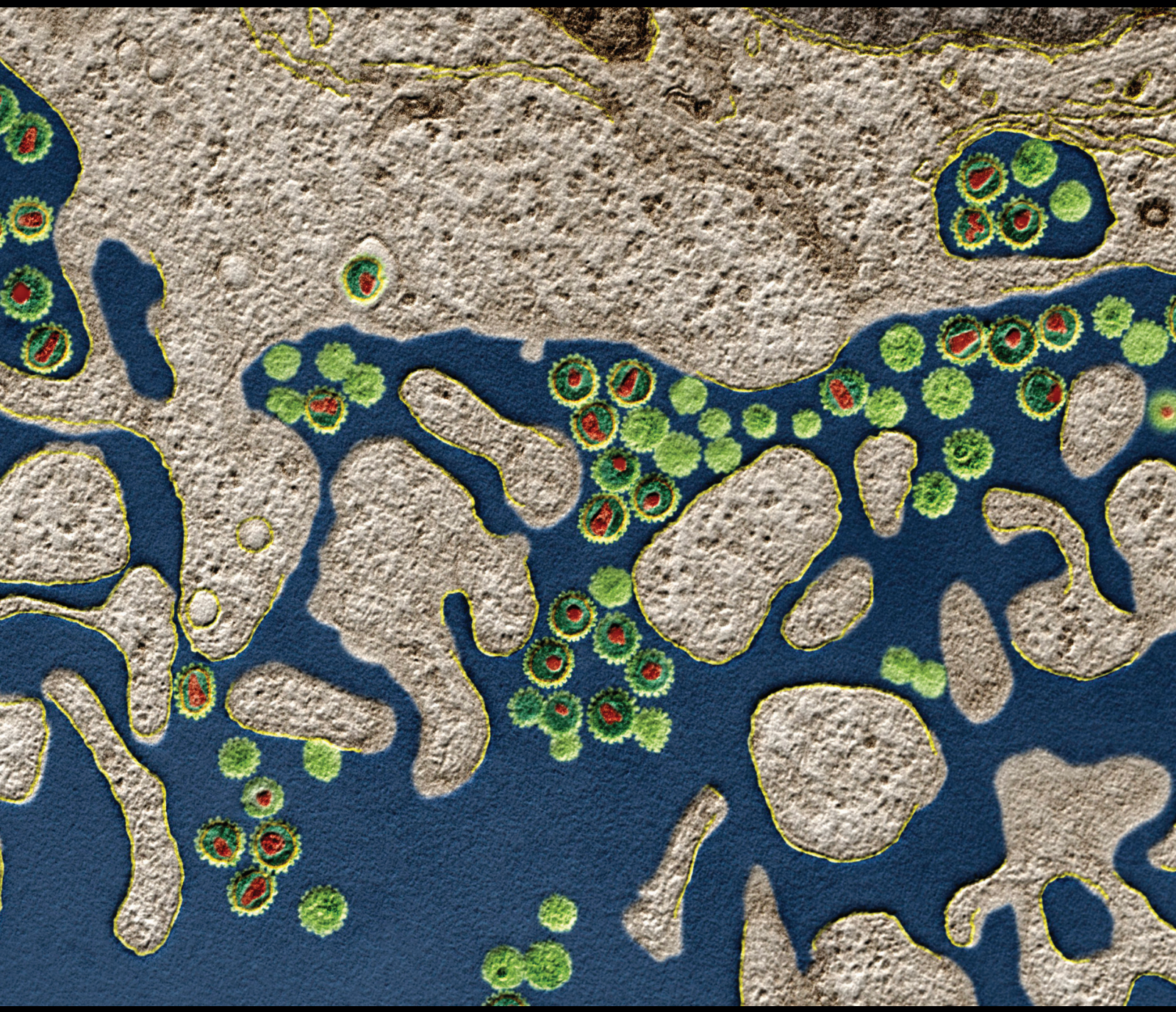
Journal of Immunology Research

Advances in Computational Immunology

Guest Editors: Francesco Pappalardo, Vladimir Brusic, Marzio Pennisi, and Guanglan Zhang



Advances in Computational Immunology

Journal of Immunology Research

Advances in Computational Immunology

Guest Editors: Francesco Pappalardo, Vladimir Brusic, Marzio Pennisi, and Guanglan Zhang

Copyright © 2015 Hindawi Publishing Corporation. All rights reserved.

This is a special issue published in "Journal of Immunology Research." All articles are open access articles distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Editorial Board

Bartholomew D. Akanmori, Congo
Stuart Berzins, Australia
Kurt Blaser, Switzerland
Federico Bussolino, Italy
Nitya G. Chakraborty, USA
Robert B. Clark, USA
Mario Clerici, Italy
Nathalie Cools, Belgium
Mark J. Dobrzanski, USA
Nejat K. Egilmez, USA
Eyad Elkord, United Kingdom
Steven E. Finkelstein, USA
Luca Gattinoni, USA
David E. Gilham, UK
Douglas C. Hooper, USA
Eung-Jun Im, USA
Hidetoshi Inoko, Japan
Peirong Jiao, China
Taro Kawai, Japan
Hiroshi Kiyono, Japan
Shigeo Koido, Japan
Herbert K. Lyerly, USA
Enrico Maggi, Italy
Mahboobeh Mahdavinia, USA
Eiji Matsuura, Japan
C. J. M. Melief, The Netherlands
Chikao Morimoto, Japan
Hiroshi Nakajima, Japan
Toshinori Nakayama, Japan
Paola Nistico, Italy
Ghislain Opdenakker, Belgium
Clelia M. Riera, Argentina
Luigina Romani, Italy
Aurelia Rughetti, Italy
Takami Sato, USA
Senthamil Selvan, USA
Naohiro Seo, Japan
Ethan M. Shevach, USA
George B. Stefano, USA
Trina J. Stewart, Australia
Jacek Tabarkiewicz, Poland
Ban-Hock Toh, Australia
Joseph F. Urban, USA
Xiao-Feng Yang, USA
Qiang Zhang, USA

Contents

Hindawi Publishing Corporation
Journal of Immunology Research
Volume 2015, Article ID 170920, 3 pages
http://dx.doi.org/10.1155/2015/170920

Editorial

Advances in Computational Immunology

Francesco Pappalardo,[1] Vladimir Brusic,[2] Marzio Pennisi,[3] and Guanglan Zhang[4]

[1]*Department of Drug Sciences, University of Catania, 95100 Catania, Italy*
[2]*Nazarbayev University, Astana 010000, Kazakhstan*
[3]*Department of Mathematics and Computer Science, University of Catania, 95100 Catania, Italy*
[4]*Boston University, Boston, MA 02201-2021, USA*

Correspondence should be addressed to Francesco Pappalardo; francesco.pappalardo@unict.it

Received 13 December 2015; Accepted 13 December 2015

Copyright © 2015 Francesco Pappalardo et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Computational immunology and immunological bioinformatics are firm and quickly growing research fields. Whereas the former aims to develop mathematical and/or computational methods to study the dynamics of cellular and molecular entities during the immune response [1–4], the latter focuses on proposing methods to investigate big genomic and proteomic immunological-related datasets and predict new knowledge mainly by statistical inference and machine learning algorithms.

The glut of data produced by high-throughput instrumentation, notably genomics, transcriptomics, epigenetics, and proteomics methods, requires computational tools for acquisition, storage, and analysis of immunological data.

The exploitation of such a huge amount of immunological data usually requires its conversion into computational problems, their solution using mathematical and computational approaches, and then the translation of the obtained results into immunologically meaningful interpretations.

In this special issue, we take an interest from mathematicians, bioinformaticians, computational scientists, and engineers together with experimental immunologists to present and discuss latest developments in different subareas of computational immunology, ranging from databases applications to computational vaccine design, modelling, and simulation and their application to basic and clinical immunology.

The review from N. Sepúlveda et al. calls attention to serology data in conjunction with mathematical modelling in providing a powerful approach to inform on malaria transmission intensity and putative changes over time. Their conclusions show that an interesting idea with public health potential is to use a panel of multidisease antibodies that can be instrumental to know what the infectious agents are in circulation in a given population and their putative dynamics. This idea has not been tested in practice, but definitely will require the extension of classical mathematical models to fully account the immunological interaction between different diseases.

In their paper W. Schreiner and colleagues illustrate that molecular dynamics was used to simulate large molecules of the immune system (major histocompatibility complexes, T-cell receptors, and coreceptors). To characterize the relative orientation and movements of domains local coordinate systems (based on principal component analysis) were generated and directional cosines and Euler angles computed. As a most interesting result, they found that the presence of the coreceptor seems to influence the dynamics within the protein complex, in particular the relative movements of the two α-helices, Gα1 and Gα2.

It is assessed that the application of personalized medicine requires integration of different data to determine each patient's unique clinical constitution. The automated analysis of medical data is a growing field where different machine learning techniques are used to minimize the time consuming task of manual analysis.

In the paper contributed by C.-M. Svensson et al., the authors investigate the interobserver variability of image data comprising fluorescently stained circulating tumor cells and its effect on the performance of two automated classifiers, a random forest and a support vector machine. They found that uncertainty in annotation between observers limited

the performance of the automated classifiers, especially when it was included in the test set on which classifier performance was measured.

Therapeutic protein products (TPP) have been widely used to treat a variety of human diseases, including cancer, hemophilia, and autoimmune diseases. However, TPP can induce unwanted immune responses that can impact both drug efficacy and patient safety. The presence of aggregates is of particular concern as they have been implicated in inducing both T-cell independent and T-cell dependent immune responses. L. Yin and collaborators used mathematical modelling to evaluate several mechanisms through which aggregates of TPP could contribute to the development of immunogenicity. Their computational analyses suggest that aggregates are unlikely to induce T-cell independent antibody responses through BCR cross-linking. In contrast, aggregates could contribute to immunogenicity via the T-cell dependent pathway by inducing the presentation of high affinity epitopes that may not be present in nonaggregated TPP and/or by enhancing danger signal to maturate dendritic cells and activate T-cells.

A. K. Irin et al. investigate computational modelling approaches on epigenetic factors in neurodegenerative and autoimmune diseases and their mechanistic analysis. The authors examine the major milestones in epigenetics research in the context of diseases and various computational approaches developed in the last decades to unravel new epigenetic modifications. However, there are limited studies that systematically link genetic and epigenetic alterations of DNA to the aetiology of diseases, they said. In this work, A. K. Irin and coauthors show how disease-related epigenetic knowledge can be systematically captured and integrated with heterogeneous information into a functional context using Biological Expression Language (BEL). This novel methodology, based on BEL, enables the integration of epigenetic modifications such as DNA methylation or acetylation of histones into a specific disease network.

In the paper by G. Bocharov et al., the authors show how the modelling approaches can be implemented to address diverse aspects of immune system functioning under normal conditions and in response to LCMV and, importantly, make quantitative predictions of the outcomes of immune system perturbations. This may highlight that data-driven applications of meaningful mathematical models in infection biology remain a challenge.

MHC α-helices form the antigen-binding cleft and are of particular interest for immunological reactions. To monitor these helices in molecular dynamics simulations, the paper contributed by R. Ribarics et al. applied a parsimonious fragment-fitting method to trace the axes of the α-helices. Each resulting axis was fitted by polynomials in a least-squares sense and the curvature integral was computed. To find the appropriate polynomial degree, the method was tested on two artificially modelled helices, one performing a bending and another a hinge movement. They found that second-order polynomials retrieve predefined parameters of helical motion with minimal relative error.

There are at present few tools available to assist with the determination and analysis of B-cell lineage trees from next-generation sequencing data. The paper from W. D Lees and A. J. Shepherd presents two utilities that support automated large-scale analysis and the creation of publication-quality results. The tools are available on the web and are also available for download so that they can be integrated into an automated pipeline. These utilities can be used with any suitable phylogenetic inference method and with any antibody germline library and hence are species-independent.

Vaccination is historically one of the most important medical interventions for the prevention of infectious disease. Previously, vaccines were typically made of rather crude mixtures of inactivated or attenuated causative agents. However, over the last 10–20 years, several important technological and computational advances have enabled major progress in the discovery and design of potently immunogenic recombinant protein vaccine antigens. L. Liljeroos and colleagues discuss three key breakthrough approaches that have potentiated structural and computational vaccine design. They illustrate the growing power of combining sequencing, structural, and computational approaches and discuss how this may drive the design of novel immunogens suitable for future vaccines urgently needed to increase the global prevention of infectious disease.

MIrExpress is a new database which takes advantage of the information theory, as well as the Pearson linear correlation method, to measure the linear correlation, nonlinear correlation, and their hybrid of cell-specific gene coexpressions in immune cells. In the work from J. Wang et al., the authors describe this database that totally includes 16 human cell groups, involving 20,283 human genes. The expression data and the calculated correlation results from the database are interactively accessible on the web page and can be implemented for other related applications and researches.

Publically available influenza data are a valuable resource for computational analyses with applications in vaccine design. Similarly, existing bioinformatics tools provide the means for extraction of information and new knowledge. However, to utilize the full potential of these resources, data preprocessing must be performed and analytical tools must be carefully combined into well-defined workflows. C. Simon et al. describe FluKB, a knowledge-based system focusing on data and analytical tools for influenza vaccine discovery. The main goal of FluKB is to provide access to curated influenza sequence and epitope data and enhance the analysis of influenza sequence diversity and the analysis of targets of immune responses. FluKB consists of more than 400,000 influenza protein sequences, known epitope data (357 verified T-cell epitopes, 685 HLA binders, and 16 naturally processed MHC ligands), and a collection of 28 influenza antibodies and their structurally defined B-cell epitopes.

Francesco Pappalardo
Vladimir Brusic
Marzio Pennisi
Guanglan Zhang

References

[1] S. Motta and F. Pappalardo, "Mathematical modeling of biological systems," *Briefings in Bioinformatics*, vol. 14, no. 4, Article ID bbs061, pp. 411–422, 2013.

[2] F. Pappalardo, A. Palladini, M. Pennisi, F. Castiglione, and S. Motta, "Mathematical and computational models in tumor immunology," *Mathematical Modelling of Natural Phenomena*, vol. 7, no. 3, pp. 186–203, 2012.

[3] F. Pappalardo, P. Zhang, M. Halling-Brown et al., "Computational simulations of the immune system for personalized medicine: state of the art and challenges," *Current Pharmacogenomics and Personalized Medicine*, vol. 6, no. 4, pp. 260–271, 2008.

[4] F. Pappalardo, D. Flower, G. Russo, M. Pennisi, and S. Motta, "Computational modelling approaches to vaccinology," *Pharmacological Research*, vol. 92, pp. 40–45, 2015.

Hindawi Publishing Corporation
Journal of Immunology Research
Volume 2015, Article ID 210675, 10 pages
http://dx.doi.org/10.1155/2015/210675

Research Article

Relative Movements of Domains in Large Molecules of the Immune System

Wolfgang Schreiner,[1] Rudolf Karch,[1] Reiner Ribarics,[1] Michael Cibena,[1] and Nevena Ilieva[2]

[1]*Section of Biosimulation and Bioinformatics, Center for Medical Statistics, Informatics and Intelligent Systems (CeMSIIS), Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria*
[2]*Institute of Information and Communication Technologies (IICT), Bulgarian Academy of Sciences, 25 A, Acad. G. Bonchev Street, 1113 Sofia, Bulgaria*

Correspondence should be addressed to Wolfgang Schreiner; wolfgang.schreiner@meduniwien.ac.at

Received 28 August 2015; Accepted 26 November 2015

Academic Editor: Francesco Pappalardo

Copyright © 2015 Wolfgang Schreiner et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Molecular dynamics was used to simulate large molecules of the immune system (major histocompatibility complex class I, presented epitope, T-cell receptor, and a CD8 coreceptor.) To characterize the relative orientation and movements of domains local coordinate systems (based on principal component analysis) were generated and directional cosines and Euler angles were computed. As a most interesting result, we found that the presence of the coreceptor seems to influence the dynamics within the protein complex, in particular the relative movements of the two α-helices, Gα_1 and Gα_2.

1. Introduction

The interaction between major histocompatibility complexes (MHCs) and T-cell receptors (TCRs) plays a key role in triggering adaptive immune responses. TCRs bind the highly polymorphic MHC proteins that present peptide fragments (p) derived from the host proteome, pathogens, or tumour antigens on the cell surface. TCR and pMHC represent the core of the immunological synapse that in turn comprises many proteins, both membrane-bound and in the cytosol that could relay signals and/or act as an adjustable screw to fine-tune TCR sensitivity. Cluster of differentiation 8 (CD8) is a transmembrane, mostly heterodimeric glycoprotein that functions as a coreceptor for the TCR. It is mainly expressed by cytotoxic T-cells (T_C), but is also found on natural killer cells, cortical thymocytes, and dendritic cells [1]. The extracellular domain of CD8 binds to the α_3-domain of the MHC class I heavy chain [2]. It is well-known that CD8 and CD4 coreceptors are able to enhance T-cell responses to antigen stimulation [3–5]. Also, when subjected to an immune response, $CD8^+$ T-cells can substantially increase in sensitivity by the mechanism of functional avidity maturation, that is, maturation of strength of multivalent antigen-antibody binding [6–8].

According to the literature, the major mechanism for stabilizing TCR-pMHC interaction by CD8 is the CD8-MHC interaction that increases the TCR-pMHC rebinding probability [9, 10]. A less obvious mechanism is stated by Borger et al. [11] and includes affected binding rates of TCR-pMHC. They propose a two-stage reversible reaction mechanism of pMHC with either TCR or CD8, similar to the mechanism found by Liu et al. [12].

Molecular dynamics (MD) simulations of TCR/pMHC (PDB ID: 3KPS) and TCR/pMHC (3KPS) plus CD8$\alpha\alpha$ homodimer were performed. The topology of the TCR/pMHC complex with and without a CD8 coreceptor is shown in Table 1 and Figure 1. We chose to monitor the relative movements of MHC α-helices, Gα_1 and Gα_2, with and without the presence of CD8 (as these helices constitute part of the binding cleft for the peptide) and of the MHC α_3-domain relative to the whole CD8 (since the α_3-domain is the binding site for the CD8-coreceptor).

Table 1: Molecules and their secondary structural elements.

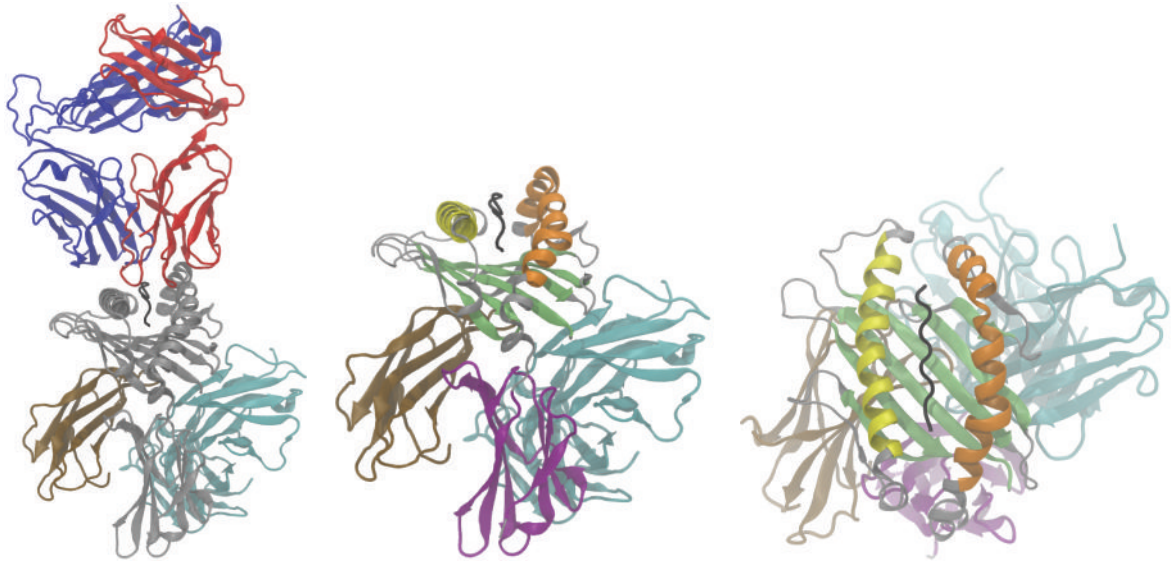
(a)

Chain	Type	Length in C_α	C_α index
Chain A	MHC	276	1–276
Chain B	β_2-microglobulin	99	277–375
Chain C	Peptide	9	376–384
Chain D	TCR, α-chain	201	385–585
Chain E	TCR, β-chain	241	586–826
Chain F	CD8 α_1	114	827–940
Chain G	CD8 α_2	114	941–1054

(b)

Secondary structures	Chain	Length in C_α	C_α index
α-helix Gα_1	A	25	59–83
α-helix Gα_2	A	31	141–170
α_3-domain	A	92	184–275
β-sheet	A	52	2–13, 21–29, 30–37, 93–103, 110–118, 124–127
TCR α_{var}	D	104	385–488
TCR β_{var}	E	117	586–702

Structural elements are given in terms of consecutive numbers of C_α atoms, renumbered throughout the whole modelled TCR/pMHC/CD8 complex, as if the complex as a whole was taken from one single PDB file.
Molecule 1 (TCR/pMHC): Chain A–Chain E. Molecule 2 (TCR/pMHC/CD8): Chain A–Chain G.



(a) (b) (c)

Figure 1: Structure description. (a) Cartoon representation of the TCR/pMHC/CD8 system: MHC (grey), β_2-microglobulin (ochre), peptide (black), TCR α-chain (red), TCR β-chain (blue), and CD8 α_1 and CD8 α_2 (cyan). (b, c) Cartoon representation of the pMHC/CD8 complex: α-helix Gα_1 (yellow), α-helix Gα_2 (orange), α_3-domain (purple), β-sheet (lime), and CD8 α_1 and CD8 α_2 (cyan).

2. Methods

2.1. Molecular Modelling. The structure of LC13 TCR, ABCD3 peptide, and MHC (TCR/pMHC) of HLA-B*44:05 type has been resolved (PDB-ID: 3KPS, www.pdb.org). Also, molecular structures of CD8 coreceptors bound to MHCs are available (PDB-ID: 1AKJ).

To our knowledge, a TCR/pMHC/CD8 complex has not yet been cocrystallized. To model the TCR/pMHC/CD8 complex we localized the CD8/MHC binding site in the 1AKJ crystal structure by finding all C_α atoms of the MHC within the range of 0.8 nm to CD8. Structures of TCR/pMHC (3KPS) and MHC/CD8 (1AKJ) were merged into one file, both MHC binding sites superimposed so as to minimize

RMSD in a least-squares sense and the MHC molecule from the MHC/CD8 complex deleted to get the TCR/pMHC/CD8 complex; see Figure 1.

2.2. Molecular Dynamics Simulations. Molecular dynamics simulations were performed with GROMACS 4.0.7 [13] using the gromos53a6 force field. The whole system counts about 274000 atoms, including the solvent (protein atoms only: about 8500), within a simulation box sized $13 \times 13.5 \times 16.5\,\mathrm{nm}^3$, to ensure 2 nm minimal distance between the protein atoms and the box walls, and with periodic boundary conditions imposed. The solvent was described with the SPC water model [14], the system neutralized at a salt concentration of 0.15 mol/L, and its energy was minimized by the steepest-descent method. The temperature was then gradually increased to 310 K within a 100 ps position-restraint simulation. Temperature was controlled by a Berendsen-thermostat with a time constant of 0.1 ps and the pressure controlled by a Berendsen-barostat set to 1 bar with a time constant of 0.5 ps, both chosen for being the most efficient in the beginning of the simulation. Constraints on all bonds were imposed with the LINCS algorithm [15], and the particle mesh Ewald (PME) method [16] was used to compute the long-range electrostatic interactions, with van der Waals and Coulomb cutoff radii of 1.4 nm. For the MD simulation runs of 200 ns with a time step of 5 fs, enabled by using virtual sites for hydrogen atoms, the thermostat was set to v-rescale, with the same time constants in order to guarantee the generation of a proper canonical ensemble [13]. Coordinates were written every 50 ps, giving thus rise to 4000 frames. Prior to the analysis of domain movements, translational and rotational motions relative to the energy-minimized protein structure were removed.

2.3. Relative Location of Domains. Within the biomolecules we consider the C_α atoms in the backbones of protein chains. These C_α atoms are addressed via their indices to define domains or even subdomains; see Table 1. We define a first domain, V, by enumerating the C_α atoms contained; for example, to select the α-helix Gα_1 we have $V = \{59, 60, \ldots, 83\}$. Similarly, we define a second domain, W. Note that domains may consist of several parts such as the β-sheet; see Table 1.

Considering a domain V containing N_V C_α atoms with Cartesian coordinates $\mathbf{x}_i(f)$ in MD-frame f, the coordinates of its geometrical center are given by

$$\bar{\mathbf{x}}_V(f) = \frac{1}{N_V}\sum_{i\in V}\mathbf{x}_i(f). \tag{1}$$

The distance between the centers of two domains, V and W, in MD-frame f is then

$$d(f) = \left\|\bar{\mathbf{x}}_W(f) - \bar{\mathbf{x}}_V(f)\right\|. \tag{2}$$

Both domains are shifted with their mean C_α coordinates into the origin before defining their relative orientation.

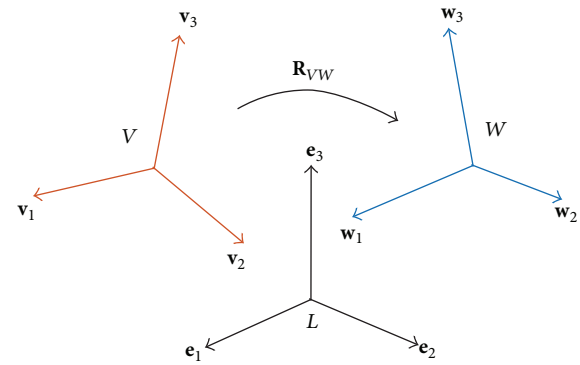


Figure 2: Relative orientation of two submolecular domains. Orthonormal eigenvectors for domains V and W and standard basis for the laboratory system L. Rotation matrix $\mathbf{R}_{VW}$ transforms eigenvectors from domain V into domain W. Note that eigenvectors share the same coordinate system origin. For better visualization, eigenvectors are displayed as if they had different origins.

