine uptake assay is sensitive and it has been extensively used in the past. Nevertheless, it requires a radioactive reagent, no data on cell phenotype, frequency or secreted cytokines are provided and 3–5 culture days are needed. Finally, cells are sacrificed and cannot be used any further. The classical format is the 96 well plate, with no possible extension to the 384 well format because of the lack of appropriate harvesters on the market. For these reasons, lymphoproliferation is loosing followers, even though most of the initial epitope mapping data were obtain with this technique. Lymphoproliferation resurged when the CFSE dye dilution method was established. A major advantage in this case is that the labeling dye is diluted between daughter cells at each cell division and thus fluorescence quantitation can estimate the number of duplications the cells went through. An accurate report with methodological details has been recently published [88]. A comparison of proliferation versus ICS for T epitope mapping has also been reported [89].

5.4. Activation Markers Induced on Specific T Cells. Several activation markers have been used for identification, enumeration, and selection of antigen specific CD4 and CD8 T cells. These markers for antigen activated T cells include CD137 for detection of CD8 cells [90] and CD154 for detection of CD4 cells [91]. Coexpression of CD25 and CD134 also identifies antigen specific CD4 T cells [92]. Therefore, the same markers can also be applied for epitope mapping. The use of these markers is advantageous on one side, as it combines functional activity with phenotypic parameters, but it may be cumberson for HTP epitope mapping.

5.5. ELIspot. T cells that secrete a given cytokine (IFNg most of the times) as a consequence of antigen recognition on APC can be detected by ELIspot. In this technique, originally developed to detect antibody secreting B cells [93], lymphocyte cultures are performed in 96 well plates, whose bottom is a nitrocellulose or PVDF membrane coated with a primary anticytokine antibody. The secreted cytokine is captured by the solid phase antibody in the proximity of the producing T-cell. Thus, antigen specific T cells can be revealed as a spot by a second biotinylated antibody, followed by a streptavidin-enzyme conjugate and the appropriate substrate that is enzymatically cleaved so to generate an insoluble product [94]. Spots that form are eventually enumerated by visual inspection under low magnification.

This assay requires simple instrumentation and thus it is broadly used. The development of dedicated scanners for spot identification and enumeration of positive cells made it applicable to large-scale screening. A further improvement is the simultaneous analysis of two cytokines produced by a single cell using two detection reagents labeled with different fluorochromes [95]. This enables enumeration of single or double producing T cells. Two limitations of ELIspot are represented by the fact that T-cell phenotype cannot be defined, unless separated subsets are used, and that the actual number of secreting cells can be underestimated. In fact, if more than one specific T-cell interacts with the same APC, one spot may contain more than one specific cell. The 384 well format can be proposed for ELIspot thanks to the availability of appropriate scanners.

5.6. Intracytoplasmic Cytokine Staining (ICS). In this technique T cells are stimulated with antigen and APC. Responding specific T cells synthesize cytokines that are retained in the cytoplasm because of the addition of drugs that inhibit secretion. Intracytoplasmic cytokines are stained with fluorescent antibodies after cell permeabilization with a detergent. This technique, developed in the late '90 s [96] proved very reliable and informative [21, 97]. The main advantage of ICS is that numerous phenotypic markers can be tested on specific T cells and different cytokines can be detected in a single cell using monoclonals labeled with different colors, the only limit being on the flow cytometer used for the analysis (and on the proficiency of the flow cytometrist). Thus, multiparametric and polychromatic flow cytometry identifies antigen specific T cells and simultaneously evaluates different effector functions [98–100]. The assay can be run in round bottom 96 well plates, in combination with automatic samplers that feed the flow cytometer. In the perspective of further extending the assay to the 384 well plate format, HTP-ICS analysis may be foreseen in the near future. Limitations of ICS are represented by the cost of reagents and instrumentation, by the fact that cells are cultured under non-physiological conditions with an inhibitor of protein secretion and by the fact that synthesized cytokines cannot be quantitated. The analysis of mean fluorescence intensity for the stained cytokines, though, may provide some indications for a relative semiquantitative evaluation.

5.7. Cytokine Secretion and Cell Surface Capture (CSC). In this assay, T cells are activated using culture conditions in larger volumes (tubes, 24 well plates, flasks). After 18–36 hours stimulation, cytokine secreting T cells are identified with an anti-CD45 antibody conjugated with an anticytokine antibody. During a short additional incubation step, secreted cytokines are captured by the bispecific antibody that decorates all PBMC. The cytokine captured on the membrane of the producing T-cell is revealed by a second antibody specific for the same cytokine, that is, directly or indirectly fluorescent for flow cytometric analysis or sorting. Alternatively, the second antibody can be recognized by magnetic beads for preparative separation of specific T cells on magnetic columns [101]. CSC is convenient for ex vivo selection of specific T lymphocytes and it has been used for selection of virus specific T cells [102, 103] that can be subsequently reinfused for adoptive immunoreconstitution [104, 105]. Cytokine positive cells can also be stained with additional markers for extensive phenotyping, but since numerous staining steps are needed, this system is not advisable for screening of peptide panels.

5.8. Cytokine Secretion and Well Surface Capture (Cell-ELISA). has been reported in 96 well plates [106] and recently proposed, as a variation on the theme of ELIspot and ELISA, in 384 or 1536 well plates [107, 108]. Secreted cytokines, as readout of T-cell activation, can be tested in culture supernatants by conventional ELISA. This procedure can be simplified and miniaturized using cell-ELISA, an assay in which wells are coated with the first anticytokine antibody. Medium and antigens are dispensed automatically and the plates are frozen for later use. When needed, plates are thawed and seeded with PBMC. During the 18–36 hours incubation, specific T cells responding to antigens or peptides secrete the tested cytokine that is captured by the solid phase antibody. Plates are eventually frozen before final processing. For development, plates are processed following a conventional miniaturized ELISA protocol for a quantitative evaluation of produced cytokines using standard curves. Before development, culture supernatants can be saved for cytokine profiling using different HTP approaches, like multiparametric flow cytometric assays or microarrays. Major advantages of this method are that an HTP approach can be taken for handling the multiwell plates.

In addition, secreted cytokines are quantitated and the 1536 well plate format permits the use of as few as 10^4 PBMC per tested antigen. This implies some considerations regarding the frequency of antigen specific T cells. In fact, as shown in Figure 3, a reasonable frequency of memory T cells specific for a recall antigen can range between 0.1% and 1%. Lower frequencies should be expected for T cells specific for a single epitope. Therefore, each culture well should contain a sufficient number of specific cells that results in a positive signal. Sensitivity of cell-ELISA has been estimated around 0.01% specific T cells diluted into autologous non specific T cells by using established lines as standards or PBMC with known frequencies of specific cells [107]. This sensitivity range is similar, if not higher, than other conventional methods including MM, ICS, ELIspot, lymphoproliferation. Therefore, in order to be confident that a detectable number of antigen specific T cells is seeded in each well, we estimate that 10^4 PBMC per well is the lowest threshold. This consideration applies to scaling down of all types of lymphocyte based assays for evaluation of memory responses. An intermediate format that proved convenient was the 384 small well plate, in which each well has a culture volume of 20 uL instead of 40 uL. Therefore, 2×10^4 PBMC can be seeded instead of 4×10^4, leaving the same cell density at the bottom of the culture well.

This obvious limitation due to reduction of the number of tested T cells applies to all assays, even if positive cells can be tested individually in principle. The issue of background noise should be considered, since it is possible

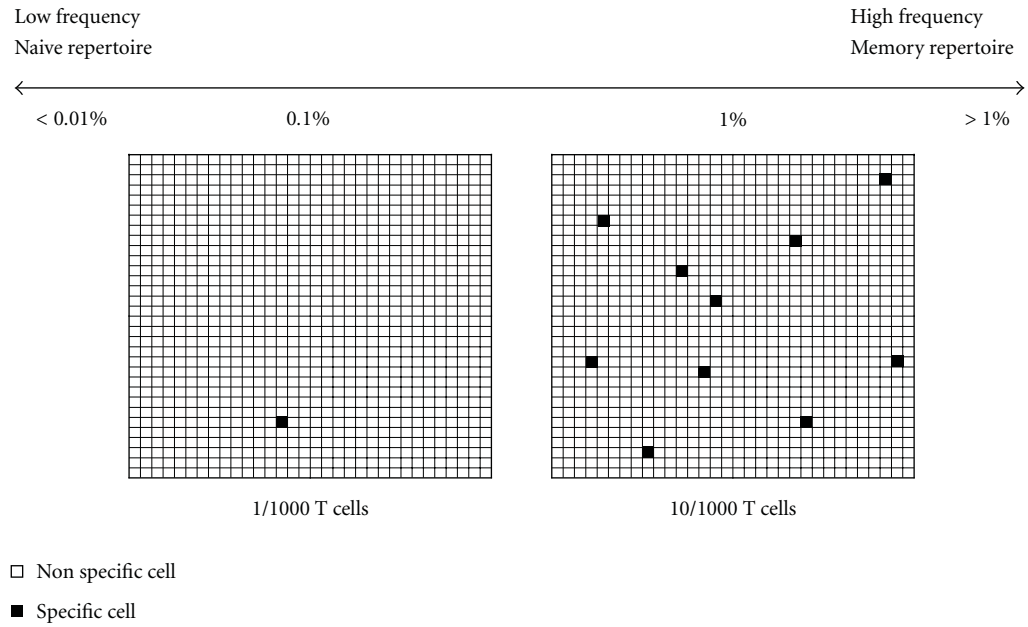


Figure 3: Frequency of specific T cells and culture size. The frequency of antigen specific T cells must be kept in mind when scaling down culture conditions. Reasonable frequencies for specific T cells approximately range from >1% to <0.1% for memory responses to recall protein antigens. Responses to individual epitopes may be at lower frequencies. Frequencies of specific T cells in the naïve repertoire may be <0.01%. Panels represent 1000 lymphocytes containing 1 specific cell (0.1%, left panel) or 10 specific cells (1%, right panel). In case 1000 cells containing 1 specific cell are seeded in a microwell, the specific T-cell may or may not be dispensed according to Poissonian distribution and the signal provided by one single cell may not be detected. In case 1000 cells containing 1% specific cells are seeded, 10 specific cells can provide a detectable positive signal. Therefore, it can be assumed that the smallest number of cells per well should be $>10^3$ cells. This reasoning applies to any T-cell assay adapted to 384 or 1536 well plates. Obviously, for very low frequency responses, more cells need to be seeded, but in this case the intrinsic sensitivity of the assay should be checked to establish a frequency threshold. Unfortunately, reliable cellular standards are not available, unless established antigen specific T-cell lines are used, as discussed in [108].

to detect ex vivo activated T cells in most lymphocyte samples from healthy subjects, with remarkable variability. Therefore, HTP techniques conflict with low frequencies of responding cells. In case a naïve repertoire needs to be analyzed, one possibility can be to go through a preliminary expansion phase with a pool of peptides and then screen the expanded population at a later stage. This obviously requires a second sample of autologous APC, that can be PBMC or EBV transformed B cells for both CD4 and CD8 responses. The generation of antigen specific human T-cell lines from a naïve repertoire for epitope mapping has long been known [109, 110], but a recent appealing technique has opened new perspectives. This "artificial lymph node" technique, described as lymphoid tissue equivalent module, permits T-cell priming in vitro under optimal conditions (www.vaxdesign.com/). This is relevant in human studies, in which antigens cannot be used in vivo for priming like in animal studies. Nevertheless, studies using this approach and evaluation of the system have not yet been published.

After collection of supernatants for cytokine testing, T cells can be saved from wells in which cytokine production was detected for further expansion to generate established T-cell lines. An additional advantage of cell-ELISA is that it can be applied in large scale, collaborative studies. In fact, batches of plates containing predispensed antigen panels can be prepared in the central laboratory and distributed to peripheral laboratories. Here plates are thawed and seeded with PBMC. After incubation, the plates are frozen again and shipped back to the central laboratory for development. This permits the design of multicenter studies on epitope mapping in patients or vaccinees, for whom evaluation of specific cellular immunity would not be feasible with other methods. Simplified culture conditions including HEPES buffer, that avoids the use of a CO2 atmosphere , and a water bath or heating block in place of an incubator, in addition to the possibility of using dry plates for easier storage and shipment, make the assay appealing for use in collaborative studies in resource poor settings. A recent study validated cell-ELISA demonstrating its reproducibility [111]. The possibility of using frozen PBMC and whole blood was also demonstrated. It should be noted that the phenotype of responding cells cannot be defined, unless purified CD4 and CD8 subsets are tested [107] and that the frequency of antigen specific T cells cannot be determined by cell-ELISA. Nevertheless, the number of specific T cells can be indirectly inferred by the amount of secreted IFNg, that has been estimated to be in the range of 1–3 pg per antigen specific cell during a 24 hours culture (Li Pira et al. ms in prep).

6. Conclusions

T-cell epitope mapping allows to identify immunodominant regions on antigenic proteins. Most assays require functional T lymphocytes. Since extended peptide panels are generally tested, large amounts of PBMC are required. Therefore, efforts have been made to reduce the size of culture conditions, independently of the assay selected to define antigen specific responses. The HTP approach is based on miniaturized formats, now available as 384 and 1536 well plates, that require automation to dispense peptides and cells and for final processing. Also dedicated scanners are needed to scan these plates. Several assays described in this review can be adapted to 384 well plates, like ICS and ELIspot, but only cell-ELISA was demonstrated to be feasible in 1536 well plates. This resulted in miniaturized cultures that contained 10^4 PBMC per well. This figure may represent the lowest scaling down threshold that still grants the presence of a detectable number of specific T cells, considering reasonable concentrations of specific memory T cells around 0.1% or less.

Miniaturized assays described here can be considered for two main applications. The first one is HTP epitope mapping, that is, testing extended peptide panels with PBMC obtained from reasonable blood samples. The second one is analysis of specific cellular immunity, that is, testing T-cell responses towards a limited array of antigen preparations (proteins or peptide libraries) using PBMC obtained from small blood samples. It should be noted that many of the assays described here can also be scaled up for preparative purposes to purify virus specific T cells for adoptive immunoreconstitution, as recently reviewed [112]. In conclusion, T-cell epitope mapping is relevant for vaccine development, for dissection of the quality of response in vaccinees or in infected subjects and for definition of peptide libraries as screening tools of cellular immunity. Thus, we foresee continuous improvements in miniaturization and automation of assays to provide an increasing amount of peptide sequence information to be deposited in dedicated databases.

Abbreviations

Th: T helper lymphocyte;
CTL: cytolytic T lymphocyte;
HTP: high throughput;
MM: multimer;
APC: antigen presenting cell;
MHC: major histocompatibility complex;
ICS: intracytoplasmic cytokine staining;
CSC: cytokine secretion and capture;
ELISA: enzyme linked immuno-sorbent assay.

Acknowledgments

This work was supported by Grants from CIPE (Rome, 2007), Regione Liguria (Genoa, 2007); Ministry of Health, Rome (Finalized Project 2008); Ministry of Higher Education, Rome (FIRB RBNEOI-RB9B.003, FIRB RBIP064CRT-006); Italian Health Institute, Rome (AIDS 40G.33, 45G.17, 50G.25); EU, Bruxelles (LSHB-CT-2005-018680).

References

[1] A. Sette and B. Peters, "Immune epitope mapping in the post-genomic era: lessons for vaccine development," *Current Opinion in Immunology*, vol. 19, no. 1, pp. 106–110, 2007.
[2] R. B. Salit, W. M. Kast, and M. P. Velders, "Ins and outs of clinical trials with peptide-based vaccines," *Front Biosci*, vol. 7, pp. 204–213, 2002.
[3] J. M. Vyas, A. G. Van der Veen, and H. L. Ploegh, "The known unknowns of antigen processing and presentation," *Nature Reviews Immunology*, vol. 8, no. 8, pp. 607–618, 2008.
[4] P. E. Jensen, "Recent advances in antigen processing and presentation," *Nature Immunology*, vol. 8, no. 10, pp. 1041–1048, 2007.
[5] E. S. Trombetta and I. Mellman, "Cell biology of antigen processing in vitro and in vivo," *Annual Review of Immunology*, vol. 23, pp. 975–1028, 2005.
[6] C. Watts and S. Powis, "Pathways of antigen processing and presentation," *Reviews in Immunogenetics*, vol. 1, no. 1, pp. 60–74, 1999.
[7] P. A. Van der Merwe and S. J. Davis, "Molecular interactions mediating T cell antigen recognition," *Annual Review of Immunology*, vol. 21, pp. 659–684, 2003.
[8] N. Frahm, K. Yusim, T. J. Suscovich, et al., "Extensive HLA class I allele promiscuity among viral CTL epitopes," *European Journal of Immunology*, vol. 37, no. 9, pp. 2419–2433, 2007.
[9] A. M. Masemola, T. N. Mashishi, G. Khoury, et al., "Novel and promiscuous CTL epitopes in conserved regions of Gag targeted by individuals with early subtype C HIV type 1 infection from Southern Africa," *Journal of Immunology*, vol. 173, no. 7, pp. 4607–4617, 2004.
[10] P. T. P. Kaumaya, S. Kobs-Conrad, S. Y. H. Seo, et al., "Peptide vaccines incorporating a 'promiscuous' T-cell epitope bypass certain haplotype restricted immune responses and provide broad spectrum immunogenicity," *Journal of Molecular Recognition*, vol. 6, no. 2, pp. 81–94, 1993.
[11] P. C. Ho, D. A. Mutch, K. D. Winkel, et al., "Identification of two promiscuous T cell epitopes from tetanus toxin," *European Journal of Immunology*, vol. 20, no. 3, pp. 477–483, 1990.
[12] Y. Adar, Y. Singer, R. Levi, et al., "A universal epitope-based influenza vaccine and its efficacy against H5N1," *Vaccine*, vol. 27, no. 15, pp. 2099–2107, 2009.
[13] S. Létourneau, E.-J. Im, T. Mashishi, et al., "Design and preclinical evaluation of a universal HIV-1 vaccine," *PLoS ONE*, vol. 2, no. 10, article e984, 2007.
[14] J. L. Greenstein, V. C. Schad, W. H. Goodwin, et al., "A universal T cell epitope-containing peptide from hepatitis B surface antigen can enhance antibody specific for HIV gp120," *Journal of Immunology*, vol. 148, no. 12, pp. 3970–3977, 1992.
[15] A. Kumar, R. Arora, P. Kaur, V. S. Chauhan, and P. Sharma, ""Universal" T helper cell determinants enhance immunogenicity of a Plasmodium falciparum merozoite surface antigen peptide," *Journal of Immunology*, vol. 148, no. 5, pp. 1499–1505, 1992.
[16] P. Panina-Bordignon, A. Tan, A. Termijtelen, S. Demotz, G. Corradin, and A. Lanzavecchia, "Universally immunogenic T

cell epitopes: promiscuous binding to human MHC class II and promiscuous recognition by T cells," *European Journal of Immunology*, vol. 19, no. 12, pp. 2237–2242, 1989.

[17] H.-H. Bui, J. Sidney, K. Dinh, S. Southwood, M. J. Newman, and A. Sette, "Predicting population coverage of T-cell epitope-based diagnostics and vaccines," *BMC Bioinformatics*, vol. 7, article 153, 2006.

[18] J. Longmate, J. York, C. La Rosa, et al., "Population coverage by HLA class-I restricted cytotoxic T-lymphocyte epitopes," *Immunogenetics*, vol. 52, no. 3-4, pp. 165–173, 2001.

[19] G. Li Pira, L. Bottone, F. Ivaldi, et al., "Identification of new Th peptides from the cytomegalovirus protein pp65 to design a peptide library for generation of CD4 T cell lines for cellular immunoreconstitution," *International Immunology*, vol. 16, no. 5, pp. 635–642, 2004.

[20] G. Li Pira, L. Bottone, F. Ivaldi, et al., "Generation of cytomegalovirus (CMV)-specific CD4 T cell lines devoid of alloreactivity, by use of a mixture of CMV-phosphoprotein 65 peptides for reconstitution of the T helper repertoire," *Journal of Infectious Diseases*, vol. 191, no. 2, pp. 215–226, 2005.

[21] F. Kern, I. P. Surel, C. Brock, et al., "T-cell epitope mapping by flow cytometry," *Nature Medicine*, vol. 4, no. 8, pp. 975–978, 1998.

[22] M. Provenzano, S. Selleri, P. Jin, et al., "Comprehensive epitope mapping of the Epstein-Barr virus latent membrane protein-2 in normal, non tumor-bearing individuals," *Cancer Immunology, Immunotherapy*, vol. 56, no. 7, pp. 1047–1063, 2007.

[23] J. Duraiswamy, J. M. Burrows, M. Bharadwaj, et al., "Ex vivo analysis of T-cell responses to Epstein-Barr virus-encoded oncogene latent membrane protein 1 reveals highly conserved epitope sequences in virus isolates from diverse geographic regions," *Journal of Virology*, vol. 77, no. 13, pp. 7401–7410, 2003.

[24] A. M. Leen, A. Christin, M. Khalil, et al., "Identification of hexon-specific CD4 and CD8 T-cell epitopes for vaccine and immunotherapy," *Journal of Virology*, vol. 82, no. 1, pp. 546–554, 2008.

[25] L. A. Veltrop-Duits, B. Heemskerk, C. C. Sombroek, et al., "Human $CD4^+$ T cells stimulated by conserved adenovirus 5 hexon peptides recognize cells infected with different species of human adenovirus," *European Journal of Immunology*, vol. 36, no. 9, pp. 2410–2423, 2006.

[26] N. Frahm, T. Korber, C. M. Adams, et al., "Consistent cytotoxic-T-lymphocyte targeting of immunodominant regions in human immunodeficiency virus across multiple ethnicities," *Journal of Virology*, vol. 78, no. 5, pp. 2187–2200, 2004.

[27] G. M. Lauer, K. Ouchi, R. T. Chung, et al., "Comprehensive analysis of $CD8^+$-T-cell responses against hepatitis C virus reveals multiple unpredicted specificities," *Journal of Virology*, vol. 76, no. 12, pp. 6104–6113, 2002.

[28] B. Livingston, C. Crimi, M. Newman, et al., "A rational strategy to design multiepitope immunogens based on multiple Th lymphocyte epitopes," *Journal of Immunology*, vol. 168, no. 11, pp. 5499–5506, 2002.

[29] M. B. Joshi, A. A. Gam, R. A. Boykins, et al., "Immunogenicity of well-characterized synthetic Plasmodium falciparum multiple antigen peptide conjugates," *Infection and Immunity*, vol. 69, no. 8, pp. 4884–4890, 2001.

[30] A. S. De Groot, L. Marcon, E. A. Bishop, et al., "HIV vaccine development by computer assisted design: the GAIA vaccine," *Vaccine*, vol. 23, no. 17-18, pp. 2136–2148, 2005.

[31] A. J. McMichael and T. Hanke, "HIV vaccines 1983–2003," *Nature Medicine*, vol. 9, no. 7, pp. 874–880, 2003.

[32] I. Cebere, L. Dorrell, H. McShane, et al., "Phase I clinical trial safety of DNA- and modified virus Ankara-vectored human immunodeficiency virus type 1 (HIV-1) vaccines administered alone and in a prime-boost regime to healthy HIV-1-uninfected volunteers," *Vaccine*, vol. 24, no. 4, pp. 417–425, 2006.

[33] A. S. De Groot, J. McMurry, L. Marcon, et al., "Developing an epitope-driven tuberculosis (TB) vaccine," *Vaccine*, vol. 23, no. 17-18, pp. 2121–2131, 2005.

[34] R. E. M. Toes, R. C. Hoeben, E. I. H. van der Voort, et al., "Protective anti-tumor immunity induced by vaccination with recombinant adenoviruses encoding multiple tumor-associated cytotoxic T lymphocyte epitopes in a string-of-beads fashion," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 26, pp. 14660–14665, 1997.

[35] J. L. Whitton, N. Sheng, M. B. A. Oldstone, and T. A. McKee, "A "string-of-beads" vaccine, comprising linked minigenes, confers protection from lethal-dose virus challenge," *Journal of Virology*, vol. 67, no. 1, pp. 348–352, 1993.

[36] L.-L. An and J. L. Whitton, "A multivalent minigene vaccine, containing B-cell, cytotoxic T- lymphocyte, and Th epitopes from several microbes, induces appropriate responses in vivo and confers protection against more than one pathogen," *Journal of Virology*, vol. 71, no. 3, pp. 2292–2302, 1997.

[37] W. Fischer, S. Perkins, J. Theiler, et al., "Polyvalent vaccines for optimal coverage of potential T-cell epitopes in global HIV-1 variants," *Nature Medicine*, vol. 13, no. 1, pp. 100–106, 2007.

[38] B. T. M. Korber, C. Brander, B. F. Haynes, et al., *HIV Molecular Immunology*, Los Alamos National Laboratory, Los Alamos, NM, USA, 2007.

[39] A. J. McMichael and R. E. Phillips, "Escape of human immunodeficiency virus from immune control," *Annual Review of Immunology*, vol. 15, pp. 271–296, 1997.

[40] A. S. De Groot, P. M. Knopf, and W. Martin, "De-immunization of therapeutic proteins by T-cell epitope modification," *Developments in Biologicals*, vol. 122, pp. 171–194, 2005.

[41] A. S. De Groot and D. W. Scott, "Immunogenicity of protein therapeutics," *Trends in Immunology*, vol. 28, no. 11, pp. 482–490, 2007.

[42] F. Kern, G. LiPira, J. W. Gratama, F. Manca, and M. Roederer, "Measuring Ag-specific immune responses: understanding immunopathogenesis and improving diagnostics in infectious disease, autoimmunity and cancer," *Trends in Immunology*, vol. 26, no. 9, pp. 477–484, 2005.

[43] F. Sieker, A. May, and M. Zacharias, "Predicting affinity and specificity of antigenic peptide binding to major histocompatibility class I molecules," *Current Protein and Peptide Science*, vol. 10, no. 3, pp. 286–296, 2009.

[44] H. H. Lin, G. L. Zhang, S. Tongchusak, E. L. Reinherz, and V. Brusic, "Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research," *BMC Bioinformatics*, vol. 9, supplement 12, article S22, 2008.

[45] J. C. Tong, T. W. Tan, and S. Ranganathan, "Methods and protocols for prediction of immunogenic epitopes," *Briefings in Bioinformatics*, vol. 8, no. 2, pp. 96–108, 2007.

[46] H.-H. Bui, A. J. Schiewe, H. Von Grafenstein, and I. S. Haworth, "Structural prediction of peptides binding to MHC class I molecules," *Proteins*, vol. 63, no. 1, pp. 43–52, 2006.

[47] W. Martin, H. Sbai, and A. S. De Groot, "Bioinformatics tools for identifying class I-restricted epitopes," *Methods*, vol. 29, no. 3, pp. 289–298, 2003.

[48] A. S. De Groot, A. Bosma, N. Chinai, et al., "From genome to vaccine: in silico predictions, ex vivo verification," *Vaccine*, vol. 19, no. 31, pp. 4385–4395, 2001.

[49] H.-G. Rammensee, J. Bachmann, N. P. N. Emmerich, O. A. Bachor, and S. Stevanović, "SYFPEITHI: database for MHC ligands and peptide motifs," *Immunogenetics*, vol. 50, no. 3-4, pp. 213–219, 1999.

[50] C. G. P. Roberts, G. E. Meister, B. M. Jesdale, J. Lieberman, J. A. Berzofsky, and A. S. De Groot, "Prediction of HIV peptide epitopes by a novel algorithm," *AIDS Research and Human Retroviruses*, vol. 12, no. 7, pp. 593–610, 1996.

[51] M. V. Larsen, C. Lundegaard, K. Lamberth, S. Buus, O. Lund, and M. Nielsen, "Large-scale validation of methods for cytotoxic T-lymphocyte epitope prediction," *BMC Bioinformatics*, vol. 8, article 424, 2007.

[52] P.-M. Kloetzel and F. Ossendorp, "Proteasome and peptidase function in MHC-class-I-mediated antigen presentation," *Current Opinion in Immunology*, vol. 16, no. 1, pp. 76–81, 2004.

[53] K. L. Rock, I. A. York, T. Saric, and A. L. Goldberg, "Protein degradation and the generation of MHC class I-presented peptides," *Advances in Immunology*, vol. 80, pp. 1–70, 2002.

[54] B. Peters, S. Bulik, R. Tampe, P. M. Van Endert, and H.-G. Holzhütter, "Identifying MHC class I epitopes by predicting the TAP transport efficiency of epitope precursors," *Journal of Immunology*, vol. 171, no. 4, pp. 1741–1749, 2003.

[55] S. Tenzer, B. Peters, S. Bulik, et al., "Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding," *Cellular and Molecular Life Sciences*, vol. 62, no. 9, pp. 1025–1037, 2005.

[56] M. V. Larsen, C. Lundegaard, K. Lamberth, et al., "An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency, and proteasomal cleavage predictions," *European Journal of Immunology*, vol. 35, no. 8, pp. 2295–2303, 2005.

[57] V. Brusic, V. B. Bajic, and N. Petrovsky, "Computational methods for prediction of T-cell epitopes—a framework for modelling, testing, and applications," *Methods*, vol. 34, no. 4, pp. 436–443, 2004.

[58] N. Petrovsky and V. Brusic, "Virtual models of the HLA class I antigen processing pathway," *Methods*, vol. 34, no. 4, pp. 429–435, 2004.

[59] S. Lata, M. Bhasin, and G. P. Raghava, "MHCBN 4.0: a database of MHC/TAP binding peptides and T-cell epitopes," *BMC Research Notes*, vol. 2, aricle 61, 2009.

[60] V. A. Walshe, C. K. Hattotuwagama, I. A. Doytchinova, et al., "Integrating in silico and in vitro analysis of peptide binding affinity to HLA-Cw*0102: a bioinformatic approach to the prediction of new epitopes," *PloS One*, vol. 4, no. 11, article e8095, 2009.

[61] M. Schirle, T. Weinschenk, and S. Stevanović, "Combining computer algorithms with experimental approaches permits the rapid and accurate identification of T cell epitopes from defined antigens," *Journal of Immunological Methods*, vol. 257, no. 1-2, pp. 1–16, 2001.

[62] U. Gowthaman and J. N. Agrewala, "In silico tools for predicting peptides binding to HLA-class II molecules: more confusion than conclusion," *Journal of Proteome Research*, vol. 7, no. 1, pp. 154–163, 2008.

[63] M. Wulf, P. Hoehn, and P. Trinder, "Identification of human MHC class I binding peptides using the iTOPIA- epitope discovery system," *Methods in Molecular Biology*, vol. 524, pp. 361–367, 2009.

[64] R. Vita, B. Peters, and A. Sette, "The curation guidelines of the immune epitope database and analysis resource," *Cytometry A*, vol. 73, no. 11, pp. 1066–1070, 2008.

[65] B. Peters and A. Sette, "Integrating epitope data into the emerging web of biomedical knowledge resources," *Nature Reviews Immunology*, vol. 7, no. 6, pp. 485–490, 2007.

[66] R. Vita, L. Zarebski, J. A. Greenbaum, et al., "The immune epitope database 2.0," *Nucleic Acids Research*, vol. 38, pp. D854–D862, 2009.

[67] S. J. Rodda, "Peptide libraries for T cell epitope screening and characterization," *Journal of Immunological Methods*, vol. 267, no. 1, pp. 71–77, 2002.

[68] F. Kiecker, M. Streitz, B. Ay, et al., "Analysis of antigen-specific T-cell responses with synthetic peptides—what kind of peptide for which purpose?" *Human Immunology*, vol. 65, no. 5, pp. 523–536, 2004.

[69] F. Kern, T. Bunde, N. Faulhaber, et al., "Cytomegalovirus (CMV) phosphoprotein 65 makes a large contribution to shaping the T cell repertoire in CMV-exposed individuals," *Journal of Infectious Diseases*, vol. 185, no. 12, pp. 1709–1716, 2002.

[70] M. L. Precopio, T. R. Butterfield, J. P. Casazza, et al., "Optimizing peptide matrices for identifying T-cell antigens," *Cytometry A*, vol. 73, no. 11, pp. 1071–1078, 2008.

[71] N. J. Maeji, A. M. Bray, and H. M. Geysen, "Multi-pin peptide synthesis strategy for T cell determinant analysis," *Journal of Immunological Methods*, vol. 134, no. 1, pp. 23–33, 1990.

[72] J. C. Reece, D. L. McGregor, H. M. Geysen, and S. J. Rodda, "Scanning for T helper epitopes with human PBMC using pools of short synthetic peptides," *Journal of Immunological Methods*, vol. 172, no. 2, pp. 241–254, 1994.

[73] R. M. Valerio, A. M. Bray, R. A. Campbell, et al., "Multipin peptide synthesis at the micromole scale using 2-hydroxyethyl methacrylate grafted polyethylene supports," *International Journal of Peptide and Protein Research*, vol. 42, no. 1, pp. 1–9, 1993.

[74] A. Kramer, U. Reineke, L. Dong, et al., "Spot synthesis: observations and optimizations," *Journal of Peptide Research*, vol. 54, no. 4, pp. 319–327, 1999.

[75] U. Reineke, R. Volkmer-Engert, and J. Schneider-Mergener, "Applications of peptide arrays prepared by the spot-technology," *Current Opinion in Biotechnology*, vol. 12, no. 1, pp. 59–64, 2001.

[76] T. Nahtman, A. Jernberg, S. Mahdavifar, et al., "Validation of peptide epitope microarray experiments and extraction of quality data," *Journal of Immunological Methods*, vol. 328, no. 1-2, pp. 1–13, 2007.

[77] B. Hemmer, C. Pinilla, J. Appel, J. Pascal, R. Houghten, and R. Martin, "The use of soluble synthetic peptide combinatorial libraries to determine antigen recognition of T cells," *Journal of Peptide Research*, vol. 52, no. 5, pp. 338–345, 1998.

[78] M.-H. Sung, Y. Zhao, R. Martin, and R. Simon, "T-cell epitope prediction with combinatorial peptide libraries," *Journal of Computational Biology*, vol. 9, no. 3, pp. 527–539, 2002.

[79] M. Sospedra, C. Pinilla, and R. Martin, "Use of combinatorial peptide libraries for T-cell epitope mapping," *Methods*, vol. 29, no. 3, pp. 236–247, 2003.

[80] J. D. Altman, P. A. H. Moss, P. J. R. Goulder, et al., "Phenotypic analysis of antigen-specific T lymphocytes," *Science*, vol. 274, no. 5284, pp. 94–96, 1996.

[81] S. P. Perfetto, P. K. Chattopadhyay, and M. Roederer, "Seventeen-colour flow cytometry: unravelling the immune system," *Nature Reviews Immunology*, vol. 4, no. 8, pp. 648–655, 2004.

[82] M. Cobbold, N. Khan, B. Pourgheysari, et al., "Adoptive transfer of cytomegalovirus-specific CTL to stem cell transplant patients after selection by HLA-peptide tetramers," *Journal of Experimental Medicine*, vol. 202, no. 3, pp. 379–386, 2005.

[83] M. Toebes, M. Coccoris, A. Bins, et al., "Design and use of conditional MHC class I ligands," *Nature Medicine*, vol. 12, no. 2, pp. 246–251, 2006.

[84] W. W. Kwok, J. A. Gebe, A. Liu, et al., "Rapid epitope identification from complex class-II-restricted T-cell antigens," *Trends in Immunology*, vol. 22, no. 11, pp. 583–588, 2001.

[85] V. Cecconi, M. Moro, S. Del Mare, P. Dellabona, and G. Casorati, "Use of MHC class II tetramers to investigate $CD4^+$ T cell responses: problems and solutions," *Cytometry A*, vol. 73, no. 11, pp. 1010–1018, 2008.

[86] D. S. Chen, Y. Soen, T. B. Stuge, et al., "Marked differences in human melanoma antigen-specific T cell responsiveness after vaccination using a functional microarray," *PLoS Medicine*, vol. 2, no. 10, article e265, 2005.

[87] J. D. Stone, W. E. Demkowicz Jr., and L. J. Stern, "HLA-restricted epitope identification and detection of functional T cell responses by using MHC-peptide and costimulatory microarrays," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 10, pp. 3744–3749, 2005.

[88] P. K. Wallace, J. D. Tario Jr., J. L. Fisher, S. S. Wallace, M. S. Ernstoff, and K. A. Muirhead, "Tracking antigen-driven responses by flow cytometry: monitoring proliferation by dye dilution," *Cytometry A*, vol. 73, no. 11, pp. 1019–1034, 2008.

[89] L. Tesfa, H.-D. Volk, and F. Kern, "Comparison of proliferation and rapid cytokine induction assays for flow cytometric T-cell epitope mapping," *Cytometry A*, vol. 52, no. 1, pp. 36–45, 2003.

[90] M. Wolfl, J. Kuball, W. Y. Ho, et al., "Activation-induced expression of CD137 permits detection, isolation, and expansion of the full repertoire of $CD8^+$ T cells responding to antigen without requiring knowledge of epitope specificities," *Blood*, vol. 110, no. 1, pp. 201–210, 2007.

[91] M. Frentsch, O. Arbach, D. Kirchhoff, et al., "Direct access to $CD4^+$ T cells specific for defined antigens according to CD154 expression," *Nature Medicine*, vol. 11, no. 10, pp. 1118–1124, 2005.

[92] J. J. Zaunders, M. L. Munier, N. Seddiki, et al., "High levels of human antigen-specific $CD4^+$ T cells in peripheral blood revealed by stimulated coexpression of CD25 and CD134 (OX40)," *Journal of Immunology*, vol. 183, no. 4, pp. 2827–2836, 2009.

[93] C. C. Czerkinsky, L. A. Nilsson, H. Nygren, et al., "A solid-phase enzyme-linked immunospot (ELISPOT) assay for enumeration of specific antibody-secreting cells," *Journal of Immunological Methods*, vol. 65, no. 1-2, pp. 109–121, 1983.

[94] C. Czerkinsky, G. Andersson, H.-P. Ekre, L.-A. Nilsson, L. Klareskog, and O. Ouchterlony, "Reverse ELISPOT assay for clonal analysis of cytokine production. I. Enumeration of gamma-interferon-secretion cells," *Journal of Immunological Methods*, vol. 110, no. 1, pp. 29–36, 1988.

[95] A. Gazagne, E. Claret, J. Wijdenes, et al., "A Fluorospot assay to detect single T lymphocytes simultaneously producing multiple cytokines," *Journal of Immunological Methods*, vol. 283, no. 1-2, pp. 91–98, 2003.

[96] S. L. Waldrop, C. J. Pitcher, D. M. Peterson, V. C. Maino, and L. J. Picker, "Determination of antigen-specific memory/effector $CD4^+$ T cell frequencies by flow cytometry: evidence for a novel, antigen-specific homeostatic mechanism in HIV-associated immunodeficiency," *Journal of Clinical Investigation*, vol. 99, no. 7, pp. 1739–1750, 1997.

[97] A. W. Sylwester, B. L. Mitchell, J. B. Edgar, et al., "Broadly targeted human cytomegalovirus-specific $CD4^+$ and $CD8^+$ T cells dominate the memory compartments of exposed subjects," *Journal of Experimental Medicine*, vol. 202, no. 5, pp. 673–685, 2005.

[98] S. C. De Rosa, F. X. Lu, J. Yu, et al., "Vaccination in humans generates broad T cell cytokine responses," *Journal of Immunology*, vol. 173, no. 9, pp. 5372–5380, 2004.

[99] M. L. Precopio, M. R. Betts, J. Parrino, et al., "Immunization with vaccinia virus induces polyfunctional and phenotypically distinctive $CD8^+$ T cell responses," *Journal of Experimental Medicine*, vol. 204, no. 6, pp. 1405–1416, 2007.

[100] R. A. Seder, P. A. Darrah, and M. Roederer, "T-cell quality in memory and protection: implications for vaccine design," *Nature Reviews Immunology*, vol. 8, no. 4, pp. 247–258, 2008.

[101] H. Brosterhus, S. Brings, H. Leyendeckers, et al., "Enrichment and detection of live antigen-specific $CD4^+$ and $CD8^+$ T cells based on cytokine secretion," *European Journal of Immunology*, vol. 29, no. 12, pp. 4053–4059, 1999.

[102] G. Rauser, H. Einsele, C. Sinzger, et al., "Rapid generation of combined CMV-specific $CD4^+$ and $CD8^+$ T-cell lines for adoptive transfer into recipients of allogeneic stem cell transplants," *Blood*, vol. 103, no. 9, pp. 3565–3572, 2004.

[103] G. Li Pira, F. Ivaldi, G. Tripodi, M. Martinengo, and F. Manca, "Positive selection and expansion of cytomegalovirus-specific CD4 and CD8 T cells in sealed systems: potential applications for adoptive cellular immunoreconstitution," *Journal of Immunotherapy*, vol. 31, no. 8, pp. 762–770, 2008.

[104] T. Feuchtinger, S. Matthes-Martin, C. Richard, et al., "Safe adoptive transfer of virus-specific T-cell immunity for the treatment of systemic adenovirus infection after allogeneic stem cell transplantation," *British Journal of Haematology*, vol. 134, no. 1, pp. 64–76, 2006.

[105] S. Mackinnon, K. Thomson, S. Verfuerth, K. Peggs, and M. Lowdell, "Adoptive cellular therapy for cytomegalovirus infection following allogeneic stem cell transplantation using virus-specific T cells," *Blood Cells, Molecules, and Diseases*, vol. 40, no. 1, pp. 63–67, 2008.

[106] D. M. McKinney, R. Skvoretz, M. Qin, et al., "Characterization of an in situ IFN-gamma ELISA assay which is able to detect specific peptide responses from freshly isolated splenocytes induced by DNA minigene immunization," *Journal of Immunological Methods*, vol. 237, no. 1-2, pp. 105–117, 2000.

[107] G. Li Pira, F. Ivaldi, L. Bottone, and F. Manca, "High throughput functional microdissection of pathogen-specific T-cell immunity using antigen and lymphocyte arrays," *Journal of Immunological Methods*, vol. 326, no. 1-2, pp. 22–32, 2007.

[108] G. Li Pira, F. Ivaldi, C. Dentone, et al., "Evaluation of antigen-specific T-cell responses with a miniaturized and automated method," *Clinical and Vaccine Immunology*, vol. 15, no. 12, pp. 1811–1818, 2008.

[109] F. Manca, J. Habeshaw, and A. Dalgleish, "The naive repertoire of human T helper cells specific for gp120, the envelope glycoprotein of HIV," *Journal of Immunology*, vol. 146, no. 6, pp. 1964–1971, 1991.

[110] F. Manca, D. Fenoglio, M. T. Valle, et al., "Human $CD4^+$ T cells can discriminate the molecular and structural context of T epitopes of HIV gp120 and HIV p66," *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology*, vol. 9, no. 3, pp. 227–237, 1995.

[111] G. Li Pira, F. Ivaldi, P. Moretti, M. Risso, G. Tripodi, and F. Manca, "Validation of a miniaturized assay based on IFNg secretion for assessment of specific T cell immunity," *Journal of Immunological Methods*, vol. 355, no. 1-2, pp. 68–75, 2010.

[112] G. Li Pira, M. Kapp, F. Manca, and H. Einsele, "Pathogen specific T-lymphocytes for the reconstitution of the immunocompromised host," *Current Opinion in Immunology*, vol. 21, no. 5, pp. 549–556, 2009.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 396847, 9 pages
doi:10.1155/2010/396847

Research Article

MHC I Stabilizing Potential of Computer-Designed Octapeptides

Joanna M. Wisniewska,[1] Natalie Jäger,[1] Anja Freier,[2] Florian O. Losch,[2] Karl-Heinz Wiesmüller,[3] Peter Walden,[2] Paul Wrede,[4] Gisbert Schneider,[1,5] and Jan A. Hiss[1,5]

[1] *Institute of Organic Chemistry and Chemical Biology, Johann Wolfgang Goethe-Universität, Siesmayerstraße. 70, 60323 Frankfurt am Main, Germany*
[2] *Clinical Research Group Tumor Immunologie, Skin Cancer Center Charité, Clinic of Dermatology, Charité-Universitätsmedizin Berlin, Charitéplatz 1, 10117 Berlin, Germany*
[3] *Institut für Organische Chemie, Universität Tübingen, Auf der Morgenstelle 18, 74076 Tübingen, Germany*
[4] *Institute of Molecular Biology and Bioinformatics, Charité-Universitätsmedizin Berlin, Arnimallee 22, 14195 Berlin, Germany*
[5] *Institute of Pharmaceutical Sciences, ETH Zürich, Wolfgang-Pauli-Street. 10, 8093 Zürich, Switzerland*

Correspondence should be addressed to Jan A. Hiss, jan.hiss@pharma.ethz.ch

Received 28 September 2009; Revised 27 January 2010; Accepted 8 March 2010

Academic Editor: Yongqun Oliver He

Experimental results are presented for 180 in silico designed octapeptide sequences and their stabilizing effects on the major histocompatibility class I molecule H-2K^b. Peptide sequence design was accomplished by a combination of an ant colony optimization algorithm with artificial neural network classifiers. Experimental tests yielded nine H-2K^b stabilizing and 171 nonstabilizing peptides. 28 among the nonstabilizing octapeptides contain canonical motif residues known to be favorable for MHC I stabilization. For characterization of the area covered by stabilizing and non-stabilizing octapeptides in sequence space, we visualized the distribution of 100,603 octapeptides using a self-organizing map. The experimental results present evidence that the canonical sequence motives of the SYFPEITHI database on their own are insufficient for predicting MHC I protein stabilization.

1. Introduction

Cell surface presentation of peptides by major histocompatibility complex I (MHC I) is prerequisite for the initiation of an adaptive immune response [1] and knowledge of MHC-binding peptides is required for the development of vaccines and immunomonitoring protocols for cell-mediated immunity. MHC I molecules are integral membrane proteins that bind peptides with a length of eight up to thirteen amino acids for presentation to CD8$^+$ T lymphocytes [2, 3]. Peptide binding to MHC I stabilizes the MHC-peptide structure at the cell surface of antigen presenting cells. Binding of an octapeptide to an MHC I molecule is defined by the recognition of the peptide by the MHC molecule and its binding affinity [4]. In consequence the binding of the octapeptide leads to stabilization of the MHC-peptide complex on the cell surface. Complex stability is critically influenced by the amino acid sequence of the bound peptide [5], for which Rammensee and coworkers suggested allele-specific canonical sequence motifs [3]. For the octapeptides presented by the mouse MHC I H-2K^b this sequence motif (the canonical or SYFPEITHI motif) is defined as X-X-(Y)-X-[Y/F]-X-X-[L, M, I, V]. Positions three, five, and eight are also referred to as "anchor positions" [6].

For characterization of the H-2K^b stabilizing and nonstabilizing sequence space we designed a diverse set of octapeptides. To explore extensions and alternatives to the known canonical motif, the set of designed octapeptides included sequences containing the full, partial, or no canonical motif. To generate new octapeptides that stabilize H-2K^b we applied an Ant Colony Optimization (ACO) [7] algorithm in combination with neural network classifiers. Artificial neural networks (ANNs) [8] were trained using a set of 423 octapeptides with known H-2K^b stabilizing effect as determined in cellular stabilization assays [9]. The resulting

machine learning classifiers served as fitness function for the ACO algorithm. Navigation through sequence space containing 20^8 possible octapeptides was realized by the ACO *meta*-heuristic which is deduced from social insect behavior [7, 10]. ACO is a probabilistic technique that is not susceptible to dominant ultimate solutions but, due to its "swarm intelligence" based on numerous autonomous agents, open for broad and distributed optimization [11]. New peptide sequences were generated with the ACO algorithm and presented to the trained ANNs for fitness evaluation. During this optimization process the peptide sequences were iteratively adapted according to the ANN fitness score. Finally, the designed octapeptides were synthesized and their stabilization effect was tested experimentally.

We present evidence that rational peptide design utilizing ACO is feasible and leads to novel bioactive peptides with minimal experimental effort. Here, the focus is on the *de novo* design of peptides with a specific MHC I stabilization effect. While some of the designed peptides conform to the known canonical motif for H-2K^b stabilizing peptides, we also show that the degree to which the peptide sequence matches the motif alone is insufficient for prediction of MHC I stabilization. We designed peptides with the complete canonical sequence motif but lacking detectable stabilizing effect. For visualization of the transition between stabilizing and nonstabilizing octapeptides we present a projection of peptide sequence space on a self-organizing map (SOM). This form of representation facilitates the identification of clusters of stabilizing peptides based on their physicochemical properties.

2. Materials and Methods

2.1. Data Set. Training data were compiled from the public databases (AntiJen [12], EPIMHC [13], IEDB [14], MHCBN [15]) and literature sources [16, 17]. The complete dataset contained 423 octapeptides with 242 positive (stabilizing) and 181 negative (nonstabilizing) examples. The annotation of octapeptides as stabilizing and nonstabilizing mouse MHC I protein H-2K^b was based on published experimental data. EC_{50} values below 10 μM were regarded as H-2K^b stabilizing, greater EC_{50} values as nonstabilizing.

2.2. Sequence Encoding. Each residue of an octapeptide was encoded by five different sets of molecular descriptors (See supplementary material (Suppl. 1) available online at doi:10.1155/2010/396847) . The combination of amino acid descriptors served as input for the ANNs. The dimension of the input originated from the coding of each amino acid of the octapeptide by each descriptor.

2.3. The Ant Colony Optimization. The ACO algorithm was implemented using the Java programming language V.1.6 (Sun Microsystems, Inc., Santa Clara, CA, USA). Our ACO algorithm is defined by three consecutive steps: sequence design, path evaluation, and pheromone update, as previously described by Jäger et al. [18]. Peptide design by ACO was terminated when the pheromone concentration had been constant for 10,000 iterations. Together the three steps represent a single iteration of the algorithm (one generation of ants). Ants are computational agents with individual memory coded via "pheromone concentrations". While moving through the search space each ant generates a path corresponding to a new octapeptide. All ants of one generation move independent of each other on individual paths. The resulting paths were evaluated by a fitness function implemented as ANNs. Communication between subsequent generations of ants is achieved through the modification of pheromone concentrations ("stigmergy" [19, 20]). The pheromone matrix represents the collective memory of an ant colony. Only the path with the highest fitness obtained a pheromone update. The advantage of the ACO algorithm is that agents need no information about the complete problem to propose a solution, in our case the complete possible sequence space containing 20^8 octapeptides.

2.4. Artificial Neural Network Fitness Function. Fully connected feedforward networks with a single hidden layer and one output neuron (all neurons with sigmoidal activation) were implemented using Matlab (version 7.4.0.287 R2007a, The Mathsworks Inc.; neural networks toolbox version 5.0.2). The outputs of five ANNs were combined as input for a jury network [21]. The output of the jury served as fitness value (or "score") for the ACO algorithm, which adopted values of the interval $]0, 1[$. Details on the network architecture were described previously [17].

2.5. Stabilization Assay. The stabilization assay was performed as described by Brock et al. [9] using TAP-deficient RMA-S cells (mutagenized Rauscher virus-induced T lymphoma cells of mouse origin) [22]. The cells were cultured in DMEM (Gibco-BRL, Karlsruhe, Germany) with 10% FCS (Sigma-Aldrich, Steinheim, Germany) at 37°C with 8% CO_2. For accumulation of peptide-free MHC I proteins at the cell surface, the cells were cultured for 16 hours at 26°C. The cells were incubated with the peptides in 10 serial dilutions of 100 to 5.6×10^{-4} μg/mL at room temperature for 1 hour, followed by 1-hour incubation at 37°C for denaturation of peptide-free MHC I proteins. The stabilized MHC I proteins were visualized and quantified by flow cytometry using the H-2K^b specific monoclonal antibody B8.24.3 [23] purified in the laboratory from hybridoma culture supernatant by protein G affinity chromatography (Pierce, Darmstadt, Germany) and an R-Phycoerythrin-conjugated anti-mouse antibody (Dianova GmbH, Hamburg, Germany) as secondary reagent. The stabilizing effect of the peptides was determined as mean fluorescence intensity (MFI). The EC_{50} value is the peptide concentration that is required for half-maximal stabilization of the MHC I molecules at the cell surfaces (half-maximal MFI). All peptides were custom-synthesized by EMC microcollections GmbH (Tübingen, Germany).

2.6. SYFPEITHI Score (S-Score). The *S*-score was calculated using the public web server at URL: http://www.syfpeithi.de/

(version July 2009). The *S*-score [6] indicates how well a peptide sequence matches the canonical motif.

2.7. IEDB-ANN Score. For calculation of the Immune Epitope Database- (IEDB-)ANN score the public web server at URL: http://tools.immuneepitope.org/analyze/html/mhc_binding.html (version 2009-09-01) was used. IEDB offers several prediction tools for peptide binding to MHC I molecules (artificial neural networks (ANNs), average relative binding (ARB), stabilized matrix method (SMM), SMM with a peptide: MHC binding energy covariance matrix (SMMPMBEC), scoring matrices derived from combinatorial peptide libraries (comblib_sidney2008), consensus) [24]. The IEBD-ANN method [25] was chosen because it has been determined to be qualitatively best performing [26]. The IEDB-ANN scores are predicted IC_{50} values.

2.8. Self-Organizing Map (SOM). For visualization of the peptide distribution in a high-dimensional descriptor space we used planar SOMs [27] as implemented in the molmap software package [28, 29]. The trained SOM performs a nonlinear mapping from the original descriptor space onto a two-dimensional map. Each data point is assigned to one of a defined number of receptive fields (neurons) of the SOM. SOM training was performed as described previously [30].

3. Results and Discussion

We report the design and examination of 180 octapeptides (Table 1; Suppl. 2) in a cellular MHC I stabilization assay. The ability of an octapeptide to stabilize MHC I was specified as EC_{50} value, which is defined as the peptide concentration required for half-maximal stabilization of the MHC I proteins at the cell surface by the test peptide. Nine of the in silico designed octapeptides exhibited a stabilizing effect (Table 1, Seq. 1–9), and 171 were nonstabilizing (Table 1, Seq. 10–50; Suppl. 2, Seq. 51–180). Six of the nine stabilizing octapeptides had EC_{50} values below 10 μM (Table 1, Seq. 4–9) (i.e., strong MHC I stabilization). Three octapeptides (Table 1, Seq. 1–3) can be regarded as medium stabilizers ($EC_{50} > 20\,\mu$M), two of which completely matched the canonical motif (Table 1, Seq. 1 and 2) with EC_{50} values of 24 μM and 20 μM. Six of the nine octapeptides that correspond to the canonical motif in only two of the three anchor positions yielded EC_{50} values below 10 μM. Peptide 3 had an EC_{50} of 25 μM. Peptide 8 (WKFIFDPV) conforming to the SYFPEITHI motif in two positions (underlined) was the most potent peptide with an EC_{50} of 0.4 μM. Peptide 9 (FHHAHRTV) obeys the canonical motif in just one anchor position but was still among the best stabilizers with an EC_{50} value of 9 μM.

The SYFPEITHI score (*S*-score) is used as a computed index for prediction of stabilizing abilities of peptides for specific MHC molecules [6]. A high value indicates strong stabilizing effects. The *S*-score for the positive control in our experiments (SIINFEKL from ovalbumin [4]) is 25. The *S*-score of a known nonstabilizing octapeptide (LSPFPFDL an endogenous MHC I H2-L^d epitope [31]) is 13. The computed *S*-scores for the nine stabilizing octapeptides were between 8 and 27 (mean = 20 ± 6) reflecting their stabilizing effect (outlier: octapeptide 9 with an *S*-score of 8). Peptides 4, 5, and 7, while exhibiting EC_{50} values similar to sequence 9 ($EC_{50} = 9\,\mu$M), have more than two times greater *S*-scores (*S*-score = 22, *S*-score = 20, *S*-score = 17 (Table 1, Seq. 4, 5, 7)). A possible explanation for the deviation between the SYFPEITHI score and the actual binding behavior could be the anchor position assignment. Sequences 4, 5, and 7 completely fulfill the canonical motif while sequence 9 fulfills it in only one position. Thus the degree of correspondence to the canonical motif is well represented by the *S*-score but does not necessarily reflect the actual binding behavior. This suggests that alternative sequence motifs might confer strong stabilization effects or that the binding motive concept needs to be extended.

We then compared our experimental results to predictions of the Immune Epitope Database (IEDB) [24]. The database offers several prediction methods of which, according to Peters et al. [26], IEDB-ANN [25] is the best performing. For the nine binding peptides found by us, the Pearson correlation [32] between the IC_{50} values predicted by IEDB-ANN for mouse H2-K^b and our measured EC_{50} values is −0.34, which indicates moderate negative correlation. Using the activity cutoff of $IC_{50} < 500$ nM for "medium activity" [25], the IEDB-ANN method correctly predicts four of nine sequences as binding peptides (Table 1, Seq. 1–3, 6).

The remaining 171 octapeptides showed no detectable stabilizing effect at a maximal experimental peptide concentration of 100 μg/mL and were therefore defined as nonstabilizing (Table 1, Seq. 10–50; cf. Suppl. 2, Seq. 51–180). The nonstabilizing octapeptides can be grouped into four categories according to the degree of fulfillment of the canonical SYFPEITHI motif:

Category (i): three canonical anchor amino acids: 12 octapeptides (Table 1, Seq. 10–21),

Category (ii): two canonical anchor amino acids: 16 octapeptides (Table 1, Seq. 22–37),

Category (iii): one canonical anchor amino acid: 23 octapeptides (Table 1, Seq. 38–42; Suppl. 2, Seq. 50–68),

Category (iv): no canonical anchor amino acids: 120 octapeptides (Table 1, Seq. 42–50; Suppl. 2, Seq. 69–180).

For octapeptides of category (i) high *S*-scores were computed in the range between 22 and 28 (mean = 25 ± 2) suggesting a stabilizing ability of the octapeptides. In comparison to the *S*-scores of the nine stabilizing octapeptides, category (i) sequences had higher *S*-scores thus erroneously predicting an even stronger MHC I stabilizing effect. Category (ii) peptides obtained a mean *S*-score of 19 ± 2 still indicating possible MHC I stabilization. Notably, none of these octapeptides had a stabilizing effect in our experiments. The *S*-scores of category (iii) peptides (mean = 11 ± 2) are in agreement with the lack of a stabilizing effect. For category (iv) peptides the computed *S*-scores (mean = 1 ± 1) perfectly agreed with the experimental results obtained for these 120 sequences.

Table 1: Experimentally tested octapeptides. Sequences are given in single-letter code; motif anchor positions fulfilling the SYFPEITHI motif are underlined. *S*-score: SYFPEITHI score [6]. IEDB-ANN score: predicted IC_{50} values by IEDB-ANN method [25]. <500 nM: recommended cutoff for IEDB-ANN score [25] for binders predicted with intermediate affinity; *active*: IEDB-ANN score <500 nM; 0: IEDB-ANN score >500 nM. EC_{50} values correspond to the peptide concentration required for 50% of maximal MHC I protein stabilization; values in brackets are standard deviations ($N = 3$). n.d.: not detectable.

	Peptide number	Sequence	*S*-score	IEDB-ANN score [μM]	<500 nM	Experimental EC_{50} [μM]
Stabilizing peptides	1	FRYPYKTL	27	0.5	active	24.0 (±11)
	2	FRYIYHTL	27	0.2	active	20.0 (±4.0)
	3	FHWDYRGL	22	0.4	active	25.0 (±11)
	4	WRFKYDNL	22	2.3	0	7.0 (±4.0)
	5	WRFVYWRL	20	1.2	0	8.0 (±4.0)
	6	WRFIYFNL	21	0.4	active	3.6 (±1.2)
	7	WDFKFDSV	17	4.7	0	9.0 (±4.0)
	8	WKFIFDPV	16	3.0	0	0.4 (±0.1)
	9	FHHAHRTV	8	10.5	0	9.0 (±4.0)
Nonstabilizing peptides of category (i)	10	FRYEYRSL	28	0.1	active	n.d.
	11	WRYIYHSI	22	2.5	0	n.d.
	12	HRYVYRNI	24	0.4	active	n.d.
	13	FHYAYRSV	23	0.1	active	n.d.
	14	YRYKYDRL	27	0.5	0	n.d.
	15	WRYQYDNL	27	0.7	0	n.d.
	16	WRYRYWSL	27	0.6	0	n.d.
	17	WRYNYDPL	26	1.7	0	n.d.
	18	WRYHYDPL	26	2.4	0	n.d.
	19	WKYQYDNL	27	0.3	active	n.d.
	20	WKYIFDPV	22	1.2	0	n.d.
	21	WKYPFDPV	22	2.0	0	n.d.
Nonstabilizing peptides of category (ii)	22	FRYLYKNA	17	1.2	0	n.d.
	23	FRYVWRTL	18	1.7	0	n.d.
	24	FHYLYHTA	17	0.9	0	n.d.
	25	FRYPYHTP	17	17.8	0	n.d.
	26	FRWEYRGL	22	0.7	0	n.d.
	27	FRHIYRTI	18	9.3	0	n.d.
	28	FRHGYRQI	18	12.6	0	n.d.
	29	FHWAYHTV	17	1.2	0	n.d.
	30	WRWLYKGV	16	4.3	0	n.d.
	31	WRFPYDQL	21	6.6	0	n.d.
	32	WRFKYDPL	21	3.3	0	n.d.
	33	WRFPYDKL	21	6.8	0	n.d.
	34	WRFVYDNL	21	1.9	0	n.d.
	35	WKFKFDPV	17	2.8	0	n.d.
	36	WKINFDPV	17	8.2	0	n.d.
	37	TTEWYTKI	18	3.4	0	n.d.
5 of 23 nonstabilizing peptides of category (iii)	38	RGEVFTAT	13	24.0	0	n.d.
	39	FHYDHRNA	9	6.5	0	n.d.
	40	HRWVFWQP	11	26.0	0	n.d.
	41	SIKNFHYY	12	13.6	0	n.d.
	42	FRHDYHSP	11	29.2	0	n.d.

Table 1: Continued.

	Peptide number	Sequence	*S*-score	IEDB-ANN score [μM]	<500 nM	Experimental EC_{50} [μM]
7 of 120 nonstabilizing peptides of category (iv)	43	TVEQGVTQ	1	37.5	0	n.d.
	44	FHHGHNVP	0	36.8	0	n.d.
	45	QDGHEIHR	1	38.4	0	n.d.
	46	TVEQGVTQ	1	37.5	0	n.d.
	47	GVDQSYLK	0	38.1	0	n.d.
	48	SVENPILR	2	33.7	0	n.d.
	49	NEGWTIHR	1	38.8	0	n.d.
	50	SVDHSIFK	2	36.1	0	n.d.

For category (iii) 5 of 23 octapeptides are given as examples. The remaining 18 octapeptides of category (iii) are listed in Suppl. 2 (Seq. No. 51–68). For category (iv) 7 of 120 ostapeptides are given as examples. The remaining 113 octapeptides of category (iv) are listed in Suppl. 2 (Seq. No. 69–180).

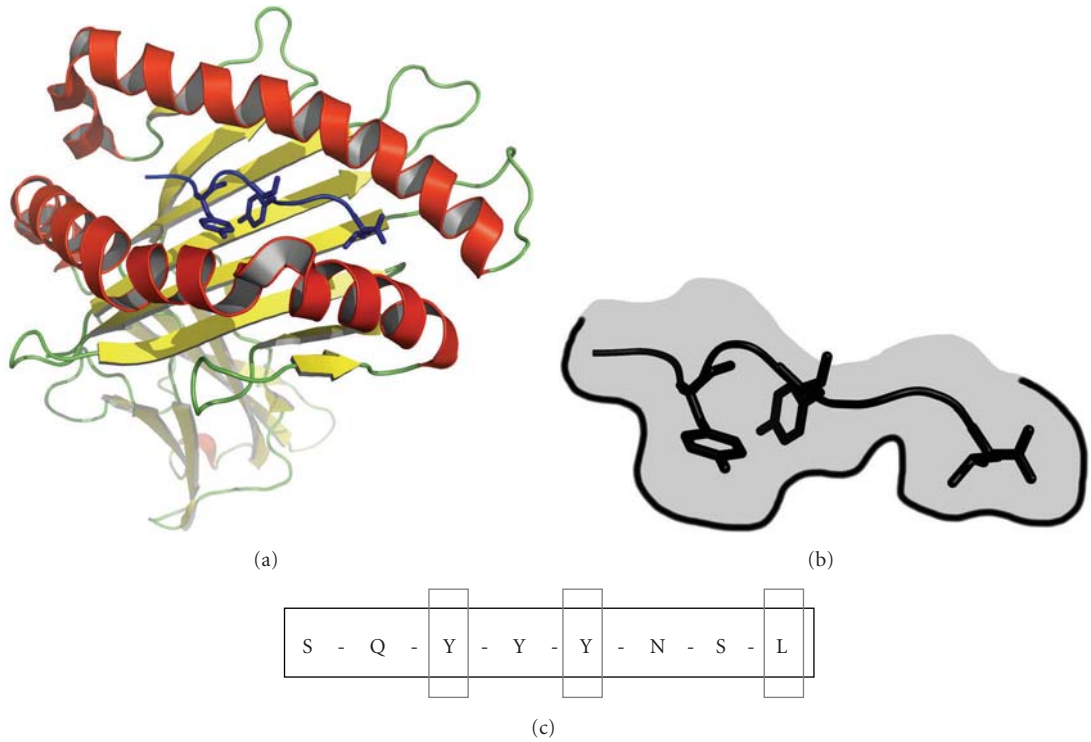


Figure 1: Structural model of murine MHC I molecule H2-K^b with bound octapeptide SQYYYNSL. (a) Crystal structure model of H2-K^b with bound octapeptide SQYYYNSL (PDB entry 2clv [33]). Top view of the peptide binding groove: *red*: alpha helices; *yellow*: beta sheets; *blue*: octapeptide SQYYYNSL, amino acids of the anchor positions are shown as sticks. (b) Schematic side view of the peptide binding groove with the bound octapeptide SQYYYNSL; Amino acid side chains at the anchor positions are shown as sticks. (c) The octapeptide SQYYYNSL with boxed anchor positions of the canonical motif.

The IEDB-ANN method [25] predicts four sequences as "binding", which were determined as "nonbinding" in our experiments (Table 1, category (ii), Seq. 10, 12, 13, 19). The remaining 37 negative sequences are correctly predicted as "nonbinding" (Table 1, categories (ii)–(iv)). Compared to the *S*-score index, the IEDB-ANN method is better suited for identifying nonbinding sequences that contain only a partial canonical motif (categories (ii) and (iii)). Despite these differences, both software tools (*S*-score and IEDB-ANN method) can be recommended for identification of negative (inactive) sequences lacking the canonical motif (categories (iii)–(iv)). Based on these limited data, quantitative IC_{50} predictions of binding/nonbinding peptides by this software seem to be of limited accuracy but qualitative prediction is acceptable.

The experimental results for the 180 designed octapeptides allowed us to reassess the canonical motif. We found 28 inactive octapeptides that conform to the motif in all three (category (i)) or two residue positions (category (ii)). This corroborates the results of Zhong et al. [16] reporting

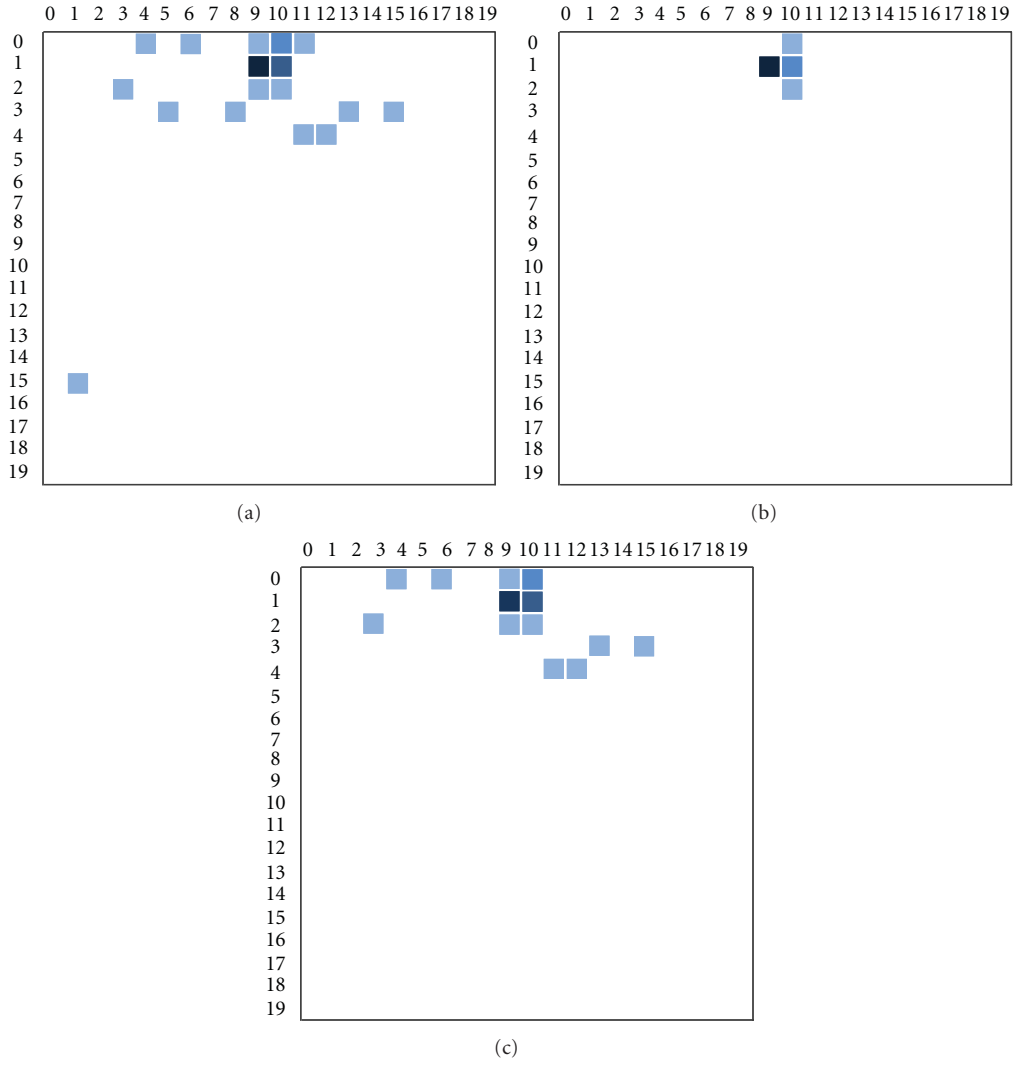


Figure 2: Self-organizing map (SOM) projection of MHC I H2-K^b stabilizing peptides. The SOM was trained with 100,603 octapeptides (603 peptides with known H-$2K^b$ binding affinity and 100,000 octamer sequences randomly generated according to the amino acid frequency of *M. musculus* proteins). Only the distribution of the known stabilizing peptides is shown. The color of a neuron reflects its relative occupancy (white: empty; light blue: slightly occupied; dark blue: most occupied). (a) Distribution of the known 251 H-$2K^b$ stabilizing octapeptides. (b) Distribution of 80 stabilizing octapeptides with amino acid sequences corresponding to the canonical motif in all three anchor positions. (c) Distribution of 152 stabilizing octapeptides with amino acid sequences corresponding to the canonical motif in two residue positions.

one nonstabilizing octapeptide with the canonical motif, and Hiss et al. [17] reporting four nonstabilizing octapeptides corresponding to the motif. Our data suggest that the canonical motif alone is insufficient for predicting MHC I stabilization. Octapeptides stabilize the MHC I molecules by binding into the peptide binding groove which is framed by two alpha helices on top of a eight-stranded beta sheet [2] (Figure 1(a)). Amino acids at sequence position three and five, favorably tyrosine or phenylalanine, can form aromatic interactions with MHC I residues facing the binding groove (Figure 1(b)), which could explain why canonical occupancy often corresponds to stabilizing peptides [34]. In addition, the aliphatic residue at position eight interacts with aliphatic amino acids in a deep pocket of the MHC I peptide binding canyon (Figure 1(b)). Octapeptides that conform completely to the canonical motif but show no stabilizing effect indicate that other amino acids besides the three anchor residues are important for the stabilizing effect. Amino acids at nonmotif positions could interfere with the favorable effects of the three anchor residues and lead to a nonstabilizing peptide.

To visualize the distributions of stabilizing and nonstabilizing octapeptides, we trained a SOM [28, 29] to obtain a two-dimensional map of the peptide distribution. The SOM represents the peptides based on their physicochemical properties coded by a multidimensional vector. Adjacent regions of a given peptide on the SOM represent peptides with similar physicochemical properties. The SOM was trained with a total of 100,603 octapeptides. We randomly generated 100,000 octamer sequences according to the amino acid frequency found in known mouse proteins to mimic

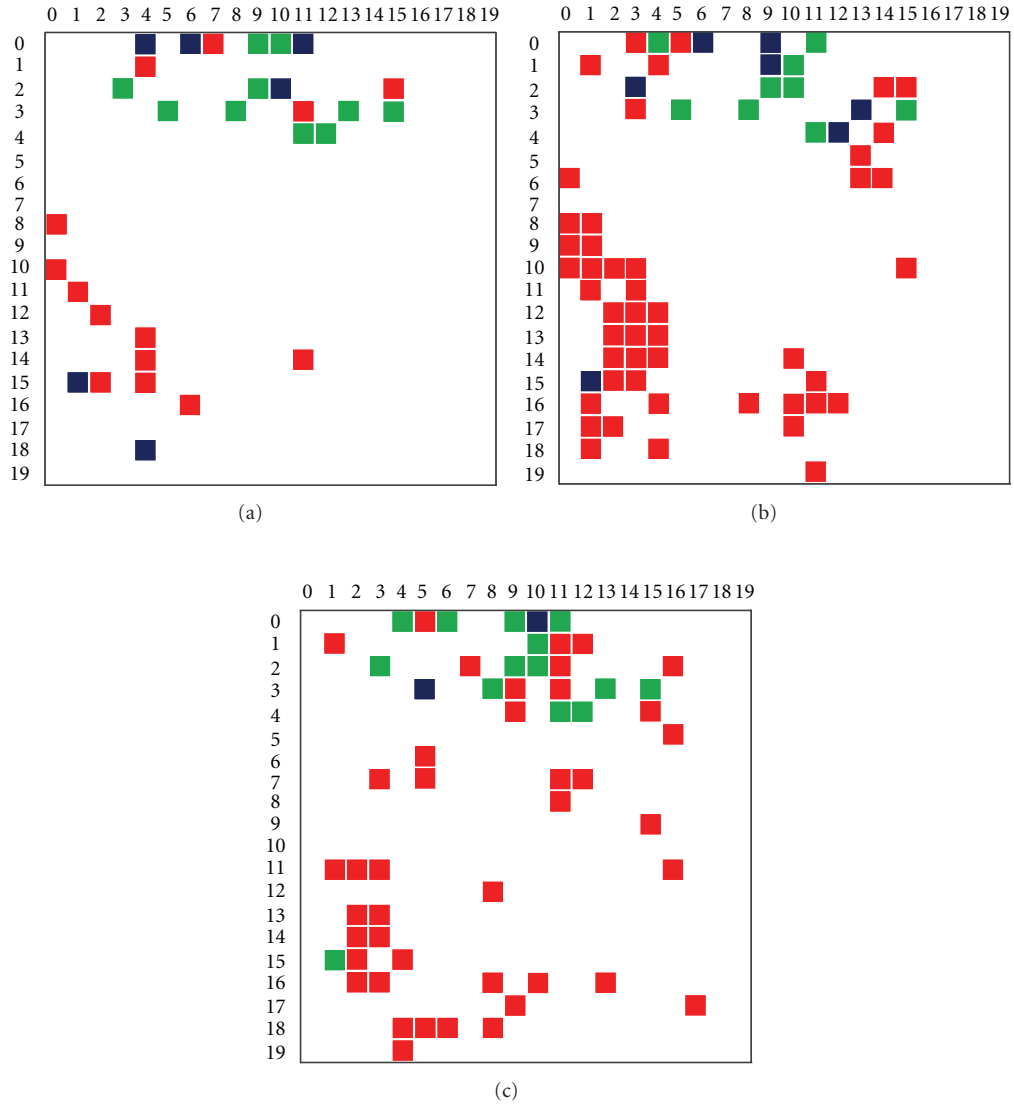


Figure 3: Self-organizing map (SOM) projection of MHC I stabilizing peptides. The SOM was trained with 100,603 octapeptides (603 peptides with known H-2K^b binding affinity and 100,000 octamer sequences randomly designed according to the amino acid frequency of *M. musculus* proteins). Neurons are colored according to the MHC I stabilizing ability of the octapeptides (red: nonstabilizing; blue: mixed; green: stabilizing). (a) Projection of the known 251 stabilizing octapeptides and 24 nonstabilizing octapeptides with amino acid sequences corresponding to the canonical motif in all three positions. (b) Projection of the known 251 stabilizing octapeptides and 93 nonstabilizing octapeptides with amino acid sequences corresponding to the canonical motif in two positions. (c) Projection of the known 251 stabilizing octapeptides and 59 nonstabilizing octapeptides with amino acid sequences corresponding to the canonical motif in only one position.

murine sequence space. For the remaining 603 octapeptides the H-2K^b binding affinities were known from published experimental data (training data set: 423 octapeptides with 242 stabilizing and 181 nonstabilizing peptides and own experimental results: 180 octapeptides with 9 stabilizing and 171 nonstabilizing peptides). The 251 octapeptides with H-2K^b binding affinity are highlighted on the trained SOM presented with Figure 2(a). It is noteworthy that 241 of the 251 stabilizing peptides form a "stabilizing cluster" on the map (neurons 9–11/0, 9-10/1, 9-10/2), which indicates that these peptides are more similar to each other than to the randomly generated octapeptides. The highest density of stabilizing peptides is located in neuron (9/1) which contains 180 sequences. The outlier neuron (1/15) contains octapeptide 9 (FHHAHRTV), a stabilizing octapeptide with a canonical residue in only one anchor position.

The distribution of stabilizing octapeptides fulfilling the canonical motif in all three anchor positions (80 sequences) is presented in Figure 2(b). All 80 octapeptides are located in a "stabilizing cluster". Of the 152 sequences complying in only two anchor positions with the canonical motif 145 are located in this "stabilizing cluster" (Figure 2(c)). The remaining seven of the 152 sequences, which are not located in the "stabilizing cluster", are located in neurons

framing the "stabilizing cluster". The canonical motif is thus overrepresented in the "stabilizing cluster". Although the known active octapeptide sequences constitute an island on the SOM implying similar physicochemical properties, our experimental results suggest that the canonical motif represents only a, albeit maybe dominant, fraction of the MHC I stabilizing sequences (Table 1, Seq. 10–21).

The SOM presented in Figures 3(a)–3(c) presents the distribution of sequences containing only stabilizing (green), only nonstabilizing (red), or containing both stabilizing and nonstabilizing octapeptides (blue). The locations of all 251 stabilizing octapeptides are shown; Figure 3(a) additionally includes category (i) nonstabilizing octapeptides, Figure 3(b) category (ii), and Figure 3(c) category (iii) peptides. The majority (54%) of the nonstabilizing octapeptides of category (i) (24 sequences) is located in neurons surrounding the "stabilizing cluster", implying similarity in terms of the peptide representation by physicochemical properties (Figure 3(a)). Notably, the three neurons (9/1), (10/2), and (11/0) also contain nonstabilizing peptides (blue-colored neurons). Four motif-conform nonstabilizing octapeptides (WRYNYDPL, FRYEYRSL, HRYVYRNI, YRYKYDRL) are located in neuron (9/1) which contains the highest number of stabilizing octapeptides (180 sequences). The remaining (46%) nonstabilizing octapeptides of category (i) populate the lower left quadrant of the SOM. As illustrated in Figure 3(b), this area becomes more densely occupied when the 93 octapeptides of category (ii) are included: 75% of these peptides are located in this area of the SOM. Only two sequences of category (ii) can be found in the "stabilizing cluster": FRYVWRTL and TTEWYTKI (neurons (9/0) and (9/1), Figure 3(b)). Apparently, category (iii) octapeptides are scattered in sequence space (Figure 3(c)). Only one nonstabilizing octapeptide of this category can be found in the "stabilizing cluster" (neuron (10/0), Figure 3(c)).

In summary, we have identified nine stabilizing octapeptides, two of which conform to the canonical motif in all three anchor positions, six fulfill two anchor requirements, and one sequence complies with the canonical motif in just one position. The majority of the designed and tested octapeptides (171 sequences) had no MHC I stabilizing effect. Twelve of the nonstabilizing octapeptides completely conform to the canonical motif, and sixteen fulfill the motif at two anchor positions. 23 octapeptides comply with the canonical motif at one residue position and exhibit no stabilizing effect. The remaining 120 octapeptides share no residue position with this motif. Since the experimental results reported here were not included in the SOM training, the resulting map provides a physicochemically defined distribution of stabilizing and nonstabilizing octapeptides in sequence space. Apparently, the stabilizing octapeptides constitute an island in octapeptide sequences space. Still, nonstabilizing octapeptides are colocated in this area. These nonstabilizing samples fulfill three or two residue positions of the canonical motif. The SOM clusters stabilizing peptides in a section of sequence space with similar physicochemical properties. A hint towards additional stabilizing clusters could be sequence 9 which is not clustered together with the other stabilizing peptides. Furthermore, neurons adjacent to the "stabilizing cluster" also contain MHC I stabilizing sequences, which indicates that the epitope motif concept need to be extended in order to cover and predict alternative stabilizing peptides.