2.3.1. Rigid Axes within Deformable Domains. To quantify the relative orientation of two domains one has to bear in mind that each domain is deformed from time step to time step of an MD simulation and hence no unique frame of reference can easily be assigned as is possible for a rigid body. Often a principal component analysis (PCA) is performed to obtain three main characteristic axes of a given domain [17–19]. However, principal components are not fully rigorously defined: for example, the orientation of the first PC-eigenvector may swap by nearly 180° for virtually the identical coordinates, just because of numerical noise. Similarly, the second and third PC-eigenvectors may interchange roles from time to time for an atomic domain almost cylindrical in shape.

In order to avoid such artifacts we refrained from computing principal components repeatedly for each MD time step and adopted the following procedure, see Figure 2:

(1) For domain V we select a reference frame, k_V, from the whole trajectory. This is done by computing the sum of RMSD-displacements of V relative to itself [20] over all frames of the trajectory and adopting the frame k_V with minimum sum of RMSD. This frame is in a geometrical sense considered most "central" for domain V within the whole trajectory.

(2) The C_α coordinates of domain V at frame k_V are subjected to a PCA, yielding the orthogonal matrix $\mathbf{T}_V(k_V) = [\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3]_{k_V}$ of three orthonormal eigenvectors $\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ of the covariance matrix of the coordinates of C_α atoms in domain V at frame k_V. These vectors define a reference frame (i.e., a local coordinate system) for domain V.

(3) Steps (1) and (2) are performed also for domain W, yielding the respective central frame k_W with an eigenvector matrix $\mathbf{T}_W(k_W) = [\mathbf{w}_1\ \mathbf{w}_2\ \mathbf{w}_3]_{k_W}$ and a local coordinate system for W.

2.3.2. Computing Robust Relative Orientations. Given all MD-frames f of a trajectory, the relative orientation of domains V and W is computed as follows:

(a) For each frame f we compute the transformation of all coordinates $\mathbf{x}_V(k_V)$ of domain V from its position within the central frame into its position at frame f according to minimum RMSD using Kabsch's method [20].

(b) The rotational part $\mathbf{R}_V(f)$ of the above transformation is applied to the eigenvectors of domain V at frame k_V (the reference frame) to obtain the position of the eigenvectors of V at frame f: $\mathbf{T}_V(f) = [\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3]_f$.

(c) Steps (a) and (b) are performed also for domain W, yielding $\mathbf{R}_W(f)$ and transformed eigenvectors $\mathbf{T}_W(f) = [\mathbf{w}_1\ \mathbf{w}_2\ \mathbf{w}_3]_f$ for each frame f.

(d) For each frame f we note that the directional relation between the two sets of eigenvectors $[\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3]_f$ and $[\mathbf{w}_1\ \mathbf{w}_2\ \mathbf{w}_3]_f$ can be represented via a rotation matrix $\mathbf{R}_{VW}(f)$ as

$$\mathbf{T}_W(f) = \mathbf{R}_{VW}(f) \cdot \mathbf{T}_V(f) \tag{3}$$

which also characterizes the relative orientation of both domains. From (3) we obtain

$$\mathbf{R}_{VW}(f) = \mathbf{T}_W(f) \cdot \mathbf{T}_V^{-1}(f) = \mathbf{T}_W(f) \cdot \mathbf{T}_V^{T}(f) \tag{4}$$

since the inverse of an orthogonal matrix equals its transpose. Rewriting the matrix $\mathbf{R}_{VW}(f)$ in terms of its column vectors

$$\mathbf{R}_{VW}(f) = \left[\begin{pmatrix} r_{1x} \\ r_{1y} \\ r_{1z} \end{pmatrix} \begin{pmatrix} r_{2x} \\ r_{2y} \\ r_{2z} \end{pmatrix} \begin{pmatrix} r_{3x} \\ r_{3y} \\ r_{3z} \end{pmatrix} \right] \tag{5}$$

the Euler angles in x-convention [21] between the two sets of the orthonormal eigenvectors $[\mathbf{v}_1\ \mathbf{v}_2\ \mathbf{v}_3]_f$ and $[\mathbf{w}_1\ \mathbf{w}_2\ \mathbf{w}_3]_f$ can be read as

$$\begin{aligned} \alpha &= \arccos\left(\frac{r_{3y}}{\sqrt{1 - r_{3z}^2}} \right), \\ \beta &= \arccos\left(r_{3z}\right), \\ \gamma &= \arccos\left(\frac{r_{2y}}{\sqrt{1 - r_{3z}^2}} \right) \end{aligned} \tag{6}$$

with all quantities depending on frame f (dependencies suppressed in the notation).

As a result, $d(f)$, $\alpha(f)$, $\beta(f)$, and $\gamma(f)$ define the relative spatial relation between domains V and W for MD frame f.

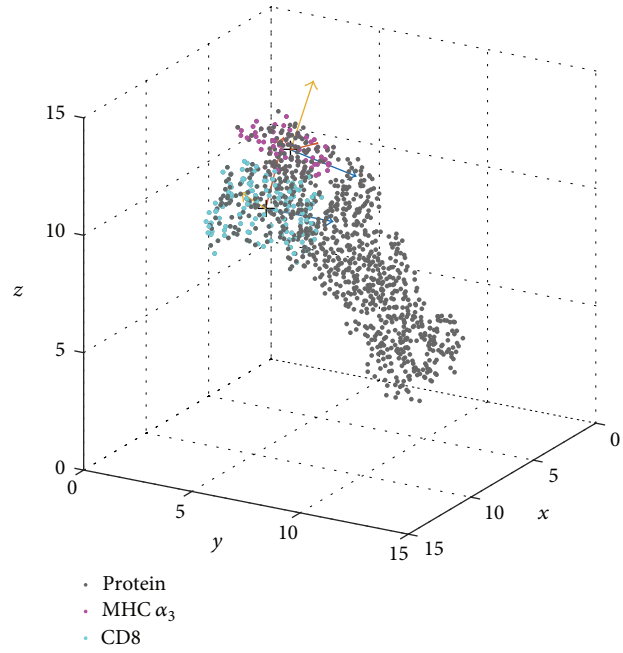


Figure 3: Domains CD8 and MHC α_3 with local eigenvectors. Domain CD8 is shown in cyan, α_3 in purple, and the remaining parts of the MHC as well as the TCR (labelled "protein") in black. Eigenvectors ($\mathbf{v}_1$, $\mathbf{v}_2$, $\mathbf{v}_3$ and $\mathbf{w}_1$, $\mathbf{w}_2$, $\mathbf{w}_3$, resp.) are shown for the first frame of the trajectory and colored ($\mathbf{v}_1$, $\mathbf{w}_1$: blue, $\mathbf{v}_2$, $\mathbf{w}_2$: red, $\mathbf{v}_3$, $\mathbf{w}_3$: yellow) for each of the domains.

3. Results

3.1. Relative Movements CD8-MHC. As a first pair of domains we considered the CD8 homodimer (CD8 α_1, CD8 α_2) as domain V and domain α_3 of the MHC (see Table 1) as domain W. Domain α_3 is the binding site for CD8 (see Figure 3) and therefore most interesting regarding relative movements.

3.1.1. Relative Distance. The distance d between domains, (2), was computed over the time; see Figure 4. This distance, initially around 2.75 nm (as modelled, based on the crystallographic structure), gradually decreases to about 2.55 nm during the first 50 ns of the simulation. Apparently, dynamics lets CD8 get somewhat closer to the TCR/pMHC complex.

3.1.2. Relative Orientation. For each domain, the sum of RMSD of each frame to all other frames of the trajectory was computed to determine the central (reference) frames, k_V and k_W; see Figure 5.

Relative positions of both domains were then computed for each frame (4000 all in all) as outlined above leading to the following results; see Section 2.3.2.

A first and direct measure is provided by the orientation cosines between corresponding eigenvectors; see Figure 6(a). They clearly reflect an initial phase different from the remaining trajectory, corresponding to the phase of decreasing interdomain distance, already seen in Figure 4. Interestingly, this effect is not at all revealed by the Euler angles; see Figure 6(b).

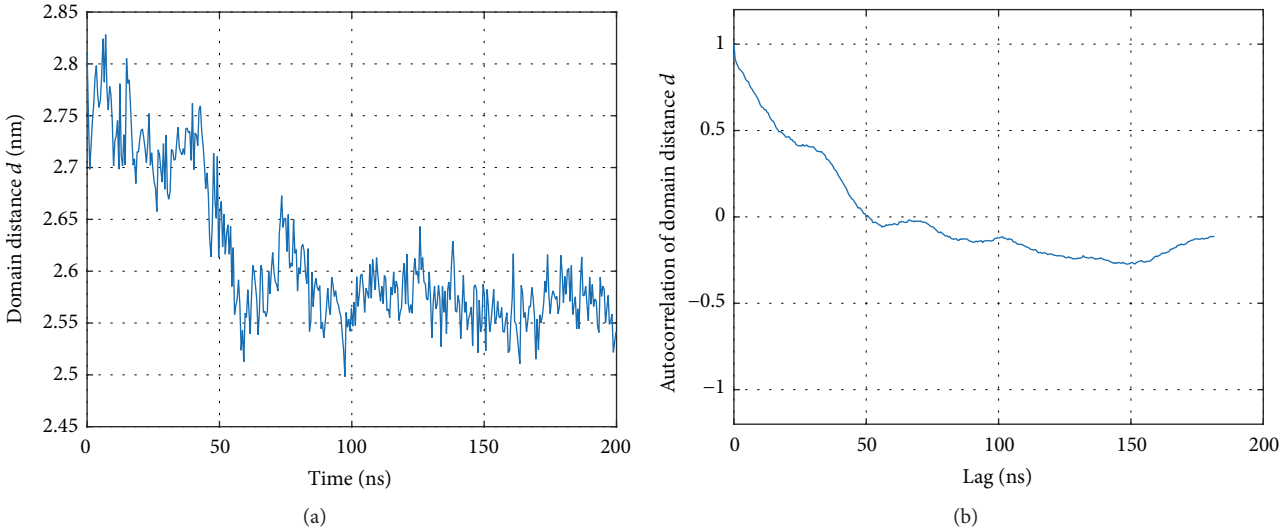


FIGURE 4: Distance between CD8 and MHC α_3. (a) Distances computed according to (2). (b) Autocorrelation of distance values.

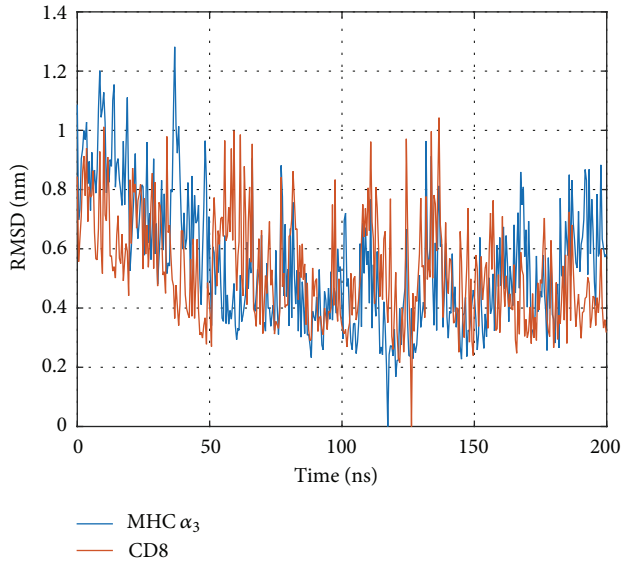


FIGURE 5: RMSD from central frame of each domain to all other frames of trajectory. Note that RMSD is zero for the respective reference frames against themselves by definition ($k_V = 237 \equiv 117.30$ ns and $k_W = 255 \equiv 126.25$ ns).

To further characterize the evolution of the geometry of the complex with time we computed the autocorrelation functions (ACFs) for cosines between eigenvectors and for Euler angles; see Figure 7(b). Euler angles exhibited relatively short autocorrelations, passing through zero already at approximately 20 ns, and oscillating around zero for larger time lags.

A completely different picture results from inspecting the ACFs of the cosines between corresponding eigenvectors; see Figure 7(a). While the relative orientation of the eigenvectors corresponding to the largest eigenvalues, $\mathbf{v}_1$ and $\mathbf{w}_1$, has only a short memory (the ACF passes through zero around 20 ns), a massively prolonged memory is seen for both smaller eigenvectors $\mathbf{v}_2$ versus $\mathbf{w}_2$ and $\mathbf{v}_3$ versus $\mathbf{w}_3$: they exhibit a long negative tail and do not become stochastic throughout the whole simulation time. This reflects the fact already seen in the time course itself (Figure 6(a)): a long equilibration phase extends up to about 100 ns (which amounts to half the simulation time).

3.2. Relative Movements of Two MHC α-Helices. The two α-helices of the MHC, Gα_1 and Gα_2, together with a β-floor form the binding cleft for the peptide. To evaluate their relative motion we chose the helices as domains V and W, located the central (reference) frames ($k_V = 279 \equiv 138.4$ ns and $k_W = 345 \equiv 171.2$ ns) of each, and computed corresponding eigenvectors; see Figure 8(b). It is interesting to see that the first eigenvectors $\mathbf{v}_1(k_V)$ and $\mathbf{w}_1(k_V)$ have similar directions and orientations, when computed for the central frames. As opposed to this, they show almost opposite directions when computed from the first frame of the trajectory; see Figure 8(a). These opposite directions of eigenvectors are formally obtained from the very same matrix algebra, although the domains themselves have by no means turned upside down, as one can see by simple inspection. We have addressed this issue in Section 2 and display an example here. Our method of fitting reference domains via the Kabsch method has been designed to avoid these effects. In fact, the eigenvectors $\mathbf{v}_1(1)$ and $\mathbf{w}_1(1)$ we actually use for frame 1 (and for all other frames) have orientations similar to $\mathbf{v}_1(k_V)$ and $\mathbf{w}_1(k_V)$, except for the actual, relative movements of both domains.

Relative motions of Gα_1 and Gα_2 are characterized by cosines (see Figure 9(a)) and Euler angles (Figure 9(b)) between eigenvectors attached to each domain. Cosines as well as Euler angles reveal a correlation in movements of eigenvectors 2 and 3. These eigenvectors point away from the axis of the helices at right angles, and their correlated changes indicate a synchronized oscillating "rolling" of both helices.

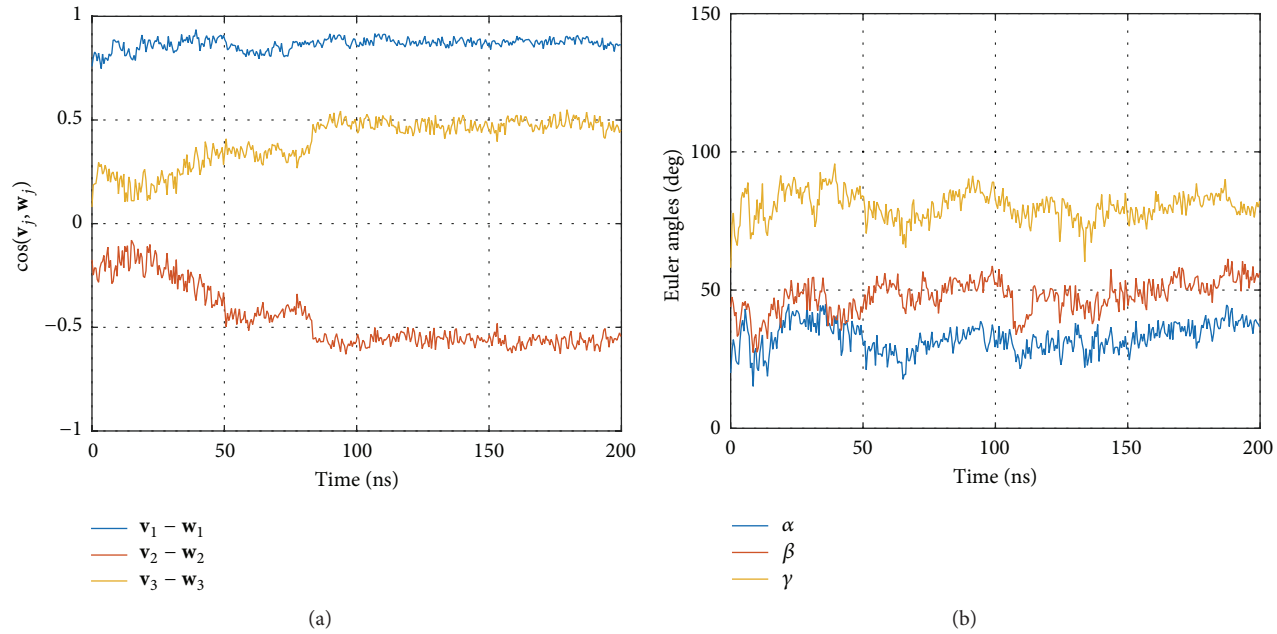


FIGURE 6: Relative Orientation of CD8 and MHC α_3. (a) Cosine values between corresponding eigenvectors (see legend box) of both domains over time. Every 10th frame is plotted. (b) Euler angles for the same data.

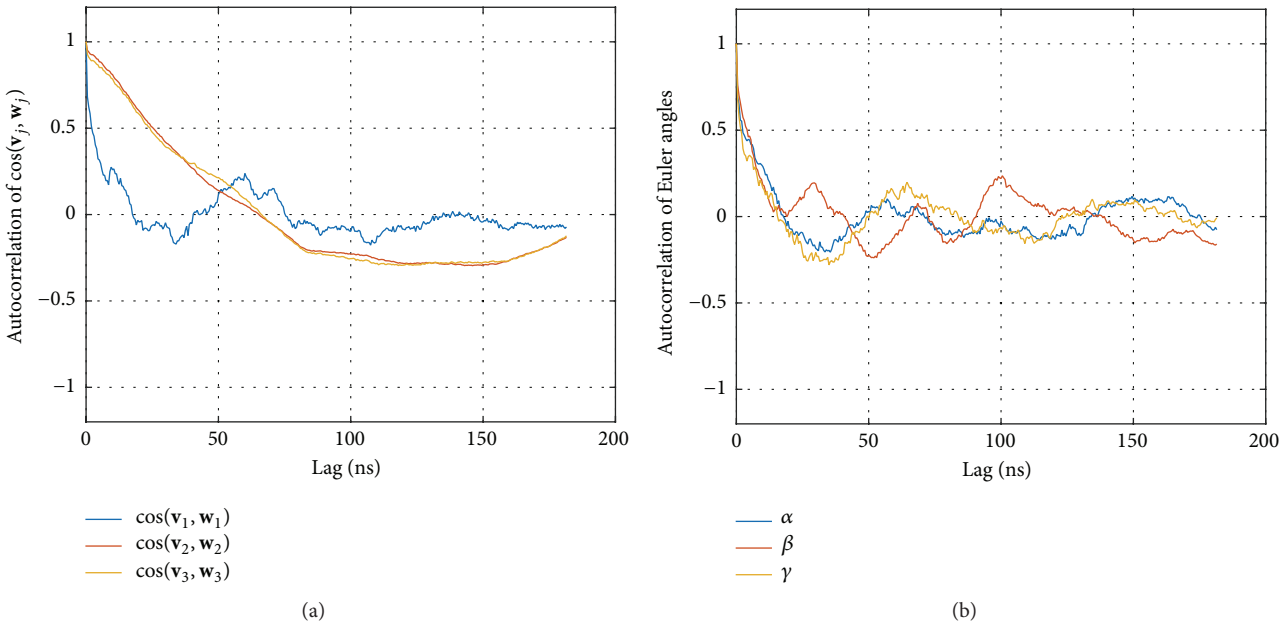


FIGURE 7: Autocorrelation of relative directions of CD8 and MHC α_3. (a) Autocorrelation of cosine values between corresponding eigenvectors (see legend box) of both domains over time. Every 10th frame is plotted. (b) Autocorrelation of Euler angles for the same data.

In fact this kind of movement is also evident when visually inspecting the trajectories in VMD [22].

The above finding is nicely quantified via the correlation coefficient $\rho = 0.79$ (with $p < 0.01$ and $N_{\text{frames}} = 503$); see Figure 10(a).

3.3. Impact of CD8 Presence. We have also analyzed the relative motions of Gα_1 and Gα_2 from our MD simulation of TCR/pMHC/CD8 with CD8 attached to the MHC and compared it with the above results without CD8. Interestingly, the relative motions do not differ significantly (figures not shown); however, the correlation of "rolling" oscillations is almost lost in the presence of CD8: correlation coefficient $\rho = 0.6$ (with $p < 0.01$ and $N_{\text{frames}} = 406$); see Figure 10(b). To check if this difference in correlation coefficients is statistically significant, we computed the 95%-confidence intervals [23], resulting in $[0.755, 0.821]$ for $\rho = 0.79$

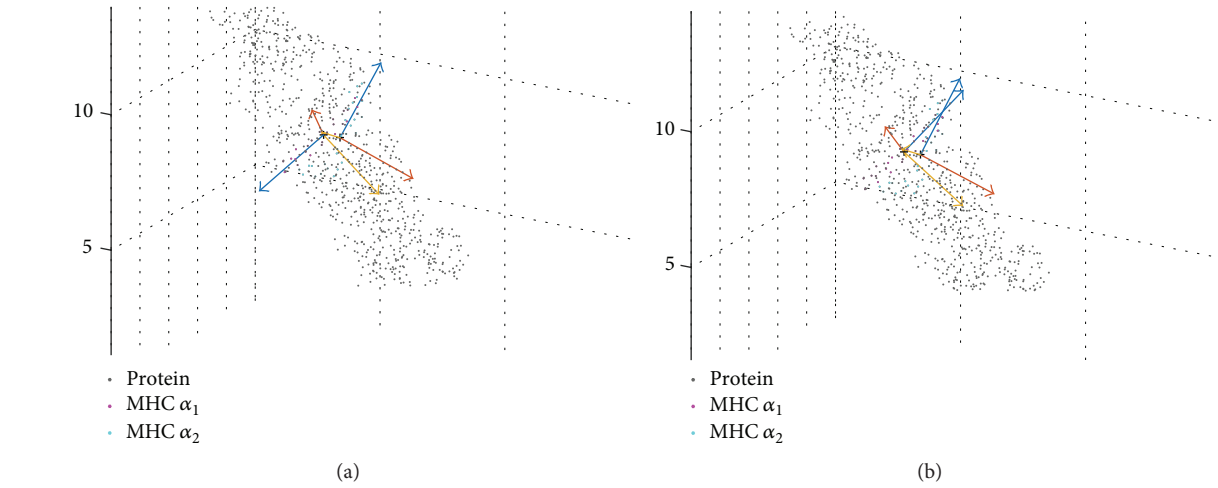


FIGURE 8: Eigenvectors switching orientations between frames. (a) Eigenvectors of first frames. (b) Eigenvectors of central frames k_V and k_W. If calculated from the atomic coordinates, the first eigenvectors ($\mathbf{v}_1$, $\mathbf{w}_1$: blue) result with almost opposite orientations in the first frame but very similar orientations in the central frames of each domain. Note that eigenvectors refer to different frames k_V and k_W, respectively, but atoms of the molecule are plotted only once. Coloring: $\mathbf{v}_1$, $\mathbf{w}_1$: blue; $\mathbf{v}_2$, $\mathbf{w}_2$: red, and $\mathbf{v}_3$, $\mathbf{w}_3$: yellow.

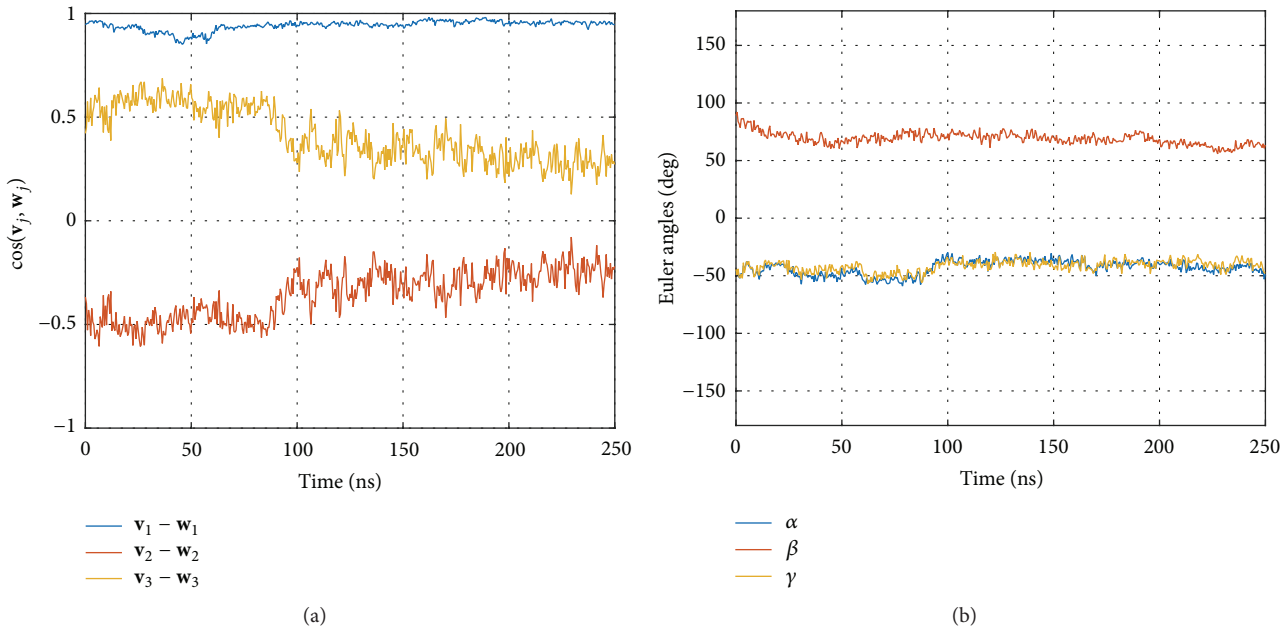


FIGURE 9: Relative movements of helices Gα_1 and Gα_2 in absence of CD8. (a) Cosines between corresponding eigenvectors. (b) Euler angles between corresponding eigenvectors.

($N_{\text{frames}} = 503$) and $[0.553, 0.659]$ for $\rho = 0.6$ ($N_{\text{frames}} = 406$). These intervals do not overlap and the difference in correlation coefficients may thus be considered statistically significant.

4. Discussion

The methodological parts of this work describe a new computational technique to obtain relative orientations of intramolecular domains. In the application parts this method is used to analyze the molecular dynamics of two systems, TCR/pMHC and TDC/pMHC/CD8, respectively.

4.1. Methods to Characterize Relative Orientations. There is a plethora of ways to characterize relative movements of intramolecular domains (e.g., [17–19, 24–27]). The most direct one is to compute average distances between groups of atoms, as they change over time. We did this by computing the distance between CD8 and MHC α_3; see Figure 4. By appropriate selection of target groups, some basic information can be obtained also on relative orientations.

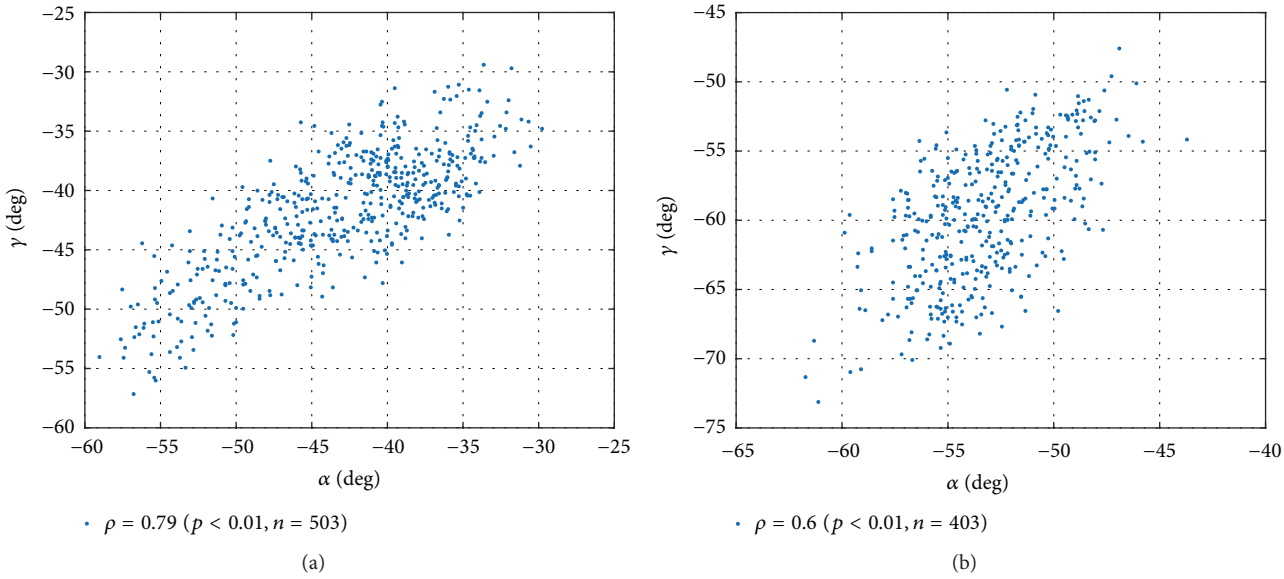


FIGURE 10: The presence of CD8 changes the relative movement of domains. (a) Scatter plot of Euler angles α and γ between domains MHC α_1 and α_2 in the absence of CD8. (b) Scatter plot of the Euler angles α and γ with CD8 present.

In this work we present a more sophisticated approach by attaching local coordinate systems (of eigenvectors) to each domain and calculating the rotational relations between them. It is a well-known drawback of eigenvector-based techniques that the eigenvector *orientation* is not well defined and may suddenly switch into almost opposite directions. Up to now, this was in most cases mended by some logical condition in the code selecting the appropriate orientation with reference to some atoms that have to be individually specified. These drawbacks are even more severe when eigenvectors are computed from internally deformable sets of data points, such as MD-frames.

We have solved this problem by computing eigenvectors only once (for each domain) from a very specific frame (the central frame), see Section 2.3.1. In this first step, the orientation of both systems of eigenvectors may be corrected if desired. Thereafter, relations will remain stable without any intervention on the side of the researcher. Stability of local coordinate systems is achieved by fitting the atoms within each domain and carrying along the eigenvectors accordingly. This results in robust relative orientations.

Note that although the method seems computationally demanding, it is in fact fast. The reason is that the Kabsch algorithm is a direct matrix operation, not an iterative procedure, as one might expect. It requires no larger computational effort than singular value decomposition.

Once mutual orientations have been computed (rotation matrix), some attention was given to select suitable quantities to characterize relative orientations. We presented cosines between corresponding eigenvectors, since they can easily be interpreted. In addition, we used Euler angles as a well-known concept for relative orientations of rigid bodies.

4.2. Unifying Orientations. PCA as such yields eigenvectors unique in directions but ambiguous in orientation. For example, the first eigenvector $\mathbf{v}_1$ may also result as $\mathbf{v}_1^* = -\mathbf{v}_1$, merely as a consequence of minute numerical noise in data substantially equal. The same is true for $\mathbf{v}_2$. The third eigenvector always has to satisfy

$$\mathbf{v}_3 = \mathbf{v}_1 \times \mathbf{v}_2 \tag{7}$$

and is hence determined by $\mathbf{v}_1$ and $\mathbf{v}_2$. When comparing the relative orientation of two domains via their eigenvectors, orientation ambiguity has to be coped with, which could be done as follows:

(i) After performing a PCA for the reference frame k_V of domain V, results are manually inspected and $\mathbf{v}_1$ given a well-defined orientation within domain V. This can be accomplished by selecting two C_α atoms and setting the orientation of $\mathbf{v}_1$ such that a positive cosine of the angle between $\mathbf{v}_1$ and the vector joining the two C_α atoms is obtained.

(ii) The same is done for $\mathbf{v}_2$, with a second pair of appropriate C_α atoms selected. $\mathbf{v}_3$ is computed via the cross product of $\mathbf{v}_1$ and $\mathbf{v}_2$; see above.

In order to arrive at standardized eigenvectors allowing for comparison between different MD-runs the resulting eigenvectors should be reoriented (if necessary) after performing PCA for the reference frame k_W of domain W: a criterion for reorientation could be positive cosines with the eigenvectors of domain V:

$$\begin{aligned} \mathbf{v}_1 \cdot \mathbf{w}_1 &\geq 0, \\ \mathbf{v}_2 \cdot \mathbf{w}_2 &\geq 0. \end{aligned} \tag{8}$$

Note that flipping the orientation of eigenvectors between different frames of a trajectory is avoided intrinsically by our procedure (fitting of domains rather than repeatedly

performing PCA to successive frames). However, initially selecting appropriate and definite orientations is only guaranteed by the above precautions.

4.3. Application of the New Methods to CD8 Coreceptor Dynamics. Inspecting the distance between CD8 and MHC α_3 over time (Figure 4(a)) we clearly observe an equilibration phase pertaining up to 50 ns, during which distance decreases. This observation suggests that more extensive MD simulations than those presented in this work are necessary to reliably investigate equilibrium properties of this system. This finding is also supported by the autocorrelation function, which passes through zero around 60 ns and shows a long negative tail afterwards; see Figure 4(b).

The existence of an equilibration phase raises the importance of a comparison of the two presentation methods employed: cosines between eigenvectors and Euler angles (see Figure 6). Clearly, cosines reflect the fact that there is a relaxation phase, while the Euler angles do not. Interestingly, the eigenvectors corresponding to the main extensions ($\mathbf{v}_1$, $\mathbf{w}_1$) fail to indicate the relaxation but ($\mathbf{v}_2$, $\mathbf{w}_2$) and ($\mathbf{v}_3$, $\mathbf{w}_3$) clearly do. This indicates that the relaxation is made up of some rotation around the major axes, $\mathbf{v}_1$ and $\mathbf{w}_1$, respectively.

This surprising finding is supported by comparing the corresponding autocorrelation functions; see Figure 7. Euler angles (Figure 7(b)) lose memory already after 20 ns and the ACF then oscillates around zero. As opposed to this, cosines between eigenvectors corresponding to the smaller components (($\mathbf{v}_2$, $\mathbf{w}_2$) and ($\mathbf{v}_3$, $\mathbf{w}_3$)) show excessively long autocorrelations, pertaining throughout the whole simulation time. This finding, even more than the 50 ns equilibration of distance, indicates the necessity of additional MD simulations to arrive at a more adequate sampling.

As a most interesting result, we found that the presence of CD8 seems to influence the dynamics within the MHC, in particular the relative movements of the two α-helices, Gα_1 and Gα_2. In this case, Euler angles proved the more sensitive tool. Negative correlation between cosines of ($\mathbf{v}_2$, $\mathbf{w}_2$) and ($\mathbf{v}_3$, $\mathbf{w}_3$) was in both cases beyond $\rho < -0.9$, and presence or absence of CD8 did not make much difference. For the Euler angles, however, we obtained the nice reduction of movement-correlation induced by the presence of CD8; see Figure 10.

All in all, both sets of orientation parameters presented (cosines and Euler angles) seem to have their merits and weaknesses that have to be explored in many more situations to arrive at a comprehensive judgment.

Conflict of Interests

Reiner Ribarics is an employee and stockholder of Gilead Sciences. The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Wolfgang Schreiner and Rudolf Karch contributed equally.

Acknowledgments

MD simulations were run on the IBM-BlueGene/P at Bulgarian NCSA. Partial support by Bulgarian Science Fund and Austrian Academic Exchange Programme under Grants DNTS-A 01-2/2013 and WTA-BG 06/2013 is acknowledged. Karin Schlangen helped with the statistical analysis. The software for the analysis is available on request from the authors.

References

[1] A. S.-Y. Leong, K. Cooper, and F. J. W.-M. Leong, *Manual of Diagnostic Cytology*, Greenwich Medical Media, 2nd edition, 2003.

[2] L. Devine, J. Sun, M. R. Barr, and P. B. Kavathas, "Orientation of the Ig domains of CD8$\alpha\beta$ relative to MHC class I," *Journal of Immunology*, vol. 162, no. 2, pp. 846–851, 1999.

[3] N. Jiang, J. Huang, L. J. Edwards et al., "Two-stage cooperative T cell receptor-peptide major histocompatibility complex-CD8 trimolecular interactions amplify antigen discrimination," *Immunity*, vol. 34, no. 1, pp. 13–23, 2011.

[4] M. Krogsgaard, Q.-J. Li, C. Sumen, J. B. Huppa, M. Huse, and M. M. Davis, "Agonist/endogenous peptide-MHC heterodimers drive T cell activation and sensitivity," *Nature*, vol. 434, no. 7030, pp. 238–243, 2005.

[5] M. N. Artyomov, M. Lis, S. Devadas, M. M. Davis, and A. K. Chakraborty, "CD4 and CD8 binding to MHC molecules primarily acts to enhance Lck delivery," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 39, pp. 16916–16921, 2010.

[6] M. R. von Essen, M. Kongsbak, and C. Geisler, "Mechanisms behind functional avidity maturation in T cells," *Clinical and Developmental Immunology*, vol. 2012, Article ID 163453, 8 pages, 2012.

[7] M. K. Slifka and J. L. Whitton, "Functional avidity maturation of $CD8^+$ T cells without selection of higher affinity TCR," *Nature Immunology*, vol. 2, no. 8, pp. 711–717, 2001.

[8] L. J. Walker, A. K. Sewell, and P. Klenerman, "T cell sensitivity and the outcome of viral infection," *Clinical and Experimental Immunology*, vol. 159, no. 3, pp. 245–255, 2010.

[9] G. F. Gao, Z. Rao, and J. I. Bell, "Molecular coordination of $\alpha\beta$ T-cell receptors and coreceptors CD8 and CD4 in their recognition of peptide-MHC ligands," *Trends in Immunology*, vol. 23, no. 8, pp. 408–413, 2002.

[10] M. A. Daniels and S. C. Jameson, "Critical role for CD8 in T cell receptor binding and activation by peptide/major histocompatibility complex multimers," *Journal of Experimental Medicine*, vol. 191, no. 2, pp. 335–346, 2000.

[11] J. G. Borger, R. Zamoyska, and D. M. Gakamsky, "Proximity of TCR and its CD8 coreceptor controls sensitivity of T cells," *Immunology Letters*, vol. 157, no. 1-2, pp. 16–22, 2014.

[12] B. Liu, S. Zhong, K. Malecek et al., "2D TCR-pMHC-CD8 kinetics determines T-cell responses in a self-antigen-specific TCR system," *European Journal of Immunology*, vol. 44, no. 1, pp. 239–250, 2014.

[13] B. Hess, C. Kutzner, D. van der Spoel, and E. Lindahl, "GROMACS 4: algorithms for highly efficient, load-balanced, and scalable molecular simulation," *Journal of Chemical Theory and Computation*, vol. 4, no. 3, pp. 435–447, 2008.

[14] H. J. C. Berendsen, J. P. M. Postma, W. F. Van Gunsteren, and J. Hermans, "Interaction models for water in relation to protein hydration," in *Intermolecular Forces*, pp. 331–342, 1981.

[15] B. Hess, "P-LINCS: a parallel linear constraint solver for molecular simulation," *Journal of Chemical Theory and Computation*, vol. 4, no. 1, pp. 116–122, 2008.

[16] U. Essmann, L. Perera, M. L. Berkowitz, T. Darden, H. Lee, and L. G. Pedersen, "A smooth particle mesh Ewald method," *The Journal of Chemical Physics*, vol. 103, no. 19, pp. 8577–8593, 1995.

[17] A. Amadei, A. B. M. Linssen, and H. J. Berendsen, "Essential dynamics of proteins," *Proteins*, vol. 17, pp. 412–425, 1993.

[18] J. S. Hub and B. L. de Groot, "Detection of functional modes in protein dynamics," *PLoS Computational Biology*, vol. 5, no. 8, Article ID e1000480, 2009.

[19] H. J. Kim, M. Y. Choi, H. J. Kim, and M. Llinás, "Conformational dynamics and ligand binding in the multi-domain protein PDC109," *PLoS ONE*, vol. 5, no. 2, article e9180, 2010.

[20] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica Section A*, vol. 32, no. 5, pp. 922–923, 1976.

[21] H. Goldstein, C. P. Poole, and J. L. Safko, *Classical Mechanics*, 2nd Impr Ed, Pearson, Upper Saddle River, NJ, USA, 3rd edition, 2006.

[22] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 33–38, 1996.

[23] J. Cohen, P. Cohen, S. G. West, and L. S. Aiken, *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, Lawrence Erlbaum, Mahwah, NJ, USA, 3rd edition, 2003.

[24] T. Ichiye and M. Karplus, "Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations," *Proteins: Structure, Function and Genetics*, vol. 11, no. 3, pp. 205–217, 1991.

[25] B. Hischenhuber, H. Havlicek, J. Todoric, S. Höllrigl-Binder, W. Schreiner, and B. Knapp, "Differential geometric analysis of alterations in MH α-helices," *Journal of Computational Chemistry*, vol. 34, no. 21, pp. 1862–1879, 2013.

[26] B. Hischenhuber, H. Havlicek, J. Todoric, S. Höllrigl-Binder, W. Schreiner, and B. Knapp, "Corrigendum: differential geometric analysis of alterations in MH alpha-helices," *Journal of Computational Chemistry*, vol. 34, no. 32, p. 2834, 2013.

[27] S. Bernhard and F. Noé, "Optimal identification of semi-rigid domains in macromolecules from molecular dynamics simulation," *PLoS ONE*, vol. 5, no. 5, Article ID e10491, 2010.

Hindawi Publishing Corporation
Journal of Immunology Research
Volume 2015, Article ID 738030, 21 pages
http://dx.doi.org/10.1155/2015/738030

Review Article

Current Mathematical Models for Analyzing Anti-Malarial Antibody Data with an Eye to Malaria Elimination and Eradication

Nuno Sepúlveda,[1,2] Gillian Stresman,[1] Michael T. White,[3,4,5] and Chris J. Drakeley[1]

[1]*London School of Hygiene and Tropical Medicine, Keppel Street, London WC1E 7HT, UK*
[2]*Centro de Estatística da Universidade de Lisboa, Faculdade de Ciências, Universidade de Lisboa, Bloco C6, Piso 4, Campo Grande, 1749-016 Lisboa, Portugal*
[3]*MRC Centre for Outbreak Analysis and Modelling, Department of Infectious Disease Epidemiology, Imperial College London, Medical School Building, Norfolk Place, London W2 1PG, UK*
[4]*Division of Population Health and Immunity, Walter and Eliza Hall Institute, 1G Royal Parade, Parkville, VIC 3052, Australia*
[5]*Department of Medical Biology, The University of Melbourne, Parkville, VIC 3010, Australia*

Correspondence should be addressed to Nuno Sepúlveda; nuno.sepulveda@lshtm.ac.uk

Received 28 August 2015; Accepted 19 October 2015

Academic Editor: Francesco Pappalardo

Copyright © 2015 Nuno Sepúlveda et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

The last decade has witnessed a steady reduction of the malaria burden worldwide. With various countries targeting disease elimination in the near future, the popular parasite infection or entomological inoculation rates are becoming less and less informative of the underlying malaria burden due to a reduced number of infected individuals or mosquitoes at the time of sampling. To overcome such problem, alternative measures based on antibodies against specific malaria antigens have gained recent interest in malaria epidemiology due to the possibility of estimating past disease exposure in absence of infected individuals. This paper aims then to review current mathematical models and corresponding statistical approaches used in antibody data analysis. The application of these models is illustrated with three data sets from Equatorial Guinea, Brazilian Amazonia region, and western Kenyan highlands. A brief discussion is also carried out on the future challenges of using these models in the context of malaria elimination.

1. Introduction

Malaria is a global health problem with more than 1 billion people estimated to be at risk. This infectious disease is caused by *Plasmodium* parasites transmitted to humans through bites of infected Anopheles mosquitos. Geographically, *Plasmodium falciparum* (*P. falciparum*) parasites predominate in sub-Saharan Africa while *Plasmodium vivax* (*P. vivax*) is the major infectious agent in South America and Southeast Asia. According to the latest World Malaria Report [1], disease mortality and risk have been steadily decreasing in the last decade to the point that many countries are already targeting malaria elimination and eradication [2–5]. This decreasing trend in malaria transmission intensity, although highly beneficial to the affected populations, brings additional challenges to disease surveillance and elimination (reviewed in [6]). One of these challenges is related to the use of the current metrics of malaria risk in populations where disease transmission intensity is low and potentially affected by seasonal effects. The popular parasite rate is determined by the proportion of infected individuals at time of the survey. However, in low transmission settings, this measure is critically affected by the different performance of current diagnostic tools to detect the presence of infection while screening asymptomatic individuals. Another difficulty of using such measure is the high chance of finding only a few infected individuals in the sample, thus, having limited power to discriminate disease hotspots from other less-affected sites, as demonstrated in studies from Brazil [7] or Somalia [8]. The entomological inoculation rate is yet another popular

measure of malaria risk. It is defined by the frequency at which people are bitten by infectious mosquitoes, thus, being informative on the direct interaction between the human and mosquito populations. The gold standard to estimate this measure is to use human-landing catches where mosquitoes are caught as they attempt to land on the exposed limbs of field workers [9, 10]. Although alternative methods exist in the literature, the estimation of the entomological inoculation rate is in general a laborious and time-consuming task in low transmission settings owing to a low number of infected mosquitoes [11]. It is also affected by seasonal effects and mosquito population dynamics and the degree of mosquito attractiveness to the human hosts or the chemicals used in the study [11].

To tackle the limitations of the above malaria risk measures, alternative indicators based on antibodies against different malaria antigens have been proposed [12] and tested in different epidemiological contexts [7, 8, 13–16]. The rationale of using antibody data is that the antibody concentrations in the serum are a direct correlate of parasite exposure, thus, providing information on current and recent infections. The temporal stability in antibody concentrations is an important advantage to reduce any seasonal effect on malaria transmission. In seroepidemiological studies, the most popular antibodies are those against the blood-stage apical membrane antigen-1 (AMA1) and merozoite surface protein-1 (MSP1) [7, 8, 13–16] owing to their broad immunogenicity and putative role in malaria vaccine development [17, 18]. Recent research identified other parasite targets [19, 20] but these remain to be tested in different epidemiological settings. Experimentally, antibody quantification is usually done by means of traditional enzyme linked immunosorbent assays [21]. Optical densities or titres in arbitrary units are then used for the subsequent data analysis. The most popular approach is to first define the serological status, seropositive or seronegative, of each individual. One then calculates the so-called seroprevalence that is defined by the proportion of seropositive individuals in the sample. Several studies showed an increased resolution of seroprevalence in discriminating sites with different endemicity levels in relation to parasite rate [7, 8]. Further analysis is then carried out in order to estimate current malaria transmission intensity. Since seroprevalence tends to increase with age as a result of augmenting immunity against malaria parasites, different stochastic models can be constructed for the data using age as a proxy of time. The common assumption to all these models is that individuals transit between seronegativity and seropositivity states upon malaria exposure or absence of it. In this scenario, one typically estimates the rate by which seronegative individuals become seropositive, the so-called seroconversion rate (SCR). SCR was found to correlate well with the parasite rate [13] or the entomological inoculation rate [12], thus, capturing the underlying malaria transmission intensity. Moreover, SCR also strongly correlates to the annual parasite index (the number of confirmed cases during 1 year/population under surveillance) $\times 1000$—usually calculated by official health authorities [7].

This paper aims to review the mathematical and statistical aspects underlying the analysis of antibody data for inferring malaria transmission intensity. Special attention will be given to current methods aiming to define seropositivity and the subsequent mathematical models for estimating SCR under different epidemiological settings: stable malaria transmission intensity, abrupt reduction in SCR due to a malaria control intervention, change in SCR due to a putative age-dependent behavior, detection of migration effects, and detection of individual level heterogeneity through a set of covariates. Models for antibody acquisition using antibody titres themselves will also be described. Three different data sets from Bioko Island in Equatorial Guinea [15], Jacarecanga from the Brazilian Amazonia region [7], and western highlands from Kenya [22] are used to illustrate the application of these models to real-world problems. Finally, future analytical challenges will be discussed in the context of malaria elimination and eradication.

2. Mathematical Approaches to Analyzing Serology Data

2.1. Defining Seropositivity. In practice, there are two popular approaches to determine the serological status of an individual. The first approach uses an additional sample of nonexposed individuals in order to determine the distribution of the antibody levels referring to the underlying seronegative population. Statistically, the antibody levels of this sample are usually log transformed in order to approximately obtain a Gaussian distribution for the data. The serological classification of each individual in the sample is done by the 3σ rule for Gaussian distributions described in any introductory textbook of statistics. In more detail, this rule defines the range of antibody levels containing a 0.999 probability under the assumption of a Gaussian random variable for the data. One then classifies the individuals as seropositive if the respective antibody levels exceed the mean plus 3 times the standard deviation of the seronegative population, otherwise the individuals are considered as seronegative. This simple approach, despite ensuring a high probability of correctly classifying exposed individuals, has the disadvantage of underestimating seroprevalence.

The second approach focuses on the data under analysis only. The basic assumption is that the sample is composed of a mixture of latent seronegative or seropositive populations. The respective data is then analyzed by the so-called two-component mixture Gaussian model invoking a Gaussian distribution with average value μ_0 and standard deviation σ_0 for the seronegative population and another one with average value μ_1 and standard deviation σ_1 for the seropositive population. For independent and identically distributed random sample of n individuals, the corresponding sampling distribution is described by the following equation:

$$f\left(\{x_i\} \mid \mu_0, \mu_1, \sigma_0, \sigma_1, \pi\right) = \prod_{i=1}^{n} \left[(1-\pi)\, f_{N(\mu_0,\sigma_0)}\left(x_i\right) + \pi f_{N(\mu_1,\sigma_1)}\left(x_i\right)\right], \quad (1)$$

where x_i is the antibody level of the ith individual in the sample, $f_{N(\mu_0,\sigma_0)}(x_i)$ and $f_{N(\mu_1,\sigma_1)}(x_i)$ are probability

density functions of the Gaussian distributions associated with seronegative and seropositive populations, respectively, and π is the probability of sampling a seropositive individual from the population. Maximum likelihood estimation is facilitated by using the expectation-maximization (EM) algorithm that can be found in the mixtools package for the R software [23]. The next stage of the analysis is to assign each individual to each corresponding serological population. Again, one can use the 3σ rule as described above [14]. An alternative way to perform such classification is to jointly use the probabilities of classifying an individual with antibody level x as either seropositive or seronegative and then specify appropriate cut-off values to determine the serological status of each individual. The probabilities of classifying an individual with antibody level x as seropositive and seronegative are, respectively, given by

$$p_{+|x} = \frac{\pi f_{N(\mu_1,\sigma_1)}(x)}{(1-\pi) f_{N(\mu_0,\sigma_0)}(x) + \pi f_{N(\mu_1,\sigma_1)}(x)}, \qquad p_{-|x} = 1 - p_{+|x}. \tag{2}$$

The classification rule of the ith individual in the sample is then described as follows:

$$C_i = \begin{cases} \text{seronegative}, & \text{if } x_i \le c^- \\ \text{indeterminate}, & \text{if } c^- < x_i < c^+ \\ \text{seropositive}, & \text{if } x_i \ge c^+, \end{cases} \tag{3}$$

where c^- and c^+ are the cut-off values in the antibody distribution that ensure a given classification probability, for instance, 90%. Note that individuals with antibody levels between c^- and c^+ are deemed indeterminate due to the uncertainty in the corresponding serological classification. Besides checking whether model assumptions hold true on the data under analysis, an additional assessment of the quality of the classification rule is to report the size of this indeterminate region and the proportion of indeterminate individuals in the sample.

Example 1 (Bioko Island). In 2004 the health authority of Equatorial Guinea launched integrated treatment and mosquito control programs in the Bioko Island. After 4 years of their initiation, a large cross-sectional survey was conducted at 18 sentinel sites in the island in order to assess the impact of these programs on malaria transmission [15]. IgG antibody levels of 6400 individuals were measured for *P. falciparum* AMA1 by ELISA. The antibody levels as measured by arbitrary titres range from −116.3 to 2618.9, suggesting a wide breadth of immune responses to this malaria antigen (Figure 1(a)). The average antibody level was 390.8 while the standard deviation was estimated at 457.4. As expected from data of a malaria endemic region, the corresponding quantile-quantile plot showed a strong departure of the data in relation to the Gaussian distribution due to presence of recently or currently exposed individuals with high antibody levels (Figure 1(b)). By fitting the above two-component Gaussian mixture model to the data, the serological status of each individual was determined by (3) with $c^- = 97.0$ and $c^+ = 200.8$ (Figure 1(c)). These cut-off values suggested that 31.2% and 56.1% of the sample consisted of seronegative and seropositive individuals, respectively. The remaining 12.7% of the sample had unclear serological classification (Table 1).

The above Gaussian mixture model can be extended to the setting where there are more than two components. Immunologically, such extension is in line with the notion that antibody levels can be boosted by frequent malaria exposure [24]. In this scenario, each component can be interpreted as corresponding to a specific degree of malaria exposure: not exposed, once exposed, twice exposed, three times exposed, and so forth.