4. Conclusions

Our study confirms and extends the epitope motif concept for MHC-binding peptides proposed by Rammensee and coworkers [6]. We found octapeptides that lack key anchor residues but still exhibit a pronounced MHC I stabilization ability well comparable to peptides that fully conform to the canonical sequence motif. We also present a number of motif-conform but nonstabilizing peptides. This two findings clearly demonstrate that the canonical sequence motif alone is no sufficient criterion for the MHC I stabilizing peptides.

Acknowledgments

The authors are grateful to Norbert Dichter for technical assistance. This study was supported by the Beilstein-Institut zur Förderung der Chemischen Wissenschaften and the Hermann-Willkomm-Stiftung, Frankfurt, Germany. J. M. Wisniewska and N. Jäger contributed equally to this study. G. Schneider and J. A. Hiss share senior authorship.

References

[1] A. R. M. Townsend, J. Rothbard, and F. M. Gotch, "The epitopes of influenza nucleoprotein recognized by cytotoxic T lymphocytes can be defined with short synthetic peptides," *Cell*, vol. 44, no. 6, pp. 959–968, 1986.

[2] P. J. Bjorkman, M. A. Saper, and B. Samroui, "The foreign antigen binding site and T cell recognition regions of class I histocompatibility antigens," *Nature*, vol. 329, no. 6139, pp. 512–518, 1987.

[3] H.-G. Rammensee, K. Falk, and O. Rötzschke, "Peptides naturally presented by MHC class I molecules," *Annual Review of Immunology*, vol. 11, pp. 213–244, 1993.

[4] A. Sette, A. Vitiello, B. Reherman, et al., "The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes," *The Journal of Immunology*, vol. 153, pp. 5586–5592, 1994.

[5] R. C. Su and R. G. Miller, "Stability of surface H-2K^b, H-2D^b, and peptide-receptive H-2K^b on splenocytes," *Journal of Immunology*, vol. 167, pp. 4869–4877, 2001.

[6] H.-G. Rammensee, J. Bachmann, N. P. N. Emmerich, O. A. Bachor, and S. Stevanovic, "SYFPEITHI: database for MHC ligands and peptide motifs," *Immunogenetics*, vol. 50, no. 3-4, pp. 213–219, 1999.

[7] M. Dorigo and G. D. Caro, "Ant colony optimization: a new meta-heuristic," in *Proceedings of the Congress on Evolutionary Computation*, pp. 1470–1477, 1999.

[8] D. E. Rumelhart and J. L. McClelland, *Parallel Distributed Processing*, MIT Press, Cambridge, Mass, USA, 1986.

[9] R. Brock, K.-H. Wiesmüller, G. Jung, and P. Walden, "Molecular basis for the recognition of two structurally different major histocompatibility complex/peptide complexes by a single T-cell receptor," *Proceedings of the National Academy of Sciences of*

the United States of America, vol. 93, no. 23, pp. 13108–13113, 1996.

[10] J. Kennedy and R. Eberhart, "Particle swarm optimization," in *Proceedings of the IEEE International Conference on Neuronal Networks*, vol. 4, pp. 1942–1948, 1995.

[11] M. Dorigo and T. Stützle, *Ant Colony Optimization*, MIT Press, Cambridge, Mass, USA, 2004.

[12] M. J. Blythe, I. A. Doytchinova, and D. R. Flower, "JenPep: a database of quantitative functional peptide data for immunology," *Bioinformatics*, vol. 18, no. 3, pp. 434–439, 2002.

[13] P. A. Reche, H. Zhang, J.-P. Glutting, and E. L. Reinherz, "EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology," *Bioinformatics*, vol. 21, no. 9, pp. 2140–2141, 2005.

[14] B. Peters, J. Sidney, P. Bourne, et al., "The immune epitope database and analysis resource: from vision to blueprint," *PLoS Biology*, vol. 3, article no. 99, 2005.

[15] M. Bhasin, H. Singh, and G. P. S. Raghava, "MHCBN: a comprehensive database of MHC binding and non-binding peptides," *Bioinformatics*, vol. 19, no. 5, pp. 665–666, 2003.

[16] W. Zhong, P. A. Reche, C.-C. Lai, B. Reinhold, and E. L. Reinherz, "Genome-wide characterization of a viral cytotoxic T lymphocyte epitope repertoire," *The Journal of Biological Chemistry*, vol. 278, no. 46, pp. 45135–45144, 2003.

[17] J. A. Hiss, A. Bredenbeck, F. O. Losch, P. Wrede, P. Walden, and G. Schneider, "Design of MHC I stabilizing peptides by agent-based exploration of sequence space," *Protein Engineering, Design and Selection*, vol. 20, no. 3, pp. 99–108, 2007.

[18] N. Jäger, J. M. Wisniewska, J. A. Hiss, et al., "Attractors in sequence space: agent-based exploration of MHC I binding peptides," *Molecular Informatics*, vol. 29, no. 1-2, pp. 65–74, 2010.

[19] P. P. Grassé, "The automatic regulations of collective behavior of social insect and "stigmergy"," *Journal de Psychologie Normale et Pathologique*, vol. 57, pp. 1–10, 1960.

[20] E. Bonabeau, M. Dorigo, and G. Theraulaz, *Swarm Intelligence: from Natural to Artificial Systems*, Oxford University Press, New York, NY, USA, 1999.

[21] P. Baldi and S. Brunak, *Bioinformatics: The Machine Learning Approach*, MIT Press, Cambridge, Mass, USA, 2nd edition, 2001.

[22] H.-G. Ljunggren and K. Kärre, "Host resistance directed selectively against H-2 deficient lymphoma variants. Analysis of the mechanism," *Journal of Experimental Medicine*, vol. 162, no. 6, pp. 1745–1759, 1985.

[23] G. Köhler, K. Fischer-Lindahl, and C. Heusser, "Characterization of a monoclonal anti-H-2K^b antibody," *The Immune System*, vol. 2, pp. 202–208, 1981.

[24] R. Vita, L. Zarebski, J. A. Greenbaum, et al., "The immune epitope database 2.0," *Nucleic Acids Research*, vol. 38, pp. D854–D862, 2010.

[25] M. Nielsen, C. Lundegaard, P. Worning, et al., "Reliable prediction of T-cell epitopes using neural networks with novel sequence representations," *Protein Science*, vol. 12, no. 5, pp. 1007–1017, 2003.

[26] B. Peters, H.-H. Bui, S. Frankild, et al., "A community resource benchmarking predictions of peptide binding to MHC-I molecules," *PLoS Computational Biology*, vol. 2, no. 6, pp. 574–584, 2006.

[27] T. Kohonen, "Self-organized formation of topologically correct feature maps," *Biological Cybernetics*, vol. 43, no. 1, pp. 59–69, 1982.

[28] G. Schneider and P. Wrede, "Artificial neural networks for computer-based molecular design," *Progress in Biophysics and Molecular Biology*, vol. 70, no. 3, pp. 175–222, 1998.

[29] G. Schneider, M. Hartenfeller, M. Reutlinger, Y. Tanrikulu, E. Proschak, and P. Schneider, "Voyages to the (un)known: adaptive design of bioactive compounds," *Trends in Biotechnology*, vol. 27, no. 1, pp. 18–26, 2009.

[30] P. Schneider, Y. Tanrikulu, and G. Schneider, "Self-organizing maps in drug discovery: compound library design, Scaffold-Hopping, repurposing," *Current Medicinal Chemistry*, vol. 16, no. 3, pp. 258–266, 2009.

[31] K. Udaka, T. J. Tsomides, and H. N. Eisen, "A naturally occurring peptide recognized by alloreactive CD8$^+$ cytotoxic T lymphocytes in association with a class I MHC protein," *Cell*, vol. 69, no. 6, pp. 989–998, 1992.

[32] K. Pearson, "Mathematical contributions to the theory of evolution. III. Regression, heredity and panmixia," *Philosophical Transactions of the Royal Society of London*, pp. 253–318, 1896.

[33] N. Auphan-Anezin, C. Mazza, A. Guimezanes, et al., "Distinct orientation of the alloreactive monoclonal CD8 T cell activation program by three different peptide/MHC complexes," *European Journal of Immunology*, vol. 36, no. 7, pp. 1856–1866, 2006.

[34] D. R. Madden, "The three-dimensional structure of peptide-MHC complexes," *Annual Review of Immunology*, vol. 13, pp. 587–622, 1995.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 705821, 8 pages
doi:10.1155/2010/705821

Research Article

MHC Class II Binding Prediction—A Little Help from a Friend

Ivan Dimitrov,[1] Panayot Garnev,[1] Darren R. Flower,[2] and Irini Doytchinova[1]

[1] *Faculty of Pharmacy, Medical University of Sofia, 2 Dunav st., 1000 Sofia, Bulgaria*
[2] *Life and Health Sciences, Aston University, Aston Triangle, Birmingham B4 7ET, UK*

Correspondence should be addressed to Irini Doytchinova, idoytchinova@pharmfac.net

Received 27 September 2009; Revised 20 January 2010; Accepted 22 February 2010

Academic Editor: Yongqun Oliver He

Vaccines are the greatest single instrument of prophylaxis against infectious diseases, with immeasurable benefits to human wellbeing. The accurate and reliable prediction of peptide-MHC binding is fundamental to the robust identification of T-cell epitopes and thus the successful design of peptide- and protein-based vaccines. The prediction of MHC class II peptide binding has hitherto proved recalcitrant and refractory. Here we illustrate the utility of existing computational tools for in silico prediction of peptides binding to class II MHCs. Most of the methods, tested in the present study, detect more than the half of the true binders in the top 5% of all possible nonamers generated from one protein. This number increases in the top 10% and 15% and then does not change significantly. For the top 15% the identified binders approach 86%. In terms of lab work this means 85% less expenditure on materials, labour and time. We show that while existing caveats are well founded, nonetheless use of computational models of class II binding can still offer viable help to the work of the immunologist and vaccinologist.

1. Introduction

Vaccines continue to have an enormous and unprecedented positive impact on humanity and its wellbeing. Hundreds of millions of human lives have been saved since the first vaccine was discovered: Edward Jenner's smallpox vaccine in 1796 [1]. Yet the need to develop and deploy new vaccines has never been more urgent. Infectious disease causes about 25% of global deaths, particularly in children under five. The leading annual causes of death are 2.9 millions for tuberculosis; 2.5 million for diarrhoeal illnesses, especially rotaviruses; a rapidly escalating 2.3 million for HIV/AIDS; and 1.08 million deaths for malaria. There are no effective vaccines for HIV and Malaria, and the only vaccine available for tuberculosis is of limited utility. Consider also the 35 new, previously unknown, infectious diseases identified in the past 25 years: ebola, SARS, Dengue, West Nile fever, and potentially pandemic H5N1 influenza among them.

Historically, vaccines have been attenuated whole pathogen vaccines such as BCG for TB or Sabin's Polio vaccine. Issues of safety have led to the development of other strategies for vaccine development, separately focusing on antigen and epitope vaccines. The epitope is the minimal structure able to evoke an immune response. It is the immunological quantum that lies at the heart of immunity. Epitope-based vaccines have the advantage that many sequences able to induce autoimmunity or adverse reactions can be eliminated. Such vaccines are intrinsically safer: they contain no viable microorganisms and cannot induce microbial disease. However, several significant obstacles must be overcome before epitope-based vaccines can reach the market en masse. One such obstacle is MHC polymorphism.

Major histocompatibility complex (MHC) proteins, also known as human leukocyte antigens (HLA), are glycoproteins which bind within the cell short peptides, also called epitopes, derived from host and/or pathogen proteins, and present them at the cell surface for inspection by T-cells. T cell recognition is a fundamental mechanism of the adaptive immune system by which the host identifies and responds to foreign antigens [2].

There are two classes of MHC molecules: class I and class II. MHC class I molecules typically present peptides from proteins synthesized within the cell (endogenous processing pathway). MHC class II proteins primarily present peptides derived from endocytosed extracellular proteins (exogenous processing pathway). Both classes of MHC proteins are extremely polymorphic. More than 3500 molecules are listed

in IMGT/HLA database [3]. MHC class I proteins are encoded by three loci: HLA-A, HLA-B, and HLA-C. MHC class II proteins also are encoded by three loci: HLA-DR, HLA-DQ, and HLA-DP. The peptide binding site of class I proteins has a closed cleft, formed by a single protein chain (α-chain) [4]. Usually, only short peptides of 8 to 11 amino acids bind in an extended conformation. In contrast, the cleft of class II proteins is open-ended, allowing much longer peptides to bind, although only 9 amino acids actually occupy the site. The class II cleft is formed by two separate protein chains: α and β [4]. Both clefts have binding pockets, corresponding to primary and secondary anchor positions on the binding peptide. The combination of two or more anchors is called a motif. The experimental determination of motifs for every allele is prohibitively expensive in terms of labor, time, and resources. The only practical and useable alternative is a bioinformatics approach.

Many bioinformatics methods exist to predict peptide-MHC binding [5]. Experimentally determined affinities data have formed the basis of many peptide-MHC binding prediction methods, able effectively to discriminate binding-from nonbinding peptides. Such methods include so-called *motifs*, as well as highly sophisticated computer science algorithms—artificial neural networks [6], HMMs [7], and support vector machines (SVMs) [8]—and methods derived from computational chemistry, such as QSAR analysis [9] and structure-based approaches [10].

MHC-binding motifs are an easily understood method of epitope identification. They generate many false-positives and many false-negatives. SVMs are based on statistical theory that seeks to induce a dichotomy of distinct classes. HMMs model systems by assuming them to be Markov processes with unknown parameters. An HMM profile can determine those sequences which exhibit binder-like characteristics. QSAR techniques can refine the peptide interactions within the MHC class I groove by optimizing individual residue-to-residue pairs.

Several sophisticated methods have been created to resolve the dynamic variable-length problem inherent within the class II prediction. Such methods include an iterative "meta-search" algorithm, Ant Colony search, Gibbs sampling algorithm, and multiobjective evolutionary algorithm. Certain new approaches have significantly outperformed more traditional methods [11]. No single method yet proposes a wholly satisfying and satisfactory solution to this dilemma. The efficiency demonstrated by these algorithms is often very different for different class II alleles and there is little overlap between peptide rankings generated by these methods. This has resulted in much pessimism regarding the usefulness of a computational approach. Here, we seek to address this issue.

In the present study, a set of 167 proteins containing 4540 epitopes binding to HLA-DRB1 alleles was compiled and used to test the predictive ability of several publically-available servers for MHC binding predictions. Our aim was not to compare the performance of the available servers, per se. This has been undertaken several times already. Rather, in the context of tasks regularly undertaken by immunologists and vaccinologists, we wish to illustrate the utility of existing computational tools for in silico prediction of peptides binding to class II MHCs. We show that while existing caveats are wellfounded, nonetheless use of computational models of class II binding can still yield viable help to the work of immunologists and vaccinologists.

2. Materials and Methods

2.1. Test Set Used in the Study. At the time of evaluation (December 2009), the Immune Epitope Database [12] contained 10 925 peptides of different length binding to HLA-DRB1 alleles. For the purpose of our study we extracted only binders annotated with an identified source protein. After removing the duplicate sequences and proteins containing unknown amino acids, the final set consisted of 4540 binders, belonging to 167 proteins, and binding to 12 widely spread HLA-DRB1 alleles. The alleles used in the study were DRB1*0101 (2051 binders), DRB1*0301 (190 binders), DRB1*0401 (392 binders), DRB1*0404 (159 binders), DRB1*0405 (244 binders), DRB1*0701 (336 binders), DRB1*0802 (153 binders), DRB1*0901 (160 binders), DRB1*1101 (275 binders), DRB1*1201 (24 binders), DRB1*1302 (243 binders), and DRB1*1501 (313 binders).

2.2. Servers for MHC Class II Binders Prediction Used in the Study. The servers for MHC class II binding prediction used in our assessment were selected on the basis of matching to the following criteria: computational or machine-learning method-based, free web access, and the ability to predict binding to at least 10 of the 12 HLA-DRB1 alleles considered in the study (Table 1).

ProPred [13] predicts MHC class II binding peptides using the quantitative matrix-based pocket profiles of Sturniolo et al. [14]. RANKPEP [15] uses position-specific scoring matrices (PSSM) or profiles which represent the observed sequence-weighted frequency of all amino acids in every position of a sequence alignment. IEDB-ARB [16] is a matrix-based prediction method where the peptide binding score is calculated by multiplying the relative contribution coefficients for each amino acid at each peptide position. The IEDB-SMM align method [17] is based on an integrated alignment and motif identification algorithm and predicts directly peptide binding affinities. MHC2Pred [18] is a SVM-based prediction server. EpiTOP [19] is a newly developed method for MHC class II binding prediction based on proteochemometrics [20]. It is a matrix-based method which considers both peptide and protein binding site amino acids contributions. NetMHCII and NetMHCIIpan are ANN-based methods. NetMHCIIpan takes into account both peptide and MHC sequence information [21].

Most of the servers do not predict binding to all DRB1 alleles used in the test sets. Only servers IEDB, NetMHCIIpan, and EpiTOP make predictions for all 12 DRB1 alleles. Obviously, each server was only evaluated using the alleles it predicts.

Table 1: Servers for MHC class II binders prediction used in the study.

Server	Method	URL
NetMHCII	ANN[a]	http://www.cbs.dtu.dk/services/NetMHCII/
NetMHCIIpan	ANN	http://www.cbs.dtu.dk/services/NetMHCIIpan/
ProPred	QM[b]	http://www.imtech.res.in/raghava/propred/
RANKPEP	QM	http://bio.dfci.harvard.edu/RANKPEP/
IEDB-ARB	QM	http://tools.immuneepitope.org/analyze/html/mhc_II_binding.html
IEDB-SMM	QM	http://tools.immuneepitope.org/analyze/html/mhc_II_binding.html
EpiTOP	QM	http://www.pharmfac.net/EpiTOP/
MHC2Pred	SVM[c]	http://www.imtech.res.in/raghava/mhc2pred/

[a]ANN: artificial neural networks, [b]QM: quantitative matrix, [c]SVM: support vector machine.

3. Performance Evaluation

The evaluation was performed under conditions similar to those an experimental immunologist might use: the complete protein sequences were submitted to a server and the results recorded. Five cutoffs were used to categorise the result: top 5%, 10%, 15%, 20%, and 25% of the predicted binding nonamers. ProPred only returns the top 10% of predictions; for RANKPEP this limitation is 20%. An identified binding peptide was considered to be any nonamer identical in sequence to a nonamer from the set of known binders, and originating from the same protein. Identified binders are shown as a percentage of all binders (*sensitivity*). Although many of the methods give quantitative predictions, in the evaluation study they were used as classification methods.

4. Results

*4.1. HLA-DRB1*0101 Binders Predictions.* The test subset of peptides binding to HLA-DRB1*0101 consisted of 2051 binders. Four of the servers (NetMHCII, NetMHCpan, RANKPEP, and EpiTOP) recognize more than 60% of them in the top 10% (Figure 1(a)). The number of the identified binders increases in the next cutoff steps reaching 93% by NetMHCpan, 91% by EpiTOP, and 88% by NetMHCII and RANKPEP at the top 25%.

*4.2. HLA-DRBI*0301 Binders Predictions.* One hundred and ninety binders from the test set bind to HLA-DRB1*0301. More than half of them are recognized by NetMHCpan, NetMHCII, and ProPred even at the top 5% of the predicted best binders. Sensitivity above 80% is achieved by NetMHCpan, NetMHCII, and EpiTOP at cutoff of 15% (Figure 1(b)).

*4.3. HLA-DRBI*0401 Binders Predictions.* The subset of HLA-DRB1*0401 binders consisted of 392 binders. Most of the servers present well at the top 5% level recognizing more than a half of the binders (Figure 1(c)). At the 20% level NetMHCII, and NetMHCIIpan present best identifying 96% and 95% of the binders, respectively, followed by EpiTOP and RANKPEP (91%).

*4.4. HLA-DRBI*0404 Binders Predications.* The binders to HLA-DRB1*0404 in the test set were 159. Only NetMHCIIpan achieves more than 60% sensitivity at the top 5% cutoff. More than 80% of the known binders are recognized by NetMHCIIpan at the top 10% level and by RANKPEP, EpiTOP, and NetMHCII at the top 15% (Figure 1(d)).

*4.5. HLA-DRB1*0405 Binders Predictions.* The test set contained 244 binders to HLA-DRB1*0405. The performance of the servers on this allele was very similar to HLA-DRB1*0404. NetMHCIIpan and NetMHCII recognize more than 60% of the binders in the top 5% of the predicted best binders. Sensitivity of 80% is achieved by NetMHCIIpan at the top 10% level and by RANKPEP, EpiTOP, and NetMHCII at the top 15% (Figure 1(e)).

*4.6. HLA-DRB1*0701 Binders Predictions.* Three hundred forty four are the binders to HLA-DRB1*0701 in the test set. NetMHCII performs best here identifying more than 60% of the known binders at the top 5% level and 80% of them at the top 10% (Figure 1(f)). Sensitivity of 80% is achieved by NetMHCIIpan, RANKPEP, and EpiTOP at the top 15%. MHC2Pred does not predict affinity to this allele.

*4.7. HLA-DRB1*0802 Binders Predictions.* The binders to HLA-DRB1*0802 in the test set were 153. NetMHCIIpan, NetMHCII, and ProPred achieve more than 60% sensitivity at the top 5% cutoff (Figure 1(g)). A sensitivity higher than 80% has NetMHCIIpan at top 10% level, and NetMHCII and EpiTOP—at top 15%. RANKPEP is not trained to predict binders to this allele

*4.8. HLA-DRB1*0901 Binders Predictions.* The test set contained 160 binders to HLA-DRB1*0901. NetMHCIIpan recognizes 68% of the known binders to this allele in the top 5% (Figure 1(h)). Sensitivity of 80% is achieved by NetMHCII, NetMHCIIpan, EpiTOP, and RANKPEP at the top 15% level. ProPred does not make predictions for this allele.

*4.9. HLA-DRB1*1101 Binders Predictions.* Two hundred seventy five peptides in the test set happen to bind to HLA-DRB1*1101. NetMHCII and NetMHCIIpan recognize 60%

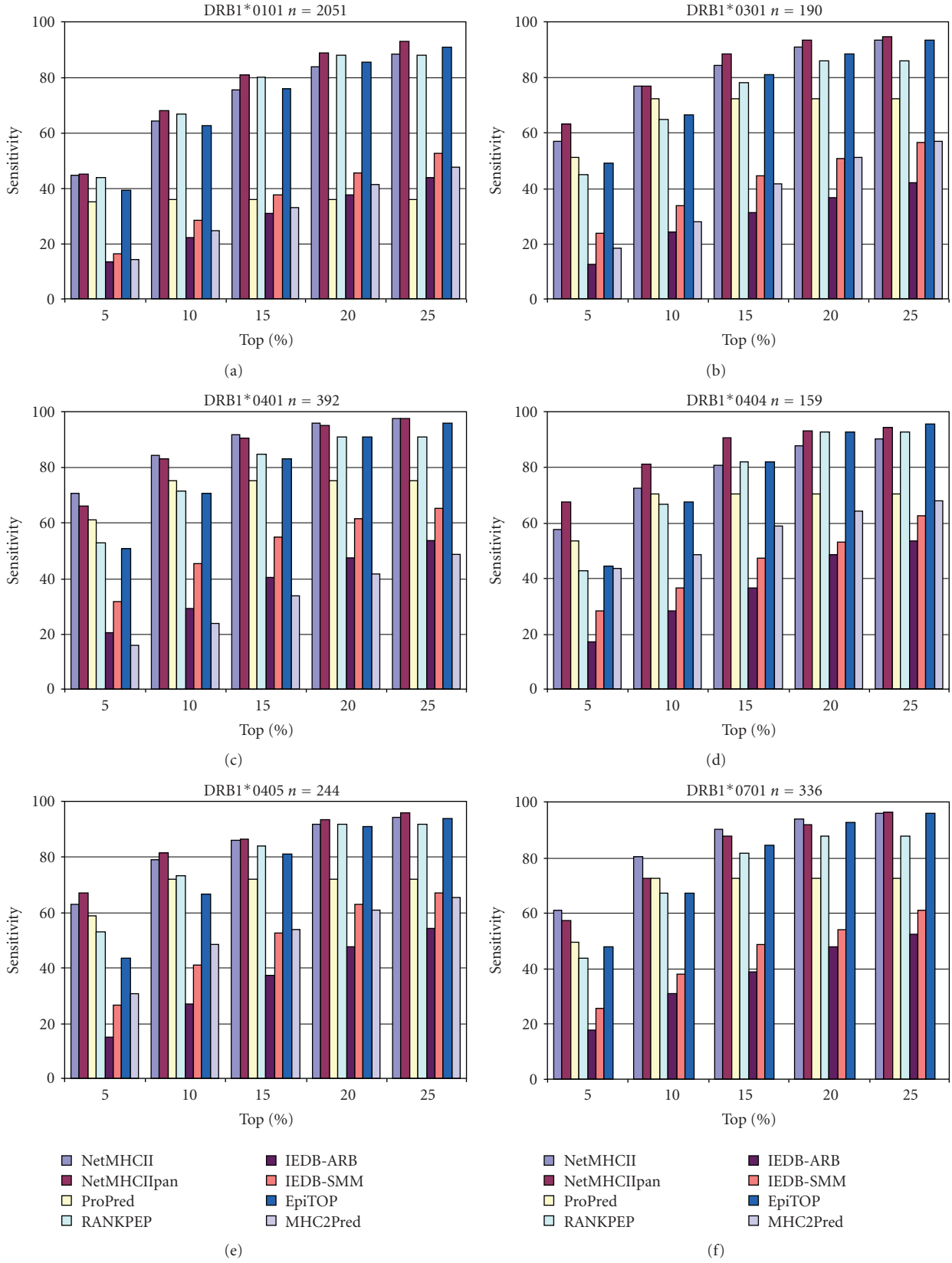


Figure 1: Continued.

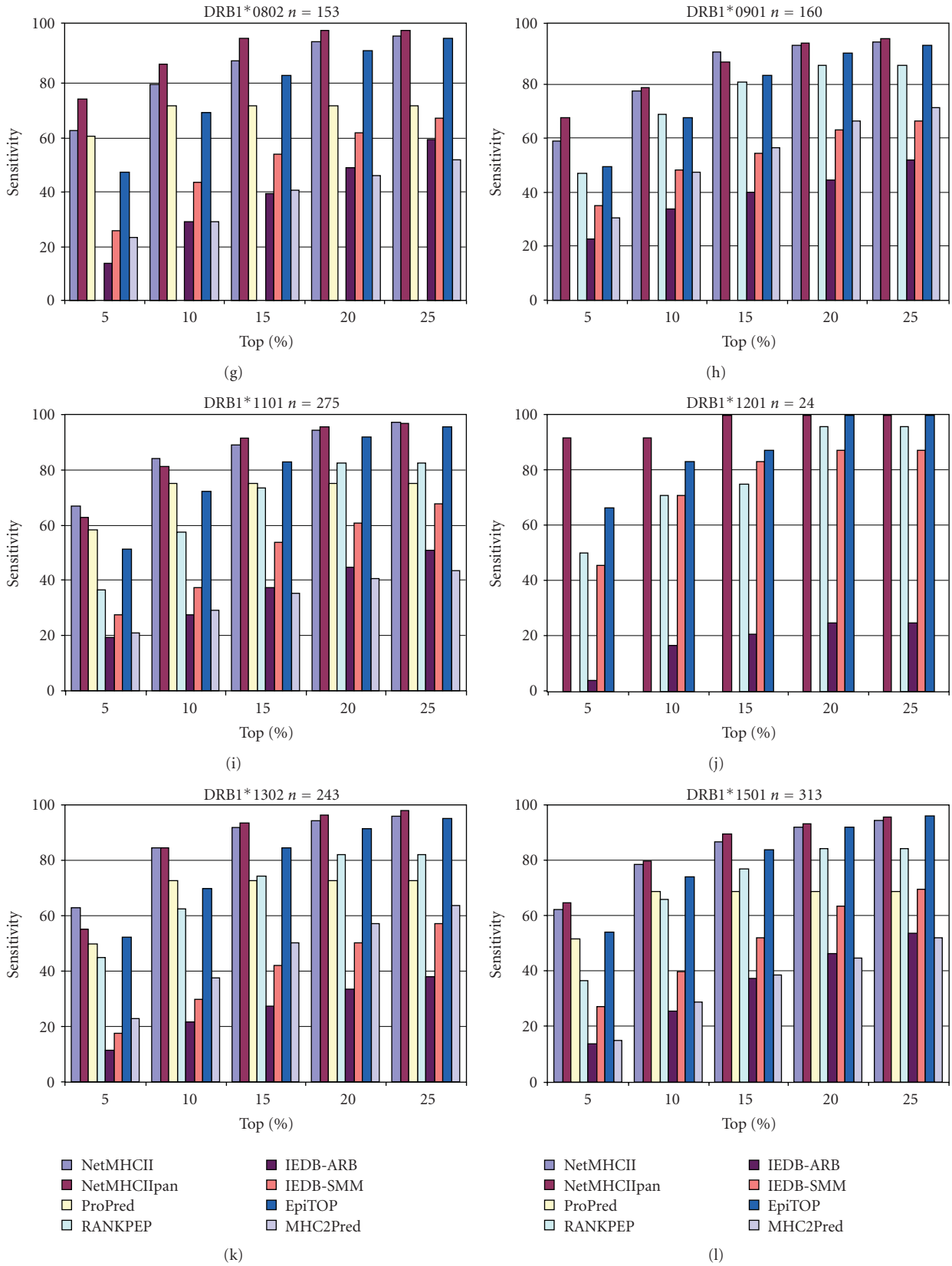


Figure 1: Number of identified binders (*sensitivity*) in the top 5%, 10%, 15%, 20% and 25% of all overlapping nonamers generated from a protein: (a) DRB1*0101, (b) DRB1*0301, (c) DRB1*0401, (d) DRB1*0404, (e) DRB1*0405, (f) DRB1*0701, (g) DRB1*0802, (h) DRB1*0901, (i) DRB1*1101, (j) DRB1*1201, (k) DRB1*1302, (l) DRB1*1501.

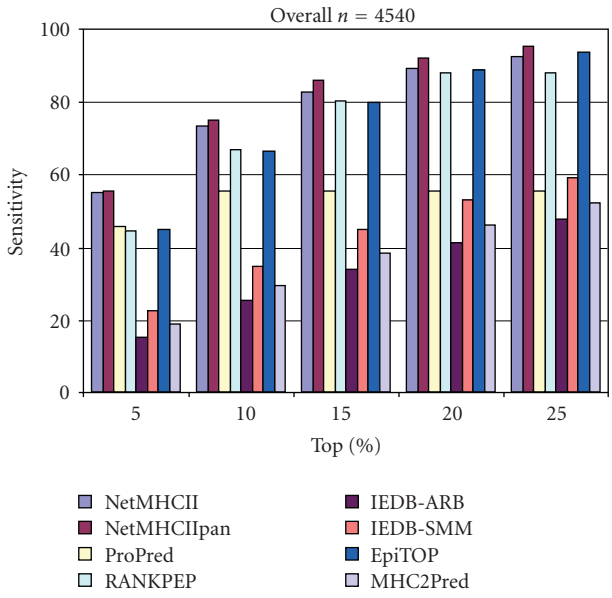


Figure 2: Overall HLA-DRB1 binders prediction.

of them at the top 5% and 80% of them at the top 10% of the predicted best binders (Figure 1(i)).

*4.10. HLA-DRB1*1201 Binders Predictions.* Only 24 peptides are the binders to HLA-DRB1*1201 in the test set. NetMHCII performs best here identifying 92% of the known binders at the top 5% level and 100% of them at the top 15% (Figure 1(j)). Second best is EpiTOP with sensitivity of 83% at the top 10%. NetMHCII, ProPred, and MHC2Pred do not predict affinity to this allele.

*4.11. HLA-DRB1*1302 Binders Predictions.* The binders to HLA-DRB1*1302 in the test set were 243. Only NetMHCIIpan achieves more than 60% sensitivity at the top 5% cutoff (Figure 1(k)). A sensitivity higher than 80% has NetMHC and NetMHCIIpan at top 10% level, and EpiTOP—at top 15%.

*4.12. HLA-DRB1*1501 Binders Predictions.* The subset of peptides binding to HLA-DRB1*1501 consisted of 313 binders. Sixty percent of them are recognized by NetMHCII and NetMHCIIpan at the top 5% cutoff (Figure 1(l)). NetMHCIIpan reaches 80% sensitivity at the top 10%, while NetMHCII and EpiTOP—at the top 15%.

4.13. Overall HLA-DRB1 Binders Predictions. Half of the binders are identified within the top 5% by NetMHCIIpan and NetMHCII (Figure 2). Sensitivity of 80% is achieved by NetMHCIIpan, NetMHCII, RANKPEP, and EpiTOP within the top 15%. This performance is maintained in the top 20% and top 25% and top 25%; see supplementary material available online at doi: 10.1165/2010/705821.

5. Discussion

The peptides presented by MHC class II molecules are derived predominantly from extracellular proteins (not cytosolic as in MHC class I), which are mainly of bacterial origin. They are endocytosed by professional antigen presenting cells (APCs) such as dendritic cells, macrophages and B-cells, digested in lysosomes by cathepsin S [22], and bound by class II molecules in subcellular vesicles. Then the complex peptide-class II molecule is expressed on the cell surface and interacts exclusively with $CD4^+$T cells (helper T-cells, T_HC). T_H cells help to trigger an appropriate immune response which may include localized inflammation and swelling due to recruitment of phagocytes or may lead to a full-force antibody-mediated immune response due to the activation of B cells.

MHC binding is the stage of antigen presentation which we understand best in both the class I and class II pathways. It is also the most discriminating stage within the presentation-recognition pathways. That is why most of the methods for T-cell epitope prediction are in practice methods for MHC binding prediction. The successful prediction of MHC class II binding is more difficult than the successful prediction of class I binding. The main difficulty is the unrestricted length of class II epitopes. Compared with MHC class I binders, which are limited up to 11 amino acids, though sometimes longer, the open-ended class II binding site does not constrain peptide lengths, allowing binding of peptides consisted of up to or more than 25 amino acids. However, as X-ray structures show, the class II binding site is always occupied by nine amino acids, with the rest of the peptide protruding at both sides. Thus, class II prediction methods need to identify the binding nonamer for each sequence and then develop a predictive model. Most of the models consider the principle "one nonamer per binding sequence". Experimental data, however, suggests the possible existence of multiple registers with different nonameric core regions within a binding peptide, each serving as recognition sites for MHC class II molecules [23]. Our previous work indicated that when an easily distinguished good binder is not available in the peptide sequence, the binding affinity is a degenerate average of affinities from several binding subsequences [24]. The multiple registers of binding peptides and the degenerate recognition might explain the lower predictive ability for MHC class II compared to those for MHC class I [25].

The present comparative study was provoked by two emergent themes in the immunoinformatics literature concerning the development of T-cell epitope predictive methods. The first includes papers describing new predictive methods. Most such studies tend to favour the method they invent and this inevitably influences the choice of test data and the way in which that test is conducted. Thus, the test set is never truly independent and the reason for this is the immanent bias associated with the process of selecting test data.

The second trend, which inspired us to perform this study, was the prevalent negative attitude of many scientists to the efforts of the immunoinformatic community, which has sought to develop methods that will facilitate

experimental lab work reducing significantly the practitioners' laborious times and resource. Where data are sufficiently abundant, methods aimed at predicting class I MHC binding seem to work well [26]. However, other types of prediction have the reputation of working poorly. for example, the structure-driven prediction of class I and class II T-cell epitopes [27]. More recently, several comparative studies have shown that, in particular, the prediction of class II T-cell epitopes is suboptimal [28–30]. This has led to unnecessary pessimism regarding the utility of such approaches; here we seek to redress this unnecessarily cautious attitude and perception.

For different reasons, reliable prediction across the board remains elusive, and will continue so for some time. Nonetheless, even extant predictive informatic methodologies, when applied shrewdly and combined synergistically with other approaches, can deliver clear and useful benefits.

Methods are limited by data. What is needed are properly designed data sets which can properly sample and explore the multidimensional space accessed by all congeneric molecules under examination. No data-driven method can go beyond the training data: all methods are better at interpolating than extrapolating. It is only by having excellent and general data that we can hope for general and excellent models.

Except in rare cases, data is usually multi-dimensional, and each dimension will typically be correlated, to a greater or lesser extent, with one or more other dimensions. Together, these many dimensions delineate a space: a space of structural variation or variation of properties. If our data is itself of sufficient quality and provides a good enough coverage of the space, then straightforward methods drawn from, say, computer science—of which there are indeed very many—are now of sufficient accuracy to generate models of high predictive accuracy.

The quality, quantity, and availability of data must increase and improve, particularly for class II. Nonetheless, existing methods built on extant data can, in spite of its inherent imperfections, still be useable and useful, as we show here. Prediction is captive to its underlying data. Bias within the data places strict limitations upon the interpretability and generality of models derived from it. In general, for MHC-peptide binding experiments, the sequences of peptides studied are very biased in terms of amino acid composition, often favouring hydrophobic sequences. This arises, in part, from preselection processes that result in self-reinforcement. Binding motifs are often used to reduce the experimental burden of epitope discovery. Very sparse sequence patterns are matched and the corresponding subset of peptides tested, with an enormous reduction in sequence diversity.

In this study, we choose a test set of 4540 known binders to HLA-DRB1 alleles from the Immune Epitope Database [12]. The binders were of different length and originate from known source proteins. The number of the source proteins was 167. Peptides bound to 12 widely spread HLA-DRB1 alleles. The evaluation was performed under conditions similar to those which an experimental immunologist might use: the complete sequence of a protein of interest is submitted to an available web server and the results recorded. Five thresholds were used: top 5%, 10%, 15%, 20%, and 25% of all overlapping nonamers generated from a protein. The identified binders were presented as a percent of each allele binders (*sensitivity*) used in the study.

The results from our evaluation are unexpectedly encouraging. Half of the known binders are identified within the top 5% by two servers: NetMHCIIpan and NetMHCII (Figure 2). For some alleles, the sensitivity at this level reaches, 92% (NetMHCIIpan for HLA-DRB1*1201). The sensitivity increases in the top 10% and 15% and then does not alter significantly. Most of the servers achieve sensitivity of 80% at the top 15%. In terms of lab work, this means 85% less expenditure on expended on materials, labour, and time. Apart from NetMHCIIpan and NetMHCII, two other servers—RANKPEP and EpiTOP—perform well. The moderate performance of IEDB-ARB, IEDB-SMM, and MHC2Pred, which are well known and widely used servers for MHC class II binding prediction, is surprising. One possible explanation could be the high levels of specificity, predefined in this evaluation through the cutoffs of top 5% to top 25% of the predicted best binders. The aim of these high levels of specificity is to avoid the great number of false positives often generated by quantitative predictions.

The results we present here for MHC binding prediction illustrate the usefulness of computational tools for the everyday work of the immunologist. These results go a long way to refuting - and refuting eloquently - the apparent skepticism among many experimental immunologists and vaccinologists regarding efforts by immunoinformatics and immunoinformaticians to design and develop highly predictive computational tools for the in silico identification of T-cell epitopes. Accurate prediction remains vital for the future of vaccine informatics and for vaccinology as a whole. It is important to realize what can be and what cannot be done.

In future work, we will use the specific results and general know-how generated in this study to inform data-fusion approaches to improve class II peptide-MHC binding prediction. In particular, we will explore the use of optimised voting algorithms to generate a viable meta-predictor, which unites the output of several prediction methods in an intelligent manner so that the combined output is more accurate and more reliable than any individual prediction program. Such approaches have been widely employed in other areas and even in immunoinformatics: Trost et al. have addressed class I binding [31], while Karpenko et al. have used this approach to predict class II MHCs binding [32]. We will seek to capitalise upon the as-yet-unrealised potential of such approaches.

What vaccine informatics offers are tools that can become important components of a deeper, broader experimental and clinical endeavour. The models we explore above are tools of true utility replete with practical real-world applications. Vaccinology and immunology, as disciplines, need only embrace such methods; in their turn such techniques will liberate the vaccinologist and immunologist from the drudgery of uninformed experimentation, allowing them to design better, faster, smarter ways of discovery of new reagents, diagnostics, and vaccines.

Acknowledgment

The research was supported by The National Science Fund of Ministry of Education and Science, Bulgaria (Grant 02-115/2008). The supplementary material provides detailed results for HLA-DRB1 binding prediction using a test set of 4540 known binders.

References

[1] D. R. Flower, "Vaccines: their place in history," in *Bioinformatics for Vaccinology*, pp. 1–54, Wiley-Blackwell, Oxford, UK, 2008.

[2] D. R. Flower, "Vaccines: how they work," in *Bioinformatics for Vaccinology*, pp. 73–112, Wiley-Blackwell, Oxford, UK, 2008.

[3] J. Robinson, M. J. Waller, P. Parham, et al., "IMGT/HLA and IMGT/MHC: sequence databases for the study of the major histocompatibility complex," *Nucleic Acids Research*, vol. 31, no. 1, pp. 311–314, 2003.

[4] C. A. Janeway, P. Travers, M. Walport, and J. D. Capra, "The recognition of antigen," in *Immunobiology: The Immune System in Health and Disease*, pp. 79–194, 1999.

[5] D. R. Flower, "Vaccines: data driven prediction of binders, epitopes and immunogenicity," in *Bioinformatics for Vaccinology*, pp. 167–216, Wiley-Blackwell, Oxford, UK, 2008.

[6] K. Gulukota and C. DeLisi, "Neural network method for predicting peptides that bind major histocompatibility complex molecules," *Methods in Molecular Biology*, vol. 156, pp. 201–209, 2001.

[7] H. Noguchi, R. Kato, T. Hanai, et al., "Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules," *Journal of Bioscience and Bioengineering*, vol. 94, no. 3, pp. 264–270, 2002.

[8] J. Wan, W. Liu, Q. Xu, Y. Ren, D. R. Flower, and T. Li, "SVRMHC prediction server for MHC-binding peptides," *BMC Bioinformatics*, vol. 7, article 463, 2006.

[9] I. A. Doytchinova, V. Walshe, P. Borrow, and D. R. Flower, "Towards the chemometric dissection of peptide—HLA-A*0201 binding affinity: comparison of local and global QSAR models," *Journal of Computer-Aided Molecular Design*, vol. 19, no. 3, pp. 203–212, 2005.

[10] M. N. Davies, C. K. Hattotuwagama, D. S. Moss, M. G. B. Drew, and D. R. Flower, "Statistical deconvolution of enthalpic energetic contributions to MHC-peptide binding affinity," *BMC Structural Biology*, vol. 6, article 5, 2006.

[11] J. Salomon and D. R. Flower, "Predicting Class II MHC-Peptide binding: a kernel based approach using similarity scores," *BMC Bioinformatics*, vol. 7, article 501, 2006.

[12] C. P. Toseland, D. J. Taylor, H. McSparron, et al., "AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical and cellular data," *Immunome Research*, vol. 1, article 4, 2005.

[13] H. Singh and G. P. S. Raghava, "ProPred: prediction of HLA-DR binding sites," *Bioinformatics*, vol. 17, no. 12, pp. 1236–1237, 2001.

[14] T. Sturniolo, E. Bono, J. Ding, et al., "Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices," *Nature Biotechnology*, vol. 17, no. 6, pp. 555–561, 1999.

[15] P. A. Reche, J.-P. Glutting, H. Zhang, and E. L. Reinherz, "Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles," *Immunogenetics*, vol. 56, no. 6, pp. 405–419, 2004.

[16] H.-H. Bui, J. Sidney, B. Peters, et al., "Automated generation and evaluation of specific MHC binding predictive tools: ARB matrix applications," *Immunogenetics*, vol. 57, no. 5, pp. 304–314, 2005.

[17] M. Nielsen, C. Lundegaard, and O. Lund, "Prediction of MHC class II binding affinity using SMM-align, a novel stabilization matrix alignment method," *BMC Bioinformatics*, vol. 8, article 238, 2007.

[18] http://www.imtech.res.in/raghava/mhc2pred.

[19] http://www.pharmfac.net/EpiTOP/.

[20] I. Dimitrov, P. Garnev, D. R. Flower, and I. Doytchinova, "Peptide binding to the HLA-DRB1 supertype: a proteochemometrics analysis," *European Journal of Medicinal Chemistry*, vol. 45, no. 1, pp. 236–243, 2010.

[21] M. Nielsen, C. Lundegaard, T. Blicher, et al., "Quantitative predictions of peptide binding to any HLA-DR molecule of known sequence: NetMHCIIpan," *PLoS Computational Biology*, vol. 4, no. 7, Article ID e1000107, 2008.

[22] L. Delamarre, M. Pack, H. Chang, I. Mellman, and E. S. Trombetta, "Differential lysosomal proteolysis in antigen-presenting cells determines antigen fate," *Science*, vol. 307, no. 5715, pp. 1630–1634, 2005.

[23] J. C. Tong, G. L. Zhang, T. W. Tan, J. T. August, V. Brusic, and S. Ranganathan, "Prediction of HLA-DQ3.2β ligands: evidence of multiple registers in class II binding peptides," *Bioinformatics*, vol. 22, no. 10, pp. 1232–1238, 2006.

[24] I. A. Doytchinova and D. R. Flower, "Towards the in silico identification of class II restricted T-cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction," *Bioinformatics*, vol. 19, no. 17, pp. 2263–2270, 2003.

[25] V. Brusic, G. Rudy, M. Honeyman, J. Hammer, and L. Harrison, "Prediction of MHC class II-binding peptides using an evolutionary algorithm and artificial neural network," *Bioinformatics*, vol. 14, no. 2, pp. 121–130, 1998.

[26] B. Peters, H. H. Bui, S. Frankild, et al., "A community resource benchmarking predictions of peptide binding to MHC-I molecules," *PLoS Computational Biology*, vol. 2, no. 6, article e65, 2006.

[27] B. Knapp, U. Omasits, S. Frantal, and W. Schreiner, "A critical cross-validation of high throughput structural binding prediction methods for pMHC," *Journal of Computer-Aided Molecular Design*, vol. 23, no. 5, pp. 301–307, 2009.

[28] Y. EL-Manzalawy, D. Dobbs, and V. Honavar, "On evaluating MHC-II binding peptide prediction methods," *PLoS ONE*, vol. 3, no. 9, article e3268, 2008.

[29] H. H. Lin, G. L. Zhang, S. Tongchusak, E. L. Reinherz, and V. Brusic, "Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research," *BMC Bioinformatics*, vol. 9, supplement 12, article S22, 2008.

[30] U. Gowthaman and J. N. Agrewala, "In silico tools for predicting peptides binding to HLA-class II molecules: more confusion than conclusion," *Journal of Proteome Research*, vol. 7, no. 1, pp. 154–163, 2008.

[31] B. Trost, M. Bickis, and A. Kusalik, "Strength in numbers: achieving greater accuracy in MHC-I binding prediction by combining the results from multiple prediction tools," *Immunome Research*, vol. 3, no. 1, article 5, 2007.

[32] O. Karpenko, L. Huang, and Y. Dai, "A probabilistic meta-predictor for the MHC class II binding peptides," *Immunogenetics*, vol. 60, no. 1, pp. 25–36, 2008.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 101932, 7 pages
doi:10.1155/2010/101932

Research Article

SAROTUP: Scanner and Reporter of Target-Unrelated Peptides

Jian Huang, Beibei Ru, Shiyong Li, Hao Lin, and Feng-Biao Guo

Key Laboratory for Neuroinformation of Ministry of Education, School of Life Science and Technology, University of Electronic Science and Technology of China, Chengdu 610054, China

Correspondence should be addressed to Jian Huang, hj@uestc.edu.cn

Received 30 September 2009; Accepted 29 January 2010

Academic Editor: Yongqun Oliver He

As epitope mimics, mimotopes have been widely utilized in the study of epitope prediction and the development of new diagnostics, therapeutics, and vaccines. Screening the random peptide libraries constructed with phage display or any other surface display technologies provides an efficient and convenient approach to acquire mimotopes. However, target-unrelated peptides creep into mimotopes from time to time through binding to contaminants or other components of the screening system. In this study, we present SAROTUP, a free web tool for scanning, reporting and excluding possible target-unrelated peptides from real mimotopes. Preliminary tests show that SAROTUP is efficient and capable of improving the accuracy of mimotope-based epitope mapping. It is also helpful for the development of mimotope-based diagnostics, therapeutics, and vaccines.

1. Introduction

In 1985, Smith pioneered phage display technology, an in vitro methodology and system for presenting, selecting and evolving proteins and peptides displayed on the surface of phage virion [1]. Since then, phage display has developed rapidly and become an increasingly popular tool for both basic research such as the exploration of protein-protein interaction networks and sites [2–4], and applied research such as the development of new diagnostics, therapeutics, and vaccines [5–10]. Usually, the protein used to screen the phage display library is termed as target and the genuine partner binding to the target is called template. Peptide mimicking the binding site on the template and binding to the target is defined as mimotope, which was first introduced by Geysen et al. [11]. One type of the most frequently used targets is monoclonal antibody. In this situation, the template is the corresponding antigen inducing the antibody, and the mimotope is a mimic of the genuine epitope. In fact, the original definition of mimotope given by Geysen et al. goes "A mimotope is defined as a molecule able to bind to the antigen combining site of an antibody molecule, not necessarily identical with the epitope inducing the antibody, but an acceptable mimic of the essential features of the epitope [11]." Mimotopes and the corresponding epitope are considered to have similar physicochemical properties and spatial organization. The mimicry between mimotopes and genuine epitope makes mimotopes reasonable solutions to epitope mapping, network inferring, and new diagnostics, therapeutics, and vaccines developing.

Powered by phage display technology, mimotopes can be acquired in a relatively cheap, efficient and convenient way, that is, screening phage-displayed random peptides libraries with a given target. However, not all phages selected out are target-specific, because the target itself is only one component of the screening system [12]. From time to time, phages reacting with contaminants in the target sample or other components of the screening system such as the solid phase (e.g., plastic plates) and the capturing molecule (e.g., streptavidin, secondary antibody) rather than binding to the actual target are recovered with those target-specific binders (displaying mimotopes) during the rounds of panning. Peptides displayed on these phages are called target-unrelated peptides (TUP), a term coined recently by Menendez and Scott in a review [12].

The results from phage display technology might be a mixture of target-unrelated peptides and mimotopes, and it can be difficult to discriminate TUP from mimotopes since the binding assays used to confirm the affinity of peptides for the target often employ the same components as the

initial panning experiment [12]. Therefore, target-unrelated peptides might be taken into study as mimotopes if the researchers are not careful enough. Undoubtedly, this will make the conclusion of the study dubious. Several such examples have been discussed in references [12, 13]. Obviously, target-unrelated peptides are not appropriate candidates for the development of new diagnostics, therapeutics, and vaccines. For mimotope-based epitope mapping, target-unrelated peptides are main noise. If TUP is included in the mapping, the input data is improper and the result might be misleading [14]. There are now quite a few programs for mimotope based epitope mapping, none of them, however, has a procedure to scan, report and exclude target-unrelated peptides [15–23].

In this study, we describe a web server named SAROTUP, which is an acronym for "Scanner And Reporter Of Target-Unrelated Peptides". SAROTUP was coded with Perl as a CGI program and can be freely accessed and used to scan peptides acquired from phage display technology. It is capable of finding, reporting, and precluding possible target-unrelated peptides, which is very helpful for the development of mimotope-based diagnostics, therapeutics, and vaccines. The power and efficiency of SAROTUP was also demonstrated by preliminary tests in the present study.

2. Materials and Methods

2.1. Compilation of TUP Motifs. Recently, Menendez and Scott reviewed a collection of target-unrelated peptides recovered in the screening of phage-displayed random peptide libraries with antibodies [12]. They divided their collection into several categories according to the component of the screening system to which target-unrelated peptides bind. They also derived one or more TUP motifs for each category. Very recently, Brammer et al. reported a completely new type of target-unrelated peptides [13]. In the review of Menendez and Scott, target-unrelated selection is due to the binding to contaminants or components other than target; however, in the report of Brammer et al., target-unrelated selection is due to a coincident point mutation in the phage library [12, 13]. We compiled a set of 23 TUP motifs from the above two references [12, 13], including 12 motifs specific for the capturing agents, 5 motifs specific for the constant region of antibody, 3 motifs specific for the screening solid phase, 2 motifs specific for the contaminants in the target sample, and 1 motif for a mutation in phage library (Table 1). All motifs are presented in patterns according to Prosite format [24].

2.2. Implementation of SAROTUP. The SAROTUP was implemented as a free online service, powered by Apache and Perl. Three pages are designed and integrated into a tabbed web interface with cascading style sheets codes. The core program of SAROTUP was *sar.pl*, a CGI script coded with Perl. In this script, the 23 TUP motifs were converted to regular expressions, which were then used to match each input peptide sequence.

2.3. Construction of Test Data Sets. We constructed two-test data sets from [12, 13, 15–23, 25, 26]. The first data set contains 8 cases; 6 of them are sourced from test cases used in extant programs for mimotope-based epitope mapping [15–23]; the left 2 are cases studies published recently [25, 26]. As shown in Table 2, the target of each case in the first data set is monoclonal antibody and the structure of corresponding antigen-antibody complex has been resolved, which is used to derive its structural epitope as the golden standard for evaluation. For each case, there is one or more sets of peptides recovered from phage display technology. These peptides have been used in mimotope-based epitope mapping by other researchers. We scanned each set of peptides with SAROTUP. If target-unrelated peptides were found, a new panel of peptides excluding TUP was produced. The old and the new panel of peptides were then used to predict epitope using Mapitope or PepSurf [15, 21, 22]. Finally, the results were compared to show if SAROTUP could improve the performance of mimotope-based epitope mapping.

The second data set is composed of 100 peptides in raw sequence format. It has two groups. The first group has 77 sequences compiled from the first data set without any known TUP motifs; the second group has 23 sequences sourced from [12, 13] with various TUP motifs. The mixture of the two groups of sequences made the second data set, which was then used as the sample input and can be used to evaluate the efficiency of SAROTUP.

3. Results and Discussion

3.1. Web Interface of SAROTUP. As a free online service, the web interface of SAROTUP has successfully been implemented as a tabbed web page. The left tab is the default page, providing a brief introduction to this web service. The right tab is a more detailed help page. Click the middle tab will display a web form. The upper section of the form is for basic input (Figure 1). The users can either paste a set of peptide sequences in the text box or upload a sequence file to the SAROTUP server for scanning. As shown in Figure 1, a panel of peptides in raw sequence format taken from the b12 test case was pasted in the text box. Besides the raw sequences, SAROTUP also supports peptides in FASTA format. However, only the standard IUPAC one-letter amino acid codes are accepted at present.

The lower section of the form has a series of options (Figure 2). It includes three drop lists for the screening target, screened library, and screening solid phase, respectively. It also has two groups of check boxes for the capturing reagents and contaminants in the target sample or screening system. By default, SAROTUP will scan each peptide against all the known 23 TUP motifs. However, the users can customize their scan according to their experiment at this section.

After the users submit their request, the scanning results of SAROTUP will be displayed on the middle tabbed page. If any target-unrelated peptides are found, they will be reported in a table. At the same time, a new panel of peptides excluding target-unrelated peptides is produced and can be

Table 1: Known patterns of target-unrelated peptides.

TUP Category	TUP Pattern	Mechanism in brief
Capturing agents	H-P-[QM], G-D-[WF]-x-F, W-x-W-L, E-P-D-W-[FY], D-V-E-x-W-[LIV]	Binding to streptavidin
	W-x-P-P-F-[RK]	Binding to biotin
	W-[TS]-[LI]-x(2)-H-[RK]	Binding to Protein A
	R-T-[LI]-[TS]-K-P, [LFW]-x-F-Q, W-I-S-x(2)-D-W, Q-[LV]-[LV]-Q, RTYK	Binding to secondary antibody
Constant region of antibody (the target)	S-S-[IL], GELVW, G-[LI]-T-D-[WY], [RHK]-P-S-P, P-S-P-[RK]	Binding to the Fc fragment
Screening solid phase	W-x(2)-W, WHWRLPS, F-H-x(2)-W	Binding to plastic
Contaminants in the target sample	F-H-E-x-W-P-[ST]	Binding to contaminant bovine serum albumin
	QSYP	Binding to contaminant bovine IgG
Phage mutation	HAIYPRH	Growing faster than other phages

Table 2: A summary of the first test data set for SAROTUP.

Target	Template	Complex	Peptides	Source
17b	HIV gp120 envelope glycoprotein (gp120)	1GC1	11	[15]
trastuzumab	human receptor tyrosine-protein kinase erbB-2 (HER2)	1N8Z	5	[20]
82D6A3	human von Willebrand factor (vWF)	2ADF	5	[19]
13b5	HIV-1 capsid protein p24	1E6J	14	[15]
BO2C11	human coagulation factor VIII	1IQD	27	[19]
cetuximab	human epidermal growth factor receptor	1YY9	4	[20]
80R	SARS-coronavirus spike protein S1	2GHW	42 + 18	[26]
b12	HIV gp120 envelope glycoprotein (gp120)	2NY7	2 + 32 + 19	[25]

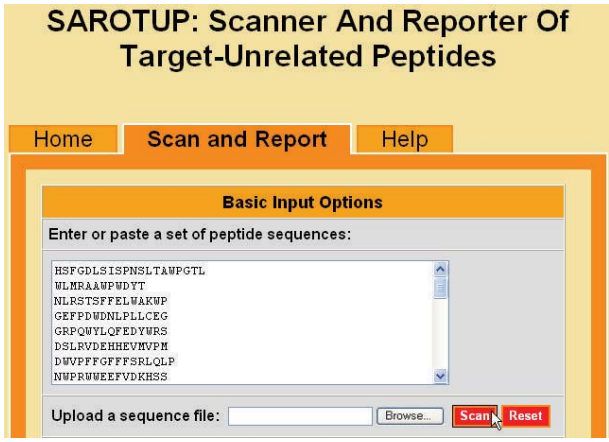


Figure 1: Snapshot of the upper section of SAROTUP.

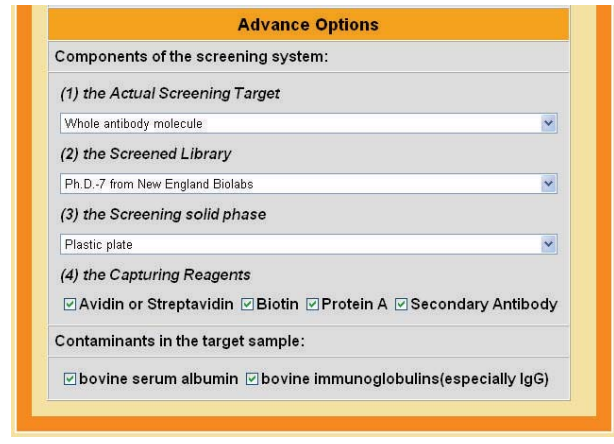


Figure 2: Snapshot of the lower section of SAROTUP.

downloaded from the hyperlink created by the SAROTUP server (Figure 3). The file of the new panel of peptides will be stored on the server for a month and then automatically deleted.

We have tested SAROTUP on the Internet Explorer (version 6.0), Mozilla Firefox (version 3.5.2), and Google Chrome (version 3.0). Although SAROTUP looks a little bit different among different browsers, it works normally on all browsers tested.

3.2. Power of SAROTUP. As shown in Table 2, the first test data set has 11 panels of peptides acquired from phage display libraries screened with 8 targets. In the 11 panels of peptides SAROTUP scanned, there were target-unrelated

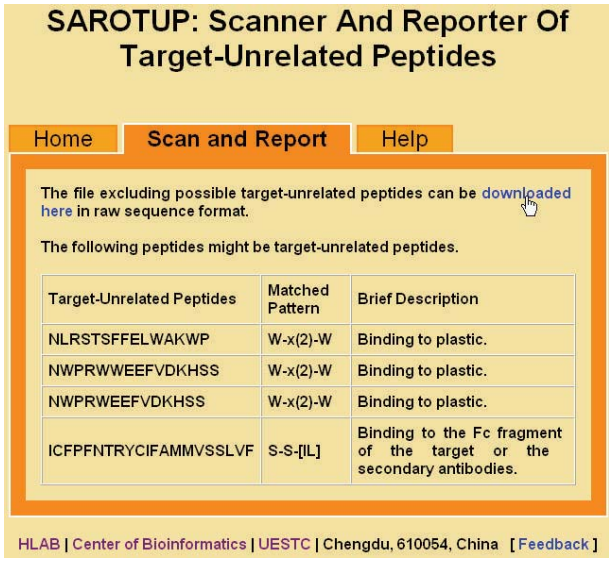


Figure 3: Snapshot of SAROTUP result page. Target-unrelated peptides in the b12 test case are reported in the table. The new panel of peptides excluding the target-unrelated peptides can be downloaded from the hyperlink.

peptides in 3 panels from cetuximab, 80R, and b12 test case, respectively (Table 3). This result suggested it was not rare that target-unrelated peptides sneaked into biopanning results and then were taken as mimotopes in study. In all, 7 target-unrelated peptides were found; 4 of them were due to binding to plastic; the left 3 were due to binding to the Fc fragment (Table 3).

For the above 3 cases, the genuine epitopes recognized by cetuximab, 80R, and b12 monoclonal antibodies are compiled according to the CED records [27] and PDBsum entries[28]. Mapitope or PepSurf [15, 21, 22] were used to perform mimotope-based epitope prediction with or without SAROTUP procedure. For Mapitope and PepSurf algorithm, the library type was set to "random"; the stop codon modification was set to "none"; and all other options were in default. The cluster with best score was taken as the predicted epitope. In the cetuximab case, PepSurf was used because there are only four or three peptides in the panel, statistically too few for Mapitope. In the case of 80R and b12, Mapitope was used because many peptides in the two cases exceeding the length limit of PepSurf, that is, 14 amino acids. If a predicted residue is identical with a residue in the true epitope, it is underlined (Table 4).

As shown in Table 4, the number of true positives improved from zero to four in the cetuximab case with SAROTUP procedure. When it came to the b12 case, the number of true positives increased from one to eight. SAROTUP did not improve the number of true positives in the 80R case when the parameters are same to the cetuximab and b12 cases. However, when the distance parameter was adjusted from default (i.e., 9 Å) to 10 Å, SAROTUP did increase the number of true positive residues from eight to eleven. These results indicate: (1) epitope prediction based on mimotope will be interfered if target-unrelated peptides are taken as mimotopes; (2) SAROTUP can improve the performance of mimotope based epitope mapping through cleaning the input data.

We also scanned the second data set to evaluate the efficiency of SAROTUP. The second data set has 100 peptides, varying from 6 to 22 residues long. Suppose that matching each pattern to each peptide manually costs 10 seconds, then it would take a researcher more than 6 hours (23,000 seconds) to look through the second data set for target-unrelated peptides, even if he is as prompt during the whole period. However, it took only one second for SAROTUP to complete this work. Besides, a table of target-unrelated peptides and a new panel of peptides excluding TUP was produced at the same time by SAROTUP. It is true that some target-unrelated peptides can be identified through control and binding competition experiments. However, using SAROTUP first will certainly save a lot of labor, money, and time for researchers in this area.

3.3. Extending of SAROTUP. Although the target of all tests described previously were monoclonal antibodies, SAROTUP can be customized and used in scanning the results from phage display technology using other targets such as enzymes and receptors. This is because their screening systems are similar. For the same reason, we can also expect that SAROTUP will extend its use to other similar in vitro evolution techniques, such as ribosome display [29–31], yeast display [32], and bacterial display [33–35].

Furthermore, SAROTUP will not only benefit the mimotope-based epitope mapping, but also the development of new diagnostics, therapeutics, and vaccines. Target-unrelated peptides are not appropriate candidates for mimotope based diagnostics, therapeutics, and vaccines, since they are mimics to components or contaminants of the screening system rather than target. Therefore, it is reasonable to find and exclude possible target-unrelated peptides from the candidate list of new diagnostics, therapeutics, and vaccines. Take the cetuximab as an example. Riemer et al. screened a phage-displayed random peptides library with the cetuximab and got four different peptides, that is, QFDLSTRRLK, QYNLSSRALK, VWQRWQKSYV, and MWDRFSRWYK [36]. As described previously, we scanned the four "mimotopes" with SAROTUP and the result suggested that the peptide VWQRWQKSYV might be a TUP. Indeed, the dot blot analysis of Riemer et al. showed that QYNLSSRALK bound the cetuximab with high affinity but VWQRWQKSYV was less reactive with the cetuximab [36]. Trying to develop a mimotope vaccine, Riemer et al. synthesized two-vaccine constructs with the peptide QYNLSSRALK and VWQRWQKSYV, respectively. After immunization mice with these constructs, they found that either the cetuximab or the antibodies induced by the QYNLSSRALK vaccine construct inhibited the growth of A431 cancer cells significantly. The inhibition of the antibodies induced by the VWQRWQKSYV vaccine construct however, was not statistically significant when compared with the inhibition caused by the isotype control antibody [36].

Table 3: Target-unrelated peptides in the first test data set.

Target	Target-unrelated peptides	Mechanism
cetuximab	V**W**QR**W**QKSYV	Binding to plastic
80R	CE**SSL**CLMYSLGPPA,	Binding to the Fc fragment
	YSTP**SSI**LDTHPLYK	Binding to the Fc fragment
b12	NLRSTSFFEL**W**AK**W**P	Binding to plastic
	N**W**PR**W**WEEFVDKHSS	Binding to plastic
	N**W**PR**W**EEFVDKHSS	Binding to plastic
	ICFPFNTRYCIFAMMV**SSL**VF	Binding to the Fc fragment

Table 4: Mimotope-based epitope prediction with or without SAROTUP procedure.

Target	Prediction without SAROTUP procedure	Genuine epitope	Prediction with SAROTUP procedure
cetuximab	N134, E136, S137, I138, Q139, W140, R141, Q164, K185, L186, T187, K188	P349, R353, L382, Q384, Q408, H409, Q411, F412, V417, S418, I438, S440, G441, K443, K465, I467, S468, N469, G471, N473	K375, I401, R403, R405, T406, K407, Q408, H409, G410, Q411, F412, D436
80R	H445, V458, P459, F460, S461, P462, D463, G464, K465, P466, C467, T468, P469, P470, A471, L472, N473, C474, Y475	R426, S432, Y436, K439, Y440, Y442, P469, P470, A471, L472, C474, Y475, W476, L478, N479, D480, G482, Y484, T485, T486, T487, G488,Y491, Q492	L443, R444, H445, I455, S456, N457, V458, P459, F460, S461, P462, D463, G464, K465, P466, C467, T468, P469, P470, A471, L472, N473, C474, Y475
b12	I108, C109, S110, L111, D113, Q114, S115, L116, K117, P118, C119, V120, P206, K207, V208, S209, F210, E211, P212, I213, P214, I251, R252, P253, I424, N425, M426, W427, C428, K429, V430	N280, A281, S365, G366, G367, D368, P369, I371, V372, T373, Y384, N386, P417, R419, V430, G431, K432, T455, R456, G472, G473, D474, M475	W95, T232, F233, N234, T236, S257, L260, N262, G263, S264, L265, A266, E267, E268, E269, V270, V271, T290, S291, S364, S365, G366, G367, D368, P369, E370, I371, V372, T373, T450, S481

3.4. Cautions in Using SAROTUP. SAROTUP must be used with caution since it is a tool only based on pattern matching at present. There are a lot of target-unrelated peptides bearing no known motifs [12]. As these TUPs are not embedded in SAROTUP at present, it is possible that a true TUP cannot be detected by SAROTUP. To reduce this kind of false negatives, we are constructing a database for target-unrelated peptides and mimotopes. Besides the motif-based search, the database-based search can find out the known TUP without known motifs.

It is also possible that a SAROTUP predicted target unrelated peptide is actually target-specific. To decrease this kind of false positives, the users should customize the scan according to their experiment at the section of advance options. For example, the user should select "antibody without Fc fragment" as the target if Fab was used in biopanning; this will prevent SAROTUP from reporting peptides bearing the Fc-binding motifs as TUP. As described above, SAROTUP in future will also provide an exact match tool based on database search. In this way, a match might mean that different research groups have isolated the same peptide with a variety of targets. It is obvious that this peptide can hardly be a true target binder. Thus, the false positive rate of SAROTUP can be decreased further when its new feature become available.

At last, we must point out that the controlled experiment is still the gold standard to distinguish TUPs from the specific mimotopes. The report of SAROTUP should be verified with experiment.

4. Conclusions

SAROTUP, a web application for scanning, reporting and excluding target-unrelated peptides has been coded with Perl. It helps researchers to predict epitope more accurately based on mimotopes. It is also useful in the development of diagnostics, therapeutics, and vaccines. To our knowledge, SAROTUP is the first web tool for TUP detecting and data cleaning. It is very convenient for the community to access SAROTUP through http://immunet.cn/sarotup/.

Acknowledgments

The authors are grateful to the anonymous reviewers for their valuable suggestions and comments, which have led to the improvement of this paper. This work was supported in part by the National Natural Science Foundation of China under the Grant 30600138 and the Scientific Research Foundation of UESTC for Youth under the Grant JX0769.

References

[1] G. P. Smith, "Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface," *Science*, vol. 228, no. 4705, pp. 1315–1317, 1985.

[2] J. K. Scott and G. P. Smith, "Searching for peptide ligands with an epitope library," *Science*, vol. 249, no. 4967, pp. 386–390, 1990.

[3] A. H. Y. Tong, B. Drees, G. Nardelli, et al., "A combined experimental and computational strategy to define protein interaction networks for peptide recognition modules," *Science*, vol. 295, no. 5553, pp. 321–324, 2002.

[4] G. Thom, A. C. Cockroft, A. G. Buchanan, et al., "Probing a protein-protein interaction by in vitro evolution," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 20, pp. 7619–7624, 2006.

[5] L. F. Wang and M. Yu, "Epitope identification and discovery using phage display libraries: applications in vaccine development and diagnostics," *Current Drug Targets*, vol. 5, no. 1, pp. 1–15, 2004.

[6] A. B. Riemer and E. Jensen-Jarolim, "Mimotope vaccines: epitope mimics induce anti-cancer antibodies," *Immunology Letters*, vol. 113, no. 1, pp. 1–5, 2007.

[7] V. A. Petrenko, "Evolution of phage display: from bioactive peptides to bioselective nanomaterials," *Expert Opinion on Drug Delivery*, vol. 5, no. 8, pp. 825–836, 2008.

[8] L. Zhao, Z. Liu, and D. Fan, "Overview of mimotopes and related strategies in tumor vaccine development," *Expert Review of Vaccines*, vol. 7, no. 10, pp. 1547–1555, 2008.

[9] H. Thie, T. Meyer, T. Schirrmann, M. Hust, and S. Dubel, "Phage display derived therapeutic antibodies," *Current Pharmaceutical Biotechnology*, vol. 9, no. 6, pp. 439–446, 2008.

[10] R. Knittelfelder, A. B. Riemer, and E. Jensen-Jarolim, "Mimotope vaccination—from allergy to cancer," *Expert Opinion on Biological Therapy*, vol. 9, no. 4, pp. 493–506, 2009.

[11] H. M. Geysen, S. J. Rodda, and T. J. Mason, "A priori delineation of a peptide which mimics a discontinuous antigenic determinant," *Molecular Immunology*, vol. 23, no. 7, pp. 709–715, 1986.

[12] A. Menendez and J. K. Scott, "The nature of target-unrelated peptides recovered in the screening of phage-displayed random peptide libraries with antibodies," *Analytical Biochemistry*, vol. 336, no. 2, pp. 145–157, 2005.

[13] L. A. Brammer, B. Bolduc, J. L. Kass, K. M. Felice, C. J. Noren, and M. F. Hall, "A target-unrelated peptide in an M13 phage display library traced to an advantageous mutation in the gene II ribosome-binding site," *Analytical Biochemistry*, vol. 373, no. 1, pp. 88–98, 2007.

[14] J. Huang, M. Xia, H. Lin, and F. Guo, "Information loss and noise inclusion risk in mimotope based epitope mapping," in *Proceedings of the 3rd International Conference on Bioinformatics and Biomedical Engineering (iCBBE '09)*, Beijing, China, 2009.

[15] D. Enshell-Seijffers, D. Denisov, B. Groisman, et al., "The mapping and reconstitution of a conformational discontinuous B-cell epitope of HIV-1," *Journal of Molecular Biology*, vol. 334, no. 1, pp. 87–101, 2003.

[16] B. M. Mumey, B. W. Bailey, B. Kirkpatrick, A. J. Jesaitis, T. Angel, and E. A. Dratz, "A new method for mapping discontinuous antibody epitopes to reveal structural features of proteins," *Journal of Computational Biology*, vol. 10, no. 3-4, pp. 555–567, 2003.

[17] I. Halperin, H. Wolfson, and R. Nussinov, "SiteLight: binding-site prediction using phage display libraries," *Protein Science*, vol. 12, no. 7, pp. 1344–1359, 2003.

[18] A. Schreiber, M. Humbert, A. Benz, and U. Dietrich, "3D-Epitope-Explorer (3DEX): localization of conformational epitopes within three-dimensional structures of proteins," *Journal of Computational Chemistry*, vol. 26, no. 9, pp. 879–887, 2005.

[19] V. Moreau, C. Granier, S. Villard, D. Laune, and F. Molina, "Discontinuous epitope prediction based on mimotope analysis," *Bioinformatics*, vol. 22, no. 9, pp. 1088–1095, 2006.

[20] J. Huang, A. Gutteridge, W. Honda, and M. Kanehisa, "MIMOX: a web tool for phage display based epitope mapping," *BMC Bioinformatics*, vol. 7, article 451, 2006.

[21] I. Mayrose, T. Shlomi, N. D. Rubinstein, et al., "Epitope mapping using combinatorial phage-display libraries: a graph-based algorithm," *Nucleic Acids Research*, vol. 35, no. 1, pp. 69–78, 2007.

[22] I. Mayrose, O. Penn, E. Erez, et al., "Pepitope: epitope mapping from affinity-selected peptides," *Bioinformatics*, vol. 23, no. 23, pp. 3244–3246, 2007.

[23] Y. X. Huang, Y. L. Bao, S. Y. Guo, Y. Wang, C. G. Zhou, and Y. X. Li, "Pep-3D-search: a method for B-cell epitope prediction based on mimotope analysis," *BMC Bioinformatics*, vol. 9, article 538, 2008.

[24] N. Hulo, A. Bairoch, V. Bulliard, L. Cerutti, et al., "The 20 years of PROSITE," *Nucleic Acids Research*, vol. 36, database issue, pp. D245–D249, 2008.

[25] E. M. Bublil, S. Yeger-Azuz, and J. M. Gershoni, "Computational prediction of the cross-reactive neutralizing epitope corresponding to the monoclonal antibody b12 specific for HIV-1 gp120," *FASEB Journal*, vol. 20, no. 11, pp. 1762–1774, 2006.

[26] N. Tarnovitski, L. J. Matthews, J. Sui, J. M. Gershoni, and W. A. Marasco, "Mapping a neutralizing epitope on the SARS coronavirus spike protein: computational prediction based on affinity-selected peptides," *Journal of Molecular Biology*, vol. 359, no. 1, pp. 190–201, 2006.

[27] J. Huang and W. Honda, "CED: a conformational epitope database," *BMC Immunology*, vol. 7, article 7, 2006.

[28] R. A. Laskowski, "PDBsum new things," *Nucleic Acids Research*, vol. 37, database issue, pp. D355–D359, 2009.

[29] R. W. Roberts and J. W. Szostak, "RNA-peptide fusions for the in vitro selection of peptides and proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 23, pp. 12297–12302, 1997.

[30] L. Gold, "mRNA display: diversity matters during in vitro selection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 9, pp. 4825–4826, 2001.

[31] D. Lipovsek and A. Pluckthun, "In-vitro protein evolution by ribosome display and mRNA display," *Journal of Immunological Methods*, vol. 290, no. 1-2, pp. 51–67, 2004.

[32] E. T. Boder and K. D. Wittrup, "Yeast surface display for screening combinatorial polypeptide libraries," *Nature Biotechnology*, vol. 15, no. 6, pp. 553–557, 1997.

[33] S. Stahl and M. Uhlen, "Bacterial surface display: trends and progress," *Trends in Biotechnology*, vol. 15, no. 5, pp. 185–192, 1997.

[34] G. Georgiern, C. Stathopoulos, P. S. Daugherty, A. R. Nayak, B. L. Iverson, and R. Curtiss III, "Display of heterologous proteins on the surface of microorganisms: from the screening of combinatorial libraries to live recombinant vaccines," *Nature Biotechnology*, vol. 15, no. 1, pp. 29–34, 1997.

[35] P. Samuelson, E. Gunneriusson, P. A. Nygren, and S. Stahl, "Display of proteins on bacteria," *Journal of Biotechnology*, vol. 96, no. 2, pp. 129–154, 2002.

[36] A. B. Riemer, H. Kurz, M. Klinger, O. Scheiner, C. C. Zielinski, and E. Jensen-Jarolim, "Vaccination with cetuximab mimotopes and biological properties of induced anti-epidermal growth factor receptor antibodies," *Journal of the National Cancer Institute*, vol. 97, no. 22, pp. 1663–1670, 2005.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 208371, 6 pages
doi:10.1155/2010/208371

Research Article

Assessment of the Genetic Diversity of *Mycobacterium tuberculosis esxA, esxH,* and *fbpB* Genes among Clinical Isolates and Its Implication for the Future Immunization by New Tuberculosis Subunit Vaccines Ag85B-ESAT-6 and Ag85B-TB10.4

Jose Davila,[1] Lixin Zhang,[1] Carl F. Marrs,[1] Riza Durmaz,[2] and Zhenhua Yang[1]

[1] *Department of Epidemiology, School of Public Health, University of Michigan, 1415 Washington Heights, Ann Arbor, MI 48109-2029, USA*
[2] *Department of Clinical Microbiology, Medical Faculty, Inonu University, 44280 Malatya, Turkey*

Correspondence should be addressed to Zhenhua Yang, zhenhua@umich.edu

Received 2 November 2009; Accepted 15 April 2010

Academic Editor: Anne S. De Groot

The effort to develop a tuberculosis (TB) vaccine more effective than the widely used Bacille Calmette-Guérin (BCG) has led to the development of two novel fusion protein subunit vaccines: Ag85B-ESAT-6 and Ag85B-TB10.4. Studies of these vaccines in animal models have revealed their ability to generate protective immune responses. Yet, previous work on TB fusion subunit vaccine candidate, Mtb72f, has suggested that genetic diversity among *M. tuberculosis* strains may compromise vaccine efficacy. In this study, we sequenced the *esxA, esxH,* and *fbpB* genes of *M. tuberculosis* encoding ESAT-6, TB10.4, and Ag85B proteins, respectively, in a sample of 88 clinical isolates representing 57 strains from Ark, USA, and 31 strains from Turkey, to assess the genetic diversity of the two vaccine candidates. We found no DNA polymorphism in *esxA* and *esxH* genes in the study sample and only one synonymous single nucleotide change (C to A) in *fbpB* gene among 39 (44.3%) of the 88 strains sequenced. These data suggest that it is unlikely that the efficacy of Ag85B-ESAT-6 and Ag85B-TB10.4 vaccines will be affected by the genetic diversity of *M. tuberculosis* population. Future studies should include a broader pool of *M. tuberculosis* strains to validate the current conclusion.

1. Introduction

The need for an improved vaccine against tuberculosis (TB) has never been more urgent. One in three people today are infected with *Mycobacterium tuberculosis*, the causative agent of TB, and worldwide approximately three million people die from TB annually. The currently available TB vaccine, Bacille Calmette-Guérin (BCG), has failed to consistently protect against the most contagious form of the disease, adult pulmonary TB, despite its widespread use [1–4]. Developing a new vaccine, which may serve as a booster or a replacement for BCG, is of critical importance in the fight against worldwide TB-related morbidity and mortality [4, 5].

Of the various vaccine candidates proposed, fusion subunit vaccines have received considerable attention in the recent literature, especially those composed of antigenic proteins ESAT-6, Ag85B, and TB10.4 [3–8]. It appears that the multiple epitopes that fusion subunit vaccines offer makes them more effective than single-peptide vaccines in interacting with the complexity of the host immune response against TB and the genetic restriction imposed by major histocompatibility complex molecules [3, 9]. Two fusion subunit vaccines, Ag85B-ESAT-6 and Ag85B-TB10.4, which are the focus of the present study, have been found to induce protective cell-mediated immunity in animal models [3, 6, 7, 9, 10]. Ag85B-ESAT-6 is currently in expanded Phase I studies in which the vaccine is tested in BCG-vaccinated, latently infected, and individuals from TB endemic regions [4]. As these candidates move forward in or toward clinical trials, it will be critically important to evaluate their protective potential as global vaccines via bioinformatic approaches built upon the comparative genomics of the pathogen population and the immunomics of the host population.