Under the assumption of a known number of components for the data (say $K + 1$), the corresponding sampling distribution is given by

$$f(\{x_i\} \mid \{\mu_k, \sigma_k, \pi_k\}) = \prod_{i=1}^{n} \left[\sum_{k=0}^{K} \pi_k f_{N(\mu_k,\sigma_k)}(x_i) \right], \tag{4}$$

where $\mu_0 < \mu_1 < \mu_2 < \cdots < \mu_K$ are the averages of the population not exposed, once exposed, twice exposed, ..., and K times exposed, respectively, $\sigma_0, \sigma_1, \sigma_2, \ldots, \sigma_K$ are the corresponding standard deviations, and $\pi_0, \pi_1, \pi_2, \ldots, \pi_K$ are the corresponding sampling probabilities (with $\pi_0 = 1 - \sum_{k=1}^{K}$). The conditional classification probabilities of seropositive and seronegative individuals given antibody level x can be generalized as follows:

$$p_{+|x} = \frac{\sum_{k=1}^{K} \pi_k f_{N(\mu_k,\sigma_k)}(x)}{\sum_{k=0}^{K} \pi_k f_{N(\mu_k,\sigma_k)}(x_i)}, \qquad p_{-|x} = 1 - p_{+|x}. \tag{5}$$

The corresponding classification rule is also given by (3) but now the cut-off values must be recalculated according to these new classification probabilities. As for the two-component Gaussian mixture model, maximum likelihood estimation via EM algorithm can also be performed to estimate the unknown parameters $\{\mu_k, \sigma_k, \pi_k,\ k = 0, \ldots, K\}$. Starting this estimation algorithm with different initial conditions is recommended to obtain the correct convergence to the global maxima of the log-likelihood function.

An important question in practice is to know how many components one must consider to describe the data well. In terms of maximum likelihood estimation, this question can be answered by using the profile likelihood method. This method proceeds as follows: (i) start the analysis with $K = 1$, (ii) obtain the corresponding maximum likelihood estimates and then calculate the respective value of the likelihood function, (iii) add another component into the analysis and repeat step (ii), and (iv) keep increasing the number of components until reaching a realistic maximum value for that parameter. The optimal number of components is the one providing the maximum value of all maximum likelihood values calculated for each number of components considered in the analysis. The profile likelihood method, despite estimating the total number of components, brings potential problems

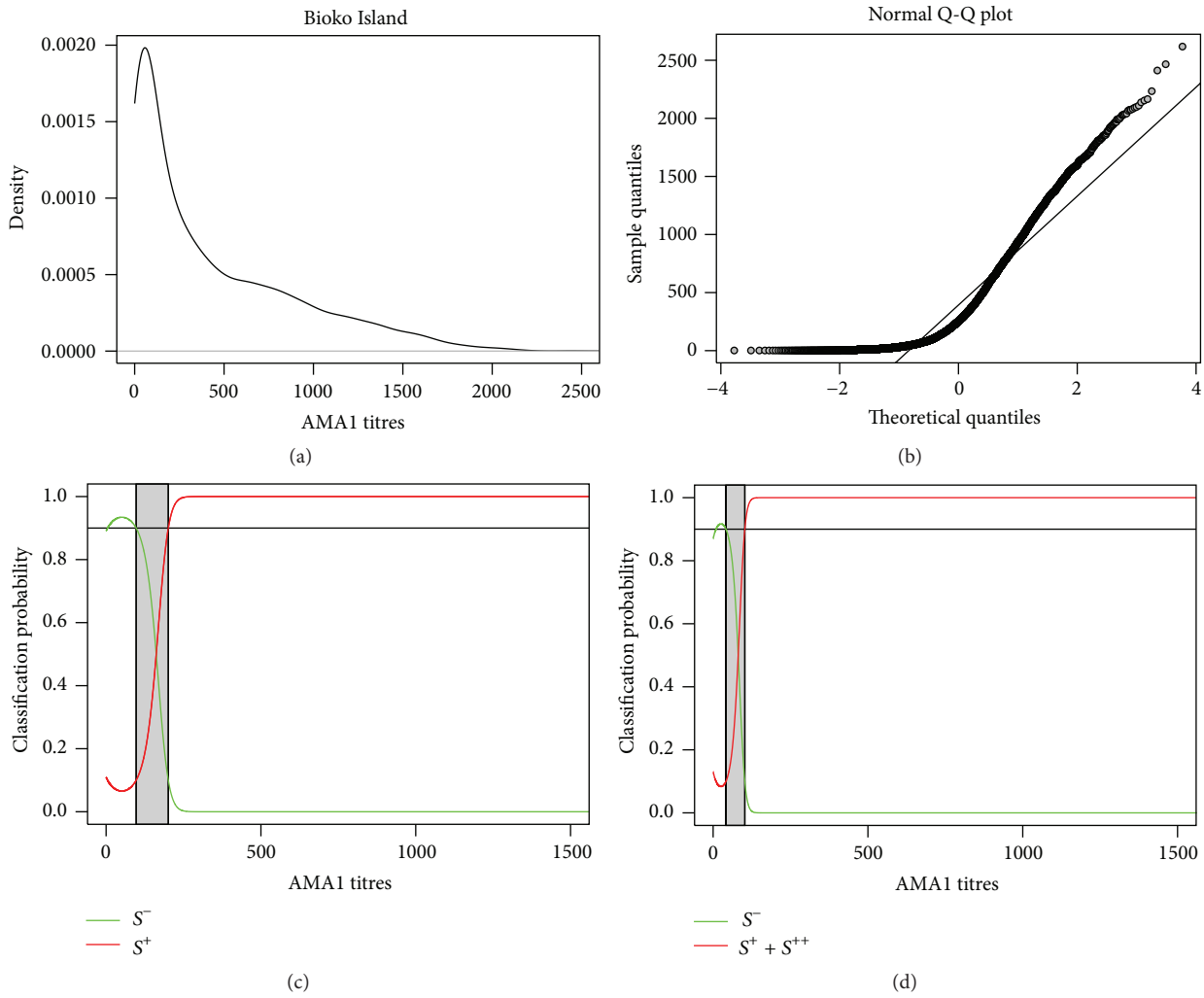


Figure 1: Determining seropositivity of anti-AMA1 antibodies from Bioko Island. (a) Probability density plot for the titre data. (b) Gaussian (or Normal) quantile-quantile plot for the data. (c) Classification probability curves predicted by the two-component Gaussian mixture model. (d) Classification probability curves predicted by the best three-component Gaussian mixture model where the intermediate component refers to a seropositive population.

of model overfitting and uncertainty in the classification rule. Overfitting can be obtained by considering a model with too many components. This problem can be surpassed by using different information measures with the aim of weighting the quality of the data fitting with the intrinsic complexity of a model. The most popular information measure is the Akaike's information criterion (AIC) defined by twice the absolute value of the log-likelihood function evaluated at the maximum likelihood estimates (measuring the quality of the respective data fitting) plus twice the total number of estimated parameters (estimating the intrinsic model complexity). Since models are penalised in this criterion as function of the total number of parameters, one should choose the model that shows the lowest AIC value among all models tested. Uncertainty in the classification rule can arise from data where the different serological populations are tight together in the antibody distribution. A simple solution is to choose the model with the highest likelihood but implying a sufficiently clear serological classification of the individuals in the sample.

An additional difficulty in using a Gaussian mixture model with more than two components is the ambiguity in linking each component to the corresponding serological status. Let us consider the three-component mixture model for the moment. In this setting, the components with the lowest and highest average titres are easily interpreted as related to seronegative and seropositive populations, respectively. On the one hand, the component with intermediate average titres can be interpreted as a seronegative population if one assumes two populations with different genetic backgrounds. This interpretation agrees with studies from Burkina Faso where the Fulani typically have higher antibody concentrations at baseline in comparison to other ethnic groups living in the same area [25, 26]. On the other hand, this intermediate component can also be interpreted as a seropositive population under the assumption of immunity boosting upon recurrent

Table 1: Gaussian mixture modelling analyses for determining seropositivity to AMA1 titre data in a sample of around 6400 individuals from Bioko Island using 90% as the cut-off value for the correct classification probability.

Number of components	AIC[a]	Mean (SD)[b]	Definition of S^+ and S^-	Cut-off values[c]		Classification probabilities[d]		
				c^-	c^+	P_{S^-}	P_{ind}	P_{S^+}
2	84601.2	59.3 (48.4)	$S^- = 1, S^+ = 2$	95.9	202.9	31.2	12.7	56.1
		668.1 (450.4)						
3	83395.2	35.8 (26.8)	$S^- = 1, S^+ = 2, 3$	44.6	109.8	19.3	13.8	66.8
		214.0 (115.4)	$S^- = 1, 2, S^+ = 3$	103.8	515.2	32.3	33.3	34.4
		848.3 (425.6)						
4	82887.4	14.1 (9.1)	$S^- = 1, S^+ = 2, 3, 4$	NA	37.2	—	17.0	83.0
		64.7 (32.4)	$S^- = 1, 2, S^+ = 3, 4$	34.0	149.5	16.1	22.1	61.9
		252.2 (120.6)	$S^- = 1, 2, 3, S^+ = 4$	135.4	560.3	36.4	31.3	32.3
		873.2 (420.6)						

[a]The best model is the one providing the lowest estimated value.
[b]Mean and standard deviation (SD) are for each Gaussian component in the model ordered by the corresponding average titres.
[c]c^- and c^+ are the cut-off values for determining the seronegative and seropositive populations, respectively.
[d]P_{S^-}, P_{ind}, and P_{S^+} are the estimated classification probabilities of seronegative, indeterminate, and seropositive individuals, respectively.

malaria exposure as described above. This and the component with the highest average concentrations are then related to exposed and boosted populations, respectively. Similar reasoning can easily be applied to the scenario of a higher number of components. For that one just needs to consider the putative existence of more than one seronegative and seropositive population. In absence of additional information about the populations under study, it is difficult to resolve the ambiguity about component interpretation. A possible solution is to first understand how the performance of the classification is affected by changing the interpretation of the components and then make a judgement call upon the reasonability of the corresponding results.

Example I (Bioko Island continued). Previous analysis was extended to fit Gaussian mixture models with more than two components. Models with three and four components seemed to describe the data better than the one with two components only, according to AIC (Table 1). Despite providing a good balance between data fitting and model complexity, the four-component models implied high percentages of individuals with unclear classification (>22%). The best model would appear to be the one with three components where the second and the third components were interpreted as referring to seropositive populations. This model improves the quality of the data fitting and implied a percentage of individuals with unclear classification (14%) similar to the one obtained from the two-component model. Comparing to previous results for the two-component model, the inclusion of a third additional component suggested that the seropositive population could in fact be split into exposed and boosted individuals with average antibody titres of 214.0 and 848.3, respectively. The new cut-off values for the classification rule led to the classification of 19.3% and 66.8% of the sampled individuals as seronegative and seropositive, respectively.

Recent research has highlighted the great potential of using Bayesian approaches in Gaussian mixture models. The major advantage of these methods is to provide a coherent and elegant analytical framework for estimating the total number of components from the data. Since this number is unknown quantity, it is considered as random variable with a given probability distribution before conducting data analysis the so-called prior distribution. Bayes theorem then allows linking the prior distributions for all unknown parameters with the sampling distribution of the data. As a result, prior distributions of the parameters are updated by the data, giving rise to the so-called posterior distributions. These latter distributions are then the core of the Bayesian statistical inference. The current success of Bayesian approaches is intimately related to the use of powerful simulation methods in order to determine the posterior distributions given the data. In Gaussian mixture models, the Markov Chain Monte Carlo with reversible jumps is a popular choice for posterior estimation [27]. Similar simulation algorithm can theoretically be applied to multivariate Gaussian mixture models [28]. These models are particularly suitable for analyzing data of more than one malaria antigen simultaneously (e.g., for analyzing AMA1 and MSP1 data together). A Bayesian two-component mixture model using arbitrary probability distributions for the latent populations was proposed for classifying fever and nonfever malaria cases according to the underlying parasitaemia [29, 30]. Up to now a single seroepidemiological study in malaria [14] is known to have analyzed data via Bayesian methods and, thus, little can be said about their performance in practice.

2.2. Detecting Stable Malaria Transmission Intensity Using Seropositivity Data. After classifying individuals into their serological status, the corresponding data analysis proceeds by estimating stochastic models that aim to inform about the underlying malaria transmission intensity. The most popular models belong to the class of the reversible catalytic models (RCMs) [31–33]. When applied to serological data from infectious diseases that do not induce long-lasting immunity, such as the case of malaria, these models assume that age

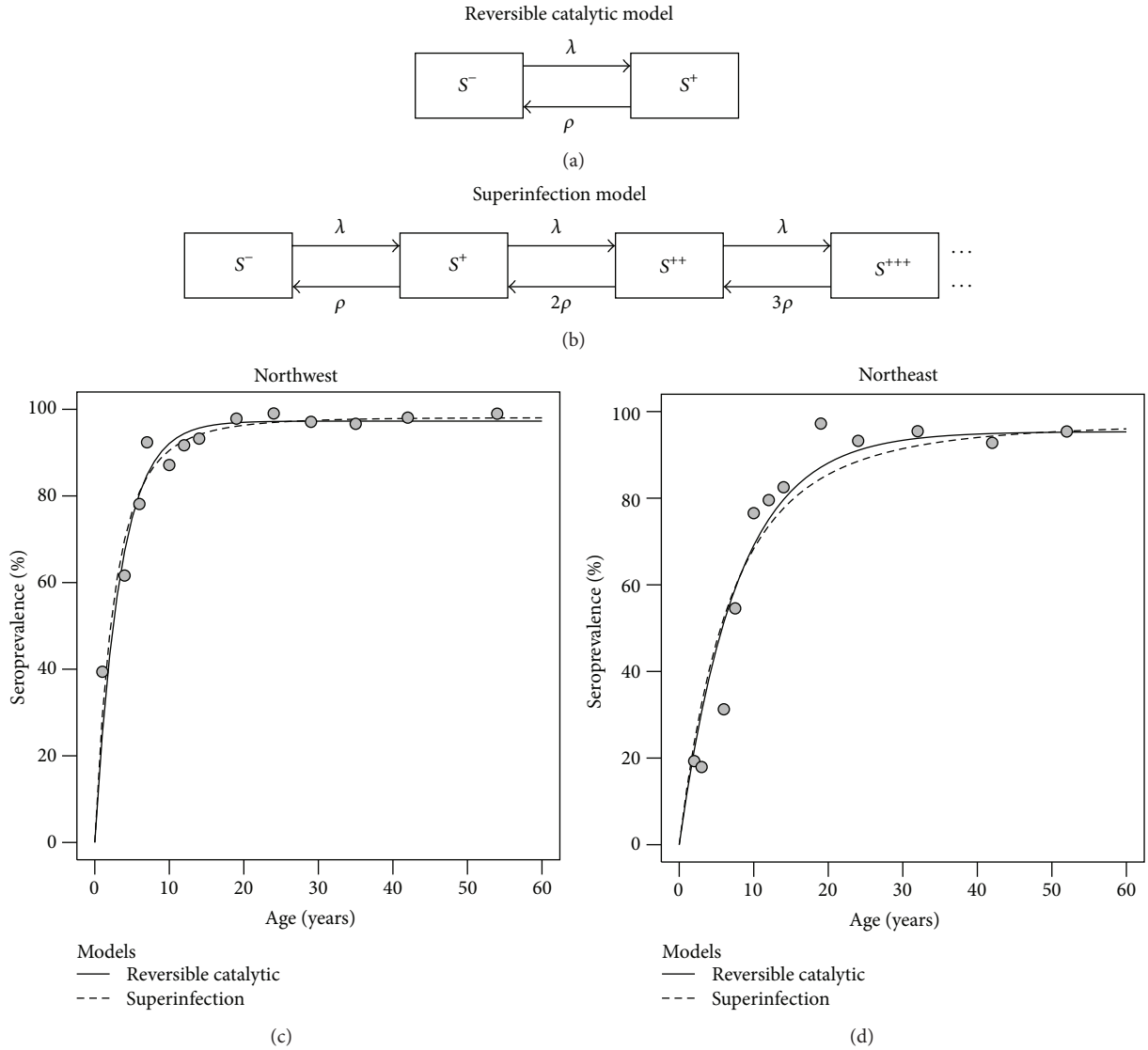


Figure 2: Analysis of seropositivity data. (a) Compartmental representation of the reversible catalytic model where individuals transit between seronegative and seropositive states with rates λ (SCR) and ρ (SRR). (b) Compartmental representation of the superinfection model in which there are multiple seropositive states owing to immunity boosting upon recurrent malaria infections. (c) Analysis of seropositivity AMA1 data from northwest region of Bioko Island under the assumption of stable malaria transmission over time. (d) Similar data analysis for northeast region of Bioko Island. In plots (c) and (d), the dots represent the observed seroprevalence of distinct age groups by splitting the sampled age distribution into 7.5% centiles. To plot each seroprevalence, the median value of each age group was used.

is deemed an appropriate proxy of the historical time so that data from each individual can be seen as a random realization of a seroconversion-seroreversion stochastic law. More precisely, individuals are born as seronegative but can be converted into seropositive upon malaria exposure. In the absence of frequent malaria exposure, individuals can revert to a seronegative state (Figure 2(a)). Mathematically, this idea can be described as a Markov chain model where one must specify the average rates by which the individuals become seropositive and return to the seronegative, the seroconversion, and seroreversion rates (SCRs and SRRs), respectively. Epidemiologically, SCR is related to the underlying disease transmission intensity as it correlates well with typical malariometrics, such as parasite rate or entomological inoculation rate. It is also related to (host) factors affecting antibody production. In contrast, SRR reflects host factors (e.g., genetics or age) affecting antibody decay in absence of malaria infection.

The simplest model for the data is to assume stable and constant malaria transmission intensity over time. A fixed SRR is also assumed because seropositivity data has limited power to describe variations in that parameter. For mathematical simplicity, the seroconversion-seroreversion dynamics of each individual is easily described by a Markov

chain with two states, seronegative (S^-) and seropositive (S^+). The resulting RCM is described by the following probability of an individual aged t being seropositive:

$$p_{S^+}(t) = \frac{\lambda}{\lambda+\rho}\left(1 - e^{-(\lambda+\rho)t}\right), \quad (6)$$

where λ and ρ are the SCR and SRR, respectively. It is worth noting that the above probability is an increasing function of age reaching a plateau at $\lambda/(\lambda+\rho)$ when age goes to infinite.

The above model can be extended to the so-called superinfection model (SIM), where immunity boosting can occur owing to recurrent malaria infections [24]. In line with the Gaussian mixture models with more than 2 components for antibody titre data, the notion of boosting can be translated into distinct seropositive states, for instance, S^+, S^{++}, and S^{+++}, depending on the cumulative level of malaria exposure (Figure 2(b)). In particular, a seronegative individual becomes a first-order seropositive upon a malaria infection. This same individual while still being first-order seropositive can evolve to a second-order seropositive upon an additional malaria exposure and so forth. A practical implication of this idea is a longer sojourn time in the seropositive state(s) in relation to the one predicted by RCM. Moreover, since there are multiple latent seropositive states, the estimates of the seroconversion rate tend to be higher in this model than in its reversible catalytic counterpart for the same data. The probability of an individual aged t being at any seropositive state is now given by

$$p_{S^*}(t) = 1 - e^{-(\lambda/\rho)(1-e^{-\rho t})}, \quad (7)$$

where S^* represents the set of all possible seropositive states an individual can belong to. More details on the corresponding mathematical derivation can be found elsewhere [24]. In practice, the application of this model to real-world problems shows limitations in terms of estimation [34]. On the one hand, SIM and RCM are approximately equivalent to each other in low transmission settings due to the rarity of boosting events. On the other hand, seroreversion is a rare event in high transmission settings due to boosting. Thus, for the matter of simplicity, a model considering seroconversion only is more reasonable for that situation. Interestingly, (6) and (7) when $\rho \to 0$ can be rewritten as the classical complementary log-log model [35]:

$$\log\left[-\log\left(1 - p_{S^+}(t)\right)\right] = \log\lambda + \log t. \quad (8)$$

Despite having limited application in malaria research [36], the complementary log-log model has been used in non-malaria immunological settings where a single immunization is thought to exert a permanent seropositive phenotype [37, 38].

With respect to model estimation, seropositive data adjusted for age is organized as a two-way frequency table with A rows and two columns, where A is the total number of different age values in the sample and the two columns refer to the serological status of the individuals (i.e., seronegative and seropositive). In this data format, the sampling distribution is assumed to be a Binomial-product sampling distribution, an independent Binomial distribution per age value and probability of success given by the model under fitting; that is,

$$f\left(\{m_t\} \mid \{n_t\}, \lambda, \rho\right) = \prod_{t=1}^{A} \binom{n_t}{m_t} [\pi(t)]^{m_t} [1 - \pi(t)]^{n_t - m_t}, \quad (9)$$

where m_t and n_t are the frequency of seropositive and all individuals aged t years, respectively, and $\pi(t)$ is the expected seroprevalence at age t described by (6), (7), or (8) if estimating RCM, SIM, or the complementary log-log model, respectively. Maximum likelihood estimation can be applied to the data. Stata and R scripts for data fitting are currently available from the authors upon request.

Example I (Bioko Island continued). As mentioned earlier, the cross-sectional survey from Bioko Island consisted of 18 sentinel sites spread over the island. To increase statistical power, the corresponding data was analyzed by considering 5 major geographical regions: northeast, northwest, southeast, southwest, and Malabo. A comprehensive analysis of this data set can be found in the original study report [15]. For illustrative purposes, the statistical analysis was carried out on data from northeast and northwest regions specifically. According to the seropositivity determination step, there are 1332 and 877 individuals with an assigned serological status from northwest and northeast regions, respectively. The corresponding overall seroprevalence was estimated at 86.7% (95% CI: 84.8%–88.5%) and 69.9% (95% CI: 66.7%–72.9%). These estimates are higher than the ones reported in the original study (69.2% and 46.6%, resp.) because this study used a two-component Gaussian mixture model for titre data, thus, predicting a higher cut-off value for seropositivity (Table 1). As expected from a malaria endemic area, the seroprevalence increased with age in both regions (Figures 2(c) and 2(d)). With respect to northwest region, both models described the seroprevalence curve well (Figure 2(c)). However, SIM provided a slightly better fit to the data than RCM (log-likelihood = −69.71 and −71.31, resp.), a result in line with the use of a three-component mixture model for seropositivity determination. Also in agreement with theoretical expectations was the higher SCR obtained from SIM in relation to the one predicted by RCM (0.359 versus 0.286; Table 2). For northeast region, the overall seroprevalence is decreased, thus, implying lower SCR estimates for RCM and SIM (0.124 and 0.139 for RCM and SIM, resp.). Although RCM showed a better fit to the data than its SIM counterpart (log-likelihood = −96.87 and −102.95, resp.), both models overestimated the seroprevalence of young aged individuals (up to 10 years ago) (Figure 2(d)) and, thus, they could not be considered as good candidate models for the data. Such overestimation suggested that young aged and older individuals have different serological dynamics that cannot easily be captured by a stable malaria transmission assumption. An easy explanation is the putative reduction in malaria transmission intensity after the initiation of malaria

control programs in 2004 in the island. This and other related topics will be explored in the following section.

2.3. Detecting Heterogeneity in Malaria Transmission Intensity Using Seropositivity Data. A unique advantage of using serology data is the possibility of detecting heterogeneity in disease transmission across different epidemiological situations. This advantage has been demonstrated in several studies where seroprevalence taken as a function of age qualitatively changes at a given age value. Such change might be attributed to an abrupt reduction in malaria transmission after the initiation of a malaria control or elimination program [15, 16]. Similar phenomenon was found for Trachoma [39] or Chagas disease [40]. Another possible explanation for that change is related to distinct malaria risk between young and older individuals owing to behavioral factors [41]. A third and last explanation is the occurrence of migration waves over time [42], as observed in Chagas disease [43]. A detailed description of these scenarios follows.

2.3.1. Detecting Historical Changes in Malaria Transmission. The commitment of many national health authorities in reducing or targeting elimination in future years brings future challenges in assessing the real impact of the designed interventions on the target populations. This assessment can be made by analyzing seropositivity data conveniently. For that one assumes there was an abrupt reduction of malaria transmission intensity at some time point before data collection. It is expected that an abrupt reduction in malaria transmission intensity would translate in a similar effect on the SCR. Sampled individuals are then split according to their date of birth in relation to the calendar time when the reduction in malaria transmission intensity actually occurred. More precisely, the serological history of individuals born before that reduction contemplates a first time period where the past SCR operates followed by a second period where the current SCR sets the rules. In contrast, individuals born after the reduction would lie down on that second time period and, thus, their serological dynamics are simply described by previous RCM and SIM for stable malaria transmission.

To calculate the seroprevalence of an individual with age t that experienced a reduction in malaria transmission intensity at time τ before sample collection, one must consider the sum of two probabilities associated with the following mutually exclusive events: (i) an individual became seropositive between birth and $t-\tau$ and remained so after that and (ii) an individual remained seronegative between birth and $t-\tau$ and became seropositive after that. Since RCM can be formulated as a two-state Markov chain, the seroprevalence for individuals with age t is calculated by multiplying the vector of probabilities associated with an individual being seropositive and seronegative at time $t-\tau$ (see (6)) by the probability transition matrix of the second Markov chain associated with the current SCR and evaluated at time τ. The resulting expected seroprevalence is then given by

$$p_{S^+}(t) = \begin{cases} \theta_2\left(1-e^{-\gamma_2\tau}\right)+\theta_1\left(1-e^{-\gamma_1(t-\tau)}\right)e^{-\gamma_2\tau}, & \text{if } t>\tau \\ \theta_2\left(1-e^{-\gamma_2 t}\right), & \text{if } t\le\tau, \end{cases} \quad (10)$$

where $\theta_i = \lambda_i/(\lambda_i+\rho)$, $\gamma_i = \lambda_i+\rho$, $i = 1, 2$, λ_1 and λ_2 are the past and current SCR under the restriction of $\lambda_2 < \lambda$. Similar argument can be applied to the superinfection model, leading to the following seroprevalence:

$$p_{S^+}(t) = \begin{cases} 1-e^{-(\lambda_1/\rho)(e^{-\rho\tau}-e^{-\rho t})-(\lambda_2/\rho)(1-e^{-\rho\tau})}, & \text{if } t>\tau \\ 1-e^{-(\lambda_2/\rho)(1-e^{-\rho t})}, & \text{if } t\le\tau. \end{cases} \quad (11)$$

With respect to parameter estimation, the sampling distribution is again assumed to be a Binomial-product distribution ((9), where $\pi(t)$ is described by (10) or (11)). To estimate all parameters (λ_1, λ_2, ρ, and τ) via maximum likelihood, a profile likelihood approach is usually applied to the data under analysis: (i) set $\tau = 1$, (ii) determine the respective maximum likelihood estimates for the remaining parameters, (iii) calculate the corresponding log-likelihood function, (iv) increase one unit to τ and repeat steps (ii-iii), and (v) keep increasing τ until reaching the maximum expected value for that parameter. The overall maximum likelihood estimates are those associated with the value of τ that provides the maximum value of all log-likelihood values. Although statistically sound, this method tends to overestimate the true change point (i.e., estimates located further in past than they should), even if using a large sample size (our own results). This suggests that seropositivity data might not have enough information to estimate that parameter with high precision. Therefore, the interpretation of a specific estimate for the reduction time point should be done with caution. In practice, models assuming a stable or an abrupt reduction in malaria transmission intensity must compare to each other for the same data. A log-likelihood ratio test can then be applied to the corresponding results using the following test statistic under the null hypothesis:

$$L = (-2)\times\left(\Lambda_{\text{stable}}-\Lambda_{\text{reduction}}\right) \rightsquigarrow \chi^2_{(2)}, \quad (12)$$

where Λ_{stable} and $\Lambda_{\text{reduction}}$ are the log-likelihood functions evaluated at the maximum likelihood estimates for the models assuming a stable or an abrupt reduction in malaria transmission intensity, respectively, and $\chi^2_{(2)}$ is a Chi-square distribution with the two degrees of freedom resulting from the difference in the total number of parameters of the models under testing (λ and ρ versus λ_1, λ_2, ρ, and τ). For a 5% significance level, p values < 0.05 show evidence for a significant change in disease transmission.

With the increasing complexity of the models under analysis, statistical inference via maximum likelihood methods becomes more cumbersome due to possible lack of convergence of the numerical algorithms leading to maximum likelihood estimates [15] and the inaccuracy of large sample approximations for the confidence intervals and test statistics [34]. These problems can be surpassed by using Bayesian inference. In this approach, each parameter in a model has an associated prior distribution that, in turn, is updated with the observed data by means of Bayes theorem. The resulting distribution is in the core of Bayesian inference and called posterior distribution. Posterior mean and median of this

Table 2: Maximum likelihood estimates for seroconversion and seroreversion rates (SCRs and SRRs, resp.) of antibodies against AMA1 expected for northwest and northeast regions of Bioko Island using the reversible catalytic and superinfection models (RCMs and SIM, resp.) under the assumptions of constant malaria transmission intensity over time and an abrupt reduction in malaria transmission at a given change time point before data collection.

Region	Model	Malaria transmission	SCR (95% CI)	SRR (95% CI)	log-likelihood
Northwest	RCM	Constant	0.286 (0.249, 0.328)	0.008 (0.005, 0.015)	−71.31
	SIM	Constant	0.359 (0.307, 0.419)	0.091 (0.069, 0.120)	−69.71
Northeast	RCM	Constant	0.124 (0.109, 0.141)	0.006 (0.004, 0.011)	−96.87
	SIM	Constant	0.139 (0.119, 0.163)	0.039 (0.028, 0.056)	−102.95
	RCM	Abrupt reduction (change point = 6)	0.274 (0.200, 0.376) 0.077 (0.058, 0.100)	0.009 (0.005, 0.014)	−84.25
	SIM	Abrupt reduction (change point = 6)	0.900 (0.431, 1.879) 0.098 (0.075, 0.129)	0.150 (0.097, 0.232)	−83.37

distribution are two possible Bayesian estimates for the same parameter. Credible intervals are the Bayesian equivalent to the confidence intervals of classical statistics and calculated by the appropriate quantiles of the posterior distribution that ensure a given probability mass (i.e., 95%). Model comparison can be performed via AIC or other Bayesian information measures, such as the Deviance Information Criterion (DIC) [44]. Theoretically, DIC is defined by the posterior mean of the deviance function (twice the absolute value of log-likelihood function) plus the effective number of parameters of a given model. In turn, the effective number of parameters is calculated by the difference between the posterior mean of the deviance function and the same function evaluated at the posterior means of the parameters. Likewise with AIC, one should choose the model that shows the lowest DIC value among all models tested.

In general, there are two major difficulties in performing Bayesian analysis. The first one relates to how to choose the prior distributions for the unknown parameters. One solution is to use noninformative prior distributions in situations where prior information about the parameters of interest is limited or scarce. Popular choices for noninformative prior distributions are the uniform distribution for parameters defining probabilities or Gaussian distributions with mean 0 and sufficiently large standard deviation for parameters defined in real space. In contrast, if one has strong prior beliefs about the parameters of interest, informative prior distributions can be elicited. Prior elicitation is generally based on a convenient probability distribution (e.g., a Gaussian distribution) upon which one determines the corresponding prior parameters—the so-called hyperparameters—by conjugating the expected prior mean with a set of prior quantiles set for that distribution. Although informative prior distributions are in line with the permanent dialogue between inductive and deductive reasoning intrinsic to the scientific method, most researchers adopt a conservative strategy to data analysis by using noninformative prior distributions for the unknown parameters. The second difficulty concerns the calculation of the posterior distributions. However, this is greatly reduced by the powerful Markov Chain Monte Carlo (MCMC) that, virtually, can deal with any kind of model complexity. In practice, R/Jags is an easy-to-use package for MCMC computing. Illustrative scripts for the above RCM and SIM are available from the authors upon request.

Example I (Bioko Island continued). As highlighted earlier, the fits of RCM and SIM assuming stable malaria transmission intensity suggested a variation in malaria risk between younger and older individuals living in the northeast region of the island (Figure 2(d)). Such variation might be attributed to a reduction in malaria transmission intensity owing to a known malaria control initiative launched in 2004. To test this hypothesis, RCM and SIM with an abrupt reduction in malaria transmission intensity were fitted to the data via maximum likelihood estimation. The most likely reduction point for both models was 6 years before data collection (Figure 3(a)); the corresponding 95% confidence intervals were 4.2–8.4 and 4.8–7.7 for RCM and SIM, respectively. Both models were in close agreement with the data visually (Figure 3(b)) and better than the previous ones assuming stable transmission, according to likelihood ratio test (p values < 0.001). SIM led to a higher log-likelihood value than its RCM counterpart (Table 2) and, thus, it might be deemed the best model for the data. Again, this result is consistent with the choice of three-component Gaussian mixture model for the corresponding titre data. Previous and current SCRs were estimated at 0.900 and 0.098 for SIM and at 0.274 and 0.077 for RCM. These implied a reduction in malaria transmission intensity of around 89% and 72% for SIM and RCM, respectively. Note the putative overestimation of the time point for the reduction event (6 years before sampling versus the time when the Bioko malaria control initiative started). This result is in line with ongoing research where the profile likelihood method overestimated the true change point from simulated data typically found in African population (our own results).

2.3.2. Detecting Changes in Malaria Exposure due to Age-Dependent Behaviors. A very similar age-adjusted seroprevalence curve to previous case can be found for populations where older individuals have a higher malaria transmission intensity compared to the one for younger individuals due to an age-dependent behavior factor. A typical example is the commute of adults to working sites that are malaria transmission hotspots in contrast to children and adolescents

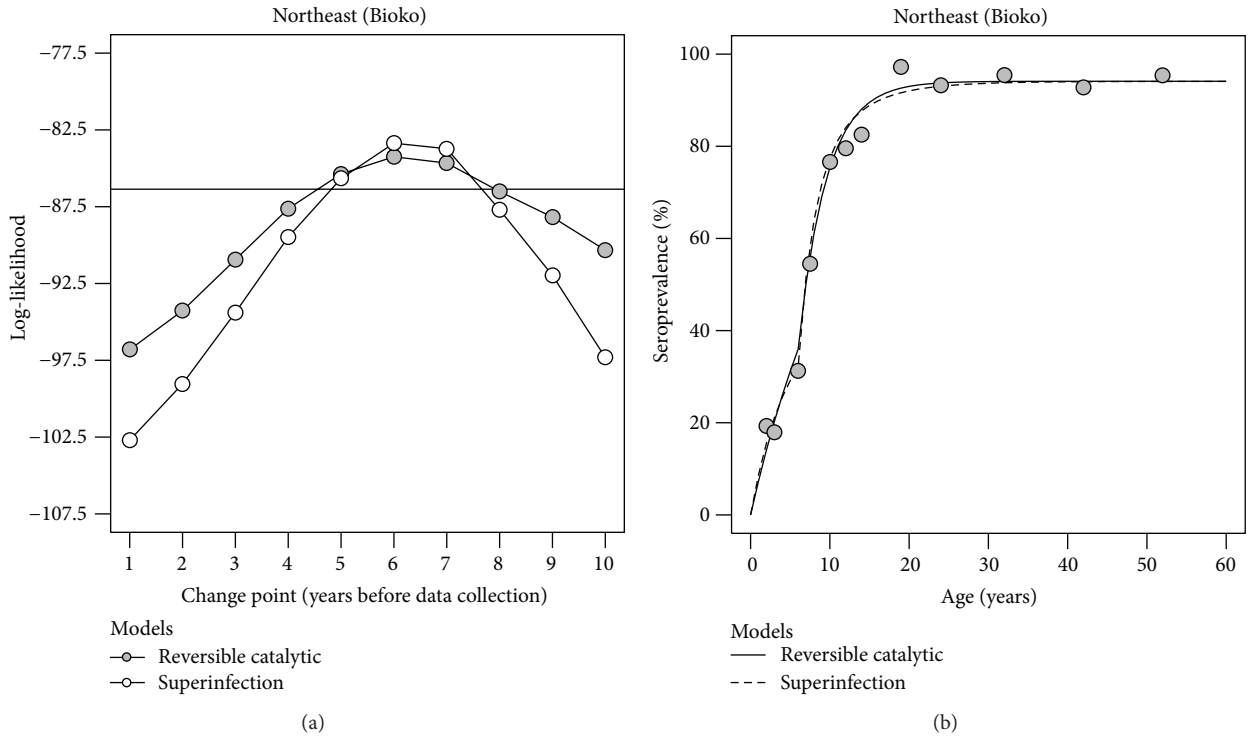


Figure 3: AMA1 seropositivity data analysis of northeast region from Bioko Island under the assumption of a past abrupt reduction in malaria transmission intensity. (a) Profile likelihood plot to estimate the best change point for the reversible catalytic model, where the solid and dashed lines refer to the log-likelihood value for the model assuming a stable transmission intensity and the cut-off value accepting that model at a 5% significance level, respectively. (b) Maximum likelihood fits of the reverse catalytic and superinfection models assuming an abrupt reduction in malaria transmission intensity estimated to have occurred 6 years before sampling.

who do not travel to those sites. This situation was reported for some populations living in the forests of Cambodia and Indonesia [41, 45]. The above RCM and SIM are easily translated to this new situation. More precisely, both younger and older individuals share the same SCR until a certain age. Then SCR abruptly increases to a new level in a similar way as previous case. Therefore, (10) for an abrupt reduction in malaria transmission intensity can be adapted as follows:

$$p_{S^+}(t) = \begin{cases} \theta_1\left(1-e^{-\gamma_1\tau}\right)+\theta_2\left(1-e^{-\gamma_2(t-\tau)}\right)e^{-\gamma_1\tau}, & \text{if } t>\tau \\ \theta_1\left(1-e^{-\gamma_1 t}\right), & \text{if } t\le\tau, \end{cases} \quad (13)$$

where $\theta_i = \lambda_i/(\lambda_i+\rho)$, $\gamma_i = \lambda_i+\rho$, $i = 1, 2$, λ_1 and λ_2 are the SCR for younger and older individuals, respectively, under the restriction of $\lambda_1 < \lambda_2$. For the superinfection assumption, the resulting model can be expressed as follows:

$$p_{S^+}(t) = \begin{cases} 1-e^{-(\lambda_2/\rho)(e^{-\rho\tau}-e^{-\rho t})-(\lambda_1/\rho)(1-e^{-\rho\tau})}, & \text{if } t>\tau \\ 1-e^{-(\lambda_1/\rho)(1-e^{-\rho t})}, & \text{if } t\le\tau. \end{cases} \quad (14)$$

Parameter estimation and model comparison can be performed via maximum likelihood and Bayesian methods as described above for the models with an abrupt change in malaria transmission intensity.

Example II (Jacareacanga, Brazil). A recent study was conducted in the Brazilian Amazonia region [7] where *P. vivax* is currently the major malaria threat opposed to what occurred in the past where *P. falciparum* infections predominated. A total of around 1300 individuals were sampled from 7 different municipalities in Pará state. Previous analysis suggested stable malaria transmission for *P. vivax* infections but detected a putative abrupt reduction of *P. falciparum* transmission intensity estimated to have occurred around 25–30 years before sampling. Although this change is in line with the intensification of malaria control initiatives by Brazilian health authorities in the area, alternative explanations were also discussed but not formally tested. More precisely, gold mining is one of the key economic activities in the area but also an important risk factor for malaria transmission. Mining was also a determinant factor of the known migration wave from nonendemic states to the region since 1970s. In this line of thought, the detection of a change occurred 25–30 years before sampling might be confounded by the increased malaria risk of the older population that are typically miners. This hypothesis is now tested against the one assuming an abrupt reduction of malaria transmission intensity. The analysis is focused on the *P. falciparum* seropositivity data from the municipality of Jacareacanga where the past reduction in SCR seemed more pronounced (i.e.,

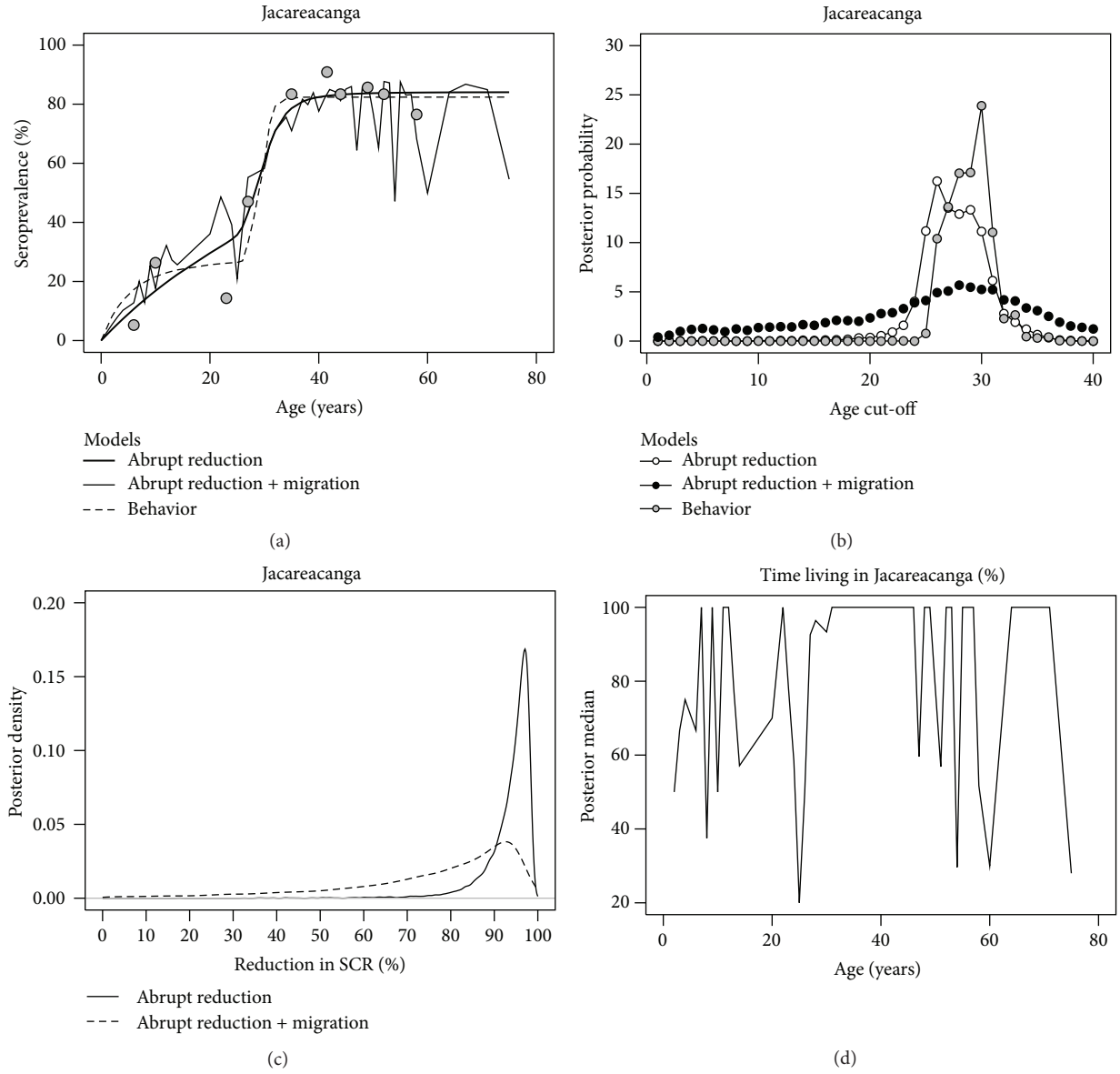


Figure 4: Analysis of *P. falciparum* seropositivity data from Jacareacanga (Brazil) using Bayesian methods. (a) Seroprevalence curves as predicted by RCMs assuming an abrupt reduction in SCR with and without migration and assuming a behavioral factor dependent on a given age cut-off, where dots represent the observed seroprevalence for age groups by splitting the age distribution in deciles. (b) Posterior distributions for the age cut-off for the models mentioned in (a). (c) Posterior probability densities for the reduction in SCR assuming or not migration effects. (d) Posterior median for the fraction of time living in the area in relation to the corresponding age of the individuals, as expected from RCM assuming migration effects and an abrupt reduction in SCR.

from 0.514 to 0.017 [7]). Data under analysis comprised a total of 172 individuals of which 2.3% were infected with malaria parasites at the day of the survey. The seroprevalence for any *P. vivax* and *P. falciparum* antigens was estimated at 69.2% and 59.3%, respectively, using a two-component Gaussian mixture model for the corresponding titre data. These estimates suggested a high malaria endemicity for that municipality as issued by the Brazilian authority for malaria control but using the recorded annual parasite index. In contrast to previous example, Bayesian methods were alternatively applied to the data using the following noninformative prior distributions for the parameters: (i) Gaussian distributions with mean 0 and standard deviation 103 for all SCRs and SRRs in log scale and (ii) a discrete uniform distribution between 1 and 40 for the age cut-off. Since seropositivity data was previously derived from a two-component Gaussian mixture model, this analysis is based on the RCM only. MCMC simulation via R/Jags package was used to obtain the posterior estimates for the parameters; a long chain of 1,050,000 iterations was generated where the first 50,000 were discarded as the burn-in period and a lag of 100 was used to reduce correlation between simulated values. As previously reported, the model assuming an abrupt reduction in SCR captures data well (Figure 4(a)). However, there was some

degree of uncertainty on the time in which that reduction had occurred (Figure 4(b)). The posterior mean and median were consistent with a sudden drop in *P. falciparum* transmission intensity 28 years before data collection (e.g., around 1980). The posterior mean for past and current SCR was 0.436 and 0.019, respectively, while the corresponding posterior medians were 0.386 and 0.019 (Table 2). In agreement with a Bayesian analysis using noninformative prior distribution, these posterior estimates implied a reduction in SCR in the same order of magnitude to that obtained in the original study. The model assuming a behavioral factor also fitted the data well (Figure 4(a)) with a slightly higher age cut-off for the occurrence of such behavior (around 29 years old). Again, there was some uncertainty associated with the age value where that behavior becomes epidemiologically relevant. In absence of that putative behavioral factor, the baseline SCR was estimated at 0.051 or 0.046 if one chooses the posterior mean or median, respectively (Table 2). This SCR increased to the posterior mean and median of 0.654 and 0.693 at older ages. DIC was then used to compare both models. The respective DIC estimates are 89.72 and 96.67 for the RCMs assuming an abrupt reduction and a change in SCR due to an age-dependent behavioral factor. Since the best model is the one that shows the lowest value for DIC, the change in SCR seemed to be better explained by an abrupt reduction in *P. falciparum* transmission intensity rather than the existence of a putative risk factor dependent on age, such as those related to gold mining activities in the heart of the Amazonia forest.

2.4. Detecting Migration Effects on Malaria Exposure. Up to now all models were analyzed under the assumption of stable populations with no migration effects. This assumption is reasonable in most studies because individuals living shortly in a given study area are typically excluded from the survey. However, in the current era of facilitated movement between populations, it might be difficult to recruit locally born individuals only, thus, affecting the estimation of the SCR. In one extreme, the easiest migration setting is the importation of malaria cases to nonendemic regions where there are no sufficient conditions for efficient malaria transmission. In this case, there is no strong rationale to use any of the above models since SCR would reflect the disease transmission intensity of the places where the sampled individuals come from. On the other extreme, migration from nonendemic region to endemic ones might introduce bias on SCR estimates. More precisely, at the time of migration, individuals are immunologically naive to malaria parasites in comparison to those with the same age but locally born in the region. This is the case of the history of malaria in Brazil where a gold rush in 1970s led to the migration of thousands of people from nonendemic states to the heart of the Amazonia forest [42]. Such migration caused an increase in population size and malaria cases in the region. Another known example from the literature is the founder effect in a Peruvian community affected by Chagas disease where locally born individuals and founders have distinct seroprevalence histories [43].

Until now age was considered a proxy of the total exposure time of each individual to malaria antigens. In the situation where individuals migrated from a nonendemic region to an endemic one, age used in all above models is simply replaced with the total residence time of each individual in the endemic area if available. In practice, such information is not routinely collected, thus, requiring additional estimation. Without lack of generality, let us focus on the simple RCM with stable malaria transmission intensity. Similar argument can be applied to the remaining models. As seen earlier, the expected seroprevalence curve is given by (6). The same model including migration effects is described as follows:

$$p_{S^+}(t) = \frac{\lambda}{\lambda + \rho}\left(1 - e^{-(\lambda+\rho)t^*}\right), \tag{15}$$

where t and t^* are the age and the total residence time of an individual living in an endemic area, respectively. In absence of information of the total residence time, the estimation of t^* can be done by considering $t^* = t \times p_t$, where $p_t \in (0, 1]$ is the proportion of residence time of individuals with age t. In practice, estimation of each p_t might be cumbersome by maximum likelihood methods. Firstly, the above RCM and SIM are intrinsically nonlinear and these additional unknown parameters might lead to convergence problems of the respective maximization algorithms. Secondly, sample information might be insufficient to provide accurate estimation of the residence time of each individual. Alternatively, Bayesian methods can overcome some of these limitations. As mentioned earlier, Bayesian inference is nowadays facilitated by the availability of powerful MCMC simulators that can estimate any kind of statistical model. Moreover, Bayesian inference can also coherently integrate external information on the residence time by describing the prior distribution accordingly. In so doing, one can consider the following family of prior distributions for p_t:

$$P[p_t = x] = \begin{cases} p_0, & \text{if } x = 1, \\ (1 - p_0)\,\dfrac{x^{\alpha_t - 1}(1 - x)^{\beta_t - 1}}{\mathrm{Be}(\alpha_t, \beta_t)}, & \text{if } 0 < x < 1, \end{cases} \tag{16}$$

where p_0 is the prior probability of an individual with age t being locally born and p_t for a migrant is modeled a priori by a Beta distribution with hyperparameters α_t and β_t. If little information is known about the migrant population, one can specify $\alpha_t = 1$ and $\beta_t = 1$ in order to obtain uniform distribution. Note that the parameter p_t is a priori allowed to vary with age. This is particularly useful to capture migration effects of specific age groups, such as adults who tend to migrate for work reasons.

Example II (Jacareacanga, Brazil, continued). As mentioned above, the history of malaria in Brazil is intimately related to a gold rush in 1970s from nonendemic regions to endemic ones in the heart of the Amazonia forest [42]. Since mining is the main economic activity of Jacarecanga, it is possible that the above past and current SCR estimates can be improved in order to take into account any past migration effects. Unfortunately data concerning time of residence were not consistently recorded across individuals and study sites

Table 3: Bayesian analysis of *P. falciparum* seropositivity data from Jacearecanga where RCM_{red}, $RCM_{red+mig}$, and $RCM_{behavior}$ denote the reversible catalytic models assuming an abrupt reduction in SCR only, an abrupt reduction together with migration effects, and a change in SCR due to a behavioral factor dependent on a given age cut-off.

Model	Parameter	Posterior estimates		
		Mean	Median	95% credible interval[a]
RCM_{red}	Past SCR	0.436	0.386	0.099–0.948
	Current SCR	0.019	0.019	0.009–0.033
	Time elapsed since reduction	27.6	28.0	22.0–33.0
$RCM_{red+mig}$	Past SCR	0.292	0.192	0.052–0.916
	Current SCR	0.038	0.037	0.013–0.067
	Time elapsed since reduction	24.5	26.0	4.0–39.0
$RCM_{behavior}$	Baseline SCR	0.051	0.046	0.019–0.106
	Risk SCR	0.654	0.693	0.153–0.988
	Age cut-off	28.9	29.0	26.0–33.0

[a]Credible interval based on 2.5% and 97.5% quantiles of the respective posterior distribution.

in the original study and, thus, the parameters p_ts were directly estimated from seropositivity data. The above data analysis was then extended to the situation of RCM assuming an abrupt reduction in the disease transmission intensity together with putative migration effects described by (10). Little information was known about the migrant population and, thus, a prior uniform distribution for the parameters p_t's was specified for the Bayesian analysis. With respect to the prior probability p_0, the Brazilian Office for Geography and Statistics states that 14.4% of the population living in Jacareacanga in the 2010 census were not born in the north states comprising the Amazonia region [46]. Therefore, it seemed unlikely that the percentage of the migrant population from Jacareacanga was lower than 15%. To understand the impact of p_0 on the subsequent inferences, different values for that hyperparameter were tested, specifically, 0.25, 0.50, 0.75, and 0.9. The best one appeared to be 0.75 because it implied the highest posterior median and mean of the log-likelihood function (results not shown). The introduction of migration effects in the RCM with an abrupt reduction resulted in a seroprevalence curve with a more complex pattern (Figure 4(a)). However, this higher complexity in the seroprevalence curves augmented the uncertainty associated with the posterior distributions of the time in which that reduction in SCR actually occurred and of the ratio between current and past SCR (Figures 4(b) and 4(c)). Adjusting for putative migration effects, the posterior median and mean for the changing point are around 24.5 and 26 years before sampling, two estimates close to the previous ones assuming no migration (around 28 years; Table 3). For the reduction in SCR itself, the respective posterior mean and medians are 75.8% and 75.9%, two estimates slightly more conservative than those obtained for the RCM with no migration effects (93.5% and 95.3%, resp.). Finally, notwithstanding the limited sample size, it was possible to borrow information from the sample in order to update the prior distributions of the residence time of each individual (Figure 4(d)). Many individuals could be assumed as locally born in Jacareacanga because the respective posterior median for the fraction of time living in area was close to 100% (Figure 4(d)). In the remaining cases, there was evidence for the presence of migrants in the sample. In conclusion, although model complexity increased uncertainty of the subsequent parameter estimation, the results provided a more realistic snapshot of the *P. falciparum* malaria history of Jacareacanga.