Bioinformatics approaches are invaluable to the development of effective vaccine candidates. Comparative genomics of the pathogen population, for instance, allows vaccine candidates that are potentially ineffective due to genetic diversity of the pathogen population to be discredited before they reach the costly stages of clinical trials. In other words, bioinformatic approaches can provide information based on which a rational selection of clinical trial sites can be made. As for subunit vaccines, comparative genomics can help analyze whether antigenic targets are conserved among infectious strains of an organism in order to ensure their protective efficacy across diverse pathogen populations circulating in different geographic region [8].

Although previous studies have suggested that *M. tuberculosis* has a relatively stable genome in comparison with other bacteria [11, 12], recent genomic studies have revealed biologically significant variation among clinical strains [13]. Hebert and colleagues, for instance, revealed considerable genetic variation in the PPE18 gene of *M. tuberculosis*, with important implications for the ability of the Mtb72f vaccine candidate to provoke protective immunity against diverse populations of *M. tuberculosis* [14]. Furthermore, the interaction of the genetic variation of the PPE18 component of Mtb72f with the allelic variation of human MHC-II DRB1 proteins negatively affects vaccine epitope binding to DRB1 proteins [15]. Taken in the context of vaccine development, revelations like these are crucial to the survival of vaccine candidates as potential clinical vaccines. A similar comparative genomics study on Ag85B-ESAT-6 and Ag85B-TB10.4 subunit vaccines may provide useful information for predicting the protective efficacy of these candidates in the pre- or early stages of their clinical evaluation.

Little information has been documented on the genetic variation of the genes encoding for ESAT-6, Ag85B, and TB10.4 proteins. If any of these three genes is highly variable, the protective efficacy of Ag85B-ESAT-6 and Ag85B-TB10.4 subunit vaccines might be compromised on the global stage. To further investigate the ability of these two-vaccine candidates in recognizing naturally occurring *M. tuberculosis* strains, we investigate the genetic diversity of the *esxA*, *esxH*, and *fbpB* genes of *M. tuberculosis* that encode for the components of the two new subunit vaccines in a sample of 88 *M. tuberculosis* strains collected from Turkey and Arkansas, USA.

2. Materials and Methods

2.1. M. tuberculosis Isolates. The clinical strains used in the present study are from Ark, USA, and Turkey. Following the work of Herbert and colleagues, the isolates were selected to represent different geographical regions, including Arkansas and Malatya, Turkey, to assess the impact of regional genetic variability on the two subunit vaccine candidates [14]. Each of the selected isolates represents a different strain of *M. tuberculosis* with a distinct IS*6110* restriction fragment polymorphism (RFLP) pattern with more than five bands or a distinct combination of a common IS*6110* RFLP pattern with five or less bands and a unique spoligo typing pattern (Table 1). The rationale for including isolates from two geographical regions was to discern the potential impact of genetic variation on future vaccination with Ag85B-ESAT-6 and Ag85B-TB10.4 in separate populations.

Our initial intent was to analyze the same set of clinical isolates ($n = 225$) used by Hebert and colleagues in their work on PPE18 and *pepA* [14]. However, after initial sequencing of a randomly selected subset ($n = 41$) of the 225 isolates revealed no genetic variation in the ESAT-6, TB10.4, and Ag85B genes, we selected only 47 of the remaining 84 isolates that had previously shown variation in the PPE18 gene ($n = 47$) for the current study. This decision was made with the consideration of cost-effective lab procedure, reasoning that local genetic variations would be indicative of broader genomic variability. The 88 study isolates represent 57 different strains from Ark, USA, and 31 distinct strains from Turkey. The 88 study isolates shared 49 different spoligotypes that represent 47 different spoligo international types, as determined by using the query tool of the fourth international spoligotyping database (SpolDB4) [16]. The present sample represents 80 of 84 strains (95.2%) that showed PPE18 variation in Hebert's study.

2.2. PCR of esxA, esxH, and fbpB Genes. The three genes under study were amplified using Invitrogen Platinum Taq PCRx Polymerase kit (Invitrogen, Carlsbad, CA) with primers published previously [17]. The primers used for *esxA*, encoding the ESAT-6 protein were esxA-F (5′-GCA-ATCCGGCGGCTCCACCAG-3′) located 533 bp upstream of the *esxA* gene and esxA-R (5′-TCGGCCGCCATGACA-ACCTCTC-3′) located 124 bp downstream from the end of the *esxA* gene. The primers used for *esxH*, encoding the TB10.4 protein were esxH-F (5′-GAGAGGGGGAGG-CGACGGCTTACC-3′) located 411 bp upstream of the esxH gene, and esxH-R (5′-TCCCCGCCCCAATGGTTTCAGC-3′) located 86 bp downstream from the end of the *esxH* gene. Finally, the primers used for *fbpB*, encoding Ag85B protein were fbpB-F2 (5′-ACTCGGCTAACTGGCTGGTGC-3′) located 217 bp upstream of the *fbpB* gene, and fbpB-R (5′-CATACCGCCATACCGTTTGTGAGC-3′) located 164 bp downstream from the end of the *fpbB* gene. The inclusion of the regions flanking the *esxA*, *esxH*, and *fbpB* genes allowed further confirmation that the PCR products were specific. The positive control in all the PCR reactions was *M. tuberculosis* H37Rv, and the negative control was PCR-grade water. The 50 μL PCR reaction mixture used was composed of 5 μL of 10 × PCRx amplification buffer, 1.5 μL of 50 mM $MgSO_4$, 1 μL of 10 mM deoxyribonucleoside triphosphate mixture, 20 pmol of each primer in 1 μL, 0.5 μL of Invitrogen Platinum Taq polymerase mixture, 4 μL of a DNA solution containing 50 ng of DNA template, and 40 μL of PCR-grade water. The thermocycling program used for all genes was one cycle at 94°C for 1 minute; 30 cycles of 94°C for 30 seconds, 62°C for 30 seconds, and 72°C for 2.5 minutes; and a final cycle of 72°C for 10 minutes. The sizes of the PCR products were verified by 1.0% (wt/vol) agarose gel electrophoresis in 1 × Tris-borate-EDTA buffer.

Table 1: Genotyping data of the 88 study isolates of *M. tuberculosis* that were originated from Arkansas, USA, and Turkey.

Isolate no.*	IS*6110* Copy no.	IS*6110* RFLP Pattern ¥	Spoligotype	Isolate no.*	IS*6110* Copy no.	IS*6110* RFLP Pattern ¥	Spoligotype
SA143	14	01046	000000000003771	SA208	03	00370	700036777760471
SA033	18	01063	000000000003771	SA738	07	05660	777777744720771
SA604	18	04495	000000000003771	SA204	10	00636	074377607760700
SA327	21	00237	000000000003771	SA565	09	01321	377737774020731
SA374	22	02623	000000000003771	SA710	17	02661	737377677760771
SA413	22	02660	000000000003771	SA221	01	00129	774377777413771
SA193	14	00638	677737607760771	SA253	14	00619	776377613760771
SA759	12	05774	677777477413771	SA052	03	00403	777776777560601
SA639	15	04804	703777740003771	SA333	04	01658	777776777560601
SA085	11	01336	776177607760771	SA597	07	04433	777776777760761
SA557	11	04222	776177607760771	SA286	08	01330	777777000020771
SA378	09	02620	777777770000000	SA198	03	00370	777777777613771
SA652	08	01646	777777774020771	SA173	19	00401	Not Available
SA053	11	00499	777777774020771	TK035	09	TK035	000000004020771
SA475	05	04430	777777777720771	TK036	08	TK036	000000007760771
SA010	09	00472	777777777720771	ZD073	15	ZD073	703777740003031
SA637	10	04802	777777777720771	TK013	03	TK013	777777404760771
SA213	13	00640	777777777720771	TK091	03	TK091	777777404760771
SA364	13	00807	777777777720771	TK110	08	TK110	777777740460771
SA025	11	00325	777777777760731	ZD039	08	ZD039	777777774020771
SA310	10	00315	776177607760771	ZD056	09	ZD056	777777774020771
SA285	06	01329	777777777760771	ZD061	08	ZD061	777777777720771
SA169	09	00326	777777777760771	ZD063	08	ZD063	777777777720771
SA120	12	00372	777777777760771	ZD034	09	ZD034	777777777720771
SA102	12	00538	777777777760771	TK086	10	TK086	777777777720771
SA471	14	06474	777777777760771	ZD013	10	ZD013	777777777720771
SA131	15	00627	777777777760771	ZD020	11	ZD020	777777777720771
SA157	06	00314	777777557760771	ZD071	11	ZD071	777777777720771
SA030	13	00570	777777607560771	TK065	05	TK065	777777777760771
SA706	10	05458	777777777760471	TK103	08	TK103	777777777760771
SA111	02	00016	777776777760771	TK095	09	TK095	777777777760771
SA352	03	00064	777776777760771	ZD062	09	ZD062	777777777760771
SA019	04	00432	777776777760771	TK030	10	TK030	777777777760771
SA170	06	00327	777776777760771	TK105	11	TK105	777777777760771
SA341	03	00694	777776777760601	ZD093	13	ZD093	757777777760771
SA020	05	00392	777776777760601	TK003	06	TK003	037637777760771
SA596	07	04432	777776777760601	ZD041	06	ZD041	777777764020731
SA526	01	00195	777777776413771	ZD031	09	ZD031	717777777760771
SA162	04	00319	700076777760700	ZD024	09	ZD024	770000770000000
SA778	08	06019	376177607760771	TK007	08	TK007	777767777760671
SA038	11	00496	777777777320771	TK016	07	TK016	037227760160771
SA512	12	00832	776377770760771	TK011	08	TK011	700377777760771
SA189	13	00598	777763777760771	TK112	09	TK112	777207777777661
SA430	01	00129	763777777413771	TK097	05	TK097	Not Available

*Isolate numbers starting with SA refer to strains from Arkansas, USA. Isolate numbers starting with ZD or TK refer to strains from Turkey.

¥ The IS*6110* RFLP patterns of the Arkansas strains were indicated using the national designations assigned by the National Tuberculosis Genotyping and Sentinel Surveillance Network Program of the Centers for Disease Control and Prevention. The IS*6110* RFLP patterns of the Turkish strains were designated using the isolate number in which a given IS*6110* RFLP pattern was observed for the first time.

2.3. Automated DNA Sequencing. PCR products were sequenced to identify any insertions/deletions or single nucleotide polymorphisms (SNPs) in the *esxA*, *esxH*, and *fbpB* genes in the selected isolates. PCR products were purified using Invitrogen PureLink PCR Purification Kit, according to manufacturer instructions (Invitrogen, Carlsbad, CA). DNA sequencing was performed with Applied Biosystems DNA sequencers 3700 and 3730 at the University of Michigan Sequencing Core, using the same primers that were used for the PCR of the three genes. The sequences of *esxA*, *esxH*, and *fbpB* of the study strains were compared to those of *M. tuberculosis* laboratory reference strain H37Rv (GenBank accession number BX842575) using the BLAST and BLAST2 nucleotide sequence alignment program of the National Center for Biotechnology Information (http://www.ncbi.nlm.nih.gov/BLAST) and the Sequencher 4.9 DNA Sequence Assembly software (Demo version) of Gene Codes Corporation (www.genecodes.com). Primary DNA transcripts were translated to amino acid sequences using the *in silico* simulation of molecular biology experiment of the University of Basque Country (http://insilico.ehu.es/).

3. Results and Discussion

3.1. Genetic Diversity of esxA, esxH, and fbpB and Corresponding Amino Acid Sequences. Among the 88 strains investigated, genetic analysis of *esxA* and *esxH* in this study revealed no nucleotide polymorphisms in the genes encoding for ESAT-6 and TB10.4 proteins. Of the 88 strains, 38 (43.2%) belong to principal genetic group 1, 29 (33.0%) belong to principal genetic group 2, and 21 (23.9%) belong to principal genetic group 3. The principle genetic groups were defined by SNPs in the *katG* and *gyrA* genes as described previously by Sreevatsan and colleagues [12]. Unlike the study by Herbert et al., where genetic group 1 strains were found to have the highest frequency of DNA polymorphisms in the PPE18 protein, a component of the Mtb72f vaccine [14], strains in all of the three principal genetic groups showed no DNA variations in both *esxA* and *esxH*. This observation suggests that these gene regions might be conserved among *M. tuberculosis* strains of different geographic origins and among different genetic groups of the pathogen. In fact, Gey Van Pittius and colleagues have recently posited interspecies conservation of the ESAT-6 gene region as part of a novel Gram-positive secretion system with distant homologues in *Bacillus subtilis, Bacillus anthracis, Staphylococcus aureus,* and *Clostridium acetobutylicum* [18].

The analysis of *fbpB,* the gene encoding for Ag85B revealed only one synonymous C to A SNP, located at position 714 bp of the gene sequence, among 39 (44.3%) of the 88 strains sequenced. Double strand sequencing was conducted on the 39 isolates to confirm the existence of this SNP. Although this SNP had no effect on the amino acid sequence of the peptide when translated, it is indicative of an allelic variation in the *M. tuberculosis* gene pool. Of the 39 strains, 15 (17.0%) belong to principal genetic group 1, 11 (12.5%) belong to principal genetic group 2, and 13 (14.7%) belong to principal genetic group 3. Furthermore, the SNP was not found to be associated with any specific geographic origin of the study strains, suggesting that this particular nucleotide polymorphism is of ancestral origin. These data suggest that *M. tuberculosis* Ag85B antigen is highly conserved in, at least, certain populations of *M. tuberculosis* clinical strains.

3.2. Implications for Immunization. TB remains one of the deadliest infectious diseases of our times. Despite widespread use of the BCG vaccine, the disease continues to claim 2-3 million lives per year. The need for a new vaccine has never been more urgent. In order to gain insight into the efficacy of two new vaccine candidates, two fusion proteins combining Ag85B and ESAT-6, and Ag85B and TB10.4 [6, 9, 10], respectively, the gene regions encoding for ESAT-6, TB10.4, and Ag85B proteins were analyzed for their variability. Experiments involving these two candidates have revealed them to be effective in generating protective immunity in animal models [6, 9, 10].

Yet, while animal models have played an important role in the development of new TB vaccines so far, they are not always representative of the internal human biological environment. As Flynn noted earlier, an important drawback of the murine model is that the pathology of pulmonary TB in mice is quite different from that in humans [19]. Specifically, the heterogeneity of granuloma types observed in the human host is not displayed in the mouse lung [19]. Similar difficulties arise when using other animal models. In the case of bovine infection, the pathology of TB is quite similar to human host response in granulomatous reactions, but differs with respect to cavitation [20]. Non-human primate models also represent the human pathology of TB quite well, but like cattle are limited by economic and infrastructural factors [20].

Furthermore, current preclinical studies of new TB vaccines' protection against *M. tuberculosis* infection in animal models do not take the population diversity of *M. tuberculosis* into consideration. However, as Hebert and colleagues noted previously [14], genetic diversity of *M. tuberculosis* genes can be found among clinical isolates, and such diversity may have important implications for the efficacy of the new vaccines. Thus, comparative genomics of the pathogen population stands as an additional useful tool for pre-clinical evaluation of new vaccines, providing information complementary to those from current *in vivo* and *in vitro* studies. Given the resource-demanding nature of clinical trials, comparative genomics serves as a method for predicting the potential protection of proposed vaccines candidates in the general population. Hebert and colleagues' work revealed that the PPE18 protein, part of the Mtb72f subunit vaccine, was quite variable among isolates collected from Turkey and Arkansas [14]. Analyzing the variability of antigens targeted by potential vaccines in a diverse set of isolates at the genomic level may indeed allow researchers to avoid developing a vaccine that is only variably effective like the current BCG. The findings of such study can also inform the rational selection of the study populations, with a consideration of

covering the diverse pathogen populations in clinical trials of new vaccines.

Our observation that ESAT-6, TB10.4, and Ag85B proteins were highly conserved in our study sample comprising strains from two geographically distant regions and three different principal genetic groups suggest that it is unlikely that the efficacy of Ag85B-ESAT-6 and Ag85B-TB10.4 subunit vaccines will be affected by the genetic diversity of *M. tuberculosis* population. Thus, the protective efficacy of these two novel vaccine candidates may have a wider reach than Mtb72b vaccine, which contains a highly variable antigen of *M. tuberculosis* [14]. However, our findings also indicate the need for further bioinformatics research on the three genes investigated and their specific interaction with the host immune system. While highly conserved genes are indicative of homologous protein antigens, they may also suggest a lack of selective pressure by the host immune system and thus a lack of recognition on behalf of the host's immune response. Previous examples of the inability of animal models to accurately represent the human host environment suggest that the degree to which the human host system interacts with these important peptides remains to be studied.

The goal of pre-clinical evaluation of these two vaccines may be furthered by future studies that include a larger sample of isolates from a greater range of geographic origins, making the data even more representative of the diversity of *M. tuberculosis* worldwide. The finding of this study that *esxA*, *esxH*, and *fbpB* were conserved across 88 clinical strains from Arkansas and Turkey does not confirm that they are conserved globally. Including a larger and more genetically diverse sample of isolates would address the two primary limitations of this study—the number and diversity of the isolates used.

Another factor that must be taken into account is the potential impact that host diversity may have on the global coverage of these two new TB vaccines. This study looked at the diversity of pathogen genes coding for vaccine proteins, but even uniformly conserved proteins may fail to induce protective immunity if host diversity impedes their ability to effectively bind effector immune cells. McNamara and colleagues provide an eloquent, bioinformatic approach to studying the impact that host diversity may have on Mtb27f vaccine coverage by analyzing the allelic variation of human class II MHC DRB1 proteins and its impact on proper vaccine epitope binding [15]. A similar study on Ag85B-ESAT-6 and Ag85B-TB10.4 may provide insightful information on host diversity and its impact on the coverage of these vaccines.

With these future directions in mind, the results of the present study represent an important first step in the pre-clinical bioinformatic assessment of Ag85B-ESAT-6 and Ag85B-TB10.4 vaccine candidates, and the impact that genetic diversity among their respective antigenic protein targets has on their potential success as global vaccines. The finding that *esxA, esxB,* and *fbpB* genes are highly conserved in two distinct populations suggests that Ag85B-ESAT-6 and Ag85B-TB10.4 vaccine candidates may be effective in geographically distinct areas of the world.

Acknowledgments

This study was supported by Grant NIH-R01-AI151975 from the National Institutes of Health and the Research Fund of the Office of the Vice President for Research of the University of Michigan. The Arkansas isolates and the genotyping data of the study isolated used in this study were kindly provided by Dr. Joseph H. Bates at the Arkansas Department of Health and Dr. Donald M. Cave at the University of Arkansas for Medical Sciences.

References

[1] B. R. Bloom and P. E. M. Fine, Eds., *The BCG Experience: Implications for Vaccines against Tuberculosis*, ASM Press, Washington, DC, USA, 1994.

[2] P. Andersen, "Tuberculosis vaccines—an update," *Nature Reviews Microbiology*, vol. 5, no. 7, pp. 484–487, 2007.

[3] L. Brandt, M. Elhay, I. Rosenkrands, E. B. Lindblad, and P. Andersen, "ESAT-6 subunit vaccination against Mycobacterium tuberculosis," *Infection and Immunity*, vol. 68, no. 2, pp. 791–795, 2000.

[4] C. Aagaard, J. Dietrich, M. Doherty, and P. Andersen, "TB vaccines: current status and future perspectives," *Immunology and Cell Biology*, vol. 87, no. 4, pp. 279–286, 2009.

[5] D. F. Hoft, "Tuberculosis vaccine development: goals, immunological design, and evaluation," *The Lancet*, vol. 372, no. 9633, pp. 164–175, 2008.

[6] J. Dietrich, C. Aagaard, R. Leah, et al., "Exchanging ESAT6 with TB10.4 in an Ag85B fusion molecule-based tuberculosis subunit vaccine: efficient protection and ESAT6-based sensitive monitoring of vaccine efficacy," *Journal of Immunology*, vol. 174, no. 10, pp. 6332–6339, 2005.

[7] A. S. Mustafa, Y. A. Skeiky, R. Al-Attiyah, M. R. Alderson, R. G. Hewinson, and H. M. Vordermeier, "Immunogenicity of Mycobacterium tuberculosis antigens in Mycobacterium bovis BCG-vaccinated and M. bovis-infected cattle," *Infection and Immunity*, vol. 74, no. 8, pp. 4566–4572, 2006.

[8] A. L. Sorensen, S. Nagai, G. Houen, P. Andersen, and A. B. Andersen, "Purification and characterization of a low-molecular-mass T-cell antigen secreted by Mycobacterium tuberculosis," *Infection and Immunity*, vol. 63, no. 5, pp. 1710–1717, 1995.

[9] A. W. Olsen, A. Williams, L. M. Okkels, G. Hatch, and P. Andersen, "Protective effect of a tuberculosis subunit vaccine based on a fusion of antigen 85B and ESAT-6 in the aerosol guinea pig model," *Infection and Immunity*, vol. 72, no. 10, pp. 6148–6150, 2004.

[10] J. A. M. Langermans, T. M. Doherty, R. A. W. Vervenne, et al., "Protection of macaques against Mycobacterium tuberculosis infection by a subunit vaccine based on a fusion protein of antigen 85B and ESAT-6," *Vaccine*, vol. 23, no. 21, pp. 2740–2750, 2005.

[11] M. Marmiesse, P. Brodin, C. Buchrieser, et al., "Macro-array and bioinformatic analyses reveal mycobacterial 'core' genes, variation in the ESAT-6 gene family and new phylogenetic markers for the Mycobacterium tuberculosis complex," *Microbiology*, vol. 150, no. 2, pp. 483–496, 2004.

[12] S. Sreevatsan, X. Pan, K. E. Stockbauer, et al., "Restricted structural gene polymorphism in the Mycobacterium tuberculosis complex indicates evolutionarily recent global dissemination," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 18, pp. 9869–9874, 1997.

[13] I. Filliol, A. S. Motiwala, M. Cavatore, et al., "Global phylogeny of Mycobacterium tuberculosis based on single nucleotide polymorphism (SNP) analysis: insights into tuberculosis evolution, phylogenetic accuracy of other DNA fingerprinting systems, and recommendations for a minimal standard SNP set," *Journal of Bacteriology*, vol. 188, no. 2, pp. 759–772, 2006.

[14] A. M. Hebert, S. Talarico, D. Yang, et al., "DNA polymorphisms in the pepA and PPE18 genes among clinical strains of Mycobacterium tuberculosis: implications for vaccine efficacy," *Infection and Immunity*, vol. 75, no. 12, pp. 5798–5805, 2007.

[15] L. McNamara, Y. He, and Z. Yang, "Using epitope predictions to evaluate efficacy and population coverage of the Mtb72f vaccine for tuberculosis," *BMC Immunology*, vol. 11, no. 1, article 18, 2010.

[16] K. Brudey, J. R. Driscoll, L. Rigouts, et al., "Mycobacterium tuberculosis complex genetic diversity: mining the fourth international spoligotyping database (SpolDB4) for classification, population genetics and epidemiology," *BMC Microbiology*, vol. 6, article 23, 2006.

[17] R. Hershberg, M. Lipatov, P. M. Small, et al., "High functional diversity in Mycobacterium tuberculosis driven by genetic drift and human demography," *PLoS Biology*, vol. 6, no. 12, article e311, pp. 2658–2671, 2008.

[18] N. C. Gey Van Pittius, J. Gamieldien, W. Hide, G. D. Brown, R. J. Siezen, and A. D. Beyers, "The ESAT-6 gene cluster of Mycobacterium tuberculosis and other high G+C Gram-positive bacteria," *Genome Biology*, vol. 2, no. 10, p. RESEARCH0044, 2001.

[19] J. L. Flynn, "Lessons from experimental Mycobacterium tuberculosis infections," *Microbes and Infection*, vol. 8, no. 4, pp. 1179–1188, 2006.

[20] A. S. Dharmadhikari and E. A. Nardell, "What animal models teach humans about tuberculosis," *American Journal of Respiratory Cell and Molecular Biology*, vol. 39, no. 5, pp. 503–508, 2008.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 910524, 14 pages
doi:10.1155/2010/910524

Review Article

Benchmarking B-Cell Epitope Prediction for the Design of Peptide-Based Vaccines: Problems and Prospects

Salvador Eugenio C. Caoili

Department of Biochemistry and Molecular Biology, College of Medicine, University of the Philippines Manila, 547 Pedro Gil Street, Ermita, Manila 1000, Philippines

Correspondence should be addressed to Salvador Eugenio C. Caoili, badong@post.upm.edu.ph

Received 18 September 2009; Revised 12 December 2009; Accepted 18 February 2010

Academic Editor: Yongqun Oliver He

To better support the design of peptide-based vaccines, refinement of methods to predict B-cell epitopes necessitates meaningful benchmarking against empirical data on the cross-reactivity of polyclonal antipeptide antibodies with proteins, such that the positive data reflect functionally relevant cross-reactivity (which is consistent with antibody-mediated change in protein function) and the negative data reflect genuine absence of cross-reactivity (rather than apparent absence of cross-reactivity due to artifactual masking of B-cell epitopes in immunoassays). These data are heterogeneous in view of multiple factors that complicate B-cell epitope prediction, notably physicochemical factors that define key structural differences between immunizing peptides and their cognate proteins (e.g., unmatched electrical charges along the peptide-protein sequence alignments). If the data are partitioned with respect to these factors, iterative parallel benchmarking against the resulting subsets of data provides a basis for systematically identifying and addressing the limitations of methods for B-cell epitope prediction as applied to vaccine design.

1. Introduction

The timely development of new vaccines is imperative to address the complex and rapidly evolving global burden of disease [1–7]. Vaccines typically induce protective immunity by eliciting antibodies that neutralize the biological activity of proteins (e.g., bacterial exotoxins) [6]. These proteins comprise B-cell epitopes, that is, molecular substructures whose defining feature is their capacity for binding by antibodies. In turn, each B-cell epitope comprises spatially proximate amino acid residues or atoms thereof [8]; but its physical boundaries cannot be precisely delineated due to the limited specificity of molecular recognition by antibodies [9].

A peptide may induce antipeptide antibodies that cross-react with a cognate protein; if the antibodies neutralize the biological activity of the protein and thereby confer protective immunity, the peptide is a candidate vaccine component [6]. Such peptides are routinely designed to contain B-cell epitopes that have been predicted (i.e., presumptively identified) through computational analysis of cognate protein sequence or higher-order structure [3, 10]. For this application, the refinement of methods to predict B-cell epitopes necessitates benchmarking against empirical data [8].

Empirical data for benchmarking B-cell epitope prediction are customarily organized into individual records, each of which contains three key components, namely structural data on an immunogen, structural data on an antigen, and data on the outcome of an antibody-antigen binding assay [11–13]; the immunogen (e.g., peptide or protein conjugate thereof) induces antibodies while the antigen (e.g., cognate protein or biological source thereof) is used in the assay to determine the binding capacity of the antibodies. In many cases, the only structural data available are the sequences of both the immunogen and antigen while the outcome of the assay is expressed as either positive or negative binding even when the original outcome variable (e.g., inhibition of biological activity) is continuous rather than dichotomous. For a single record containing these minimal data, the task actually benchmarked is the exhaustive identification of putative epitopes as sequences that are predicted to both induce antibodies as part of the immunogen and act as

targets for binding by the antibodies as part of the antigen. If the immunogen is found to contain at least one such putative epitope, positive binding is the predicted outcome of the assay; otherwise, negative binding is the predicted outcome of the assay.

In the discussion of approaches to benchmark B-cell epitope prediction, a major source of confusion is the superficial parallelism between cross-reaction of antipeptide antibodies with proteins and cross-reaction of antiprotein antibodies with peptides. B-cell epitope prediction for both types of cross-reaction may be benchmarked against data in records of the same format, with the core of each record containing data on an immunogen, an antigen and the outcome of an antibody-antigen binding assay; but the roles of peptide and cognate protein are reversed for the latter type of cross-reaction, wherein cognate protein serves as immunogen while peptide serves as antigen for the binding assay. The physicochemical ramifications of this difference [14] imply that cross-reaction of antiprotein antibodies with peptides is mechanistically irrelevant to peptide vaccination and, by extension, that data on this type of cross-reaction are inappropriate for benchmarking B-cell epitope prediction where the intended application is the design of peptide-based vaccines [15].

Against an unprecedentedly large set of empirical data from high-throughput peptide-scanning experiments, benchmarking has revealed apparent underperformance of methods for B-cell epitope prediction that are based solely on sequence [12]. This outcome has long been anticipated from the gross oversimplification of modeling proteins as if they were unidimensional entities [16]. However, the data used for the analysis are irrelevant to peptide vaccination because they pertain exclusively to cross-reaction of antiprotein antibodies with peptides [15]; furthermore, the analysis itself neglects the multiplicity of factors that complicate B-cell epitope prediction, which merit closer scrutiny considering the pitfalls of reductionism in vaccine design [17–20]. In light of the fact that conclusions drawn from benchmarking are highly dataset-dependent [21], the present work explores the ensuing problems and suggests how to avoid them through judicious selection and partitioning of empirical data.

2. Conceptual Basis

B-cell epitope prediction can be employed to arrive at a computational result on the capacity of antipeptide antibodies to cross-react with a protein, but a definitive empirical result is established by observing for evidence of actual cross-reaction in a real system [8]. The essence of benchmarking is appraisal of the computational result against the empirical result: If these two results are in agreement, the computational result is deemed true; otherwise, it is deemed false. By convention, each result is either positive if it affirms cross-reaction or negative if it negates cross-reaction. Hence, the computational result falls into one of four mutually exclusive categories, namely true-positive, true-negative, false-positive and false-negative (hereafter denoted by TP, TN, FP and FN, resp.) [22, 23].

Benchmarking entails computation of sensitivity as TP/(TP + FN) and specificity as TN/(TN + FP), where each algebraic symbol represents the number of computational results falling within the denoted category [22, 23]. Sensitivity and specificity both range from zero to one, and would both be equal to one for a perfect predictive method; but in practice, either may be greater than zero only if the other is less than one. The accuracy of B-cell epitope prediction is increased by simultaneously increasing both sensitivity and specificity, or by increasing either without decreasing the other.

The mathematical definitions of sensitivity and specificity clarify the inherent problems of attempting to benchmark B-cell epitope prediction against irrelevant empirical results; insofar as the design of peptide-based vaccines is concerned, the most obvious of these results are from experiments that do not even simulate vaccination with peptides (e.g., where antibodies are never elicited by peptides in the first place). The designation of such results as either positive or negative is meaningless; at worst, it leads to erroneous appraisal of computational results that translates to miscalculation of both sensitivity and specificity. Consequently, methods for B-cell epitope prediction can be either underrated or overrated, thereby compromising efforts to assess their performance and address their limitations accordingly.

3. Selection of Data for Benchmarking

B-cell epitope prediction has been largely benchmarked against data acquired by probing antiprotein antibodies for cross-reactivity with peptides [11, 12]; yet for the design of vaccine peptides, reliable data are acquired only by probing antipeptide antibodies for cross-reactivity with proteins. Cross-reaction of antipeptide antibodies with proteins is distinct from cross-reaction of antiprotein antibodies with peptides [24], for which reason antibodies elicited by a protein do not necessarily cross-react with a peptide even if antibodies elicited by the peptide cross-react with the protein [25, 26]; because of this phenomenological asymmetry, benchmarking against data on antiprotein antibodies inevitably leads to misclassification of computational results that apply to antipeptide antibodies, with true-positives misclassified as false-positives (risking calculation of erroneously low values for both sensitivity and specificity) and false-negatives misclassified as true-negatives (risking calculation of erroneously high values for both sensitivity and specificity). Methods for B-cell epitope prediction may thus be either underrated or overrated for peptide-based vaccine design if they are benchmarked against data on antiprotein antibodies. Similar problems can arise with data acquired using monoclonal antibodies: If a monoclonal antipeptide antibody fails to cross-react with a protein, this result by itself may not be representative of a polyclonal antibody response in vivo [27]. Such sampling errors are avoided by using a sufficiently large panel of different monoclonal antibodies [27], which is virtually equivalent to polyclonal antibody. Therefore, the fundamental criterion for selecting empirical data is their generation through

experiments that are adequate to detect cross-reactions of polyclonal antipeptide antibodies with proteins; but other criteria must also be invoked to avoid the problems of functionally irrelevant cross-reactivity (in relation to positive data) and apparent absence of cross-reactivity (in relation to negative data).

Functionally relevant cross-reactivity is cross-reaction of antipeptide antibodies with proteins that alters protein function (e.g., inhibiting enzymes). This is implicit in cross-protective immunogenicity, that is, the capacity of vaccine peptides to induce antipeptide antibodies that confer protective immunity by cross-reacting with proteins [9, 10]. In contradistinction, functionally irrelevant cross-reactivity is cross-reaction of antipeptide antibodies with proteins that does not alter protein function. This is conceptually analogous to apparent cross-reaction, which stems from the notion that cross-reaction of antipeptide antibodies with proteins is either genuine if the proteins are in native form or merely apparent if the proteins are denatured [28–30]; but the focus on functional correlates irrespective of underlying protein conformations obviates the classical dichotomy between native and denatured proteins, which is increasingly difficult to reconcile with the emerging paradigm of structural and functional versatility among proteins [31–34] that encompasses intrinsic protein disorder [35–39] as well as coupled protein folding and binding [40–43]. Positive data that reflect only functionally irrelevant cross-reactivity compromise benchmarking if they lead to misclassification of true-negatives as false-negatives and false-positives as true-positives. This problem is avoided if all the positive data used for benchmarking have been validated by assays that detect antibody-mediated change in protein function (e.g., enzyme inhibition by antibodies), as opposed to immunoassays that merely detect antibody-protein binding without regard to protein function (e.g., binding of an enzyme by antibodies without regard to its catalytic activity) [15, 44].

Whereas the interpretation of positive data is confounded by functionally irrelevant cross-reactivity, the interpretation of negative data is confounded by apparent absence of cross-reactivity, that is, failure to detect cross-reactions of antipeptide antibodies with proteins that is due to artifactual masking of B-cell epitopes rather than genuine lack of capacity for binding. Apparent absence of cross-reactivity occurs in the setting of solid-phase immunoassays for which proteins are immobilized onto solid surfaces (e.g., by passive adsorption) in a way that renders B-cell epitopes physically inaccessible to paratopes. This problem accounts for conflicting results observed among different solid-phase immunoassays [45, 46] and between solid-phase and fluid-phase immunoassays [46, 47]; it is avoided if all the negative data used for benchmarking have been validated by fluid-phase immunoassays (e.g., immunoprecipitation) for which proteins are dispersed in the fluid phase prior to their encounter with antibodies [46], provided that artifactual masking of B-cell epitopes has not occurred by alternative mechanisms (e.g., through interactions with blocking reagents that might have been used to attenuate nonspecific antibody binding).

Taken together, the preceding considerations suggest the following approach to the selection of data for benchmarking: Admit only those data that pertain to cross-reactivity of polyclonal antipeptide antibodies with proteins, choosing the positive data that reflect antibody-mediated change in protein function and the negative data that are from fluid-phase immunoassays.

Data thus selected are suitable for benchmarking B-cell epitope prediction in support of applications for which peptide-based immunogens are designed to induce antipeptide antibodies that cross-react with native proteins; these applications include active and passive immunization for prophylaxis against and treatment of disease, and also the production of antipeptide antibodies as immunoaffinity reagents for protein purification (e.g., in the preparation of functional recombinant gene products). The same data must be used with caution for benchmarking B-cell epitope prediction where antipeptide antibodies are produced as diagnostic probes to detect proteins (e.g., of pathogens in clinical specimens). For this purpose, use of the data is appropriate only if the proteins are in native form when encountered by the antibodies; otherwise, both sensitivity and specificity are likely to be inaccurately estimated due to protein denaturation, as may occur during the collection, storage and processing of biological samples. This problem arises in relation to negative data on antipeptide antibodies that fail to bind native proteins yet bind denatured forms of the same proteins [28]; if the envisioned diagnostic procedure relies even partly on the detection of denatured protein, benchmarking against these negative data risks misclassification of positive predictions as false rather than true and of negative predictions as true rather than false.

As for data that pertain to cross-reactivity of antiprotein antibodies with peptides, these data are suitable for benchmarking where the intended application is the design of peptide-based diagnostic probes to detect antiprotein antibodies (e.g., for serodiagnosis of infectious diseases). For this application, published B-cell epitope prediction algorithms are found wanting as they perform only marginally better than random [12, 48, 49], although those based on three-dimensional structure outperform those based on sequence alone [48, 49]. As a B-cell epitope contains residues that must simultaneously interact with a paratope for binding to occur, the superior performance of the structure-based algorithms is plausibly realized through bypassing inaccurate sequence-based prediction of both spatial proximity among protein residues and their accesibility to antibodies [50, 51]; yet the hitherto unrealized success of these algorithms is at least partly due to misplaced emphasis on structure. Application of the structure-based algorithms is focused on an assumed protein structure supposedly encountered by antiprotein antibodies in the course of immunization; if this structure (e.g., from protein crystallography) differs from the actual immunogenic structure (e.g., of denatured protein in vivo), inaccurate predictions may be inevitable [10]. Published B-cell epitope prediction algorithms are therefore incomplete in that they fail to explicitly model the immunogenic structure of proteins in vivo. Attempts to address this deficiency by structural modeling are presently impossible

to validate given the lack of experimental data on protein structure in vivo; while structures have been elucidated for complexes consisting of proteins already bound by antibodies or fragments thereof, the actual structure of the proteins in vivo as they are encountered by antibodies is unknown [10]. Granting that this impasse is somehow resolved as new structural data become available, the problem of underperformance could still persist due to an exclusive focus on protein structure prior to binding by antibodies considering that their binding may bring even solvent-inaccessible antigen residues into contact with paratope residues [19, 52]; if so, B-cell epitope prediction might be substantially improved only through detailed computational modeling of antibody-antigen binding itself (e.g., by means of molecular docking analyses that allow for mutually induced fit between antibody and antigen, for antiprotein antibodies both reacting with proteins and cross-reacting with peptides). By the same token, a similarly elaborate computational approach might be required for peptide-based vaccine design in order to accurately predict B-cell epitopes for antipeptide antibodies both reacting with peptides and cross-reacting with proteins.

At any rate, valid claims regarding the practical utility of a method for B-cell epitope prediction are based on benchmarking against data that correspond well to the intended application; where this application is peptide-based vaccine design, the results of benchmarking are open to question if any of the data pertain to antiprotein rather than antipeptide antibodies or can be explained by either functionally irrelevant cross-reactivity or apparent absence of cross-reactivity.

4. Partitioning of Selected Data

Data selected as described above are heterogeneous in view of multiple factors that complicate B-cell epitope prediction (e.g., factors that define key structural differences between immunizing peptides and their cognate proteins). If the data are partitioned with respect to these factors, parallel benchmarking against the resulting subsets of data opens the possibility of discovering context-dependent variations in predictive performance. Knowledge of such variations could aid in identifying and appropriately addressing the limitations of methods to predict B-cell epitopes, thereby improving efficiency in both utilization and refinement of these methods.

From a purely physicochemical perspective, the data can be partitioned with respect to factors that are correlated with structural similarity between immunizing peptides and their cognate proteins. Though antibodies to a peptide may cross-react with a protein of apparently unrelated sequence, the likelihood of cross-reaction diminishes with decreasing sequence similarity [53]; and even where sequences are exactly matched, cross-reaction fails to occur if conformations are sufficiently dissimilar [54]. Cross-reaction thus tends to be disfavored by structural differences between peptides and proteins at the levels of both sequence and conformation.

Peptide and protein sequences are conventionally represented as strings of symbols for the twenty canonical proteinogenic amino acids, such that a difference in sequence is inferred from mismatched symbols in a pairwise sequence alignment; but this approach overlooks more subtle differences that are nonetheless relevant to the cross-reaction of antipeptide antibodies with proteins. A case in point is the mismatching of backbone charges between peptides and their cognate proteins: even if a peptide and a protein segment appear to share exactly the same sequence, they can differ from one another in terms of electrical charge at their N- or C-terminal ends. Ordinarily, both ends of the peptide backbone are charged, as when the peptide is derived from the cognate protein by hydrolysis of peptide bonds; in contrast, the corresponding ends of the protein segment lack backbone charges if they are at internal sequence positions of the protein. If a charged end of the peptide backbone corresponds to an internal sequence position of the protein, the charge on that end is unmatched in the protein. Unmatched charges of this nature can be eliminated by chemically blocking the ends of the peptide backbone (e.g., by acetylating the N-terminal amino group and amidating the C-terminal carboxyl group); but such blocking can itself result in an unmatched charge if a blocked end of the peptide backbone corresponds to an unblocked end of the protein backbone. An unmatched charge can also result from variability in the protonation state of the histidine sidechain at physiologic pH; antibodies elicited by peptides bearing this sidechain often bind the peptides only if it is unprotonated [55], yet it is frequently protonated as part of a folded protein [56–58]. As the placement of charges on epitopes is critical for binding by antibodies [55, 59], the backbone and sidechain charge mismatches just described may preclude cross-reaction [60, 61].

Apart from charge mismatches, conventional sequence analysis also overlooks structural differences related to the propensity of cysteine for oxidative cross-linkage via disulfide bond formation to yield cystine. As cysteine and cystine are chemically nonequivalent, a paratope optimized for binding a peptide that contains cystine in place of cysteine might fail to cross-react with a cognate protein that contains cysteine in place of cystine. Moreover, loops of residues are conformationally constrained by formation of intramolecular disulfide bonds between cysteines, with loops of identical sequence possibly adopting different conformations; although cyclization (i.e., formation of a covalently closed loop) thermodynamically favors binding by antibody through a decrease in conformational entropy [62], a paratope optimized for binding a loop in a particular conformation may fail to cross-react with a loop of identical sequence in a different conformation [9].

While conformational differences between a peptide and its cognate protein tend to disfavor cross-reaction, elimination of such differences by itself cannot ensure cross-reaction; even if the peptide closely resembles a segment of the protein at the levels of both sequence and conformation, the structural placement of the segment in the whole protein poses steric barriers to binding by the antipeptide antibodies if prospective interaction surfaces on the protein

are buried or located within concavities that are inaccessible to paratopes [63]. As these barriers are overcome through structural adjustment to realize induced fit between antigen and antibody, cross-reaction may actually be favored more by conformational flexibility (i.e., dynamic disorder) of both peptide and cognate protein rather than their similarity in the sense of rigid conformation (e.g., expressed as root-mean-square atomic displacements between superposed crystallographic structures) [9, 52].

The variety of structural differences between immunizing peptides and their cognate proteins is increased by both chemical treatment of peptides and posttranslational modification of proteins. Chemical treatment of peptides is widely applied to elicit strong antipeptide antibody responses [64]. This commonly involves conjugation of peptides to carrier proteins using covalent linker reagents such as glutaraldehyde [65]. The aldehyde groups of glutaraldehyde react with amino groups of lysine sidechains and N-terminal residues of immunizing peptides, potentially creating charge mismatches between the peptides (whose amino groups are covalently modified to uncharged derivatives) and their cognate proteins (whose amino groups are positively charged by virtue of protonation). A greater number of charge mismatches may be produced by carbodiimide reagents, such as 1-ethyl-3-(3-dimethylaminopropyl)-carbodiimide, which covalently modify both amino and carboxyl groups to uncharged derivatives [66]. More generally, the chemical moieties invariably formed through conjugation can themselves induce antibodies that dominate the immune response but fail to cross-react with protein [65]. Posttranslational modification of proteins (e.g., glycosylation) poses essentially the same problem as chemical treatment of peptides: Both processes can give rise to structural differences between immunizing peptides and their cognate proteins, thereby disfavoring cross-reaction [67].

Beyond structural differences between peptides and their cognate proteins at the level of single molecules, cross-reaction may be disfavored by protein localization in supramolecular assemblies such as biological membranes. Antipeptide antibodies that cross-react with a protein in solution may fail to cross-react with the same protein when it is physically associated with a membrane [68]. Possible mechanisms for such context-dependent cross-reactivity are steric shielding (e.g., along transmembrane protein segments) and even membrane-induced conformational transitions that themselves may be contingent upon membrane composition (e.g., lipid content) [69]. As an intact membrane is itself a mechanical barrier to the passage of macromolecules, proteins and segments thereof sequestered within a membrane-bound compartment are inaccessible to antibodies from outside the compartment unless the membrane has been adequately permeabilized (e.g., with organic solvent or surfactant) [70–78].

A corollary to these complexities of antibody-antigen interactions in biological systems is that various features of functionally relevant cross-reactivity are themselves factors for data partitioning. Functionally relevant cross-reactivity has already been introduced herein using the example of enzyme inhibition through binding by cross-reactive antipeptide antibodies. The enzyme could be a soluble single-domain protein for which B-cell epitope prediction would be much more straightforward than if the enzyme were membrane-bound and comprised multiple protein subunits that each contained multiple domains; in the latter case, separate polypeptide chains might even contribute residues to form a single neotope, that is, a B-cell epitope whose existence depends on the integrity of protein quaternary structure. B-cell epitope prediction is far more complicated for the neutralization of viruses and other infectious agents, in which case functionally relevant cross-reactivity may reflect emergent properties of the host-pathogen system that are left unaccounted for by analyses of the isolated pathogen or structural components thereof [79]. Synergy among pathogen virulence factors suggests the potential benefit of including more than one antigen in a vaccine; but it is difficult to predict which antigen combinations might afford protection against disease, as the interactions between immune responses to different antigens range from synergistic to antagonistic [80]. Much of this problem is due to uncertainty regarding the biological consequences of antibody-antigen binding, which tend to be highly context-dependent in ways that defy the intuitive notion of neutralization capacity as a monotonically increasing function of both antibody concentration and affinity for antigen [81]. For instance, occupancy of pathogen surfaces by antibodies often blocks host-pathogen surface interactions critical for infection while facilitating immune clearance of pathogens, yet such occupancy may itself lead to enhancement of infection (e.g., by mediating viral entry into host cells expressing receptors for the Fc portions of antibodies [82]). Divergent biological effects may also result from the binding of antigen by antibodies with similar affinities but via different molecular mechanisms (e.g., irreversibly inactivating virus only when binding induces certain conformational changes in capsid proteins [83]). The unifying theme of such phenomena is the coevolution of pathogens and the host responses against them that generates diverse antibodies to impair pathogen survival processes (e.g., disrupting various stages of viral replication cycles [84]) while it simultaneously decreases pathogen vulnerabilities to these antibodies (e.g., by transient assembly of viral surface neotopes only at the time that they are required to mediate entry into a host cell, thereby minimizing their exposure to neutralizing antibodies [85]). Taking soluble single-domain proteins from non-pathogen sources as a reference class of antigens for which B-cell epitope prediction is presumably least difficult, each additional complicating feature (e.g., quaternary structure, membrane association, origin from a pathogen) represents a factor for data partitioning.

From a broader biological perspective, numerous other factors merit investigation as well (e.g., genetic background, environmental influences, physiological status, mode of immunization). Carried to the extreme, data selected for benchmarking might be so extensively partitioned that each empirical result is placed in a class of its own; while it would not be of immediate practical value for benchmarking due to the consequent sparseness of the data, such partitioning

would properly emphasize the fact that each empirical result is the product of a unique combination of circumstances.

The most compelling experimental observations in support of data partitioning as outlined above are discordant results indicating that cross-reaction of antipeptide antibodies with proteins is critically dependent on details of immunogen chemistry other than the sequences shared by immunizing peptides and their cognate proteins (insofar as these sequences are conventionally represented in terms of the twenty canonical proteinogenic amino acids). This is exemplified by discordant results obtained for the model hexapeptide of sequence IRGERA, which corresponds to the C-terminal residues 130–135 of histone H3; conjugation of this peptide to ovalbumin yields immunogen that induces rabbit polyclonal antipeptide antibodies capable of binding histone H3 if glutaraldehyde is used as a covalent linker reagent, but not if carbodiimide is used instead of glutaraldehyde [86]. Other discordant results have been documented among monoclonal antipeptide antibodies to polyhistidine, which are used to detect recombinant proteins bearing His-tags (i.e., sequences of consecutive histidine residues that often facilitate protein purification by affinity chromatography on metal-chelate resin columns); using different His-tagged cognate proteins (each bearing either C-terminal or N-terminal hexameric His-tag), the observed pattern of cross-reaction with the proteins varies considerably among the antibodies contrary to the expectation of uniformly positive cross-reaction [87]. Yet another notable case of discordant results concerns a peptide corresponding to a covalently closed loop (residues 24–41, with N- and C-terminal cysteines linked by a disulfide bond) in toxin alpha of *Naja nigricola*; antibodies capable of neutralizing the toxin are elicited by the peptide in a cyclic but not a linear form [88].

Such discordant results bring to attention the problem of discordant benchmark data, particularly where predictive methods consider only protein sequence or structure (as is the usual case); unless a predictive method is sophisticated enough to properly utilize other pertinent data (e.g., on conjugation chemistry), overtly discordant benchmark data necessarily limit the apparent performance of the method because the same prediction (i.e., either positive or negative) is rendered for any two discordant results, such that the prediction must always be deemed incorrect (i.e., either false-positive or false-negative) for one of these two results. The problem may persist in a latent form even after excluding overtly discordant benchmark data if, for example, any of the remaining data have been derived from studies wherein conjugation chemistry produced charge mismatches between immunizing peptides and cognate proteins, in which case discordance may be manifest as negative results that could otherwise have been positive had creation of the charge mismatches been avoided [15]; if a predictive method yielded positive predictions that were benchmarked against the negative results, the predictions would be labeled as false-positive even though they could just as well be labeled as true-positive.

The problem of discordant benchmark data is addressed by partitioning the benchmark dataset into two or more subsets, of which the primary subset is defined by excluding data whose interpretation is complicated by one or more experimental conditions known to disfavor production of antipeptide antibodies that cross-react with proteins; among these conditions are those resulting in charge mismatches between immunizing peptide and cognate protein (e.g., by virtue of peptide synthesis and postsynthetic chemical treatment [15]). Revisiting the above-mentioned example of the peptide IRGERA with cognate protein histone H3 [86], the primary subset could include a positive result for the glutaraldehyde-treated peptide (which would likely mimic the C-terminus of histone H3 due to loss of positive charge on isoleucine with maintenance of the negative charges on glutamate and alanine) but not a discordant negative result for the carbodiimide-treated peptide (which would likely differ markedly from the C-terminus of histone H3 due to loss of the negative charges). Recalling the influence of histidine protonation on the binding of antipeptide antibodies [55], the variable protonation state of histidine at physiologic pH [56–58] and the discordant results for cross-reactivity of anti-polyhistidine antibodies with His-tagged proteins [87], data on histidine-rich and possibly all histidine-containing sequences might best be excluded from the primary subset. Likewise, data on cysteine-containing sequences might best be excluded from the primary subset given the potential difficulty of ascertaining disulfide linkage and loop conformation between cysteine residues in peptides and proteins, notwithstanding the potential for cross-protective immunogenicity of certain cysteine-containing peptides [88].

By thus partitioning the benchmark data, the problem of discordant benchmark data is mitigated within the primary subset; and by benchmarking against the primary subset, the complexity of the predictive task is restricted for assessment of performance under relatively few and simple constraints (e.g., protein structure). Compared with benchmarking against the unpartitioned benchmark data, this decreases the risk of underrating predictive methods. Actual performance is expectedly poorer in the presence of any additional constraints (e.g., charge mismatches between immunizing peptides and their cognate proteins), but these can often be avoided in practice (e.g., through appropriate synthesis and postsynthetic chemical treatment of peptides); more importantly, poorer performance in their presence suggests the possibility of improving performance through more accurate modeling of their effects.

5. Implications for B-Cell Epitope Prediction

The overall process just described for selecting and subsequently partitioning data drastically limits the sizes of datasets for benchmarking compared with the entire body of available B-cell epitope data. This is evident on searching the Immune Epitope Database (IEDB), a comprehensive online repository of curated empirical data on epitopes [89, 90]. Over sixty-thousand individual B-cell responses are represented in IEDB, comprising positive and negative subsets of comparable size. By selecting the polyclonal

antipeptide responses associated with antibody binding leading to change in biological activity (a surrogate qualifier for antibody-mediated change in protein function), about a thousand positive responses are found; by selecting the polyclonal antipeptide responses analyzed by either immunoprecipitation or radioimmunoassay (which represent fluid-phase immunoassays), about a hundred negative responses are found. For both positive and negative responses, the discrepancies between numbers prior to and after selection largely reflect the exclusion of data acquired by probing antiprotein antibodies for cross-reactivity with peptides in high-throughput peptide-scanning experiments based on PEPSCAN technology [91–93]; the more pronounced discrepancy with negative responses probably also reflects a bias towards solid-phase immunoassays due to their greater convenience compared with fluid-phase immunoassays. These discrepancies highlight a relative scarcity of selected data and the extent to which any subsequent partitioning of these data further limits the sizes of datasets for benchmarking.

Besides facilitating the selection of data for benchmarking, IEDB also supports partitioning of the selected data with respect to factors that are represented as customizable input fields on the web-based user interface for searching among B-cell responses (http://www.immuneepitope.org/advancedQueryBcell.php). These factors include various aspects of host biology (e.g., species, sex, age) and contextual qualifiers for assayed antibodies (e.g., source material, heavy chain type). As for factors relevant to structural differences between immunizing peptides and their cognate proteins, the partitioning of data requires supplementary information that is accessible via links within IEDB to external databases of the National Center for Biotechnology Information (NCBI) [94] for the original references in published literature as well as annotated records on cognate proteins and their biological sources. As a general rule, primary sources from literature must be reviewed to ascertain the structures of both an immunizing peptide and its cognate protein before structural differences between the two can be thoroughly evaluated; this calls for attention to details of peptide chemistry (encompassing synthesis and postsynthetic covalent modification) and protein structural biology (encompassing posttranslational modification and molecular localization).

In principle, supplementary data on peptide chemistry and protein structural biology could be curated and distributed within the framework of a revised IEDB; but given the currently limited amount of selected data, a more expedient alternative is extraction of readily available data from IEDB (e.g., raw sequences of peptides and proteins) combined with independent manual curation of the supplementary data (e.g., from literature retrieved via links within IEDB). The relative scarcity of selected data motivates their provisional partitioning with respect to a few physicochemically plausible factors. To a first approximation, each factor may be qualitatively defined as the presence or absence of a certain feature posited to complicate B-cell epitope prediction on the basis of reasoning from physicochemical principles. This is illustrated using the specific example of histidine content as a factor defined by the presence of histidine in an immunizing peptide or the corresponding segment of the cognate protein: In the presence of histidine, epitope prediction is complicated by the difficulty of ascertaining the sidechain protonation state of histidine at physiologic pH. Other factors that could likewise be considered are cysteine content (defined by analogy to histidine content); unmatched backbone charge between immunizing peptide and cognate protein (due to the manner of peptide synthesis or postsynthetic chemical treatment); structural difference between immunizing peptide and cognate protein due to posttranslational modification (other than oxidation of cysteine to cystine); and localization of cognate protein in a supramolecular assembly (e.g., biological membrane or other multimeric aggregate structure). Partitioning the data with a number n of such factors yields a maximum of 2^n non-overlapping populated datasets. Of these datasets, the most valuable is that which is free of all posited complicating features (e.g., histidine and cysteine). It is logically the main reference dataset for initial benchmarking of methods for B-cell epitope prediction to assess their performance characteristics prior to refinement; if parallel benchmarking of a predictive method against the other datasets reveals poorer performance relative to this dataset, refinement of the method can focus on those factors found to be associated with performance deficits. Cycles of refinement based on parallel benchmarking could be repeated as deemed necessary to address the performance deficits; and as the body of selected data grows with the availability of new empirical results, the entire analysis from initial data partitioning onward could itself be repeated to increase its statistical power and broaden its scope to subsume additional factors.

From a practical standpoint, factors for data partitioning appear amenable to alternative casting as exclusion criteria for data selection, thus obscuring the crucial distinction between data selection and data partitioning. For instance, the factor of unmatched backbone charge between immunizing peptides and their cognate proteins appears amenable to alternative casting as a criterion for excluding data during data selection on the grounds that the attempt at molecular mimicry is flawed [15]; yet the attempt may nevertheless succeed [47, 95–100], which is not at all surprising given the molecular mimicry of protein epitopes by peptides of apparently unrelated sequence [101, 102]. Excluding data is tantamount to rejecting them as meaningless, whereas retaining them acknowledges their validity. By excluding data that arguably could be retained instead, the conceivable domain of B-cell epitope prediction is artificially restricted to problems of arbitrarily limited complexity, which creates the mistaken impression that more complex problems are either nonexistent or intractable. For example, if posttranslational modification were a criterion for excluding data on the grounds that it complicates B-cell epitope prediction [14, 67], posttranslationally modified sequences might be avoided altogether in the design of peptide-based vaccines; yet the synthesis of peptides that structurally mimic these sequences is presently feasible [103], and the expanding repertoire of techniques for inducing antibodies that bind haptens [104] offers the prospect of peptide-based vaccines

to induce protective antibody responses directed against a diverse array of posttranslationally modified epitopes.

In retrospect, the performance of methods to predict B-cell epitopes has yet to be rigorously evaluated for the design of peptide-based vaccines. At the same time, the long and unbroken history of failed attempts to develop clinically proven and commercially viable peptide-based vaccines [105, 106] points to the inadequacy of the underlying design strategies. A common albeit unjustified working assumption of these strategies is the preponderance of protein segments whose sequences can, as isolated peptide-based immunogens, mimic neutralization epitopes to induce protective antibody-mediated immunity. Typical neutralization epitopes are discontinuous; each comprises atoms that are not all located on residues of a single contiguous sequence [106]. Such an epitope may span one or more contiguous sequences, but each of these sequences as an isolated peptide immunogen may fail to induce neutralizing antipeptide antibodies. This outcome is likely if binding of cognate protein by antipeptide antibodies is impeded by steric barriers, especially where the paratopes are structurally optimized for binding peptide sequences in conformations unlike those of the corresponding sequences in native protein [54]. To induce neutralizing antipeptide antibodies, immunization with peptides identical in sequence to cognate protein segments is probably a reasonable approach only in exceptional cases, as when the cognate protein segments are conformationally unconstrained N- or C-terminal sequences entirely accessible to antibodies in biologically relevant structural contexts (e.g., on extracellular surfaces). In other cases, neutralizing antipeptide antibodies might be elicited, if at all, only by mimotopes lacking obvious sequence similarity to neutralization epitopes recognized by antiprotein antibodies [102]. Where neutralizing antipeptide antibodies can be elicited, protective immunity might be possible only with the synergistic action of such antibodies directed against multiple distinct binding sites in vivo, in analogy to synergistic protective immunity conferred by a combination of monoclonal antibodies to different neutralization epitopes [106]. Clinically informative prediction of peptides that induce neutralizing antibodies itself requires a systems view to correctly evaluate potential for both antibody-antigen binding in vivo and the possible biological consequences thereof (e.g., neutralization of viral infectivity versus enhancement of viral infection). If ever the routine design of safe and efficacious peptide-based vaccines becomes feasible, it may necessitate combining mimotopes of various neutralization epitopes for synergistic cross-protective immunogenicity, based on thorough knowledge of pathophysiological mechanisms that vaccination seeks to suppress, disrupt or otherwise circumvent.

In any event, the development of B-cell epitope prediction methods to design peptide-based vaccines remains encumbered by the problems of benchmarking discussed herein. Presently, the most challenging of these problems is the paucity of available benchmark data; in this regard, the current situation is reminiscent of the period during which the classic sequence-based methods for B-cell epitope prediction were initially developed [107]. Since then, accumulation of data gleaned from numerous studies has been the default strategy for building benchmark datasets [11, 89]. While this strategy could eventually yield ample positive data that reflect functionally relevant cross-reactivity, historical trends suggest that it would be much less effective for negative data that reflect genuine absence of cross-reactivity, owing to an overall bias towards generation of positive rather than negative data. This bias follows from the prioritization of predicted B-cell epitopes for evaluation as immunizing peptides, in keeping with the aim of peptide-based vaccine development; with the negative data generated as unintended byproducts, incentive is lacking for their confirmation as genuine using fluid-phase immunoassays.

Yet, even assuming future abundance of both positive and negative benchmark data, their use as such is subject to the criticism that they are for the most part defined by dichotomization of continuous outcome variables (e.g., the extent to which biological activity is attenuated by antibody binding, or the fraction of antigen immunoprecipitated in a fluid-phase immunoassay). This dichotomization is accomplished by applying some invariably arbitrary cut point (i.e., threshold value), which is often implicit and unknown (e.g., as determined by the limit of detection for a qualitative assay). Dichotomization of continuous data, which permits calculation of sensitivity and specificity, entails potentially significant loss of both information and statistical power [108]; this is worsened by failure to use optimal cut points [108], which are unlikely to have been consistently used among different studies from which benchmark data are pooled.

To avoid dichotomization of continuous data, benchmarking must be performed without resorting to calculation of sensitivity and specificity. Such an alternative approach could be developed by supplanting dichotomous benchmark data with continuous dose-response data on antibody-mediated modulation of biological activity, thereby redefining the predictive task as computational estimation of biological activity as a function of antibody concentration and other variables. For antibody-mediated attenuation of biological activity, the dose-response data could be expressed as the observed fraction f_{obs} of residual activity, given by

$$f_{\text{obs}} = \frac{A_{\text{obs}}}{A_{\text{max}}}, \tag{1}$$

where A_{obs} is the observed activity at some specified antibody concentration and A_{max} is the maximal activity in the absence of antibody, holding constant all variables other than antibody concentration; the corresponding predicted fraction f_{pre} of residual activity would be calculated for each antibody concentration used to obtain f_{obs}, and benchmarking would be performed by evaluating the correlation between f_{pre} and f_{obs}. For a perfect predictive method, plotting f_{pre} against f_{obs} would yield points all falling on the diagonal line defined by $y = x$, with Pearson correlation coefficient of 1.

To demonstrate how f_{pre} might be estimated, a simple conceptual model is that of a reversible bimolecular association reaction between a catalytically active enzyme E and an inhibitory antigen-binding antibody fragment I bearing a single paratope that binds a unique epitope on E

to yield a catalytically inactive complex EI. At equilibrium, the association constant K_a is given by the law of mass action as

$$K_a = \frac{[EI]}{[E][I]}, \tag{2}$$

where each symbol with enclosing square brackets ([]) denotes the molar concentration of the corresponding species. If EI is initially absent, (2) may be rewritten as

$$K_a = \frac{[EI]}{([E]_0 - [EI])([I]_0 - [EI])}, \tag{3}$$

where the subscript of 0 denotes initial value. If catalytic activity is directly proportional to [E], f_{pre} may be computed as

$$f_{pre} = 1 - \frac{[EI]}{[E]_0} \tag{4}$$

by analogy to (1). If values of $[E]_0$, $[I]_0$ and K_a are available, solving (3) for [EI] allows calculation of f_{pre} according to (4). Both $[E]_0$ and $[I]_0$ may be known empirically while K_a can be computed as

$$K_a = \exp\left(-\frac{\Delta G^\circ}{RT}\right), \tag{5}$$

where ΔG° is the standard free energy change for the association of E with I to form EI, R is the gas constant and T is the absolute temperature; if the structure of E is known, protein structural energetics [109–111] provides a means to estimate ΔG° from anticipated changes in solvent-accessible surface area (ASA) for putative B-cell epitopes of E upon their binding by paratopes [14, 15].

The main disadvantage of reliance on (3) is the assumption of an equilibrium state that may never be reached in practice; in particular, f_{pre} may overestimate f_{obs} if EI is thermodynamically stable but formed very slowly. This error might be avoided by using an alternative approach based on a kinetic description of [EI], such as

$$\frac{d[EI]}{dt} = k_{on}([E]_0 - [EI])([I]_0 - [EI]) - k_{off}[EI], \tag{6}$$

where t is reaction time, with k_{on} and k_{off} being the respective rate constants for association and dissociation. If values of k_{on} and k_{off} are also available in addition to $[E]_0$ and $[I]_0$, solving (6) for [EI] at the appropriate time t (e.g., for preincubation of E with I before assaying catalytic activity) allows calculation of f_{pre} using (4). Estimation of k_{on} could be attempted on the basis of transition state theory [112], as

$$k_{on} = \left(\frac{k_B T}{h}\right)\exp\left(-\frac{\Delta G^\ddagger}{RT}\right), \tag{7}$$

where k_B is the Boltzmann constant, h is the Planck constant and $\Delta G^\ddagger$ is the free energy change of transition state formation; in turn, k_{off} could be estimated using the relationship

$$K_a = \frac{k_{on}}{k_{off}} \tag{8}$$

in conjunction with (5) and structural-energetic methods for calculating ΔG°. The value of $\Delta G^\ddagger$ can be crudely approximated as an energetic penalty for structural adjustment of E upon binding by I, as calculated for the unfolding of a putative B-cell epitope on E [15]; better approximation of $\Delta G^\ddagger$ might be feasible through detailed computational modeling of the antibody-antigen binding process, which may occur as a multistep series of conformational changes [113].

If benchmarking were to be redefined as evaluating the correlation between f_{pre} and f_{obs}, data selection would be redefined as the admission of continuous dose-response data comprising biological activity data (from which f_{obs} could be computed) and other pertinent empirical data (e.g., reactant concentrations and protein structure, from which f_{pre} could be computed); meanwhile, data partitioning would be applicable not only as already described for dichotomous benchmark data (e.g., with respect to charge mismatches between immunizing peptides and cognate proteins), but also with respect to the manner in which f_{pre} is computed. More elaborate quantitative models than suggested by (6) describe real antibody-antigen interactions [81]. Typical antibody molecules each bear two or more paratopes while an antigen molecule may bear two or more similar if not identical B-cell epitopes; cooperative antibody-antigen binding phenomena do occur, and antibody-mediated modulation of biological activity may be nonlinearly related to the extent of antibody binding (e.g., where neutralization of viral infectivity is enhanced by antibody-mediated aggregation of virions, which may occur only near the equivalence point of antibody-virion interaction [84]). Under these various circumstances, uniformly accurate prediction is unlikely if based on a first-approach method for computing f_{pre}. Data partitioning could therefore be performed according to the computational complexity demanded by modeling of the antibody-antigen binding process and its functional consequences, such that relatively simple methods for computing f_{pre} are benchmarked against data on correspondingly simple systems (e.g., a single-domain enzyme having but a single active site); this could facilitate refinement of the methods that possibly extends their applicability to more complex systems (e.g., an enzyme having multiple active sites, or a virion having multiple binding sites for receptors on host cells).

At a deeper level, the underlying logic of B-cell epitope prediction methods may itself provide additional insights on how problems might be avoided through data selection and partitioning. As applied to the design of peptide-based vaccines, the impact of selecting data on excessively long immunizing peptides is, for example, clarified by examining the implications of a structural-energetic approach to B-cell epitope prediction [14, 15]; this models immunodominance among B-cell epitopes of immunizing peptides as a thermodynamically determined hierarchical steric-exclusion phenomenon, based on the premise that antibodies are preferentially elicited by an immunodominant epitope due to its higher affinity for antibody relative to other epitopes with which it physically overlaps [15]. If an immunizing peptide contains exactly one predicted immunodominant epitope, any observed functionally relevant cross-reactivity

can be readily attributed to the epitope; but if the peptide contains more than one such epitope, a problem of ambiguous attribution arises that can lead to erroneous benchmarking results (e.g., if functionally relevant cross-reactivity is incorrectly attributed to a functionally irrelevant epitope for computation of f_{pre}). This implies the existence of an upper limit on the length of an immunizing peptide for data on the peptide to be reliably informative; if immunizing peptides are assumed to be completely unfolded such that each hexapeptide sequence thereof is regarded as a candidate epitope [14, 15], ambiguous attribution seems possible for an immunizing peptide as short as 12 residues (in which case two predicted immunodominant epitopes might occur in tandem) and may be unavoidable for an immunizing peptide as short as 17 residues (in which case two predicted immunodominant epitopes may be accommodated regardless of how the first one is positioned), and even if an immunizing peptide is so short that it contains only one predicted immunodominant epitope, ambiguous attribution is still a problem if the epitope sequence occurs in nonidentical structural contexts as part of the cognate protein (e.g., as sequence repeats within a single protein domain, each of which yields a distinct value of f_{pre}). Data selection could avoid such problems of ambiguous attribution by admitting data on functionally relevant cross-reactivity only where the immunizing peptide contains exactly one predicted immunodominant epitope sequence that does not occur in nonidentical structural contexts as part of the cognate protein; this would also avoid the problem of having to experimentally define the structural boundaries of epitopes, which is impossible to accomplish by analyzing polyclonal antibodies through the use of immunoassays [9]. Subsequent partitioning of the data thus selected could be performed with respect to repetition of the predicted immunodominant epitope sequence in the cognate protein (e.g., in a soluble symmetric homodimer, for which the sequence occurs in two structurally identical contexts that yield the same value of f_{pre}); such repetition may enable the formation of large immune complexes through cross-linkage of antibodies by antigen, which could complicate the computational analysis of antibody-antigen binding (e.g., by leading to antibody-mediated aggregation of virions).

Progress in B-cell epitope prediction might therefore be realized through development of quantitative methods that is guided by physicochemical principles; but in the final analysis, success is an unrealistic expectation for peptide-based vaccine design without a comprehensive contextual basis for application-specific computational modeling of immune responses and their biological consequences. Even if a computational method were developed that could accurately predict functionally relevant cross-reactivity in general (e.g., as benchmarked by evaluating f_{pre} against f_{obs} for a wide variety of systems), only profound context-specific knowledge (e.g., on molecular recognition of host cell receptors by pathogen virulence factors) would ensure correct identification of peptide sequences that could induce protective immunity against a particular disease (e.g., by mimicking appropriate neutralization epitopes). This knowledge would encompass aspects of host immunobiology such as immune tolerance that result from the interplay of numerous genetic and environmental factors in both healthy and diseased states; immune tolerance deserves special attention in this regard, as problems could arise due to either immune tolerance itself (e.g., complicating the prediction of immunogenic epitopes) or vaccine-induced loss thereof (e.g., leading to autoimmunity or other forms of hypersensitivity [114, 115]). Perhaps the most serious defect of attempts to design peptide-based vaccines has been the naive understanding of B-cell epitope prediction as a generic procedure that requires only physicochemical knowledge of antigens.

6. Conclusions

The development of peptide-based vaccines demands refinement of methods to predict B-cell epitopes that is based on benchmarking against judiciously selected and partitioned empirical data. The data are meaningful for this purpose if they pertain to cross-reactivity of polyclonal antipeptide antibodies with proteins, with the positive data reflecting functionally relevant cross-reactivity and the negative data reflecting genuine absence of cross-reactivity. These data can be partitioned with respect to multiple factors that complicate B-cell epitope prediction, of which the most important are those that define key structural differences between immunizing peptides and their cognate proteins. Parallel benchmarking against the resulting subsets of data is a foundational strategy to discover context-dependent variations in the performance of methods for B-cell epitope prediction, serving to identify and address the limitations of these methods through iterative cycles of benchmarking and refinement. Ultimately, this could lead to the establishment of an integrative computational platform for B-cell epitope prediction that reliably guides the design of peptide-based vaccines. In the meantime, B-cell epitope prediction must transcend its de facto role of a rudimentary screening tool whose utility is severely limited by a narrow reductionist view of vaccine design.

Acknowledgment

This work was supported by a grant (PGMA-SEGS) from the Commission on Higher Education of the Philippine Government.

References

[1] S. J. Marshall, "Developing countries face double burden of disease," *Bulletin of the World Health Organization*, vol. 82, no. 7, p. 556, 2004.

[2] A. S. De Groot and R. Rappuoli, "Genome-derived vaccines," *Expert Review of Vaccines*, vol. 3, no. 1, pp. 59–76, 2004.

[3] B. Korber, M. LaBute, and K. Yusim, "Immunoinformatics comes of age," *PLoS Computational Biology*, vol. 2, no. 6, article e71, 2006.

[4] A. S. Fauci, "Emerging and re-emerging infectious diseases: influenza as a prototype of the host-pathogen balancing act," *Cell*, vol. 124, no. 4, pp. 665–670, 2006.

[5] D. Belpomme, P. Irigaray, A. J. Sasco, et al., "The growing incidence of cancer: role of lifestyle and screening detection (Review)," *International Journal of Oncology*, vol. 30, no. 5, pp. 1037–1049, 2007.

[6] A. W. Purcell, J. McCluskey, and J. Rossjohn, "More than one reason to rethink the use of peptides in vaccine design," *Nature Reviews Drug Discovery*, vol. 6, no. 5, pp. 404–414, 2007.

[7] C. G. Gay, R. Zuerner, J. P. Bannantine, et al., "Genomics and vaccine development," *Revue Scientifique et Technique*, vol. 26, no. 1, pp. 49–67, 2007.

[8] J. A. Greenbaum, P. H. Andersen, M. Blythe, et al., "Towards a consensus on datasets and evaluation metrics for developing B-cell epitope prediction tools," *Journal of Molecular Recognition*, vol. 20, no. 2, pp. 75–82, 2007.

[9] M. H. V. Van Regenmortel, "What is a B cell epitope?" in *Epitope Mapping Protocols*, U. Reineke and M. Schutkowski, Eds., pp. 3–20, Humana Press, New York, NY, USA, 2nd edition, 2009.

[10] M. H. V. Van Regenmortel, "Immunoinformatics may lead to a reappraisal of the nature of B cell epitopes and of the feasibility of synthetic peptide vaccines," *Journal of Molecular Recognition*, vol. 19, no. 3, pp. 183–187, 2006.

[11] J. L. Pellequer, E. Westhof, and M. H. V. Van Regenmortel, "Predicting location of continuous epitopes in proteins from their primary structures," *Methods in Enzymology*, vol. 203, pp. 176–201, 1991.

[12] M. J. Blythe and D. R. Flower, "Benchmarking B cell epitope prediction: underperformance of existing methods," *Protein Science*, vol. 14, no. 1, pp. 246–248, 2005.

[13] J. E. Larsen, O. Lund, and M. Nielsen, "Improved method for predicting linear B-cell epitopes," *Immunome Research*, vol. 2, article 2, 2006.

[14] S. E. C. Caoili, "A structural-energetic basis for B-cell epitope prediction," *Protein and Peptide Letters*, vol. 13, no. 7, pp. 743–751, 2006.

[15] S. E. C. Caoili, "Immunization with peptide-protein conjugates: impact on benchmarking B-cell epitope prediction for vaccine design," *Protein and Peptide Letters*, vol. 17, no. 3, pp. 386–398, 2010.

[16] M. H. V. Van Regenmortel and J. L. Pellequer, "Predicting antigenic determinants in proteins: looking for unidimensional solutions to a three-dimensional problem?" *Peptide Research*, vol. 7, no. 4, pp. 224–228, 1994.

[17] M. H. V. Van Regenmortel, "Molecular recognition in the post-reductionist era," *Journal of Molecular Recognition*, vol. 12, no. 1, pp. 1–2, 1999.

[18] M. H. V. Van Regenmortel, "Pitfalls of reductionism in the design of peptide-based vaccines," *Vaccine*, vol. 19, no. 17–19, pp. 2369–2374, 2001.

[19] M. H. V. Van Regenmortel, "Reductionism and the search for structure-function relationships in antibody molecules," *Journal of Molecular Recognition*, vol. 15, no. 5, pp. 240–247, 2002.

[20] M. H. V. Van Regenmortel, "Biological complexity emerges from the ashes of genetic reductionism," *Journal of Molecular Recognition*, vol. 17, no. 3, pp. 145–148, 2004.

[21] J. Huang, W. Honda, and M. Kanehisa, "Predicting B cell epitope residues with network topology based amino acid indices," *Genome Informatics*, vol. 19, pp. 40–49, 2007.

[22] J. A. Swets, "Measuring the accuracy of diagnostic systems," *Science*, vol. 240, no. 4857, pp. 1285–1293, 1988.

[23] O. Lund, M. Nielsen, C. Lundegaard, C. Kesmir, and S. Brunak, *Immunological Bioinformatics*, MIT Press, Cambridge, Mass, USA, 1st edition, 2005.

[24] L. Craig, P. C. Sanschagrin, A. Rozek, S. Lackie, L. A. Kuhn, and J. K. Scott, "The role of structure in antibody cross-reactivity between peptides and folded proteins," *Journal of Molecular Biology*, vol. 281, no. 1, pp. 183–201, 1998.

[25] N. Green, H. Alexander, A. Olson, et al., "Immunogenic structure of the influenza virus hemagglutinin," *Cell*, vol. 28, no. 3, pp. 477–487, 1982.

[26] M. Z. Atassi and C. R. Young, "Discovery and implications of the immunogenicity of free small synthetic peptides: powerful tools for manipulating the immune system and for production of antibodies and T cells of preselected submolecular specificities," *Critical Reviews in Immunology*, vol. 5, no. 4, pp. 387–409, 1985.

[27] P. V. Barnett, D. J. Rowlands, and N. R. Parry, "Characterization of monoclonal antibodies raised against a synthetic peptide capable of inducing a neutralizing response to human rhinovirus type 2," *Journal of General Virology*, vol. 74, no. 7, pp. 1295–1302, 1993.

[28] W. G. Laver, G. M. Air, R. G. Webster, and S. J. Smith-Gill, "Epitopes on protein antigens: misconceptions and realities," *Cell*, vol. 61, no. 4, pp. 553–556, 1990.

[29] C. Schwab and H. R. Bosshard, "Caveats for the use of surface-adsorbed protein antigen to test the specificity of antibodies," *Journal of Immunological Methods*, vol. 147, no. 1, pp. 125–134, 1992.

[30] L. Leder, H. Wendt, C. Schwab, et al., "Genuine and apparent cross-reaction of polyclonal antibodies to proteins and peptides," *European Journal of Biochemistry*, vol. 219, no. 1-2, pp. 73–81, 1994.

[31] R. M. Williams, Z. Obradovi, V. Mathura, et al., "The protein non-folding problemml: amino acid determinants of intrinsic order and disorder," in *Proceedings of the Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, pp. 89–100, 2001.

[32] V. N. Uversky, "Protein folding revisited. A polypeptide chain at the folding-misfolding-nonfolding cross-roads: which way to go?" *Cellular and Molecular Life Sciences*, vol. 60, no. 9, pp. 1852–1871, 2003.

[33] A. L. Fink, "Natively unfolded proteins," *Current Opinion in Structural Biology*, vol. 15, no. 1, pp. 35–41, 2005.

[34] A. K. Dunker, C. J. Oldfield, J. Meng, et al., "The unfoldomics decade: an update on intrinsically disordered proteins," *BMC Genomics*, vol. 9, supplement 2, p. S1, 2008.

[35] A. K. Dunker, J. D. Lawson, C. J. Brown, et al., "Intrinsically disordered protein," *Journal of Molecular Graphics and Modelling*, vol. 19, no. 1, pp. 26–59, 2001.

[36] S. Vucetic, C. J. Brown, A. K. Dunker, and Z. Obradovic, "Flavors of protein disorder," *Proteins*, vol. 52, no. 4, pp. 573–584, 2003.

[37] P. Radivojac, L. M. Iakoucheva, C. J. Oldfield, Z. Obradovic, V. N. Uversky, and A. K. Dunker, "Intrinsic disorder and functional proteomics," *Biophysical Journal*, vol. 92, no. 5, pp. 1439–1456, 2007.

[38] A. Szilagyi, D. Gyorffy, and P. Zavodszky, "The twilight zone between protein order and disorder," *Biophysical Journal*, vol. 95, no. 4, pp. 1612–1626, 2008.

[39] L. M. Espinoza-Fonseca, "Reconciling binding mechanisms of intrinsically disordered proteins," *Biochemical and Biophysical Research Communications*, vol. 382, no. 3, pp. 479–482, 2009.

[40] H. J. Dyson and P. E. Wright, "Coupling of folding and binding for unstructured proteins," *Current Opinion in Structural Biology*, vol. 12, no. 1, pp. 54–60, 2002.

[41] A. B. Sigalov, A. V. Zhuravleva, and V. Y. Orekhov, "Binding of intrinsically disordered proteins is not necessarily accompanied by a structural transition to a folded form," *Biochimie*, vol. 89, no. 3, pp. 419–421, 2007.

[42] P. E. Wright and H. J. Dyson, "Linking folding and binding," *Current Opinion in Structural Biology*, vol. 19, no. 1, pp. 31–38, 2009.

[43] E. Hazy and P. Tompa, "Limitations of induced folding in molecular recognition by intrinsically disordered proteins," *ChemPhysChem*, vol. 10, no. 9-10, pp. 1415–1419, 2009.

[44] Y. M. Tang, G.-F. Chen, P. A. Thompson, et al., "Development of an antipeptide antibody that binds to the C-terminal region of human CYP1B1," *Drug Metabolism and Disposition*, vol. 27, no. 2, pp. 274–280, 1999.

[45] M. A. Schmidt, B. Raupach, M. Szulczynski, and J. Marzillier, "Identification of linear B-cell determinants of pertussis toxin associated with the receptor recognition site of the S3 subunit," *Infection and Immunity*, vol. 59, no. 4, pp. 1402–1408, 1991.

[46] Z. Gechtman, A. Belleli, S. Lechpammer, and S. Shaltiel, "The cluster of basic amino acids in vitronectin contributes to its binding of plasminogen activator inhibitor-1: evidence from thrombin-, elastase- and plasmin-cleaved vitronectins and anti-peptide antibodies," *Biochemical Journal*, vol. 325, no. 2, pp. 339–349, 1997.

[47] P. L. Simon, V. Kumar, J. S. Lillquist, et al., "Mapping of neutralizing epitopes and the receptor binding site of human interleukin 1 beta," *The Journal of Biological Chemistry*, vol. 268, no. 13, pp. 9771–9779, 1993.

[48] R. Reitmaier, "Review of immunoinformatic approaches to in-silico B-cell epitope prediction," *Nature Precedings*, 2007.

[49] U. Reimer, "Prediction of linear B-cell epitopes," *Methods in Molecular Biology*, vol. 524, pp. 335–344, 2009.

[50] U. Kulkarni-Kale, S. Bhosle, and A. S. Kolaskar, "CEP: a conformational epitope prediction server," *Nucleic Acids Research*, vol. 33, supplement 2, pp. W168–W171, 2005.

[51] V. Batori, E. P. Friis, H. Nielsen, and E. L. Roggen, "An in silico method using an epitope motif database for predicting the location of antigenic determinants on proteins in a structural context," *Journal of Molecular Recognition*, vol. 19, no. 1, pp. 21–29, 2006.

[52] E. D. Getzoff, H. M. Geysen, and S. J. Rodda, "Mechanisms of antibody binding to a protein," *Science*, vol. 235, no. 4793, pp. 1191–1196, 1987.

[53] S. Deroo and C. P. Muller, "Antigenic and immunogenic phage displayed mimotopes as substitute antigens: applications and limitations," *Combinatorial Chemistry and High Throughput Screening*, vol. 4, no. 1, pp. 75–110, 2001.

[54] S.-W. W. Chen, M. H. V. Van Regenmortel, and J.-L. Pellequer, "Structure-activity relationships in peptide-antibody complexes: implications for epitope prediction and development of synthetic peptide vaccines," *Current Medicinal Chemistry*, vol. 16, no. 8, pp. 953–964, 2009.

[55] A. Kondo, H. Takamatsu, S. Katoh, and E. Sada, "Adsorption equilibrium in immunoaffinity chromatography with antibodies to synthetic peptides," *Biotechnology and Bioengineering*, vol. 35, no. 2, pp. 146–151, 1990.

[56] J. Sancho, L. Serrano, and A. R. Fersht, "Histidine residues at the N- and C-termini of alpha-helices: perturbed pKas and protein stability," *Biochemistry*, vol. 31, no. 8, pp. 2253–2258, 1992.

[57] R. Loewenthal, J. Sancho, and A. R. Fersht, "Histidine-aromatic interactions in barnase: elevation of histidine pKa and contribution to protein stability," *Journal of Molecular Biology*, vol. 224, no. 3, pp. 759–770, 1992.

[58] S. P. Edgcomb and K. P. Murphy, "Variability in the pKa of histidine side-chains correlates with burial within proteins," *Proteins*, vol. 49, no. 1, pp. 1–6, 2002.

[59] P. Nakra, V. Manivel, R. A. Vishwakarma, and K. V. S. Rao, "B cell responses to a peptide epitope. X. Epitope selection in a primary response is thermodynamically regulated," *Journal of Immunology*, vol. 164, no. 11, pp. 5615–5625, 2000.

[60] P. Motte, G. Alberici, M. Ait-Abdellah, and D. Bellet, "Monoclonal antibodies distinguish synthetic peptides that differ in one chemical group," *Journal of Immunology*, vol. 138, no. 10, pp. 3332–3338, 1987.

[61] T. C. Liang, W. Luo, J.-T. Hsieh, and S.-H. Lin, "Antibody binding to a peptide but not the whole protein by recognition of the C-terminal carboxy group," *Archives of Biochemistry and Biophysics*, vol. 329, no. 2, pp. 208–214, 1996.

[62] T. Zhang, E. Bertelsen, and T. Alber, "Entropic effects of disulphide bonds on protein stability," *Nature Structural Biology*, vol. 1, no. 7, pp. 434–438, 1994.

[63] J. Novotny, M. Handschumacher, and E. Haber, "Antigenic determinants in proteins coincide with surface regions accessible to large probes (antibody domains)," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 83, no. 2, pp. 226–230, 1986.

[64] D. C. Hancock, N. J. O'Reilly, and G. I. Evan, "Synthesis of peptides for use as immunogens," *Methods in Molecular Biology*, vol. 80, pp. 69–79, 1998.

[65] J. P. Briand, S. Muller, and M. H. V. Van Regenmortel, "Synthetic peptides as antigens: pitfalls of conjugation methods," *Journal of Immunological Methods*, vol. 78, no. 1, pp. 59–69, 1985.

[66] T. E. Adrian, "Production of antisera using peptide conjugates," *Methods in Molecular Biology*, vol. 73, pp. 239–249, 1997.

[67] W. J. Gullick, "Production of antisera to synthetic peptides," *Methods in Molecular Biology*, vol. 32, pp. 389–399, 1994.

[68] F. T. Liang, J. M. Nowling, and M. T. Philipp, "Cryptic and exposed invariable regions of VlsE, the variable surface antigen of *Borrelia burgdorferi* sl," *Journal of Bacteriology*, vol. 182, no. 12, pp. 3597–3601, 2000.

[69] A. C. M. Ferreon, Y. Gambin, E. A. Lemke, and A. A. Deniz, "Interplay of alpha-synuclein binding and conformational switching probed by single-molecule fluorescence," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 14, pp. 5645–5650, 2009.

[70] H. C. Haspel, M. G. Rosenfeld, and O. M. Rosen, "Characterization of antisera to a synthetic carboxyl-terminal peptide of the glucose transporter protein," *The Journal of Biological Chemistry*, vol. 263, no. 1, pp. 398–403, 1988.

[71] H.-Y. Wang, L. Lipfert, C. C. Malbon, and S. Bahouth, "Site-directed anti-peptide antibodies define the topography of the beta-adrenergic receptor," *The Journal of Biological Chemistry*, vol. 264, no. 24, pp. 14424–14431, 1989.

[72] M. C. Dinauer, E. A. Pierce, R. W. Erickson, et al., "Point mutation in the cytoplasmic domain of the neutrophil p22-phox cytochrome b subunit is associated with a nonfunctional NADPH oxidase and chronic granulomatous disease," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 88, no. 24, pp. 11231–11235, 1991.

[73] M. Bruss, R. Hammermann, S. Brimijoin, and H. Bonisch, "Antipeptide antibodies confirm the topology of the human norepinephrine transporter," *The Journal of Biological Chemistry*, vol. 270, no. 16, pp. 9197–9201, 1995.

[74] A. Baude, Z. Nusser, E. Molnar, R. A. J. McIlhinney, and P. Somogyi, "High-resolution immunogold localization of AMPA type glutamate receptor subunits at synaptic and non-synaptic sites in rat hippocampus," *Neuroscience*, vol. 69, no. 4, pp. 1031–1055, 1995.

[75] R. Kuroda, J.-Y. Kinoshita, M. Honsho, J.-Y. Mitoma, and A. Ito, "In situ topology of cytochrome b5 in the endoplasmic reticulum membrane," *Journal of Biochemistry*, vol. 120, no. 4, pp. 828–833, 1996.

[76] N. N. Dewji and S. J. Singer, "Cell surface expression of the Alzheimer disease-related presenilin proteins," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 18, pp. 9926–9931, 1997.

[77] J. Yu and M. Wessling-Resnick, "Structural and functional analysis of SFT, a stimulator of Fe transport," *The Journal of Biological Chemistry*, vol. 273, no. 33, pp. 21380–21385, 1998.

[78] G. Lambert, M. Traebert, N. Hernando, J. Biber, and H. Murer, "Studies on the topology of the renal type II NaPi-cotransporter," *Pflugers Archiv European Journal of Physiology*, vol. 437, no. 6, pp. 972–978, 1999.

[79] P. D. Taylor, T. Day, D. Nagy, G. Wild, J.-B. Andre, and A. Gardner, "The evolutionary consequences of plasticity in host-pathogen interactions," *Theoretical Population Biology*, vol. 69, no. 3, pp. 323–331, 2006.

[80] J. N. Wilson, D. J. Nokes, and N. J. Dimmock, "Analysis of the relationship between immunogenicity and immunity for viral subunit vaccines," *Journal of Medical Virology*, vol. 64, no. 4, pp. 560–568, 2001.

[81] P. J. Klasse and Q. J. Sattentau, "Occupancy and mechanism in antibody-mediated neutralization of animal viruses," *Journal of General Virology*, vol. 83, no. 9, pp. 2091–2108, 2002.

[82] D. R. Burton, E. O. Saphire, and P. W. H. I. Parren, "A model for neutralization of viruses based on antibody coating of the virion surface," *Current Topics in Microbiology and Immunology*, vol. 260, pp. 109–143, 2001.

[83] P. W. H. I. Parren and D. R. Burton, "The antiviral activity of antibodies in vitro and in vivo," *Advances in Immunology*, vol. 77, pp. 195–262, 2001.

[84] S. A. Reading and N. J. Dimmock, "Neutralization of animal virus infectivity by antibody," *Archives of Virology*, vol. 152, no. 6, pp. 1047–1059, 2007.

[85] M. B. Zwick and D. R. Burton, "HIV-1 neutralization: mechanisms and relevance to vaccine design," *Current HIV Research*, vol. 5, no. 6, pp. 608–624, 2007.

[86] J.-P. Briand, C. Barin, M. H. V. Van Regenmortel, and S. Muller, "Application and limitations of the multiple antigen peptide (MAP) system in the production and evaluation of anti-peptide and anti-protein antibodies," *Journal of Immunological Methods*, vol. 156, no. 2, pp. 255–265, 1992.

[87] N. Debeljak, L. Feldman, K. L. Davis, R. Komel, and A. J. Sytkowski, "Variability in the immunodetection of His-tagged recombinant proteins," *Analytical Biochemistry*, vol. 359, no. 2, pp. 216–223, 2006.

[88] M. Leonetti, L. Pillet, B. Maillere, et al., "Immunization with a peptide having both T cell and conformationally restricted B cell epitopes elicits neutralizing antisera against a snake neurotoxin," *Journal of Immunology*, vol. 145, no. 12, pp. 4214–4221, 1990.

[89] R. Vita, K. Vaughan, L. Zarebski, et al., "Curation of complex, context-dependent immunological data," *BMC Bioinformatics*, vol. 7, article 341, Article ID 341, 2006.

[90] R. Vita, B. Peters, and A. Sette, "The curation guidelines of the immune epitope database and analysis resource," *Cytometry A*, vol. 73, no. 11, pp. 1066–1070, 2008.

[91] J. M. Carter, "Epitope mapping of a protein using the Geysen (PEPSCAN) procedure," *Methods in Molecular Biology*, vol. 36, pp. 207–223, 1994.

[92] J. M. Carter and L. Loomis-Price, "B cell epitope mapping using synthetic peptides," *Current Protocols in Immunology*, chapter 9, unit 9.4, 2004.

[93] U. Reineke and R. Sabat, "Antibody epitope mapping using SPOT peptide arrays," *Methods in Molecular Biology*, vol. 524, pp. 145–167, 2009.

[94] A. D. Baxevanis, "Searching the NCBI databases using Entrez," *Current Protocols in Human Genetics*, chapter 6, unit 6.10, 2006.

[95] J. Bouhnik, F.-X. Galen, and J. Menard, "Production and characterization of human renin antibodies with region-oriented synthetic peptides," *The Journal of Biological Chemistry*, vol. 262, no. 6, pp. 2913–2918, 1987.

[96] A. J. Sytkowski and K. A. Donahue, "Immunochemical studies of human erythropoietin using site-specific anti-peptide antibodies. Identification of a functional domain," *The Journal of Biological Chemistry*, vol. 262, no. 3, pp. 1161–1165, 1987.

[97] I. Harari, A. Donohue-Rolfe, G. Keusch, and R. Arnon, "Synthetic peptides of Shiga toxin B subunit induce antibodies which neutralize its biological activity," *Infection and Immunity*, vol. 56, no. 6, pp. 1618–1624, 1988.

[98] S. Harshman, J. E. Alouf, O. Siffert, and F. Baleux, "Reaction of staphylococcal alpha-toxin with peptide-induced antibodies," *Infection and Immunity*, vol. 57, no. 12, pp. 3856–3862, 1989.

[99] C. H. Pontzer, J. K. Russell, and H. M. Johnson, "Localization of an immune functional site on staphylococcal enterotoxin A using the synthetic peptide approach," *Journal of Immunology*, vol. 143, no. 1, pp. 280–284, 1989.

[100] A. Bavoso, A. Ostuni, J. De Vendel, et al., "Aldehyde modification of peptide immunogen enhances protein-reactive antibody response to toxic shock syndrome toxin-1," *Journal of Peptide Science*, vol. 12, no. 12, pp. 843–849, 2006.

[101] C. D. Partidos, "Peptide mimotopes as candidate vaccines," *Current Opinion in Molecular Therapeutics*, vol. 2, no. 1, pp. 74–79, 2000.

[102] C. D. Partidos and M. W. Steward, "Mimotopes of viral antigens and biologically important molecules as candidate vaccines and potential immunotherapeutics," *Combinatorial Chemistry and High Throughput Screening*, vol. 5, no. 1, pp. 15–27, 2002.

[103] G. Sabatino and A. M. Papini, "Advances in automatic, manual and microwave-assisted solid-phase peptide synthesis," *Current Opinion in Drug Discovery and Development*, vol. 11, no. 6, pp. 762–770, 2008.

[104] R. Lemus and M. H. Karol, "Conjugation of haptens," *Methods in Molecular Medicine*, vol. 138, pp. 167–182, 2008.

[105] D. Hans, P. R. Young, and D. P. Fairlie, "Current status of short synthetic peptides as vaccines," *Medicinal Chemistry*, vol. 2, no. 6, pp. 627–646, 2006.

[106] M. H. V. Van Regenmortel, "Synthetic peptide vaccines and the search for neutralization B cell epitopes," *The Open Vaccine Journal*, vol. 2, pp. 33–44, 2009.

[107] T. P. Hopp and K. R. Woods, "Prediction of protein antigenic determinants from amino acid sequences," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 78, no. 6 I, pp. 3824–3828, 1981.

[108] V. Fedorov, F. Mannino, and R. Zhang, "Consequences of dichotomization," *Pharmaceutical Statistics*, vol. 8, no. 1, pp. 50–61, 2009.

[109] K. P. Murphy and E. Freire, "Thermodynamics of structural stability and cooperative folding behavior in proteins," *Advances in Protein Chemistry*, vol. 43, pp. 313–361, 1992.

[110] K. P. Murphy, "Predicting binding energetics from structure: looking beyond ΔG°," *Medicinal Research Reviews*, vol. 19, no. 4, pp. 333–339, 1999.

[111] S. P. Edgcomb and K. P. Murphy, "Structural energetics of protein folding and binding," *Current Opinion in Biotechnology*, vol. 11, no. 1, pp. 62–66, 2000.

[112] M. E. Mummert and E. W. Voss Jr., "Transition-state theory and secondary forces in antigen-antibody complexes," *Biochemistry*, vol. 35, no. 25, pp. 8187–8192, 1996.

[113] L. C. James and D. S. Tawfik, "Structure and kinetics of a transient antibody binding intermediate reveal a kinetic discrimination mechanism in antigen recognition," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 36, pp. 12730–12735, 2005.

[114] F. Verdier, "Non-clinical vaccine safety assessment," *Toxicology*, vol. 174, no. 1, pp. 37–43, 2002.

[115] J. Descotes, G. Ravel, and C. Ruat, "Vaccines: predicting the risk of allergy and autoimmunity," *Toxicology*, vol. 174, no. 1, pp. 45–51, 2002.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 832341, 8 pages
doi:10.1155/2010/832341

Methodology Report

Proposing Low-Similarity Peptide Vaccines against *Mycobacterium tuberculosis*

Guglielmo Lucchese, Angela Stufano, and Darja Kanduc

Department of Biochemistry and Molecular Biology "Ernesto Quagliariello", University of Bari, Bari 70126, Italy

Correspondence should be addressed to Darja Kanduc, d.kanduc@biologia.uniba.it

Received 23 September 2009; Revised 2 December 2009; Accepted 24 March 2010

Academic Editor: Yongqun Oliver He

Using the currently available proteome databases and based on the concept that a rare sequence is a potential epitope, epitopic sequences derived from *Mycobacterium tuberculosis* were examined for similarity score to the proteins of the host in which the epitopes were defined. We found that: (i) most of the bacterial linear determinants had peptide fragment(s) that were rarely found in the host proteins and (ii) the relationship between low similarity and epitope definition appears potentially applicable to T-cell determinants. The data confirmed the hypothesis that low-sequence similarity shapes or determines the epitope definition at the molecular level and provides a potential tool for designing new approaches to prevent, diagnose, and treat tuberculosis and other infectious diseases.

1. Introduction

Mycobacterium tuberculosis is the causative agent of tuberculosis (TB) and is a major infectious pathogen. A resurgence of tuberculosis has marked the beginning of the third millennium, with about 8 million new cases and 1.8 million deaths annually worldwide attributed to *M. tuberculosis* [1]. More recently, the World Health Organization's Global Tuberculosis Control Report 2009 estimated 9.27 million cases of TB in 2007 [2], an increase from 9.24 million cases in 2006, 8.3 million cases in 2000, and 6.6 million cases in 1990. These numbers underscore the need to develop and characterize rational and effective TB therapies and diagnostic reagents. Currently, the *M. bovis* bacillus Calmette-Guérin (BCG) vaccine is used against TB. BCG has excellent immuno-potentiating properties, but it is not uniformly effective [3], and efforts to develop a more effective *M. bovis*-derived vaccine against TB have not had satisfactory results [4, 5]. Moreover, the emergence of multiple-drug-resistant TB [6] and the increasing prevalence of mycobacterial disease in people with acquired immunodeficiency syndrome [7] add urgency to the need for an effective vaccine against TB.

Recent studies have provided convincing evidence that antibodies can protect against mycobacterial infections [8]. Passive immunization using mono- and polyclonal antibodies and humoral immune responses against specific mycobacterial antigens in experimental animals have shown that antibodies may play an anti-TB role [9–11]. The epitopic characterization of mycobacterial antigens is performed antigen-by-antigen, searching for specific peptide features and epitope-paratope interaction domains for each mycobacterial protein antigen [12–14]. In essence, the common molecular basis and the fundamental rules underlying the immune function of specific peptide units remain unexplained [15, 16]. Little is known about the structure-function relationship(s) of protein sequences in immunology. This lack of knowledge may underlie the confusion surrounding neutralizing versus nonneutralizing antibodies, harmful vaccine side effects, and imprecise diagnostic tools.

Within this framework, we propose that rare peptide sequences, that is, peptide motifs rarely found in host proteins, elicit an immune response. Unique protein sequences are more likely to evoke an immune response than are highly repeated motifs, which would be immunogenically silenced by tolerance. We have previously validated the relationship between sequence rarity and peptide immunoreactivity in cancer, autoimmunity, and infectious disease models. We

reported several examples of specific targeting of peptide regions with no or limited sequence similarity to the host proteome, including mono- and polyclonal antibodies raised against EC Her-2/Neu oncoprotein [17], desmoglein-3 [18], melan-a/MART-1 [19], high-molecular-weight melanoma-associated-antigen [20, 21], tyrosinase [22, 23], prostate-specific-antigen [24], HPV16 E7 [25], HCV [26], and influenza A [27] proteins. In addition, a review of the epitope mapping literature revealed that most peptide epitopes obey the low-similarity rule [28], further supporting our hypothesis. Taken together, our data [17–27] and findings from other research [28] suggest that unique peptide sequences are more likely to be immunogenic than are highly repeated peptide sequences, which may be immuno-tolerated.

In the present study, we investigated the low-similarity hypothesis in *M. tuberculosis* antigen epitopes with the aim of defining immunogenic peptide sequences that may be effective for anti-TB immunotherapy. We studied validated *M. tuberculosis* epitopes currently cataloged in the Immune Epitope Database and Analysis Resources (IEDB) (http://www.immuneepitope.org/) [29, 30]. We report that most of the mycobacterial epitopes involved in the immune response are characterized by a low level of similarity to the host proteome.

2. Methods

The IEDB contains hundreds of mycobacterial epitopes derived from various strains and with different characteristics [30]. Specifically, the IEDB contains B- and T-cell mycobacterial epitopes that are (a) linear or conformational; (b) derived from *M. tuberculosis*, *M. bovis, M. Avium,* and *M. leprae*; (c) of different lengths (up to 36 aa) [31]; and (d) have sequences that were negative, positive, or produced mixed negative/positive results in immunoassays. The present report focused on B-cell linear epitopes derived from *M. tuberculosis* that were positive in immunoassays.

These mycobacterial epitopes were analyzed for similarity to the host proteins with the aim of identifying identical amino acid groupings that were common sequences [32]. The amino acid groupings were linear pentapeptides. Pentapeptides were used because the canonical literature indicates that five to six amino acids are a sufficient minimal determinant for an epitope-paratope interaction, and thus a pentapeptide can act as an immune unit and play a crucial role in cell immunoreactivity and antigen-antibody recognition [33, 34].

Fifty-six *M. tuberculosis* epitopes cataloged as immuno-positive at IEDB resource were dissected into pentamer motifs offset by one residue: that is, WDEDG, DEDGE, EDGEK, DGEKR, and so forth. Then, each 5-mer was used as a probe to scan the entire host proteome searching for identical matches using the PIR protein database and PIR perfect peptide match program (http://pir.georgetown.edu/pirwww/) [35]. The number of matches of each mycobacterial pentamer to the host proteome indicates the similarity score that can range from no matches to hundreds of matches. Following previous experimental data, a pentapeptide fragment that has about five (or fewer) perfect matches to the host proteome was considered to be a low-similarity sequence, that is, a rare fragment [17–28].

3. Results

3.1. Low-Similarity Sequences and B-Cell Epitopes. We examined the linear *M. tuberculosis*-derived B-cell epitopes cataloged in the IEDB [30], using pentapeptide probes as described above and report the similarity scores (Table 1).

As shown in Table 1, most of the mycobacterial B-cell epitopes contain pentapeptides that are rare or absent in the host proteins. Furthermore, preliminary screening revealed that a few B-cell epitopes were also T-cell epitopes (Table 1, footnote 5) [36, 49–52].

The relationship between low-similarity sequences and mycobacterial epitopes was further confirmed by extending the similarity analysis to the COOH domain of the mycobacterial DNAK protein. The DNAK chaperon protein is a member of the highly conserved heat shock protein family and is a protein antigen involved in the immune response to mycobacteria [53, 54]. In particular, the DNAK carboxy-terminus domain is highly immunogenic [41, 55] and hosts several of the epitopes described in the IEDB database (Table 1) [41]. Thus, the COOH domain of the mycobacterial DNAK protein was suitable for verifying the similarity-immunogenicity link. In brief, the primary sequence of the DNAK COOH-terminus domain (aa 500–609) was dissected into pentamers that were probed against the human proteome by the computer-assisted similarity analysis described in the Methods section. Then, the epitopic sequences were allocated along the mycobacterial-versus-human similarity profile.

Figure 1 shows the similarity profile of the DNAK COOH-terminus protein to the human proteome and the localization of the linear epitopes along the antigenic sequence. It can be seen that the carboxy terminal sequence of the mycobacterial chaperon protein has pentapeptides occurring frequently in the human proteome as well as unique fragments. Five pentamers (i.e., TWRIGY, WRIGY, VTGHW, TGHWR, and GHWRC) were unique to the DNAK COOH-terminus domain and these accounted for 5% of the total DNAK pentamers. The remaining 100 pentapeptides were repeated at different extent throughout the human proteome and occurred a total of 2468 times. Furthermore, and most importantly, Figure 1 shows that the peptide epitopes recognized by the human sera, as demonstrated by Elsaghier et al. [41], were among the mycobacterial peptide fragments that had a low similarity to the human proteome. That is, the unique pentapeptide signatures of the DNAK carboxy-terminus domain coincided with the epitopic motifs involved in immune recognition, indicating a direct relationship between low similarity and the humoural antibody response. Conversely, *M. tuberculosis* peptides not responsive to the human sera had a high number of pentamer matches to the human proteome (Figure 1, black arrows numbered from 6 to 15).

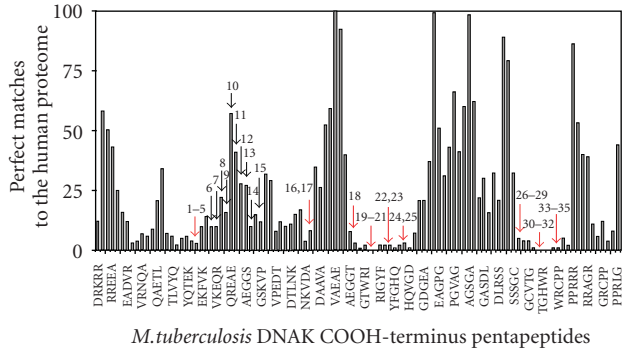
Table 1: Similarity analysis of *M. tuberculosis* B-cell epitopes to the host proteome.

Epitope Source Name	IEDB ID[1]	Epitope Sequence[2]	Number of Matches[3]	Host Proteome[4]	Refs.
10 kDa chaperonin	3293	wdedGEKRIpldvae	4	H	[12, 36]
10 kDa chaperonin	6508	PDTAKekpqegtvva[5]	3	H	[12, 36]
10 kDa chaperonin	3294	ldvaegdtvIYSKY[5]	2	H	[12, 36]
10 kDa chaperonin	1011756	gtvvavgpGRWDEdge[5]	3	H	[37–39]
10 kDa chaperonin	1011758	GRWDEdgekripldva	3	H	[37–39]
10 kDa chaperonin	3290	etttasglviPDTAK[5]	3	H	[12, 36, 38]
10 kDa chaperonin	3289	dkilvQANEAettta	1	H	[12, 36]
6 kDa ESAT[5]	8105	yqgvQQKWD[5]	1	M	[40]
6 kDa ESAT	8086	eQQWNFagieaaa[5]	0	M	[40]
6 kDa ESAT	8090	DEGKQsltk[5]	2	M	[40]
6 kDa ESAT	8106	aaAWGGSgsEAYQGvQQKWData[5]	3, 1, 1	M	[40]
DNAK	1052559	lvyQTEKF	4	H	[41]
DNAK	1052560	vyQTEKFv	4	H	[41]
DNAK	1052558	yQTEKFvk	4	H	[41]
DNAK	1052561	qTEKFVke	3	H	[41]
DNAK	1052562	TEKFVkeq	3	H	[41]
DNAK	1052583	lNKVDAav	4	H	[41]
DNAK	1052584	NKVDAava	4	H	[41]
DNAK	1052589	aeaEGGTW	3	H	[41]
DNAK	1052591	aegGTWRI	2	H	[41]
DNAK	1052593	ggTWRIGY	0	H	[41]
DNAK	1052594	WRIGYfgh	0	H	[41]
DNAK	1052595	riGYFGHq	2	H	[41]
DNAK	1052596	igYFGHQv	1	H	[41]
DNAK	1052597	GYFGHqvg	2	H	[41]
DNAK	1052598	YFGHQvgd	1	H	[41]
DNAK	24658	HQVGDgea	1	H	[41]
DNAK	1052627	lrsSSGCV	5	H	[41]
DNAK	1052628	rssSGCVT	4	H	[41]
DNAK	1052629	sssGCVTG	4	H	[41]
DNAK	1052630	ssgCVTGH	1	H	[41]
DNAK	1052631	cvTGHWRc	0	H	[41]
DNAK	1052632	VTGHWrcp	0	H	[41]
DNAK	1052633	TGHWRCpp	0	H	[41]
DNAK	1052549	HWRCPprr	1	H	[41]
DNAK	1052550	WRCPPrrr	1	H	[41]
ESAT-6-like protein ESXB	6090	laqeagNFERIsgdl[5]	2	H	[42]
ESAT-6-like protein ESXB	6090	laqeagNFERIsgdl	1	M	[42]
Immunogenic protein MPT64	1043665	nitsATYQSaipprgtqavv	1, 0	H	[43]
Immunogenic protein MPT64	1043668	hptttyKAFDWdqayrkpit	0	H	[43]
Lipoprotein LPQH	1034449	vtgsvvCTTAAgnvniaigg[5]	2	M	[44]
Lipoprotein LPQH	1107625	vtGSVVC[5]	5	M	[44]
Lipoprotein LPQH	1107627	VNKSFei	1	M	[44]
Lipoprotein LPQH	1052656	VKRGLtvavagaailvagls	2	M	[45]
Lipoprotein LPQH	1034455	taspgaasgpkvvidGKDQN	4	M	[44]
Lipoprotein LPQH	1018364	dMANPMspvnksfeievtcs	0	M	[44]
MTP40 protein	1013654	gprlyGEMTMqgtrkprpsgp	2	H	[46]
MTP40 protein	1011740	ttlGMHCGsfgsapsng	0	H	[46]
MTP40 protein	1011751	mlgtGTPNRarINFNC	3, 3	H	[46]
MTP40 protein	1011736	MLGNApsvvpnTTLGM	4, 3	H	[46]
MTP40 protein	1011754	INFNCevWSNVSetisgprly	3, 2	H	[46]

Table 1: Continued.

Epitope Source Name	IEDB ID[1]	Epitope Sequence[2]	Number of Matches[3]	Host Proteome[4]	Refs.
Phosphate-binding protein 1	1127065	lfnlWGPAFherypnvtita	1	M	[44]
Phosphate-binding protein 1	1019382	kqdpegWGKSPgfgttvdfp	1	M	[47]
Phosphate-binding protein 1	1107628	GPAFHer	1	M	[44]
Phosphate-binding protein 1	1123943	gnggMVTGCaetpgcvayig	1	M	[48]
Phosphate-binding protein 1	1107630	gfasKTPANqais	1	M	[44]

[1]IEDB ID Reference Dataset of Mycobacterial Immune Epitopes: http://iedb.zendesk.com/entries/18171-reference-datasets-of-mycobacterial-immune-epitopes.
[2] Epitope sequence with amino acid sequence in one-letter code. Low-similarity pentapeptides in capital letters.
[3]Number of times the low-similarity pentapeptide fragment(s) occur(s) in the set of proteins that comprehensively form the host proteome (see under Methods and [17–27]).
[4]Host proteome: H, human; M, murine.
[5]Sequences entirely or partially contained in *M. tuberculosis* T-cell epitopes [36, 49–52].
[6]ESAT, early secretory antigenic target.



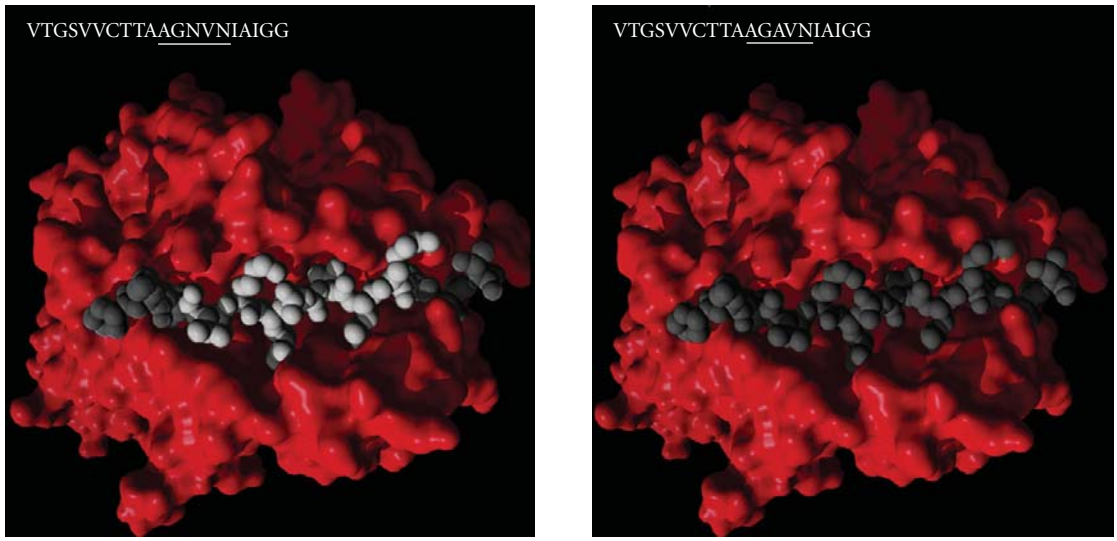
M.tuberculosis DNAK COOH-terminus pentapeptides

Figure 1: Similarity profile of the mycobacterial DNAK carboxy-terminus domain versus the human proteome at the pentapeptide level. The columns indicate the number of DNAK pentapeptide occurrences in the human proteome. DNAK corresponds to hsp70 protein, UniProt accession: P0A5B9; length: 609 aa. The numbered arrows indicate the location of mycobacterial peptide sequences immunoassayed with human sera [41]. Black arrows refer to the immuno-negative peptide sequences corresponding to IEDB ID: (6) 1212577 (aa FVKEQREA); (7) 1212579 (aa VKEQREAE); (8) 1212581 (aa KEQREAEG); (9) 1212582 (aa EQREAEGG); (10) 1212585 (aa QREAEGGS); (11) 1212586 (aa REAEGGSK); (12) 1212589 (aa EAEGGSKV); (13) 1212590 (aa AEGGSKVP); (14) 1212593 (aa EGGSKVPE); (15) 1212595 (aa GGSKVPED). Red arrows refer to the positive epitopic sequences corresponding to IEDB ID: (1) 1052559; (2) 1052560; (3) 1052558; (4) 1052561; (5) 1052562; (16) 1052583; (17) 1052584; (18) 1052589; (19) 1052591; (20) 1052593; (21) 1052594; (22) 1052595; (23) 1052596; (24) 1052597; (25) 1052598; (26) 24658; (27) 1052627; (28) 1052628; (29) 1052629; (30) 1052630; (31) 1052631; (32) 1052632; (33) 1052633; (34) 1052549; (35) 1052550. See Table 1 for amino acid sequence details. IEDB ID Reference Dataset of Mycobacterial Immune Epitopes: http://iedb.zendesk.com/entries/18171-reference-datasets-of-mycobacterial-immune-epitopes.

3.2. Low-Similarity Sequences and T-Cell Epitopes. The low-similarity *M. tuberculosis* B-cell epitopes hosted T-cell epitopes [36, 49–52] in several cases (Table 1). These data suggest that low-similarity influences the T- as well as the B-cell immune response, a finding that might be used in defining T-cell epitopes. Characterization of T-cell epitopes is complicated by different sets of rules governing the immune response. The problems include, but are not limited to, epitope hierarchy, immunoprevalence and immunodominance [56], variable epitopic length, the epitope major histocompatibility complex (MHC) binding potential and the difficulty of correctly predicting MHC binding activities in the majority of the test-set peptides [57], and the degenerate T-cell receptor recognition [22].

In this intricate context, the low-similarity hypothesis may provide insight for the structural characterization of immunogenic T-cell epitopic sequences. We know that the optimal peptide length for MHC binding is between nine and fifteen amino acids, and the central 5-6 residues contribute the majority of the specific contacts [59]. Thus, as previously proposed [19], the structural features that define an immunogenic T-cell epitope may be characterized by the MHC binding terminal residues flanking a central low-similarity core formed by 5-6 residues. This model is illustrated in Figure 2 using an immunodominant mycobacterial T-cell epitope validated and described by Ivanyi et al. [58]. The immunodominant p61–80/PT19 mycobacterial epitope (VTGSVVCTTAAGNVNIAIGG) has a critical core formed by five residues (AGNVN, aa 71–75): a single amino acid substitution with N73A impairs T-cell immunogenicity of the target p61–80/PT19 epitope. Our similarity analysis reveals that this substitution causes a shift in the proteomic similarity level of the five core residues (aa 71–75) that changes from one match (aa sequence AGNVN) to eight matches (aa sequence AGAVN) and "hides" the 5-mer core from the immune system.

The model illustrated in Figure 2 is supported by reports that the HFMPT pentapeptide is a minimal antigenic determinant for MHC class I-restricted T lymphocytes [60], and the KYVKQ pentapeptide is a minimal antigenic determinant for CD4(+) T-cell clones [61]. Moreover, the proteomic similarity values for HFMPT and KYVKQ were no matches and one match, respectively.



(a) (b)

Figure 2: The immunogenic impact of a single amino acid substitution in the five core residues (AGNVN, aa 71–75) of the immunodominant p61–80/PT19 mycobacterial epitope (VTGSVVCTTAAGNVNIAIGG). The substitution of N73A impairs T immunogenicity for the target epitope and changes the proteomic similarity level of the five core residues (aa 71–75 underlined) from one match (aa sequence AGNVN) to eight matches (aa sequence AGAVN). (a) The immunogenic MHC-peptide complex containing the central low-similarity pentapeptide AGNVN (in white); (b) The nonimmunogenic MHC-peptide complex containing the higher similarity pentapeptide AGAVN (in gray). Immunoreactivity and mutagenesis experiments are described in [58]. The figures are modified versions of the original drawings from the website http://www.iayork.com/ by kind authorization of Professor Ian York, University of Michigan.

4. Discussion

In this study, we explored the IEDB resources to define and characterize immunogenic mycobacterial epitopes that could be used for effective anti-TB immunotherapy. We observed that the immunological information carried by the *M. tuberculosis* antigens was localized in rare peptide fragments, confirming the relationship between low-similarity and immunogenicity. Sequences containing pentapeptide fragment(s) with low similarity to the host proteome appear to be those involved in the humoral antibody recognition of the mycobacterial antigen and may also influence T-cell immunoreactivity too.

Our findings provide a method for investigating the immune potential of the mycobacterial proteome, and thus, may help elucidate *M. tuberculosis* immunobiology, one of the primary challenges in current tuberculosis research. This is an important issue, particularly because the epitopic characterization of mycobacterial antigens advances slowly, antigen-by-antigen [12–14].

Moreover, mycobacterial antigenic motifs rarely found in host proteins are more likely to serve as potential immunogenic antigens. Use of short peptide modules rather than full-length mycobacterial antigens in vaccines may increase specificity and efficacy. It has been understood since the 1980s that chemically synthesized small peptides can induce antibodies that react with intact proteins [62]. Furthermore, peptide-based vaccines have been successfully used against several pathologies and infectious diseases. For example, tumour antigen-derived peptides evoked potent antitumour immunity in the murine melanoma M-3 [63]; cytotoxic CD4+ and CD8+ T lymphocytes generated by mutant p21-ras (12Val) peptide vaccination selectively killed autologous tumour cells carrying this mutation [64]; Her-2/neu peptide (aa 657–665) is an immunogenic epitope of Her-2/neu oncoprotein with potent antitumour properties [65]; synthetic peptides identified as antigenic sites on the S1 subunit of pertussis toxin induced especially high antibody titres against native pertussis toxin in mice [66]; a linear peptide containing minimal T- and B-cell epitopes of *Plasmodium falciparum* circumsporozoite protein provided protection against a transgenic sporozoite challenge [67]. These findings indicate that the use of exact immunogenic mycobacterial peptide sequences may provide the most effective active and passive immunotherapeutic anti-TB approaches. In addition, a major obstacle to antibody-based anti-TB therapy is the risk of adverse effects that range from *Lupus vulgaris* to granulomatous hepatitis [68–72], possibly caused by cross-reactivity [73]. The use of low-similarity peptides may allow the development of effective vaccines without adverse side-effects [74]. In conclusion, the findings of the present study may provide guidance in the analysis, identification, and utilization of the B- and T-cell response in vaccination against mycobacterial and other infectious diseases.

Acknowledgments

This study was made possible by the free access to the Immune Epitope Database at the National Health Institutes,

Bethesda, MD. We thank Professor Ian York, University of Michigan, for allowing us to use his figures. We are indebted to the Referees for precious comments and suggestions. Of course, we remain responsible for any remaining errors.

References

[1] E. L. Corbett, C. J. Watt, N. Walker, et al., "The growing burden of tuberculosis: global trends and interactions with the HIV epidemic," *Archives of Internal Medicine*, vol. 163, no. 9, pp. 1009–1021, 2003.

[2] The World Health Organization Report, *Global Tuberculosis Control*, WHO, Geneva, Switzerland, 2009.

[3] P. E. M. Fine, "BCG: the challenge continues," *Scandinavian Journal of Infectious Diseases*, vol. 33, no. 4, pp. 243–245, 2001.

[4] K. Huygen, "DNA vaccines against mycobacterial diseases," *Future Microbiology*, vol. 1, pp. 63–73, 2006.

[5] A. Tanghe, J.-P. Dangy, G. Pluschke, and K. Huygen, "Improved protective efficacy of a species-specific DNA vaccine encoding mycolyl-transferase Ag85A from *Mycobacterium ulcerans* by homologous protein boosting," *PLoS Neglected Tropical Diseases*, vol. 2, no. 3, article e199, 2008.

[6] D. B. Meya and K. P. McAdam, "The TB pandemic: an old problem seeking new solutions," *Journal of Internal Medicine*, vol. 261, no. 4, pp. 309–329, 2007.

[7] V. Nissapatorn, I. Kuppusamy, F. P. Josephine, I. Jamaiah, M. Rohela, and A. Khairul Anuar, "Tuberculosis: a resurgent disease in immunosuppressed patients," *The Southeast Asian Journal of Tropical Medicine and Public Health*, vol. 37, supplement 3, pp. 153–160, 2006.

[8] F. Abebe and G. Bjune, "The protective role of antibody responses during *Mycobacterium tuberculosis* infection," *Clinical and Experimental Immunology*, vol. 157, no. 2, pp. 235–243, 2009.

[9] Y. Xu, B. Zhu, Q. Wang, et al., "Recombinant BCG coexpressing Ag85B, ESAT-6 and mouse-IFN-γ confers effective protection against *Mycobacterium tuberculosis* in C57BL/6 mice," *FEMS Immunology and Medical Microbiology*, vol. 51, no. 3, pp. 480–487, 2007.

[10] Y. Q. Qie, J. L. Wang, B. D. Zhu, et al., "Evaluation of a new recombinant BCG which contains mycobacterial antigen ag85B-mpt64190-198-mtb8.4 in C57/BL6 mice," *Scandinavian Journal of Immunology*, vol. 67, no. 2, pp. 133–139, 2008.

[11] S. Chang-Hong, W. Xiao-Wu, Z. Hai, Z. Ting-Fen, W. Li-Mei, and X. Zhi-Kai, "Immune responses and protective efficacy of the gene vaccine expressing Ag85B and ESAT6 fusion protein from *Mycobacterium tuberculosis*," *DNA and Cell Biology*, vol. 27, no. 4, pp. 199–207, 2008.

[12] R. Hussain, H. M. Dockrell, and T. J. Chiang, "Dominant recognition of a cross-reactive B-cell epitope in *Mycobacterium leprae* 10 K antigen by immunoglobulin G1 antibodies across the disease spectrum in leprosy," *Immunology*, vol. 96, no. 4, pp. 620–627, 1999.

[13] K. K. Singh, N. Sharma, D. Vargas, et al., "Peptides of a novel *Mycobacterium tuberculosis*-specific cell wall protein for immunodiagnosis of tuberculosis," *Journal of Infectious Diseases*, vol. 200, no. 4, pp. 571–581, 2009.

[14] L. Baassi, K. Sadki, F. Seghrouchni, et al., "Evaluation of a multi-antigen test based on B-cell epitope peptides for the serodiagnosis of pulmonary tuberculosis," *International Journal of Tuberculosis and Lung Disease*, vol. 13, no. 7, pp. 848–854, 2009.

[15] U. Gowthaman and J. N. Agrewala, "In silico tools for predicting peptides binding to HLA-class II molecules: more confusion than conclusion," *Journal of Proteome Research*, vol. 7, no. 1, pp. 154–163, 2008.

[16] M. J. Blythe and D. R. Flower, "Benchmarking B cell epitope prediction: underperformance of existing methods," *Protein Science*, vol. 14, no. 1, pp. 246–248, 2005.

[17] A. Mittelman, A. Lucchese, A. A. Sinha, and D. Kanduc, "Monoclonal and polyclonal humoral immune response to EC HER-2/neu peptides with low similarity to the host's proteome," *International Journal of Cancer*, vol. 98, no. 5, pp. 741–747, 2002.

[18] A. Lucchese, A. Mittelman, M.-S. Lin, D. Kanduc, and A. A. Sinha, "Epitope definition by proteomic similarity analysis: identification of the linear determinant of the anti-Dsg3 MAb 5H10," *Journal of Translational Medicine*, vol. 2, article 43, 2004.

[19] R. Tiwari, J. Geliebter, A. Lucchese, A. Mittelman, and D. Kanduc, "Computational peptide dissection of Melan-a/MART-1 oncoprotein antigenicity," *Peptides*, vol. 25, no. 11, pp. 1865–1871, 2004.

[20] A. Mittelman, R. Tiwari, G. Lucchese, J. Willers, R. Dummer, and D. Kanduc, "Identification of monoclonal anti-HMW-MAA antibody linear peptide epitope by proteomic database mining," *Journal of Investigative Dermatology*, vol. 123, no. 4, pp. 670–675, 2004.

[21] R. Dummer, A. Mittelman, F. P. Fanizzi, G. Lucchese, J. Willers, and D. Kanduc, "Non-self-discrimination as a driving concept in the identification of an immunodominant HMW-MAA epitopic peptide sequence by autoantibodies from melanoma cancer patients," *International Journal of Cancer*, vol. 111, no. 5, pp. 720–726, 2004.

[22] A. Lucchese, J. Willers, A. Mittelman, D. Kanduc, and R. Dummer, "Proteomic scan for tyrosinase peptide antigenic pattern in vitiligo and melanoma: role of sequence similarity and HLA-DR1 affinity," *Journal of Immunology*, vol. 175, no. 10, pp. 7009–7020, 2005.

[23] J. Willers, A. Lucchese, A. Mittelman, R. Dummer, and D. Kanduc, "Definition of anti-tyrosinase MAb T311 linear determinant by proteome-based similarity analysis," *Experimental Dermatology*, vol. 14, no. 7, pp. 543–550, 2005.

[24] A. Stufano and D. Kanduc, "Proteome-based epitopic peptide scanning along PSA," *Experimental and Molecular Pathology*, vol. 86, no. 1, pp. 36–40, 2009.

[25] D. Kanduc, A. Lucchese, and A. Mittelman, "Individuation of monoclonal anti-HPV16 E7 antibody linear peptide epitope by computational biology," *Peptides*, vol. 22, no. 12, pp. 1981–1985, 2001.

[26] D. Kanduc, L. Tessitore, G. Lucchese, A. Kusalik, E. Farber, and F. M. Marincola, "Sequence uniqueness and sequence variability as modulating factors of human anti-HCV humoral immune response," *Cancer Immunology, Immunotherapy*, vol. 57, no. 8, pp. 1215–1223, 2008.

[27] G. Lucchese, A. Stufano, and D. Kanduc, "Proteome-guided search for influenza A B-cell epitopes," *FEMS Immunology and Medical Microbiology*, vol. 57, no. 1, pp. 88–92, 2009.

[28] D. Kanduc, ""Self-nonself" peptides in the design of vaccines," *Current Pharmaceutical Design*, vol. 15, no. 28, pp. 3283–3289, 2009.

[29] B. Peters, J. Sidney, P. Bourne, et al., "The immune epitope database and analysis resource: from vision to blueprint," *PLoS Biology*, vol. 3, no. 3, article e91, 2005.

[30] M. J. Blythe, Q. Zhang, K. Vaughan, et al., "An analysis of the epitope knowledge related to Mycobacteria," *Immunome Research*, vol. 3, no. 1, article 10, 2007.

[31] P. Chakhaiyar, Y. Nagalakshmi, B. Aruna, K. J. R. Murthy, V. M. Katoch, and S. E. Hasnain, "Regions of high antigenicity within the hypothetical PPE major polymorphic tandem repeat open-reading frame, Rv2608, show a differential humoral response and a low T cell response in various categories of patients with tuberculosis," *Journal of Infectious Diseases*, vol. 190, no. 7, pp. 1237–1244, 2004.

[32] A. C. W. May, "Percent sequence identity: the need to be explicit," *Structure*, vol. 12, no. 5, pp. 737–738, 2004.

[33] M. B. A. Oldstone, "Molecular mimicry and immune-mediated diseases," *FASEB Journal*, vol. 12, no. 13, pp. 1255–1265, 1998.

[34] G. Lucchese, A. Stufano, B. Trost, A. Kusalik, and D. Kanduc, "Peptidology: short amino acid modules in cell biology and immunology," *Amino Acids*, vol. 33, no. 4, pp. 703–707, 2007.

[35] C. W. Wu, L.-S. Yeh, H. Huang, et al., "The protein information resource," *Nucleic Acids Research*, vol. 31, no. 1, pp. 345–347, 2003.

[36] R. Hussain, F. Shahid, S. Zafar, M. Dojki, and H. M. Dockrell, "Immune profiling of leprosy and tuberculosis patients to 15-mer peptides of *Mycobacterium leprae* and *M. tuberculosis* GroES in a BCG vaccinated area: implications for development of vaccine and diagnostic reagents," *Immunology*, vol. 111, no. 4, pp. 462–471, 2004.

[37] B. Chua-Intra, S. Peerapakorn, N. Davey, et al., "T-cell recognition of mycobacterial GroES peptides in Thai leprosy patients and contacts," *Infection and Immunity*, vol. 66, no. 10, pp. 4903–4909, 1998.

[38] B. Chua-Intra, J. Ivanyi, A. Hills, J. Thole, C. Moreno, and H. M. Vordermeier, "Predominant recognition of species-specific determinants of the GroES homologues from *Mycobacterium leprae* and *M. tuberculosis*," *Immunology*, vol. 93, no. 1, pp. 64–72, 1998.

[39] B. Chua-Intra, R. J. Wilkinson, and J. Ivanyi, "Selective T-cell recognition of the N-terminal peptide of GroES in tuberculosis," *Infection and Immunity*, vol. 70, no. 3, pp. 1645–1647, 2002.

[40] M. Harboe, A. S. Malin, H. S. Dockrell, et al., "B-cell epitopes and quantification of the ESAT-6 protein of *Mycobacterium tuberculosis*," *Infection and Immunity*, vol. 66, no. 2, pp. 717–723, 1998.

[41] A. Elsaghier, R. Lathigra, and J. Ivanyi, "Localisation of linear epitopes at the carboxy-terminal end of the mycobacterial 71 kDa heat shock protein," *Molecular Immunology*, vol. 29, no. 9, pp. 1153–1156, 1992.

[42] J. S. Spencer, H. J. Kim, A. M. Marques, et al., "Comparative analysis of B- and T-cell epitopes of *Mycobacterium leprae* and *Mycobacterium tuberculosis* culture filtrate protein 10," *Infection and Immunity*, vol. 72, no. 6, pp. 3161–3170, 2004.

[43] P. W. Roche, C. G. Feng, and W. J. Britton, "Human T-cell epitopes on the *Mycobacterium tuberculosis* secreted protein MPT64," *Scandinavian Journal of Immunology*, vol. 43, no. 6, pp. 662–670, 1996.

[44] D. P. Harris, H.-M. Vordermeier, A. Arya, K. Bogdan, C. Moreno, and J. Ivanyi, "Immunogenicity of peptides for B cells is not impaired by overlapping T-cell epitope topology," *Immunology*, vol. 88, no. 3, pp. 348–354, 1996.

[45] K. R. Ashbridge, B. T. Bäckström, H.-X. Liu, et al., "Mapping of T helper cell epitopes by using peptides spanning the 19-kDa protein of *Mycobacterium tuberculosis*: evidence for unique and shared epitopes in the stimulation of antibody and delayed-type hypersensitivity responses," *Journal of Immunology*, vol. 148, no. 7, pp. 2248–2255, 1992.

[46] J. C. Falla, C. A. Parra, M. Mendoza, et al., "Identification of B- and T-cell epitopes within the MTP40 protein of *Mycobacterium tuberculosis* and their correlation with the disease course," *Infection and Immunity*, vol. 59, no. 7, pp. 2265–2273, 1991.

[47] K. A. L. De Smet, H. M. Vordermeier, and J. Ivanyi, "A versatile system for the production of recombinant chimeric peptides," *Journal of Immunological Methods*, vol. 177, no. 1-2, pp. 243–250, 1994.

[48] H.-M. Vordermeier, D. P. Harris, C. Moreno, M. Singh, and J. Ivanyi, "The nature of the immunogen determines the specificity of antibodies and T cells to selected peptides of the 38 kDa mycobacterial antigen," *International Immunology*, vol. 7, no. 4, pp. 559–566, 1995.

[49] Y. López-Vidal, S. P. de León-Rosales, M. Castañón-Arreola, M. S. Rangel-Frausto, E. Meléndez-Herrada, and E. Sada-Díaz, "Response of IFN-γ and IgG to ESAT-6 and 38 kDa recombinant proteins and their peptides from *Mycobacterium tuberculosis* in tuberculosis patients and asymptomatic household contacts may indicate possible early-stage infection in the latter," *Archives of Medical Research*, vol. 35, no. 4, pp. 308–317, 2004.

[50] H. Shams, P. Klucar, S. E. Weis, et al., "Characterization of a *Mycobacterium tuberculosis* peptide that is recognized by human $CD4^+$ and $CD8^+$ T cells in the context of multiple HLA alleles," *Journal of Immunology*, vol. 173, no. 3, pp. 1966–1977, 2004.

[51] A. S. Mustafa, F. A. Shaban, R. Al-Attiyah, et al., "Human Th1 cell lines recognize the *Mycobacterium tuberculosis* ESAT-6 antigen and its peptides in association with frequently expressed HLA class II molecules," *Scandinavian Journal of Immunology*, vol. 57, no. 2, pp. 125–134, 2003.

[52] D. P. Harris, H. M. Vordermeier, G. Friscia, et al., "Genetically permissive recognition of adjacent epitopes from the 19-kDa antigen of *Mycobacterium tuberculosis* by human and murine T cells," *Journal of Immunology*, vol. 150, no. 11, pp. 5041–5050, 1993.

[53] D. B. Young, "The immune response to mycobacterial heat shock proteins," *Autoimmunity*, vol. 7, no. 4, pp. 237–244, 1990.

[54] T. M. Shinnick, "Heat shock proteins as antigens of bacterial and parasitic pathogens," *Current Topics in Microbiology and Immunology*, vol. 167, pp. 145–160, 1991.

[55] D. B. Young and T. R. Garbe, "Heat shock proteins and antigens of *Mycobacterium tuberculosis*," *Infection and Immunity*, vol. 59, no. 9, pp. 3086–3093, 1991.

[56] C. Oseroff, B. Peters, V. Pasquetto, et al., "Dissociation between epitope hierarchy and immunoprevalence in CD8 responses to vaccinia virus western reserve," *Journal of Immunology*, vol. 180, no. 11, pp. 7193–7202, 2008.

[57] D. A. Winkler and F. R. Burden, "Nonlinear predictive modeling of MHC class II-peptide binding using Bayesian neural networks," *Methods in Molecular Biology*, vol. 409, pp. 365–377, 2007.

[58] D. P. Harris, M. Hill, H.-M. Vordermeier, et al., "Mutagenesis of an immunodominant Y cell epitope can affect recognition of different T and B determinants within the same antigen," *Molecular Immunology*, vol. 34, no. 4, pp. 315–322, 1997.

[59] J. B. Rothbard and M. L. Gefter, "Interactions between immunogenic peptides and MHC proteins," *Annual Review of Immunology*, vol. 9, pp. 527–565, 1991.

[60] M. J. Reddehase, J. B. Rothbard, and U. H. Koszinowski, "A pentapeptide as minimal antigenic determinant for MHC class I-restricted T lymphocytes," *Nature*, vol. 337, no. 6208, pp. 651–653, 1989.

[61] B. Hemmer, T. Kondo, B. Gran, et al., "Minimal peptide length requirements for $CD4^+$ T cell clones—implications for molecular mimicry and T cell survival," *International Immunology*, vol. 12, no. 3, pp. 375–383, 2000.

[62] H. L. Niman, R. A. Houghten, L. E. Walker, et al., "Generation of protein-reactive antibodies by short peptides is an event of high frequency: implications for the structural basis of immune recognition," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 80, no. 16, pp. 4949–4953, 1983.

[63] W. Schmidt, M. Buschle, W. Zauner, et al., "Cell-free tumor antigen peptide-based cancer vaccines," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 94, no. 7, pp. 3262–3267, 1997.

[64] M. K. Gjertsen, J. Bjorheim, I. Saeterdal, J. Myklebust, and G. Gaudernack, "Cytotoxic $CD4^+$ and $CD8^+$ T lymphocytes, generated by mutant p21-ras (12Val) peptide vaccination of a patient, recognize 12Val-dependent nested epitopes present within the vaccine peptide and kill autologous tumour cells carrying this mutation," *International Journal of Cancer*, vol. 72, no. 5, pp. 784–790, 1997.

[65] A. D. Gritzapis, A. Fridman, S. A. Perez, et al., "HER-2/neu (657-665) represents an immunogenic epitope of HER-2/neu oncoprotein with potent antitumor properties," *Vaccine*, vol. 28, no. 1, pp. 162–170, 2009.

[66] P. Askelöf, K. Rodmalm, G. Wrangsell, et al., "Protective immunogenicity of two synthetic peptides selected from the amino acid sequence of *Bordetella pertussis* toxin subunit S1," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 4, pp. 1347–1351, 1990.

[67] J. M. Calvo-Calle, G. A. Oliveira, C. O. Watta, J. Soverow, C. Parra-Lopez, and E. H. Nardin, "A linear peptide containing minimal T- and B-cell epitopes of *Plasmodium falciparum* circumsporozoite protein elicits protection against transgenic sporozoite challenge," *Infection and Immunity*, vol. 74, no. 12, pp. 6929–6939, 2006.

[68] K. Farsinejad, M. Daneshpazhooh, H. Sairafi, M. Barzegar, and M. Mortazavizadeh, "*Lupus vulgaris* at the site of BCG vaccination: report of three cases," *Clinical and Experimental Dermatology*, vol. 34, no. 5, pp. e167–e169, 2009.

[69] M. Enserink, "Public health: in the HIV era, an old TB vaccine causes new problems," *Science*, vol. 318, no. 5853, p. 1059, 2007.

[70] E. Attia, "BCG vaccine-induced lupus vulgaris," *European Journal of Dermatology*, vol. 17, no. 6, pp. 547–548, 2007.

[71] A. Hristea, A. Neacsu, D. A. Ion, A. Streinu-Cercel, and F. Stăniceanu, "BCG-related granulomatous hepatitis," *Pneumologia*, vol. 56, no. 1, pp. 32–34, 2007.

[72] A. Mandavilli, "When the vaccine causes disease," *Nature Medicine*, vol. 13, no. 3, p. 274, 2007.

[73] A. Spratt, T. Key, and A. J. Vivian, "Chronic anterior uveitis following bacille Calmette-Guérin vaccination: molecular mimicry in action?" *Journal of Pediatric Ophthalmology and Strabismus*, vol. 45, no. 4, pp. 252–253, 2008.

[74] D. Kanduc, "Epitopic peptides with low similarity to the host proteome: towards biological therapies without side effects," *Expert Opinion on Biological Therapy*, vol. 9, no. 1, pp. 45–53, 2009.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 961752, 7 pages
doi:10.1155/2010/961752

Methodology Report

A Method for Individualizing the Prediction of Immunogenicity of Protein Vaccines and Biologic Therapeutics: Individualized T Cell Epitope Measure (iTEM)

Tobias Cohen,[1] Leonard Moise,[1,2] Matthew Ardito,[1] William Martin,[1] and Anne S. De Groot[1,2,3]

[1] *EpiVax, 146 Clifford Street, Providence, RI 02903, USA*
[2] *Institute for Immunology and Informatics, University of Rhode Island, 80 Washington Street, Providence, RI 02903, USA*
[3] *Alpert Medical School, Brown University, Providence, RI 02903, USA*

Correspondence should be addressed to Anne S. De Groot, annied@epivax.com

Received 25 January 2010; Revised 21 March 2010; Accepted 19 April 2010

Academic Editor: Yongqun Oliver He

The promise of pharmacogenomics depends on advancing predictive medicine. To address this need in the area of immunology, we developed the individualized T cell epitope measure (iTEM) tool to estimate an individual's T cell response to a protein antigen based on HLA binding predictions. In this study, we validated prospective iTEM predictions using data from in vitro and in vivo studies. We used a mathematical formula that converts DRB1* allele binding predictions generated by EpiMatrix, an epitope-mapping tool, into an allele-specific scoring system. We then demonstrated that iTEM can be used to define an HLA binding threshold above which immune response is likely and below which immune response is likely to be absent. iTEM's predictive power was strongest when the immune response is focused, such as in subunit vaccination and administration of protein therapeutics. iTEM may be a useful tool for clinical trial design and preclinical evaluation of vaccines and protein therapeutics.

1. Introduction

Peptide binding to HLA (MHC) is the critical first step required for a T cell response. HLA binding enables antigen presenting cells to engage T cells via the T cell receptor to initiate a cascade of events that stimulate proinflammatory responses [1, 2]. Indeed, one of the most critical determinants of protein immunogenicity is the strength of peptide binding to MHC molecules [3]. Binding of antigenic peptides to HLA is desirable for vaccine design because immunogenic antigens produce protective T cell and antibody responses. However, the same interaction is often undesired in the context of biologic drug therapies, such as monoclonal antibodies and replacement proteins, because neutralizing antibodies raised against the therapy lower drug efficacy. In several cases, immune responses to proteins administered as drugs or vaccines have been linked to a particular HLA allele, which is better able to bind peptides derived from the antigen [4–6]. Consequently, the ability to predict this relationship might be useful in clinical trial design; for example, subjects who carry specific HLA alleles could be excluded from a protein therapeutic trial. We set out to develop a statistical analysis tool, individualized T cell Epitope Measure (iTEM), which estimates the likelihood that a particular antigen will generate an immune response for a specific subject. As shown in the five case studies reported here, iTEM can be used as a benchmark to determine whether or not an individual subject is likely to respond to a given epitope or subunit protein. We conclude that iTEM scores can be used as a binary test, with a threshold over which a peptide or protein is likely to bind an individual's HLA and could potentially trigger an immune response, and below which a response is unlikely.

2. Methods

2.1. iTEM Calculations. To calculate an iTEM score we first identify putative HLA ligands and T cell epitope clusters using the EpiMatrix system [7, 8]. Input amino acid sequences are parsed into overlapping 9-mer frames. Each

frame is then evaluated for binding potential against a panel of eight common Class II alleles (DRB1*0101, DRB1*0301, DRB1*0401, DRB1*0701, DRB1*0801, DRB1*1101, DRB1*1301, and DRB1*1501) [9]. We call each frame-by-allele evaluation an EpiMatrix "assessment". EpiMatrix raw scores are normalized and reported on a "Z" scale. Assessments with a score of at least 1.64, theoretically speaking the top 5% of any given sample, are considered highly likely to bind HLA and are called "hits" [10]. The resulting dataset is then screened for regions containing more hits than we would expect to find by chance alone. For regions with a high density of hits an EpiMatrix Cluster Score is calculated. To calculate an EpiMatrix Cluster Score we sum the scores of all the hits contained in a given cluster and deduct the sum of scores we would expect to find in a randomly generated sequence of similar length. In other words the EpiMatrix Cluster Score is the deviation between the observed EpiMatrix sum of scores and the expected EpiMatrix sum of scores. The expected sum of scores can be calculated as the number of 9-mer frames contained in the target sequence times the number of alleles screened against times .05 (the expected hit rate) times 2.06 (the expected value of a Z score above 1.64).

T cell epitope clusters are promiscuous but they are not universal, and human APCs present only two DR alleles. We have observed that certain peptides stimulate immune response in some subjects better than others. In order to explain part of this observed variation we have developed the iTEM Score. iTEM scores are a special case of the EpiMatrix Cluster Score. iTEM scores describe the relationship between a particular patient's HLA haplotype (considering only two HLA-DR alleles) and the amino acid sequence of a given epitope cluster. iTEM scores are used to predict the likelihood that the amino acid sequence of an antigenic peptide will be presented by a given subject's antigen presenting cells and in turn stimulate that subject's T cells. To calculate an iTEM score for a given individual we calculate an EpiMatrix Cluster Score for each HLA allele in the haplotype. Allele-specific cluster scores of less than zero are discarded (literally set to zero), and the two allele specific cluster scores are then added together to form an iTEM score. Negative allele specific cluster scores are discarded because the binding relationship between a given peptide and a given allele is independent of the relationship between that peptide and another allele. In other words the failure of one allele to present a given peptide does not negatively affect the relationship between that peptide and any other allele, and therefore we felt it would be wrong to allow negative allele specific cluster scores to detract from accompanying positive scores. Higher iTEM scores indicate an increased likelihood of immunogenicity.

An example of an EpiMatrix report from which an iTEM score can be calculated is shown in Figure 1.

2.2. Experimental Data Sources. The six case studies reported here are based on immunogenicity measurements made in enzyme-linked immunospot (ELISpot) assays that measure interferon-gamma secretion from antigen-stimulated peripheral blood mononuclear cells (PBMCs) from humans or splenocytes from mice. Four involve vaccine candidate studies for *Mycobacterium tuberculosis* [11], *Variola major* [12], *and Francisella tularensis* [13, 14]; one involves a type 1 diabetes (T1D) autoantigen [15] and finally an angiogenesis inhibiting protein therapeutic known as FPX [16]. ELISpot responses were considered positive if the number of spots detected was greater than 50 spots per one million cells over background (1 response over background per 20,000 cells).

2.3. Correlation of iTEM Scores and ELISpot Results. Antigens from each study were grouped into one of four categories based on overall iTEM score and ELISpot result. True positives are peptide-HLA pairs with both positive iTEM scores and ELISpot results. True negatives are peptide-HLA pairs with both negative iTEM scores and ELISpot results. False positives have positive iTEM scores and negative ELISpot results, and false negatives, have negative iTEM scores and positive ELISpot results. The cutoff for a positive iTEM score was initially set at 2.06, as described above.

Analyses were rerun with a higher cutoff that we hoped would adjust for factors of immunogenicity not predicted by EpiMatrix since peptides that are more likely to bind HLA would probably be less likely to be affected by other factors during antigen processing and presentation. We arbitrarily decided on 2.5 as the higher cutoff. Standard chi-squared and linear regression analysis between iTEM scores and ELISpot results were performed by hand or using Microsoft Excel 2003, respectively. Statistical significance was defined as $P < .05$. In linear regression analysis, the intercept was constrained at zero because peptides that do not bind HLA cannot stimulate T cells above background levels. Humans bearing HLA DRB1* alleles not predicted by EpiMatrix were excluded from this study.

3. Results

3.1. Case Study 1: Tuberculosis Vaccine. In the tuberculosis vaccine study, two groups of six HLA A2/DR1 transgenic mice were immunized intranasally once with a pool of 25 peptides at 2 μg/peptide and CpG oligodeoxynucleotide 1826 inside liposomes in 50 μL, twice two weeks apart. A third group received the same set of injections except that there were only 16 peptides in the pool. Two weeks after the second immunization, the mice were sacrificed and lymphocytes were isolated from the spleens. Cells were plated on an IFN-γ ELISpot plate at 200,000 cells/well and stimulated with individual peptides in triplicate at 10 μg/mL for two days. Peptides that induced an average spot count of 50 spots/million cells over background were considered to have generated a positive response in this assay.

There were 66 peptides tested in the DR1 mice, whose splenocytes were pooled, and therefore there were 66 peptide-HLA pairs for which we calculated iTEM scores. In the peptide immunization study, when the iTEM threshold was 2.06, there were 54 positives and 12 negatives. Of the 54 positives, 38 (70%) were ELISpot positive, and of the 12 negatives, 11 (92%) were ELISpot negative. Using chi-squared analysis, we found the association between iTEM score and ELISpot result to be statistically significant ($X^2 = 16.79$, df = 1, $P < .001$). Although the linear regression

Frame start	AA sequence	Frame stop	Hydrophobicity	DRB1*0101 Z-score	DRB1*0301 Z-score	DRB1*0401 Z-score	DRB1*0701 Z-score	DRB1*0801 Z-score	DRB1*1101 Z-score	DRB1*1301 Z-score	DRB1*1501 Z-score	Hits
1	VDVFKLWLM	9	1.38								1.44	0
2	DVFKLWLMW	10	0.81					1.47				0
3	VFKLWLMWR	11	0.7								1.29	0
4	FKLWLMWRA	12	0.43			1.32			1.63			0
5	KLWLMWRAK	13	−0.31									0
6	LWLMWRAKG	14	0.08	1.98	1.67			2.29	2.8	1.7	1.34	5
7	WLMWRAKGT	15	−0.42	1.77		2.22	2.58		1.66			4
8	LMWRAKGTT	16	−0.4					1.37				0
9	MWRAKGTTG	17	−0.87	1.69	1.35	1.53				1.31	1.55	1
10	WRAKGTTGF	18	−0.77				1.66	2.05		1.33		2
11	RAKGTTGFE	19	−1.06									0
12	AKGTTGFEA	20	−0.36									0
13	KGTTGFEAH	21	−0.91									0

HLA Allele	DRB1*0101	DRB1*0301	DRB1*0401	DRB1*0701	DRB1*0801	DRB1*1101	DRB1*1301	DRB1*1501
Sum of Significant Z scores	5.44	1.67	2.22	4.24	4.34	4.46	1.7	0

0701 iTEM = 4.24 − (13*.0.05*2.06) = 2.9
1101 iTEM = 4.46 − (13*.0.05*2.06) = 3.12
0701/1101 iTEM = 2.9 + 3.12 = 6.02

Figure 1: Calculating an iTEM score. Using the EpiMatrix report for a peptide, the iTEM score is equal to the difference between the sum of significant scores for an allele (shown in the black boxes above) and the expected score for a peptide of that length. Assessments outside the top 10% (Z-scores below 1.28) are hidden for ease of viewing the more significant scores, and assessments are shaded with different levels of darkness to highlight Z-scores that are in the top 5% (1.64–2.32) or top 1% (greater than 2.32). For a subject with two alleles, the iTEM score for each allele is calculated and then added together, as shown in the three equations within the figure. The peptide constant is equal to the product of the number of frames (13 in this example), the expected frequency of hits (0.05), and the expected value for a hit (2.06). For this peptide, an HLA of DRB1*0701/DRB1*1101 would be expected to respond (iTEM score: 6.02).

did not predict the number of SFC given the iTEM score ($R^2 = 0.39$), the analysis yielded a statistically significant positive slope of 39.35 ± 6.12 ($P < .001$).

Characteristics of this study are summarized in Table 1, and the results are summarized in Table 2.

3.2. Case Study 2: Smallpox Vaccine. In the smallpox vaccine study, HLA-DR1 or DR3 transgenic mice were immunized using a DNA-prime/peptide-boost strategy. Mice were injected intramuscularly, twice, two weeks apart, with 100 μg of a plasmid DNA vaccine bearing a multiepitope gene. Two weeks following the second DNA injection, mice received 50 μg of peptides corresponding to the epitopes contained in the DNA vaccine via a subcutaneous injection of peptides in liposomes delivered intranasally, twice, two weeks apart. Two weeks after the final immunization, mice were sacrificed and spleens removed for splenocyte preparation. Cells were transferred to an IFN-γ ELISpot plate at 200,000 cells/well and stimulated with individual peptides in triplicate at 10 μg/mL for two days. Peptides that induced an average spot count of 50 spots/million cells over background were considered positive.

DR1 mice were immunized with 32 different peptides while DR3 mice were immunized with 41 different peptides; therefore, there were 73 peptide-HLA pairs for which we calculated iTEM scores. When the iTEM threshold was set to 2.5, 42 positives and 31 negatives were calculated. Of the 42 positives, 16 (38%) had positive ELISpot results and of the 31 negatives, 26 (84%) had negative ELISpot results. Using chi-squared analysis, we found the association between iTEM score and ELISpot result to be statistically significant ($X^2 = 4.20$, df = 1, $P < .05$). Again, there was no association between the SFC and iTEM score ($R^2 = 0.20$), but the linear regression analysis yielded a statistically significant positive slope of 23.92 ± 5.58 ($P < .001$).

Characteristics of this study are summarized in Table 1, and the results are summarized in Table 2.

3.3. Case Study 3: Tularemia Antigenicity. The Tularemia antigenicity study involved 23 *F. tularensis* survivors. PBMCs were isolated from whole blood and cultured for 5–20 days with 10 μg/mL of pooled putative T cell epitope peptides with IL-2 and IL-7 added every other day until 3 days before IFN-γ ELISpot assay. Cells were then transferred to ELISpots plates at 200,000 cells/well and stimulated with individual peptides or pools at 10 μg/mL in triplicate. Peptides that induced average spot counts greater than 50/million cells and more than double the negative control wells were considered positive responses.

Due to varying cell yields, PMBCs from all 23 subjects were not exposed to all 27 peptides, and only 510 peptide-HLA pairs were tested via ELISpot. Since EpiMatrix only predicts for certain HLA alleles, there were only 232 peptide-HLA pairs for which an iTEM score could be calculated. Because the study consisted of human subjects who had been naturally infected with tularemia, there was no cutoff to generate a statistically significant association between iTEM score and ELISpot result due to a very large number of false positives. Using the 2.5 iTEM cutoff, there were 181 positives and 51 negatives. Of the 181 positives, 34 (19%) had

Table 1: Design of six different studies used to validate the iTEM predictions. In three cases, inbred HLA transgenic mice were used for measurements of vaccine immunogenicity. In the other three, human subjects were recruited for evaluations of immune responses to protein antigens.

Study	Subject	Immunization	Boost
Tularemia	HLA Transgenic Mice	DNA vector IN	Peptides IN
Smallpox	HLA Transgenic Mice	DNA vector IM	Peptides SC
Tuberculosis	HLA Transgenic Mice	Peptide IN	Peptides
T1D	Humans	Natural Autoimmunity	Peptides ex vivo
FPX	Humans	Protein IV or SC	None
Tularemia	Humans	Natural Infection	None

Table 2: A summary of the results using the threshold model for iTEM's association with immunogenicity listed in descending NPV for an iTEM threshold of 2.06.

		iTEM Threshold = 2.06			iTEM Threshold = 2.5		
Study	n	P-value	PPV (%)	NPV (%)	P-value	PPV (%)	NPV (%)
Human Tularemia	232	<.02	19	95	<.10	19	92
Tuberculosis	66	<.001	70	92	<.001	71	80
Smallpox	73	<.10	36	85	<.05	38	84
FPX	30	<.001	94	71	<.001	94	71
Mouse Tularemia	78	<.10	61	63	<.10	61	63
T1D	56	<.01	84	54	<.05	83	44

positive ELISpot results and of the 51 negatives, 47 (95%) had negative ELISpot results ($X^2 = 3.48$, df = 1, $P < .10$). We could not perform linear regression analysis because the values of the negative spot counts were not reported.

Characteristics of this study are summarized in Table 1, and the results are summarized in Table 2.

3.4. Case Study 4: Tularemia Vaccine. In the tularemia vaccine study, six HLA-DR1 transgenic mice were immunized three times, at two-week intervals, with 25 μg in 50 μL intratracheally with a plasmid DNA vaccine encoding a multiepitope gene. Two weeks after the final DNA immunization, mice were immunized intratracheally, twice, two weeks apart, with 50 μg of peptides, corresponding to the epitopes contained in the vaccine construct, formulated in liposomes together with CpG oligodeoxynucleotide. Two weeks following the final immunization, mice were sacrificed and their spleens removed for splenocyte isolation. Cells were transferred to an IFN-γ ELISpot plate at 200,000 cells/well, and stimulated with individual peptides in triplicate at 10 μg/mL for two days. Peptides that induced an average spot count of 50 spots/million cells over background were considered positive.

The DNA plasmid administered to the DR1 mice contained 78 different peptide sequences; therefore, 78 peptide-HLA pairs were tested via ELISpot. There were 18 positives and 60 negatives using an iTEM score cutoff of 2.5. Of the 18 positives, 11 (61%) had positive ELISpot results and of the 60 negatives, 38 (63%) had negative ELISpot results. The large number of false negatives yielded a statistically insignificant association between iTEM score and ELISpot results ($X^2 = 3.39$, df = 1, $P < 0.10$). Even though the regression did not predict the number of SFC given the iTEM score ($R^2 = 0.41$), a linear regression analysis yielded a statistically significant positive slope of 163.43 ± 22.35 ($P < .001$).

Characteristics of this study are summarized in Table 1, and the results are summarized in Table 2.

3.5. Case Study 5: Immunogenicity of an Autoantigen. We investigated the autoantigenicity of glutamic acid decarboxylase (GAD65) T cell epitopes in diabetic patients. PBMCs, isolated from whole blood samples of six human subjects, were cultured for seven days with a pool of 14 GAD65 peptides, with IL-2 and IL-7 added to the growth medium every two days. Cells were transferred to an IFN-γ ELISpot plate at 200,000 cells/well, and stimulated with individual peptides in triplicate at 10 μg/mL for two days. Peptides that induced an average spot count of 50 spots/million cells over background were considered positive.

Due to varying cell yields, PMBC from all six subjects were not exposed to all 14 peptides, and only 67 peptide-HLA pairs were tested via ELISpot; however, since EpiMatrix only predicts for certain HLA alleles, there were only 56 peptide-HLA pairs for which an iTEM score could be calculated. While this may seem different from the immunogenicity studies carried out above, in T1D, which is mechanically similar to a peptide immunization, a single autoantigen is considered to be the target of autoreactive T cells and therefore an iTEM score of 2.06 had the most predictive power, with 43 positives and 13 negatives. Of the 43 positives, 36 (84%) had positive ELISpot results, and of the 13 negatives, 7 (54%) had negative ELISpot results. Using chi-squared analysis, we found the association between iTEM score and ELISpot result was statistically significant

($X^2 = 7.51$, df = 1, $P < .01$). Again, performing a regression did not yield a clear link between the number of SFC and iTEM score ($R^2 = 0.38$), linear regression analysis yielded a statistically significant positive slope of 25.57 ± 4.42 ($P < .001$).

Characteristics of this study are summarized in Table 1, and the results are summarized in Table 2.

3.6. Case Study 6: Immunogenicity of Protein Therapeutic. In the FPX immunogenicity study, 36 human subjects received FPX peptides intravenously and 40 others received FPX subcutaneously. PBMCs from 15 subjects were isolated from whole blood samples on day 1 before dosing and then again on days 42 and 180. Cells were cultured for 7 days with 10 μg/mL of individual FPX T cell epitope peptides (three different peptides) and then transferred to ELISpot plates to detect levels of IFN-γ and IL-4 upon restimulation with 10 μg/mL of test peptide at 200,000 cells/well. Peptides that induced an average spot count of 50 spots/million cells over background were considered positive.

45 peptide-HLA pairs were tested via ELISpot; however, since EpiMatrix only predicts for certain HLA alleles, there were only 30 peptide-HLA pairs for which an iTEM score could be calculated. Since this was a peptide immunization study, an iTEM threshold of 2.06 was used. There were 16 positives and 14 negatives according to iTEM scores. Of the 16 positives, 15 (94%) had positive ELISpot responses, and of the 14 negatives, 10 (71%) had negative ELISpot results. Using chi-squared analysis, we found the association between iTEM score and ELISpot result was statistically significant ($X^2 = 13.66$, df = 1, $P < .001$). Linear regression analysis yielded a statistically significant positive slope of 70.37 ± 11.17 ($P < .001$); however, the regression could not accurately predict the number of SFC given the iTEM score ($R^2 = 0.57$).

Characteristics of this study are summarized in Table 1, and the results are summarized in Table 2.

4. Discussion

In this study, we developed an immunoinformatic tool to predict the outcomes of experimental measurements of vaccination and immunogenicity analysis. Two methods for predicting the association between iTEM and ELISpot response gave significant correlations. These were the linear model and the threshold model.

Of the studies that could be fit to a linear model, the linear slopes derived were rather varied. The average of the five slopes was 64.51 ± 58.34. While the linear model did reach statistical significance in most of the data, its accuracy was moderate at best. The average R^2 value, which is a measurement of the accuracy of the linear model, was 0.39. The low R^2 values as well as the different slopes for each study lead us to reject the use of a linear equation to predict experimental results from in silico analysis. Given that the average negative predictive value (NPV) and positive predictive value (PPV) of all the data when a cutoff of 2.06 was used were 77% and 61%, respectively, we believe that the relationship between iTEM scores and experimental results is more accurately described by the threshold model where the magnitude of an immune response cannot be predicted beyond "positive" or "negative". In the future, it is possible that when iTEM may be able to account for more complex issues (e.g., the likelihood that an epitope will be processed) that a more linear correlation between iTEM scores and experimental results will exist.

iTEM scores correlated best with immunogenicity for studies in which a protein or small peptide was administered, but the correlation did not reach statistical significance with studies of immune responses to peptides following exposure to a pathogen. The strongest correlations were observed for the TB, T1D and FPX studies. In these studies, the average negative predictive value (NPV) was 72% and the average positive predictive value (PPV) was 83%, when the iTEM threshold was set at 2.06. Thus it appears that processing and antigen presentation do not introduce significant variation unaccounted for by EpiMatrix predictions. Since iTEM scores are generated only by examining specific peptide-HLA interactions, when these variables are minimized by using protein or peptide prime and boosts, the iTEM was a more accurate predictor of immune responses. While HLA presentation is necessary for immunogenicity, it is not sufficient. A number of factors related to the expression and processing of the antigen also influence immunogenicity. If the protein is not expressed or secreted by the pathogen during infection or not properly cleaved or transported by the host, it will not be immunogenic. Even with proper presentation, the peptide may be homologous to self, and therefore either the corresponding T cell has been deleted during thymic selection or T cells that respond to the peptide have been anergized in the periphery.