2.5. Detecting Individual Level Heterogeneity in Malaria Exposure. All above models for seropositivity data provide a broad description of the SCR at the population level. Their utility is then limited if one aims to understand more granular, individual level heterogeneities inherent to malaria transmission. For example, a recent study from Cambodia has highlighted the role of age, ethnicity, village of residence, or forest work on the seroconversion of each individual during rainy season [45]. Other examples are the effect of elevation in SCR in northeast Tanzania [12, 13] or the impact of malaria control interventions in western Kenyan highlands [22]. Although age is an intrinsic variable of the above RCM and SIM, the effect of other types of covariates affecting seropositivity suggests adopting a regression-type approach to tackle putative individual level heterogeneity in SCR. This can be easily done by considering the following log-linear model for the SCR of the ith individual with a set of p covariates $x_{1,i}, x_{2,i}, \ldots, x_{p,i}$:

$$\log \lambda_i = \beta_0 + \beta_1 x_{1,i} + \beta_2 x_{2,i} + \cdots + \beta_p x_{p,i}, \qquad (17)$$

where β_0 is the overall effect in absence of covariates and $\beta_1, \ldots, \beta_p$ are the regression coefficients associated with each covariate. This regression model is then coupled with RCM or SIM (see (6) and (7)) with stable transmission intensity but describing λ with the above model. In the unrealistic situation that malaria infections induce lifelong immunity (see (8)), the inclusion of covariates is facilitated because the resulting model is integrated in the well-known generalized linear model framework via a complementary log-log model for binary variables [47].

Since the analysis must take into account the data from each individual, previous Binomial-product for the sampling

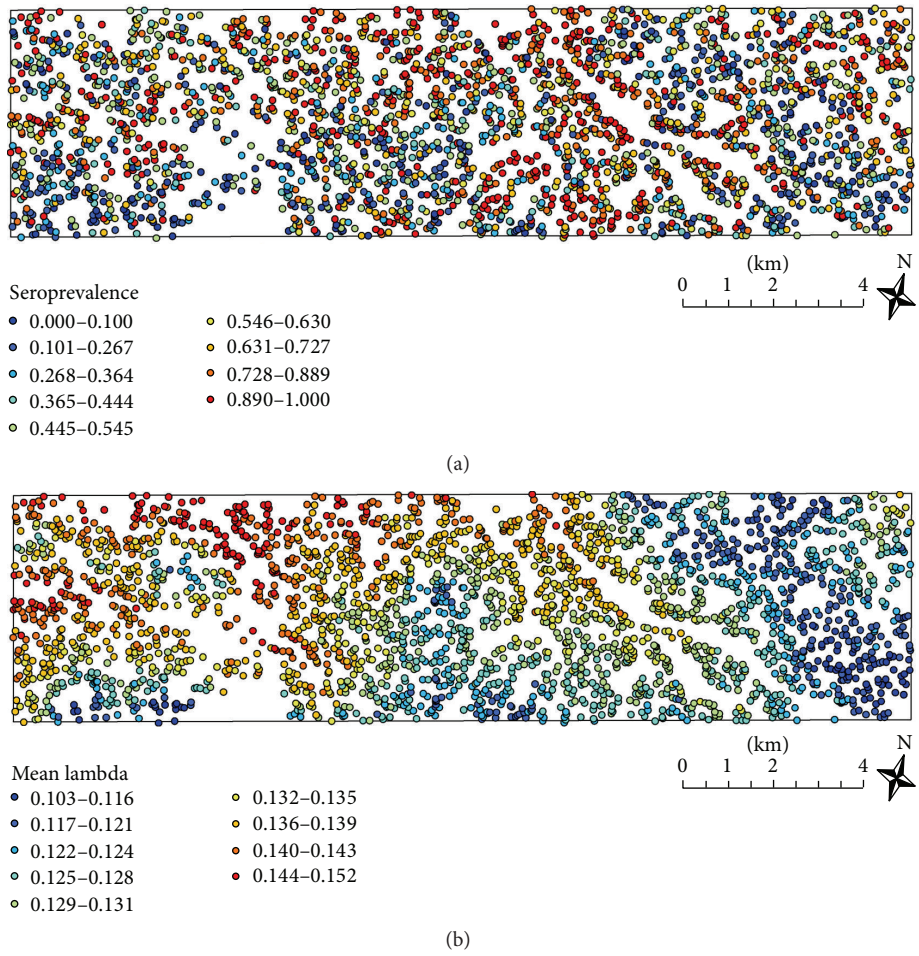


Figure 5: Maps of the western Kenyan highlands showing the distribution of the surveyed households and household level exposure. (a) Map based on the combined seroprevalence for AMA1 and MSP1 antigens. (b) Map based on the posterior mean of SCR adjusting for variations in elevation and gender and use of mosquito control interventions. Each household is represented by a circle and the shading shows the intensity of malaria exposure from blue (low) to red (high).

distribution (see (9)) is now reconverted into a Bernoulli-product, one Bernoulli distribution per individual; that is,

$$f\left(\{y_i\} \mid \{\pi_i\}\right) = \prod_{i=1}^{n} \pi_i^{y_i} \left(1 - \pi_i\right)^{1-y_i}, \tag{18}$$

where y_i is the serological status of the ith individual, π_i is the probability of the ith individual being seropositive, and n is the sample size.

In theory, maximum likelihood and Bayesian methods can be applied to estimate all unknown parameters of the above model. In practice, computationally efficient maximum likelihood methods still need to be developed and, therefore, Bayesian methods via MCMC are the most pragmatic approach to data analysis. In absence of prior information about the regression parameters, the usual choice for the respective a priori distribution is to use Gaussian distribution with mean 0 and a sufficiently large standard deviation (e.g., 100). If any prior information exists for the regression parameters, one can alternatively use any eliciting method for Bayesian regression analysis as described elsewhere [48, 49].

Example III (Rachuonyo South, Kenya). The western Kenyan highlands are currently characterized by low-level endemic and highly heterogeneous *P. falciparum* malaria transmission. To ensure high resolution to detect heterogeneity in malaria exposure, approximately one-third of the total population, around 17,500 individuals, were sampled from a 100 km^2 area in Rachuonyo South in the western Kenyan highlands [22]. The analysis was focused on *P. falciparum* seropositivity data from about 13,000 individuals with complete data. Combined seropositivity for AMA1 and MSP1 antigens was calculated using the two-component Gaussian mixture model approach for determining seropositivity to each antigen. RCM with stable malaria transmission was then fitted to the data using maximum likelihood methods. The overall SCR was estimated at 0.132, suggesting an overall seroprevalence of 55.2%. When the observed seropositivity of each individual was aggregated to the household level and plotted on a map, there is a substantial variation within the study area (Figure 5(a)).

However, the large amount of variation renders it difficult to delineate hotspots of seroprevalence. With this in mind, the previous analysis was then refined in order to take into account available information on gender, elevation, residing or not in a house that received indoor residual spraying in the previous 12 months, and sleeping or not under a bednet the previous night. Similar to the Jacareacanga example, Bayesian methods were applied to the data using noninformative prior distributions for the regression coefficients and SRR. Since a two-component Gaussian mixture model was used for determining seropositivity to each antigen, this extended analysis focused on the RCM model given by (6) where SCR was described by a log-linear regression model including the above-mentioned covariates. Posterior estimates highlighted a significant role of elevation on SCR while the remaining covariates, despite explaining some individual variation in seropositivity, were not statistically significant in the regression model (results not shown) but were maintained due to their known impact on malaria. A new map based on the posterior means of SCR for each individual aggregated to the household level was then generated (Figure 5(b)). This map suggested that significant variation in SCR exists within this 100 km^2 study area and identifies households with high SCRs. The identification of these putative hotspots of exposure may be instrumental to design future interventions in the study area.

2.6. Antibody Acquisition Models. In all models described above, the information on antibody titres is reduced to the proportions of seropositive and seronegative. Alternatively one can analyze data of antibody titres themselves using the antibody acquisition models [50, 51]. In these models, one assumes that the rate at which antibody levels are acquired can be used as a marker for transmission intensity. If an individual's antibody level A is boosted at rate $\alpha(t)$ and decays at rate r then antibody levels can be described by the following ordinary differential equation [51]:

$$\frac{dA}{dt} = \alpha(t) - rA. \tag{19}$$

When malaria transmission is constant over time, the same is assumed for the rate of generation of antibodies in response to infection; that is, $\alpha(t) = \alpha$. Under the initial condition of $A(0) = 0$, the above equation can be solved analytically to give

$$A(t) = \frac{\alpha}{r}\left(1 - e^{-rt}\right), \tag{20}$$

where t is again regarded as the age of an individual at data collection. The above model can be also extended to include the effect of maternal antibodies [51].

Likewise for seropositivity-based models, historical changes in malaria transmission intensity can also be accounted for. For example, if there was an abrupt reduction in transmission τ years before data collection such that the rate of acquisition of antibodies changed from α_1 to α_2, then the expected antibody titre of an individual with age t is

$$A(t) = \begin{cases} \dfrac{\alpha_2}{r}\left(1 - e^{-rt}\right), & t \le \tau, \\ \dfrac{\alpha_1}{r}\left(1 - e^{-r(t-\tau)}\right)e^{-rt} + \dfrac{\alpha_2}{r}\left(1 - e^{-r\tau}\right), & t > \tau. \end{cases} \tag{21}$$

The above equation is explained as follows. For individuals born after the change in transmission ($t \le \tau$), the expected antibody dynamics follow exactly as in the constant transmission scenario but with boosting rate α_2 (see (20)). For the individuals born before the change in transmission, one can partition the antibody levels into two terms, the first one referring to the expected antibody levels produced until the change point with boosting rate α_1 discounted by an exponential decay with rate r until present time and the second one referring simply to the antibody counts expected to be produced since the change point.

Equations (20) and (21) provide expressions for an individual's antibody titre as a function of age. However, in a population of individuals there is likely to be substantial variation in antibody titres. As mentioned earlier for seropositivity determination, antibody titre data are often approximately Gaussian distributed on a log scale. Therefore, when constructing the sampling distribution for the comparison of the antibody acquisition model with data, one can assume that at age t antibody data is log-Normally distributed with parameters $\mu = \log(A(t))$ and σ. Note that in this interpretation $A(t)$ is the geometric mean titre (GMT) at age t (corresponding to the mean on a log scale). For a random sample of n individuals, the respective sampling distribution is given by

$$f\left(\{x_i\} \mid \theta, n\right) = \prod_{i=1}^{n} \frac{1}{x_i \sigma \sqrt{2\pi}} e^{-(\log x_i - \log A(t_i))^2 / 2\sigma^2}, \tag{22}$$

where x_i and t_i are the antibody titres and age of the ith individual, respectively, and θ is the parameter vector associated with antibody acquisition model under analysis. This parameter vector is given by $\theta = (\alpha, r, \sigma)$ or $\theta = (\alpha_1, \alpha_2, \tau, r, \sigma)$ if fitting the model with constant malaria transmission intensity or fitting the model with an abrupt reduction in transmission, respectively. Since the above models are nonlinear, parameter estimation can be performed by nonlinear least squares available for the R software or by Bayesian methods via MCMC. For cases where the data are not well described by a log-Normal distribution, alternative sampling distributions will need to be constructed. For example, if a proportion of the population has never been exposed to malaria (i.e., their antibody titres are just background responses), then a zero-inflated log-Normal distribution could be used as an alternative sampling model. Statistical methods to fit that distribution to data can be found elsewhere [52–54].

Example 1 (Bioko Island continued). To extend previous analysis based on seropositivity data, antibody titre data from northeast and northwest regions of Bioko Island were alternatively analyzed using the above antibody acquisition models.

Table 4: Parameter estimates for antibody acquisition models applied to anti-AMA1 antibody titre data (AU: arbitrary units) from northwest and northeast region of Bioko Island. Estimates are presented as posterior medians with 95% credible intervals in brackets.

Region	Malaria transmission	α_1	α_2	τ	r	σ
Northwest	Constant	60.4 (50, 65)	—	—	0.098 (0.07, 0.12)	1.36 (1.32, 1.41)
Northeast	Constant	20.2 (18, 23)	—	—	0.053 (0.04, 0.07)	1.36 (1.31, 1.42)
	Drop	128 (65, 232)	20 (17, 23)	7.2 (6.2, 8.7)	0.16 (0.11, 0.21)	1.33 (1.28, 1.39)

The respective data is shown in Figures 6(a) and 6(b). To compare with previous results, the above antibody acquisition models assuming a constant malaria transmission intensity (Figures 6(c) and 6(d)) and an abrupt reduction in malaria transmission (Figures 6(e) and 6(f)) were fitted to each data set separately. Again, there was evidence for a constant malaria transmission intensity in the northwest region of the island (Figure 6(c)) with an average increase of antibody titres of around 60.4 units per year of exposure (Table 4). In contrast, the antibody acquisition model with constant transmission intensity showed some fitting deficiencies at younger ages for the northeast region (Figure 6(d)), which were eliminated by assuming a drop in malaria transmission intensity (Figure 6(f)). That drop in malaria transmission intensity seemed to have occurred 7 years before sampling, an estimate consistent with the one obtained from seroprevalence data (6 years before sampling; Table 2). According to posterior estimates in Table 4, the average value of antibody acquisition per year decreased from 128 to 20. These estimates suggested a reduction of 84% in malaria transmission intensity, which is in close agreement with a reduction in SCR of 89% estimated from the superinfection model (Table 2).

3. Envisioning the Future: Serology and Malaria Elimination

Malaria eradication and elimination are currently in the agenda of various countries worldwide, such as Sri Lanka [3] or Haiti [4]. With this mind, an important question naturally arises: How can one declare if elimination or eradication was actually achieved? Again, serology can help answering this question due to its capacity to detect recent malaria exposure in an apparently asymptomatic population.

As discussed above, the first step of serology data analysis is typically to determine seropositivity from titre data. However, in a malaria elimination setting, seroprevalence is supposedly low in the population, thus, making it difficult to discriminate whether the data comes from a single Gaussian distribution or from a Gaussian mixture model. In this context, the presence of a single Gaussian distribution in the data can be interpreted as indicative of a seronegative population only, thus, suggesting malaria elimination (or eradication). On the other hand, the detection of a mixture distribution indicates the presence of seropositive individuals that might be on their way to seronegativity but also might have been exposed to malaria after a putative elimination event. Hence, sample size determination before data collection ensures to some extent the accuracy of the findings. However, sample size determination is theoretically challenging in the context of mixture distributions because standard asymptotic theory for hypothesis testing does not hold. This implies using less-known statistical methods, such as the bootstrap approach proposed by McLachlan [55] or the adjusted likelihood ratio test derived by Lo et al. [56]. Limited sample size guidelines exist for these alternative methods and, therefore, future research is needed to better help designing malaria surveillance based on serology data.

Under the assumption of detecting a mixture distribution in the antibody titre data, it was shown above how to infer different sources of heterogeneity in malaria transmission intensity as long as the correct antibodies are analyzed. In this regard, recent research produced a list of putative antigens that can be used in malaria elimination studies for being highly informative on the time to most recent infection [20]. Although malaria elimination and eradication are more likely to be achieved by a slow decrease of malaria transmission intensity towards 0, a reasonable analytical starting point is to aim at using the above RCM and SIM under the assumption of an abrupt reduction in SCR at a given time point before sampling collection. In a malaria elimination setting, current SCR of these models is set at 0 while past SCR should be estimated from the data. The corresponding age-adjusted seroprevalence curve shows a distinctive pattern where children born after the time of malaria elimination are all seronegative while the remainder might or not be seropositive depending on the past malaria exposure before malaria elimination (Figure 7(a)). These latter individuals would slowly revert to a seronegative state over time. For certification purposes, these models must be compared to the ones assuming a (very low) stable malaria transmission intensity for the same data. Therefore, one needs to understand whether data has enough statistical power for detecting disease elimination. To overcome eventual problems later in a study, it is then recommended to perform sample size calculation before collecting data. This problem has recently been tackled for estimating the SCR in stable malaria transmission intensity settings [34]. General guidelines for sample size determination are difficult to put forward because estimation precision and statistical power are intimately related to the age distribution associated with a given study design. In this regard, African studies are more facilitated than those conducted elsewhere because the age distribution tends to be consistent across populations with a decreasing trend from newborns to elders (Figure 7(b)). If such distribution is combined with the expected age-adjusted seroprevalence, one can have an idea of the evolution of overall seroprevalence over time (Figure 7(c)). Since age distribution can vary from one study to another, the minimum sample size for detecting malaria elimination is better calculated by means of data simulation using the expected age distribution for the sample

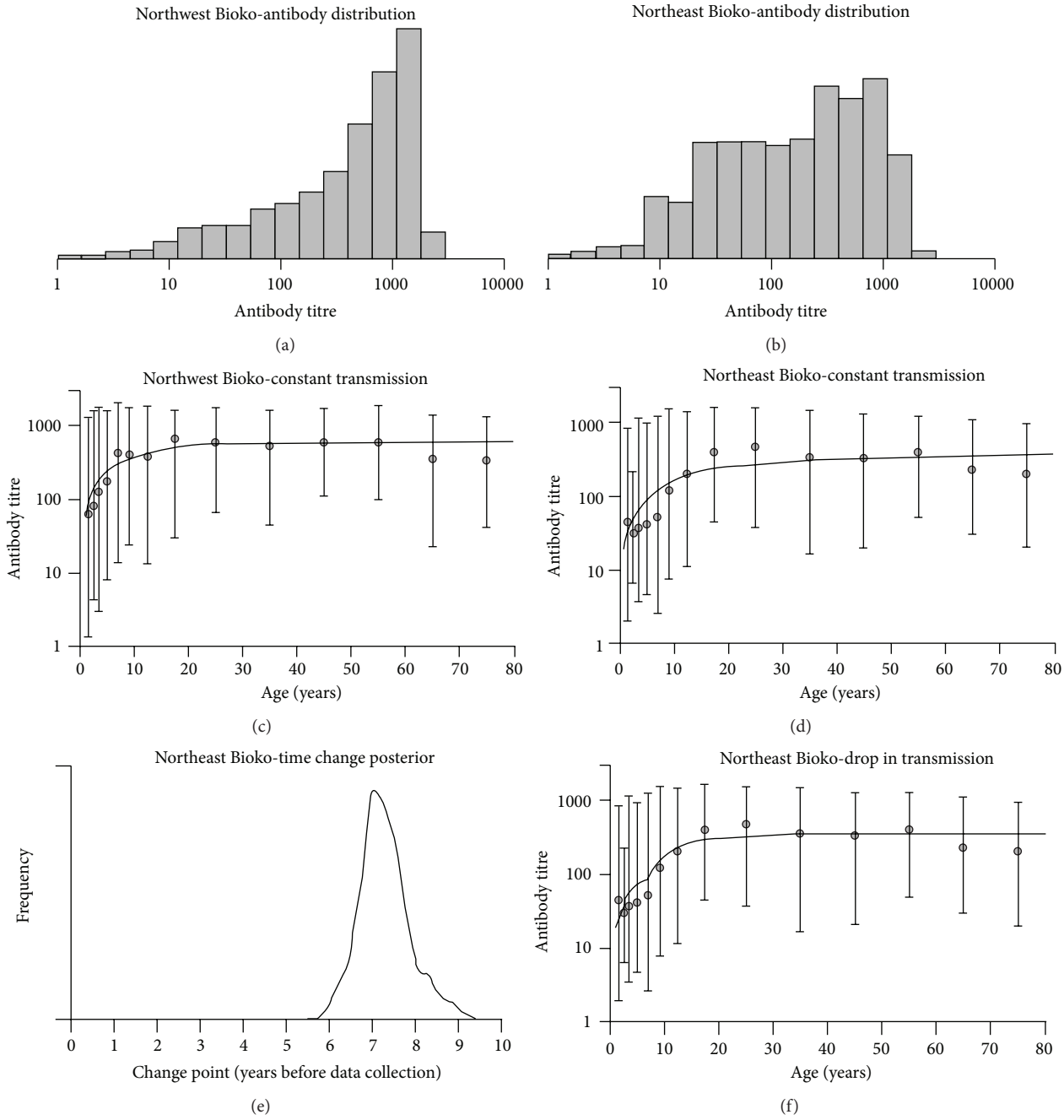


Figure 6: Data analysis of anti-AMA1 antibody titres from Bioko Island using the antibody acquisition models. (a) Sample distribution of antibody titres from northwest region. (b) Sample distribution of antibody titres from northeast region. (c) Antibody acquisition model with constant transmission applied to data from northwest Bioko. (d) Antibody acquisition model with constant transmission applied to data from northeast Bioko. (e) Posterior probability distribution of change point predicted by the antibody acquisition model applied to data from northeast region. (f) Antibody acquisition model with a drop transmission applied to data from northeast region.

together with the most likely age-adjusted seroprevalence curve for the malaria elimination. As an example, Figure 7(d) shows the power to detect malaria elimination as a function of sample size under the assumption of simple random sampling from a typical African population where the past SCR was conceptually equivalent to 0.1 infectious bites per person per year. As expected, the required sample size decreases with time of the malaria elimination event. For a 90% power, detecting malaria elimination occurring 3, 5, and 10 years before data collection requires sample sizes of 1,000, 500, and <250 individuals, respectively. It is worth noting that the choice of a particular sample strategy involves weighting ethical issues, availability of human and economic resources, and so forth, in order to be officially approved and feasible

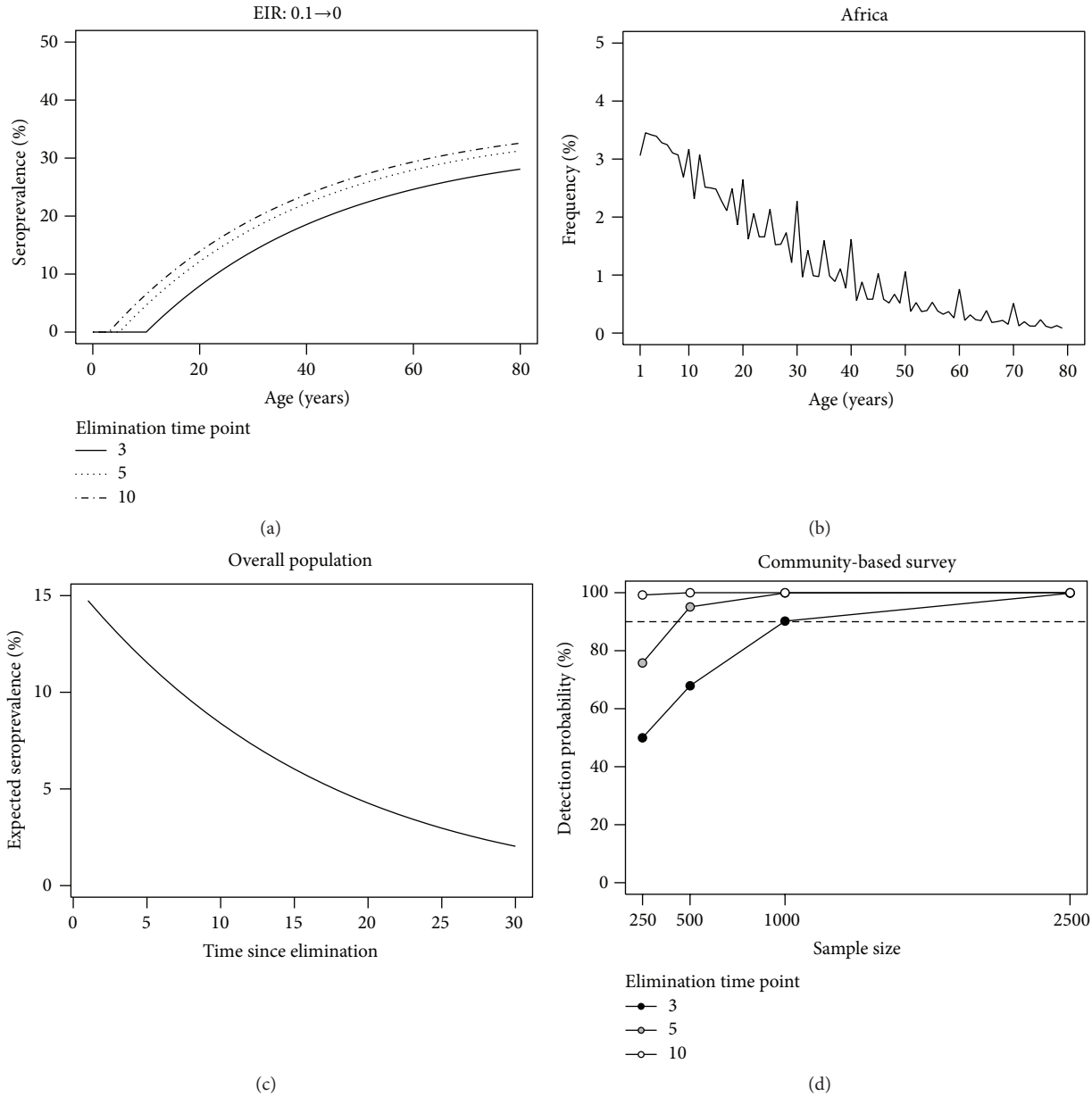


Figure 7: Certifying malaria elimination under a serology-based approach. (a) Expected seroprevalence curve from RCM assuming different elimination time points in relation to data collection. (b) Typical age distribution of an African population. (c) Expected seroprevalence in a random sample taken from a typical African population. (d) Probability to detect elimination as a function of sample size under the assumption of a community-based survey conducted in a typical African population.

in real time. Moreover, targeting or oversampling specific age groups might be alternative sampling strategies to reduce sample size. This and other issues will be tackled in a future research.