Adding a second step between immunization and exposure to the protein immunogen and antigen presentation lowered the predictive accuracy of iTEM for a data set. For example, in those studies where DNA vaccination was used, such as smallpox, the predictive accuracy of iTEM was not significant if the cutoff of 2.06 was used, but was improved by increasing the threshold to 2.5. This suggests that T cell responses to those peptides that are very likely to bind HLA as indicated by higher iTEM scores were more likely following DNA vaccination. Perhaps the addition of steps beyond HLA binding (such as protein processing following expression of the pseudoprotein from the DNA vaccine) can accidentally remove peptides otherwise capable of activating the immune system. Thus, in the smallpox vaccine study, iTEM only had a PPV of 38%; on the other hand, its NPV was 84%.

The relationship between immune response and T cell epitope is more complicated in studies where exposure of T cells to the antigen (natural infection studies) is affected by many factors including gene expression, protein processing, dose and route. Thus, for subjects exposed to *F. tularensis*, iTEM was not an accurate predictor of immune response to tularemia peptide in vitro. In this case, for example, the proteins from which the peptide epitopes were derived might not have been expressed during infection or the bacteria generated factors that downregulated HLA expression, thereby lowering the repertoire of HLA ligands [17].

iTEM's accuracy is greatest when used as a negative predictor of immune response, a feature which may be very useful for the interpretation of failed vaccine efficacy studies and for clinical trials of protein therapeutics. Using a threshold of 2.06, iTEM had an average NPV across the six case studies of nearly 80%, but a PPV of only 61%. A high NPV may be due to the fact that peptides that are unable to bind to HLA (as described by a low iTEM score) have a very small chance of being immunogenic. In contrast, a low PPV can be explained by factors affecting the protein's processing and presentation, despite its constituent epitopes, capacity to bind HLA.

4.1. Uses of iTEM in Vaccine Efficacy Studies. iTEM's ability to predict the absence of a T cell response suggests useful applications in vaccine design. iTEM analysis might be useful for vaccine studies to explain both interpeptide and intersubject variability in T cell assays and clinical trials. In theory, iTEM analysis could be used to predict whether or not a subject would fall into the 5%–20% of the population that does not respond to commonly used vaccines. If a lack of response is predicted, a patient could be exempt, thus sparing him/her from unnecessary vaccinations.

4.2. Application of iTEM to Studies of Autoimmune Diseases. Others have examined the role that an individual's HLA DR genotype plays in regulating immune responses [18]. Several autoimmune disorders have demonstrated an association with specific HLA types. T1D is associated with DR4-DQ8 and DR3-DQ2 [19], multiple sclerosis with DR15 [6], and ankylosing spondylitis and the other spondyloarthropathies are associated with B27 [20]. It would be interesting to consider whether the antigenic targets of certain autoimmune diseases, such as the acetylcholine receptor in myasthenia gravis, contain T cell epitopes that are more likely to be presented in the context of the HLA alleles that are associated with manifestation of autoimmunity, and less likely to be presented on alleles that have an inverse association with autoimmunity. Such a correlation would be evidence supporting the use of iTEM to identify auto-antigens.

4.3. Application of iTEM to Clinical Studies of Protein Therapeutics. In the context of protein therapeutics, drug developers generally agree that T cell response, which is associated with the development of antidrug antibodies, is undesirable. The ability to predict that an immune response in a particular subject is likely would be useful for identifying individuals at higher risk of developing antidrug antibodies. These individuals could be excluded from clinical trials and/or advised to avoid the use of the protein therapeutic, increasing the safety of protein therapeutics for human use.

For example, in a prospective study, Koren et al identified selected HLA types that were associated with a stronger immune response against a novel antitumor therapeutic. These HLA alleles were associated with higher iTEM scores, higher T cell responses, and higher antibody responses in the Phase I trial [16]. The same HLA restriction effect can be seen in the response to IFN-β treatment in human subjects with MS, where DRB1*0701 was overrepresented in subjects with anti-IFN-β antibodies [6].

5. Conclusion

In summary, the iTEM tool appears to be a useful method for predicting the absence of T cell response to a given vaccine or protein therapeutic. While this study has only examined CD4 epitopes, we expect that a similar, and possibly stronger correlation may exist for CD8 epitopes given that the constraints imposed by the closed-end HLA Class I binding groove [21] make class I predictions inherently much more accurate than the class II predictions [22]. Further studies of immune response using the iTEM tool will be needed before it will be able to be adapted for use by drug developers. Tools such as iTEM are likely to play an important role in the development of safer vaccines and therapeutics in the future.

References

[1] E. S. Trombetta and I. Mellman, "Cell biology of antigen processing in vitro and in vivo," *Annual Review of Immunology*, vol. 23, pp. 975–1028, 2005.

[2] P. Cresswell and A. Lanzavecchia, "Antigen processing and recognition," *Current Opinion in Immunology*, vol. 13, no. 1, pp. 11–12, 2001.

[3] C. A. Lazarski, F. A. Chaves, S. A. Jenks et al., "The kinetic stability of MHC class II: peptide complexes is a key parameter that dictates immunodominance," *Immunity*, vol. 23, no. 1, pp. 29–40, 2005.

[4] C. A. Alper, M. S. Kruskali, D. Marcus-Bagley et al., "Genetic prediction of nonresponse to hepatitis B vaccine," *The New England Journal of Medicine*, vol. 321, no. 11, pp. 708–712, 1989.

[5] T. Höhler, E. Reuss, N. Evers et al., "Differential genetic determination of immune responsiveness to hepatitis B surface antigen and to hepatitis A virus: a vaccination study in twins," *The Lancet*, vol. 360, no. 9338, pp. 991–995, 2002.

[6] M. D. F. S. Barbosa, J. Vielmetter, S. Chu, D. D. Smith, and J. Jacinto, "Clinical link between MHC class II haplotype and interferon-beta (IFN-β) immunogenicity," *Clinical Immunology*, vol. 118, no. 1, pp. 42–50, 2006.

[7] A. S. De Groot, A. Bosma, N. Chinai et al., "From genome to vaccine: in silico predictions, ex vivo verification," *Vaccine*, vol. 19, no. 31, pp. 4385–4395, 2001.

[8] A. S. De Groot, B. M. Jesdale, E. Szu, J. R. Schafer, R. M. Chicz, and G. Deocampo, "An interactive web site providing major histocompatibility ligand predictions: application to HIV research," *AIDS Research and Human Retroviruses*, vol. 13, no. 7, pp. 529–531, 1997.

[9] S. Southwood, J. Sidney, A. Kondo et al., "Several common HLA-DR types share largely overlapping peptide binding repertoires," *Journal of Immunology*, vol. 160, no. 7, pp. 3363–3373, 1998.

[10] A. S. De Groot and W. Martin, "Reducing risk, improving outcomes: bioengineering less immunogenic protein therapeutics," *Clinical Immunology*, vol. 131, no. 2, pp. 189–201, 2009.

[11] L. Moise, et al., "Immunoinformatic-Discovered Mycobacterium Tuberculosis T-Cell Epitopes are Immunogenic in HLA Transgenic Mice," 2009.

[12] L. Moise, et al., "Breadth of T cell response but not antibody critical for protection following vaccination with VennVax,

a DNA-prime, peptide-boost multi-T-cell epitope vaccine," 2009.

[13] J. A. McMurry, S. H. Gregory, L. Moise, D. Rivera, S. Buus, and A. S. De Groot, "Diversity of Francisella tularensis Schu4 antigens recognized by T lymphocytes after natural infections in humans: identification of candidate epitopes for inclusion in a rationally designed tularemia vaccine," *Vaccine*, vol. 25, no. 16, pp. 3179–3191, 2007.

[14] L. Moise, et al., "Rapid Development and Validation of a T-Cell Epitope-Based Tularemia Vaccine for F. Tularensis," 2009.

[15] A. S. De Groot, et al., "Effect of "Tregitopes" on T1D immune response in vitro and on diabetes in NOD mice," 2009.

[16] E. Koren, A. S. De Groot, V. Jawa et al., "Clinical validation of the "in silico" prediction of immunogenicity of a human recombinant therapeutic protein," *Clinical Immunology*, vol. 124, no. 1, pp. 26–32, 2007.

[17] C. M. Bosio, H. Bielefeldt-Ohmann, and J. T. Belisle, "Active suppression of the pulmonary immune response by Francisella tularensis Schu4," *Journal of Immunology*, vol. 178, no. 7, pp. 4538–4547, 2007.

[18] G. T. Nepom and H. Erlich, "MHC class-II molecules and autoimmunity," *Annual Review of Immunology*, vol. 9, pp. 493–525, 1991.

[19] C. B. Sanjeevi, S. K. Sedimbi, M. Landin-Olsson, I. Kockum, and Å. Lernmark, "Risk conferred by HLA-DR and DQ for type 1 diabetes in 0-35-year age group in Sweden," *Annals of the New York Academy of Sciences*, vol. 1150, pp. 106–111, 2008.

[20] A. Mathieu, F. Paladini, A. Vacca, A. Cauli, M. T. Fiorillo, and R. Sorrentino, "The interplay between the geographic distribution of HLA-B27 alleles and their role in infectious and autoimmune diseases: a unifying hypothesis," *Autoimmunity Reviews*, vol. 8, no. 5, pp. 420–425, 2009.

[21] K. Murphy, P. Travers, and M. Walport, *Antigen Recognition by B-Cell and T-Cell Receptors, in Janeway's Immunobiology*, Garland Science, New York, NY, USA, 2008.

[22] M. Ardito, et al., "EpiMatrix: tool for accelerated epitope selection and vaccine design," in *Proceedings of the 2nd Annual Global Vaccine Congress*, Boston, Mass, USA, 2008.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 297505, 15 pages
doi:10.1155/2010/297505

Methodology Report

Vaxign: The First Web-Based Vaccine Design Program for Reverse Vaccinology and Applications for Vaccine Development

Yongqun He,[1,2,3] Zuoshuang Xiang,[1,2,3] and Harry L. T. Mobley[2]

[1] *Unit for Laboratory Animal Medicine, University of Michigan Medical School, Ann Arbor, MI 48109, USA*
[2] *Department of Microbiology and Immunology, University of Michigan Medical School, Ann Arbor, MI 48109, USA*
[3] *Center for Computational Medicine and Biology, University of Michigan Medical School, Ann Arbor, MI 48109, USA*

Correspondence should be addressed to Yongqun He, yongqunh@med.umich.edu

Received 2 November 2009; Accepted 6 May 2010

Academic Editor: Anne S. De Groot

Vaxign is the first web-based vaccine design system that predicts vaccine targets based on genome sequences using the strategy of reverse vaccinology. Predicted features in the Vaxign pipeline include protein subcellular location, transmembrane helices, adhesin probability, conservation to human and/or mouse proteins, sequence exclusion from genome(s) of nonpathogenic strain(s), and epitope binding to MHC class I and class II. The precomputed Vaxign database contains prediction of vaccine targets for > 70 genomes. Vaxign also performs dynamic vaccine target prediction based on input sequences. To demonstrate the utility of this program, the vaccine candidates against uropathogenic *Escherichia coli* (UPEC) were predicted using Vaxign and compared with various experimental studies. Our results indicate that Vaxign is an accurate and efficient vaccine design program.

1. Introduction

Reverse vaccinology is an emerging vaccine development approach that starts with the prediction of vaccine targets by bioinformatics analysis of microbial genome sequences [1]. Predicted proteins are selected based on desirable attributes. Normal wet laboratory experiments are conducted in a later stage to test all or selected vaccine targets. Rino Rappuoli, the pioneer of reverse vaccinology [1, 2], first applied this approach to the development of a vaccine against serogroup B *Neisseria meningitidis* (MenB), the major cause of sepsis and meningitis in children and young adults [2]. In this study, bioinformatic methods were first used to screen the complete genome of a MenB strain MC58 for genes encoding putative surface-exposed or secreted proteins. These proteins were predicted to be antigenic and therefore may represent the most suitable vaccine candidates. In total, 350 novel vaccine candidates were predicted and expressed in *Escherichia coli*; 28 were found to elicit protective immunity. It took less than 18 months to identify more vaccine candidates in MenB than had been discovered during the past 40 years by conventional methods [2]. Since then, the concept of reverse vaccinology has also successfully been applied to other pathogens, including *Bacillus anthracis* [3], *Porphyromonas gingivalis* [4], *Chlamydia pneumoniae* [5], *Streptococcus pneumoniae* [6], *Helicobacter pylori* [7], and *Mycobacterium tuberculosis* [8]. Compared to a conventional vaccine development approach that starts from the wet laboratory, reverse vaccinology begins with bioinformatics analysis, which dramatically quickens the process of vaccine development.

Since reverse vaccinology was conceived and applied in a test case ten years ago, this technology has progressed dramatically. Subcellular location is still considered as one main criterion for vaccine target prediction. However, more criteria have been added. For example, since it was found that outer membrane proteins containing more than one transmembrane helix were, in general, difficult to clone and purify [2], the number of transmembrane domains for a vaccine target is often considered in bioinformatics filtering. More and more genomes are now available for each pathogenic species. It is now required to examine all completed genomes and predict vaccine targets that are conserved in all genomes. If genomes from non-pathogenic strains of the species are

also available, ideal vaccine targets are those that exist in genomes of virulent pathogen strains but are absent from the avirulent strains. To induce strong immunity and avoid autoimmunity, predicted vaccine targets are required not to have sequence similarity to proteins of hosts (e.g., human). Epitope-based vaccines have been demonstrated to induce protection against many infectious diseases [9]. To optimize epitope vaccines, it has become an essential task to predict immune epitopes from protective antigens.

While reverse vaccinology has been used for a decade, this approach is often not accessible to the general laboratory, due to the lack of software programs that are easy to use and implement. Although many individual software programs are available to aid in vaccine target prediction [10–17], they are individually developed for different purposes and contain disparate data formats and programming settings. This makes tool and data integration difficult. Successful use of these tools often requires local installation, command line execution, and substantial computational power. Many tools are not optimized for high throughput data processing. NERVE, for example, is a new enhanced reverse vaccinology environment that includes several steps of programs for reverse vaccinology [18]. NERVE aims to help save time and money in vaccine design. However, it also requires software download and database setup. In addition, NERVE does not include precomputed data of vaccine target prediction, which makes the prediction time extensive. In addition, NERVE does not perform MHC class I and II epitope predictions.

Many immunoinformatics epitope mapping tools have been developed during the last three decades [19]. For example, DeLisi and Berzofsky developed the earliest computer-driven algorithm for epitope mapping based on empirical observations of amino acid residue periodicity in T-cell epitopes [20]. The anchor-based MHC binding motifs were used for T-cell epitope identification by many researchers, such as Sette et al. in 1989 [21] and Rotzschke et al. in 1991 [22]. Matrix-based approached for T-cell epitope mapping have been developed by a number of research teams such as Sette et al. [23], Davenport et al. [24], De Groot et al. [25], and Reche et al. [15]. Many databases of MHC-binding peptides, starting from MHCPEP developed by Brusic et al. in 1994 [26] to the currently frequently used IEDB [27], have been developed for use with matrices and neural network-based epitope prediction tools.

Uropathogenic *Escherichia coli* (UPEC) is the most common cause of community-acquired urinary tract infection (UTI). Over half (53%) of all women (and 14% of men) experience at least one urinary tract infection (UTI), leading to an estimated 6.8 million annual physician visits in the United States alone, 1.3 million emergency room visits, and 246,000 hospitalizations of women with an annual cost of more than $2.4 billion [28]. Although many groups have attempted to develop vaccines against UPEC [29–33], no preparations are yet in general use in the United States. Complete and annotated genomic sequences have now been determined for four strains of extraintestinal pathogenic *E. coli* including CFT073, UTI89, 536, and F11; these UPEC strains were isolated from human cases of cystitis, pyelonephritis, and/or bacteremia. These provide a basis for predicting UPEC vaccine targets using these genome sequences based on reverse vaccinology. Recently, we have also performed several high throughput proteomic and genomic studies including *in vivo* microarray [34], proteomics of urine-grown bacteria [35], and *in vivo* induced antigen technology (IVIAT) [36]. We hypothesized that vaccine targets predicted based on genome analysis largely correlate with the results obtained from these high throughput data analyses.

Vaxign (http://www.violinet.org/vaxign/), the first web-based, publically available vaccine design system, was first introduced in the second Vaccine Congress meeting in December 2008 in Boston, MA, USA. Vaxign was demonstrated to successfully predict vaccine targets against different pathogens [37]. Since then, Vaxign has significantly been improved in terms of performance and speed. In this report, we systematically introduce the updated Vaxign prediction system, and describe how Vaxign was used to predict vaccine targets against uropathogenic *E. coli* (UPEC). Many predicted results, based on genome sequence analyses, were also confirmed by wet-lab testing and other studies based on RNA, protein, and antibody analyses.

2. Methods

2.1. Vaxign Software Components for Vaccine Target Prediction. Vaxign integrates open source tools and internally developed programs with user-friendly web interfaces. Input data for Vaxign execution are amino acid sequences from one protein or whole genomes. This Vaxign pipeline includes the following components (Figure 1).

(1) Prediction of subcellular localization. Vaxign predicts different subcellular locations using optimized PSORTb 2.0 that has a measured overall precision of 96% [10].

(2) Transmembrane domain prediction. The transmembrane helix topology analysis is performed using optimized HMMTOP based on a general hidden Markov model (HMM) decoding algorithm [11]. A profile-based hidden Markov model implemented in PROFtmb is used in Vaxign for the prediction and discrimination of bacterial transmembrane beta barrels [38]. The resulting PROFtmb method reaches an overall four-state (up-, down-strand, periplasmic-,and outer-loop) accuracy as high as 86% [38]. Since the execution of PROTtmb is very time consuming, not all proteins in all genomes in the Vaxign database were preanalyzed for transmembrane beta barrel analysis.

(3) Calculation of adhesin probability. Adhesin probability is predicted using optimized SPAAN [12]. The SPAAN prediction has a sensitivity of 89% and specificity of 100% based on a defined test set [12]. The probability of being an adhesin has a default cut-off of 0.51.

(4) Protein conservation among different genomes. This program identifies conserved sequences among more than one genome. OrthoMCL is applied to calculate the homology between different sequences [13]. The E-value of 10^{-5} is set as the default value. An internally developed reciprocal best fit method, based on BLAST, was also developed for result comparison.

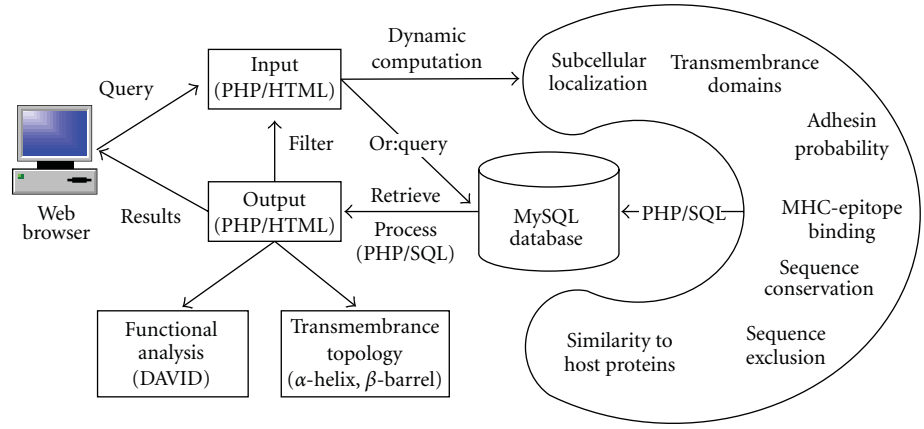


Figure 1: The Vaxign algorithm pipeline.

(5) Exclusion of sequences present in nonpathogenic strains. OrthoMCL is used to calculate the homology between predicted sequences and all proteins in a specified non-pathogenic strain genome(s) [13].

(6) Comparison of sequence similarity between predicted proteins and host (human and/or mouse) proteome. OrthoMCL is customized for this purpose.

(7) Prediction of MHC class I- and class II-binding epitopes. Vaxign uses an internally developed program Vaxitope to predict MHC class I and class II binding epitopes. Vaxitope is developed based on PSSM (Position Specific Scoring Matrix) motif prediction. The PSSMs for the prediction of peptide binders to MHC class I or II are calculated based on a position-based weighting method using the BLK2PSSM utility included in the BLIMPS package [14]. Data for generating the PSSMs came from known epitope data from the IEDB immune epitope database [27]. The P value for the predicted epitope binding to PSSMs is calculated by the MAST sequence homology search algorithm [39]. A receiver operating characteristic (ROC) curve and the values of the area under the ROC Curve (AUC) were used to calculate the accuracy of the Vaxitope prediction [40]. For the AUC analysis, the epitope data from the IEDB immune epitope database [27] were used. A leave-one-out approach was applied to test if a known epitope can be predicted on the condition that this epitope is excluded in initial generation of PSSMs.

(8) Protein functional analysis: Predicted proteins can be selected and automatically exported to the DAVID bioinformatics resources [41] for functional protein analysis.

2.2. Vaxign Server and Web Implementation. Vaxign is implemented using a three-tier architecture built on two Dell Poweredge 2580 servers which run the Redhat Linux operating system (Redhat Enterprise Linux ES 5). Users can submit database or analysis queries through the web. These queries are then processed using PHP/HTML/SQL (middle-tier, application server based on Apache) against a MySQL (version 5.0) relational database (back-end, database server), or executed in runtime based on the Vaxign algorithm pipeline. The result of each query is then presented to the user in the web browser (Figure 1). Two servers are scheduled to regularly backup each others' data.

2.3. Application of Vaxign in Prediction of UPEC Vaccine Targets. To predict vaccine targets against uropathogenic *E. coli* (UPEC) using Vaxign, four UPEC strains with fully sequenced genomes were used: strains CFT073 (RefSeq ID: NC_004431), 536 (NC_008253), UTI89 (NC_007946), and F11 (NZ_AAJU00000000). Microbial genomes and protein sequences were downloaded from NCBI RefSeq genome database [42]. To determine whether predicted antigens exist in UPEC strains but not in non-pathogenic *E. coli*, the non-pathogenic *E. coli* K-12 strain MG1655 (RefSeq ID: NC_000913) [43] was used as a control genome.

2.4. Comparison of Different Methods in UPEC Vaccine Target Prediction. The results of UPEC vaccine targets predicted by Vaxign were manually compared with results from our previous studies using microarray [34], proteomics [35], immunoproteomic analysis [36].

2.5. Verification of UPEC Vaccine Targets Predicted by Vaxign. To experimentally verify the predicted data, UPEC proteins were prepared using recombinant cloning technology. For active immunization, CBA/J mice ($N = 10$ for each group) were intranasally immunized with individual proteins combined with cholera toxin. As negative control, cholera toxin alone was also used to vaccinate mice. The vaccinated group were boosted at 7 and 14 days. One week after the final boost, control (naïve: Ctx-treated) and vaccinated mice were transurethrally challenged with 5×10^8 CFU *E. coli* CFT073. After a one-week, the efficacy of protection by individual subunit vaccines was evaluated by measuring the CFU/ml urine and CFU/g bladder or kidney tissue. The vaccine challenge experiments were reported in a recent publication [33].

3. Results

3.1. The Vaxign Algorithm for Vaccine Target Prediction. The workflow of the Vaxign pipeline is shown in Figure 1. The

Table 1: Pathogens currently analyzed by Vaxign.

species	# of Genomes	# of proteins
Bacterial species:		
(1) *Bacillus anthracis*	3	16182
(2) *Brucella*	9	28559
(3) *Campylobacter*	10	17443
(4) *Clostridium*	10	35130
(5) *Corynebacterium diphtheriae*	1	2272
(6) *Coxiella burnetii*	5	9686
(7) *Escherichia coli*	7	34592
(8) *Francisella*	9	14771
(9) *Haemophilus influenzae*	4	6735
(10) *Helicobacter pylori*	6	9261
(11) *Mycobacterium tuberculosis*	2	8178
(12) *Neisseria meningitidis*	4	7909
Virus strains:		
(13) HIV	2	33
(14) Measles virus	1	7
(15) Vaccinia virus	1	223
(16) Variola virus	1	197
(17) Yellow fever virus	1	14
Total #	76	191192

predicted features in Vaxign include protein subcellular location, transmembrane helices, adhesin probability, sequence conservation among pathogen genomes, and sequence similarity to host (human and mouse) proteomes. For those pathogens against which a strong B cell response (for antibody production) is critical, surface-exposed proteins such as secreted proteins and outer membrane proteins (especially adhesins) are ideal targets for vaccine development. For these pathogens, nonsurface proteins such as cytoplasmic or inner membrane proteins, however, may not represent good targets for vaccine development due to lack of close contact with the host cells [1, 2]. However, for the vaccine development against those pathogens where T cell response is critical, subcellular localization is not an issue since a T cell response could be directed to any protein target. It has been reported that 250 out of 600 vaccine candidates from *N. meningitidis* B failed to be cloned and expressed due to the presence of more than one transmembrane spanning region [2]. Therefore, it might also be prudent to ignore those proteins with multiple transmembrane spanning regions in the first place. The adherence of microbial pathogens to host cells is mediated by adhesins. Adhesins are essential for bacterial colonization and survival and represent possible targets for vaccine development. The conserved vaccine targets among different strains in one pathogen offer protection against these different strains. A vaccine candidate with similar sequence to the host (e.g., human or mouse) is likely to be a poor immunogen due to epitope mimicry, or if an immune response is triggered, cause autoimmunity in the host [44–46]. These aspects are considered in the Vaxign prediction pipeline (Figure 1).

During the past decades, many algorithms and software programs have been developed to address individual processes in the Vaxign vaccine design pipeline. Many software programs have been widely tested and validated. To avoid reinventing the wheel, we have incorporated many existing software programs into Vaxign as described in the Section 2. All open source programs (e.g., BLAST) have been customized. The Vaxitope (vaccine epitope prediction) is a new program that is internally developed and will be described later in this paper in more detail. One focus of the Vaxign development was to seamlessly incorporate different programs with different development styles and even program languages into a comprehensive analysis system. To achieve this goal, MySQL relational database was used to replace plain text input files typically used in original programs. In a typical scenario, output data of one program is stored in MySQL, and SQL query scripts are used to retrieve and process the data as input for another program. Each component program except Vaxitope in the Vaxign pipeline has individually been tested and validated in the literature [10–13]. The testing of Vaxitope is described below.

The Vaxign database contains precomputed prediction results using 76 genomes from 13 pathogens (Table 1). In total, 191,192 proteins have been precomputed. These data can be queried using the Vaxign web interface. A user can also input protein sequence data for dynamic computation and result output.

3.2. Vaxitope: Prediction of MHC Class I and Class II Binding Epitopes. Vaxign predicts both MHC class I and class II binding epitopes using an internally developed tool Vaxitope. Vaxitope is based on Position Specific Scoring Matrix (PSSM), a type of scoring matrix used in protein similarity searches in which amino acid substitution scores are given separately for each position in a protein multiple sequence alignment. In PSSM, a Tyr-Trp substitution at position A of an alignment may receive a very different score than the same substitution at position B. In contrast, in position-independent matrices such as the PAM and BLOSUM matrices, the Tyr-Trp substitution receives the same score no matter at what position it occurs. The general strategy of using PSSMs for prediction of MHC Class I and II binding has proven effective in RANKPEP [15].

To evaluate the performance of Vaxitope, a receiver operating characteristic (ROC) curve analysis was generated for prediction of epitopes against 40 MHC class I or II alleles (Table 2). The ROC analysis detects the ability of predictions to classify each predicted epitope peptide into MHC class I or II binding based on its comparison with existing epitope database [40]. Plotting the rates of true-positive predictions (sensitivity) as a function of the rate of false-positive predictions (1-specificity) gives an ROC curve. For example, a ROC curve based on Vaxign analysis was generated using HLA A*0201 specific PSSM (Figure 2). HLA A*0201 is one of the most studied HLA MHC Class I allele. According to the IEDB immune epitope database [27], 3216 epitopes are known to positively bind to this allele (as positive testing dataset), and 4826 epitopes cannot bind

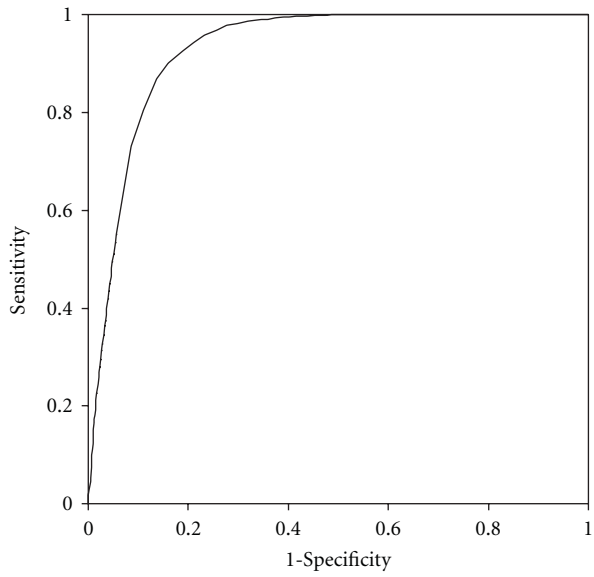


Figure 2: ROC curve analysis of epitopes binding HLA A*0201.

to this allele (as negative testing dataset). The positive HLA A*0201 alleles were used to calculate the True Positive Rate (Sensitivity). The negative alleles were used to calculate the False Positive Rate (1-Specificity) (Figure 2). The areas under the ROC curve (AUC) provide a way to measure prediction quality. An AUC of 0.5 represents random predictions, and an AUC of 1.0 indicates perfect predictions [16]. The value of the Area Under the ROC Curve (AUC) for the HLA A*0201 analysis using Vaxitope is 0.929. Our analysis of 30 alleles indicates that Vaxitope is a very specific and sensitive method for MHC Class I and II binding epitope prediction (Table 2).

It is interesting to compare Vaxign and RANKPEP since both methods are based on PSSM. If only AUC values are taken into account, our prediction results are in general better than the results predicted by RANKPEP [15]. However, the results may not be comparable, since the data required to generate PSSMs might be different. Different from RANKPEP, which uses a percentage or top number as the cut off as shown in RANKPEP [15], Vaxitope defines statistical *P*-values based on a random sequence model that assumes each position in a random sequence is generated according to the average letter frequencies of all sequences in the NCBI peptide non-redundant database [39]. Our studies indicate that the *P* value of .05 provides a cutoff with high and balanced sensitivity and specificity (Table 2). Another unique feature in Vaxitope is that it integrates with other vaccine design components in Vaxign. For example, the input sequence of Vaxitope may come from those peptides that are part of an outer membrane protein and exposed outside the bacterial membrane (Figure 3). These protein peptides are predicted by Vaxign and easily available as input data for Vaxitope. Vaxitope also allows genome-wide query on different MHC host species.

Traditional reverse vaccinology does not consider prediction of epitopes. With the *P* value cut off of .05, 1436 epitopes from *E. coli* protein Hma for 39 MHC Class I alleles in 4 hosts and 515 epitopes for 23 MHC Class II alleles have been found in 4 hosts—human, mouse, macaque, and chimpanzee. It remains a challenge to rank and optimize the epitopes for vaccine development. Possible solutions to address this challenge are described in the Discussion.

3.3. User-Friendly Vaxign Web Interface. To make Vaxign easy to use, two methods of implementation have been developed. Users can either directly query precomputed prediction results from the Vaxign database, or request Vaxign to dynamically calculate results based on the users' input sequences. The prediction data from the precomputed Vaxign database can be easily queried using our Vaxign web query interface (Figure 3).

A simple web query interface is available for querying the precomputed Vaxign results from the protein level or genome level (Figure 3). Users are prompted to set up preferred query criteria; the output data are then provided. The query of precomputed Vaxign results is fast. A typical query involved in four genomes and all the steps as shown in our UPEC use case (Figure 3) takes approximately 2–5 seconds.

The other form is dynamic Vaxign analysis, which is similar to the precomputed Vaxign except that a user is prompted to provide information for up to 300 proteins at one time. The protein information may be protein sequences using FASTA format, NCBI protein GI, or RefSeq accession number. Vaxign predicts vaccine targets based on runtime execution. It typically takes 30–60 seconds to execute all the steps in run time for one single protein. Therefore, it would take 150–300 minutes to finish analysis of 300 proteins. Once all steps are finished, the web link of the predicted results will be sent to a registered user through email.

3.4. Vaxign Predicts 22 Outer Membrane Proteins as UPEC Vaccine Targets. The genomes of all four UPEC strains (CFT073, 536, UT189, and F11) for which complete sequence data are available were analyzed by Vaxign (Figure 4). These four genomes contain 4704–5379 genes. Only outer membrane proteins (OMP) are predicted and analyzed. From the total 5379 proteins in UPEC strain CFT073, Vaxign detects 107 outer membrane proteins. Among the 107 proteins, three proteins contain more than one transmembrane helix. Vaxign further predicts 70 proteins from the 107 OMPs in strain CFT073 as possible adhesins or adhesin-like proteins [34]. These predicted adhesins are likely critical for colonization, a major challenge facing UPEC in the urinary tract. While some of these proteins, such as PapC [47], are adhesins, many of these 70 proteins (e.g., Hma, FepA) predicted to be adhesins are not typically considered as adhesins. The roles of these adhesin-like proteins in adhering to host cells require further investigation. None of these 70 proteins shows sequence similarity to any human or mouse proteins. Similar strategy was applied to obtain vaccine targets for the other three UPEC strains (Figure 4).

Ortholog analysis was then applied to obtain conserved vaccine targets from four UPEC strains. In total, 85 OMPs were found to be conserved across all four pathogenic UPEC

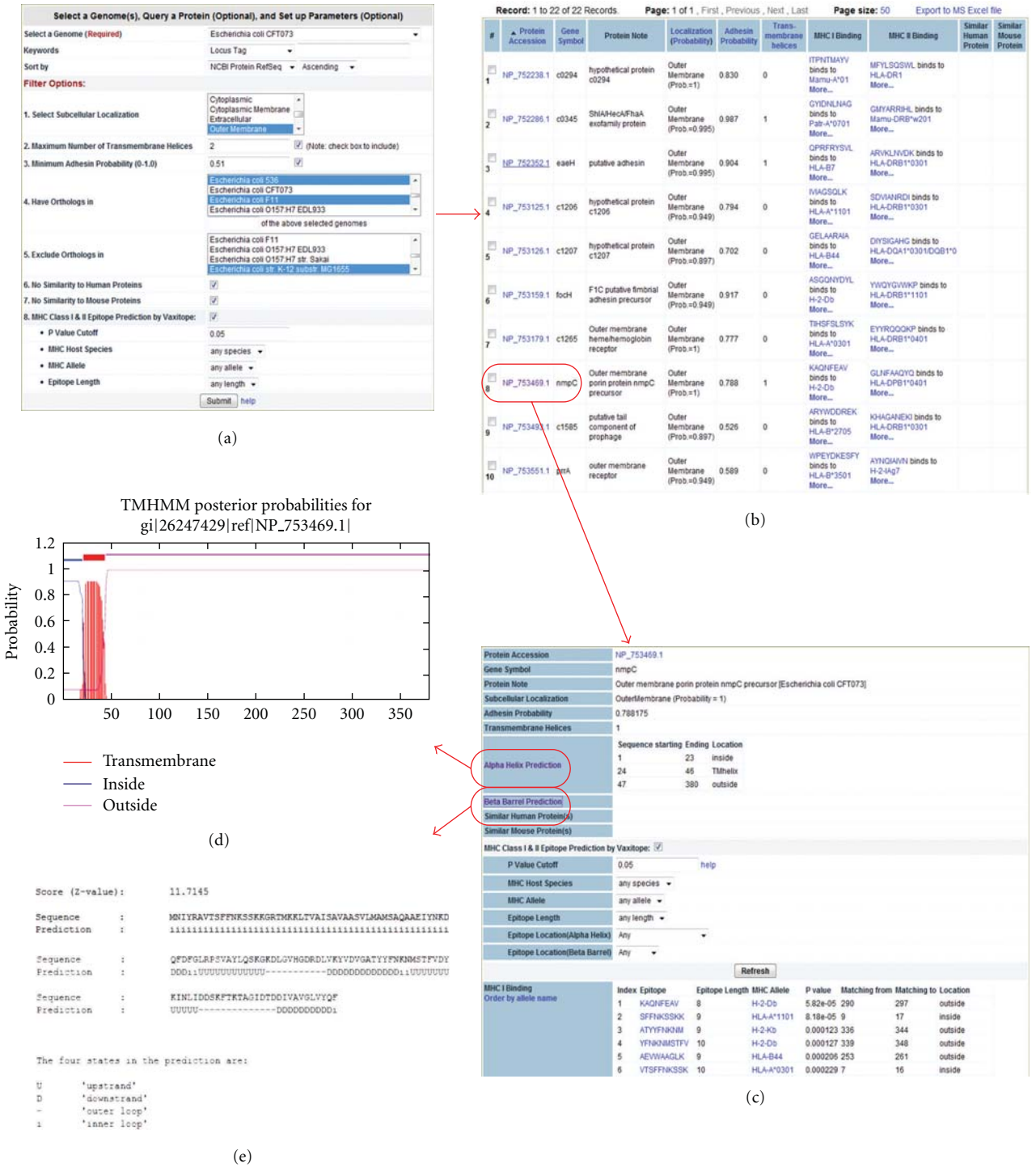


Figure 3: Vaxign prediction of UPEC vaccine targets. (a) Selection of *E. coli* CFT073 as a seed genome. Different filter options are applied: only outer membrane proteins, maximum of one transmembrane helix, minimum adhesin probability of.51, having orthologs in other three UPEC genomes but no orthologs in K-12 strain M1655, and no similarity to any human or mouse proteins. (b) In total 22 proteins were predicted to meet selected criteria. (c) Detailed results for a particular protein, for example, NmpC. MHC Class I and II epitope prediction results are also available and can be further filtered. (d) Prediction of α-helices and their possible cellular locations of NmpC. (e) Prediction of β-barrels and their possible cellular locations of NmpC.

Table 2: Epitope prediction performance by Vaxitope as measured by AUC values.

#	MHC allele	Length	AUC	Sensitivity ($P = .01$)	Specificity ($P = .01$)	Sensitivity ($P = .05$)	Specificity ($P = .05$)	Sensitivity ($P = .1$)	Specificity ($P = .1$)
1	HLA-A*0101	9	0.929	.854	.874	.99	.709	1	.621
2	HLA-A*0201	9	0.871	.298	.956	.531	.876	.792	.783
3	HLA-A*0201	10	0.913	.471	.957	.789	.874	.901	.773
4	HLA-A*0202	9	0.869	.309	.953	.658	.875	.792	.774
5	HLA-A*0202	10	0.863	.333	.949	.737	.808	.891	.684
6	HLA-A*0203	9	0.874	.304	.956	.659	.865	.828	.769
7	HLA-A*0203	10	0.867	.317	.963	.691	.827	.834	.712
8	HLA-A*0206	9	0.900	.387	.961	.73	.867	.881	.781
9	HLA-A*0206	10	0.916	.403	.957	.775	.878	.922	.782
10	HLA-A*0301	9	0.887	.445	.921	.855	.803	.959	.69
11	HLA-A*0301	10	0.868	.505	.9	.915	.706	.988	.554
12	HLA-A*1101	9	0.863	.337	.946	.672	.833	.859	.71
13	HLA-A*1101	10	0.879	.461	.924	.9	.742	.989	.627
14	HLA-A*2402	9	0.984	.727	.985	.97	.879	1	.783
15	HLA-A*3101	9	0.912	.426	.952	.813	.872	.927	.752
16	HLA-A*3101	10	0.855	.419	.905	.889	.711	.99	.563
17	HLA-A*3301	9	0.937	.495	.959	.94	.851	.989	.723
18	HLA-A*3301	10	0.905	.51	.942	.905	.755	.966	.619
19	HLA-A*6801	9	0.908	.406	.949	.841	.841	.946	.744
20	HLA-A*6801	10	0.848	.418	.9	.866	.701	.973	.564
21	HLA-A*6802	9	0.918	.446	.96	.801	.868	.922	.757
22	HLA-A*6802	10	0.913	.452	.947	.837	.825	.977	.715
23	HLA-A*6901	9	0.803	.279	.895	.674	.785	.837	.674
24	HLA-B*0702	9	0.963	.659	.966	.962	.894	.981	.849
25	HLA-B*1501	9	0.873	.514	.927	.816	.765	.939	.603
26	HLA-B*3501	9	0.838	.403	.927	.701	.775	.834	.666
27	HLA-B*5101	9	0.978	.835	.953	1	.871	1	.824
28	HLA-B*5301	9	0.989	.84	.981	.991	.896	1	.825
29	HLA-B*5801	9	0.923	.769	.933	.894	.813	.952	.702
30	H-2-Kb	8	0.936	.753	.922	.935	.744	.987	.623
31	H-2-IAd	—*	0.928	.582	.992	.705	.959	.82	.926
32	H-2-IEd	—	0.977	.903	1	.935	.935	.935	.887
33	H-2-IEg7	—	0.998	.993	.989	.993	.945	1	.893
34	H-2-IEk	—	0.940	.775	.95	.875	.9	.9	.813
35	HLA-DPB1*0401	—	0.950	.717	.978	.826	.924	.913	.87
36	HLA-DPB1*0901	—	0.978	.739	.989	.891	.913	.913	.859
37	HLA-DR1	—	0.923	.587	.972	.781	.915	.838	.834
38	HLA-DR7	—	0.990	.976	.988	.976	.905	.976	.845
39	HLA-DRB1*1101	—	0.952	.732	.978	.828	.914	.898	.831
40	HLA-DRB1*1501	—	0.951	.846	.981	.897	.942	.91	.872

Note: * means flexible length.

strains (Figure 4). Among these 85 OMPs, two proteins (NP_755264.1, NP_756232.1) are predicted to contain three transmembrane helixes. Multiple transmembrane helixes make it difficult to purify recombinant proteins [48]. Therefore, these two proteins may not be good vaccine targets as whole protein antigens. When adhesin probability is taken into account, 58 out of the 83 proteins have an adhesin probability of $\geq .051$.

Functional gene enrichment analysis was performed to classify the roles of these 58 OMPs using the software DAVID (Table 3). Only 48 genes have annotation in DAVID and thus included in the DAVID analysis. Among these 48

Table 3: Selected function annotations significantly enriched for UPEC vaccine candidates based on DAVID analysis.

Category	Term	# of Genes	%	*P*-value	Benjamini *P*-value
GO MF*	Transport activity	22	45.8	1.8*E*−14	2.7 *E*−11
Interpro	TonB-dependent receptor, beta-barrel	10	20.8	1.0 *E*−13	4.3 *E*−10
Interpro	Porin, Gram-negative type	8	16.7	1.3 *E*−12	1.8 *E*−9
GO MF	Iron ion transmembrane transporter activity	5	10.4	6.9 *E*−6	1.4 *E*−3
Interpro	Fimbrial biogenesis outer membrane	5	10.4	9.3 *E*−5	4.1 *E*−2

*: GO MF, the Molecular Function (MF) branch of the Gene Ontology (GO).

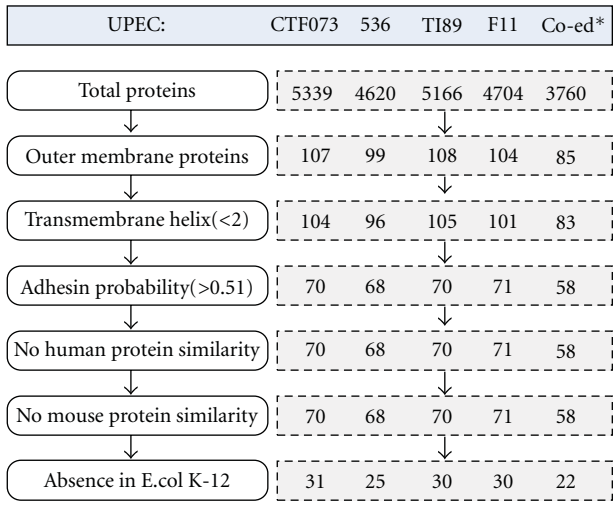


Figure 4: Prediction of UPEC vaccine targets conserved in four sequenced UPEC genomes using Vaxign. Note: * Co-ed represents the conserved proteins.

genes, significantly enriched function annotations are in the areas of transport activities, TonB-dependent receptor (beta-barrel), Gram-negative porin, iron ion transmembrane transporter activity, and fimbrial biogenesis in outer membrane (Table 3).

Of these 58 outer membrane proteins identified by Vaxign, 36 were further found to be present in the nonpathogenic *E. coli* K-12 strain MG1655 [43]. K-12 is used to remove those proteins that have been exposed to the host environment (e.g., gut) and may be tolerant by the host [49]. Only 22 proteins have been identified to be unique to the pathogenic UPEC strains (Figure 4).

A table of genes in different categories were further generated based on the Figure 4 and Table 3 and manual curation of literature data (Table 4). Eight *E. coli* proteins are predicted to contain iron-binding and iron siderophore transporter activity. Ten proteins are associated with a TonB box [50], and thus may play a role in iron acquisition by the bacterium. Another eight proteins are fimbrial biogenesis outer membrane usher proteins. Nine proteins are related to porin and ion transport. Indeed, many proteins in the list participate in transporter activity. Many lipoproteins have also been found. All of these targets would be logical selections. Many hypothetic proteins have been found with no defined functions or annotations.

3.5. Comparison of Vaxign Prediction Results and other Methods. The predicted results based on DNA sequence analysis are compared with data from transcriptomic microarray data [34], mass spectrometry proteomic studies [35, 51], and antigenicity analysis [36]. Out of 85 predicted outer membrane proteins that are conserved among four UPEC strains, 23 proteins have been found upregulated *in vivo* or in urine at the mRNA and/or protein levels (Table 4). It was found that many proteins with upregulated gene expression belong to iron ion binding proteins and porin family. However, only one protein (FimD) from fimbrial biogenesis outer membrane protein family was shown to be upregulated in DNA microarray analysis (Table 4) [34].

Five out of 14 iron binding proteins (IroN, FepA, FhuA, Hma, and ChuA) discovered by Vaxign have been found to be upregulated *in vivo* or in urine (Table 4) [34–36, 51]. Since iron metabolism is critical for UPEC pathogenesis, these proteins are important vaccine targets. Five proteins from porin family have also been found upregulated *in vivo* or in urine, including NmpC, OmpC, LamB, OmpF, and FadL (Table 4). Limited study has been performed to investigate the roles of these porin proteins in induction of protective immunity against UPEC infection.

3.6. Verification of Vaxign Predicted Results. Iron binding proteins were chosen for development of UPEC subunit vaccines. These proteins are typically outer membrane β-barrel proteins that function as receptors for iron-containing compounds. This group of proteins were predicted by Vaxign (Table 4) and significantly enriched based on gene enrichment analysis (Table 3). The antigen c2482 (renamed Hma for heme acquisition), a heme-binding protein, was first cloned and purified, and used for *in vivo* mouse testing. It was found that intranasal immunization with Hma generated an antigen-specific humoral response, antigen-specific production of IL-17 and IFN-γ, and provided significant protection against experimental infection with UPEC strain CFT073 [33].

ChuA was another heme/hemoglobin receptor that was also present in microarray & proteomics studies (Table 4) [34, 35]. Our experimental studies found that recombinant ChuA induced severe sickness in mice. Mice that recovered from the ChuA vaccination were challenged with strain CFT073, but were not protected (data not shown).

IroN has been found to be a protective antigen [49]. However, our study did not find significant protection

Table 4: Conserved UPEC outer membrane proteins predicted by Vaxign.

#	Protein RefSeq	Symbol	Adhesin	TMH	Not in K-12	Micro-array	Prote-omics	Protein Name
Iron ion binding and iron/siderophore transporter activity								
1	NP_752820.1	ybiL	0.857	1		–	–	Catecholate siderophore receptor fiu precursor (TonB-dependent receptor fiu) (Ferric iron uptake protein)
2	NP_754328.1	c2436	0.473	0		–	–	putative pesticin receptor precursor (tonB-dependent receptor)
3	NP_754406.1	c2518	0.83	0	X	–	–	TonB dependent receptor
4	NP_752238	c0294	0.83	0	X	–	–	TonB dependent receptor
5	NP_753164.1	iroN	0.672	0		+	+	Siderophore receptor iron
6	NP_752600.1	fepA	0.792	0		+	–	ferrienterobactin receptor (TonB-dependent receptor)
7	NP_752135.1	fhuA	0.746	0		+	+	ferrichrome outer membrane transporter
8	NP_754374.1	Hma (c2482)	0.772	0	X	+	+	outer membrane heme acquisition protein
9	NP_756170	chuA	0.846	0	X	+	+	Outer membrane heme/hemoglobin receptor
10	NP_753551.1	prrA	0.589	0	X	–	–	Putative TonB-dependent outer membrane receptor
11	NP_753179	c1265	0.777	0	X	–	–	Outer membrane heme/ hemoglobin receptor
12	NP_753125	c1206	0.794	0	X	–	–	Outer membrane heme/haemoglobin receptor
13	NP_755646	c3775	0.79	0	X	–	–	putative iron compound receptor
14	NP_753820.1	yddB	0.765	0		–	–	hypothetical protein c1924 (tonB-dependent receptor family)
Fimbrial biogenesis outer membrane usher protein								
15	NP_757244.1	fimD	0.744	1		+	–	Outer membrane usher protein fimD precursor
16	NP_757034	papC_2	0.674	1	X	–	–	PapC protein
17	NP_755465	papC	0.666	0	X	–	–	PapC protein
18	NP_754524.1	yehB	0	0		–	–	Outer membrane usher protein yehB precursor
19	NP_752120.1	htrE	0.643	0		–	–	Putative outer membrane usher protein
20	NP_753830.1	c1934	0.856	1		–	–	Outer membrane usher protein fimD precursor
21	NP_754765.1	yfcU	0.563	0	X	–	–	Fimbrial export usher family protein
22	NP_753156.1	focD	0.854	0		–	–	F1C fimbrial usher
23	NP_756076.1	ycbS	0.559	1		–	–	Outer membrane usher protein ycbS precursor
24	NP_753159	focH	0.917	0	X	–	–	F1C putative fimbrial adhesin precursor
Porin and ion transport								
25	NP_753469.1	nmpC	0.788	1	X	–	+	Outer membrane porin protein nmpC precursor

Table 4: Continued.

#	Protein RefSeq	Symbol	Adhesin	TMH	Not in K-12	Micro-array	Prote-omics	Protein Name
				Porin and ion transport				
26	NP_754644.1	ompC	0.688	0		−	+	Outer membrane porin protein C
27	NP_756858.1	lamB	0.806	0		−	+	maltoporin
28	NP_752996.1	ompF	0.614	0		−	+	Outer membrane protein F (Porin family)
29	NP_754240.1	c2348	0.759	0	X	−	−	Outer membrane porin protein nmpC precursor
30	NP_752325.1	phoE	0.729	0		−	−	Outer membrane phosphoporin protein E
31	NP_753724.1	ompN	0.751	0		−	−	Outer membrane protein N precursor
32	NP_754275.1	c2383	0.597	0	X	−	−	Outer membrane protein N precursor
33	NP_754771.1	fadL	0.871	0		−	+	Long-chain fatty acid outer membrane transporter
34	NP_756025.1	hofQ	0.186	0		+	−	Outer membrane porin HofQ
				Other transport proteins				
35	NP_756748.1	c4894	0.81	0	X	−	−	Nucleoside-specific channel-forming protein tsx precursor
36	NP_756500.1	c4642	0.694	0		−	−	Putative outer membrane Protein yieC
37	NP_753669.1	c1765	0.437	0		−	−	Partial putative outer membrane channel protein
38	NP_752455.1	tsx	0.833	0		−	+	Nucleoside-specific channel-forming protein tsx precursor
				Lipoproteins				
39	NP_756849.1	yjbH	0.651	0		−	−	Lipoprotein yjbH precursor
40	NP_755008.1	yfiB	0.564	0		−	−	Putative outer membrane lipoprotein
41	NP_756936.1	yjcP	0.185	0		−	−	Putative outer membrane efflux protein MdtP
42	NP_752589.1	cusC	0.516	0		−	−	Copper/silver efflux system outer membrane protein CusC
43	NP_754925.1	yfhM	0.224	0		−	−	Lipoprotein yfhM precursor
44	NP_756232.1	yiaD	0.526	3		+	−	Putative outer membrane lipoprotein
				Other outer membrane proteins				
45	NP_752286.1	c0345	0.987	1	X	−	−	ShlA/HecA/FhaA exofamily protein
46	NP_752352	eaeH	0.904	1	X	−	−	Putative adhesin
47	NP_753126	c1207	0.702	0	X	−	−	Hypothetical protein
48	NP_753493	c1585	0.526	0	X	−	−	Putative tail component of prophage
49	NP_754912	c3030	0.71	1	X	−	−	SinI-like protein
50	NP_756286	c4424	0.99	0	X	−	−	Putative adhesin
51	NP_752116.1	yadC	0.81	1		−	−	Putative fimbrial-like adhesin protein

Table 4: Continued.

#	Protein RefSeq	Symbol	Adhesin	TMH	Not in K-12	Micro-array	Prote-omics	Protein Name
				Other outer membrane proteins				
52	NP_752162.1	yaeT	0.637	0		+	+	Outer membrane protein assembly factor YaeT
53	NP_752163.1	hlpA	0.587	0		−	−	Periplasmic chaperone
54	NP_752339.1	yagX	0.571	0		−	−	Hypothetical protein c0402
55	NP_752642.1	crcA	0.739	1		−	−	Palmitoyl transferase for Lipid A
56	NP_752830.1	ompX	0.818	1		+	+	Outer membrane protein X
57	NP_753262.1	flgK	0.84	0		−	−	Flagellar hook-associated protein FlgK
58	NP_753627.1	ompW	0.848	0		−	+	Outer membrane protein W
59	NP_753695.1	ompG	0.618	0		−	−	Outer membrane protein G precursor
60	NP_754014.1	ydiY	0.708	0		−	−	Hypothetical protein c2120
61	NP_754081.1	yeaF	0.852	0		+	+	MltA-interacting protein precursor
62	NP_754523.1	yehA	0.847	0		−	−	Hypothetical protein c2635
63	NP_754661.1	yfaL	0.949	0		−	−	Adhesin
64	NP_755530.1	c3655	0.962	0		−	+	Antigen 43 precursor
65	NP_756601.1	pldA	0.756	0		−	−	Phospholipase A
66	NP_754578.1	cirA	0.51	0		−	−	Colicin I receptor
67	NP_755652.2	tolC	0	0		−	+	Outer membrane channel protein
68	NP_752583.1	ompT	0	0		+	+	Outer membrane protease
69	NP_756783.1	btuB	0	0		−	+	Vitamin B12/cobalamin outer membrane transporter
70	NP_754246.1	fliF	0.213	2		−	−	Flagellar MS-ring protein
71	NP_755960.1	yheF	0.239	0		−	−	General secretion pathway protein D precursor
72	NP_752585.1	nfrA	0.298	0		−	−	Bacteriophage N4 receptor, outer membrane subunit
73	NP_753902.1	uidC	0.304	1		−	−	Putative outer membrane porin protein
74	NP_754656.1	c2770	0.348	0		−	−	Hypothetical protein c2770
75	NP_757165.1	ytfM	0.4	0		−	−	Hypothetical protein c5318
76	NP_753730.1	ydbH	0.401	1		−	−	Hypothetical protein c1828
77	NP_756204.1	yhjL	0.408	0		−	−	Cellulose synthase subunit BcsC
78	NP_751977.1	c0021	0.418	1		−	−	Hypothetical protein c0021
79	NP_757166.1	ytfN	0.426	1		−	−	Hypothetical protein c5319
80	NP_754559.2	yohG	0.427	0		+	−	Multidrug resistance outer membrane protein MdtQ
81	NP_756598.1	c4739	0.43	0		−	−	Hypothetical protein c4739
82	NP_753585.1	ychP	0.435	0		−	−	Hypothetical protein c1680
83	NP_754913.1	c3031	0.448	0		−	−	SinH-like protein
84	NP_752491.1	ybaU	0.464	1		−	−	Peptidyl-prolyl cis-trans isomerase (rotamase D)
85	NP_755264.1	c3389	0.282	3	X	−	−	Hypothetical protein

Note: The protein RefSeq numbers are from UPEC strain CFT073. All 85 proteins are conserved across total four UPEC genomes. This table also shows the adhesin probability, number of transmembrane helixes, and absence in nonpathogenic K-12 MG1655. TMH, transmembrane alpha helix prediction. Microarray, transcriptomic mRNA results (+ for up-regulation *in vivo*). Proteomics, protein expression results (+ for up-regulation in urine or *in vivo*).

stimulated by IroN [33]. This protein also exists in *E. coli* K-12, which may bring a discussion about whether it is needed to use this cutoff.

Two other proteins, IreA (NP_757022.1, c5174) and IutA (NP_755498.1), were also tested based on six independent screens [33]. Both are putative iron-regulated outer membrane virulence proteins. Our studies found that IreA and IutA were able to independently stimulate protective immunity in mouse bladder against challenge with UPEC strain CFT073 [33]. These two proteins were not shown in our final list of vaccine candidates predicted by our Vaxign analysis pipeline because they were filtered out due to their absence in the other three UPEC genomes.

4. Discussion

Vaxign is the first web-based vaccine design software program freely available for the purpose of facilitating reverse vaccinology. Vaxign optimizes the conditions and performance of many public tools and provides new programs in a way optimal for analyzing high throughput data. The seamless integration makes Vaxign a user-friendly environment specific for reverse vaccinology. Our analysis indicates that Vaxign specifically and sensitively predicts known vaccine targets and also provides new vaccine target candidates deserving further wet lab confirmation. Vaxign is expected to become a publically available web-based program for vaccine researchers to efficiently design vaccine targets and develop vaccines using a rationale reverse vaccinology strategy.

To test whether Vaxign is capable of predicting those protective antigens that have been validated based on wet laboratory experiments, we have curated the literature and obtain a list of proteins and used Vaxign to analyze those protective antigens. Vaxign has also been used to predict vaccine targets using other bacteria such as *Brucella* spp., *Neisseria meningitides*, and *Mycobacterium tuberculosis*. Our studies indicated that Vaxign predicted results are consistent with existing reports [37].

We showed in this report that Vaxign can be successfully used for prediction of UPEC vaccine candidates. While UPEC FimH was reported to be a protective antigen [52], it was not included in our list of predicted genes (Table 4). FimH is predicted by Vaxign as an adhesin with an adhesin probability of .96. This prediction is consistent with current knowledge about this protein [52]. Based on an X-ray structure analysis, FimH is folded into two domains of the all-beta class connected by a short extended linker [53]. FimH was not shown in our final predicted list since its subcellular localization was predicted unknown (Probability = .2). If only a high adhesin probability is considered, FimH would be included in our prediction list. This also indicates different Vaxign options selected by a user would change the results. However, we identified another protein in that complex (FimD) (Table 4). Vaxign identified IroN, Hma, and ChuA (Table 4) which were selected as possible protective antigens after lengthy experimental assessment [33]. Our study found that Hma induced protection in mice from transurethral challenge with UPEC. Another independent study indicated that subcutaneous immunization with denatured IroN conferred significant protection against renal, but not bladder, urinary tract infection in a mouse model [54].

While recombinant ChuA induced severe sickness in mice, the immunized mice did not protect against virulent UPEC infection. This sickness was probably due to its Heme-binding activity. The possible release of high levels of inflammatory cytokines and innate immune response might lead to mouse death. It is likely that ChuA contains some immunodominant T cell epitope(s) that activates effector (inflammatory) T cell immunity [46]. In many cases, subdominant epitopes that induce subdominant responses may be important components of an effective immune defence [19]. Immunization with subdominant but optimal epitopes can often induce T cell responses that are more effective than immunodominant epitopes. A more advanced *in silico* prediction would be able to predict and optimize epitopes for vaccine development.

Our study also indicated that microarray and proteomics gene expression data were complementary to DNA sequence-based analysis in predicting vaccine targets (Table 4). Future directions of further Vaxign development may include addition of other components such as analysis of high throughput transcriptomic (e.g., DNA microarray and superarray) and proteomic data for vaccine target prediction. Predicted vaccine targets can also be analyzed based on gene annotation enrichment to further refine vaccine targets using tools such as DAVID. The gene enrichment results combined with predictions based on DNA sequence analysis as well as mRNA and protein gene expression allowed us to focus on the group of iron binding proteins for experimental testing.

More than 700 microbial genomes have been sequenced and analyzed, which provide a foundation for scientists to develop vaccines using the reverse vaccinology. Reverse vaccinology shortens the period of vaccine target discovery and evaluation to 1-2 years [1]. This new strategy also revolutionizes new vaccine development against pathogens for which the applications of Pasteur's principles have failed.

The use of proteins is a common approach for genetically engineered vaccine development. However, generating epitope vaccines has many advantages and is currently an active research area. To give the most simplified example, if only one epitope of a large protein is protective, using the peptide epitope would allow the delivery of much higher dose of the key epitope during vaccination. Therefore, prediction of a successful epitope would increase efficacy for the vaccine.

Our studies found that Vaxitope is a sensitive and specific program for predicting immune epitopes that provide good candidates for epitope vaccine development. We are in the processing of designing and evaluating epitope-based UPEC vaccines using Vaxign. We will first target to predict epitopes from antigens (e.g., Hma) that have proven able to induce protective immunity.

It often occurs that many epitopes can be predicted from one specific protein. It is often challenging to rank predicted epitopes for vaccine testing. The epitope ranking can also be used to rank proteins. Many programs, such as

EpiAssembler by EpiVax [55], allow epitope content ranking. It is known that the best T cell epitopes tend to contain "clusters" of MHC binding motifs, and the clustering is highly correlated with the immunogenicity [46]. Therefore, it is more effective to design a peptide(s) containing clustered epitopes for induction of better immunogenicity in rational vaccine development. Promiscuous epitopes are those MHC ligands or T-cell epitopes that are recognized in the context of more than one MHC molecule and recognized by more than one T-cell clone. Many software programs, such as TEPITOPE [56], enable the computational identification of promiscuous MHC ligands. The prediction of promiscuous epitopes is also an important feature for epitope-based vaccine design.

It is often that a vaccine candidate that is effective in a mouse model is not effective in human. If the epitopes are designed for human use, the mice used for testing the epitope vaccine usually need to be transgenic. Generating HLA transgenic mice is costly and time consuming. It is possible, however, to design epitopes that are effective for both mouse and human. For example, it was reported that an epitope in human immunodeficiency virus 1 reverse transcriptase was recognized by both mouse and human cytotoxic T lymphocytes [57]. Prediction and screening of such epitopes would simplify our testing of human vaccine candidates in the mouse model.

The molecular mimicry or the cross-reactivity between self epitopes and pathogen epitopes has been found a common reason for many pathogen-induced autoimmune diseases [46]. Many pathogens, such as *Klebsiella pneumoniae*, *Proteus mirabilis*, human coronavirus, and Lyme disease spirochete *Borrelia burgdorferi* carry antigens which cross-react with human antigens [44, 46]. For example, the oligopeptide QTDRED is common to both *K. pneumoniae* and HLA-B27 nitrogenase reductase enzyme. This sequence similarity appears to cause ankylosing spondyltis. *Proteus mirabilis* hemolysin contains a molecular mimicry sequence ESRRAL that has the same shape and charge distribution as the rheumatoid arthritis susceptibility sequence EQRRAA. Antibody levels against *P. mirabilis* hemolysin and a synthetic peptide ESRRAL were significantly higher in rheumatoid arthritis patients [44]. To avoid the autoimmunity, it is important to eliminate the epitopes that are conserved. Currently Vaxign provides a genome-wide sequence similarity analysis at protein levels. Many programs, such as Conservatrix [19] and IEDB Sequence Mapping tool (http://tools.immuneepitope.org/esm/esmhelp.jsp?tab=help), have been developed to map epitope sequences. We plan to develop such epitope sequence mapping tool in Vaxign in the future.

Vaxign is part of VIOLIN, a web-based vaccine database and analysis resource [58]. The predicted vaccine targets from Vaxign will also integrate with those manually annotated vaccine data available in VIOLIN. An literature mining program based on the Vaccine Ontology (http://www.violinet.org/vaccineontology) is also being developed to facilitate automated literature data processing and inference for the purpose of retrieving valuable data for rational vaccine design.

Acknowledgments

This paper was supported by a pilot research Grant to YH and HM at the Center for Computational Medicine and Bioinformatics (CCMB) at the University of Michigan Medical School, Michigan, USA, and Public Health Service grants AI43363 and AI081062 from the National Institutes of Health.

References

[1] R. Rappuoli, "Reverse vaccinology," *Current Opinion in Microbiology*, vol. 3, no. 5, pp. 445–450, 2000.

[2] M. Pizza, V. Scarlato, V. Masignani, et al., "Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing," *Science*, vol. 287, no. 5459, pp. 1816–1820, 2000.

[3] N. Ariel, A. Zvi, H. Grosfeld et al., "Search for potential vaccine candidate open reading frames in the Bacillus anthracis virulence plasmid pXO1: in silico and in vitro screening," *Infection and Immunity*, vol. 70, no. 12, pp. 6817–6827, 2002.

[4] B. C. Ross, L. Czajkowski, D. Hocking et al., "Identification of vaccine candidate antigens from a genomic analysis of *Porphyromonas gingivalis*," *Vaccine*, vol. 19, no. 30, pp. 4135–4142, 2001.

[5] S. Montigiani, F. Falugi, M. Scarselli et al., "Genomic approach for analysis of surface proteins in *Chlamydia pneumoniae*," *Infection and Immunity*, vol. 70, no. 1, pp. 368–379, 2002.

[6] T. M. Wizemann, J. H. Heinrichs, J. E. Adamou et al., "Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection," *Infection and Immunity*, vol. 69, no. 3, pp. 1593–1598, 2001.

[7] D. N. Chakravarti, M. J. Fiske, L. D. Fletcher, and R. J. Zagursky, "Application of genomics and proteomics for identification of bacterial gene products as potential vaccine candidates," *Vaccine*, vol. 19, no. 6, pp. 601–612, 2000.

[8] J. C. Betts, "Transcriptomics and proteomics: tools for the identification of novel drug targets and vaccine candidates for tuberculosis," *IUBMB Life*, vol. 53, no. 4-5, pp. 239–242, 2002.

[9] A. S. De Groot, "Immunomics: discovering new targets for vaccines and therapeutics," *Drug Discovery Today*, vol. 11, no. 5-6, pp. 203–209, 2006.

[10] J. L. Gardy, M. R. Laird, F. Chen et al., "PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis," *Bioinformatics*, vol. 21, no. 5, pp. 617–623, 2005.

[11] L. Käll, A. Krogh, and E. L. Sonnhammer, "Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server," *Nucleic Acids Research*, vol. 35, pp. W429–W432, 2007.

[12] G. Sachdeva, K. Kumar, P. Jain, and S. Ramachandran, "SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks," *Bioinformatics*, vol. 21, no. 4, pp. 483–491, 2005.

[13] L. Li, C. J. Stoeckert Jr., and D. S. Roos, "OrthoMCL: identification of ortholog groups for eukaryotic genomes," *Genome Research*, vol. 13, no. 9, pp. 2178–2189, 2003.

[14] S. Henikoff, J. G. Henikoff, and S. Pietrokovski, "Blocks+: a non-redundant database of protein alignment blocks derived from multiple compilations," *Bioinformatics*, vol. 15, no. 6, pp. 471–479, 1999.

[15] P. A. Reche, J.-P. Glutting, H. Zhang, and E. L. Reinherz, "Enhancement to the RANKPEP resource for the prediction of peptide binding to MHC molecules using profiles," *Immunogenetics*, vol. 56, no. 6, pp. 405–419, 2004.

[16] B. Peters, H. H. Bui, S. Frankild et al., "A community resource benchmarking predictions of peptide binding to MHC-I molecules," *PLoS Computational Biology*, vol. 2, no. 6 , article e65, 2006.

[17] H.-G. Rammensee, J. Bachmann, N. P. N. Emmerich, O. A. Bachor, and S. Stevanović, "SYFPEITHI: database for MHC ligands and peptide motifs," *Immunogenetics*, vol. 50, no. 3-4, pp. 213–219, 1999.

[18] S. Vivona, F. Bernante, and F. Filippini, "NERVE: new enhanced reverse vaccinology environment," *BMC Biotechnology*, vol. 6, article 35, 2006.

[19] A. S. De Groot, H. Sbai, C. S. Aubin, J. McMurry, and W. Martin, "Immuno-informatics: mining genomes for vaccine components," *Immunology and Cell Biology*, vol. 80, no. 3, pp. 255–269, 2002.

[20] C. DeLisi and J. A. Berzofsky, "T-cell antigenic sites tend to be amphipathic structures," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 82, no. 20, pp. 7048–7052, 1985.

[21] A. Sette, S. Buus, E. Appella et al., "Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 86, no. 9, pp. 3296–3300, 1989.

[22] O. Rotzschke, K. Falk, S. Stevanovic, G. Jung, P. Walden, and H.-G. Rammensee, "Exact prediction of a natural T cell epitope," *European Journal of Immunology*, vol. 21, no. 11, pp. 2891–2894, 1991.

[23] A. Sette, J. Sidney, C. Oseroff et al., "HLA DR4w4-binding motifs illustrate the biochemical basis of degeneracy and specificity in peptide-DR interactions," *Journal of Immunology*, vol. 151, no. 6, pp. 3163–3170, 1993.

[24] M. P. Davenport, I. A. P. H. Shon, and A. V. S. Hill, "An empirical method for the prediction of T-cell epitopes," *Immunogenetics*, vol. 42, no. 5, pp. 392–397, 1995.

[25] A. S. De Groot, B. M. Jesdale, E. Szu, J. R. Schafer, R. M. Chicz, and G. Deocampo, "An interactive web site providing major histocompatibility ligand predictions: application to HIV research," *AIDS Research and Human Retroviruses*, vol. 13, no. 7, pp. 529–531, 1997.

[26] V. Brusic, G. Rudy, and L. C. Harrison, "MHCPEP: a database of MHC-binding peptides," *Nucleic Acids Research*, vol. 22, no. 17, pp. 3663–3665, 1994.

[27] B. Peters, J. Sidney, P. Bourne et al., "The immune epitope database and analysis resource: from vision to blueprint," *PLoS Biology*, vol. 3, no. 3, article e91, pp. 379–381, 2005.

[28] M. S. Litwin, C. S. Saigal, E. M. Yano et al., "Urologic diseases in America project: analytical methods and principal findings," *Journal of Urology*, vol. 173, no. 3, pp. 933–937, 2005.

[29] I. Connell, W. Agace, P. Klemm, M. Schembri, S. Mårild, and C. Svanborg, "Type 1 fimbrial expression enhances *Escherichia coli* virulence for the urinary tract," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 18, pp. 9827–9832, 1996.

[30] S. Langermann, R. Möllby, J. E. Burlein et al., "Vaccination with FimH adhesin protects cynomolgus monkeys from colonization and infection by uropathogenic *Escherichia coli*," *Journal of Infectious Diseases*, vol. 181, no. 2, pp. 774–778, 2000.

[31] D. T. Uehling, W. J. Hopkins, J. E. Elkahwaji, D. M. Schmidt, and G. E. Leverson, "Phase 2 clinical trial of a vaginal mucosal vaccine for urinary tract infections," *Journal of Urology*, vol. 170, no. 3, pp. 867–869, 2003.

[32] V. Kumar, N. K. Ganguly, K. Joshi et al., "Protective efficacy and immunogenicity of *Escherichia coli* K13 diphtheria toxoid conjugate against experimental ascending pyelonephritis," *Medical Microbiology and Immunology*, vol. 194, no. 4, pp. 211–217, 2005.

[33] C. J. Alteri, E. C. Hagan, K. E. Sivick, S. N. Smith, and H. L. T. Mobley, "Mucosal immunization with iron receptor antigens protects against urinary tract infection," *PLoS Pathogens*, vol. 5, no. 9, Article ID e1000586, 2009.

[34] J. A. Snyder, B. J. Haugen, E. L. Buckles et al., "Transcriptome of uropathogenic *Escherichia coli* during urinary tract infection," *Infection and Immunity*, vol. 72, no. 11, pp. 6373–6381, 2004.

[35] C. J. Alteri and H. L. T. Mobley, "Quantitative profile of the uropathogenic *Escherichia coli* outer membrane proteome during growth in human urine," *Infection and Immunity*, vol. 75, no. 6, pp. 2679–2688, 2007.

[36] E. C. Hagan and H. L. T. Mobley, "Uropathogenic *Escherichia coli* outer membrane antigens expressed during urinary tract infection," *Infection and Immunity*, vol. 75, no. 8, pp. 3941–3949, 2007.

[37] Z. Xiang and Y. He, "Vaxign: a web-based vaccine target design program for reverse vaccinology," *Procedia in Vaccinology*, vol. 1, no. 1, pp. 23–29, 2009.

[38] H. R. Bigelow, D. S. Petrey, J. Liu, D. Przybylski, and B. Rost, "Predicting transmembrane beta-barrels in proteomes," *Nucleic Acids Research*, vol. 32, no. 8, pp. 2566–2577, 2004.

[39] T. L. Bailey and M. Gribskov, "Combining evidence using p-values: application to sequence homology searches," *Bioinformatics*, vol. 14, no. 1, pp. 48–54, 1998.

[40] J. A. Swets, "Measuring the accuracy of diagnostic systems," *Science*, vol. 240, no. 4857, pp. 1285–1293, 1988.

[41] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2009.

[42] K. D. Pruitt, T. Tatusova, and D. R. Maglott, "NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins," *Nucleic Acids Research*, vol. 33, pp. D501–D504, 2005.

[43] F. R. Blattner, G. Plunkett III, C. A. Bloch et al., "The complete genome sequence of *Escherichia coli* K-12," *Science*, vol. 277, no. 5331, pp. 1453–1462, 1997.

[44] C. Wilson, H. Tiwana, and A. Ebringer, "Molecular mimicry between HLA-DR alleles associated with rheumatoid arthritis and Proteus mirabilis as the aetiological basis for autoimmunity," *Microbes and Infection*, vol. 2, no. 12, pp. 1489–1496, 2000.

[45] I. Nachamkin, B. M. Allos, and T. Ho, "Campylobacter species and Guillain-Barre syndrome," *Clinical Microbiology Reviews*, vol. 11, no. 3, pp. 555–567, 1998.

[46] C. A. Weber, P. J. Mehta, M. Ardito, L. Moise, B. Martin, and A. S. De Groot, "T cell epitope: friend or foe? Immunogenicity of biologics in context," *Advanced Drug Delivery Reviews*, vol. 61, no. 11, pp. 965–976, 2009.

[47] P. Naves, G. del Prado, L. Huelves et al., "Correlation between virulence factors and in vitro biofilm formation by *Escherichia coli* strains," *Microbial Pathogenesis*, vol. 45, no. 2, pp. 86–91, 2008.

[48] D. Serruto, R. Rappuoli, and M. Pizza, "Meningococcus B: from genome to vaccine," in *Genomics, Proteomics and Vaccines*, G. Grandi, Ed., pp. 185–201, John Wiley & Sons, 2004.

[49] L. Durant, A. Metais, C. Soulama-Mouze, J.-M. Genevard, X. Nassif, and S. Escaich, "Identification of candidates for a subunit vaccine against extraintestinal pathogenic *Escherichia coli*," *Infection and Immunity*, vol. 75, no. 4, pp. 1916–1925, 2007.

[50] A. G. Torres, P. Redford, R. A. Welch, and S. M. Payne, "TonB-dependent systems of uropathogenic *Escherichia coli*: aerobactin and heme transport and TonB are required for virulence in the mouse," *Infection and Immunity*, vol. 69, no. 10, pp. 6179–6185, 2001.

[51] M. S. Walters and H. L. T. Mobley, "Identification of uropathogenic *Escherichia coli* surface proteins by shotgun proteomics," *Journal of Microbiological Methods*, vol. 78, no. 2, pp. 131–135, 2009.

[52] S. Langermann, S. Palaszynski, M. Barnhart et al., "Prevention of mucosal *Escherichia coli* infection by FimH-adhesin-based systemic vaccination," *Science*, vol. 276, no. 5312, pp. 607–611, 1997.

[53] D. Choudhury, A. Thompson, V. Stojanoff et al., "X-ray structure of the FimC-FimH chaperone-adhesin complex from uropathogenic *Escherichia coli*," *Science*, vol. 285, no. 5430, pp. 1061–1066, 1999.

[54] T. A. Russo, C. D. McFadden, U. B. Carlino-MacDonald, J. M. Beanan, R. Olson, and G. E. Wilding, "The Siderophore receptor IroN of extraintestinal pathogenic *Escherichia coli* is a potential vaccine candidate," *Infection and Immunity*, vol. 71, no. 12, pp. 7164–7169, 2003.

[55] A. S. De Groot, L. Marcon, E. A. Bishop et al., "HIV vaccine development by computer assisted design: the GAIA vaccine," *Vaccine*, vol. 23, no. 17-18, pp. 2136–2148, 2005.

[56] T. Sturniolo, E. Bono, J. Ding et al., "Generation of tissue-specific and promiscuous HLA ligand databases using DNA microarrays and virtual HLA class II matrices," *Nature Biotechnology*, vol. 17, no. 6, pp. 555–561, 1999.

[57] A. Hosmalin, M. Clerici, R. Houghten et al., "An epitope in human immunodeficiency virus 1 reverse transcriptase recognized by both mouse and human cytotoxic T lymphocytes," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 87, no. 6, pp. 2344–2348, 1990.

[58] Z. Xiang, T. Todd, K. P. Ku et al., "VIOLIN: vaccine investigation and online information network," *Nucleic Acids Research*, vol. 36, no. 1, pp. D923–D928, 2008.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 361841, 7 pages
doi:10.1155/2010/361841

Review Article

A New Method for the Evaluation of Vaccine Safety Based on Comprehensive Gene Expression Analysis

Haruka Momose,[1] Takuo Mizukami,[1] Masaki Ochiai,[2] Isao Hamaguchi,[1] and Kazunari Yamaguchi[1]

[1] *Department of Safety Research on Blood and Biological Products, National Institute of Infectious Diseases, 4-7-1 Gakuen, Musashimurayama, Tokyo 208-0011, Japan*
[2] *Division of Quality Assurance, National Institute of Infectious Diseases, 4-7-1 Gakuen, Musashimurayama, Tokyo 208-0011, Japan*

Correspondence should be addressed to Kazunari Yamaguchi, kyama@nih.go.jp

Received 30 September 2009; Accepted 2 April 2010

Academic Editor: Yongqun Oliver He

For the past 50 years, quality control and safety tests have been used to evaluate vaccine safety. However, conventional animal safety tests need to be improved in several aspects. For example, the number of test animals used needs to be reduced and the test period shortened. It is, therefore, necessary to develop a new vaccine evaluation system. In this review, we show that gene expression patterns are well correlated to biological responses in vaccinated rats. Our findings and methods using experimental biology and genome science provide an important means of assessment for vaccine toxicity.

1. Introduction

Vaccination effectively enables the control of many infectious diseases. However, we cannot always avoid the problem of adverse reactions accompanied by vaccination. While most adverse reactions are mild and local, some vaccines have been associated with very rare but severe systemic reactions. Therefore, all vaccines for public use are made in compliance with Good Manufacturing Practices (GMP) to prevent safety problems. Furthermore, manufacturers must submit samples and results of their in-house tests for each vaccine batch to the national control authorities before vaccines are released into the market. Among many quality control tests, conventional animal safety tests are performed to detect vaccine toxicity because residual vaccine toxicity has the potential to cause adverse reactions. For example, the animal body weight change test is the most commonly used test to evaluate the toxicity of vaccines [1]. Although a good correlation of the body weight loss with a vaccine's toxicity has been shown [2, 3], a greater understanding of the molecular mechanisms involved in the reaction to a vaccines' toxicity is needed. We, therefore, attempted to measure animals' responses to vaccines by determining changes in gene expression profiles.

Gene expression profiling is a unique way to characterize how cells or tissues are affected by abnormal conditions. The measurement of gene expression levels upon exposure to toxicants can be used to identify toxic products, and to provide information about the mechanism of toxicity [4]. DNA microarray technology has opened the way for the parallel detection and analysis of expression patterns of thousands of genes in a single experiment. Furthermore, the development of high-quality gene arrays has allowed DNA microarray technology to become a standard tool in molecular toxicology. Recently, the field of toxicogenomics has validated the concept of gene expression profiles as "signatures" of toxicant classes [5–7]. These signatures have effectively directed the analytical search for predictive toxicant biomarkers and they have contributed to the understanding of the dynamic responses of molecular mechanisms associated with toxic responses. In fact, many studies of gene-expression profiles have now been reported in the toxicology field. For example, Hamadeh et al. reported patterns of gene expression in liver tissue taken from rats exposed to different

chemicals [8]. DNA microarray assays have also been applied to the analysis of the side effects of medicines [9]. Recently, the United States Food and Drug Administration (FDA) and the European Medicines Agency (EMEA) have, either individually or together, started to review submissions for the qualification of biomarkers for medical products for specific purposes proposed by industry [10]. The introduction of pharmacogenomics, or pharmacogenetics, to the evaluation of medicines is a global trend.

For a better understanding of the molecular toxicology regarding vaccines, DNA microarray analysis promises to be an ideal method, as has been the case for pharmaceuticals. The FDA now encourages the voluntary submission of genomic data to the FDA outside of the regular review process [11]. However, no studies similar to those described above for pharmaceuticals have yet been conducted in the field of vaccines. At the beginning of this review, we summarized the current efforts used for the control of vaccine safety using conventional animal tests. We then referred to our recent efforts using DNA microarray analysis to identify "genetic signatures" for the toxicants remaining in vaccines. Since pertussis and influenza vaccines are among the most commonly used vaccines, we tried to develop a system to evaluate the "genetic signatures" of the toxicity of these vaccines.

2. Current Vaccine Safety Test

2.1. Body Weight Change in Vaccinated Animals. To screen for general toxicity of vaccines, the body weight of vaccine-treated animals can be analyzed as the general safety test [12]. Five mL of the vaccine are injected into the peritoneum of guinea pigs weighing 300–400 g, and the weight loss experienced by the animals is analyzed at days 1, 2, 3, 4, and 7 after administration. None of the animals should show any abnormal signs; no statistically significant ($P = .01$) difference in weight loss should be observed between the treated animals and the control group on any observation day. This test has been applied to a wide variety of vaccines in a unified way, and plays an important role in ensuring the safety and consistency of vaccine batches [12]. For pertussis vaccine (inactivated whole cell formulation), the effects of vaccine treatment were also measured using test for toxicity to mouse weight gain, in addition to the general safety test. All mice were weighed on days 0, 1, 2, 3, 4, and 7 after vaccine administration. The criterion for judgment is that mean body weight 3 days after injection should be no less than that at the time of injection upon statistical analysis, and no mice showed any abnormal sign during the observation periods [12]. When the reference vaccine (RE: the inactivated whole cell pertussis vaccine) was administrated, weight loss was observed on day 1 after administration (Figure 1(a)).

2.2. Leukocytosis-Promoting Toxicity in Vaccinated Animals. To detect the toxin present in pertussis vaccines, the number of peripheral leukocytes can also be analyzed. Pertussis vaccine is injected into the peritoneum of mice at a dose of 0.5 mL. Leukocytes present in peripheral blood are then counted 3 days after injection [12]. The white blood cell (WBC) counts in peripheral blood of reference vaccine-treated mice reach approximately 2,500 cells/μL (Figure 1(b)). The standard criterion of safety for pertussis vaccine (inactivated whole cell formulation) is that the mean count of leukocytes in peripheral blood, 3 days after injection, should not exceed 10 times that before injection [12].

2.3. Leukopenic Toxicity Test in Vaccinated Animals. Quality control of influenza vaccines is performed using the general safety test and the leukopenic toxicity test (LTT), which is based on peripheral WBC counts in mice 12–18 hours after intraperitoneal injection of a vaccine. The criterion for judgment is that the leukopenic toxicity of the test sample relative to that of the toxicity reference sample should be no higher than the value corresponding to 80% of the leukocyte count of the control relative to that of the toxicity reference sample [12–14].

3. DNA Microarray-Based Safety Test

The currently used quality control and safety tests, such as the LTT and the general safety test, have been used to evaluate vaccine safety for over 50 years [3]. We are now developing a new quality control method for vaccines using DNA microarray analysis as a substitute for the conventional animal tests [15–17]. The principle of this method is to translate vaccine quality, immunogenicity, and reactogenicity, into gene expression profile data. This method is expected to be informative, rapid, and highly sensitive.

For DNA microarray analysis using vaccines, 8 week-old male rats, weighing 180–220 g, were intraperitoneally administered with 5 mL of vaccine or physiological saline (SA). Three to 6 rats were used for each group. Vaccinated rats were sacrificed to obtain whole lung, kidney, brain, and the lateral left lobe of the liver on day 1, 2, 3, and 4 postadministration (Figure 2). Tissues were immediately frozen in liquid nitrogen for storage. Thawed tissue was homogenized and poly(A)$^+$ RNA was purified from the lysate. Cyanine 5-labeled poly(A)$^+$ RNA was subjected to DNA microarray analysis. Blood was also collected, however, this could not be analyzed due to the low quality of purified RNA.

For DNA microarray analysis, a set of synthetic polynucleotides (80-mers) representing 11,468 rat transcripts and including most of the RefSeq genes deposited in the NCBI database (MicroDiagnostic, Tokyo, Japan) was arrayed on aminosilane-coated glass slides [18, 19]. Cyanine 5-labeled poly(A)$^+$ RNA was competitively hybridized on the slide with cyanine 3-labeled common reference RNA. Hybridization signals were measured, processed into primary expression ratios ([Cyanine 5-intensity obtained from each sample]/[Cyanine 3-intensity obtained from common reference RNA]), and then normalized by multiplying normalization factors calculated for each microarray feature.

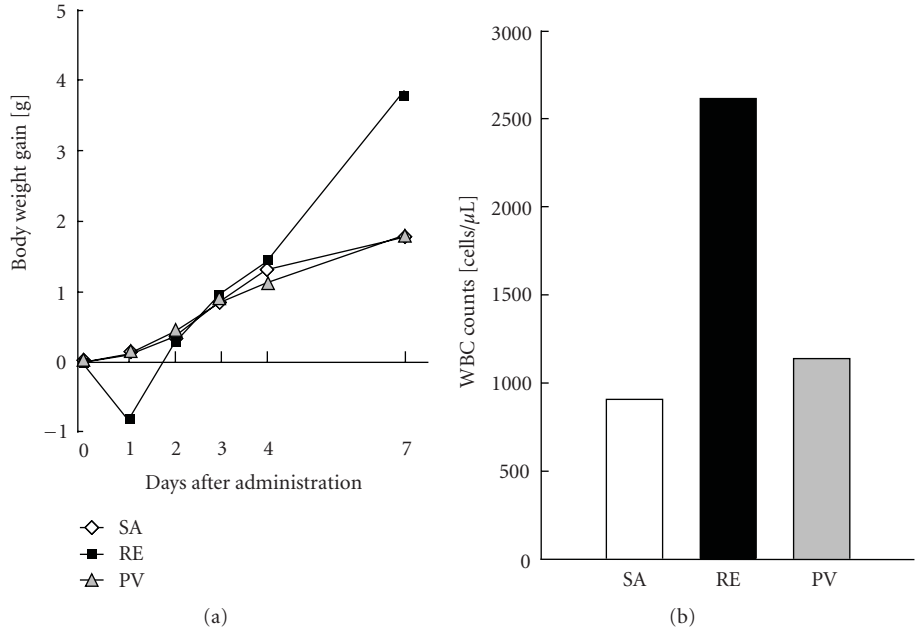


Figure 1: Safety control tests for pertussis vaccines. (a) Test for toxicity to mouse weight gain. Physiological saline (SA), an inactivated whole-cell pertussis vaccine (RE), or an acellular pertussis vaccine (PV)-administered mice were weighed on 0, 1, 2, 3, 4, and 7 days postadministration. Ten mice in each group were used, and the mean changes in body weight are indicated. (b) Leukocytosis promoting activity of various pertussis vaccines. White blood cell (WBC) counts in peripheral blood were measured 3 days after vaccine administration. Ten mice in each group were used and the mean WBC counts are indicated.

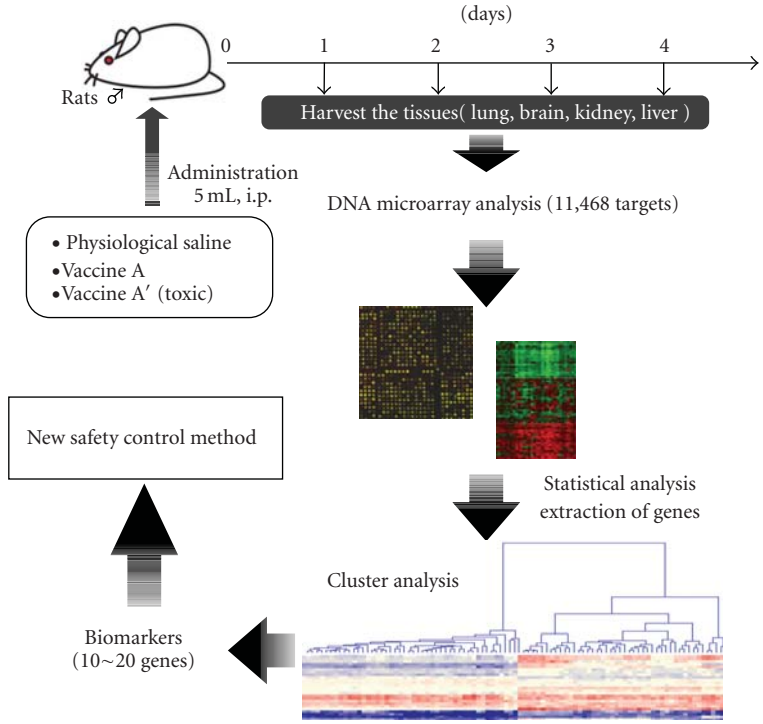


Figure 2: The gene expression analysis procedure. The detail of the procedure is described in the text.

For data processing and hierarchical cluster analysis, the primary expression ratios were converted into $\log_2$ ratios ($\log_2$ Cyanine-5 intensity/Cyanine-3 intensity). The genes with $\log_2$ ratios over 1 or under -1 in at least one sample were extracted from the primary data matrix, then subjected to two-dimensional hierarchical cluster analysis for samples and genes.

For the identification of biomarker genes for pertussis vaccines, we extracted differentially expressed genes from physiological saline and pertussis toxin-treated lung samples using the *t*-test ($P < .01$). Among the extracted genes, we further selected genes that exhibited mean average $\log_2$ ratio differences greater than 0.75 between the two sample groups [17]. For influenza vaccines, we extracted differentially expressed genes from physiological saline and inactivated whole-virion vaccine-treated lung samples using the *t*-test ($P < .005$) [16].

4. Pertussis Vaccines

Pertussis, or whooping cough, is an infectious respiratory disease caused by a Gram-negative bacillus, *Bordetella pertussis*. *Bordetella pertussis* possesses several pathogenic components, including pertussis toxin (PT) [20]. PT is known as a leukocytosis promoting factor, a major contributor to the pathogenesis of pertussis, and an antigen in immunity to pertussis [21]. At present, whole-cell pertussis vaccines and acellular pertussis vaccines containing inactivated PT are in commercial use [20].

Although pertussis vaccines are effective in the prevention of whooping cough, they have occasionally caused local reactions such as redness, swelling, and pain at the injection site. However, little is known about the overall responses to these vaccines. To address this problem, we applied DNA microarray analysis and quantification of specific genes to analyze the toxicants in pertussis vaccines [15, 17]. Three preparations, an acellular vaccine containing inactivated pertussis toxin (PV), an inactivated whole-cell vaccine (RE), and a purified pertussis toxin (PT) were prepared. RE is a reference vaccine for National Quality Control Tests of pertussis vaccines in Japan and is made from formaldehyde-inactivated *Bordetella pertussis* preparations. Physiological saline (SA) was used as a control. For comprehensive gene expression analysis, 5 mL of SA, PV, PT, and RE were each injected into 3 rats and the vaccinated tissues, lung, brain, kidney, and liver, were harvested at 1, 2, 3, and 4 days after vaccine administration. The experiments were performed twice and purified poly(A)$^+$ RNA from a total of 384 samples was subjected to DNA microarray analysis.

Of the 4 organs tested, the lung expressed genes that were extracted by DNA microarray analysis were classified sharply into clusters depending on sample treatment. From the DNA microarray analysis of vaccinated rat lungs at day 1, 13 genes for which expression levels were dynamically changed in response to PT treatment were [17] (accession numbers were updated in Table 1). Interestingly, the DNA microarray-based gene expression data correlated well with the body weight change of vaccine-treated mice (Figure 1(a)) and rats [17]. The real-time PCR quantification results of the expression levels of the 13 genes were comparable to the relative expression ratios from the DNA microarray analysis. Furthermore, cluster analysis using the 13 genes could distinguish SA- and PV-treated groups from PT- and RE-treated groups. These 13 genes are likely to be closely involved in the toxicity of pertussis vaccines. To quantify these genes in a convenient way, the QuantiGene Plex assay was applied. The QuantiGene Plex assay enabled the simultaneous analysis of the 13 genes. We evaluated the expression levels of the 13 genes in the lungs of rats vaccinated with various doses of RE. Nine genes, *S100A9*, *S100A8*, *IRF7*, *MX2*, *IFI27L*, *BEST5*, *MMP9*, *MMP8*, and *CYP2E1* (indicated in bold letters in Table 1) showed dose-dependent up-or down-regulation in response to the various doses of RE treatment. RE vaccine toxicity could be measured by the expression level in lung lysate of these 9 genes. The quantification of these 9 genes using the QuantiGene Plex assay is, we believe, a promising candidate for a new control test for pertussis vaccines.

5. Influenza Vaccines

Influenza virus triggers a highly contagious acute respiratory disease and has caused epidemics and global pandemics, partly because it possesses the capacity for gradual antigenic change in two surface antigens, hemagglutinin (HA) and neuraminidase (NA) [22]. To combat influenza, split vaccines consisting of subvirion preparations and whole-virus vaccines are manufactured using strains recommended annually by the WHO, based on the antigenic characteristics of HAs and NAs. Furthermore, the recent circulation of the highly pathogenic avian influenza A (H5N1) virus has raised concerns about the preparations for a coming influenza pandemic [23]. Many efforts are underway to develop vaccines against influenza A (H5N1).

To identify biomarkers for influenza vaccine toxicity, 3 vaccines were used: trivalent influenza HA vaccine (HAv, a split vaccine), trivalent influenza vaccine (WPv, an inactivated whole-virion vaccine), and prepandemic influenza vaccine (PDv, inactivated whole-virion (A/H5N1) absorbed onto an aluminum salt). All were produced by Kaketsuken, The Chemo-Sero-Therapeutic Research Institute, Japan. Physiological saline (SA) was used as a control. For comprehensive gene expression analysis, SA, HAv, WPv, and PDv were each injected into 5 rats, and the vaccinated tissues, lung, liver, brain, and peripheral blood, were harvested at 1, 2, 3, and 4 days after vaccine administration. Purified poly(A)$^+$ RNA from a total of 320 samples was subjected to DNA microarray analysis [16]. Based on the analysis of pertussis vaccines, described above, the gene expression profiles from lung samples were subjected to two-dimensional hierarchical cluster analysis. PDv- and WPv-treated samples at day 1 formed an independent cluster from other samples, indicating distinct profiles in gene expression of these groups. As was the case with pertussis vaccines, we tried to identify several biomarkers from the analysis of lung gene expression. The analysis of lungs from vaccinated rats at day 1 resulted in the extraction of 76 genes, whose expression levels were statistically different between SA- and

Table 1: Biomarkers for pertussis vaccine toxicity.

Category	Accession no.	Symbol	Brief description
Inflammation	NM_053587	***S100A9***	A calcium binding protein that may be associated with acute inflammatory processes, coupled with S100a8
	NM_053822	***S100A8***	May play a role in inflammatory responses such as cell motility, coupled with S100a9
	NM_019323	*MCPT9*	A serine protease expressed in mast cells, but the precise function has not yet been determined
	NM_031530	*CCL2*	A ligand for CCR2 that acts as a chemoattractant of monocytes
IFN inducible, immune response	NM_001033691	***IRF7***	Unknown
	NM_134350	***MX2***	Involved in inhibiting vesicular stomatitis virus
	NM_203410	***IFI27***	Induced by steroid hormone, IFN, and LPS in endometrium at implantation, dendric cells, and macrophases
	NM_001007694	*IFIT3*	May induced by IFN or virus infection
	Y07704	***BEST5***	Induced by IFN and involved in bone formation
Peptidoglycan metabolism	NM_031055	***MMP9***	Metalloproteinase involved in extracellular matrix remodeling, bone resorption, and immune responses
	NM_022221	***MMP8***	May play a role in appositional bone formation and regulation of the extracellular matrix
Xenobiotic metabolism	J02627	***CYP2E1***	Protects hepatocytes from stress-induced cell death
Others	NM_001106862	*NGP*	Unknown

Table 2: Biomarkers for influenza vaccine toxicity.

Category	Accession No.	Symbol	Brief description
IFN inducible gene	NM_172019	*IFI47*	Mouse homolog may be a guanine nucleotide-binding protein induced by IFN-gamma
	AF329825	*TRAFD1*	Putative TRAF-interacting zinc finger protein
	NM_019242	*IFRD1*	May be involved in proliferation of neuronal and glial precursors
IFN inducible, immune response	NM_001033691	*IRF7*	Unknown
	NM_134350	*MX2*	Involved in inhibiting vesicular stomatitis virus
Immune response	NM_172222	*C2*	Likely component of the classical pathway of the complement cascade
	NM_012708	*PSMB9*	Subunit of the proteasome complex, which may play a role in protein catabolism
	NM_032056	*TAP2*	Transports peptides into the ER lumen for binding with MHC class I molecules; plays a role in antigen processing and presentation
	NM_033098	*TAPBP*	Facilitates the binding of MHC class I molecules to the transporter associated with antigen processing (TAP) in MHC class I assembly
	NM_017264	*PSME1*	May play a role in proteasome activation
Chemokine and Cytokine function	AF065438	*LGALS3BP*	Displays differential expression in a fibroblast cell line transformed by human T-cell leukemia virus type 1 Tax protein
	NM_012977	*LGALS9*	A highly selective urate transporter/channel
	NM_053819	*TIMP1*	Acts as an inhibitor of metalloprotease activity; may play a role in vascular tissue remodeling
	NM_023981	*CSF1*	Plays a role in macrophage formation
	NM_145672	*CXCL9*	Chemokine which plays a role in the recruitment of mononuclear cells and in allograft rejection
	XM_223236	*CXCL11*	Mouse homolog is a chemokine and is involved in the immune response
Transcription activity	AJ302054	*ZBP1*	DNA binding protein; thought to bind Z-DNA, which is largely controlled by the amount of supercoiling

WPv-treated samples ($P < .005$) [16]. The cluster analysis using these 76 genes successfully distinguished WPv- and PDv-treated groups at day 1 from other groups, indicating the suitability of the 76 genes as biomarkers for influenza vaccines.

The extracted 76 genes were categorized according to function, such as interferon-inducible, chemokine and cytokine function, immune response, transcriptional activity, and so on. Among the 76 genes, 17 genes met the requirement for high expression levels and were chosen as representatives for each functional category (Table 2). Among the 17 genes, *IRF7* and *MX2* were also nominated for biomarkers of pertussis vaccine toxicity. Real-time PCR quantification results of the expression levels of the 17 genes were comparable to the relative expression ratios determined by DNA microarray analysis. We are now working to establish a rapid quantification system for these 17 biomarkers using the QuantiGene Plex assay.

6. Japanese Encephalitis Vaccines

Japanese encephalitis (JE) is a seasonal and sporadic encephalitis in East Asia caused by the JE virus. Vaccination is very important to prevent JE infection, because palliative care is the only treatment available for JE patients. Recently, a Vero cell-derived JE vaccine had been licensed in Japan as an alternative to the long-used mouse brain-derived JE vaccines. The newly developed Vero cell-derived vaccine should be at least equivalent to the mouse brain-derived vaccines, because the mouse brain-derived vaccines were considered generally safe and succeeded in the near elimination of JE in certain endemic regions. In this context, we performed DNA microarray analysis of tissues from rats administered with mouse brain-derived or Vero cell-derived JE vaccine and compared the gene expression profiles. As expected, the gene expression patterns in brain and liver were comparable between mouse brain-derived and Vero cell-derived vaccines, indicating that both vaccines possessed equivalent reactivity characteristics in rats [24].

7. Conclusions

Over recent decades, the safety control of vaccines has been assessed using several animal tests, including the body weight change test and white blood cell counts. However, conventional animal safety tests need to be improved in many aspects. For example, the number of test animals used needs to be reduced and the test period needs to be shortened. This requires the development of a new vaccine evaluation system. In this review, we showed that gene expression patterns were well correlated to the biological responsiveness of vaccinated animals. From the DNA microarray analysis of lungs from vaccinated rats, we identified 13 and 17 biomarkers to detect the toxicity of pertussis and influenza vaccines, respectively.

Furthermore, the QuantiGene Plex assay for gene expression analysis is being introduced. The QuantiGene Plex assay was revealed to be as accurate as real-time PCR and has the great benefit of being able to evaluate all biomarkers simultaneously. Using the QuantiGene Plex assay, we could rapidly and sensitively detect the gene expression changes that accompany biological reactivity in vaccinated rats.

Thus, it may be concluded that DNA microarray technology is an informative, rapid, and highly sensitive method with which to evaluate vaccine quality. Our data suggest that this new method has the potential to shorten the time for safety tests and can reduce the number of animals used. In addition, our test may contribute to the development of urgently required vaccines. Further analyses are required to confirm that gene expression changes correlate with vaccine quality.

In this review, we referred to our recent efforts of exploring new safety control methods using gene expression pattern indexes, focusing on pertussis and influenza vaccines. In the future, for the evaluation of all kinds of vaccines, microarray analysis is expected to play an important role in the new safety control test, especially for checking toxin-reactive transcripts.

Acknowledgment

This work was supported by Grants-in-Aid from the Ministry of Health, Labour and Welfare, Japan.

References

[1] Y. Horiuchi, M. Takahashi, T. Konda, et al., "Quality control of diphtheria tetanus acellular pertussis combined (DTaP) vaccines in Japan," *Japanese Journal of Infectious Diseases*, vol. 54, no. 5, pp. 167–180, 2001.

[2] M. Kurokawa, "Toxicity and toxicity testing of pertussis vaccine," *Japanese Journal of Medical Science and Biology*, vol. 37, no. 2, pp. 41–81, 1984.

[3] T. Mizukami, A. Masumi, H. Momose, et al., "An improved abnormal toxicity test by using reference vaccine-specific body weight curves and histopathological data for monitoring vaccine quality and safety in Japan," *Biologicals*, vol. 37, no. 1, pp. 8–17, 2009.

[4] T. Lettieri, "Recent applications of DNA microarray technology to toxicology and ecotoxicology," *Environmental Health Perspectives*, vol. 114, no. 1, pp. 4–9, 2006.

[5] E. K. Lobenhofer, P. R. Bushel, C. A. Afshari, and H. K. Hamadeh, "Progress in the application of DNA microarrays," *Environmental Health Perspectives*, vol. 109, no. 9, pp. 881–891, 2001.

[6] W. Pennie, S. D. Pettit, and P. G. Lord, "Toxicogenomics in risk assessment: an overview of an HESI collaborative research program," *Environmental Health Perspectives*, vol. 112, no. 4, pp. 417–419, 2004.

[7] A. H. Harrill and I. Rusyn, "Systems biology and functional genomics approaches for the identification of cellular responses to drug toxicity," *Expert Opinion on Drug Metabolism & Toxicology*, vol. 4, no. 11, pp. 1379–1389, 2008.

[8] H. K. Hamadeh, P. R. Bushel, S. Jayadev, et al., "Gene expression analysis reveals chemical-specific profiles," *Toxicological Sciences*, vol. 67, no. 2, pp. 219–231, 2002.

[9] N. Ejiri, K.-I. Katayama, N. Kiyosawa, Y. Baba, and K. Doi, "Microarray analysis on phase II drug metabolizing enzymes expression in pregnant rats after treatment with

pregnenolone-16α-carbonitrile or phenobarbital," *Experimental and Molecular Pathology*, vol. 79, no. 3, pp. 272–277, 2005.

[10] European Medicines Agency Concept Paper, "Pharmacogenomics (PG) biomarker qualification: format and data standards," in *Proceedings of the International Conference on Harmonisation of Technical Requirements for Registration of Pharmaceuticals for Human Use*, June 2008, EMEA/CHMP/190395/2008, http://www.emea.europa.eu/pdfs/human/pharmacogenetics/19039508en.pdf.

[11] F. W. Frueh, "Impact of microarray data quality on genomic data submissions to the FDA," *Nature Biotechnology*, vol. 24, no. 9, pp. 1105–1107, 2006.

[12] *Minimum Requirements for Biological Products*, National Institute of Infectious Diseases, Tokyo, Japan, 2006, http://www.nih.go.jp/niid/MRBP/files/seibutsuki_english.pdf.

[13] M. Kurokawa, S. Ishida, S. Asakawa, S. Iwasa, and N. Goto, "Toxicities of influenza vaccine: peripheral leukocytic response to live and inactivated influenza viruses in mice," *Japanese Journal of Medical Science and Biology*, vol. 28, no. 1, pp. 37–52, 1975.

[14] F. Chino, "The views and policy of the Japanese control authorities on the three Rs," *Developments in Biological Standardization*, vol. 86, pp. 53–62, 1996.

[15] I. Hamaguchi, J.-I. Imai, H. Momose, et al., "Two vaccine toxicity-related genes Agp and Hpx could prove useful for pertussis vaccine safety control," *Vaccine*, vol. 25, no. 17, pp. 3355–3364, 2007.

[16] T. Mizukami, J.-I. Imai, I. Hamaguchi, et al., "Application of DNA microarray technology to influenza A/Vietnam/1194/2004 (H5N1) vaccine safety evaluation," *Vaccine*, vol. 26, no. 18, pp. 2270–2283, 2008.

[17] I. Hamaguchi, J.-I. Imai, H. Momose, et al., "Application of quantitative gene expression analysis for pertussis vaccine safety control," *Vaccine*, vol. 26, no. 36, pp. 4686–4696, 2008.

[18] E. Ito, R. Honma, J.-I. Imai, et al., "A tetraspanin-family protein, T-cell acute lymphoblastic leukemia-associated antigen 1, is induced by the Ewing's sarcoma-Wilms' tumor 1 fusion protein of desmoplastic small round-cell tumor," *American Journal of Pathology*, vol. 163, no. 6, pp. 2165–2172, 2003.

[19] S. Kobayashi, E. Ito, R. Honma, et al., "Dynamic regulation of gene expression by the Flt-1 kinase and Matrigel in endothelial tubulogenesis," *Genomics*, vol. 84, no. 1, pp. 185–192, 2004.

[20] K. M. Edwards and M. D. Decker, "Pertussis vaccines," in *Vaccines*, S. A. Plotkin, W. A. Orenstein, and P. A. Offit, Eds., pp. 467–517, Elsevier, New York, NY, USA, 5th edition, 2008.

[21] H. Sato and Y. Sato, "Bordetella pertussis infection in mice: correlation of specific antibodies against two antigens, pertussis toxin, and filamentous hemagglutinin with mouse protectivity in an intracerebral or aerosol challenge system," *Infection and Immunity*, vol. 46, no. 2, pp. 415–421, 1984.

[22] C. B. Bridges, J. M. Katz, R. A. Levandowski, and N. J. Cox, "Inactivated influenza vaccines," in *Vaccines*, S. A. Plotkin, W. A. Orenstein, and P. A. Offit, Eds., pp. 259–290, Elsevier, New York, NY, USA, 5th edition, 2008.

[23] K. Ungchusak, P. Auewarakul, S. F. Dowell, et al., "Probable person-to-person transmission of avian influenza A (H5N1)," *The New England Journal of Medicine*, vol. 352, no. 4, pp. 333–340, 2005.

[24] H. Momose, J.-I. Imai, I. Hamaguchi, et al., "Induction of indistinguishable gene expression patterns in rats by Vero cell-derived and mouse brain-derived Japanese encephalitis vaccines," *Japanese Journal of Infectious Diseases*, vol. 63, no. 1, pp. 25–30, 2010.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 942365, 9 pages
doi:10.1155/2010/942365

Research Article

Mass Spectrometric Analysis of Multiple Pertussis Toxins and Toxoids

Yulanda M. Williamson,[1] Hercules Moura,[1] David Schieltz,[1] Jon Rees,[1] Adrian R. Woolfitt,[1] James L. Pirkle,[1] Jacquelyn S. Sampson,[2] Maria L. Tondella,[2] Edwin Ades,[2] George Carlone,[2] and John R. Barr[1]

[1] *Biological Mass Spectrometry Laboratory, National Center for Environmental Health, Centers for Disease Control and Prevention, Chamblee, GA 30341, USA*
[2] *Division of Bacterial Diseases, National Center for Immunizations and Respiratory Diseases, Centers for Disease Control and Prevention, Atlanta, GA 30333, USA*

Correspondence should be addressed to John R. Barr, jbarr@cdc.gov

Received 15 December 2009; Accepted 9 March 2010

Academic Editor: Yongqun Oliver He

Bordetella pertussis (Bp) is the causative agent of pertussis, a vaccine preventable disease occurring primarily in children. In recent years, there has been increased reporting of pertussis. Current pertussis vaccines are acellular and consist of Bp proteins including the major virulence factor pertussis toxin (Ptx), a 5-subunit exotoxin. Variation in Ptx subunit amino acid (AA) sequence could possibly affect the immune response. A blind comparative mass spectrometric (MS) analysis of commercially available Ptx as well as the chemically modified toxoid (Ptxd) from licensed vaccines was performed to assess peptide sequence and AA coverage variability as well as relative amounts of Ptx subunits. Qualitatively, there are similarities among the various sources based on AA percent coverages and MS/MS fragmentation profiles. Additionally, based on a label-free mass spectrometry-based quantification method there is differential relative abundance of the subunits among the sources.

1. Introduction

Bordetella pertussis (Bp), a gram-negative bacterium, is the causative agent of pertussis, commonly known as whooping cough [1]. Pertussis is a highly communicable respiratory infection that in the absence of vaccination primarily afflicts infants and children. Transmission occurs via direct contact with discharges from respiratory mucous membranes of infected persons. Persons who contract pertussis develop a persistent cough that includes the characteristic high-pitched "whoop" sound that can last up to eight weeks [1, 2]. Pertussis is a vaccine preventable disease but according to the World Health Organization (WHO), there are an estimated 50 million cases of pertussis per year worldwide with approximately 300,000 leading to death [3, 4]. The majority of the high mortality rates come from developing countries. In the United States acellular pertussis vaccines are administered to infants and young children [5] and in 2006 were recommended for adolescents and adults [6–8]. Acellular vaccines consist of up to five Bp components including the major virulence factor pertussis toxin (Ptx), filamentous hemagluttinin, pertactin, fimbrae 2, and/or fimbrae 3 [3]. Pertussis vaccines are generally formulated in combination with vaccine immunogens for other diseases, such as tetanus and diphtheria [9].

Ptx, a five-subunit protein complex (S1–S5) is a major bacterial exotoxin, surface adhesin, and porin, is involved in the establishment of infection in humans and has been shown to be an important component of the vaccine [10]. Ptx, a ~105 kDa A-B protein complex, comprises the protomer (A) and oligomer (B). The A component consists of S1 (~22 kDa), which contains ADP ribosylation enzymatic activity and is the most immunoreactive region of the toxin. The B portion consists of S2, S3, S4, and S5, decreasing in size from ~20 kDa to 10 kDa, respectively, and is associated with adherence to human host cells [11, 12].

From a cellular perspective, the B portion binds to cells inducing the enzymatic activity of S1 and its transversion into the cell, followed by inactivation of G proteins by S1 and subsequent initiation of an intracellular cascade of signaling events which includes factors that regulate the immune response [11]. Importantly, the Ptxs in current vaccines have been inactivated (toxoids; Ptxd) by covalent crosslinkages formed using agents such as formaldehyde or glutaraldehyde [13, 14].

Recently, the U.S. and several other countries have reported an increase in pertussis cases. Factors such as waning immunity after natural infection or immunization, improved molecular-based diagnostics, increased surveillance, or changes in circulating strain antigenicity have all been hypothesized as contributing factors [3] and have led researchers to re-examine pertussis vaccine components and preparation strategies. Changes in Ptx subunit protein sequences as well as the amounts present in the vaccine formulation could affect the overall immune response. Moreover, identification of commonalities as well as potential disparities of various Ptx and Ptxd sources on a proteomic level could provide insight into changes in population immune responses to the vaccine.

Mass spectrometry (MS), a powerful and sensitive analytical tool has been used to identify, differentiate, and characterize proteins in complex mixtures [15]. In particular, MS techniques, such as liquid chromatography-electrospray ionization (LC-ESI) tandem MS (MS/MS) or matrix-assisted laser desorption ionization time-of-flight (MALDI-TOF) MS/MS, which vary in their rapidity and sensitivity of peptide detection as a result of enzymatic cleavage, are commonly used in proteomic analysis [15, 16]. Proteins can be proteolytically cleaved by enzymes such as trypsin and the peptides generated are ionized and separated based on their mass-to-charge ratios. The peptides and the proteins, from which they originated, are identified based on comparisons between the MS/MS fragmentation patterns and protein databases. [17]. For instance, Bottero et al. used MALDI-TOF MS/MS as a means to characterize enriched Bp membrane proteins, which included the Ptx S1, from different strains [18]. Alternatively, Tumala and colleagues have used liquid chromatography MS/MS (LC-MS/MS) to accurately measure the molecular weights of Ptx subunits as well as to identify novel chemical modification sites [19].

In this study, a blind comparative mass spectrometric analysis of Ptx and Ptxd from multiple sources was performed to assess potential sequence and amino acid (AA) coverage variability as well as relative amounts of pertussis toxin subunits. This first-phase study demonstrates the sensitivity of mass spectrometry in ascertaining differences at the peptide sequence level and may provide a basis for future correlative immunoproteomic studies.

2. Methods

2.1. Pertussis Toxin and Toxoid Sources. The commercially available Ptx and Ptxd were purchased from Protein Express, Inc. (Cincinnati, OH, USA), LIST Biologicals, Inc. (Campbell, CA, USA), Sanofi Pasteur Pharmaceutical (Paris, France), or GSK Pharmaceutical (London, UK). The sources were blindly assigned a letter, in which Ptxs are represented as A or B, while the chemically inactivated Ptxd samples are denoted as C or D.

2.2. Pertussis Toxin and Toxoid Digestions. The Ptx and Ptxd sources (10 μg) were vortexed and precipitated with 100% acetonitrile (ACN), for 60 minutes, and then dried using a vacuum centrifuge. Trypsin (5 μg) (Promega Corporation, Madison, WI, USA) was added and the volume was brought to 100 μL with tryptic digestion buffer (50 mM ammonium bicarbonate adjusted to pH 8.5 with NaOH, 0.6 mM $CaCl_2$, 0.1% (w/v) potassium azide (KN_3), and 3 mg/L phenol red). The samples were incubated at 37°C for 3, 6, and 24 hours. Digested samples (250 ng total) were subsequently prepared for either MALDI-TOF MS/MS or LC-MS/MS. Additionally, in separate experiments, samples were treated with an acid cleavable detergent, RapiGest (RG; Waters Corporation, Milford, MA, USA) to assess the efficiency of tryptic digestion. 0.1% RG in tryptic digestion buffer was added to Ptx or Ptxd (10 μg) samples. The samples were next incubated at 100°C for 5 minutes followed by cooling at room temperature with the subsequent addition of trypsin (5 μg). Trypsin digestions were performed as stated above. However, prior to MS analysis, the RG was inactivated by cleavage with 0.175 M HCL at 37°C for 30 minutes, and the supernatant (obtained by centrifugation for 15 minutes at 17,300 g) was transferred to a new vial.

2.3. MS Sample Preparation. Samples for MALDI-TOF MS/MS were resuspended in an equal volume of 0.1% trifluoroacetic acid (TFA), and the peptides were separated using a C_{18} zip-tip minicolumn (Millipore Corporation, Billerica, MA, USA) and eluted in two steps with 3 μL of 30% ACN in 0.1% TFA and then 3 μL of 80% ACN in 0.1% TFA. The eluates were mixed with 10 μL α-cyano-4-hydroxycinnamic acid (α-CHCA) (Sigma Aldrich, St. Louis, MO, USA) and 0.5 μL was next spotted on to a stainless steel plate. A protein mix diluted at 1 : 10 was spotted on the plate in the appropriate wells and used for instrument calibration purposes. For LC-MS/MS, samples were simply resuspended in equal volumes of 0.1% formic acid (FA).

2.4. Mass Spectrometry

2.4.1. MALDI-TOF MS/MS. Mass spectra of each sample well were obtained by scanning from 800 to 4000 m/z in MS positive-ion reflector mode on the Applied Biosystems 4800 Proteomics Analyzer (Framingham, MA, USA). The instrument uses a nd:YAG laser at 337 nm with a 200 MHz repetition rate, and each spectrum is an average of 1000 laser shots. In MS/MS mode, the CID spectra were generated using air as the collision gas at 1 kV and an average of 500 laser shots for MS/MS acquisition. The instrument was tuned and calibrated for MS analysis using a mixture of five peptides: des-Arg1-Bradykinin (m/z 904.47), angiotensin I (m/z 1,296.69), Glu1-fibrinopeptide B (m/z 1,570.68), ACTH (1–17)(m/z 2093.08), and ACTH (18–39)(m/z 2,465.20). For MS/MS mass calibration, known

fragment ion masses observed in the Glu1-fibrinopeptide B were used. The instrument was programmed for automated MS/MS acquisition using data dependent acquisition (DDA) mode. Selected peptides were subjected to MS/MS analysis to confirm their identity through sequence determination.

All MS/MS spectra were automatically submitted to a Mascot Server through Mascot Distiller (Matrix Science Inc., London, UK; version 2.2.1.0) and searched against the entire NCBInr database or a modified NCBInr database created to search "*Bordetella*" or "pertussis" proteins with deamidation, oxidation and carbamidomethylation assigned as variable modifications. Tryptic constraints were applied, allowing up to two missed cleavages. Peptide searches were performed with an initial tolerance on mass measurement of 100 ppm in MS mode and 150 ppm in MS/MS mode. All matched peptides were verified manually and the searches combined and processed by Scaffold (Proteome Software Inc., Portland, OR, USA; version 2.03.01).

2.4.2. Nano-LC ESI-MS/MS (LTQ-Orbitrap). Protein identification was achieved by using nanoflow liquid chromatography, data-dependant tandem mass spectrometry, and database searching. A 365 μm by 75 μm fused silica pulled needle capillary (New Objective, Inc., Woburn, MA) was filled in house with 10 cm of 5 μm Symmetry 300 reverse phase packing material (Waters Inc., Bedford, MA). Protein digests were directly loaded onto the analytical column without using a trap column and introduced using an Eksigent autosampler. The gradient was delivered by an Eksigent 2D nanoLC system (Eksigent Technologies, Inc, Dublin, CA) where the mobile phase solvents consisted of: (solvent A) 0.2% formic acid (Thermo Scientific, Rockford, IL), 0.005% trifluoroacetic acid (Sigma-Aldrich) in water (Burdick and Jackson, Muskegon, MI), (solvent B) 0.2% formic acid, and 0.005% trifluoroacetic acid in acetonitrile (Burdick and Jackson). The gradient flow was set at 400 nL/min, and the profile consisted of a hold at 5% B for 5 minutes followed by a ramp to 30% B over 100 minutes then a ramp up to 90% B in 5 minutes and a hold at 90% for 2 minutes before returning to 5% B in 2 minutes and re-equilibration at 5% B for 20 minutes. Peptides were nano-electrosprayed into an LTQ Orbitrap tandem mass spectrometer (Thermo Scientific, San Jose, CA). The voltage applied to the nano-electrospray source was 2.0 kV. The mass spectrometer was programmed to perform data-dependant acquisition by scanning the mass range from m/z 400 to 1600 at a nominal resolution setting of 60,000 for parent ion acquisition in the Orbitrap. Then tandem mass spectra of doubly-charged and higher-charge states were acquired for the top 10 most intense ions. Singly-charged ions were not considered for MS/MS. All tandem mass spectra were recorded using the linear ion trap. This DDA process cycled continuously throughout the duration of the gradient.

All tandem mass spectra were extracted from the raw data file using Mascot Distiller (Matrix Science, London, UK; version 2.2.1.0) and searched using Mascot (version 2.2.0). Mascot was setup to search using the entire NCBInr database or a modified NCBInr database created to search "*Bordetella*" or "pertussis" recognized proteins in which trypsin is used as the digestion agent. Mascot was searched with a fragment ion tolerance mass of 0.80 Da and a parent ion tolerance of 200 ppm. Variable modifications for the mascot search were deamidation, oxidation, and carbamidomethylation. Scaffold was used to validate MS/MS-based peptide and protein identifications. Peptide identifications were accepted if they could be established at greater than 95.0% probability as specified by the Peptide Prophet algorithm [20]. Protein identifications were accepted if they could be established at greater than 99.0% probability and contained at least 2 identified peptides [21]. With these stringent parameters of Peptide Prophet and Protein Prophet, the false discovery rate was zero.

2.4.3. Nano-LC ESI-MS/MS (Q-Tof Premier). LC-MS/MS was carried out using a nanoAcquity ultra-performance liquid chromatography (UPLC) coupled to a QTof Premier MS system (Waters Corporation). LC separation was performed using a Symmetry C_{18} trapping column and a BEH C_{18} column (100 μm I.D. $\times$ 100 mm long with 1.7 μm packing), at a flow rate of 1.2 μL/min. Standard water/ACN gradients containing 0.1% formic acid in both solvents were used for elution, with the ACN increasing linearly from 1% to 50% in 50 minutes, and an overall cycle time of 90 minutes. The QTof utilized an MS^E (or Protein Expression) method, which involved acquiring data-independent alternating low- and high-energy scans over the m/z range 50–1990 in 0.6 second, along with lockmass data on a separate channel to obtain accurate mass measurement. The data obtained using the MS^E method were further processed using ProteinLynx Global Server v2.3, which provided statistically validated peptide and protein identification using a SwissProt-Trembl protein database, along with relative protein quantification analysis. Root mean square mass accuracies were typically within 8 ppm for the MS data, and within 15 ppm for MS/MS data. Also the false positive rate was 4%. Lastly, the relative quantification assessed in this study used an internal standard from a known protein and is considered a surrogate.

3. Results

A mass spectrometric analysis was performed to assess protein sequence homogeneity of Ptx and Ptxd. MALDI-TOF MS/MS, and LC-MS/MS approaches revealed comparable to varying AA percent composition profiles (percent is based on the number of AA (within tryptic peptides) generated divided by the total number of AA in the protein) among Ptx and Ptxd (Table 1) as well as a compilation of detected peptides (Table 2). Generally, Ptx A and B digested with trypsin generated higher AA coverages among the nontreated oligomeric subunits compared to their chemically modified Ptxd counterparts regardless of the MS system used. Additionally, trypsin digests of Ptx B generated the highest average among S2–S5 compared to all samples. Also, for both Ptxd samples there was low to no detection of S2–S5 peptides.

However, AA coverage reports for S1 on average between Ptx and Ptxd were fairly comparable with 45 and 44%, respectively. Probing further, LC-MS/MS instruments

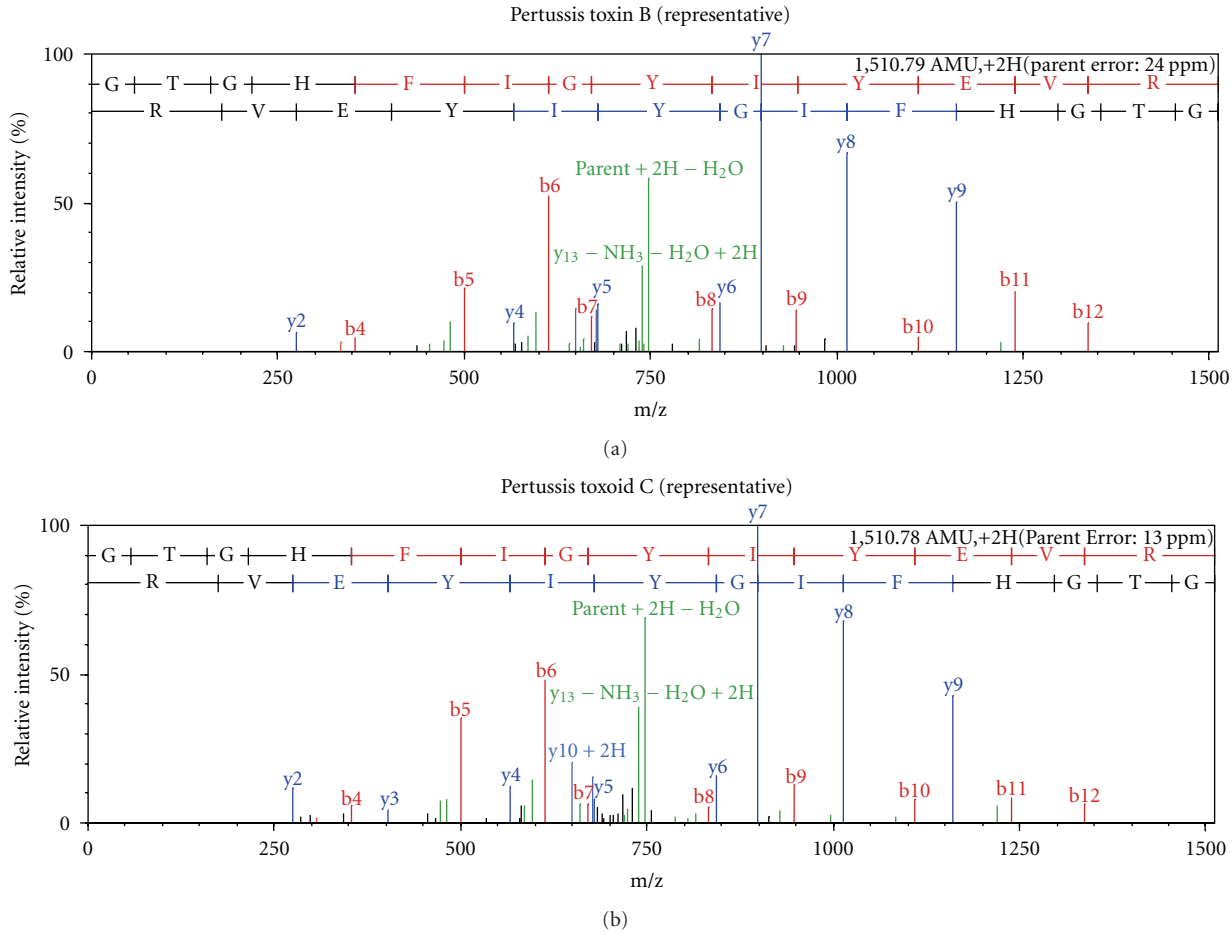


Figure 1: MS/MS spectra illustrating similar fragmentation profiles of Subunit 1 peptide GTGHFIGIYEVR from representative pertussis toxin (B) and pertussis toxoid (C). Nano Liquid Chromatography MS/MS were performed using the LTQ-Orbitrap on tryptic peptides obtained from pertussis toxin B and pertussis toxoid C. The MS/MS data was searched against the NCBI database and the peptide spectrum was analyzed [comparing relative peak intensity (0–100%) versus mass/charge (0 to 1500)] using the Scaffold software. The red and blue lines represent y- and b-ion distribution peaks, respectively.

Table 1: Comparative toxin and toxoid sequence analysis (% amino acid coverage).

	A-1	A-2	A-3	B-1	B-2	B-3	C-1	C-2	C-3	D-1	D-2	D-3
S1	30	58	26	35	69	55	36	49	61	23	41	58
S2	6	33	28	20	39	31	ND	14	16	ND	5	19
S3	ND	28	18	18	39	49	ND	12	22	ND	18	20
S4	20	53	13	41	57	42	ND	28	ND	ND	28	ND
S5	ND	ND	8	ND	ND	8	ND	ND	ND	ND	ND	8

-ND: Not Detected
-Pertussis toxin Subunits (S1–S5)
-Sources: Pertussis toxins (A and B) and Pertussis toxoids (C and D)
-Instruments: 1–MALDI-TOF MS (4800 Proteomics Analyzer), 2–ESI-MS (LTQ-Orbitrap), and 3–ESI-MS (Q-Tof)

proved to be more fruitful in terms of yielding a higher-sequence coverage and an increase in peptide detection. Specifically, S1 gave a higher-percent coverage by LC-MS/MS instruments resulting in a 1.7-fold higher-overall average as compared to the MALDI-TOF mass-spectrometric approach.

Tryptic digestions supplemented with RG resulted in an increase in sequence coverage in all samples and differentially within the subunits (Table 3). RG had a greater effect on the Ptxd giving rise to the greatest increase in sequence coverage. Ptxd D (2.2) resulted in a slightly greater overall fold change as compared to Ptxd C (1.6), in the presence of RG, while fold changes between all subunits were lower in both Ptx A (0.5) and Ptx B (1.3). Collectively for both Ptx and Ptxd, S1, S3, and S5 had marked gains in the presence of RG with a nearly 1.5, 2.5, and 10-fold increase, respectively.

As the sequence coverage increased in the presence of RG, the number of peptides detected by LC-MS/MS increased as well and to a lesser extent in MALDI-TOF (data not shown). Since the LC-MS/MS LTQ Orbitrap instrument gave the optimum sequence coverage (in particular for the toxins) in the presence of RG, a comparative assessment of peptide detection for Ptx subunits among all sources was compiled (Table 4). Table 4 reveals common and unique

Table 2: Mass spectrometry-detected peptides found among all pertussis toxins and toxoids.

	Peptides Detected	Toxin A	Toxin B	Toxoid C	Toxoid D
S1	DDPPATVYR (35–43)	X	X		X
	SCQVGSSNSAFVSTSSSR (74–91)		X	X	
	YTEVYLEHR (93–101)	X	X	X	X
	MQEAVEAER (102–110)	X	X	X	X
	GTGHFIGYIYEVR (114–126)	X	X	X	X
	ADNNFYGAASSYFEYVDTYGDNAGR (127–151)	X	X	X	X
	ILAGALATYQSEYLAHR (152–168)	X	X	X	X
	IPPENIR (170–176)	X	X	X	X
	VYHNGITGETTTTEYSNAR (181–199)	X	X	X	
	YVSQQTRANPNYTSR (200–215)			X	
	ANPNPYTSR (207–215)		X		
	SVASIVGTLVR (217–227)	X	X	X	X
	QAESSEAMAAWSER (239–252)	X	X	X	X
S2	ALTVAELR (56–63)	X		X	X
	GSGDLQEYLR (64–73)	X	X	X	X
	GWSIFALYDGTYLGGEYGGVIK (78–99)	X	X		
	DGTPGGAFDLK (100–110)	X	X	X	X
	NTGQPATDHYYSNVTATR (120–137)	X	X	X	
	LCAVFVR (146–152)				X
	SGQPVIGACTSPYDGK (153–168)	X	X	X	
	MLYLIYVAGISVR (179–191)	X			
S3	**ALFTQQGGAYGR (39–50)**	X	X	X	X
	ALTVAELR (57–64)	X		X	X
	GNAELQTYLR (65–74)	X		X	X
	QITPGWSIYGLYDGTYLGQAYGGIIK (75–100)		X	X	X
	DAPPGAGFIYR (101–111)	X	X	X	X
	ETFCITTIYK (112–121)	X	X	X	X
	TGQPAADHYYSK (122–133)	X	X	X	
	LLASTNSR (139–146)				X
	LCAVFVR (147–153)			X	
	DGQSVIGACASPYEGR (154–169)	X			
	DMYDALR (172–178)		X		
	LLYMIYMSGLAVR (180–192)	X	X		
S4	DVPYVLVK (43–50)	X			
	TNMVVTSVAMKPYEVTPTR (51–69)	X	X	X	X
	MLVCGIAAK (70–78)		X		
	LGAAASSPDAHVPFCFGK (79–96)	X	X		
	RPGSSPMEVMLR (100–111)	X	X	X	X
	QLTFEGKPALELIR (128–141)	X	X	X	
S5	NFTVQELALK (43–52)	X	X		X

-Detected peptide present in source (X-box); Detected peptide absent in source (white box)
-Common peptide (detected among all sources)
-Instruments: MALDI-TOF, LTQ-Orbitrap & Q-Tof Premier (−) Rapigest

peptides that were detected among Ptx and Ptxd sources analyzed. More importantly, an in-depth visual scan analysis of MS/MS spectra of the common peptides revealed similar fragmentation profiles (based on y- and b-ion distribution and relative intensity of fragment ions), suggestive of Ptx subunit sequence homogeneity as illustrated in Figure 1.

In addition to the determination of sequence homogeneity among sources, the relative amounts of protein as measured by relative abundance (fmol) of subunits from Ptx and Ptxd was preliminarily explored (Table 5). An LC-MS/MS approach was used, in which the Q-Tof MS^E data were employed for this assessment in which differential

Table 3: Comparative pertussis toxin and toxoid subunit sequence analysis (−/+ rapigest treatment).

Subunit	Source	− RG (% AA coverage)	# peptides detected	+ RG (% AA coverage)	# peptides detected
S1	Toxin A	58	10	57	10
	Toxin B	69	12	74	12
	Toxoid C	49	8	66	10
	Toxoid D	41	7	70	11
S2	Toxin A	33	4	31	4
	Toxin B	39	4	45	7
	Toxoid C	14	2	14	2
	Toxoid D	5	1	14	2
S3	Toxin A	28	5	49	8
	Toxin B	39	6	60	*9*
	Toxoid C	12	2	38	*6*
	Toxoid D	18	2	33	*5*
S4	Toxin A	53	4	28	2
	Toxin B	57	4	41	3
	Toxoid C	28	2	28	2
	Toxoid D	28	2	50	4
S5	Toxin A	—	—	10	1
	Toxin B	—	—	18	2
	Toxoid C	—	—	10	1
	Toxoid D	—	—	10	1

-Approach: Electrospray Ionization Mass Spectrometry-Instrument: LTQ-Orbitrap only-RG (Rapigest)

abundance was revealed among sources and the subunits. The absolute abundance of Ptx B was 7.2-fold higher among all subunits compared to Ptx A. Ptx B S4 gave the highest fold change at nearly 19 in comparison to Ptx A. With respect to Ptxd, the abundance was nearly the same for all of the subunits, with Ptxd D S1 revealing the highest fold change (1.5) versus Ptxd C.

4. Discussion

We initiated a study using mass spectrometry to determine if AA sequence differences exist between Ptx and Ptxd when examined from different sources. Using proteins (i.e., Ptx) from strains that may not share the exact same AA sequence or that have undergone different inactivation treatments could result in differences in immune effectiveness. Protein sequence variation between strains may be small (i.e., single AA addition, deletion, or substitution) or potentially more extensive with the addition or deletion of immunogenically relevant motifs or domain structures. As a result, protein sequence variance could lead to subtle structural, conformational, and/or functional differences that can be detected only through the use of sensitive technologies.

Previous MS investigations have been used for identification, differentiation, and characterization of protein complexes. For example, Nandakumar et al. [22] utilized MS techniques for identification of bacterial proteins. Protein mixtures were subject to tryptic in-solution digestion and analyzed by reverse phase nano LC-MS/MS followed by database analysis. Other investigators have used MALDI-TOF MS/MS for the analysis of peptides generated from the enzymatic cleavage of *Clostridium botulinum* C3 coenzyme [23]. Also, Yoder and colleagues analyzed tryptic digests of Vaccinia virus using MALDI-TOF, as well as a quadrupole ion trap mass spectrometer and a quadrupole TOF to identify proteins associated with infection [24]. Additionally, nano-LC-MS/MS was used to identify AA variability between two-bovine surfactant proteins [25]. In our study, an analytical and proteomic evaluation was conducted to assess any composition (i.e., AA sequence of the protein components) and formulation (i.e., how much of a specific component) disparities among Ptx and Ptxd.

Based on AA coverages and peptide detection alone, the two Ptxs and two Ptxds appear to have similar sequences, an indication that they were derived from the same origin, although different purification schemes and inactivation methods most likely were employed. The fact that the Ptx and Ptxd sources shared similar tryptic digest patterns, in terms of the peptides being generated and detected as well as spectrum profiles, suggests homogeneity. The presence of RG improved protein digestion resulting in higher-percent coverages generally across source as well as Ptx subunits. In addition, RG treatment appeared to have exposed the smallest-sized S5, resulting in an increase in coverage for this subunit. In doing so, additional peptides were detected that once again were shown to be similar among Ptx and Ptxd sources based on y- and b-ion distribution and relative intensities of the fragmented peptides. Moreover, studies by Sekura et al. [26] revealed, by amino acid analysis, that Ptx purified from Bp strain 165 was indistinguishable from Ptx prepared from the well-characterized Bp Tahoma strain. In addition, Tummala et al. confirmed Ptx sequence homogeneity among their Bp culture supernatant purified Ptx as compared with previously reported sequence information [19]. These data are confirmatory and strengthen the MS results in this study for Ptx sequence homogeneity. However, although the data suggest protein sequence homogeneity, if the sequence is indeed different but not present in the database then, upon searching, the generated peptides would not be detected resulting in a lower peptide map coverage.

Ptx is a major virulence factor and in its natural state has physiological inhibitory effects [11]. The subunits are highly immunoreactive, including the dominant S1, making it an ideal component for vaccines. To reduce the cellular lethality induced by Ptx, the Ptx is inactivated or detoxified in its final vaccine form. Inactivation occurs by chemical modification (i.e., formaldehyde or glutaraldehyde) via the formation of covalent crosslinkings on targeted AA (usually lysines). Interestingly in Ptxd samples, the oligomeric containing subunits percent AA coverage was lower. The S2–S5 subunits, rich in lysine are the subunits that are known to be heavily modified due to the chemical action exhibited by formaldehyde and/or glutaraldehyde [13, 27, 28]. As a result of modification, these peptides and subsequent MS-MS profiles would thus not be

Table 4: Peptide detection among pertussis toxin and toxoid subunits (−/+ rapigest treatment).

	− RG	+ RG
S1	GTGHFIGYIYEVR (114–126) YTEVYLEHR (93–101) MQEAVEAER (102–110) IPPENIR (170–176)	GTGHFIGYIYEVR (114–126) YTEVYLEHR (93–101) MQEAVEAER (102–110) IPPENIR (170–176) **VYHNGITGETTTTEYSNAR (181–199)** **ILAGALATYQSEYLAHR (152–168)** **MAPVIGACMAR (228–238)**
S2	GSGDLQEYLR (64–73)	GSGDLQEYLR (64–73) **NTGQPATDHYYSNVTATR (120–137)**
S3	ALFTQQGGAYGR (39–50)	ALFTQQGGAYGR (39–50) **QITPGWSIYGLYDGTYLGQAYGGIIK (75–100)** **GNAELQTYLR (65–74)** **DAPPGAGFIYR (101–111)**
S4	TNMVVTSVAMKPYEVTPTR (51–69)	TNMVVTSVAMKPYEVTPTR (51–69) **RPGSSPMEVMLR (100–111)**
S5		**NFTVQELALK (43–52)**

-**Peptides detected in +RG (rapigest) (in bold)** -Approach: Electrospray Ionization Mass Spectrometry-Instrument: LTQ-Orbitrap only

Table 5: Comparative pertussis toxin and toxoid subunit quantitative analysis.

Source	S1	S2	S3	S4	S5	Total (all subunits)
Toxin A	0.38	0.19	0.13	0.02	0.09	0.81
Toxin B	1.48	2.04	1.35	0.37	0.57	5.81
Toxoid C	1.88	0.34	0.32	ND	ND	2.54
Toxoid D	2.91	0.40	0.21	ND	0.06	3.54

-Measured as relative abundance (f mol/μL)
-ND: not detected
-Approach: Electrospray Ionization Mass Spectrometry
-Instrument: Q-Tof Premier only

detected using traditional selected database searching tools resulting in lower AA coverages.

Although formaldehyde, for instance, is an ideal reagent for chemical crosslinking, it also immobilizes the antigen (i.e., Ptx) and alters its conformation preventing a natural interaction between the antibody and antigen. In other words, formaldehyde reduces the accessibility of antibodies to their epitopes by creating steric hindrance or by altering the epitope site itself. In Ptxd samples, inactivation by formaldehyde appears to have a severe impact on peptide detection, due to crosslinking of lysine on the Ptx subunits [13, 27–29]. These subunits are most likely masked preventing trypsin, whose natural recognition sites are lysine and arginine, to effectively cleave the protein resulting in a lower AA coverage. On a cellular level, it is sensible for these subunits to be inactivated heavily, since in vivo, they make up the oligomeric complex that adheres to the human host cells which upon binding to the appropriate receptor initiates Ptx function. Blocking these cellular activities, in essence, inhibits or diminishes further downstream enzymatic activity by S1, which would allow for progression of normal immune responses by the host towards Bp. Besides, S1 has consistently had higher AA coverage reporting among all sources, which could be explained by the fact that S1 lacks lysines. Although S1 is deficient in lysine, free cellular lysine in the presence of formaldehyde would still create crosslinking bridges that lead to the detoxification [28, 29] but to a lesser extent compared to S2–S5. As a result of minimal chemical bombardment and thus potentially incomplete inactivation by formaldehyde on the toxic S1 component, residual toxicity may remain, with varying levels among Ptxd sources, could lead and be an explanative factor for differential immunity, albeit speculative.

5. Conclusions

In conclusion, this study incorporated multiple mass-spectrometric approaches to identify protein sequence similarities among Ptx and Ptxd sources, based on AA coverage profiles, peptide detection, and fragmentation profiles. Lower-percent coverage among Ptxd oligomeric subunits as compared to S1 could be explained most likely due to chemical modifications. As a result of these modifications, inactivation of Ptx subunits, as evident by failed detection of peptides, could lead to diminished immuno-cellular interactions between human-generated antibodies against chemically modified Bp antigens. In addition to the protein components, changes in the formulation, including the concentration, ratio and abundance of the Ptx subunits could translate into differences with respect to immune responses and vaccine efficacy. Moreover, information from this MS study could prove useful in improving immunological tests that use Ptx as an examining factor for vaccine evaluation and even pertussis diagnosis.

Future studies such as *de novo* sequencing or protein sequence tag analysis may be considered to validate and

strengthen the current finding that the protein sequences among all the sources are indeed homogeneous. In addition, incorporation of other enzymes in the presence of RG to expand the AA coverage and peptide detection may also be performed. Also, a comparative quantitative MS analysis assessing the functionality of Ptx and Ptxd sources may be able to determine the effect, if any, that Ptx inactivation, Ptx function, and potential residual toxicity play on vaccine efficacy such as immune responses. Lastly, in vivo animal studies may be conducted to examine the correlation between Ptx protein sequence homogeneity and immunoreactivity.

Disclaimer

The references in this paper to any specific commercial products, process, service, manufacturer, or company do not constitute an endorsement or a recommendation by the U.S. Government or the Centers for Disease Control and Prevention. The findings and conclusions in this paper are those of the authors and do not necessarily represent the views of CDC.

References

[1] M. Singh and K. Lingappan, "Whooping cough: the current scene," *Chest*, vol. 130, no. 5, pp. 1547–1553, 2006.

[2] J. B. Bobbins, R. Schneerson, B. Trollfors, et al., "The diphtheria and pertussis components of diphtheria-tetanus toxoids-pertussis vaccine should be genetically inactivated mutant toxins," *The Journal of Infectious Diseases*, vol. 191, no. 1, pp. 81–88, 2005.

[3] Q. He and J. Mertsola, "Factors contributing to pertussis resurgence," *Future Microbiology*, vol. 3, no. 3, pp. 329–339, 2008.

[4] World Health Organization, "WHO Recommended Standards for Surveillance of Selected Vaccine-preventable diseases," *WHO/V&B/02.01*, 2003.

[5] Centers for Disease Control and Prevention, "Pertussis vaccination: use of acellular pertussis vaccines among infants and young children: recommendations of the Advisory Committee on Immunization Practices (ACIP)," *Morbidity and Mortality Weekly Report*, vol. 46, no. RR-7, pp. 1–25, 1997.

[6] K. Kretsinger, K. R. Broder, M. M. Cortese, et al., "Preventing tetanus, diphtheria, and pertussis among adults: use of tetanus toxoid, reduced diphtheria toxoid and acellular pertussis vaccine recommendations of the Advisory Committee on Immunization Practices (ACIP) and recommendation of ACIP, supported by the Healthcare Infection Control Practices Advisory Committee (HICPAC), for use of Tdap among health-care personnel," *Morbidity and Mortality Weekly Report*, vol. 55, no. RR-17, pp. 1–37, 2006.

[7] K. R. Broder, M. M. Cortese, J. K. Iskander, et al., "Preventing tetanus, diphtheria, and pertussis among adolescents: use of tetanus toxoid, reduced diphtheria toxoid and acellular pertussis vaccines recommendations of the Advisory Committee on Immunization Practices (ACIP)," *Morbidity and Mortality Weekly Report*, vol. 55, no. RR-3, pp. 1–34, 2006.

[8] K. R. Powell, R. S. Baltimore, H. H. Bernstein, et al., "Prevention of pertussis among adolescents: Recommendations for use of tetanus toxoid, reduced diphtheria toxoid, and acellular pertussis (Tdap) vaccine," *Pediatrics*, vol. 117, no. 3, pp. 965–978, 2006.

[9] B. Taylor and R. Fahim, "Pasteur Merieux Connaught five-component acellular pertussis vaccine," *Biologicals*, vol. 27, no. 2, pp. 103–104, 1999.

[10] C. Locht, R. Antoine, and F. Jacob-Dubuisson, "*Bordetella pertussis*, molecular pathogenesis under multiple aspects," *Current Opinion in Microbiology*, vol. 4, no. 1, pp. 82–89, 2001.

[11] M. Tamura, K. Nogimori, S. Murai, et al., "Subunit structure of islet-activating protein, pertussis toxin, in conformity with the A-B model," *Biochemistry*, vol. 21, no. 22, pp. 5516–5522, 1982.

[12] J. Moss and M. Vaughan, "Activation of adenylate cyclase by choleragen," *Annual Review of Biochemistry*, vol. 48, pp. 581–600, 1979.

[13] S. Fowler, D. K.-L. Xing, B. Bolgiano, C.-T. Yuen, and M. J. Corbel, "Modifications of the catalytic and binding subunits of pertussis toxin by formaldehyde: effects on toxicity and immunogenicity," *Vaccine*, vol. 21, no. 19-20, pp. 2329–2337, 2003.

[14] E. S. Bamberger and I. Srugo, "What is new in pertussis?" *European Journal of Pediatrics*, vol. 167, no. 2, pp. 133–139, 2008.

[15] J. O. Lay Jr., "MALDI-TOF mass spectrometry of bacteria," *Mass Spectrometry Reviews*, vol. 20, no. 4, pp. 172–194, 2001.

[16] C. Fenselau and F. A. Demirev, "Characterization of intact microorganisms by MALDI mass spectrometry," *Mass Spectrometry Reviews*, vol. 20, no. 4, pp. 157–171, 2001.

[17] G. Chen and B. N. Pramanik, "LC-MS for protein characterization: current capabilities and future trends," *Expert Review of Proteomics*, vol. 5, no. 3, pp. 435–444, 2008.

[18] D. Bottero, M. E. Gaillard, M. Fingermann, et al., "Pulsed-field gel electrophoresis, pertactin, pertussis toxin S1 subunit polymorphisms, and surfaceome analysis of vaccine and clinical *Bordetella pertussis* strains," *Clinical and Vaccine Immunology*, vol. 14, no. 11, pp. 1490–1498, 2007.

[19] M. Tummala, P. Hu, S.-M. Lee, A. Robinson, and E. Chess, "Characterization of pertussis toxin by LC-MS/MS," *Analytical Biochemistry*, vol. 374, no. 1, pp. 16–24, 2008.

[20] A. Keller, A. I. Nesvizhskii, E. Kolker, and R. Aebersold, "Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search," *Analytical Chemistry*, vol. 74, no. 20, pp. 5383–5392, 2002.

[21] A. I. Nesvizhskii, A. Keller, E. Kolker, and R. Aebersold, "A statistical model for identifying proteins by tandem mass spectrometry," *Analytical Chemistry*, vol. 75, no. 17, pp. 4646–4658, 2003.

[22] R. Nandakumar, N. Madayiputhiya, and A. F. Fouad, "Proteomic analysis of endodontic infections by liquid chromatography-tandem mass spectrometry," *Oral Microbiology and Immunology*, vol. 24, no. 4, pp. 347–352, 2009.

[23] M. V. Muetzelburg, F. Hofmann, I. Just, and A. Pich, "Identification of biomarkers indicating cellular changes after treatment of neuronal cells with the C3 exoenzyme from *Clostridium botulinum* using the iTRAQ protocol and LC-MS/MS analysis," *Journal of Chromatography B*, vol. 877, no. 13, pp. 1344–1351, 2009.

[24] J. D. Yoder, T. S. Chen, C. R. Gagnier, S. Vemulapalli, C. S. Maier, and D. E. Hruby, "Pox proteomics: mass spectrometry analysis and identification of Vaccinia virion proteins," *Virology Journal*, vol. 3, article 10, 2006.

[25] S. Liu, L. Zhao, D. Manzanares, et al., "Characterization of bovine surfactant proteins B and C by electrospray ionization mass spectrometry," *Rapid Communications in Mass Spectrometry*, vol. 22, no. 2, pp. 197–203, 2008.

[26] R. D. Sekura, F. Fish, C. R. Manclark, B. Meade, and Y. L. Zhang, "Pertussis toxin. Affinity purification of a new ADP-ribosyltransferase," *Journal of Biological Chemistry*, vol. 258, no. 23, pp. 14647–14651, 1983.

[27] M. J. Corbel, D. K. L. Xing, B. Bolgiano, and D. J. Hockley, "Approaches to the control of acellular pertussis vaccines," *Biologicals*, vol. 27, no. 2, pp. 133–141, 1999.

[28] J. B. Robbins, R. Schneerson, J. M. Keith, J. Shiloach, M. Miller, and B. Trollors, "The rise in pertussis cases urges replacement of chemically-inactivated with genetically-inactivated toxoid for DTP," *Vaccine*, vol. 25, no. 15, pp. 2811–2816, 2007.

[29] C. Montero, "The antigen-antibody reaction in immunohistochemistry," *The Journal of Histochemistry and Cytochemistry*, vol. 51, no. 1, pp. 1–4, 2003.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 360652, 7 pages
doi:10.1155/2010/360652

Research Article

Factors Associated with Hepatitis A Vaccination Receipt in One-Year-Olds in the State of Michigan

Andrea L. Weston[1,2] and Kyle S. Enger[1]

[1] *Division of Immunization, Michigan Department of Community Health, Lansing, MI 48909, USA*
[2] *Department of Epidemiology, University of Michigan, Ann Arbor, MI 48109-2029, USA*

Correspondence should be addressed to Andrea L. Weston, weston.andrea@gmail.com

Received 25 June 2009; Accepted 16 September 2009

Academic Editor: Robert T. Chen

The objectives of this study were to determine factors associated with hepatitis A vaccination and to assess overall hepatitis A vaccination coverage levels among one-year-olds in Michigan. The study population was the first hepatitis A vaccination-eligible birth cohort ($n = 134\,226$) enrolled in the Michigan Care Improvement Registry (MCIR) after 2006 recommendations were made to routinely vaccinate all one-year-olds. All children whose first birthday occurred on or between May 1, 2006 and April 31, 2007 were included in the study population. Racial/ethnic minorities had increased odds of receiving the hepatitis A vaccination in Michigan, and Medicaid and WIC status modified this relationship. Fully understanding these relationships will be useful in targeting vaccination outreach and education programs.

1. Introduction

Hepatitis A was one of the most frequently reported vaccine-preventable diseases in the United States in the 1990s [1]. Typically, infection rates in the U.S. were higher in children under six years of age, adults over 50 years of age, American Indians, Alaska Natives, Hispanics, and lower socioeconomic groups [1–7]. Since hepatitis A vaccination introduction, infection rates in the U.S. have dropped drastically, especially in groups with the highest disease incidence historically [3–5, 8].

The Advisory Committee for Immunization Practices (ACIP) made initial recommendations in 1996 to vaccinate all high-risk groups (i.e., travelers and IV drug users), and the target populations have been expanded several times [9–11]. The 1999 recommendations included routine vaccination of children ≥2 years of age living in areas with hepatitis A incidence consistently ≥2 times the national average of 10 cases per 100 000; this included 11 states located in western and southwestern regions of the United States [10]. In May, 2006, the ACIP recommended routine vaccination of all 12–23 month old children in the United States with 2 doses at least 6 months apart [11].

In 2003, data collected from the National Immunization Survey (NIS) showed 50.9% coverage (≥1 dose) among 24 to 35 month old children in the 11 states where recommendations for routine vaccination had been made versus only 1.4% of children in the same age group in the rest of the country [2]. Some studies have found significant postrecommendation increases in vaccination receipt and decreases in hepatitis A incidence among American Indians and Alaska Natives [2, 5]. Other groups that have had significantly increased odds of receiving the hepatitis A vaccination include women, people with public (as opposed to private) health providers, children living in urban areas, and children born to mothers with lower levels of education [2, 7, 12, 13].

Michigan Care Improvement Registry (MCIR), Michigan's Immunization Information System (IIS) has collected data on hepatitis A vaccination in one-year-olds since the recommendation has been implemented in Michigan. MCIR is one of the leading IISs in participating number of children younger than 6 years, and in participating public and private vaccination provider sites [14, 15].

The primary goal of this study was to determine factors associated with hepatitis A vaccination in one-year-olds in

Michigan after the 2006 routine recommendations were implemented. It was hypothesized that being Hispanic, being American Indian, being of lower socioeconomic status (by Medicaid and WIC status), and living in nonrural areas would all increase the odds of vaccination receipt in Michigan, similar to findings in other states. A secondary objective was to assess overall hepatitis A vaccination coverage levels among one-year-olds in Michigan.

2. Materials and Methods

2.1. Study Population. The study population contained 134 226 one-year-olds in MCIR who had their first birthday on or between May 1, 2006 and April 31, 2007. These children were the first eligible cohort of one-year-olds for hepatitis A vaccination following ACIP's May 2006 recommendations for universal vaccination.

2.2. Data Source. This study uses data from the Michigan Care Improvement Registry (MCIR). At the end of 2007, more than 95% of children 19–35 months of age had at least two recorded vaccinations in MCIR, and approximately 91% of the 2255 Michigan childhood vaccination providers submitted data to MCIR from July to December 2007 [16]. Coverage estimates from the MCIR generally fall within the 95% confidence limits of Michigan's vaccination coverage measures in 19–35 month olds according to the National Immunization Survey (K. Enger, unpublished observations).

2.3. Outcome and Predictor Variables. The primary outcome variable was receipt of at least one dose of hepatitis A vaccination by the time the child turned two. The explanatory variables tested were gender, race/ethnicity, mother's age, number of months the vaccination recommendation was in effect by the child's first birthday, provider type (public versus private), Medicaid status, WIC status, rurality, and geographic region within the state, defined by MCIR region (Figure 1). Medicaid and WIC status were defined as anyone enrolled in Medicaid and/or WIC at the time the child turned one (when they were first eligible for vaccination receipt). The rurality variable was created using Categorization F of the University of Washington's Rural-Urban Commuting Area Codes (RUCAs) [17]. This classification system uses zip code and census tract commuting information to create categorizations of urban and rural areas. Categorization F created a rural category where only those areas that have little to no commuting of residents to larger areas for recreation or work are classified as rural. This can be thought of as "nonurban and nonlarge rural" [17].

2.4. Analysis. Descriptive statistics were used to describe the study population and assess vaccination coverage following the recommendations. Differences in child characteristics by vaccination status were assessed using chi-square tests for categorical variables and ANOVAs for continuous variables. Unadjusted logistic regression models were also used as a means of calculating odds ratios and 95% confidence intervals to assess the relationships between covariates and hepatitis A vaccination receipt. A main effects multivariable logistic regression model was used to determine adjusted measures of association, adjusted for all covariates with $P < .05$ at the bivariate level. All possible two-way interactions of race/ethnicity with Medicaid/WIC status and two-way interactions between race/ethnicity and provider type were assessed in other logistic regression models. The main effects model was compared with the full model (containing the interaction terms) using the likelihood ratio test to determine if an interaction effect was present.

All statistical analyses were conducted using SAS version 9.1. This study was approved by the institutional review boards of the Michigan Department of Community Health and the University of Michigan.

3. Results

There were 134 226 one-year-olds included in these analyses; 15 540 (11.6%) were excluded from multivariable analyses due to missing values for ≥1 variables. Of all MCIR-enrolled one-year-olds, 55.8% ($n = 74885$) received at least one dose of the vaccine by the time of their second birthday. Specific coverage rates varied by race/ethnicity, Medicaid and WIC status, mother's age, time the recommendation had been in effect by the child's first birthday, and geographic location (Table 1). By the time of a child's second birthday, only 0.06% ($n = 81$) received a full two-dose series.

The multivariable logistic regression model with interaction terms for race/ethnicity and Medicaid/WIC status was significantly different from the main effects model, based on the log-likelihood test with $P < .05$, and will therefore be the only multivariable model reported. In the multivariable model (Table 2), the odds of vaccination differed significantly by race/ethnicity, provider type, Medicaid/WIC enrollment, residence in a rural ZIP code, and by region within Michigan. Compared to one-year-olds who were white, each of the minority race/ethnic groups had statistically higher odds of receiving a vaccination: Asian and Asian Pacific Islander (Odds ratio [OR] = 1.75; 95% Confidence Interval [CI]: 1.58, 1.95), black (OR = 1.09; 95% CI: 1.01, 1.29), and Hispanic (OR = 1.16; 95% CI: 1.07, 1.26) (Table 2). Children visiting a public provider were more likely to receive the vaccination than children seeing a private provider (OR = 1.19; 95% CI: 1.13, 1.25) (Table 2). Children on Medicaid and WIC together had increased odds (OR = 1.38; 95% CI: 1.31, 1.44) of vaccination receipt when compared to children receiving neither service (Table 2). Children living in rural areas had decreased odds (OR = 0.88; 95% CI: 0.80, 0.96) of receiving a vaccination, when compared to children living in nonrural areas (Table 2). Northern, less-populated areas of Michigan (regions 3, 5, and 6) were less likely to be vaccinated than southern, more densely populated areas (regions 1, 2, and 4) (Table 2 and Figure 1).

Significant interaction was also present between Hispanic or black race/ethnicity and Medicaid/WIC enrollment (Table 2). Among children with Medicaid, Hispanic children had 1.83 (95% CI: 1.40, 2.41) times the odds of being

Figure 1: Regional map of the State of Michigan, as defined by Michigan Care Improvement Registry (MCIR).

vaccinated than white children. Among children with WIC, Hispanic children (OR = 1.35; 95% CI: 1.10, 1.67) and black children (OR = 1.17; 95% CI: 1.01, 1.34) had increased odds of vaccination receipt when compared to white children (Table 2). Compared to white children on both Medicaid and WIC, black children (OR = 1.17; 95% CI: 1.05, 1.29) and Hispanic children (OR = 1.32; 95% CI: 1.12, 1.56) again had increased odds of receiving the vaccination (Table 2).

4. Discussion

Hepatitis A vaccination was statistically associated with race/ethnicity, provider type, region, mother's age, time the recommendation had been in effect by the child's first birthday, and rurality in Michigan. Hispanic and black children receiving Medicaid, WIC, or both have increased odds of receiving the vaccination. These results are in contrast to the general trend of lower vaccination coverage among minority groups [18] and indicate that something may be differentiating promotion and administration of the hepatitis A vaccine from other vaccines. These results suggest that WIC and Medicaid may be promoting hepatitis A vaccination more effectively for Hispanic and black children than for white children.

Our results are consistent with the previous findings showing that Hispanic children, children with public insurance (and probably lower socioeconomic status), and children who live in nonrural areas have increased odds of hepatitis A vaccination receipt [2, 7, 12, 13]. One possible explanation for this could be that some groups of children are being targeted more because they have historically had a high risk for the disease [1]. Another possible explanation for our results could be physicians having different knowledge about the vaccine and/or the recommendation [19].

The high coverage rate (55.79%) with 1 or more doses of hepatitis A vaccination in the first cohort of one-year-olds eligible for routine vaccination is remarkable. For example, the CDC recommended that all 6 to 23 month year olds in the U.S. should be vaccinated for influenza for the 2002-2003 season, but only 7.4% of that age group received at least 1 influenza vaccination [20]. Furthermore, in the 2006-2007 influenza season, only 31.8% of that age group had received at least 1 influenza vaccination [20, 21]. In Michigan, hepatitis A coverage has also risen much more rapidly than varicella vaccination. Although varicella and hepatitis A vaccinations are both recommended at the first birthday, it took three years following varicella recommendations for coverage to reach 39.6% [22]. While less than 1% of these children received a full two-dose series by the time of their second birthday, we believe this is probably explained in large part by the recommendation stating vaccinations to be at least 6 months apart, so many children probably go longer than the 6 month minimum before receiving the second dose. It is important to note, however, that the 6-month interval between first and second doses is longer than other recommended vaccination intervals, which may lead to failure of followup of the second dose. Many children may

Table 1: Baseline characteristics, by hepatitis A vaccination receipt[a] coverage.