4. Conclusion

In conclusion, serology data in conjunction with mathematical modelling provides a powerful approach to inform epidemiologists on malaria transmission intensity and its putative changes over time. However, serology-based analysis needs to be complemented with any additional data available that would provide an external validation to the serological findings. Imagining the best model for a given data set assumes a constant malaria transmission intensity over time. Checking the official statistics of the malaria cases if available might shed some light on whether such assumption holds true in reality. Similar rationale can be applied to situations where a drop in malaria transmission intensity was detected on the data. This approach of using official data to consolidate serological finding was indeed followed by the study in the Brazilian Amazonia region here analyzed [7]. In this study,

the seroconversion rates for *P. vivax* antigens from several sites with different malaria endemicity levels were highly correlated with the corresponding annual parasite indexes compiled by the Brazilian health authorities, suggesting that the serological analysis was capturing the epidemiology of the study sites [7]. In the same line of thought, another study found that seroconversion rates were highly correlated with the parasite rates in northeast Tanzania [13]. Notwithstanding this good agreement between serological findings and other data, it is worth mentioning that the mathematical models for serology data are a simplified abstraction of the real world; a discussion about how robust these models are in practice can be found elsewhere [12]. However, in the words of the statistician George Box [57], all models are wrong but some are useful and that seems to be the case for the mathematical models here presented.

Two main limitations can be pinpointed to the use of a serological approach to malaria epidemiology. The most obvious one is related to which antigens are epidemiologically informative. With this respect, the two antigens most used in practice are MSP1 and AMA1 due to their immunogenicity together with the existence of optimised experimental protocols. Research efforts are currently carried out in order to identify the panel of antigens that would provide the best characterization of the epidemiological status quo of a population [20, 58]. However, these new identified antigens remained to be fully tested in subsequent field studies. One less obvious limitation is the putative lack of statistical power to estimate the seroreversion rate and its putative changes throughout life. This is very clear in low transmission settings where only a few seronegative individuals might result from seroreversion events. Ideally, seroreversion rate is a quantity that is best estimated via longitudinal studies. However, in practice, seroreversion rate is estimated indirectly via cross-sectional data, possibly leading to poor estimation precision as discussed in depth elsewhere [34]. Possible solutions for this problem include using prior information in the analysis or analyzing data from different populations together in order to borrow information from samples where the estimation of seroreversion rate is facilitated. Several future challenges were identified in particular in the context of malaria elimination and eradication. It is worth noting that a serology approach should not be seen as strict to malaria epidemiology with the potential of being applicable to other infectious diseases as long as these are capable of triggering an antibody-mediated immune response in the host. In particular, antibody data is particularly useful to track down the transmission intensity of some neglected tropical diseases, such as Trachoma [39], Chagas [40, 43], or Dengue [57, 59], due to their low endemicity and the lack of clinical symptoms of most infections. However, this requires a deeper knowledge of the antibodies with the highest potential of informing the underlying disease transmission intensity. An interesting idea with public health potential is to use a panel of multidisease antibodies that can be instrumental to know what the infectious agents are in circulation in a given population and their putative dynamics. This idea has not been tested in practice but definitely will require extending the above mathematical models to fully account for the immunological interaction between different diseases.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to thank Dr. Teun Bousema for providing the data set from the western Kenyan highlands to highlight the extension of the RCM to generate individual estimates of SCR, Dr. Lotus van den Hoogen for helping with the analysis of the superinfection model, and Dr. Kevin Tetteh for useful discussion on the type of antimalarial antibodies used in practice. Nuno Sepúlveda and Chris J. Drakeley acknowledge funding from the Wellcome Trust (Grant no. 091924) and Fundação para a Ciência e Tecnologia (Portugal) through the project Pest-OE/MAT/UI0006/2011.

References

[1] World Health Organization (WHO), *World Malaria Report 2014*, World Health Organization (WHO), Geneva, Switzerland, 2014.

[2] R. G. A. Feachem, A. A. Phillips, J. Hwang et al., "Shrinking the malaria map: progress and prospects," *The Lancet*, vol. 376, no. 9752, pp. 1566–1578, 2010.

[3] N. D. Karunaweera, G. N. Galappaththy, and D. F. Wirth, "On the road to eliminate malaria in Sri Lanka: lessons from history, challenges, gaps in knowledge and research needs," *Malaria Journal*, vol. 13, no. 1, article 59, 2014.

[4] S. Herrera, S. A. Ochoa-Orozco, I. J. González, L. Peinado, M. L. Quiñones, and M. Arévalo-Herrera, "Prospects for malaria elimination in mesoamerica and hispaniola," *PLOS Neglected Tropical Diseases*, vol. 9, no. 5, Article ID e0003700, 2015.

[5] J. Monteiro Rodriguez, J. O. Guintran, C. Gomes et al., "Moving to malaria elimination in Cape Verde," *Malaria Journal*, vol. 11, supplement 1, article O9, 2012.

[6] G. Stresman, T. Kobayashi, A. Kamanga et al., "Malaria research challenges in low prevalence settings," *Malaria Journal*, vol. 11, article 353, 2012.

[7] M. G. Cunha, E. S. Silva, N. Sepúlveda et al., "Serologically defined variations in malaria endemicity in Pará state, Brazil," *PLoS ONE*, vol. 9, no. 11, Article ID 0113357, 2014.

[8] T. Bousema, R. M. Youssef, J. Cook et al., "Serologic markers for detecting malaria in areas of low endemicity, Somalia, 2008," *Emerging Infectious Diseases*, vol. 16, no. 3, pp. 392–399, 2010.

[9] World Health Organization (WHO), *Malaria Entomology and Vector Control. Guide for Participants*, WHO, Geneva, Switzerland, 2013.

[10] M. Service, *Mosquito Ecology—Field Sampling Methods*, Applied Science Publishers, London, UK, 1976.

[11] L. S. Tusting, T. Bousema, D. L. Smith, and C. Drakeley, "Measuring changes in plasmodium falciparum transmission. Precision, accuracy and costs of metrics," *Advances in Parasitology*, vol. 84, pp. 151–208, 2014.

[12] P. Corran, P. Coleman, E. Riley, and C. Drakeley, "Serology: a robust indicator of malaria transmission intensity?" *Trends in Parasitology*, vol. 23, no. 12, pp. 575–582, 2007.

[13] C. J. Drakeley, P. H. Corran, P. G. Coleman et al., "Estimating medium- and long-term trends in malaria transmission by using serological markers of malaria exposure," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 14, pp. 5108–5113, 2005.

[14] M. T. Bretscher, S. Supargiyono, M. A. Wijayanti et al., "Measurement of *Plasmodium falciparum* transmission intensity using serological cohort data from Indonesian schoolchildren," *Malaria Journal*, vol. 12, article 21, 2013.

[15] J. Cook, I. Kleinschmidt, C. Schwabe et al., "Serological markers suggest heterogeneity of effectiveness of malaria control interventions on Bioko Island, Equatorial Guinea," *PLoS ONE*, vol. 6, no. 9, Article ID e25137, 2011.

[16] J. Cook, H. Reid, J. Iavro et al., "Using serological measures to monitor changes in malaria transmission in Vanuatu," *Malaria Journal*, vol. 9, no. 1, article 169, 2010.

[17] E. J. Remarque, B. W. Faber, C. H. M. Kocken, and A. W. Thomas, "Apical membrane antigen 1: a malaria vaccine candidate in review," *Trends in Parasitology*, vol. 24, no. 2, pp. 74–84, 2008.

[18] B. W. Faber, S. Younis, E. J. Remarque et al., "Diversity covering AMA1-MSP119 fusion proteins as malaria vaccines," *Infection and Immunity*, vol. 81, no. 5, pp. 1479–1490, 2013.

[19] P. D. Crompton, M. A. Kayala, B. Traore et al., "A prospective analysis of the Ab response to *Plasmodium falciparum* before and after a malaria season by protein microarray," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 15, pp. 6958–6963, 2010.

[20] D. A. Helb, K. K. Tetteh, P. L. Felgner et al., "Novel serologic biomarkers provide accurate estimates of recent *Plasmodium falciparum* exposure for individuals and communities," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 112, no. 32, pp. E4438–E4447, 2015.

[21] A. Voller, G. Huldt, C. Thors, and E. Engvall, "New serological test for malaria antibodies," *British Medical Journal*, vol. 1, no. 5959, pp. 659–661, 1975.

[22] T. Bousema, J. Stevenson, A. Baidjoe et al., "The impact of hotspot-targeted interventions on malaria transmission: study protocol for a cluster-randomized controlled trial," *Trials*, vol. 14, article 36, 2013.

[23] T. Benaglia, D. Chauveau, D. R. Hunter, and D. S. Young, "Mixtools: an R package for analyzing finite mixture models," *Journal of Statistical Software*, vol. 32, no. 6, pp. 1–29, 2009.

[24] S. Bosomprah, "A mathematical model of seropositivity to malaria antigen, allowing seropositivity to be prolonged by exposure," *Malaria Journal*, vol. 13, article 12, 2014.

[25] D. Modiano, A. Chiucchiuini, V. Petrarca et al., "Humoral response to *Plasmodium falciparum* Pf155/ring-infected erythrocyte surface antigen and Pf332 in three sympatric ethnic groups of Burkina Faso," *The American Journal of Tropical Medicine and Hygiene*, vol. 58, no. 2, pp. 220–224, 1998.

[26] D. Modiano, V. Petrarca, B. S. Sirima et al., "Different response to *Plasmodium falciparum* malaria in west African sympatric ethnic groups," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 93, no. 23, pp. 13206–13211, 1996.

[27] S. Richardson and P. J. Green, "On Bayesian analysis of mixtures with an unknown number of components," *Journal of the Royal Statistical Society, Series B: Methodological*, vol. 59, no. 4, pp. 731–792, 1997.

[28] Z. Zhang, K. L. Chan, Y. Wu, and C. Chen, "Learning a multivariate Gaussian mixture model with the reversible jump MCMC algorithm," *Statistics and Computing*, vol. 14, no. 4, pp. 343–355, 2004.

[29] P. Vounatsou, T. Smith, and A. F. M. Smith, "Bayesian analysis of two-component mixture distributions applied to estimating malaria attributable fractions," *Journal of the Royal Statistical Society C: Applied Statistics*, vol. 47, no. 4, pp. 575–587, 1998.

[30] J. Qin and D. H. Leung, "A semiparametric two-component compound mixture model and its application to estimating malaria attributable fractions," *Biometrics*, vol. 61, no. 2, pp. 456–464, 2005.

[31] H. Muench, *Catalytic Models in Epidemiology*, Harvard University Press, Cambridge, Mass, USA, 1959.

[32] J. H. Pull and B. Grab, "A simple epidemiological model for evaluating the malaria inoculation rate and the risk of infection in infants," *Bulletin of the World Health Organization*, vol. 51, no. 5, pp. 507–516, 1974.

[33] A. Bekessy, L. Molineaux, and J. Storey, "Estimation of incidence and recovery rates of *Plasmodium falciparum* parasitaemia from longitudinal data," *Bulletin of the World Health Organization*, vol. 54, no. 6, pp. 685–693, 1976.

[34] N. Sepúlveda and C. J. Drakeley, "Sample size determination for estimating antibody seroconversion rate under stable malaria transmission intensity," *Malaria Journal*, vol. 14, article 141, 2015.

[35] C. I. Bliss, "The calculation of the dosage-mortality curve," *Annals of Applied Biology*, vol. 22, no. 1, pp. 134–167, 1935.

[36] M. E. von Fricken, T. A. Weppelmann, B. Lam et al., "Age-specific malaria seroprevalence rates: a cross-sectional analysis of malaria transmission in the Ouest and Sud-Est departments of Haiti," *Malaria Journal*, vol. 13, article 361, 2014.

[37] B. G. Williams and C. Dye, "Maximum likelihood for parasitologists," *Parasitology Today*, vol. 10, no. 12, pp. 489–493, 1994.

[38] T. Bonnefoix, P. Bonnefoix, P. Verdiel, and J.-J. Sotto, "Fitting limiting dilution experiments with generalized linear models results in a test of the single-hit poisson assumption," *Journal of Immunological Methods*, vol. 194, no. 2, pp. 113–119, 1996.

[39] D. L. Martin, R. Bid, F. Sandi et al., "Serology for trachoma surveillance after cessation of mass drug administration," *PLoS Neglected Tropical Diseases*, vol. 9, no. 2, Article ID e0003555, 2015.

[40] S. Delgado, R. C. Neyra, V. R. Q. Machaca et al., "A history of Chagas disease transmission, control, and re-emergence in peri-rural La Joya, Peru," *PLoS Neglected Tropical Diseases*, vol. 5, no. 2, article e970, 2011.

[41] S. Supargiyono, M. T. Bretscher, M. A. Wijayanti et al., "Seasonal changes in the antibody responses against *Plasmodium falciparum* merozoite surface antigens in areas of differing malaria endemicity in Indonesia," *Malaria Journal*, vol. 12, no. 1, article 444, 2013.

[42] A. C. Marques, "Human migration and the spread of malaria in Brazil," *Parasitology Today*, vol. 3, no. 6, pp. 166–170, 1987.

[43] N. M. Bowman, V. Kawai, M. Z. Levy et al., "Chagas disease transmission in periurban communities of Arequipa, Peru," *Clinical Infectious Diseases*, vol. 46, no. 12, pp. 1822–1828, 2008.

[44] D. J. Spiegelhalter, N. G. Best, B. P. Carlin, and A. van der Linde, "Bayesian measures of model complexity and fit (with discussion)," *Journal of the Royal Statistical Society Series B*, vol. 64, no. 4, Article ID 583639, pp. 583–639, 2002.

[45] J. Cook, N. Speybroeck, T. Sochanta et al., "Sero-epidemiological evaluation of changes in *Plasmodium falciparum* and *Plasmodium vivax* transmission patterns over the rainy season in Cambodia," *Malaria Journal*, vol. 11, article 86, 2012.

[46] Brazilian Institute of Geography and Statistics, Census 2010, http://censo2010.ibge.gov.br/en/censo-2010.html.

[47] P. McCullagh and J. A. Nelder, *Generalized Linear Models*, Chapman & Hall, London, UK, 2nd edition, 1989.

[48] E. J. Bedrick, R. Christensen, and W. Johnson, "A new perspective on priors for generalized linear models," *Journal of the American Statistical Association*, vol. 91, no. 436, pp. 1450–1460, 1996.

[49] P. H. Garthwaite, J. B. Kadane, and A. O'Hagan, "Statistical methods for eliciting probability distributions," *Journal of the American Statistical Association*, vol. 100, no. 470, pp. 680–701, 2005.

[50] E. Pothin, *Modelling antibody responses to malaria blood stage infections: a novel method to estimate malaria transmission intensity from serological data [Ph.D. thesis]*, Imperial College London, London, UK, 2013.

[51] M. T. White, J. T. Griffin, O. Akpogheneta et al., "Dynamics of the antibody response to *Plasmodium falciparum* infection in african children," *Journal of Infectious Diseases*, vol. 210, no. 7, pp. 1115–1122, 2014.

[52] X.-H. Zhou and W. Tu, "Comparison of several independent population means when their samples contain log-normal and possibly zero observations," *Biometrics*, vol. 55, no. 2, pp. 645–651, 1999.

[53] L. Tian, "Inferences on the mean of zero-inflated lognormal data: the generalized variable approach," *Statistics in Medicine*, vol. 24, no. 20, pp. 3223–3232, 2005.

[54] N. Li, D. A. Elashoff, W. A. Robbins, and L. Xun, "A hierarchical zero-inflated log-normal model for skewed responses," *Statistical Methods in Medical Research*, vol. 20, no. 3, pp. 175–189, 2011.

[55] G. J. McLachlan, "On bootstrapping the likelihood ratio test for two component normal mixture model," *Applied Statistics*, vol. 36, no. 3, pp. 318–324, 1987.

[56] Y. Lo, N. R. Mendell, and D. B. Rubin, "Testing the number of components in a normal mixture," *Biometrika*, vol. 88, no. 3, pp. 767–778, 2001.

[57] C. Ochieng, P. Ahenda, A. Y. Vittor et al., "Seroprevalence of infections with dengue, rift valley fever and chikungunya viruses in Kenya, 2007," *PLoS One*, vol. 10, no. 7, Article ID e0132645, 2015.

[58] F. Lu, J. Li, B. Wang et al., "Profiling the humoral immune responses to *Plasmodium vivax* infection and identification of candidate immunogenic rhoptry-associated membrane antigen (RAMA)," *Journal of Proteomics*, vol. 102, pp. 66–82, 2014.

[59] N. Imai, I. Dorigatti, S. Cauchemez, N. M. Ferguson, and S. I. Hay, "Estimating dengue transmission intensity from seroprevalence surveys in multiple countries," *PLoS Neglected Tropical Diseases*, vol. 9, no. 4, Article ID e0003719, 2015.

Hindawi Publishing Corporation
Journal of Immunology Research
Volume 2015, Article ID 140819, 10 pages
http://dx.doi.org/10.1155/2015/140819

Research Article

MIrExpress: A Database for Gene Coexpression Correlation in Immune Cells Based on Mutual Information and Pearson Correlation

Luman Wang,[1,2] Qiaochu Mo,[1] and Jianxin Wang[1,3]

[1]*School of Information, Beijing Forestry University, Beijing 100083, China*
[2]*Department of Natural Science in Medicine, Peking University Health Science Center, Beijing 100191, China*
[3]*Center for Computational Biology, Beijing Forestry University, Beijing 100083, China*

Correspondence should be addressed to Jianxin Wang; wangjx@bjfu.edu.cn

Received 29 May 2015; Accepted 9 November 2015

Academic Editor: Francesco Pappalardo

Copyright © 2015 Luman Wang et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Most current gene coexpression databases support the analysis for linear correlation of gene pairs, but not nonlinear correlation of them, which hinders precisely evaluating the gene-gene coexpression strengths. Here, we report a new database, MIrExpress, which takes advantage of the information theory, as well as the Pearson linear correlation method, to measure the linear correlation, nonlinear correlation, and their hybrid of cell-specific gene coexpressions in immune cells. For a given gene pair or probe set pair input by web users, both mutual information (MI) and Pearson correlation coefficient (r) are calculated, and several corresponding values are reported to reflect their coexpression correlation nature, including MI and r values, their respective rank orderings, their rank comparison, and their hybrid correlation value. Furthermore, for a given gene, the top 10 most relevant genes to it are displayed with the MI, r, or their hybrid perspective, respectively. Currently, the database totally includes 16 human cell groups, involving 20,283 human genes. The expression data and the calculated correlation results from the database are interactively accessible on the web page and can be implemented for other related applications and researches.

1. Introduction

In recent years, the advance of microarray technology has provided amounts of information for us to observe the expression levels of genes together. Based on the increasing availability of gene expression data, public gene expression repositories were successfully constructed, such as the GEO [1] database and ArrayExpress [2]. These supply more opportunities to study gene expressional correlation (gene coexpression). Gene coexpression may reveal general functional tasks and regulatory mechanisms; moreover, it may identify novel genes to be involved in certain diseases. In addition, in the related fields of biology, many studies illustrated that the dependencies of gene expression can reflect the normal and dysfunctional biological processes and furthermore make us understand the underlying molecular mechanisms [3, 4]. It is difficult, however, for biologists without bioinformatics background to retrieve the gene coexpression information effectively and efficiently. For such, there in the field of plant biology are many coexpression databases, such as PLANEX [5], ATTED-II [6], Cop [7], TEGD [8], and PlantCART [9], the information in which was derived from large-scale gene expression data. Besides those, several coexpression databases peculiarly for mammals recently have been established and widely used by researchers and have thus accelerated the coexpression analysis process in the field of bioinformatics. COXPRESdb [10] was constructed with gene expression data from 63 human tissues and it utilizes the correlation rank to compare the coexpression strengths among multiple species. The database GeneFriends [11] adopted the same approach as COXPRESdb to construct coexpression maps based on transcriptome sequencing (RNA-seq) gene data instead of microarray gene data. HGCA [12] was constructed based on gene expression data from about two thousand samples of various cells and tissues. The overall correlation in gene expression was identified in

this database across multiple tissues, or mixed tissues and cells, without meeting the necessity of coexpression in the same cell type. Immuco [13] is a cell-specific database in which gene expression values in each cell type across various conditions are provided, as well as gene coexpression and correlation information. Though these databases have been constructed successfully and are able to meet users' needs to some extent, they capture only the linear coexpression relationships between different genes by the Pearson correlation coefficient (value r). In fact, r with small absolute value of two genes does not necessarily mean that the two genes are independent, since nonlinear relationship may exist in the gene coexpression data [14]. In particular, two variables with a vanishing correlation coefficient may be heavily dependent, as illustrated in the later example in this paper (see Figure 2). The mutual information (MI) is able to measure the mutual dependence of two random variables, particularly in terms of positive, negative, and nonlinear correlations [15], and in comparison with Pearson correlation coefficient, it may provide a criterion better and more general to investigate gene coexpression. And in recent years, the mutual information is regarded as a common way to detect dependencies between different genes. Steuer et al. initiated the mutual information approach [16] for one specific gene dataset to analyze intergene dependencies.

Bioconductor is an open source software which provides the key function in Affymetrix array analysis in the R software environment (http://www.r-project.org/) [17], and Meyer et al. [18] developed a package "minet" in Bioconductor, in which a powerful tool is provided to calculate the mutual information between different gene pairs. Based on a publicly available dataset *Saccharomyces cerevisiae* [19] including 2,467 genes, Butte and Kohane applied the mutual information to measure gene-gene interaction and obtained the result that the mode of MI was about 0.7. Consequently, 22 relevance networks were constructed when the threshold of information (TMI) was set to 1.3 [20]. With gene expression data from various environments, the mutual information approach [21] was employed to reconstruct regulatory networks of relationships.

In spite of the many researches and applications mentioned above about mutual information for gene correlations, few publications related to mutual information focus on immune cells. Since the mutual information should be calculated for each gene thoroughly connected to every other gene for correlation [20], the amount of correlation coefficients is tremendous and grows significantly with increasing number of genes. Thus, most publications applied the mutual information algorithms to measure coexpression on public sample datasets or testing datasets that includes much fewer genes than initial datasets.

In order to investigate the expression correlation of immune genes, we constructed a database named MIrExpress (http://wjx.bjfu.edu.cn/MIrExpress) including 41,477 probe sets for 20,283 human genes with each of the 16 cell types in immune cells to reflect the linear and nonlinear correlation of cell-specific gene coexpression profiles across multiple experimental conditions, aided by both Pearson correlation coefficient (r) and mutual information value (MI). Through a web interface, the database exhibits the scatter plot of the cooccurrence signal values of any two probe pairs to illustrate the extent and strength of correlation. For a given gene pair, not only is the MI given through the web interface, but its rank expressed in percentage is also presented in all the gene pairs, that is, about 8.6×10^8 pairs for each dataset. Besides, it is the same case for the Pearson correlation value r. Both the values and ranks of MI and r are displayed and contrasted graphically. In the querying web pages, the top 10 most relevant genes of an input gene can be listed with the perspective of Pearson correlation, mutual information, and their hybrid, respectively.

2. Materials and Methods

2.1. Data Preparation and Preprocessing. Gene Expression Omnibus (GEO) founded by National Center for Biotechnology Information (NCBI) in July 2000 is the largest public database to date for gene expression data (http://www.ncbi.nlm.nih.gov/geo/) [22]. In this paper, the SOFT format annotation files in GEO database were downloaded from the platform GPL570 for human cells. According to the SOFT files, samples related to immune cells were screened and sorted by cell types. Based on cell-specific sample ID, the raw gene expression data in CEL format files were downloaded from the GEO database using the GEOquery package [23] in R language environment, each expression data containing a single value describing the signal intensity for each probe set on the array.

In order to help improve the efficiency of the data analyzing process, the functions in the packages of Bioconductor were performed on the gene expression data. Firstly, package "simpleaffy" was used to discard the samples with extreme values in order to control the quality of raw data including scale factor, background level, percentage of genes which are called present, and $3'/5'$ ratios as the QC metrics [24]. After quality control, 6,909 human samples for 293 GEO series were selected as they were done in Immuco database [13]. Secondly, in the package "affy," MAS 5.0 algorithms including background correction, normalization, and summary were applied to generally process qualified gene expression data which are allowed for comparison among the gene expression data of samples from different experiments [25, 26]. After that, 41,477 probe sets for human organism were retained for later gene coexpression correlation analysis while about 15% of the samples were discarded due to quality control.

2.2. Calculation of Pearson Correlation Coefficient. Pearson correlation coefficient (r) is a measure of the linear correlation between two probe sets X and Y, which can be denoted by $r_{X,Y}$ and calculated as follows:

$$r_{X,Y} = \frac{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right)}{\sqrt{\sum_{i=1}^{n} \left(X_i - \overline{X}\right)^2 \sum_{i=1}^{n} \left(Y_i - \overline{Y}\right)^2}}, \quad (1)$$

where n is the number of samples from different experiments and X_i and Y_i are the expression profiles' values of probe sets X and Y in the ith sample, respectively.

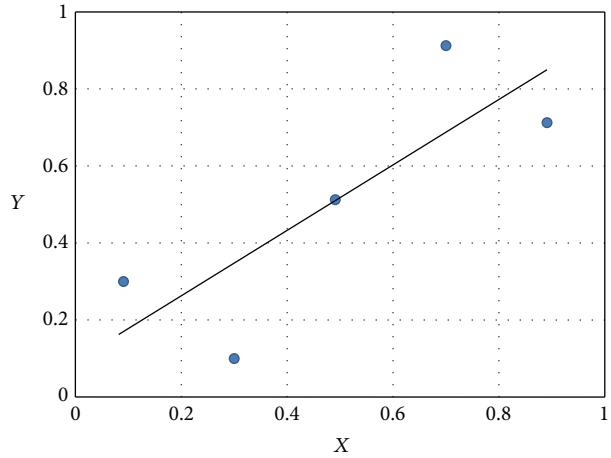


Figure 1: A scatter plot shows expression data of probe sets X and Y for dataset $[(0.1, 0.3), (0.3, 0.1), (0.5, 0.5), (0.7, 0.9), (0.9, 0.7)]$.