Characteristic	Percentage with characteristic ($n = 134\,226$)	Coverage (%) with ≥1 dose of the vaccination	P-value*	Unadjusted odds ratio (95% CI)[d]
Entire population	100.0	55.8%	NA	
Categorical variables				
Gender, female	48.8%	56.1%	.08	1.02 (1.00, 1.04)
Race/Ethnicity				
American Indian	0.6%	58.8%	.14	1.12 (0.96, 1.29)
Asian/Pacific Islander	3.2%	62.6%	<.001	1.32 (1.24, 1.40)
Black (non-Hispanic)	16.6%	62.5%	<.001	1.37 (1.33, 1.41)
Hispanic	6.2%	65.3%	<.001	1.51 (1.44, 1.58)
White (non-Hispanic)	67.2%	53.4%	<.001	0.67 (0.65, 0.69)
Mother's Age				
<20 years	9.0%	57.5%	<.001	1.08 (1.05, 1.12)
20–34 years	72.5%	55.9%	<.001	1.02 (1.00, 1.04)
>35 years	12.6%	53.0%	<.001	0.88 (0.85, 0.91)
Provider type, Public				
Public	6.4%	60.7%	<.001	1.17 (1.12, 1.22)
Private	41.4%	56.9%	<.001	0.86 (0.82, 0.90)
Medicaid/WIC enrollment				
Medicaid only[b]	6.2%	57.1%	.01	1.06 (1.01, 1.11)
WIC only[b]	10.6%	55.3%	.19	0.98 (0.94, 1.01)
Both Medicaid and WIC[b]	30.3%	64.1%	<.001	1.64 (1.60, 1.68)
Neither Medicaid nor WIC[b]	52.9%	51.0%	<.001	0.66 (0.65, 0.67)
Rural residence	3.6%	51.5%	<.001	0.79 (0.74, 0.83)
MCIR Region				
Region 1	45.9%	58.5%	<.001	1.10 (1.08, 1.12)
Region 2	24.4%	58.7%	<.001	1.09 (1.05, 1.11)
Region 3	5.8%	51.5%	<.001	0.78 (0.74, 0.82)
Region 4	10.1%	59.6%	<.001	1.11 (1.07, 1.15)
Region 5	6.9%	49.5%	<.001	0.71 (0.68, 0.74)
Region 6	2.4%	47.3%	<.001	0.66 (0.62, 0.71)
Continuous variable	Mean for unvaccinated population (SD)	Mean for vaccinated population (SD)	P-value	Unadjusted odds ratio (95% CI)
No. of months recommendation in effect[c]	5.5 (3.2)	5.8 (3.2)	<.001	1.03 (1.02, 1.03)

[a]Receipt of vaccination means the child has received one or more doses by the time of their 2nd birthday.
[b]Medicaid and WIC enrollment by the time of the child's first birthday.
[c]No. months = number of months vaccination recommendation was in effect by child's 1st birthday.
[d]Children with the given characteristic compared with children lacking the given characteristic.
*P-values are from chi-square tests for categorical variables and ANOVA for continuous variables.

be visiting a doctor according to the American Academy of Pediatrics' well child visit schedule [23], and if a child visits a doctor just shortly before they are eligible to receive the second vaccination, they may go many months before seeing a doctor again, and the opportunity to vaccinate with the second dose may be missed.

The rapid uptake of hepatitis A vaccination may be due to the existence of the vaccine well before recommendations were implemented, improved vaccination promotion (including assessment of hepatitis A in MCIR), the degree of use prior to generalized recommendations for all children, perceived differences in severity between varicella and hepatitis disease or some combination of these. Another reason for the rapid uptake could be MCIR itself. As one article put it concisely, "IISs are among the most mature public health information systems that bridge the public health/clinical care divide" [15]. Within the last 20 years, vaccination schedules for children have gone from providing protection from seven infectious diseases, and a total of 11 doses in the 1980s, to providing protection from 16

Table 2: Adjusted odds ratios and 95% confidence intervals of hepatitis A vaccination receipt[a] among 1-year-olds given by the final logistic regression model*.

Variable	Odds Ratio	95% CI
No. months[b]	1.02	(1.02, 1.03)
Race/Ethnicity		
White (non-Hispanic)	Reference	Reference
American Indian	1.24	(0.89, 1.74)
Asian/Pacific Islander	1.75	(1.58, 1.95)
Black (non-Hispanic)	1.09	(1.01, 1.17)
Hispanic	1.14	(1.01, 1.29)
Mother's age		
20–34 years	Reference	Reference
<20 years	1.06	(1.00, 1.13)
>35 years	0.98	(0.93, 1.03)
Provider Type		
Private	Reference	Reference
Public	1.19	(1.13, 1.25)
Medicaid and WIC enrollment		
Neither service	Reference	—
Medicaid only[c]	1.14	(1.06, 1.23)
WIC only[c]	0.98	(0.91, 1.05)
Both services[c]	1.38	(1.31, 1.44)
Geographic location		
Rural residence	0.88	(0.80, 0.96)
MCIR regions 1, 2 and 4	Reference	Reference
MCIR region 3	0.75	(0.70, 0.81)
MCIR region 5	0.75	(0.70, 0.80)
MCIR region 6	0.59	(0.53, 0.66)
Interaction Variables	OR	95% CI
Medicaid enrollment x		
Hispanic	1.83	(1.40, 2.41)
WIC enrollment x		
Black	1.17	(1.01, 1.34)
Hispanic	1.35	(1.10, 1.67)
Medicaid & WIC enrollment x		
Black	1.17	(1.05, 1.29)
Hispanic	1.32	(1.12, 1.56)

[a]Receipt of vaccination means the child has received one or more doses before their 2nd birthday.
[b]No. months = number of months since the May 1, 2006 *vaccination* recommendation of the child's 1st birthday.
[c]Medicaid and WIC enrollment on the child's first birthday.
*Logistic regression model run with a sample size of 118686, due to missing/unknown values; all two-way interactions for race/ethnicity and Medicaid/WIC enrollment and race/ethnicity and provider type were tested for, but only those with significant P-values (<.05) were included in the final model.

infectious diseases and a total of 30 to 40 doses in 2007 [24]. IISs allow all providers to have access to the most up to date vaccination information for every child to whom they are providing care, which helps keep track of such a complex vaccination schedule [24].

One reason cited for low vaccination rates is often missed opportunity [25]. If a child is in to see a physician for a reason other than routine vaccinations, the physician may not take that opportunity to vaccinate the child. MCIR and other IISs play an important role in reducing missed opportunity because they can be used in to produce vaccination reminders when it is time for a child to come in, and recall notices can be produced when a child is late for his/her vaccination [25]. Such reminder and recall notices can also be generated to send directly to homes, so parents are aware that a child is due for a vaccination [25]. Studies looking at offices where reminders and recalls are regularly used have an overall increase in vaccination coverage for children from all different population subgroups [25].

As can be seen in Table 1, region 6 had quite a remarkably lower vaccination rate that region 4. We believe this may be partially due to region 6 being more rural and being "whiter."

However, these things were adjusted for in multivariable analyses, and the difference in vaccination receipt still persisted. Lower vaccination receipt in region 6 may also be explained in part by differences in the culture and/or beliefs between providers and residents in different regions. For example, there may be difference in beliefs about the need for or effectiveness of vaccines between regions, but MCIR does not collect any information on this. We also do not have any data on vaccine promotion by region, and this may be a factor influencing differences between regions.

4.1. Strengths and Limitations. A strong point of this study is the high quality of the data from MCIR. At the end of 2007, more than 95% of children 19–35 months of age had at least two recorded vaccinations in MCIR, and approximately 91% of the 2255 Michigan childhood vaccination providers submitted data to MCIR from July to December 2007 [16]. With such high participation rates, and a large sample size, we are confident that Michigan's population of one-year-olds was well represented in this data set.

The MCIR has some limitations that tend to bias its vaccination coverage measures downward. Although reporting of childhood vaccinations is legally required [26] in Michigan, some vaccinations go unreported. Newly required vaccinations may also be subject to more incomplete reporting or more data entry error. MCIR is not always notified when children move out of Michigan. Therefore, if an unvaccinated child leaves the state and is subsequently vaccinated, that child could still be considered an unvaccinated Michigan resident in MCIR, further biasing vaccination coverage downward. Actual vaccination coverage levels in Michigan are probably slightly higher than estimates from MCIR.

Bias in vaccination reporting to MCIR is a concern because of inherent differences in the types of people who visit doctors, and in the type of doctors who consistently update MCIR information. Although participation in MCIR is good overall, local health departments tend to participate more fully than private health care providers, and some race/ethnic groups may be more likely to visit public providers. For example, a larger percentage of American Indian, Hispanic, and black one-year-olds received their vaccinations from public providers than white one-year-olds. In addition, southeast Michigan has had lower MCIR participation than the rest of the state; since 83% of the residents of the city of Detroit (which is in southeastern Michigan) are black [27], we cannot entirely exclude differential reporting of vaccinations based on race, geographic region, type of health care provider, or insurance status. It is difficult to know how these biases might affect our results.

The multivariable analyses in this study were limited to one-year-olds who were born in Michigan because some of the explanatory variables were provided to MCIR via Michigan's electronic birth certificate. If a child was born outside Michigan, MCIR would lack electronic birth certificate information, which would lead to missing values for some of the explanatory variables, causing that record to be excluded from the model.

Although this analysis found significant associations between hepatitis A vaccination and several factors, the final model has low pseudo-R^2 (Nagelkerke $R^2 = 0.01$). This indicates that other variables that were not included in our analysis may be needed in order to fully explain the variation in hepatitis A vaccination coverage. This indicates that further research may be necessary to determine more important factors associated with vaccination. Likely candidates include socioeconomic status and education, which we were not able to measure directly.

4.2. Conclusions and Future Directions. These data provide a baseline for hepatitis A vaccination coverage in Michigan one-year-olds following the recommendation of hepatitis A vaccination for children nationwide. As shown in this study, IISs are a valuable tool, as the data gathered within them can be used to see if there are specific subgroups of a population not getting a vaccination or getting a vaccination at higher rates than the rest of the population [25]. Data extracted from MCIR can also be used by public health officials to direct vaccine-specific education and outreach programs directly towards underserved populations [25]. Studies have shown when vaccination education programs are paired with other strategies like reminder/recall, vaccination coverage increases for all subgroups of a population [25].

Further outreach efforts may be needed to increase hepatitis A vaccination levels among those not receiving the vaccination, and these efforts should be specialized to this vaccine as its trends appear different from other vaccines. Michigan will continue to monitor trends in hepatitis A vaccination coverage. Future studies could examine potential underlying causes for increased and decreased odds of vaccination receipt in specific groups.

Acknowledgments

A. Weston was funded by the University of Michigan School of Public Health Department of Epidemiology as part of an externship with the Michigan Department of Community Health. The data was presented as a poster at the 2008 Annual Michigan Epidemiology Conference in Lansing Michigan, and as a talk at the July 2007 West Michigan Epi-Exchange Meeting. The authors would like to acknowledge Patricia Vranesich, Lynda Lisabeth and JoLynn Montgomery for their assistance in editing the manuscript. They have no disclosures or conflicts of interest.

References

[1] B. P. Bell, C. N. Shapiro, M. J. Alter, et al., "The diverse patterns of hepatitis A epidemiology in the United States—implications for vaccination strategies," *Journal of Infectious Diseases*, vol. 178, no. 6, pp. 1579–1584, 1998.

[2] J. J. Amon, N. Darling, A. E. Fiore, B. P. Bell, and L. E. Barker, "Factors associated with hepatitis A vaccination among children 24 to 35 months of age: United States, 2003," *Pediatrics*, vol. 117, no. 1, pp. 30–33, 2006.

[3] S. R. Bialek, D. A. Thoroughman, D. Hu, et al., "Hepatitis A incidence and hepatitis A vaccination among American Indians and Alaska Natives, 1990–2001," *American Journal of Public Health*, vol. 94, no. 6, pp. 996–1001, 2004.

[4] Committee on Infectious Diseases, "Hepatitis A vaccine recommendations," *Pediatrics* , vol. 120, no. 1, pp. 189–199, 2007.
[5] A. E. Fiore, A. Wasley, and B. P. Bell, "Prevention of hepatitis A through active or passive immunization: recommendations of the Advisory Committee on Immunization Practices (ACIP)," *Morbidity and Mortality Weekly Report*, vol. 55, no. RR-7, pp. 1–23, 2006.
[6] C. N. Shapiro, P. J. Coleman, G. M. McQuillan, M. J. Alter, and H. S. Margolis, "Epidemiology of hepatitis A: seroepidemiology and risk groups in the USA," *Vaccine*, vol. 10, supplement 1, pp. S59–S62, 1992.
[7] A. Wasley, A. Fiore, and B. P. Bell, "Hepatitis A in the era of vaccination," *Epidemiologic Reviews*, vol. 28, no. 1, pp. 101–111, 2006.
[8] A. Werzberger, B. Mensch, D. R. Nalin, and B. J. Kuter, "Effectiveness of hepatitis A vaccine in a former frequently affected community: 9 years' followup after the Monroe field trial of VAQTA®," *Vaccine*, vol. 20, no. 13-14, pp. 1699–1701, 2002.
[9] CDC, "Prevention of hepatitis A through active or passive immunization," *Morbidity and Mortality Weekly Report*, vol. 45, no. RR-15, pp. 1–30, 1996.
[10] CDC, "Prevention of hepatitis A through active or passive immunization: recommendations of the advisory committee on immunization practices (ACIP)," *Morbidity and Mortality Weekly Report*, vol. 48, no. RR-12, pp. 1–37, 1999.
[11] CDC, "Prevention of hepatitis A through active or passive immunization: recommendations of the advisory committee on immunization practices (ACIP)," *Morbidity and Mortality Weekly Report*, vol. 55, no. RR-7, pp. 1–23, 2006.
[12] A. Fiore, L. C. Baxter, B. P. Bell, et al., "Hepatitis A 2004 vaccination in children. Methods and findings of a survey in two states," *American Journal of Preventive Medicine*, vol. 33, no. 4, pp. 346–352, 2007.
[13] E. C. Owen, K. M. Peddecord, W. Wang, et al., "Hepatitis A vaccine uptake in San Diego County: Hispanic children are better immunized," *Archives of Pediatrics and Adolescent Medicine*, vol. 159, no. 10, pp. 971–976, 2005.
[14] American Immunization Registry Association, "Registry profile—Michigan childhood immunization registry," January 2008, http://www.immregistries.org/public.php/ImmRegs/regMain.php?MainButton=Details&id=25.
[15] A. R. Hinman, G. A. Urquhart, and R. A. Strikas, "Immunization information systems: National Vaccine Advisory Committee Progress report, 2007," *Journal of Public Health Management and Practice*, vol. 13, no. 6, pp. 553–558, 2007.
[16] The National Center for Immunization and Respiratory Diseases, "Section IX. Immunization registr," January–December 2007, http://www2a.cdc.gov/nip/irar/grantee/giregistry2007.asp?projectID=522556.
[17] Rural Health Research Center, "Rural-urban commuting area codes," July 2007, http://depts.washington.edu/uwruca/.
[18] K. G. Wooten, L. T. Luman, and L. E. Barker, "Socioeconomic factors and persistent racial disparities in childhood vaccination," *American Journal of Health Behavior*, vol. 31, no. 4, pp. 434–445, 2007.
[19] S. Sabnis, A. J. Pomeranz, and J. Mao, "Physician beliefs and practices regarding the use of hepatitis A vaccine," *Wisconsin Medical Journal*, vol. 106, no. 4, pp. 211–214, 2007.
[20] T. A. Santibanez, J. M. Santoli, C. B. Bridges, and G. L. Euler, "Influenza vaccination coverage of children aged 6 to 23 months: the 2002-2003 and 2003-2004 influenza seasons," *Pediatrics*, vol. 118, no. 3, pp. 1167–1175, 2006.
[21] CDC, "Influenza vaccination coverage among children aged 6–23 months. United States 2006-2007 Influenza Season," *Morbidity and Mortality Weekly Report*, vol. 57, no. 38, pp. 1039–1043, 2008.
[22] CDC, "National Immunization Survey (NIS)—children only," October 2008, http://www.cdc.gov/vaccines/stats-surv/imz-coverage.htm#nis.
[23] American Academy of Pediatrics CoPaAM, "Recommendations for preventive pediatric health care," *Pediatrics*, vol. 96, no. 2, pp. 373–374, 1995.
[24] K. M. Hillenbrand, "What is going on with vaccines: keeping up with the childhood immunization schedule," *Journal of Public Health Management and Practice*, vol. 13, no. 6, pp. 544–552, 2007.
[25] Task Force on Community Preventive Services, "Recommendations regarding interventions to improve vaccination coverage in children, adolescents, and adults," *American Journal of Preventive Medicine*, vol. 18, no. 1, supplement 1, pp. 92–96, 2000.
[26] State Office of Administrative Hearings and Rules, Department of Community Health community public health agency childhood immunization registry, October 2008, http://www.state.mi.us/orr / emi/ admincode.asp?AdminCode=.asp?AdminCode = Single & Admin_Num = 32500161 & Dpt = CH &RngHigh=.
[27] U.S. Census Bureau. Detroit city, Michigan. American Fact Finder 2007, January 2008, http://factfinder.census.gov/home/saff/main.html?_lang=en.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 916525, 10 pages
doi:10.1155/2010/916525

Research Article

Immunization Milestones: A More Comprehensive Picture of Age-Appropriate Vaccination

Steve G. Robison,[1] Samantha K. Kurosky,[1] Collette M. Young,[1] Charles A. Gallia,[2] and Susan A. Arbor[2]

[1] *Immunization Program, State of Oregon Department of Human Services, 800 NE Oregon St., Suite 370, Portland, OR 97232, USA*
[2] *Division of Medical Assistance Programs, State of Oregon Department of Human Services, 500 Summer St. NE, E-35, Salem, OR 97301, USA*

Correspondence should be addressed to Steve G. Robison, steve.g.robison@state.or.us

Received 28 October 2009; Revised 7 February 2010; Accepted 4 March 2010

Academic Editor: Robert T. Chen

A challenge facing immunization registries is developing measures of childhood immunization coverage that contain more information for setting policy than present vaccine series up-to-date (UTD) rates. This study combined milestone analysis with provider encounter data to determine when children either do not receive indicated immunizations during medical encounters or fail to visit providers. Milestone analysis measures immunization status at key times between birth and age 2, when recommended immunizations first become late. The immunization status of a large population of children in the Oregon ALERT immunization registry and in the Oregon Health Plan was tracked across milestone ages. Findings indicate that the majority of children went back and forth with regard to having complete age-appropriate immunizations over time. We also found that immunization UTD rates when used alone are biased towards relating non-UTD status to a lack of visits to providers, instead of to provider visits on which recommended immunizations are not given.

1. Introduction

A challenge to increasing early childhood immunizations on a state or local level is the limited ability of standard summary measures of up-to-date (UTD) rates to identify barriers to improvement. Typically, early childhood UTD rates are based on the number of doses of recommended vaccines, by individual antigen or in total, that a cohort of children receive by either a fixed age or a fixed date of assessment [1]. For the National Immunization Survey (NIS), this represents the proportion of 19- to 35-month-old children having all recommended doses for up to 7 vaccine types. Similarly, health plans utilize a Healthcare Effectiveness Data and Information Set (HEDIS) immunization measure, which counts the number of doses received by 24 months of age [2, 3]. However, a low UTD rate among a population, derived using these measures, does not aid in determining why the rate is low [4]. The growing complexity of the early childhood immunization schedule, with up to 19 vaccine doses recommended across at least 6 visits by age 2, means that there are many points of time and many reasons by which children can fall behind on immunizations [5]. While immunization summary UTD rates for 2-year-olds can identify general problems, there is also a need for more detailed assessment tools to describe local immunization coverage and more specific vaccine usage [6].

One alternative method of evaluation is to consider age-appropriate vaccinations. A substantial body of prior work exists comparing summary UTD measures against more specific assessments either for complete antigen series or combinations when actually due without considering late catch-up [7–9], or in comparison of individual antigens and shots to when they are expected or late [10]. To the extent that early childhood immunizations are often used as a proxy for the quality of early childhood routine care, the timeliness of immunizations is a relevant measure—delayed or lagging immunizations may reflect other issues with early childhood care.

One perspective on age-appropriate immunizations is to track children's progress through immunization milestones between birth and age 2 [11]. Immunization milestones are the ages at which recommended immunizations first become late according to the schedule developed by the Advisory Committee on Immunization Practices (ACIP). These milestones occur at 3, 5, 7, 16, 19, and 24 months of age. When immunizations are tracked using this approach, children's progression through milestones unfolds as a story of falling behind and catching up with recommended doses. This is a beneficial method as it facilitates identifying provider failure to give all or some of the immunizations that are due at healthcare encounters (missed opportunities) or parental failure to bring children to providers for vaccination-eligible encounters (missed visits). The prevalence of missed opportunities and missed visits at each milestone age can guide immunization interventions. However, basing milestone analysis exclusively on immunization record data may misclassify missed opportunities as missed visits, and shift the apparent burden of children who are not appropriately immunized for their age from providers to parents. This can occur because healthcare encounters during which no vaccinations are administered will not be captured into immunization record datasets.

This study provides an example of using a milestone approach to assess a specific population and their progression through early childhood immunizations, where both payor-administrative and state-level immunization information system (IIS) data are available. Combining children's healthcare encounter information from payer records with their immunization records gives a more accurate assessment at each milestone age of the effect of missed opportunities and missed visits on age-appropriate immunizations and overall immunization rates.

2. Methods

2.1. Data Sources. The study population consisted of a birth cohort of Oregon children enrolled in the Oregon Health Plan (OHP) and whose immunization records were in the Oregon ALERT Immunization Information System (ALERT IIS). ALERT IIS immunization records were merged with provider encounter records from the OHP for this population. The ALERT IIS is a statewide immunization registry which receives immunization records from 97% of Oregon private healthcare providers and 100% of the immunization records from public providers. The OHP is Oregon's public healthcare plan that provides healthcare coverage for children in families living below 185% of the federal poverty level and covers both those with traditional Medicaid eligibility as well as an expanded State Children's Health Insurance Program (SCHIP) population. The majority of OHP-enrolled children are placed in commercially available managed care plans. Immunization records for OHP children are available both from OHP collected records for billing and encounters as well as from direct provider record submissions to ALERT.

2.2. Data Setup. For this study, ALERT IIS records and OHP encounter records were selected from their respective data systems for children born in 2005. Immunization records and encounter data were restricted to those received through the child's 24th month of age. Children's records were merged across the two data systems based on the child's name, date of birth, and county of residence. The matching process was based on the observation that within the 2005 Oregon birth cohort, a combination of name and date of birth was over 99.9% unique, with almost all exceptions resolving with the inclusion of residence. This high-probability matching process was selected over the usual process ALERT uses to incorporate OHP and other administrative data, wherein a hard-match is required also on additional information such as address or phone number.

Immunization records were selected for the six vaccines (including combination vaccines) included in the recommended "4:3:1:3:3:1" series. The 4:3:1:3:3:1 series consists of 4 diphtheria, tetanus toxoid, and acellular pertussis (DTaP); 3 poliovirus (IPV); 1 measles, mumps, and rubella (MMR), 3 *Haemophilus influenzae* type b (Hib); 3 hepatitis B (HepB), and 1 varicella. (VAR) vaccines. Records for these vaccines were reviewed for appropriate age and interval between doses, according to the 2007 ACIP Immunization Schedule. Doses given too early or with insufficient spacing between doses to be considered valid were removed from the analysis.

To ensure that records of encounters would be available to compare with immunizations at all of the milestone ages, children with limited enrollment or nonenrollment at key ages were excluded from the final analysis dataset. The study population was restricted to children enrolled in OHP within 30 days of birth, with a cumulative total of at least 365 days of enrollment by age 2, and continuous enrollment across the key period of 15 to 18 months, when the 4th dose of DTaP is due. Because OHP enrollment generally occurs in 12-month blocks, the majority of children meeting the above requirements were also continuously enrolled through their second birthday. Children born outside Oregon were excluded since possibly both early encounters and immunizations would not be reported to ALERT or OHP. Children with only a birth dose of hepatitis B and no other vaccines in ALERT also were excluded.

The OHP requires health plans and providers to submit detailed encounter records on enrolled children for all services received. For OHP-enrolled children, a subset of vaccination-eligible encounters was created from all encounters, based on a review of ICD-9 and CPT coding in OHP encounter records. A vaccination-eligible encounter was defined as an encounter occurring in a nonemergent or noninpatient setting with a medical provider, and with either a CPT procedure code indicating that routine care or evaluation was performed, or an ICD-9 diagnostic code indicating that the purpose of the encounter was consistent with routine care and immunization evaluation. These criteria were used to identify not only visits that providers would define as well-child visits, but also other visits during which immunizations could have been given, and include nonemergent "sick" visits. In a few cases, ALERT had a record of an immunization visit for which there was no

matching OHP encounter record. This was usually found to reflect free vaccinations without administration fees at sites outside of those normally reporting to OHP, such as school clinics, and some public health departments. These visits were also counted as vaccination-eligible encounters. Also if CPT codes for vaccine administration were found in OHP data without other evidence of a shot-eligible encounter, these were taken as indicating that a vaccine encounter occurred.

2.3. Analysis. The definitions of milestone periods were taken from Luman and Chu [11], and reflect the dates at which recommended immunizations are first late according to the 2007 ACIP schedule. The milestone periods occur at the start of 3 months, 5 months, 7 months, 16 months, 19 months, and 24 months of age. Immunization (UTD) status at each milestone was evaluated for the timely receipt of all doses due in the 4:3:1:3:3:1 series by age 2. Additionally children were categorized according to whether they had vaccination-eligible encounters in the period prior to each milestone. In cases in which a non-UTD child at a milestone had multiple encounters in the prior period, and received vaccinations at some encounters and not at others, they were counted as having a vaccination visit. Schematically, this classification is presented in Table 1.

At each milestone age, immunization status was compared with the immunization status at the previous milestone age to determine whether children had remained UTD, remained non-UTD, fallen behind due to a missed visit, fallen behind due to a missed opportunity, or caught up with the immunization schedule. Children remained UTD if they were UTD at the prior milestone age and UTD at the current milestone age. Children remained non-UTD if they were non-UTD at the prior milestone age and non-UTD at the current milestone age. This classification is presented in Table 2.

Children fell behind due to a missed visit if they were UTD at the prior milestone age, non-UTD at the current milestone age, and had no record of a valid healthcare visit or immunization record during the period in between. Children fell behind due to a missed opportunity if they were UTD at the prior milestone age, non-UTD at the current milestone age, and had a vaccination-eligible healthcare visit during the period in between. Children caught up if they were non-UTD at the prior milestone age and UTD at the current milestone age. An exception to the missed opportunity calculation is for the 16-month milestone, which includes the first MMR vaccine. The MMR is not valid before 12 months of age, so encounters between 7 months and 12 months were not counted as potential missed opportunities.

As a check on the completeness of immunization visits represented by this merged dataset, a Lincoln-Peterson capture-recapture method was used to estimate the percentage of immunization visits for the study population not captured by the ALERT IIS either by provider records or by OHP administrative records. The total number of immunization visits, both captured and uncaptured, was estimated by

$$N = \frac{[(A+1)\times(B+1)]}{(AB-1)}, \quad (1)$$

where N is the total estimated number of visits, A is the number of visits captured by provider reports in ALERT, B is the number of visits captured by OHP billing and administrative reports, and AB is the number of visits captured in both by date.

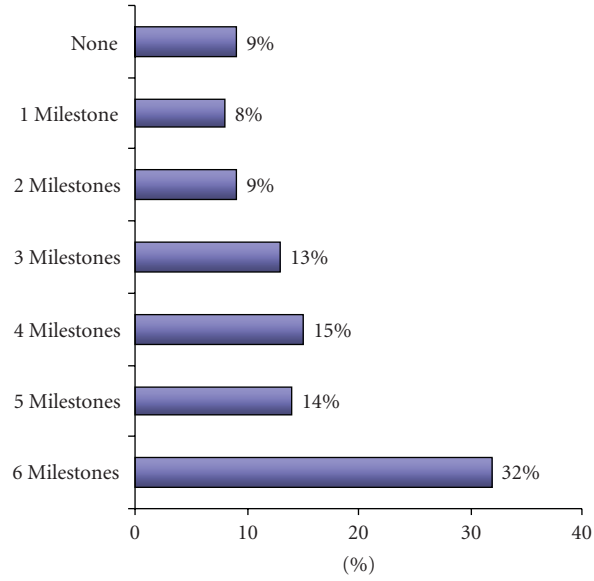
The principal assessment tool of this study is a time-based progression of young children across milestones and age-appropriate immunizations, presented in a novel form for easier depiction of change between milestones. The data by milestone are presented for whether children were complete on age-appropriate immunizations along with categories for catching up and falling behind by milestone, and by missed opportunities versus missed visits for non-UTD children. Finally, a comparison of age-appropriate milestone results is made to a summary UTD measure for the 4:3:1:3:3:1 immunization series assessed at 24 to 35 months of age.

3. Results

Of 20,411 children born in 2005 who were enrolled in the Oregon Health Plan for some period of time, 13,199 met the requirements to be counted among the study population. The reasons and scope of exclusions are presented in Table 3. The numbers of those excluded are listed in the order they were excluded and do not reflect the total prevalence of each criterion in the population; for example, of the 1,004 children excluded for being born out of Oregon, the majority would also have been excluded for length of enrollment. The primary reason for exclusions from the study population is nonenrollment after 1 year of age, so that no encounter data would be reported to OHP. Also 2% of the study population met all enrollment criteria except that they did not have any reported immunizations. These may represent true nonimmunized children, and were also excluded.

The count of total immunization visits across both datasets was 74,919. The capture-recapture estimate of total immunization visits for the study population was 76,087. Thus the dataset appears relatively complete for all immunization visits of the study population, with the combination of ALERT IIS provider reports and OHP administrative data capturing 98.4% of estimated immunization visits among the study population.

As shown in Figure 1, only 32% of children had all ACIP recommended immunizations on time at all milestones, while 14% were not complete at only one milestone, 15% were not complete at 2 milestones, and 9% were not complete at any milestone. Also 41% of children had vaccination-eligible encounters in all of the periods before each milestone, and 35% of children were missing encounters in only one of the periods before milestones. Another 14% were missing encounters in 2 periods, and 10% were missing encounters in 3 or more periods.



For milestone status based on having ACIP recommended immunizations in the DTaP, HepB, Hib, MMR, polio, and varicella series by 3, 5, 7, 16, 19 & 24 months.
Source: Oregon ALERT IIS

Figure 1: Children by Count of Milestones with All Age-Appropriate Immunizations.

Table 1: Immunization Encounter Classification.

	Had immunizations by milestone?		
Had provider encounters?	Yes, UTD	Yes, Non-UTD	No
Yes	Immunization Visit	Missed Opportunity	Missed Opportunity
No	—	—	Missed Visit

Where UTD is up to date status for ACIP recommended immunizations in the DTaP, HepB, Hib, Polio, and Varicella series by each milestone age.

Table 2: Immunization Status Categorization.

	At prior milestone	
At current milestone	UTD	Non-UTD
Up to date (UTD)	Remain UTD	Catch Up
Non-UTD	Fall Behind	Stay Non-UTD

UTD is up to date status for ACIP recommended immunizations in the DTaP, HepB, Hib, Polio, and Varicella series.

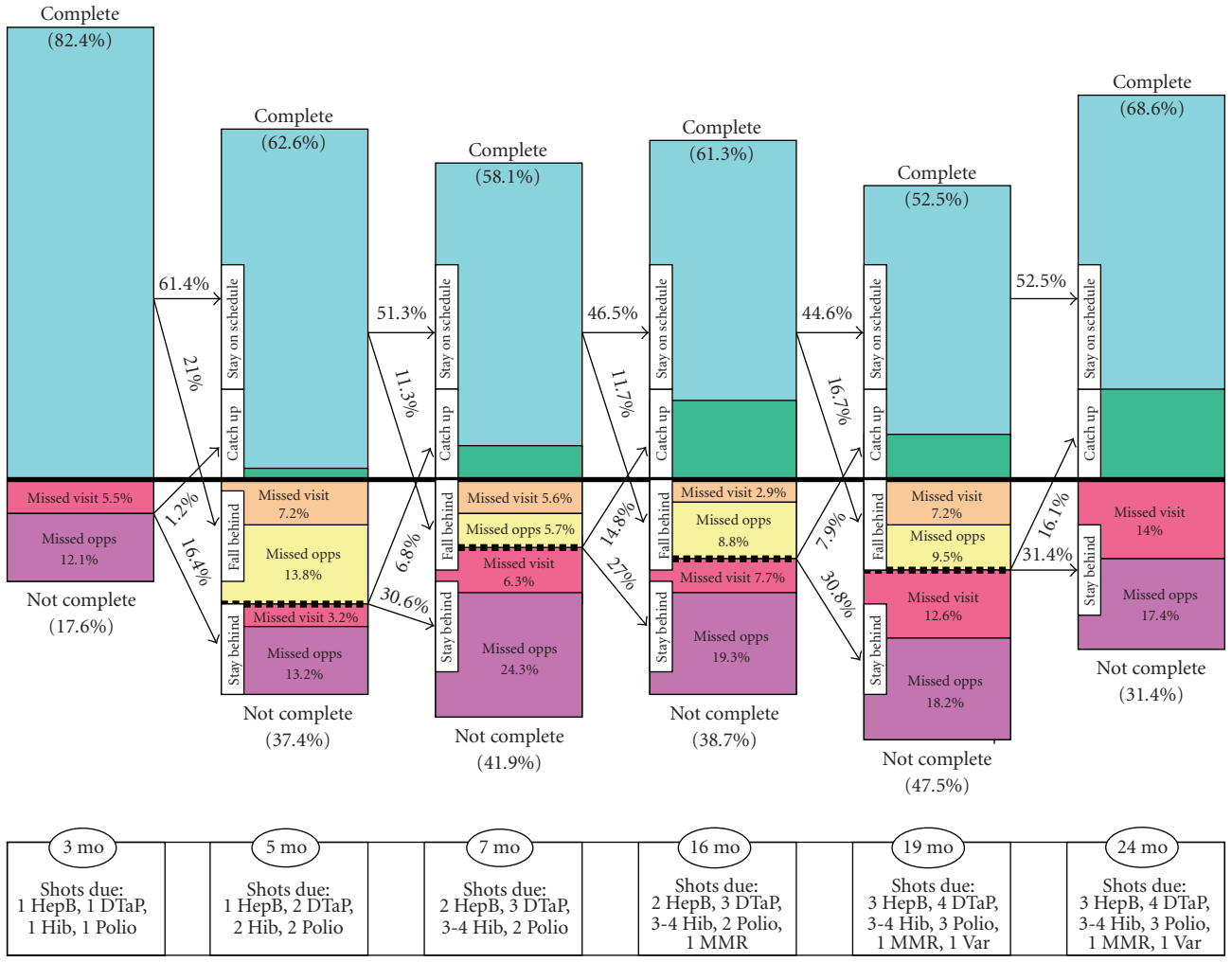
Table 3: Population Size and Exclusions.

OHP children in ALERT:	20,411
Exclusions (in order applied)	
Not born in Oregon	1,004
<180 days of enrollment by age 2	1,123
<365 days of enrollment by age 2	583
Not enrolled within 30 days of birth	345
Not enrolled after 1 year of age	2,884
Not enrolled through 15–18-month period	1,009
No immunizations reported post birth	264
Total exclusions	7,212
Total study population	13,199

OHP is the Oregon Health Plan ALERT is the ALERT Immunization Information System.

Rates of completeness of age-appropriate immunizations per the ACIP schedule varied among the milestone ages, from a high of 82.4% at 3 months of age to a low of 52.5% at 19 months of age. For the final milestone at 24-months, no further vaccinations were due, and the final 24 month completion rate for the study population was 68.6%, with 16.1% catching up from the prior 19-month milestone. The pattern of completion, falling behind, and catching up by milestone period is presented in Figure 2.

While 17.6% of the study population had fallen behind by the 3-month milestone, representing a late start on immunizations, the most salient episode of falling behind occurred at the 5-month milestone, where 21.0% of children fell behind. In this analysis, children fell behind at all milestone ages at which new immunizations were added. The 19-month milestone adds 4 antigens beyond the 16-month milestone, including the fourth DTaP and varicella; and the risk of falling behind between these milestones is calculated from Table 2 as (16.7/61.3) = 27.2%. This is interpreted as, for those who are on schedule at 16 months, 27.2% will fall behind by 19 months. The largest total percentage



Completeness is based on ACIP recommendations by age for diphtheria, tetanus toxoid, and pertussis (DTaP); poliovirus (IPV); measles, mumps, and rubella (MMR), *Haemophilus influenzae* type b (Hib); hepatitis B (HepB), and varicella (VAR) vaccines.
Source: Oregon ALERT IIS

Figure 2: Immunization Completeness At Milestone Ages (for Two Year Olds in 2007, $N = 13,199$).

of children without age-appropriate immunizations, 47.5%, also occurred at the 19-month milestone. Overall, the percentages of children who were not complete by milestone from Table 2 with missed opportunities was 68.5% at 3 months, 72.2% at 5 months, 71.7% at 7 months, 72.5% at 16 months, 58.4% at 19 months, and 60.0% at 24 months.

Another approach to interpreting the reasons for children falling behind is to further examine missed visits, missed opportunities, and vaccination visits at each milestone for those who were not complete. Figure 3 describes how children who are not complete for age-appropriate immunizations at any milestone have either missed visits or missed opportunities, where missed opportunities are divided between provider encounters with no shots received versus encounters where some shots are received.

The percentage of noncomplete children per milestone with provider encounters on which some shots were received (vaccination visits) ranged from a high of 50.8% prior to the 3-month milestone to a low of 15.2% before the 24-month milestone. The percentage of children with provider encounters and no shots, and without vaccination visits, in each period is a measure of the amount by which immunization-record-only data would misclassify missed opportunities as missed visits. This potential misclassified percentage of noncomplete children having vaccination-eligible encounters with no reported vaccinations ranged from a high of 44.8% at the 24-month milestone, to a low of 21.4% at the 5-month milestone. At the 3-month milestone, noncomplete children were evenly divided between those with no encounters on record during the period (31.4%), those with encounters but no vaccinations (34.1%), and those with encounters during which some vaccinations were given (34.5%).

As a final analysis, the results by milestone for encounters and completeness were stratified by children's status on a summary UTD measure for having all shots in a 4:3:1:3:3:1

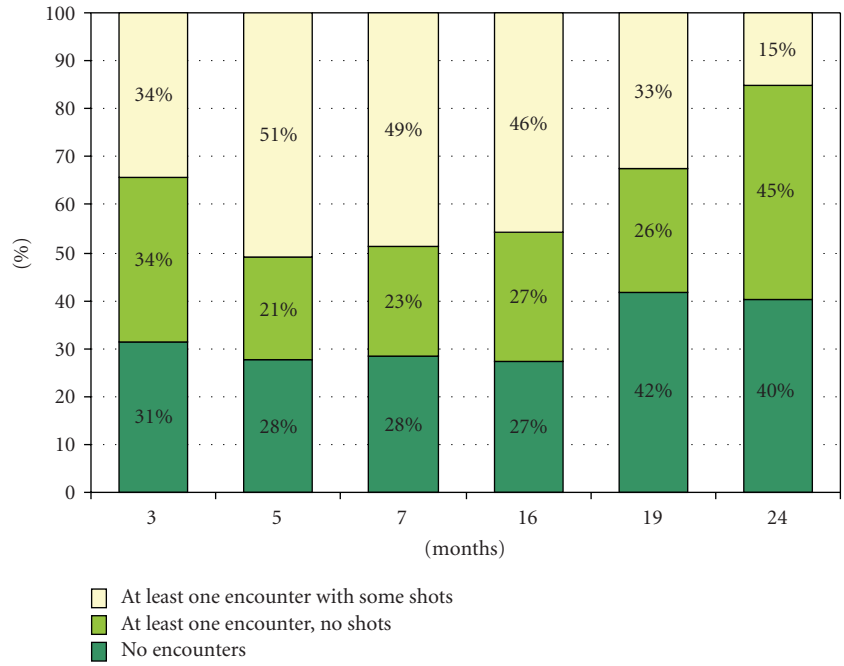


Figure 3: Provider Encounters by Milestones for Children Lacking Complete Age-Appropriate Immunizations.

series by age 24 to 35 months, and using a fixed date of assessment. Overall 77.8% of the study population were UTD by 24 to 35 months for the 4:3:1:3:3:1 series. The comparison across milestones for children who were UTD is presented in Figure 4.

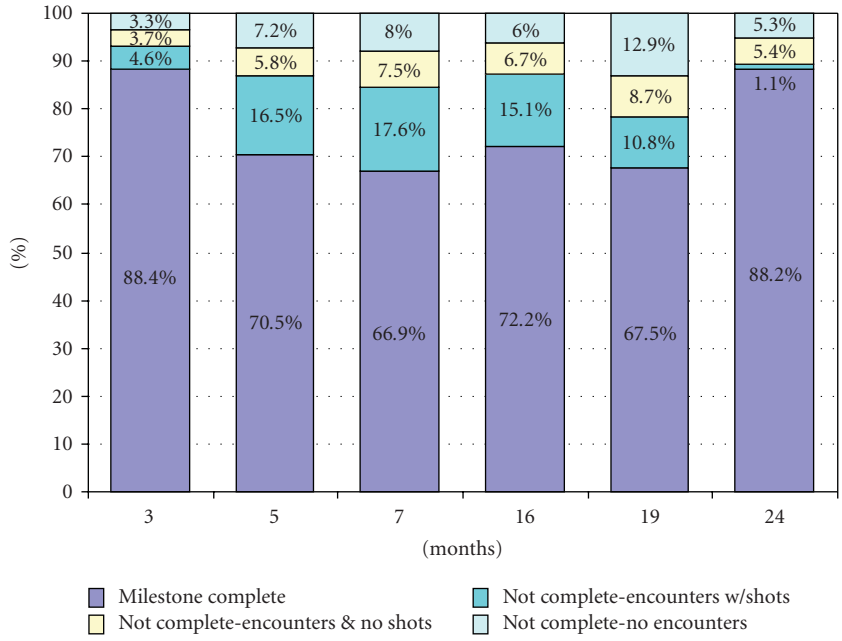
Of the children who were UTD by the date of assessment, only 68% were complete for age-appropriate immunizations by the 19-month milestone, 72% were complete by the 16-month milestone, 67% were complete at the 7-month milestone, and 71% were complete at the 5-month milestone. The majority of those who were not complete at any milestone, with the exception of the 24-month milestone, also had encounters with providers on which some shots were given.

Figure 5 presents the same analysis across milestones for the 22.2% of the study population who were not UTD for the 4:3:1:3:3:1 series at 24 to 35 months.

Children not UTD at 24–35 months were also not complete for age-appropriate immunizations at 19 and 24 months by definition of which shots were required at these milestones. Overall in Figure 5 the majority of non-UTD children at 24 to 35 months who also were not complete at milestones had substantial volumes of encounters with providers. An analysis using only shot-record data, however, would reach, falsely, the conclusion that the majority of non-UTD children were missing provider encounters at each milestone.

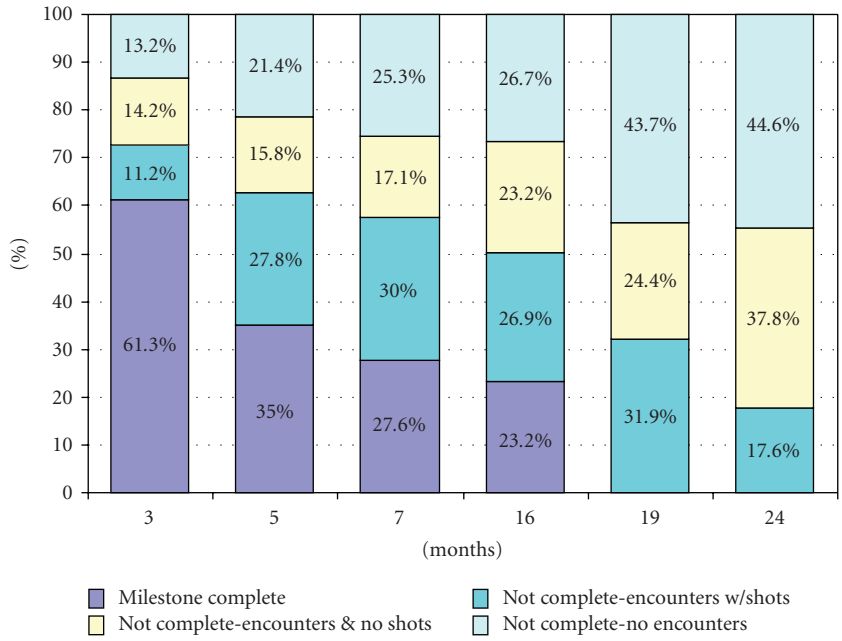
4. Discussion

The reality of children's immunizations in the present study population is a story of falling behind and catching up with recommended immunizations. This finding is consistent with prior application of a milestone age approach [11]. What this study adds to the understanding of milestones and immunizations is a more accurate representation of how missed opportunities and missed visits contribute to children not being up-to-date for recommended vaccines. The present finding that for many children, periods of missed visits in immunization record data are actually periods of missed opportunities is a first in the analysis of larger, population-based data systems such as immunization registries. This finding should lead at least to caution in assigning reasons regarding why children are not up-to-date according to either state-level immunization registries or other immunization record data, including the National Immunization Survey. The importance of this is that the prevalence of either true missed visits or true missed opportunities should lead to different interventions to improve immunization rates. Focusing on methods to improve rates of missed visits, such as reminder-recalls to parents of children who appear to have missing vaccinations and visits, may be of limited utility if the greater issue is that the parents have brought their children in to providers across milestones without receiving needed vaccinations. As a recommendation to



For milestone completeness based on having ACIP age-appropriate immunizations in the DTaP, HepB, Hib, MMR, polio, and varicella series by 3, 5, 7, 16, 19 & 24 months.
Source: Oregon ALERT IIS

Figure 4: Milestone Completeness and Provider Encounters for Children UTD for 4:3:1:3:3:1 Series by 24 to 35 Months.



For milestone completeness based on having ACIP age-appropriate immunizations in the DTaP, HepB, Hib, MMR, polio, and varicella series by 3, 5, 7, 16, 19 & 24 months.
Source: Oregon ALERT IIS

Figure 5: Milestone Completeness and Provider Encounters for Children Non-UTD for 4:3:1:3:3:1 Series by 24 to 35 Months.

correct this problem, immunization record data for at least a sample of covered children should be compared with encounter records from billing and administrative sources before considering appropriate immunization interventions. Also, data collected from samples of provider records for immunization assessment should include basic information on all encounters, whether vaccinations were given or not.

In this study, the majority of children who were not catching up at each milestone were having encounters with providers. These encounters were potentially ones in which missing vaccinations could be administered; however, converting these missed opportunities to vaccination visits may be difficult. Provider reluctance to administer vaccinations during sick or other nonroutine visits is a known barrier to improving immunization rates, and may be difficult to change [12–14]. The type of encounter may also be a barrier to receiving vaccinations in many clinics and healthcare providers; for example, in urgent care encounters, when limited time and a press of higher priority needs make it difficult to include review of records and delivery of vaccinations [15–17]. While parental reluctance in such circumstances is likely a factor, at least one study has found that the barrier in such visits is more likely to be provider-based than parental [18]. Also, reimbursement levels may not be sufficient to encourage providers to expand immunizations outside of well-child visits [19]. Finally parents who bring their children in for sick visits or other encounters without immunizations may easily believe that their child has received all needed care, including immunizations. Most parents of children who are not UTD believe their child has received all needed immunizations and may not understand the difference between well-child and other types of encounters [20]. Solutions to this problem may lie in the redesign of early childhood care encounters within clinics and healthcare providers that are concerned about their immunization rates, to deemphasize urgent care or access to short, single-purpose visits in favor of longer appointments during which aspects of routine care such as immunizations are also reviewed.

A strength of the present study is the combination of administrative data on all encounters with immunization records reported separately to the ALERT IIS. This approach could potentially serve as a standard for the evaluation of immunizations given to health plan participants and public populations in areas that have strong immunization information systems such as ALERT. A similar approach by Dombkowski et al. [21] has previously demonstrated the utility of combining registry and Medicaid data in Michigan for assessing missed opportunities to vaccinate asthmatics against influenza. The potential for missed opportunities to be misclassified as missed visits when conducting milestone analysis solely from immunization record data without all encounters should lead to caution in interpreting the balance of responsibility between parents and providers for children not being up-to-date. Also the present study does not address the extent to which parental reluctance to accept all age-appropriate immunizations may limit the ability of provider-based interventions to improve milestone immunization completeness.

From the perspective of a state immunization program with concerns for improving immunization rates, the development of a roadmap showing where and how children are falling behind is invaluable for setting policy. While national measurements such as the NIS can identify variations in rates between states, state-level programs have been left on their own to identify what in-state factors are affecting their rates [6]. Also because immunization levels are often taken as a measure for overall quality of care in early childhood [1], counting doses by age two is not as strong a proxy as is the checking of timely receipt of age-appropriate immunizations across the entire period from birth to age two. A risk of solely depending on immunization results at age two is that UTD and non-UTD status may be taken as discrete categories, irrespective of age-appropriate history. Searching for explanatory factors for these two categories may be misleading for developing an understanding of where barriers exist and where interventions are needed. This is illustrated in the present study by the observation that only a minority of children were consistently on schedule at all milestone ages, and that the majority fell behind at one or more point in receiving immunizations. Falling behind by milestone period is a more useful concept for intervention than final UTD status. A useful model then is that most children are at great risk of falling behind at many points, and that their final status reflects the work that providers do to catch them up to standard.

The concept of milestone ages and the charting of children's progress through the milestones, as advanced by Luman and Chu [11], provides such a roadmap for use by local programs. Local variations in patterns of falling behind and catching up, however, argue for analyzing milestones with available local or state-level data to determine where problems are most salient. Yet, while specific findings may differ, the present study confirms the utility of the milestone approach for a local population.

The present study is representative only of a single state population, and of children enrolled through the Oregon Health Plan. Because children in the OHP are generally enrolled in the same health plans, with the same networks of providers and benefits, as privately insured children, their encounters are potentially similar to the wider state population. Overall in 2007, the OIP estimated that among all OHP-enrolled 2-year-olds, the UTD rate for a 4:3:1:3:3:1 series was 75.2%, as compared to 72.9% among non-OHP-enrolled children. However, the present study population also reflects a group with stable, long-term enrollment in OHP. Children with short-term enrollment or who disenrolled after age 1 are not represented and may have substantially different patterns of falling behind and catching up to recommended immunizations. Also it is expected that individual health plans under the OHP are a significant factor in the receipt of age-appropriate immunizations; however, this information was not included in the present analysis dataset.

Another limitation on the present results is that the definition of encounters was deliberately set broadly, to reflect any encounters in which vaccinations could have been delivered as opposed to well-child visits, during which immunization screening should be routine. As such, the

possibility of raising immunization rates by converting all missed opportunities here should be taken as an upper figure to what is possible. Institutional, scheduling, reimbursement and parental acceptance are all potential factors on what proportion of encounters without vaccinations could incorporate immunization screening.

Also, encounters were not stratified by type of provider or principal reason for each encounter. While some research suggests that provider type is not a key factor for immunization performance when the volume of well-child visits is taken account [22], no provider information was included in the present study dataset to confirm or identify other relevant provider features. Whether encounters without immunizations are due to children using a spectrum of different provider types, to parental reluctance, or are due to use of settings such as urgent care in place of scheduled well-child visits cannot be determined from the data of this study.

5. Conclusion

The milestone approach to evaluating early childhood immunizations provides a useful perspective for understanding the time-based progression of children through immunization periods. However, the results of this study warrant some caution in the use of immunization record data only in assessing failure to have age-appropriate immunizations because of the chance of misclassification of missed opportunities by providers as missed visits by parents. Nevertheless, for local assessment by public agencies or health plans, and for the design of interventions to improve immunization rates, looking at the patterns by which children fall behind or catch up on immunizations at milestone periods is a valuable next step beyond the count of vaccine doses received by age two.

Acknowledgments

This study was in part funded under CDC Grant no. 280540/09. The authors also wish to thank Diana Bartlett, Laura Pabst, Jim Gaudino, Holly Groom, and Elizabeth Luman for their work in review of this research.

References

[1] L. Rodewald, E. Maes, J. Stevenson, B. Lyons, S. Stokley, and P. Szilagyi, "Immunization performance measurement in a changing immunization environment," *Pediatrics*, vol. 103, no. 4, part 2, pp. 889–897, 1999.

[2] National Committee for Quality Assurance (NCQA), *HEDIS® 2009: Healthcare Effectiveness Data & Information Set. Vol. 1: Narrative*, vol. 1, National Committee for Quality Assurance (NCQA), Washington, DC, USA, 2008.

[3] P. J. Smith, D. C. Hoaglin, M. P. Battaglia, M. Khare, and L. E. Barker, "Statistical methodology of the National Immunization Survey, 1994–2002," *Vital and Health Statistics. Series 2, Data Evaluation and Methods Research*, no. 138, pp. 1–55, 2005.

[4] J. H. Glauber, "The immunization delivery effectiveness assessment score: a better immunization measure?" *Pediatrics*, vol. 112, no. 1, part 1, pp. e39–e45, 2003.

[5] Centers for Disease Control and Prevention, "Recommended immunization schedules for persons 0–18 years—United States, 2007," *Morbidity and Mortality Weekly Report*, vol. 55, no. 51, pp. Q1–Q4, 2007.

[6] A. Shefer, J. Santoli, and J. A. Singleton, "Measuring vaccination coverage—where are we now and where are we going?" *Journal of Public Health Management and Practice*, vol. 13, no. 6, pp. 541–543, 2007.

[7] K. J. Dombkowski, P. M. Lantz, and G. L. Freed, "The need for surveillance of delay in age-appropriate immunization," *American Journal of Preventive Medicine*, vol. 23, no. 1, pp. 36–42, 2002.

[8] K. J. Dombkowski, P. M. Lantz, and G. L. Freed, "Risk factors for delay in age-appropriate vaccination," *Public Health Reports*, vol. 119, no. 2, pp. 144–155, 2004.

[9] G. H. Dayan, K. M. Shaw, A. L. Baughman, et al., "Assessment of delay in age-appropriate vaccination using survival analysis," *American Journal of Epidemiology*, vol. 163, no. 6, pp. 561–570, 2006.

[10] E. T. Luman, L. E. Barker, M. M. McCauley, and C. Drews-Botsch, "Timeliness of childhood immunizations: a state-specific analysis," *American Journal of Public Health*, vol. 95, no. 8, pp. 1367–1374, 2005.

[11] E. T. Luman and S. Y. Chu, "When and why children fall behind with vaccinations. Missed visits and missed opportunities at milestone ages," *American Journal of Preventive Medicine*, vol. 36, no. 2, pp. 105–111, 2009.

[12] L. E. Rodewald, P. G. Szilagyi, S. G. Humiston, R. Barth, R. Kraus, and R. F. Raubertas, "A randomized study of tracking with outreach and provider prompting to improve immunization coverage and primary care," *Pediatrics*, vol. 103, no. 1, pp. 31–38, 1999.

[13] J. M. Santoli, P. G. Szilagyi, and L. E. Rodewald, "Barriers to immunization and missed opportunities," *Pediatric Annals*, vol. 27, no. 6, pp. 366–374, 1998.

[14] C. S. Minkovitz, A. D. Belote, S. M. Higman, J. R. Serwint, and J. P. Weiner, "Effectiveness of a practice-based intervention to increase vaccination rates and reduce missed opportunities," *Archives of Pediatrics and Adolescent Medicine*, vol. 155, no. 3, pp. 382–386, 2001.

[15] P. G. Szilagyi, L. E. Rodewald, S. G. Humiston, et al., "Reducing missed opportunities for immunizations: easier said than done," *Archives of Pediatrics and Adolescent Medicine*, vol. 150, no. 11, pp. 1193–1200, 1996.

[16] S. S. Sabnis, A. J. Pomeranz, P. S. Lye, and M. M. Amateau, "Do missed opportunities stay missed? A 6-month follow-up of missed vaccine opportunities in inner city Milwaukee children," *Pediatrics*, vol. 101, no. 5, p. e5, 1998.

[17] M. F. Daley, B. L. Beaty, J. Barrow, et al., "Missed opportunities for influenza vaccination in children with chronic medical conditions," *Archives of Pediatrics and Adolescent Medicine*, vol. 159, no. 10, pp. 986–991, 2005.

[18] S. L. Udovic, T. A. Lieu, S. B. Black, P. M. Ray, G. T. Ray, and H. R. Shinefield, "Parent reports on willingness to accept childhood immunizations during urgent care visits," *Pediatrics*, vol. 102, no. 4, p. e47, 1998.

[19] T. K. McInerny, W. L. Cull, and B. K. Yudkowsky, "Physician reimbursement levels and adherence to american academy of pediatrics well-visit and immunization recommendations," *Pediatrics*, vol. 115, no. 4, pp. 833–838, 2005.

[20] E. T. Luman, S. Stokley, D. Daniels, and R. M. Klevens, "Vaccination visits in early childhood: just one more visit to be fully vaccinated," *American Journal of Preventive Medicine*, vol. 20, no. 4, supplement 1, pp. 32–40, 2001.

[21] K. J. Dombkowski, M. M. Davis, L. M. Cohn, and S. J. Clark, "Effect of missed opportunities on influenza vaccination rates among children with asthma," *Archives of Pediatrics and Adolescent Medicine*, vol. 160, no. 9, pp. 966–971, 2006.

[22] A. Guttmann, D. Manuel, P. T. Dick, T. To, K. Lam, and T. A. Stukel, "Volume matters: physician practice characteristics and immunization coverage among young children insured through a universal health plan," *Pediatrics*, vol. 117, no. 3, pp. 595–602, 2006.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 426479, 13 pages
doi:10.1155/2010/426479

Methodology Report

Literature-Based Discovery of IFN-γ and Vaccine-Mediated Gene Interaction Networks

Arzucan Özgür,[1] Zuoshuang Xiang,[2,3,4] Dragomir R. Radev,[1,5] and Yongqun He[2,3,4]

[1] *Department of Electrical Engineering and Computer Science, University of Michigan, Ann Arbor, MI 48109, USA*
[2] *Unit for Laboratory Animal Medicine, University of Michigan, Ann Arbor, MI 48109, USA*
[3] *Department of Microbiology and Immunology, University of Michigan, Ann Arbor, MI 48109, USA*
[4] *Center for Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI 48109, USA*
[5] *School of Information, University of Michigan, Ann Arbor, MI 48109, USA*

Correspondence should be addressed to Yongqun He, yongqunh@med.umich.edu

Received 1 November 2009; Accepted 8 March 2010

Academic Editor: Rino Rappuoli

Interferon-gamma (IFN-γ) regulates various immune responses that are often critical for vaccine-induced protection. In order to annotate the IFN-γ-related gene interaction network from a large amount of IFN-γ research reported in the literature, a literature-based discovery approach was applied with a combination of natural language processing (NLP) and network centrality analysis. The interaction network of human IFN-γ (Gene symbol: IFNG) and its vaccine-specific subnetwork were automatically extracted using abstracts from all articles in PubMed. Four network centrality metrics were further calculated to rank the genes in the constructed networks. The resulting generic IFNG network contains 1060 genes and 26313 interactions among these genes. The vaccine-specific subnetwork contains 102 genes and 154 interactions. Fifty six genes such as TNF, NFKB1, IL2, IL6, and MAPK8 were ranked among the top 25 by at least one of the centrality methods in one or both networks. Gene enrichment analysis indicated that these genes were classified in various immune mechanisms such as response to extracellular stimulus, lymphocyte activation, and regulation of apoptosis. Literature evidence was manually curated for the IFN-γ relatedness of 56 genes and vaccine development relatedness for 52 genes. This study also generated many new hypotheses worth further experimental studies.

1. Introduction

In 1965 Wheelock et al. first reported Interferon-gamma-(IFN-γ-) like virus inhibitor induced in supernatant fluid of cultures of fresh human leukocytes following incubation with phytohemagglutinin [1]. In early 1970s, IFN-γ was further studied, and its name was eventually designated. IFN-γ is the only type II IFN family member. It is secreted by activated immune cells—primarily T and NK cells, but also B-cells, NKT cells, and professional antigen presenting cells. IFN-γ has been widely studied and found critical in anti-infectious host defense, inflammatory conditions, cancer, and autoimmune diseases [1, 2]. The most striking phenotype from mice lacking either IFN-γ or its receptor has increased susceptibility to the infections of bacterial and viral pathogens [3]. IFN-γ is also critical for tumor immunosurveillance as assessed using spontaneous, transplantable, and chemical carcinogen-induced experimental tumors. Additionally, IFN-γ is found important in leukocyte homing, cellular adhesion, immunoglobulin class switching, T helper cell polarity, antigen presentation, cell cycle arrest and apoptosis, neutrophil trafficking, and NK cell activation [1, 4, 5].

The induction of IFN-γ response is critical for successful development of vaccines against various viruses and intracellular bacteria, for example, human immunodeficiency virus (HIV) [6, 7], *Mycobacterium tuberculosis* [8–10], *Leishmania* spp. [11, 12], and *Brucella* spp. [13, 14]. The IFN-γ analysis is widely used for the quantification and characterization of the HIV-specific $CD8^+$ T cell responses [6]. It is a marker used as a representative function of cytotoxic T cells to quantify the HIV-specific cellular immune response. IFN-γ is required for protection against mycobacterial infection [15]. *M. tuberculosis*-stimulated whole-blood production of IFN-γ, although imperfect, is the best available correlate

of protective immunity to *M. tuberculosis* in humans [8]. In humans, complete IFN-γR deficiency is associated with frequent infection and ultimately death from the attenuated *M. tuberculosis* BCG vaccine [16]. The inability to secrete IFN-γ or the development of auto-antibodies neutralizing endogenous IFN-γ resulted in the death of a patient by overwhelming mycobacterium infection [17].

Today IFN-γ is ranked as one of the most important endogenous regulators of immune responses. Thousands of relevant papers have been published. However, a comprehensive understanding of how it works and what other factors it interacts with is still largely unclear. Although IFN-γ is essential for protective immunity, animal and human studies have found that IFN-γ alone is not sufficient for the prevention of TB disease [8]. Therefore, it would be very interesting to investigate what other genes or gene interaction networks are needed to stimulate protective immunity. However, due to so-complicated roles of IFN-γ in different conditions, it is challenging to annotate the interaction network of IFN-γ such that it becomes increasingly suitable to interpret its role in various diseases [1].

One of the greatest challenges that the researchers in the biomedical domain face is that most of the knowledge remains hidden in the unstructured text of the published articles. Currently, there are over 19 million publications indexed in PubMed (http://www.ncbi.nlm.nih.gov/pubmed/) and both the total number of publications and the growth rate of the number of publications are increasing exponentially [18, 19]. Given the current amount and the growth rate of the biomedical literature, it is difficult or impossible for biomedical scientists to keep up with the relevant publications. For example, a search in PubMed for "ifn-gamma OR interferon-gamma" returned 75464 articles as of October, 2009. Even if a researcher is only interested in the relatedness of IFN-γ to vaccine development and restricts his search to "vaccine AND (ifn-gamma OR interferon-gamma)", the number of articles retrieved was 7536, which is still too high for reading manually. There are a number of manually curated databases that store protein interactions, such as the Molecular INTeraction database (MINT) [20], the Biomolecular Interaction Network Database (BIND) [21], and the Human Protein Reference Database (HPRD) [22]. Many databases also summarize results from publications about gene-disease relationships, such as the Online Mendelian Inheritance in Man (OMIM) [23], the *Brucella* Bioinformatics Portal (BBP) [24], and the Pathogen-Host Interaction Data Integration and Analysis System (PHIDIAS) [25]. However, it usually takes a lot of time and effort before new discoveries are included in these databases.

To systemically analyze the network of IFN-γ with other genes, an internally developed literature-based discovery approach based on literature mining and network centrality analysis was applied [26]. This literature-based discovery methodology in [26] was shown to be effective in identifying prostate cancer-related genes. To discover genes relevant to IFN-γ and vaccine development, IFN-γ was used as the single seed gene, and all the article abstracts available in PubMed were used as the text knowledge source. The interactions of IFN-γ and its neighbors from abstracts in PubMed were first extracted using a natural language processing (NLP) and machine learning (ML) based method. Two gene interaction networks were eventually built using the automatically extracted interactions. The first network is the generic IFN-γ (IFNG) network, which is the network of interactions of IFN-γ and its neighbors. The second network is the vaccine-specific subgraph of the first network, which is built using only the interactions that are extracted from vaccine relevant sentences. Next, the topologies of the networks were analyzed using the degree, eigenvector, betweenness, and closeness network centrality measures.

To the best of our knowledge, this is the first study that integrates text mining with network analysis in the vaccine informatics domain. The literature-based discovery approach that we have introduced in [26] has been successfully adapted and expanded to discover genes related to IFN-γ and vaccine development. The literature-mined IFN-γ and IFNG-vaccine-mediated networks were systematically analyzed using network centrality metrics. The results support our hypothesis that the central genes in the two IFN-γ networks are related to the functions of IFN-γ and part of the gene list is important for vaccine development. Many predicted genes and gene networks are good candidates for further IFN-γ and vaccine development studies. In this paper, we describe the overall method design and the results.

2. Methods

The high-level system description for predicting IFN-γ and vaccine-associated genes is shown in Figure 1. The approach is described in more detail in the following subsections.

2.1. Literature Corpus. To construct the literature-mined IFN-γ gene interaction network, all article abstracts available in PubMed are used. The sentences of the abstracts were obtained from the BioNLP database in the National Center for Integrative Biomedical Informatics (NCIBI; http://ncibi.org/), which were generated using the MxTerminator [27] sentence boundary detection tool.

2.2. Gene Name Identification and Normalization. Genia Tagger (http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/), whose developers report an F-score performance of 71.37% for biological named entity recognition [28], was used to identify the gene names in the sentences. Consider the example sentence "[IL-2] and [IL-15] induced the production of [IL-17] and [IFN-γ] in a dose dependent manner by PBMCs" taken from the abstract of [29]. The gene names, which were correctly identified by Genia Tagger, are enclosed in square brackets.

One of the greatest challenges in biomedical text processing is that a gene might have several different synonyms. For example, the IFN-γ gene can occur in text as IFN-gamma, IFNG, IFNGamma, interferon-gamma, or interferon gamma. Similarly, the IL2 gene can occur in text as IL2, IL-2, or interleukin 2. If the gene names that correspond to

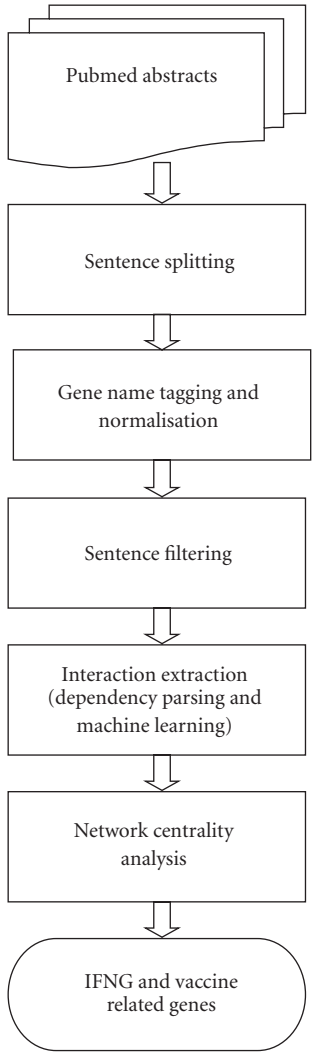


Figure 1: Description of the literature-based discovery system for identifying IFN-γ and vaccine-related genes.

the same gene are not normalized, each different synonym will be represented as a separate node in the gene-interaction network as shown in Figure 2. With five different synonyms for IFN-γ and three different synonyms for IL2, 15 different edges can be obtained although they actually represent the same edge (interaction). Therefore, a dictionary-based approach was used to normalize the gene names tagged by Genia Tagger so that each gene is represented by a single node in the interaction network. HUGO Gene Nomenclature Committee (HGNC) database (http://www.genenames.org/) [30] was used as the dictionary for gene names and their synonyms. As of October, 2009 the database contains 28240 approved gene records. Each tagged gene name was unified with its corresponding approved gene symbol. In the HGNC database, the official gene symbol for the IFN-γ gene is listed as IFNG, and the description is listed as "interferon, gamma". The database does not include any synonyms for the gene. However, IFN-γ is frequently mentioned in text with the names that are shown in Figure 2. Therefore, we included these names to the HGNC dictionary as synonyms for IFN-γ.

2.3. Sentence Filtering. The potential interaction sentences were selected from the abstracts in PubMed that have "human" in the MeSH heading, before applying the text mining method to extract the IFN-γ (IFNG) gene-interaction network from the literature. A list of 826 interaction keywords such as binds, bound, interacts, activates, inhibits, and phosphorylates was compiled from the literature (the list of interaction keywords is available at: http://clair.si.umich.edu/clair/ifngnet/interaction_keywords.txt). Our assumption is that a sentence that describes an interaction between a pair of genes should contain an interaction keyword and at least two distinct normalized gene names. The sentences that do not meet this requirement were filtered out.

The IFNG gene-interaction network was built in two steps. In the first step, the genes that interact with IFNG (or called the neighbors of IFNG) were extracted. The number of sentences that contain IFNG or one of its synonyms (case-insensitive match) and are from abstracts that have "human" in the MeSH headings is 73024. A filter program was further performed to filter out those sentences that do not have at least one interaction keyword and at least two distinct normalized gene names, one of which is IFNG. As a result, 26876 sentences were obtained with our interaction extraction module for identification of the genes that interact with IFNG. The interaction extraction module extracted 1059 neighbors of IFNG.

In the second step, the interactions among the neighbors of IFNG were extracted. There are over 9 million sentences that are from abstracts which have "human" in the MeSH headings and contain at least one of the IFNG neighbors or their synonyms. Out of these, the sentences for further processing by the interaction extraction module are those that have at least one interaction keyword, and at least two distinct normalized gene names, which were identified as neighbors of IFNG in the first step. In total, 422566 sentences met these criteria and were further processed by the interaction extraction module, which is described in the next subsection.

2.4. Gene Interaction Extraction from the Literature. The interaction extraction task was formulated as a classification task, where each sentence is classified as a possible interaction between a given gene pair. The support vector machines (SVMs) [31] were used as our classification algorithm with features extracted from the dependency parse trees of the sentences, which capture the semantic predicate-argument dependencies among the words. The Stanford Parser (http://nlp.stanford.edu/software/lex-parser.shtml) was used to obtain the dependency parse trees of the sentences [32].

Figure 3 shows the dependency parse tree for the example sentence "IL-2 and IL-15 induced the production of IL-17 and IFN-γ in a dose dependent manner by PBMCs". The nodes of the tree represent the words of the sentence and the edges represent the types of the dependencies among the

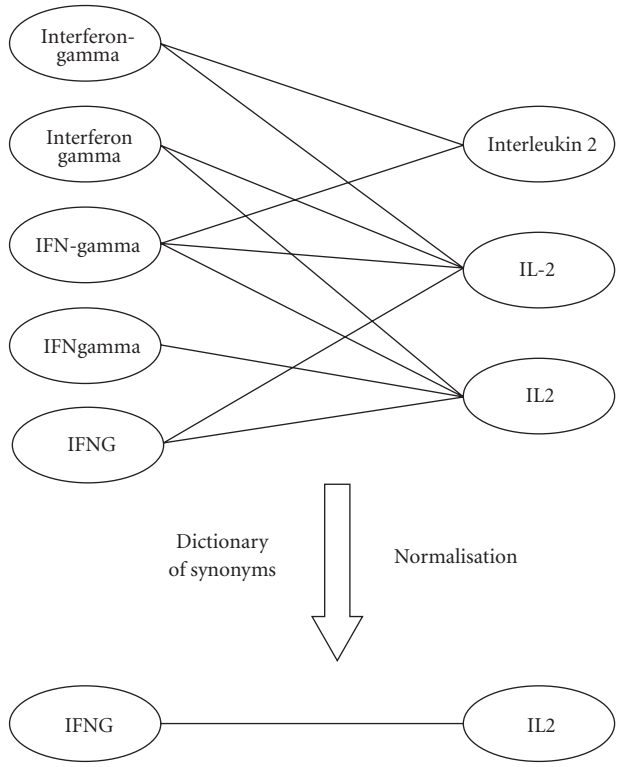


Figure 2: Gene name normalization example.

words. For example, "IL-2" is the noun subject "nsubj" of "induced". There are four gene names in the sentence. The sentence describes an interaction between the gene pairs "IL-2 and IL-17", "IL-2 and IFN-γ", "IL-15 and IL-17", and "IL-15 and IFN-γ". It does not describe an interaction between the gene pairs "IL-2 and IL-15" and "IL-17 and IFN-γ".

The shortest path between each gene pair from the dependency tree of the sentence was used with SVM. The motivating assumption is that the path between two gene names in a dependency tree is a good description of the semantic relation between them in the corresponding sentence. For example, the path between the interacting gene pair "IL-2 and IL-17" is "nsubj induced dobj production prep_of" and the path between the noninteracting pair "IL-2 and IL-15" is "conj_and".

The similarity between two dependency paths was indicated based on the word-based edit distance, which is defined as the minimum number of word insertion, deletion, or substitution operations needed to transform the first path to the second. For example, the edit distance between the paths "nsubj induced dobj production prep_of" and "conj_and" is five, since the first path can be transformed to the second one by deleting four words (nsubj, induced, dobj, and production) and substituting one word, that is, substituting prep_of with conj_and. The more similar two paths are (smaller edit distance), the more likely they belong to the same class; that is, either both describe or both do not describe an interaction for the corresponding gene pairs. The path edit distance measure between two paths p_i and p_j was converted into a path similarity function as follows:

$$\text{edit\_sim}\left(p_i, p_j\right) = e^{-\gamma(\text{edit\_distance}\ (p_i, p_j))}. \quad (1)$$

This path similarity measure was integrated as a kernel function to SVM by plugging it in the SVM$^{\text{light}}$ package (http://www.svmlight.joachims.org/) [31].

This interaction extraction approach was introduced in [33] and was shown that it achieves the state-of-the-art results (55.61% F-score performance for the AIMED data set (ftp://ftp.cs.utexas.edu/pub/mooney/bio-data/) and 84.96% F-score performance for the CB data set). We have successfully applied this approach to extract the interactions of the prostate cancer relevant genes in [26] and to provide annotations for the BioCreative Meta-Server by classifying abstracts as describing a protein interaction or not in [34]. To extract the interactions of IFNG and its neighbors, the system was trained by combining the AIMED and the CB data sets. The preprocessed data sets are available at http://clair.si.umich.edu/clair/biocreative/datasets/.

2.5. Network Centrality Analysis. Gene interactions can be represented as a network, where the genes are represented as nodes, and an interaction between a pair of genes is represented with an edge connecting the corresponding nodes. This representation allows the analysis of interactions from a graph theory and complex networks perspective,

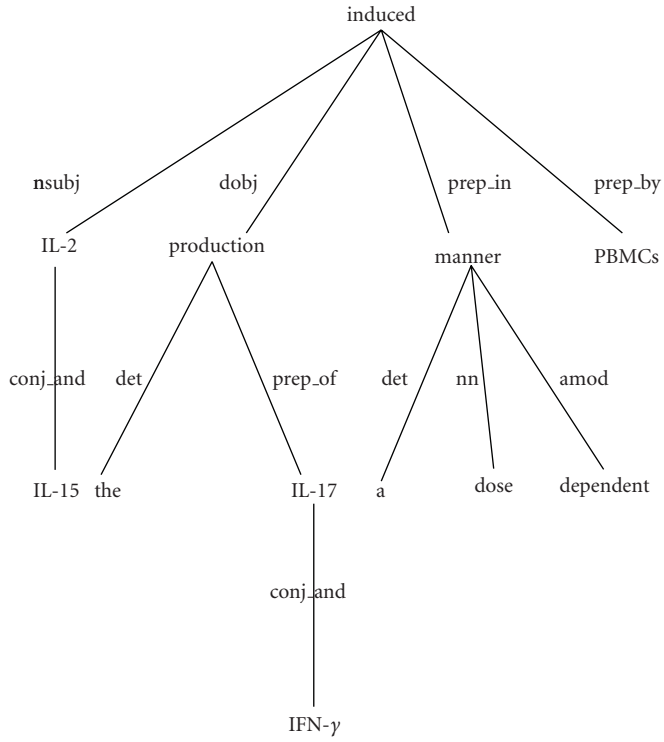


Figure 3: The dependency parse tree of the sentence "IL-2 and IL-15 induced the production of IL-17 and IFN-γ in a dose dependent manner by PBMCs."

which can give biologists a variety of new insights. For example, Schwikowski et al. used a majority-rule method that assigns to a protein the function that occurs most commonly among its neighbors and reported an accuracy of 70% for the yeast protein interaction network [35]. Similarly, Spirin and Mirny used the protein interaction networks to discover molecular modules that function as a unit in certain biological processes by identifying subgraphs that are densely connected within themselves but sparsely connected with the rest of the network [36].

Another network feature that can reveal important principles underlying the biological systems is the centrality of a node, which defines the relative importance of the node in the graph. The importance of a node can be defined in different ways. Degree centrality is defined as the number of edges incident to the node (i.e., the number of neighbors that a node has) [37]. It measures the extent of influence that a node has on the network. The more neighbors a node has, the more important it is.

In degree centrality each neighbor contributes equally to the centrality of a node. However, all the connections of a node are not always equally important. This notion is defined as "prestige" in social networks. The prestige of a person does not only depend on the number of acquaintances he has but also on who his acquaintances are (i.e., how prestigious they are). Eigenvector centrality assigns each node a centrality that not only depends on the quantity of the connections but also on their importance. The eigenvector centrality of a node is proportional to the sum of the centralities of its neighbors [38].

Closeness centrality of a node is defined as the inverse sum of the distances from the node to the other nodes in the network [37]. The closer a node to the other nodes in the network, the more important it is.

Betweenness centrality of a node is defined as the proportion of the shortest paths between all the pairs of nodes in the network that pass through the node in interest [37]. A node is considered important if it occurs on many shortest paths between other nodes. This characterizes the control of a node over the information flow of the network.

Centrality measures have originally been developed and used in nonbiological domains. For example, the web pages in the popular search engine Google are ranked by using the Pagerank algorithm, which is based on eigenvector centrality [39]. A number of recent studies have successfully applied centrality measures in biological domains. For example, Jeong et al. used degree centrality to predict lethal mutations in the yeast protein interaction network [40]. They showed that the network is tolerant to random errors, whereas errors related to the most central proteins cause lethality. Similarly, Joy et al. [41] and Hahn and Kern [42] have found that there is an association between the betweenness centrality and the essentiality of a gene, where an essential gene is a gene that causes the organism to die when it malfunctions. Recently, we have applied centrality measures to predict genes relevant to prostate cancer [26]. We were able to identify genes, which

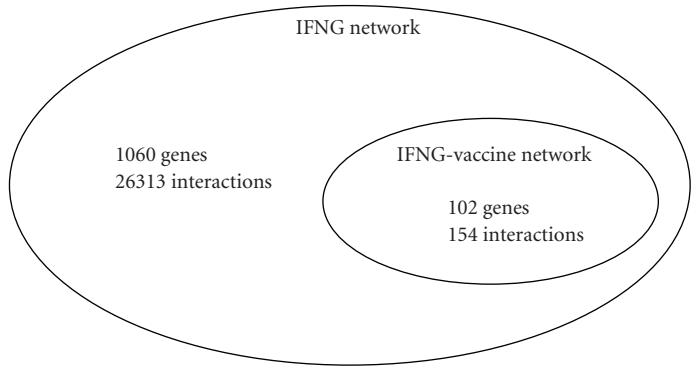


Figure 4: Summary of the IFNG network and its vaccine-specific subnetwork.

are not marked as being related to prostate cancer by the curated databases such as the Online Mendelian Inheritance in Man (OMIM) and the Human Prostate Gene Database (PGDB) [43] even though there are recent articles that confirm the association of these genes with the disease.

In this study the IFNG interaction network was analyzed from graph centrality perspective. IFNG and its neighbors are represented as nodes and there is an edge between two genes if an interaction between them from the literature has been extracted. The gene names in the network are normalized and represented with their official HGNC symbols. The vaccine-specific subgraph of this network contains only the interactions that have been extracted from sentences that contain the term "vaccin", which is the root form of the vaccine-related terms such as vaccine, vaccines, vaccination, and vaccinated. Therefore, the edges in this subgraph are all vaccine specific. Analysis of this IFNG-vaccine network helps us to understand the genes and interactions that play important roles in both the vaccine and IFNG network. Since IFNG is one of the most important immune factors and critical for vaccine development, we hypothesized that genes central in the generic IFNG and IFNG-vaccine networks might be important for vaccine development. The results presented in the next section support the hypothesis.

2.6. Gene Annotation Enrichment Analysis. The web-based DAVID bioinformatics program was used to perform the gene annotation enrichment analysis [44].

3. Results

3.1. Topological Properties of the Networks. Our program detected 1060 nodes (genes including IFNG and its neighbors) linked by 26313 edges (interactions) (Figure 4). Since all the genes in the IFNG network are connected to IFNG, the diameter of the network (the longest of the shortest paths between the pairs of genes in the interaction network) is 2 and the average shortest path length (the average of the shortest paths between all genes in the network) is 1.95. The clustering coefficient of the network is 0.4933, which is an order of magnitude higher than the clustering coefficient of a random network with the same number of nodes (0.0473). The clustering coefficient [45] of a node describes how well connected a node's neighbors are and is defined as the number of connections between this node's neighbors divided by the number of possible connections between them. The clustering coefficient of a network is the average of the clustering coefficients of the nodes in the network. The IFNG network is a small-world network [45], characterized by having a small average shortest path length and a clustering coefficient that is significantly higher than that of a random network with the same number of nodes. The IFNG network is a scale-free network, which is characterized by having a power-law degree distribution, $P(k) \sim k^{-\gamma}$, where $P(k)$ is the probability that a randomly selected node will have a degree (i.e., number of connections) of k [46].

In scale-free networks most nodes make only a few connections, while a small set of nodes (known as hubs) have very large number of links. This is different from random networks, which follow Poisson distribution, where majority of the nodes have degrees close to the average degree of the network. The exponent (γ) of the power-law degree distribution of the IFNG network is 2.15. The graph of the IFNG network is shown in Figure 1 of the supplementary material available online at doi:10.1155/2010/426479

The IFNG and vaccine-associated network (IFNG-vaccine network) is a much smaller subset of the generic IFNG network. This small subnetwork contains 102 genes and 154 interactions (Figure 4). Since the IFNG-vaccine network is built by removing the edges that are not associated with "vaccine" from the IFNG network, some of the genes that were connected in the IFNG network are not connected in the IFNG-vaccine network. In total, the IFNG-vaccine network contains 84 genes that are interconnected and 18 genes that are separated from this largest connected component of 84 genes (Figure 5). Also, the diameter of the IFNG-vaccine network and the average shortest path length are larger than those of the IFNG network. The diameter of the IFNG-vaccine network is 9 and the average shortest path length is 3.55. The IFNG-vaccine network still possesses the small-world property with a relatively small average shortest path length and a clustering coefficient (0.2218) that is significantly higher than the clustering coefficient of a random network with the same number of nodes (0.0388). The network is scale-free with a power-law

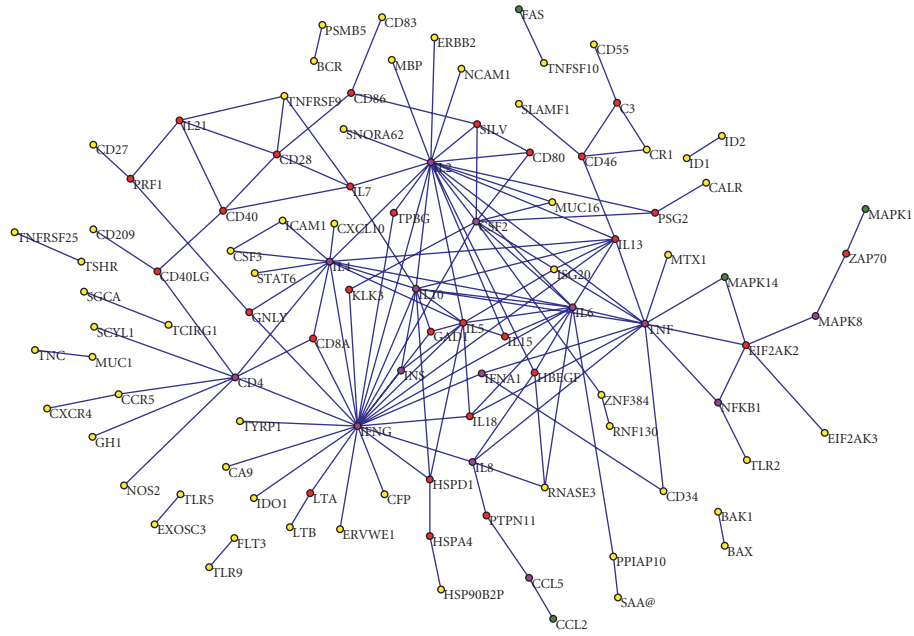


FIGURE 5: The graph of the IFNG-vaccine network extracted from the literature. The network consists of 102 nodes (genes) and 154 edges (interactions). All the edges in the network are associated with the term "vaccine" and its variants. The purple nodes are the genes that are central in both the generic and the IFNG-vaccine networks. The red nodes are the genes that are central only in the IFNG-vaccine network. The green nodes are the genes that are central only in the generic IFNG network. The rest of the nodes are shown in yellow.

degree distribution with exponent 2.37. The small-world and scale-free characteristics of the generic IFNG and the IFNG-vaccine networks are consistent with the topological properties of previously studied biological networks [26, 40, 47, 48] and nonbiological networks such as the Internet [49] and social networks [45].

3.2. Lists of Genes Are Predicted and Sorted by Centrality Analyses. All the genes in the two networks (generic IFNG network and IFNG-vaccine network) are sorted based on centrality analyses. Supplementary File 1 lists the rankings of all the genes in the generic IFNG network and Supplementary File 2 lists the rankings of all the genes in the IFNG-vaccine network. IFNG is not included in these rankings, since it is trivially ranked highest by all the centrality measures in both networks due to the fact that the networks are specific to IFNG. The most central genes (the genes ranked among the top 25 by at least one of the centrality measures) are analyzed in more detail in Table 1. These genes (a total of 56 genes) are predicted to be associated with IFNG and relevant for vaccine development. Literature evidence was manually curated for the IFNG association (IFNG-Ref column in Table 1) and the vaccine development relatedness (Vaccine-Ref column in Table 1) of these genes.

It is interesting that in the generic IFNG network, all centrality measures find the same 23 genes among the top 25, although the ranking might change slightly (Table 1). For example, IL10 is ranked 5th by degree and closeness centralities, but 4th by eigenvector and betweenness centralities. Since all the genes in the generic IFNG network are connected to IFNG, the distance (shortest path length) between a pair of genes is at most two. In other words, the distance between a pair of genes is one if they are directly connected to each other and it is two if they are not directly connected to each other (i.e., they are connected through IFNG). Therefore, in this network, the more genes a gene is connected to (higher degree centrality), the less distant it is to the other genes (higher closeness centrality). So, the degree and closeness centralities produce the same rankings for the generic IFNG network. For the IFNG-vaccine network, the top 25 genes sorted based on centrality analyses overlapped with the sorted results from the generic IFNG network.

Three different levels of prediction are available based on the comparison between the generic IFNG network and the more specific IFNG-vaccine network.

(1) Genes Ranked High in Both Networks. Thirteen genes were ranked among the top 25 in both networks by at least one of the centrality measures. Among these 13 genes, 8 genes are central by all centrality measures in both networks: TNF, IL6, IL8, IL10, IL4, IL2, CSF2, and CD4. These genes are well studied in both generic IFNG research and vaccine specific research. The ranking may change in both networks. For example, IL2 was ranked top 1 in the IFNG-vaccine network, while it was ranked top 7-8 in the generic IFNG network based on different centrality scores. This is probably due to the fact that the role of IL2 in vaccine research has widely been recognized and studied in more depth in the vaccine context.

Among the 13 genes in this group, five genes (NFKB1, MAPK8, INS, IFNA1, and CCL5) were ranked high in the

Table 1: Predicted 56 genes related to IFN-γ and vaccine networks.

Gene	Generic IFNG Network					IFNG-vaccine Network				
	D	E	B	C	IFNG-Ref	D	E	B	C	Vaccine-Ref
TNF	1	1	1	1	3132506	2	3	2	7	16446013
NFKB1	2	2	2	2	9888423	—	23	—	—	16971487
IL6	3	3	3	3	1719090	3	4	7	3	10225849
IL8	4	5	6	4	8473010	10	13	10	9	11378044
IL10	5	4	4	5	8102388	6	8	11	2	10930151
IL4	6	6	5	6	2136895	4	2	4	4	8519092
MAPK1	7	9	9	7	15307176	—	—	—	—	19428911
IL2	8	7	8	8	6429853	1	1	1	1	8459207
VEGFA*	9	10	10	9	12816689	—	—	—	—	17502972
TP53*	10	8	7	10	16391798	—	—	—	—	18846387
BCL2*	11	13	13	11	11064392	—	—	—	—	19389797
AKT1*	12	11	12	12	11135576	—	—	—	—	19107122
MAPK8	13	14	14	13	18950753	—	—	15	—	19428911
INS	14	12	11	14	8383325	—	—	—	16	19203100
MAPK14	15	15	18	15	10700460	—	—	—	—	19428911
CSF2	16	18	17	16	11665752	7	6	6	6	19459853
FAS	17	17	16	17	10895367	—	—	—	—	15979942
CCL2	18	19	19	18	9407497	—	—	—	—	19833737
IFNA1	19	16	15	19	11449378	—	—	—	13	19667099
EGFR*	20	20	23	20	17362940	—	—	—	—	19178753
JUND*	21	21	22	21	10070035	—	—	—	—	19124729
KITLG*	22	24	—	22	7540064	—	—	—	—	-
CCL5	23	23	21	23	8921438	—	24	—	—	15827150
CD4	24	22	20	24	15173593	9	5	3	12	17298856
EGF*	25	25	—	25	18160214	—	—	—	—	16357522
CRP	—	—	24	—	10675363	—	—	—	—	16395099
STAT3*	—	—	25	—	7488223	—	—	—	—	-
IL5	—	—	—	—	9432015	5	7	20	8	11138639
IL13	—	—	—	—	12670721	8	9	5	5	12232042
IL7	—	—	—	—	7594482	11	14	12	17	17496983
EIF2AK2	—	—	—	—	11342638	12	10	8	—	19596385
CD28	—	—	—	—	7634349	13	12	—	—	12594842
HSPD1	—	—	—	—	12407015	14	19	16	14	12218165
SILV	—	—	—	—	11839572	15	20	17	23	11459172
IL21	—	—	—	—	14657853	16	17	—	—	16785513
IL18	—	—	—	—	8666798	17	—	—	10	19467215
HBEGF	—	—	—	—	9062364	18	25	—	21	10729731
CD46	—	—	—	—	15307176	19	11	9	—	11757799
CD40	—	—	—	—	7554483	20	16	—	—	11403919
PSG2	—	—	—	—	2516715	21	—	22	—	11155821
GAD1	—	—	—	—	9703171	22	—	—	18	12421990
IL15	—	—	—	—	9834271	23	—	—	22	16785513
C3	—	—	—	—	1337336	24	15	—	—	19477524
PRF1	—	—	—	—	19651871	25	22	19	—	15214037
ZAP70	—	—	—	—	11034358	—	18	23	—	—
CD40LG	—	—	—	—	10769003	—	21	18	—	11403919
GNLY	—	—	—	—	17382591	—	—	13	19	10644038
PTPN11	—	—	—	—	12270932	—	—	14	—	—
CD86	—	—	—	—	9836505	—	—	21	—	12594842

Table 1: Continued.

Gene	Generic IFNG Network					IFNG-vaccine Network				
	D	E	B	C	IFNG-Ref	D	E	B	C	Vaccine-Ref
CCR5	—	—	—	—	9616137	—	—	24	—	16672545
HSPA4	—	—	—	—	18442794	—	—	25	—	11779704
TPBG	—	—	—	—	16630022	—	—	—	11	16630022
KLK3	—	—	—	—	16000955	—	—	—	15	19171173
CD8A	—	—	—	—	1904117	—	—	—	20	18425263
CD80	—	—	—	—	7537534	—	—	—	24	10498243
LTA	—	—	—	—	3102976	—	—	—	25	15908422

Note: The genes ranked among the top 25 by the centrality measures (D: Degree; E: Eigenvector; B: Betweenness; C: Closeness) in the generic IFNG network or the IFNG-vaccine network. The genes are represented with their official HGNC symbols. Literature evidences for the relatedness of the genes to IFNG (IFNG-Ref) and to vaccine development (Vaccine-Ref) are manually curated. "—" indicates that the gene is not ranked among the top 25 by the corresponding centrality measure in the corresponding network or no literature evidence was found.

IFNG network by all measures but only high in the IFNG-vaccine network by certain centrality measures. For example, MAPK8 (mitogen-activated protein kinase 8; Aliases: JNK, JNK1, SAPK1) was ranked high by all centrality metrics in the IFNG network, whereas it was ranked high by only the betweenness centrality metric in the IFNG-vaccine network (Table 1). The high betweenness score was reflected by the fact that MAPK8 connects the two genes (ZAP70 and MAPK1) to the rest of the network (Figure 5). In the generic IFNG network, 322 other genes are directly connected to MAPK8 (Figure 6). Many of these genes (e.g., NFKB1, IL4, and CD40) also exist in the IFNG-vaccine network (Figure 5) although they do not directly interact with MAPK8. However, the majority of these 322 genes (e.g., TLR4 and IL1B) are not in the IFNG-vaccine network. It is reasonable to suggest that many of these genes that were found in the IFNG-MAPK8 network (Figure 6) but not in the IFNG-vaccine network (Figure 5) may also be important for vaccine specific network through an interaction with MARK8. Therefore, the comparison between these two networks may lead to hypothesis of new genes involved in vaccine specific immune network, some of which deserve further experimental verifications.

(2) Genes Ranked High in the Generic IFNG Network but Not in the IFNG-Vaccine Network. In total 14 genes are included in this group. Nine out of these 14 genes were not found in the IFNG-vaccine network (Supplementary File 2). These genes are labeled with "$*$" in Table 1. These genes have not been well studied in the vaccine context. However, since these genes are strongly associated with IFNG, it is likely that each of these genes may also play an important role in vaccine-induced protective immune network. For example, as one of the 14 genes, the serine/threonine kinase AKT1 is a key regulator of cell proliferation and death. AKT1 regulates lymphocyte apoptosis and Th1 cytokine propensity [50]. IFNG is a representative cytokine in Th1 response that is crucial for induction of vaccine-induced protection. Therefore, it is reasonable to hypothesize that AKT1 plays an important role in regulated vaccine-induced protective immune responses.

Among the 14 genes in this group, five genes (MAPK1, MAPK14, FAS, CCL2, and CRP) were found in the IFNG-vaccine network but not ranked high based on any centrality analysis. For example, FAS is a critical gene in regulation of programmed cell death through the FAS pathway. FAS (TNF receptor superfamily, member 6; Aliases: CD95, APO-1) has been found to play an important role in promoting an appropriate effector response following vaccinations against *Helicobacter pylori* [51], hepatitis C virus [52], and cancer [53]. Since FAS is well studied and ranked top in the generic IFNG network, more knowledge about its interactions with other genes shown from the generic IFNG network provides valuable basis for further analysis of FAS-related, vaccine-specific interaction network.

(3) Genes Ranked High in the IFNG-Vaccine Network but Not in the Generic IFNG Network. In total, 29 genes that were ranked among the top 25 in the IFNG-vaccine network based on at least one of the centrality scores are not ranked among the top 25 in the generic IFNG network (Table 1). These genes may be more vaccine-specific and play relatively less important roles in many other IFNG-regulated immune systems (e.g., cell cycle). It is also possible that some of these genes are very important for other IFNG-related immune functions. In that case, the data for these genes obtained from vaccine research may provide supportive results for expanded studies. One important set of these 29 genes cover many interleukins including IL5, IL7, IL13, IL15, IL18, and IL21. For example, interleukin-18 (IL18) is a newly discovered cytokine with profound effects on T-cell activation. IL18 can possibly be used as a strong vaccine adjuvant [54]. The new knowledge obtained from IL18 in vaccine research may be applied to other IFNG-related immune systems.

3.3. Gene Annotation Enrichment Shows Various Immune Responses Regulated by IFN-γ. The 56 genes ranked among the top 25 by at least one of the centrality methods in one or both networks were used for gene enrichment analysis using DAVID [44]. These genes were classified in various immune mechanisms such as response to extracellular stimulus,

Table 2: Gene annotation enrichment among top predicted genes in the generic IFNG and the IFNG-vaccine networks.

Category	Term	Count	*P*-Value	FDR
GOTERM_BP_ALL	GO:0050896~response to stimulus	43	$2.99E-22$	$5.71E-19$
GOTERM_BP_ALL	GO:0007154~cell communication	39	$5.74E-13$	$1.10E-09$
GOTERM_BP_ALL	GO:0007165~signal transduction	35	$9.70E-11$	$1.86E-07$
GOTERM_BP_ALL	GO:0006950~response to stress	29	$7.14E-20$	$1.37E-16$
GOTERM_BP_ALL	GO:0030154~cell differentiation	28	$6.94E-13$	$1.33E-09$
GOTERM_BP_ALL	GO:0006952~defense response	26	$7.12E-23$	$1.36E-19$
GOTERM_BP_ALL	GO:0006955~immune response	26	$8.88E-18$	$1.70E-14$
GOTERM_BP_ALL	GO:0008283~cell proliferation	23	$9.37E-16$	$1.70E-12$
GOTERM_BP_ALL	GO:0008219~cell death	23	$2.28E-15$	$4.46E-12$
GOTERM_BP_ALL	GO:0006915~apoptosis	22	$9.38E-15$	$1.78E-11$
GOTERM_BP_ALL	GO:0007242~intracellular signaling cascade	19	$4.85E-07$	$9.27E-04$
GOTERM_BP_ALL	GO:0001775~cell activation	18	$5.86E-19$	$1.12E-15$
GOTERM_BP_ALL	GO:0006954~inflammatory response	17	$8.26E-16$	$1.49E-12$
GOTERM_BP_ALL	GO:0046649~lymphocyte activation	14	$1.77E-14$	$3.38E-11$
GOTERM_BP_ALL	GO:0006468~protein amino acid phosphorylation	14	$2.60E-07$	$4.98E-04$
GOTERM_BP_ALL	GO:0006807~nitrogen compound metabolic process	13	$4.47E-08$	$8.56E-05$
GOTERM_BP_ALL	GO:0042110~T cell activation	12	$9.02E-14$	$1.73E-10$
GOTERM_BP_ALL	GO:0048534~hemopoietic or lymphoid organ development	12	$4.57E-11$	$8.74E-08$
GOTERM_CC_ALL	GO:0005576~extracellular region	29	$5.33E-18$	$8.27E-15$
GOTERM_MF_ALL	GO:0005125~cytokine activity	19	$5.30E-21$	$9.51E-18$
GOTERM_MF_ALL	GO:0008083~growth factor activity	12	$6.77E-12$	$1.21E-08$
KEGG_PATHWAY	hsa04060:Cytokine-cytokine receptor interaction	23	$7.90E-16$	$9.77E-13$
KEGG_PATHWAY	hsa04620:Toll-like receptor signaling pathway	13	$3.12E-10$	$3.91E-07$
KEGG_PATHWAY	hsa04660:T cell receptor signaling pathway	12	$2.04E-09$	$2.57E-06$
KEGG_PATHWAY	hsa04630:Jak-STAT signaling pathway	11	$2.99E-06$	0.003745

lymphocyte activation, and regulation of apoptosis (Table 2). These gene annotation enrichment results are correlated with current knowledge about IFN-γ [1, 4, 5]. It further demonstrates the capability of our literature-based discovery approach in correctly extracting genes related to IFN-γ.

4. Discussion

Our method is different from many other literature mining approaches. To extract the gene interactions from the text, an SVM classifier was used in our approach with features extracted from the dependency parse trees of the sentences [33]. A dependency parse tree captures the semantic predicate-argument relationships among the words of a sentence. Compared to the traditional cooccurrence and pattern-matching-based information extraction methods, our method allows us to make more syntax-aware inferences about the roles of the genes in a sentence. Our method of integrating literature mining with network analysis was first introduced in 2008 to study prostate cancer [26]. In that study, 15 genes related to prostate cancer were chosen as seed genes, and 48245 articles from PubMed Central (PMC) Open Access (http://ncbi.nlm.nih.gov/pmc/about/openftlist.html) were processed to build the network of their interactions. Genes that are not marked as being related to prostate cancer by the curated OMIM or PGDB [43] databases were identified even though there are recent articles that confirm their association to the disease. In this current study, only one gene (IFNG) was used as the seed gene, and 19 million papers in PubMed were analyzed. Therefore, our method of literature-based discovery can be generalized and used in different applications. Since the vaccine is emphasized, the vaccine term and its variation terms in our NLP analysis were used (act like gene in our approach). This approach is new in this type of analysis.

Our analysis discovered a large number of genes that interact with IFNG and genes important for both IFNG and vaccine. Many of these genes have been studied but never been collected for systematic network analysis. Current databases contain limited information about IFNG gene interaction network. The Michigan Molecular Interactions (MiMI) is a repository that includes interaction data from over 10 databases such as the Database of Interacting Proteins (DIP), the Human Protein Reference Database (HPRD), and the Biomolecular Interaction Network Database (BIND) [55]. As of October 2009, MiMI contains only 12 genes that interact with IFNG and 27 interactions among these genes. Our IFNG gene interaction network contains more than 80-fold of genes that interact with IFNG. While the correctness of all these interactions require further confirmation,

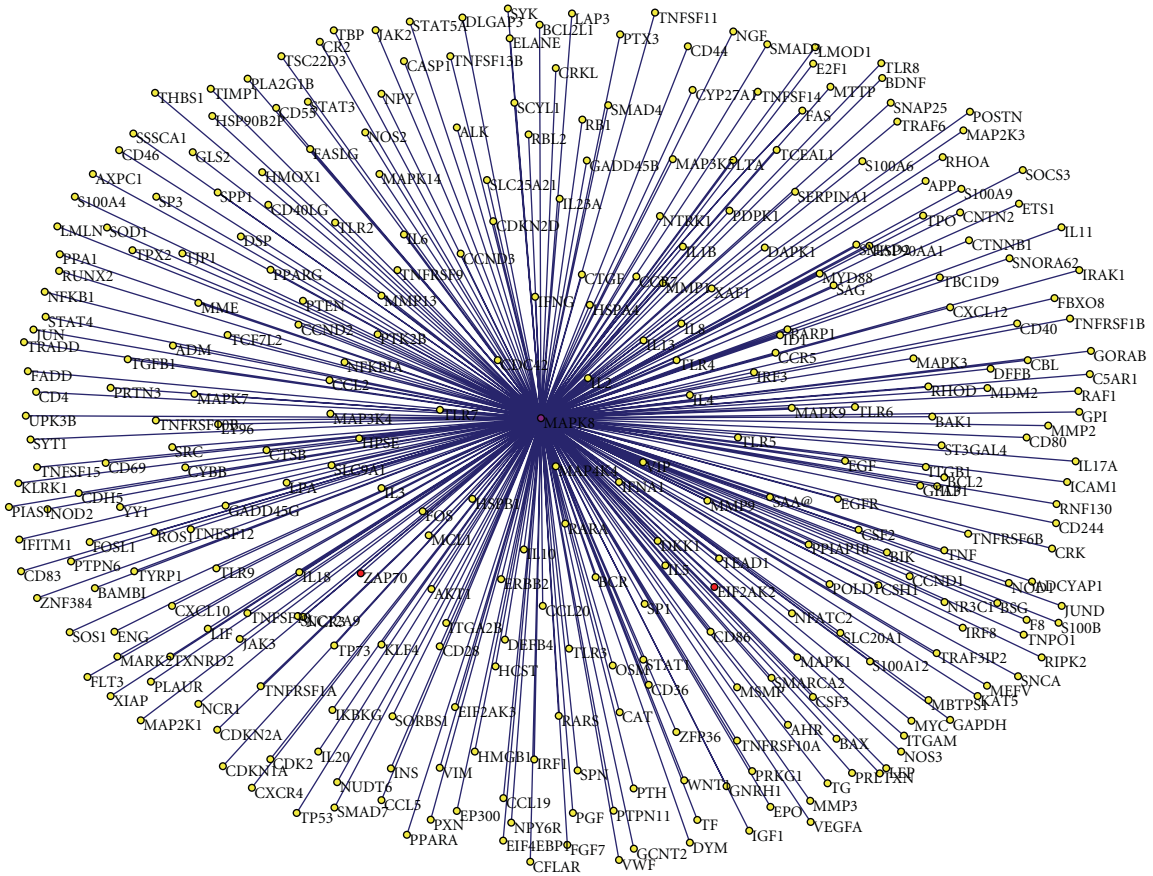


Figure 6: Interactions of MAPK8 with other genes in the generic IFNG network. MAPK8 is shown in purple. The two genes (ZAP70 and EIF2AK2) that MAPK8 also interacts in the IFNG-vaccine network are shown in red.

our manual confirmation of selected 56 interactions (Table 1) has already demonstrated the power of our literature-based discovery method. Since IFNG is an important immune regulator for vaccine-induced protective immunity, the systematical analysis of vaccine-induced IFNG-regulated gene network is critical to understand vaccine-induced immune mechanism and support rational vaccine design. Our selective analyses of the IFNG-vaccine subnetwork showed that genes potentially important for vaccine research can be predicted. Many predicted genes and gene networks deserve further experimental verifications.

Our study demonstrated that MAPK8 is an important component of the generic IFNG network (Table 1, Figures 5 and 6). MAPK8 is a member of the mitogen-activated protein (MAP) kinase family. MAPK8 is important for many cellular processes such as cell proliferation, apoptosis, and differentiation. IFNG and MAPK8 regulate each other depending on different experimental conditions [56–60]. For example, the IFNG inhibits the activation of MAPK8 in macrophages and many other cells through the production of nitric oxide [56]. However, IFNG activates JNK activation and both contribute to apoptosis in lymphocyte cells through the regulation of the reactive oxygen species (ROS) production [57]. Meanwhile, the JNK stress-activated MAPK signal transduction pathway is required for IFNG production for T helper 1 (Th1) effector cells [58]. The inhibition of MAPK8 results in marked reduction of IFNG transcription in activated Jurkat T cells [59]. The activation of JNK pathway also mediates the production of IFNG in human breast tumor cells [60]. Our study shows that MAPK8 interacts with 322 genes which also interact individually with IFNG in the generic IFNG network (Figure 6). The finding of such a large number of interactive genes suggests that the interactions among MAPK8, IFNG, and the other genes may regulate many different biological processes. Based on our GO enrichment analysis (data not shown), the 322 genes that interact with both MAPK8 and IFNG (Figure 6) cover a variety of different biological processes, such as response to external stimulus, inflammatory response, cell proliferation, programmed cell death, and cytokine activity. It is interesting that only two genes (ZAP70 and EIF2AK2) among the 323 genes in the IFNG-MAPK8 network were found to connect to MAPK8 in the IFNG-vaccine network (Figure 5). Since many other genes (e.g., NFKB1, IL4, and CD40) in the IFNG-MAPK8 network also exist in the IFNG-vaccine network (Figure 5), it is possible that more genes act in the vaccine-specific gene network through their interactions with both MAPK8 and IFNG. It is also likely that many genes shown in the IFNG-MAPK8 network but not in the IFNG-vaccine network may contribute to vaccine-induced protective immunity.

Future work includes development of a web server to store the analyzed data and provide a user-friendly web interface to query and visualize the analyzed data. We expect to provide such a user-friendly web interface for the analyses of IFNG and IFNG-vaccine gene networks in 2010. It is noted that the interactions shown in our networks may be specific for certain conditions. The interactions may not be true when experimental conditions change. One future research is to link individual interactions to specific conditions. It will provide us a more comprehensive view of the IFNG and vaccine networks. Our literature mining method will also be applied to analyze other IFNG and vaccine-related interaction networks in other animal species (e.g., mouse, rat, and cattle).

Acknowledgments

This research is supported by NIH Grants R01AI081062 and U54-DA-021519. The authors appreciate Alex Ade's support for their access to the BioNLP database in the National Center for Integrative Biomedical Informatics (NCIBI).

References

[1] A. Billiau and P. Matthys, "Interferon-γ: a historical perspective," *Cytokine and Growth Factor Reviews*, vol. 20, no. 2, pp. 97–113, 2009.

[2] T. Wieder, H. Braumuller, M. Kneilling, B. Pichler, and M. Rocken, "T cell-mediated help against tumors," *Cell Cycle*, vol. 7, no. 19, pp. 2974–2977, 2008.

[3] K. Schroder, P. J. Hertzog, T. Ravasi, and D. A. Hume, "Interferon-γ: an overview of signals, mechanisms and functions," *Journal of Leukocyte Biology*, vol. 75, no. 2, pp. 163–189, 2004.

[4] D. J. Gough, D. E. Levy, R. W. Johnstone, and C. J. Clarke, "IFNγ signaling—does it mean JAK-STAT?" *Cytokine and Growth Factor Reviews*, vol. 19, no. 5-6, pp. 383–394, 2008.

[5] H. Takayanagi, K. Sato, A. Takaoka, and T. Taniguchi, "Interplay between interferon and other cytokine systems in bone metabolism," *Immunological Reviews*, vol. 208, pp. 181–193, 2005.

[6] H. Streeck, N. Frahm, and B. D. Walker, "The role of IFN-γ Elispot assay in HIV vaccine research," *Nature Protocols*, vol. 4, no. 4, pp. 461–469, 2009.

[7] K. Kedzierska and S. M. Crowe, "Cytokines and HIV-1: interactions and clinical implications," *Antiviral Chemistry and Chemotherapy*, vol. 12, no. 3, pp. 133–150, 2001.

[8] H. A. Fletcher, "Correlates of immune protection from tuberculosis," *Current Molecular Medicine*, vol. 7, no. 3, pp. 319–325, 2007.

[9] J. L. Flynn, "Immunology of tuberculosis and implications in vaccine development," *Tuberculosis*, vol. 84, no. 1-2, pp. 93–101, 2004.

[10] G. A. W. Rook, G. Seah, and A. Ustianowski, "*M. tuberculosis*: immunology and vaccination," *European Respiratory Journal*, vol. 17, no. 3, pp. 537–557, 2001.

[11] P. Mansueto, G. Vitale, G. Di Lorenzo, G. B. Rini, S. Mansueto, and E. Cillari, "Immunopathology of leishmaniasis: an update," *International Journal of Immunopathology and Pharmacology*, vol. 20, no. 3, pp. 435–445, 2007.

[12] M. T. M. Roberts, "Current understandings on the immunology of leishmaniasis and recent developments in prevention and treatment," *British Medical Bulletin*, vol. 75-76, no. 1, pp. 115–130, 2005.

[13] Y. He, R. Vemulapalli, and G. G. Schurig, "Recombinant *Ochrobactrum anthropi* expressing *Brucella abortus* Cu,Zn superoxide dismutase protects mice against B. abortus infection only after switching of immune responses to Th1 type," *Infection and Immunity*, vol. 70, no. 5, pp. 2535–2543, 2002.

[14] Y. He, R. Vemulapalli, A. Zeytun, and G. G. Schurig, "Induction of specific cytotoxic lymphocytes in mice vaccinated with *Brucella abortus* RB51," *Infection and Immunity*, vol. 69, no. 9, pp. 5502–5508, 2001.

[15] R. S. Wallis, T. M. Doherty, P. Onyebujoh, et al., "Biomarkers for tuberculosis disease activity, cure, and relapse," *The Lancet Infectious Diseases*, vol. 9, no. 3, pp. 162–172, 2009.

[16] E. Jouanguy, F. Altare, S. Lamhamedi, et al., "Interferon-γ-receptor deficiency in an infant with fatal bacille Calmette-Guerin infection," *New England Journal of Medicine*, vol. 335, no. 26, pp. 1956–1961, 1996.

[17] R. Doffinger, M. R. Helbert, G. Barcenas-Morales, et al., "Autoantibodies to interferon-gamma in a patient with selective susceptibility to mycobacterial infection and organ-specific autoimmunity," *Clinical Infectious Diseases*, vol. 38, no. 1, pp. e10–e14, 2004.

[18] B. K. Cohen and L. Hunter, "Natural language processing and systems biology," in *Artificial Intelligence Methods and Tools for Systems Biology*, pp. 147–173, Springer, Berlin, Germany, 2004.

[19] L. Hunter and K. B. Cohen, "Biomedical language processing: what's beyond PubMed?" *Molecular Cell*, vol. 21, no. 5, pp. 589–594, 2006.

[20] A. Zanzoni, L. Montecchi-Palazzi, M. Quondam, G. Ausiello, M. Helmer-Citterich, and G. Cesareni, "MINT: a molecular INTeraction database," *FEBS Letters*, vol. 513, no. 1, pp. 135–140, 2002.

[21] G. D. Bader, D. Betel, and C. W. V. Hogue, "BIND: the biomolecular interaction network database," *Nucleic Acids*, vol. 31, no. 1, pp. 248–250, 2003.

[22] T. S. Keshava Prasad, R. Goel, K. Kandasamy, et al., "Human protein reference database—2009 update," *Nucleic Acids Research*, vol. 37, database issue, supplement 1, pp. D767–D772, 2009.

[23] Online Mendelian Inheritance in Man (OMIM), http://www.ncbi.nlm.nih.gov/omim/.

[24] Z. Xiang, W. Zheng, and Y. He, "BBP: *Brucella* genome annotation with literature mining and curation," *BMC Bioinformatics*, vol. 7, no. 1, article 347, 2006.

[25] Z. Xiang, Y. Tian, and Y. He, "PHIDIAS: a pathogen-host interaction data integration and analysis system," *Genome Biology*, vol. 8, no. 7, article R150, 2007.

[26] A. Özgür, T. Vu, G. Erkan, and D. R. Radev, "Identifying gene-disease associations using centrality on a literature mined gene-interaction network," *Bioinformatics*, vol. 24, no. 13, pp. i277–i285, 2008.

[27] J. C. Reynar and A. Ratnaparkhi, "A maximum entropy approach to identifying sentence boundaries," in *Proceedings of the 5th Conference on Applied Natural Language Processing*, Washington, DC, USA, March-April 1997.

[28] Y. Tsuruoka, Y. Tateishi, J.-D. Kim, et al., "Developing a robust part-of-speech tagger for biomedical text. Advances in informatics," in *Proceedings of the 10th Panhellenic Conference on Informatics*, vol. 3746 of *Lecture Notes in Computer Science*, pp. 382–392, 2005.

[29] Y. Y. Fan and C. Y. Wu, "The role of cytokines in the production of IL-17 and IFN-gamma via the induction of normal human peripheral blood mononuclear cells and CD4(+) T cells," *Xi Bao Yu Fen Zi Mian Yi Xue Za Zhi*, vol. 23, no. 10, pp. 914–916, 2007.

[30] T. A. Eyre, F. Ducluzeau, T. P. Sneddon, S. Povey, E. A. Bruford, and M. J. Lush, "The HUGO gene nomenclature database, 2006 updates," *Nucleic Acids Research*, vol. 34, database issue, pp. D319–D321, 2006.

[31] T. Joachims, "Making large-scale support vector machine learning practical," in *Advances in Kernel Methods: Support Vector Learning*, B. Schölkopf and A. J. Smola, Eds., pp. 169–184, MIT Press, Cambridge, Mass, USA, 1999.

[32] M. C. de Marneffe, B. Maccartney, and C. D. Manning, "Generating typed dependency parses from phrase structure parses," in *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC '06)*, 2006.

[33] G. Erkan, A. Özgür, and D. R. Radev, "Semi-supervised classification for extracting protein interaction sentences using dependency parsing," in *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '07)*, pp. 228–237, 2007.

[34] F. Leitner, M. Krallinger, C. Rodriguez-Penagos, et al., "Introducing meta-services for biomedical information extraction," *Genome Biology*, vol. 9, supplement 2, article S6, 2008.

[35] B. Schwikowski, P. Uetz, and S. Fields, "A network of protein-protein interactions in yeast," *Nature Biotechnology*, vol. 18, no. 12, pp. 1257–1261, 2000.

[36] V. Spirin and L. A. Mirny, "Protein complexes and functional modules in molecular networks," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 100, no. 21, pp. 12123–12128, 2003.

[37] L. C. Freeman, "Centrality in social networks: conceptual clarification," *Social Networks*, vol. 1, no. 3, pp. 215–239, 1978.

[38] M. E. J. Newman, "The structure and function of complex networks," *SIAM Review*, vol. 45, no. 2, pp. 167–256, 2003.

[39] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: bringing order to the web," Tech. Rep., Stanford InfoLab, 1999.

[40] H. Jeong, S. P. Mason, A.-L. Barabasi, and Z. N. Oltvai, "Lethality and centrality in protein networks," *Nature*, vol. 411, no. 6833, pp. 41–42, 2001.

[41] M. P. Joy, A. Brock, D. E. Ingber, and S. Huang, "High-betweenness proteins in the yeast protein interaction network," *Journal of Biomedical Biotechnology*, vol. 2005, no. 2, pp. 96–103, 2005.

[42] M. W. Hahn and A. D. Kern, "Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks," *Molecular Biology and Evolution*, vol. 22, no. 4, pp. 803–806, 2005.

[43] L.-C. Li, H. Zhao, H. Shiina, C. J. Kane, and R. Dahiya, "PGDB: a curated and integrated database of genes related to the prostate," *Nucleic Acids Research*, vol. 31, no. 1, pp. 291–293, 2003.

[44] D. W. Huang, B. T. Sherman, and R. A. Lempicki, "Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources," *Nature Protocols*, vol. 4, no. 1, pp. 44–57, 2009.

[45] D. J. Watts and S. H. Strogatz, "Collective dynamics of 'small-world' networks," *Nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[46] R. Albert and A.-L. Barabasi, "Statistical mechanics of complex networks," *Reviews of Modern Physics*, vol. 74, no. 1, pp. 47–97, 2002.

[47] H. Chen and B. M. Sharp, "Content-rich biological network constructed by mining PubMed abstracts," *BMC Bioinformatics*, vol. 5, article 147, 2004.

[48] R. Hoffmann and A. Valencia, "Implementing the iHOP concept for navigation of biomedical literature," *Bioinformatics*, vol. 21, 2, pp. ii252–ii258, 2005.

[49] S.-H. Yook, H. Jeong, and A.-L. Barabasi, "Modeling the internet's large-scale topology," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 99, no. 21, pp. 13382–13386, 2002.

[50] U. Bommhardt, K. C. Chang, P. E. Swanson, et al., "Akt decreases lymphocyte apoptosis and improves survival in sepsis," *Journal of Immunology*, vol. 172, no. 12, pp. 7583–7591, 2004.

[51] Y. Avitzur, E. Galindo-Mata, and N. L. Jones, "Oral vaccination against Helicobacter pylori infection is not effective in mice with fas ligand deficiency," *Digestive Diseases and Sciences*, vol. 50, no. 12, pp. 2300–2306, 2005.

[52] B. Langhans, S. Schweitzer, I. Braunschweiger, M. Schulz, T. Sauerbruch, and U. Spengler, "Cytotoxic capacity of hepatitis C virus (HCV)–specific lymphocytes after in vitro immunization with HCV-derived lipopeptides," *Cytometry A*, vol. 65, no. 1, pp. 59–68, 2005.

[53] G. Shi, J. Mao, G. Yu, J. Zhang, and J. Wu, "Tumor vaccine based on cell surface expression of DcR3/TR6," *Journal of Immunology*, vol. 174, no. 8, pp. 4727–4735, 2005.

[54] C. A. Dinarello, "Interleukin-18," *Methods*, vol. 19, no. 1, pp. 121–132, 1999.

[55] V. G. Tarcea, T. Weymouth, and A. Ade, "Michigan molecular interactions r2: from interacting proteins to pathways," *Nucleic Acids Research*, vol. 37, database issue, pp. D642–D646, 2009.

[56] H.-S. Park, S.-H. Huh, M.-S. Kim, S. H. Lee, and E.-J. Choi, "Nitric oxide negatively regulates c-Jun N-terminal kinase/stress-activated protein kinase by means of S-nitrosylation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 97, no. 26, pp. 14382–14387, 2000.

[57] M. Pearl-Yafe, D. Halperin, O. Scheuerman, and I. Fabian, "The p38 pathway partially mediates caspase-3 activation induced by reactive oxygen species in Fanconi anemia C cells," *Biochemical Pharmacology*, vol. 67, no. 3, pp. 539–546, 2004.

[58] B. Lu, H. Yu, C.-W. Chow, et al., "GADD45γ mediates the activation of the p38 and JNK MAP kinase pathways and cytokine production in effector TH1 cells," *Immunity*, vol. 14, no. 5, pp. 583–590, 2001.

[59] K. Hayashi, S. Ishizuka, C. Yokoyama, and T. Hatae, "Attenuation of interferon-γ mRNA expression in activated Jurkat T cells by exogenous zinc via down-regulation of the calcium-independent PKC-AP-1 signaling pathway," *Life Sciences*, vol. 83, no. 1-2, pp. 6–11, 2008.

[60] L. Xue, G. L. Firestone, and L. F. Bjeldanes, "DIM stimulates IFNγ gene expression in human breast cancer cells via the specific activation of JNK and p38 pathways," *Oncogene*, vol. 24, no. 14, pp. 2343–2353, 2005.

Hindawi Publishing Corporation
Journal of Biomedicine and Biotechnology
Volume 2010, Article ID 856842, 8 pages
doi:10.1155/2010/856842

Methodology Report

IMMUNOCAT—A Data Management System for Epitope Mapping Studies

Jo L. Chung, Jian Sun, John Sidney, Alessandro Sette, and Bjoern Peters

Division of Vaccine Discovery, La Jolla Institute for Allergy & Immunology, La Jolla, CA 92037, USA

Correspondence should be addressed to Bjoern Peters, bpeters@liai.org

Received 30 November 2009; Accepted 7 March 2010

Academic Editor: Anne S. De Groot

To enable rationale vaccine design, studies of molecular and cellular mechanisms of immune recognition need to be linked with clinical studies in humans. A major challenge in conducting such translational research studies lies in the management and integration of large amounts and various types of data collected from multiple sources. For this purpose, we have established "IMMUNOCAT", an interactive data management system for the epitope discovery research projects conducted by our group. The system provides functions to store, query, and analyze clinical and experimental data, enabling efficient, systematic, and integrative data management. We demonstrate how IMMUNOCAT is utilized in a large-scale research contract that aims to identify epitopes in common allergens recognized by T cells from human donors, in order to facilitate the rational design of allergy vaccines. At clinical sites, demographic information and disease history of each enrolled donor are captured, followed by results of an allergen skin test and blood draw. At the laboratory site, T cells derived from blood samples are tested for reactivity against a panel of peptides derived from common human allergens. IMMUNOCAT stores results from these T cell assays along with MHC:peptide binding data, results from RAST tests for antibody titers in donor serum, and the respective donor HLA typing results. Through this system, we are able to perform queries and integrated analyses of the various types of data. This provides a case study for the use of bioinformatics and information management techniques to track and analyze data produced in a translational research study aimed at epitope identification.

1. Introduction

A crucial step for rational subunit vaccine design is the selection of antigens to include. For vaccines against infectious agents, antigens capable of inducing protective immune responses are desired. Several strategies based on genomic and proteomic approaches are being used to identify subsets of antigens that are highly expressed in general [1], on the surface [2], or during infection [3]. Antigens from these priority subsets are then followed up individually to test for their capacity to induce protective immunity. An alternative strategy that identifies protective antigens directly is to map targets of immune responses in previously infected hosts that successfully cleared the infection. This strategy is applicable whenever past infectious are known to provide protective immunity. In those cases, the capacity of antigens to induce protective immunity in a vaccine setting has been shown to correlate with the magnitude of the response against that antigen post infection [4]. Therefore, knowledge of targets of immune responses in infected hosts has high value for vaccine design against infectious diseases. Knowing immune response targets is also crucial for the development of allergy vaccines, whose goal is to modulate the pathologic immune responses of allergic individuals towards those found in non-allergics [5, 6]. Similarly, for cancer vaccines to be successful, it is necessary to identify antigens targeted by immune responses associated with tumor regression [7, 8]. In summary, identifying the targets and characteristics of immune responses in well characterized host populations enables the rational design of vaccines.

One established approach for the identification and characterization of T cell immune response is the use of peptide based epitope mapping strategies. These are especially efficient when used in combination with bioinformatics predictions of candidate peptides [9]. The identification of epitopes, the exact molecular unit of recognition within an

antigen, also provides a mechanistic understanding of cross-reactivity of immune responses for different pathogens. This has recently been applied to study T cell immunity to swine flu [10, 11], and is important when designing cross-protective vaccines.

We have participated in two recently completed large-scale T cell epitope mapping projects, one to characterize epitopes responsible for the protective immunity conveyed by the smallpox vaccine [12–15], another to characterize epitopes in Arenaviruses [16–18], which has led to the generation of a candidate for a cross protective vaccine (M. Kotturi et al., PLoS Pathogens, in press). One lesson learned from these studies is that their data management is challenging, as the epitope response patterns discovered are typically complicated [19]. Also, these studies require the integration of large amounts and various types of data collected from multiple clinical and laboratory sites. Like many other groups, we have managed these data in a collection of spreadsheets, lab notebooks, and database systems designed for a single type of experiment. While each of these provides a sufficient mechanism to capture a specific type of information, the integrated analysis of these data often becomes labor intensive. Worse, problems due to inconsistencies in nomenclature and incompleteness of datasets are often only discovered at the time of analysis rather than at the time of data entry, which can make it hard or impossible to rectify them.

One way to address these issues is to collect data, from the start, in an integrated database system which immediately connects data from different sources. While this requires more work upon data entry, compared to less formal means of data capture such as spreadsheets, it greatly reduces the effort for data analysis. A deciding factor for our group to move toward an integrated database system was the award of a large scale NIH contract that aims to identify T cell epitopes from common allergens with the ultimate goal to facilitate the development of improved allergy vaccines. This study involves tracking donors enrolled at two clinical sites, capturing complex clinical history information, and storing results from multiple experiments performed in-house and through external providers (see Figure 1 for a breakdown of the overall study). Our experience with previous studies which were less complex, suggested that the maintenance and analysis of these data would be challenging. We therefore decided to establish IMMUNOCAT, which aims to store all data relevant for an integrated analysis of epitope mapping experiments, with the ultimate goal to enable rational vaccine design.

IMMUNOCAT integrates and enhances existing information management systems in our laboratory. Two standalone web-based relational databases were previously developed. ELICAT stores and analyzes results from ELISPOT [20] experiments used to test T cells for their cytokine secretion profile in response to peptide antigens. TOPCAT was designed to store the binding affinities determined in competitive peptide:MHC binding assays [21] utilizing a TopCount instrument. The databases are linked by the use of common peptide identifiers, which are assigned to every peptide that is synthesized in our studies. For the allergen epitope mapping study, we needed to link the results of T cell reactivities stored in ELICAT with information about the donor from whom the T cells were derived. We first built a new web-based database called DONORDB to store answers from a clinical questionnaire, skin test results, and records of blood samples. Subsequently, this donor database was integrated with the existing ELICAT and TOPCAT databases (Figure 2) through the use of shared MHC allele and donation identifiers. The result is IMMUNOCAT, an integration of three web-based databases.

In the following, we describe the features and components of IMMUNOCAT, which has been in stable use in our laboratory for over a year. With this system, we are able to manage all data relevant to our epitope mapping studies in a central location and ensure their quality and integrity. Answers to integrated queries that would previously have taken significant effort can now be determined instantly, such as how many blood samples are available from donors that showed T cell reactivity to a certain peptide. In addition to the allergy epitope mapping study described here, IMMUNOCAT has been modified for use in two recently initiated epitope mapping studies for Dengue Virus and Mycobacterium Tuberculosis. This demonstrates the benefits of applying bioinformatics and information management techniques to epitope mapping studies in order to facilitate rational vaccine development.

2. User Administration

IMMUNOCAT provides different functionality to different groups of users. Each user has an individual account created by a database administrator. In order to access the database through the web, users have to identify themselves by supplying a user name and password. Different functionalities are provided based on the user group assignments. For example, there are two primary groups of users for DONORDB. The first group represents staff members at the clinical sites, who enter information from the clinical questionnaire and results from the skin test and blood draw. The second group is staff members at the laboratory site, who track the use of blood samples for each donor. Different users within each group are assigned different levels of access. For example, lab scientist supervisors have the ability to audit, add, delete, or modify the information in the system as necessary. Other lab scientists can only enter raw data and must request lab scientist supervisors to correct, for example, data entry errors. Separate projects using the DONORDB, such as the epitope mapping studies for Dengue Virus, use different web addresses for user access.

3. System Components and Features

3.1. DONORDB at Clinical Site. DONORDB is used at the clinical sites to capture information as donors are enrolled, interviewed, and undergo clinical procedures. For our ongoing allergen epitope mapping study, demographic information of the donor, such as gender, birthdate, ethnicity and parents' birthplaces (as an additional way to identify

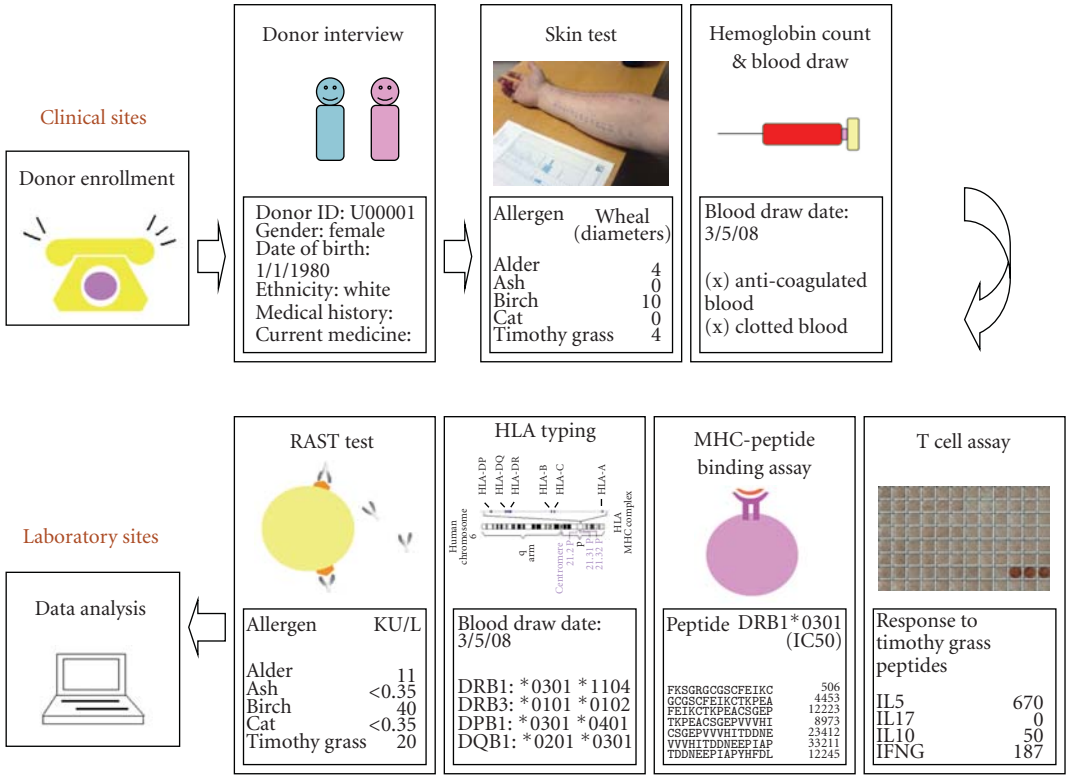


Figure 1: The allergy epitope mapping research workflow. In this study, allergic and normal donors were enrolled at two clinical sites (University of California, San Diego and National Jewish Medical Research Center). During the clinical interview, donor demographic information and medical history were recorded. Following that, allergen skin tests, hemoglobin test, and blood sample collection were performed. All the data and blood samples were sent to LIAI for further experiments and analyses. At LIAI, blood samples were tested for their immunogenicity (T cell activities) against a selected panel of peptides derived from common human allergens by ELISPOT assays [20] (T cell assays). Concurrently, the selected peptides were tested for their binding capacities to a set of common MHC molecules [21]. Peptide binding to MHC molecules is a necessary, but not sufficient, requirement for T cell recognitions [22]. Blood samples were also sent to an external company for radioallergosorbent (RAST) and human leukocyte antigens (HLA) typing test. The RAST test detects the amount of the IgE antibody in the blood that reacts with a specific allergen [23]. HLA typing determines the specific HLA types expressed by the donor, which is an indication of MHC restriction [24, 25]. The figure of skin test was adapted from the Wikimedia Commons file "File:Allergy_skin_testing.jpg". The figure of HLA typing was adapted from the Wikimedia Commons file "File:HLA.jpg".

ethnicity), is first entered. Next, a questionnaire is completed based on donor interviews (Figure 3). The questionnaire is tied to a scoring scheme that is used to classify donors into different disease categories, namely, three classifications of "allergic rhinitis" (none, possible, and probable) and "allergic asthma". In terms of clinical procedures, the results of a prick skin test for allergic reactions against 32 common allergens are recorded in terms of flare and wheal diameters. Also, the hemoglobin count of a donor is measured and recorded to determine if he/she is safely capable of donating blood. If so, a separate blood donation visit is scheduled, and the volume of blood drawn is recorded in the database before samples are shipped for further analysis to the laboratory site.

At every step, the system ensures that donors meet all enrollment criteria, by notifying clinical staff if a donor falls under one of the exclusion criteria based on the entered information. During initial enrollment, exclusion would occur if, for example, the patient states that he is currently receiving allergen immunotherapy without yet being in maintenance. After initial enrollment, exclusion would occur, for example, when the hemoglobin count is too low to allow a blood draw, or when the donor is still taking antihistamine medication right before the skin test. Integrating these questions and criteria into the database makes it easier for the clinical sites to keep track of all aspects of the enrollment criteria, and promotes consistency between sites.

Only anonymized information about the donors is entered at the clinical sites. As donors are enrolled, they get assigned a donation identifier by the system. This identifier is used to match blood samples shipped to the laboratory site with the information collected at the clinical sites. The clinical sites store the donation identifiers in their general patient records, which allow them to, for example, match repeat donations of a patient pre- and post immunotherapy, or match existing patient identifiers established at the individual sites to the DONORDB donation identifies. No personally identifiable information (such as names or social security numbers) is stored in the database, which ensures that all anonymization requirements for data analysis are met.

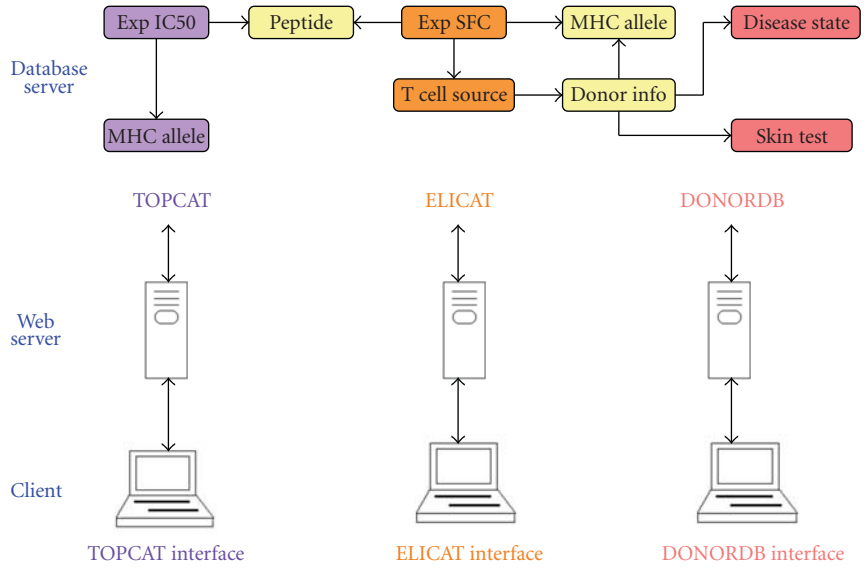


Figure 2: The architecture of IMMUNOCAT system. IMMUNOCAT is an integration of three web accessible relational databases. DONORDB is designed for data collected during donor enrollment and clinical tests. ELICAT stores the data from T cell assays performed using donor blood samples stimulated with a set of allergen derived peptides. TOPCAT stores the MHC-peptide binding data of the same set of peptides. Simplified database structures are presented at the top of the figure. Arrows represent the links between database tables. Tables shared by two databases are colored yellow. In DONORDB, donor information table is linked to skin test, disease state, and MHC allele tables. In ELICAT, table of experimental spot forming cell (SFC) is linked to peptide, T cell source, and MHC allele tables. In TOPCAT, table of experimental IC50 is linked to peptide and MHC allele tables. DONORDB is connected with ELICAT using donor IDs and MHC allele IDs. ELICAT and TOPCAT are connected with each other using shared peptide IDs.

By default, the data entry process follows the order in which the study was envisioned to be conducted. However, this order can be specifically overruled by users. At any time point, users can save completed data entry forms, and log out. They can log back in, identify the donor they were working on based on the donation ID, and continue data entry. If necessary, users can jump to different entry parts at their convenience. This was identified as critical functionality, as in clinical practice the idealized workflow outlined above is often interrupted due to changes in patients' schedules, and availability of time for data entry. Similarly, data entry requires internet access which for various reasons may not always be available. We found that enrollment information at clinical sites is therefore often recorded on paper before transferred into the database.

3.2. DONORDB at Laboratory Site. The lab scientists receive blood samples from the clinical sites, and process them to extract peripheral blood mononuclear cells (PBMCs). They use the database to track the availability of PBMCs from each donation, determine the number of vials that could initially be made from the blood shipment, record the use of vials in subsequent experiments, and track the location of the remaining vials in the freezer.

For each donor, data from two experiments performed at outside commercial laboratories are stored directly in DONORDB. The first set of data consists of IgE antibody titers against a panel of allergens determined in a RAST assay. The second set of data consists of the results from HLA typing assays, which determine the specific set of MHC molecules the donor expresses.

Simple queries can be run against DONORDB in a web-based form (Figure 4). At this point, the system provides query functions to retrieve results from skin tests and the receipt of blood samples, as well as to monitor the number of donors in the two clinical sites. For example, lab scientists can query the database and see whether the blood samples of specific donors have been received at LIAI. They can also identify the donors who were skin test positive for a particular allergen and recruited at a specific clinical site. These queries are used routinely by laboratory scientists to select suitable blood samples for experiments.

3.3. ELICAT Database. ELICAT was designed for the management of data from ELISPOT assays. In these assays [13, 18], T cells are stimulated with peptide antigens, and those that respond by producing cytokines are visualized as individual spots. These recognized peptides are potential candidates for inclusion in an allergy vaccine. The experiments are performed on 96-well plates, and the numbers of spots are counted by an automated ELISPOT reader. The raw data generated is exported to a text file containing a matrix whose elements denote the number of spots detected in each well. This raw data is automatically imported into ELICAT, where it is connected with information about experimental design previously entered through the interface shown in Figure 5. An experiment is defined as one or more ELISPOT plates run with the same layout and cells from the same source. The plate layout specifies which wells are used to

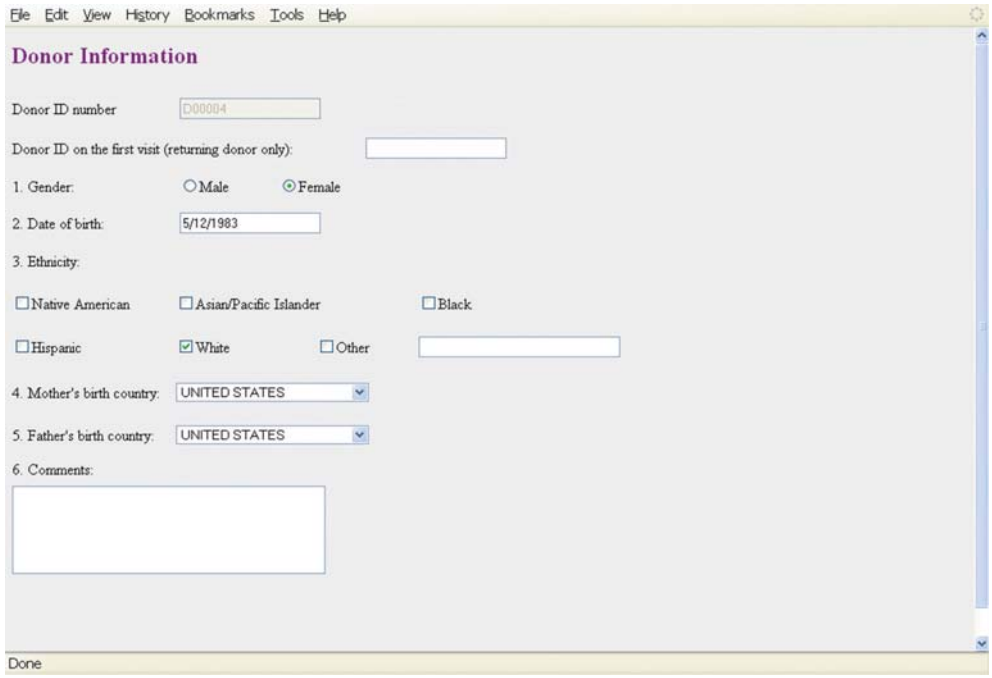
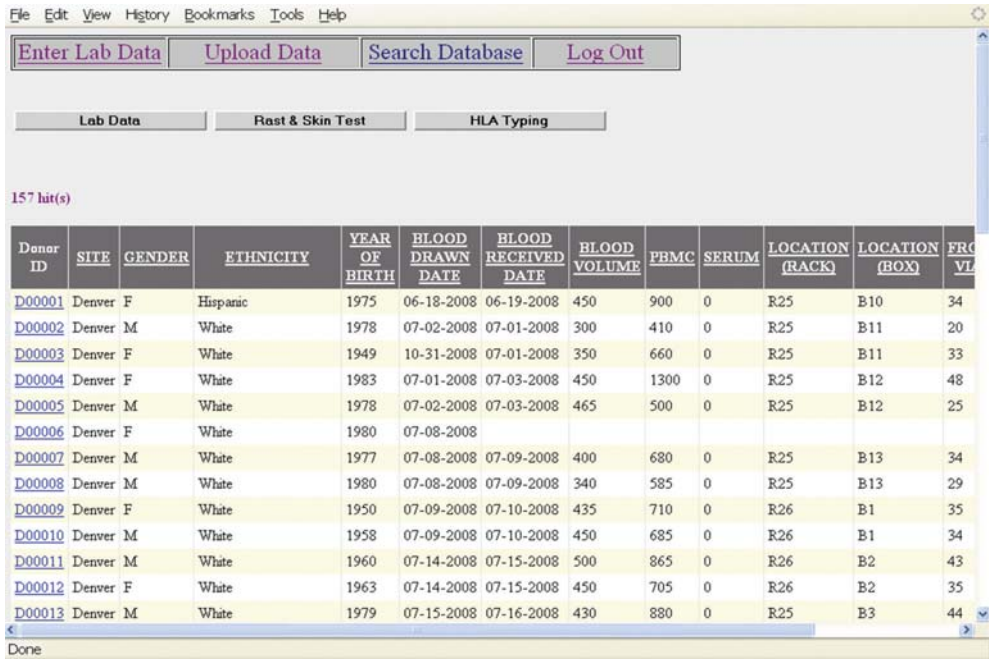


Figure 3: DONORDB: demographic information. The Figure displays a screen shot of the form used during initial enrollment of a donor.



Donor ID	SITE	GENDER	ETHNICITY	YEAR OF BIRTH	BLOOD DRAWN DATE	BLOOD RECEIVED DATE	BLOOD VOLUME	PBMC	SERUM	LOCATION (RACK)	LOCATION (BOX)	FRO VI
D00001	Denver	F	Hispanic	1975	06-18-2008	06-19-2008	450	900	0	R25	B10	34
D00002	Denver	M	White	1978	07-02-2008	07-01-2008	300	410	0	R25	B11	20
D00003	Denver	F	White	1949	10-31-2008	07-01-2008	350	660	0	R25	B11	33
D00004	Denver	F	White	1983	07-01-2008	07-03-2008	450	1300	0	R25	B12	48
D00005	Denver	M	White	1978	07-02-2008	07-03-2008	465	500	0	R25	B12	25
D00006	Denver	F	White	1980	07-08-2008							
D00007	Denver	M	White	1977	07-08-2008	07-09-2008	400	680	0	R25	B13	34
D00008	Denver	M	White	1980	07-08-2008	07-09-2008	340	585	0	R25	B13	29
D00009	Denver	F	White	1950	07-09-2008	07-10-2008	435	710	0	R26	B1	35
D00010	Denver	M	White	1958	07-09-2008	07-10-2008	450	685	0	R26	B1	34
D00011	Denver	M	White	1960	07-14-2008	07-15-2008	500	865	0	R26	B2	43
D00012	Denver	F	White	1963	07-14-2008	07-15-2008	450	705	0	R26	B2	35
D00013	Denver	M	White	1979	07-15-2008	07-16-2008	430	880	0	R25	B3	44

Figure 4: DONORDB: search functionality. The Figure shows a screen shot of the results generated when querying for all enrolled donors.

hold the tested peptides and which wells are used to hold the controls. The peptides used are specified in a separate file containing peptide identifiers.

Based on these inputs, ELICAT calculates summary metrics for the ELISPOT results for each peptide or peptide pool. The fraction of spot forming cells per million is calculated based on the number of spots detected in a well and the number of effector T cells added. Based on replicate spot counts, a P value is calculated that evaluates if the detected spots for a peptide are significantly higher than those detected in negative control wells that contained no peptides. Finally, a stimulation index, which divides the average spot count for a peptide by the average background value, is calculated.

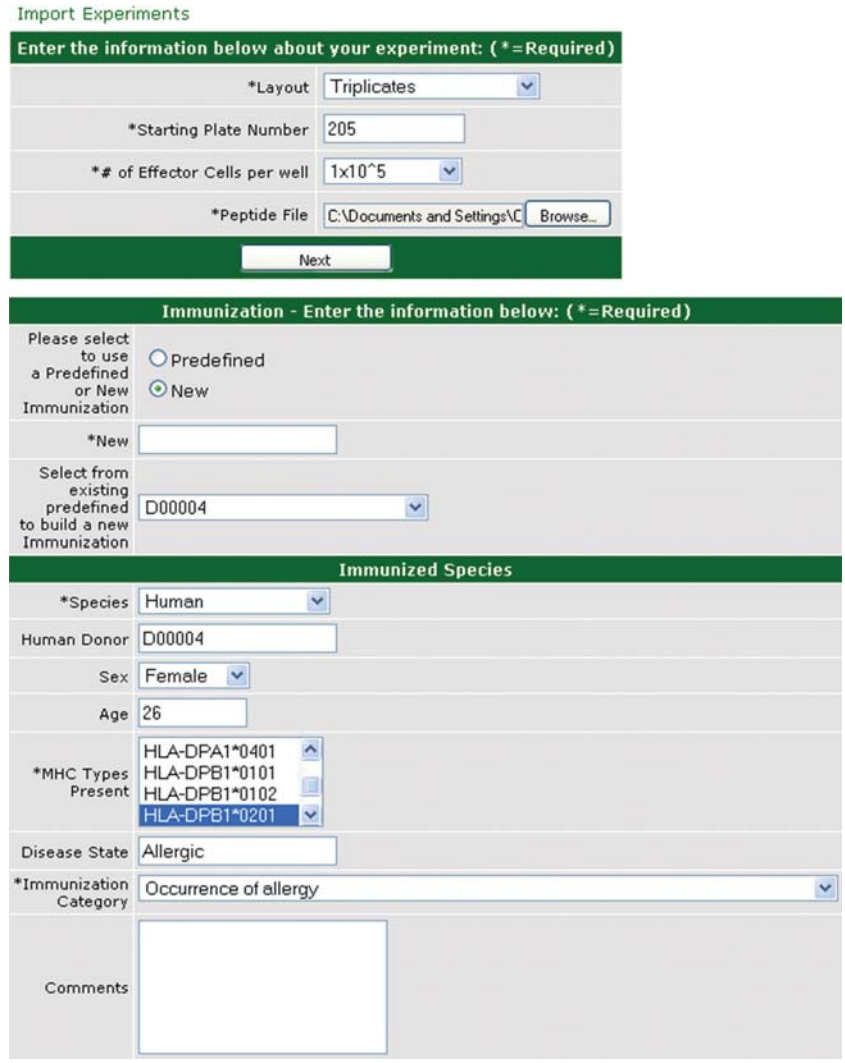


Figure 5: ELICAT database: experimental design functionality. The figure shows two screens used to capture the experimental design of an ELISPOT experiment. By specifying the donor from whom T cells used in the experiment were derived, the experiment becomes linked to the information for that donor in DONORDB, and some fields such as the HLA alleles expressed by that donor become prepopulated. ELICAT also allows the users to specify the type of cytokines and cells used in the experiments (not shown here).

In the allergen epitope mapping study, two sets of ELISPOT experiments are performed to identify peptide epitopes. First, pools of peptides are screened for reactivity with PBMCs from allergic donors. Secondly, the positive pools are deconvoluted to identify which peptides caused the response. As a cutoff for the positive response >100 spot forming cells per million, $P < .05$ and stimulation index >2 were used. The ELICAT user interface allows description of each assay, including identification of each individual peptide tested or definition of each peptide component of specific pools.

3.4. TOPCAT Database. TOPCAT was the first database established in the laboratory, and its design was largely replicated in ELICAT. TOPCAT is used to store results from assays that evaluate MHC:peptide binding affinity, a necessary, but not sufficient, requirement for T cell recognition [21]. In our assays [26, 27], the ability of the tested peptides at different concentrations to completely inhibit the binding of a radiolabeled high affinity ligand is determined. The concentration of the tested peptide at which the number of labeled ligands bound is reduced by 50% is the IC50 value. Under the condition utilized, measured IC50 values approximate the Kd value of the binding interaction [28]. The IC50 values are calculated based on radioactivity detection in a 96-well plate measured by TOPCOUNT NXT microscintillation reader. Information on the tested peptides, MHC alleles, and corresponding MHC-peptide binding affinities are stored and analyzed in the database.

All peptides used at the laboratory site are assigned a peptide identifier in the TOPCAT database, and the tubes containing the peptide are labeled accordingly. This identifies the sequence of the peptide and the protein and organism from which the peptide was derived, as well as the purity

of the synthesis. If the same peptide is synthesized multiple times, a new identifier is assigned to each unique synthesis. The ELICAT database uses the same peptide identifiers, and uses them to retrieve information about peptides from the TOPCAT system.

For the allergen epitope mapping study, peptides derived from protein sequences of common human allergens were synthesized. Each is being tested for binding affinity against a panel of 23 human MHC class II alleles. Through the integration of the three databases shown in Figure 2, it now becomes possible to directly link the allergic status of a donor, the peptides derived from the corresponding allergens, and their binding affinity to the MHC alleles present in the donor and the T cell reactivities detected towards these peptides.

4. System Design Process and Implementation

ELICAT and TOPCAT have been in routine use in the lab for more than 5 years. Regarding DONORDB, requirements were gathered in terms of what information needs to be captured for the process of donor enrollment, donor interviews, clinical procedures, blood sample shipments, and lab tests. For each step, the information desired to track progress was identified. Based on this, prototype web pages were created, and discussed with the clinical collaborators and LIAI lab scientists. Test of these prototypes was performed by both sites that analyzed use-scenarios, which led to the identification of additional requirements. These were, primarily, giving the users more flexibility in the order in which data was entered, and the capacity to store additional information. After three iterations, the prototypes were considered complete, and a functional system was implemented. The addition of more fields for new types of studies will require going through the same design process and modifications of the database and web application itself. As we gather experience in the kinds of modifications to expect, we plan to create tools that will ease making standard extensions.

All three databases are relational databases with web-browser-based interfaces and are implemented with SQL server 2005. The user interface was implemented with ASP for TOPCAT, and ASP.NET 2.0 for DONORDB and ELICAT. All three databases are hosted on a Dual-Processor Quad-Core Intel Xeon Rackmount Server. The database information is stored on a RAID 5 drive which ensures against single hard-disk failure, and is backed up daily through the LIAI network.

5. Usage

At this point, clinical information for 86 donors at the UCSD sites and 71 donors at NJMRC was deposited into DONORDB. Immunogenicity and MHC binding data for more than 60,000 peptides are stored in ELICAT and TOPCAT. We have not, so far, encountered data loss during operation. The source code of IMMUNOCAT is available for download at http://donor.liai.org/Donor_Source.zip. Users will need to have an installation of SQL Server 2005 DBMS and ASP.NET 2.0 framework, and the ability to customize and configure the database for their own purposes. The code will be maintained for at least the next four years, and updates will be made available as new versions of the systems are completed. Prospective users are highly encouraged to contact the authors with any installation problems. For studies that require extensive modifications of the code and for laboratories operating incompatible IT environments, it will be preferable to redesign the application from scratch. In those cases, the present manuscript should still be useful in identifying requirements and reusing appropriate design patterns.

6. Future Prospects

We are planning to implement more search and analysis interfaces that take advantage of the integrated data in IMMUNOCAT. Currently, such analyses are being done by manually running SQL queries, which cannot be expected of laboratory technicians.

We are in the process of extending DONORDB to handle donors from three additional epitope-mapping studies, one dealing with donors from Dengue fever endemic regions, another dealing with donors that have tuberculosis, and a third with donors recently vaccinated against smallpox. This will require modifying some of the information captured for each donor, such as the serological typing of patients for the type of dengue viruses they have been exposed to. Other aspects, such as MHC typing and blood sample collection, will remain the same. In the longer term, we are aiming to make the customization of DONORDB for individual studies easier by allowing the addition of fields from predefined modules. This could, for example, mean that different lab tests performed could be chosen from an ontology such as the Ontology for Biomedical Investigations [29], and that vaccine terms could be taken from the Vaccine Ontology [30].

Finally, one of the goals of IMMUNOCAT is to submit data from completed studies into the Immune Epitope Database [31, 32]. Currently, both TOPCAT and ELICAT have export mechanisms that provide XML formatted data that can be imported into the IEDB. These mechanisms need to be integrated and further updated to integrate the data from DONORDB, and provide comprehensive export functionality of an entire study.

Acknowledgments

This work was supported by NIH contracts HHSN27220070-0048C, HHSN272200900042C, and HHSN272200900044C, and Grant no.5 T32 AI00749-14. The authors want to thank Carla Oseroff, Ravi Kolla, Carrie Moore, David Broide, Debbie Broide, Rafeul Alam, Linda Bannister, Susanna Burr, and Howard Grey for helpful discussions.

References

[1] G. Grandi, "Antibacterial vaccine design using genomics and proteomics," *Trends in Biotechnology*, vol. 19, no. 5, pp. 181–188, 2001.

[2] P. A. Cullen and C. E. Cameron, "Progress towards an effective syphilis vaccine: the past, present and future," *Expert Review of Vaccines*, vol. 5, no. 1, pp. 67–80, 2006.

[3] C. Rollenhagen, M. Sörensen, K. Rizos, R. Hurvitz, and D. Bumann, "Antigen selection based on expression levels during infection facilitates vaccine development for an intracellular phathogen," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 23, pp. 8739–8744, 2004.

[4] M. Moutaftsi, S. Salek-Ardakani, M. Croft, et al., "Correlates of protection efficacy induced by vaccinia virus-specific $CD8^+$ T-cell epitopes in the murine intranasal challenge model," *European Journal of Immunology*, vol. 39, no. 3, pp. 717–722, 2009.

[5] D. Campbell, R. H. DeKruyff, and D. T. Umetsu, "Allergen immunotherapy: novel approaches in the management of allergic diseases and asthma," *Clinical Immunology*, vol. 97, no. 3, pp. 193–202, 2000.

[6] R. Valenta and D. Kraft, "From allergen structure to new forms of allergen-specific immunotherapy," *Current Opinion in Immunology*, vol. 14, no. 6, pp. 718–727, 2002.

[7] J. A. Berzofsky, M. Terabe, S. Oh, et al., "Progress on new vaccine strategies for the immunotherapy and prevention of cancer," *Journal of Clinical Investigation*, vol. 113, no. 11, pp. 1515–1525, 2004.

[8] R. Offringa, S. H. Van Der Burg, F. Ossendorp, R. E. Toes, and C. J. Melief, "Design and evaluation of antigen-specific vaccination strategies against cancer," *Current Opinion in Immunology*, vol. 12, no. 5, pp. 576–582, 2000.

[9] A. S. De Groot, H. Sbai, C. S. Aubin, J. McMurry, and W. Martin, "Immuno-informatics: mining genomes for vaccine components," *Immunology and Cell Biology*, vol. 80, no. 3, pp. 255–269, 2002.

[10] A. S. De Groot, M. Ardito, E. M. McClaine, L. Moise, and W. D. Martin, "Immunoinformatic comparison of T-cell epitopes contained in novel swine-origin influenza A (H1N1) virus with epitopes in 2008-2009 conventional influenza vaccine," *Vaccine*, vol. 27, no. 42, pp. 5740–5747, 2009.

[11] J. A. Greenbaum, M. F. Kotturi, Y. Kim, et al., "Pre-existing immunity against swine-origin H1N1 influenza viruses in the general human population," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 48, pp. 20365–20370, 2010.

[12] M. Moutaftsi, H.-H. Bui, B. Peters, et al., "Vaccinia virus-specific $CD4^+$ T cell responses target a set of antigens largely distinct from those targeted by $CD8^+$ T cell responses," *Journal of Immunology*, vol. 178, no. 11, pp. 6814–6820, 2007.

[13] M. Moutaftsi, B. Peters, V. Pasquetto, et al., "A consensus epitope prediction approach identifies the breadth of murine $TCD8^+$-cell responses to vaccinia virus," *Nature Biotechnology*, vol. 24, no. 7, pp. 817–819, 2006.

[14] V. Pasquetto, H.-H. Bui, R. Giannino, et al., "HLA-A*0201, HLA-A*1101, and HLA-B*0702 transgenic mice recognize numerous poxvirus determinants from a wide variety of viral gene products," *Journal of Immunology*, vol. 175, no. 8, pp. 5504–5515, 2005.

[15] C. Oseroff, F. Kos, H.-H. Bui, et al., "HLA class I-restricted responses to vaccinia recognize a broad array of proteins mainly involved in virulence and viral gene regulation," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 39, pp. 13980–13985, 2005.

[16] J. Botten, J. Alexander, V. Pasquetto, et al., "Identification of protective Lassa virus epitopes that are restricted by HLA-A2," *Journal of Virology*, vol. 80, no. 17, pp. 8351–8361, 2006.

[17] J. Botten, J. L. Whitton, P. Barrowman, et al., "HLA-A2-restricted protection against lethal lymphocytic choriomeningitis," *Journal of Virology*, vol. 81, no. 5, pp. 2307–2317, 2007.

[18] M. F. Kotturi, B. Peters, F. Buendia-Laysa Jr., et al., "The $CD8^+$T-cell response to lymphocytic choriomeningitis virus involves the L antigen: uncovering new tricks for an old virus," *Journal of Virology*, vol. 81, no. 10, pp. 4928–4940, 2007.

[19] A. Sette and B. Peters, "Immune epitope mapping in the post-genomic era: lessons for vaccine development," *Current Opinion in Immunology*, vol. 19, no. 1, pp. 106–110, 2007.

[20] C. C. Czerkinsky, L. A. Nilsson, and H. Nygren, "A solid-phase enzyme-linked immunospot (ELISPOT) assay for enumeration of specific antibody-secreting cells," *Journal of Immunological Methods*, vol. 65, no. 1-2, pp. 109–121, 1983.

[21] A. Sette, A. Vitiello, B. Reherman, et al., "The relationship between class I binding affinity and immunogenicity of potential cytotoxic T cell epitopes," *Journal of Immunology*, vol. 153, no. 12, pp. 5586–5592, 1994.

[22] J. Klein, *Natural History of the Major Histocompatibility Complex*, John Wiley & Sons, New York, NY, USA, 1986.

[23] G. Virella, *Medical Immunology*, Marcel Dekker, New York, NY, USA, 5th edition, 2001.

[24] J. B. Henry, *Clinical Diagnosis and Management by Laboratory Methods*, W.B. Saunders, Philadelphia, Pa, USA, 20th edition, 2001.

[25] V. Vengelen-Tyler and American Association of Blood Banks, *Technical Manual*, American Association of Blood Banks, Bethesda, Md, USA, 12th edition, 1996.

[26] J. Sidney, S. Southwood, C. Oseroff, M. F. del Guercio, A. Sette, and H. M. Grey, "Measurement of MHC/peptide interactions by gel filtration," *Current Protocols in Immunology*, chapter 18, unit 18.3, 2001.

[27] J. Sidney, E. Assarsson, C. Moore, et al., "Quantitative peptide binding motifs for 19 human and mouse MHC class i molecules derived using positional scanning combinatorial peptide libraries," *Immunome Research*, vol. 4, no. 1, article 2, 2008.

[28] Y. Cheng and W. H. Prusoff, "Relationship between the inhibition constant (K_1) and the concentration of inhibitor which causes 50 per cent inhibition (I_{50}) of an enzymatic reaction," *Biochemical Pharmacology*, vol. 22, no. 23, pp. 3099–3108, 1973.

[29] T. O. Consortium, "Modeling biomedical experimental processes with OBI," in *Proceedings of the the 12th Annual Bio-Ontologies Meeting (ISMB '09)*, p. 41, 2009.

[30] Y. He, "VO: vaccine ontology," in *Nature Precedings*, 2009.

[31] B. Peters and A. Sette, "Integrating epitope data into the emerging web of biomedical knowledge resources," *Nature Reviews Immunology*, vol. 7, no. 6, pp. 485–490, 2007.

[32] B. Peters, J. Sidney, P. Bourne, et al., "The immune epitope database and analysis resource: from vision to blueprint," *PLoS Biology*, vol. 3, no. 3, article e91, 2005.