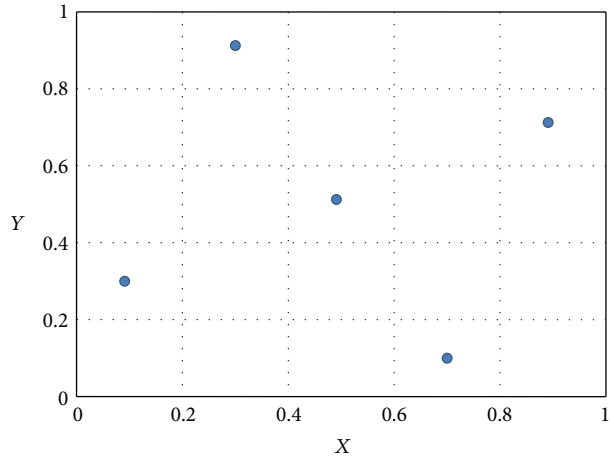


Figure 2: A scatter plot about expression data of probe sets X and Y with fixed intervals to divide the axes into discrete bins. Dataset = $[(0.1, 0.3), (0.3, 0.9), (0.5, 0.5), (0.7, 0.1), (0.9, 0.7)]$.

The r values range between −1 and +1, in which 1 implies total positive correlation, −1 total negative correlation, and 0 no correlation between the probe set pairs. The simple example in Figure 1 is a scatter plot about expression data of probe sets X and Y for a dataset, and the corresponding Pearson correlation coefficient r computed using (1) is 0.8. It indicates that there is strongly linear correlation between the probe set pairs.

2.3. Statistical Analysis about Mutual Information. The concept of entropy originates in physics, which measures the disorder of a thermodynamic systems. Shannon [27] originally devised the entropy to study the amount of information in a transmitted message and constructed information theory. So far, entropy has wide applications in various fields. Based on the theory of entropy, mutual information is applied to measure the information contained in one probe set about the other. If the mutual information of two probe sets is high, it means that it is easy to predict the expression value of one probe set according to the expression value of the other, which indicates that there may be a close relationship between genes. On the other hand, if the mutual information of two probe sets is zero, it implies that the two variables (two genes) are independent and do not correlated [14]. Based on the entropy theory, we implemented the mutual information approach to study gene coexpression.

According to the concept of mutual information, we regard a probe set as a discrete random variable and calculate the mutual information of two probe sets as the following process [21]. Suppose that A is the value range of a probe set X and A is divided by the subinterval set $\{A_i\}$, $i = 1, 2, \ldots, M$, satisfying that $\bigcup_i \{A_i\} = A$ and that $A_i \cap A_k = \emptyset$ if $i \neq k$. The entropy $H(X)$ of the probe set X can be defined as

$$H(X) = -\sum_{i=1}^{M} p(A_i) \log_2 p(A_i), \tag{2}$$

where probabilities $p(A_i)$ are approximated by the corresponding relative frequencies of occurrence in A_i and can be calculated as

$$p(A_i) \longrightarrow \frac{k_i}{N}, \tag{3}$$

where k_i denotes the number of gene expression data in the subsection A_i and N is the total number of gene expression data for the probe set X [16]. When the probability $p(A_i)$ is 1 and all other probabilities $p(A_j)$ with $i \neq j$ are zero, we get the minimum of $H(X)$, zero. In contrast, if $p(A_i) = 1/M$ for each A_i, maximum of $H(X)$ can be reached as $\log_2 M$. The joint entropy $H(X, Y)$ of two probe sets X and Y is defined as

$$H(X, Y) = -\sum_{i=1}^{M_X} \sum_{j=1}^{M_Y} p(X_i, Y_j) \log_2 p(X_i, Y_j). \tag{4}$$

Here $p(X_i, Y_j)$ denotes the joint probability that X is in subinterval set $\{A_i\}$, $i = 1, 2, \ldots, M_X$, and Y is in subinterval set $\{B_j\}$, $j = 1, 2, \ldots, M_Y$, and $p(X_i, Y_j)$ can be computed approximately as

$$p(X_i, Y_j) \longrightarrow \frac{k_{ij}}{N}. \tag{5}$$

In the above equation, k_{ij} denotes the number of gene expression data when X lies in A_X and Y in B_Y. If the probe sets X and Y are statistically independent, we can get the joint entropy $H(X, Y)$ after factorizing the joint probabilities as the following formula [16]:

$$H(X, Y) = H(X) + H(Y). \tag{6}$$

The mutual information $I(X, Y)$ between the probe sets X and Y is then defined as

$$I(X, Y) = H(X) + H(Y) - H(X, Y) \geqslant 0. \tag{7}$$

When the probe sets X and Y are statistically independent, the mutual information $I(X, Y)$ is zero according to (6) and (7). In sum, $I(X, Y)$ can be taken as measure of correlation

no matter whether the correlation is linear or nonlinear. According to (2) and (3), (7) can be rewritten as

$$\begin{aligned} I(X,Y) &= -\sum_{i=1}^{M_X} p(X_i)\log_2 p(X_i) \\ &\quad - \sum_{j=1}^{M_Y} p\left(Y_j\right)\log_2 p\left(Y_j\right) \\ &\quad + \sum_{i=1}^{M_X}\sum_{j=1}^{M_Y} p\left(X_i,Y_j\right)\log_2 p\left(X_i,Y_j\right) \\ &= \log_2 N + \frac{1}{N}\sum_{i=1}^{M_X}\sum_{j=1}^{M_Y} k_{ij}\log_2\frac{k_{ij}}{k_i k_j}. \end{aligned} \tag{8}$$

We now use the above formula of mutual information to estimate the value about MI for dataset A in Figure 1. The dataset consists of $N = 5$ data points divided into $M_X = M_Y = 5$ bins with fixed intervals and the resulted value of mutual information $I(X,Y)$ computed using (8) is 2.322.

If we change the positions of the data points in Figure 1 and rearrange them to a state as shown in Figure 2, we will find that Y is equally dependent on X as before, because an occurrence of X is equally capable of predicting the occurrence of Y as before. That is to say, the MI remains to be 2.322 without any change. The Pearson correlation coefficient r, however, changes dramatically from 0.8 to 0, which implies that X and Y are now not linearly correlated at all. This simple example indicates that MI can generally measure the dependency including both the linear and nonlinear correlation between two probe sets and overcome the drawback of Pearson correlation that takes only the linear correlation into account.

It is a simple approach to estimate the probabilities for gene expression data occurrence in each interval by (3), but it leads to overestimating the mutual information for finite-size datasets [28]. Instead of dividing the expression data range into equal intervals, we adopted an adaptive partitioning strategy to calculate mutual information between two variables (two probe sets here) [16, 29]. It means that the value range of each probe set is divided into M discrete nonoverlapping intervals, each containing approximately N/M data points. The width of each interval is thus various according to the density of data points and more occupied regions are covered with smaller intervals. For instance, let $M = 11$ and the entropy $H(X)$ and $H(Y)$ in (2) can be described as $H(X) = H(Y) = -\log_2 11$. Consequently, the mutual information $I(X,Y)$ between probe sets X and Y can be calculated as

$$I(X,Y) = 2\log_2 11 + \sum_{i=1}^{M_X}\sum_{j=1}^{M_Y}\frac{k_{ij}}{N}\log_2\frac{k_{ij}}{N}. \tag{9}$$

3. Calculation

For improved measuring effect, the Pearson correlation coefficient and mutual information can be jointly applied to evaluate the strength of gene coexpression. The Pearson correlation coefficient r reflects the linear correlation between any two genes, while the mutual information MI generally measures the dependency of one gene on another, both linearly and nonlinearly. But the range and the distribution of values for these two measures are different ($\mathrm{MI} \in [0, +\infty)$, $r \in [-1, 1]$); thus it is not suitable to compare these two measures directly to quantify the linear and nonlinear correlation between any gene pairs [16, 30], and so we adopted the rank ordering of MI and r as coexpression measure in the MIrExpress database instead. So we need to compare ordering ranks of MI and r instead of their values; that is, we need to find the ranks of MI and r in more than 8.6×10^8 values, respectively. However, it is both space-inefficient and time-inefficient to find the rank of a given value in that large amount of values. In fact, ranks in percentage are sufficient because we do not need the rank ordering information more detailed than the percentage.

In order to get the MI rank of any gene pair, all we need now is a vector $V = \{v_i\}$, $i = 0, 1, \ldots, 100$, where v_0 is the minimum mutual information and v_{100} is the maximum one, and the mutual information of approximately 1 percent gene pairs resides between v_{i-1} and v_i, $i = 1, 2, \ldots, 100$. With the vector V, we can directly search the proper interval that a given MI resides in and thus find how many mutual information values in percentage are smaller than this MI. It is by this way that we reduce the memory consumption from about $O(8.6 \times 10^8)$ to $O(101)$.

It is easy for us to save the vector V, but difficult to obtain it, for we have no prior knowledge about the distribution of the mutual information values unless we scan the whole pairs and rearrange them. That means we have to record each information value of all the gene pairs for later use, which would consume large amount of memory. Fortunately we noticed that if the expression value range is divided into 11 intervals (as we did with MIrExpress database), the mutual information of any gene pair is between 0 and 3.5, and so we equally divided the MI value range $[0, 3.5]$ into 35,000 subintervals and counted the number of pairs whose mutual information resides in a given interval. After scanning all the gene pairs, we obtained another vector with integer values $U = \{u_j\}$, $j = 1, 2, \ldots, 35000$, where u_j is the number of gene pairs whose mutual information is in the interval $[(u_j - 1)/10000, u_j/10000)$. By using this compression technique, we reduced the memory consumption from more than $O(8.6 \times 10^8)$ to $O(35{,}000)$, yet still retaining high accuracy.

Besides the above two techniques of space-saving, we designed highly time-efficient algorithms to accelerate the coexpression analysis. Firstly, the initial values (expression data) were preprocessed and only the corresponding interval information was saved. Thus, the expression of each gene in about 1,000 samples was designated to one of the 11 intervals and the variable values needed in (9) are well-prepared. The second technique was constructing a table for the $(k_{ij}/N)\log_2(k_{ij}/N)$ part in (9). We noticed that the number of different values of $(k_{ij}/N)\log_2(k_{ij}/N)$ is no more than the number of samples that is less than 2,000. And so we saved the values in a table T indexed by the integer value k_{ij},

and thus we were able to search the table instead of computing the complex function.

With the vectors V, U and the table T well-prepared, it was relatively easy for us to calculate the MI for any given gene pair and its rank in all the MI values. Firstly, the expression value pair for each sample was divided into one element of an 11-by-11 matrix. Then, we could look up all the function values of $(k_{ij}/N)\log_2(k_{ij}/N)$ in the table T and add them up according to (9) to get the MI. Finally we would look up the MI value in the vectors V and U to get the knowledge of how many MI values are smaller than this one and at which percentage this value is located.

4. Results

4.1. The Rank of MI and r. The MIrExpress database displays a global view of cell-specific gene expression profile across different experiment conditions through two-dimensional scatter plots whose axes represent the signal values of two probe sets. The scatter plot provides database users significant intuition about the general coexpression level of two genes. As is mentioned above, it is not suitable for us to directly compare MI and r to find dependency level, linear correlation level, and linear component of the dependency relation, since their value ranges and distributions are widely different. However, it is much more reasonable if we compare their value ranks, that is, where the MI and r are located in all sorted MI values and r values, respectively. For example, if the rank of an MI is 70%, then it means that 70 percent of all the MI values are smaller than this MI value.

We denote the rank of an MI in all the sorted MI values by RoMI and that of r by Ror. In immune cells, there are 4 cases for two given probe sets when we have calculated their RoMI and Ror.

(1) Both RoMI and Ror are high. For example, the MI and r of probe sets ID 201577_at (Gene Symbol: NME1) and 1053_at (Gene Symbol: RFC2) in CD4+ T cells are 0.732516 and 0.821057, respectively, and the RoMI and Ror of these two measures are both 99%. This indicates that there exists strong linear and nonlinear correlation and coexpressed relationship between these two probe sets (Figure 3(a)).

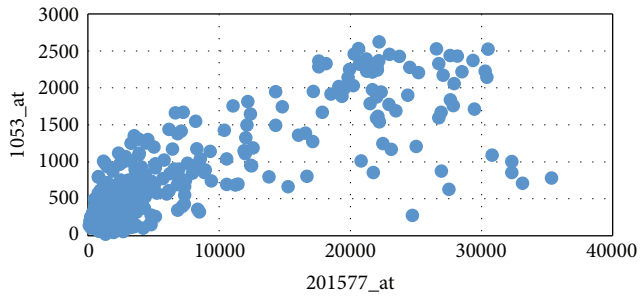
(2) The RoMI is high while the Ror is low. For example, there is strongly coexpressed relationship (total coexpression rate = 75.86%) between probe sets 1487_at (Gene Symbol: ESRRA) and 203176_s_at (Gene Symbol: TFAM) in DC cells (dendritic cells), which cannot be reflected through r value (−0.000118). If we employ MI value to measure the dependent relationship in MIrExpress database, then the MI is 0.59463 and the corresponding RoMI is 99%, a much higher rank than Ror (1%), which indicates a weak linear correlation but a strong nonlinear correlation between the probe sets (Figure 3(b)). Take CD4+ T cell, for instance, and there is the strong coexpressed relationship (total coexpression rate = 95.10%) between probe sets 219123_at (Gene Symbol: ZNF232) and 1552316_a_at (Gene Symbol: GIMAP1), which cannot be reflected by r value (−0.006148), either. But inquiring MIrExpress database, we get the RoMI and Ror as 99% and 1%, respectively (Figure 3(c)). Through these two examples we can observe that mutual information (MI) and the rank of it (RoMI) better interpret the coexpression relationship between probe sets than Pearson correlation coefficient (r) and the rank of it (Ror). So mutual information provides a more reliable and reasonable explanation of gene coexpression.

(3) Both RoMI and Ror are low. For example, the MI and r of probe sets ID 1320_at (Gene Symbol: PTPN21) and 1554627_a_at (Gene Symbol: ASCC1) in CD4+ T cells are 0.13114 and 0.00097, respectively, and the corresponding RoMI and Ror of these two measures are both 1%. It indicates that both linear and nonlinear correlation are quite weak between these probe sets (Figure 3(d)).

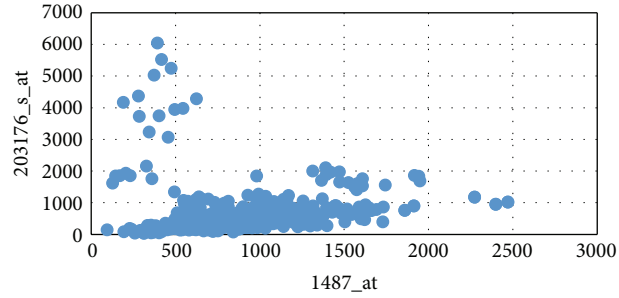
(4) The RoMI is low while the Ror is high. For example, the MI and r of probe set ID 1553169_at (Gene Symbol: LRRN4) and 234776_at (Gene Symbol: DMBX1) in CD4+ T cells are 0.13189 and 0.56614, respectively, and the corresponding RoMI and Ror are 1% and 99%, respectively (Figure 3(e)). But we notice that the absent-absent rate (AA) is 99.64% and present-present rate (PP) is 0.00%, which makes it seems true that the two probe sets are strongly linear-correlated. In fact, they are not indeed highly linear-correlated, because a high AA together with a low PP makes the Pearson correlation coefficient quite great. The low RoMI is consistent with the vanishing dependency between the two probe sets who both have low expression level in almost all samples.

4.2. Database Contents. We built the MIrExpress database (Browser/Server architecture) adopting Apache Tomcat as web server and MySQL as database server, and it provides users an easy-understanding web interface. All samples for the MIrExpress database are based on immune cells including 16 human cell groups, and the expression data of samples are chosen for Affymetrix Human Genome U133 plus 2.0 Array from GEO database. The web interface of MIrExpress database mainly includes three types of pages: page for pairwise correlation analysis (see Figure 4), page for most related genes, and page for cell-type-based overview of rank difference.

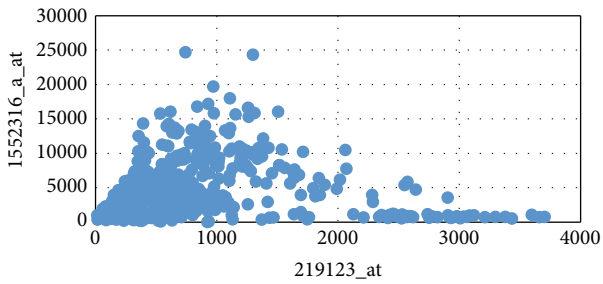
(1) Page for pairwise correlation analysis presents the general expression level of any two genes specified by users among 41,477 probe sets. Users only need to select the cell types and input two queried genes by symbol (e.g., DDR1 and RFC2) or probe sets (e.g., 1007_s_at and 1053_at) in the querying box and click the submitting button to acquire the two-dimensional scatter plot for these two genes. Meanwhile, the MI and r, together with the corresponding RoMI and Ror and their comparison, are displayed in the responding page.



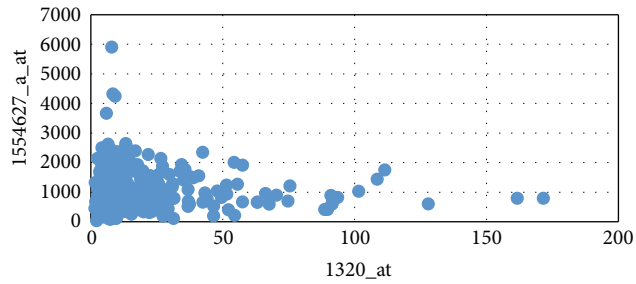
(a) CD4+ T cell; MI = 0.732516; $r = 0.821057$; RoMI = Ror = 99%; total coexpression rate = 88.38%



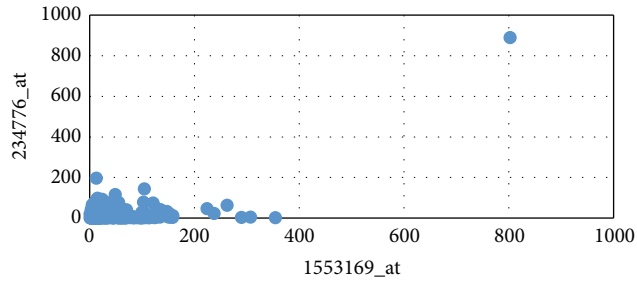
(b) DC; MI = 0.59463; $r = -0.000118$; RoMI = 99%; Ror = 1%; total coexpression rate = 75.86%



(c) CD4+ T cell; MI = 0.60838; $r = -0.00615$; RoMI = 99%; Ror = 1%; total coexpression rate = 95.10%



(d) CD4+ T cell; MI = 0.13114; $r = 0.00097$; RoMI = Ror = 1%; total coexpression rate = 1.09%



(e) CD4+ T cell; MI = 0.13189; $r = 0.56614$; RoMI = 1%; Ror = 99%; total coexpression rate = 99.64%

Figure 3: Sample applications for gene coexpression analysis. (a) NME1 and RFC2 in CD4+ T cells. (b) ESRRA and TFAM in DC cells. (c) ZNF232 and GIMAP1 in CD4+ T cells. (d) PTPN21 and ASCC1 in CD4+ T cells. (e) LRRN4 and DMBX1 in CD4+ T cells.

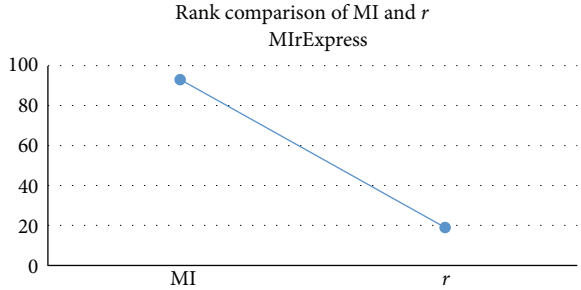
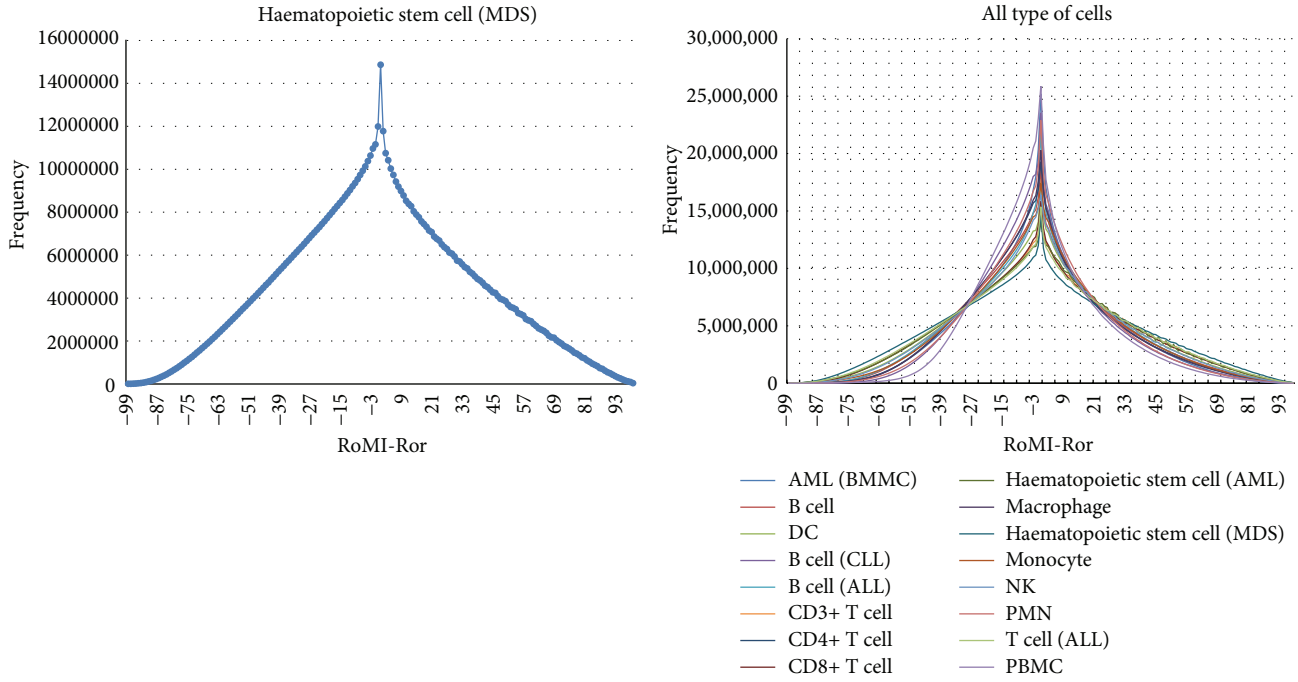


Figure 4: Page for pairwise correlation analysis. The scatter diagram of probe set pair is omitted which appears as Figures 3(a), 3(b), 3(c), 3(d), and 3(e). The species is "human"; the dataset is "CD3+ T cell"; for Gene A, the probe set ID is "1007_s_at" and the gene symbol is "DDR1"; for Gene B, the probe set ID is "1053_at" and the gene symbol is "RFC2." The Pearson's r value is −0.05538, the MI value is 1.04368, and the MIr value is 0.36477 for their hybrid.

(2) Page for most related genes lists information about the 10 most strongly correlated genes to the queried one with 3 perspectives, namely, MI, r, and their hybrid, respectively. We use MIr to denote the hybrid measure of MI and r, calculated as the follows:

$$\mathrm{MIr} = \beta \frac{r\left(X_i, X_j\right)}{\max_{k \neq i}\left(r\left(X_i, X_k\right)\right)} + \left(1 - \beta\right) \frac{\mathrm{MI}\left(X_i, X_j\right)}{\max_{k \neq i}\left(\mathrm{MI}\left(X_i, X_k\right)\right)}, \tag{10}$$

where X_i is the queried gene, X_j is any other one, and β is a coefficient often set to be round 0.5 for optimum effect. For example, if a user inputs a probe set 1007_s_at of CD3+ T cell in the selected page



(a) Haematopoietic stem cell (MDS) overview of rank difference

(b) Cell-type-based overview of rank difference for all the cell types

Figure 5: Page for cell-type-based overview of rank difference.

and submits the selection, then 30 probe sets and their gene information will be retrieved, in which 10 probe sets are most related to the queried probe set according to the r, another 10 to MI, and still another 10 to MIr (Table 1).

(3) The page for cell-type-based overview of rank difference provides for each cell type an overview of how all the RoMI-Ror values are distributed. For instance, the RoMI-Ror value distribution overview of haematopoietic stem cell (MDS) type is shown in Figure 5(a), in which the horizontal axis represents the difference between RoMI and Ror, namely, RoMI-Ror, and the vertical axis represents the frequency of a given difference RoMI-Ror which occurs. And the sum of all the frequency is 860,150,026. Specifically, MI and r have the same rank when RoMI-Ror = 0, and we observe that the frequency of this rank difference is the highest. As the difference RoMI-Ror gradually increases (or decreases) from this point, the frequency that the corresponding difference occurs gradually decreases. The rank difference (RoMI-Ror) distributions of all the 16 cell types are shown in Figure 5(b).

4.3. Managing and Expanding the Database. The website and the database are totally automatic in responding the users' query if there is no abnormity. However, human operators are required to be involved to expand the database. In fact, in order to increase the visit speed of the website, we have preprocessed all the data before they are mounted into the database, and thus all newly acquired expression data should be preprocessed by human operators with the preprocessing software, individually or in batch.

5. Conclusions and Discussions

The MIrExpress database provides an effective and novel method to observe linear and nonlinear dependencies for pairwise gene expression data under a series of experiment conditions in immune cells. To date, this cannot be achieved in other related databases about correlation of gene expression. Traditionally, standard methods, such as Pearson correlation, are used to identify gene coexpression and correlation relationships. However, in some cases, coexpression relationship exists obviously but the Pearson correlation coefficient cannot reflect the dependency, which indicates that there is nonlinear correlation between gene pairs. In this paper, we took into account the rank ordering of mutual information and Pearson correlation coefficient to generally measure the gene correlation in linear and nonlinear aspects, which better describes the gene coexpressions.

There is much room for the MIrExpress database to be improved. First, much more samples may also be incorporated to enrich the database content in order to more precisely measure the correlation in the future. Second, the more kinds of cells, especially those of animals, can be incorporated into a next version of MIrExpress to more extensively reveal coexpression relationship between gene pairs. Third, a pressing need from a variety of applications is to cluster the genes according to mutual information or its variations in order to find interesting gene groups within which the genes share common functional tasks and

Table 1: Page for most related genes.

The most relevant probe sets to 1007_s_at								
Most relevant probe sets to Gene A (according to MI)			Most relevant probe sets to Gene A (according to r)			Most relevant probe sets to Gene A (according to MIr)		
Probe set	Gene symbol	Pearson's r	Probe set	Gene symbol	Pearson's r	Probe set	Gene symbol	Pearson's r
225437_s_at	C7orf27	1.26033	223460_at	CAMKK1	0.72291	219071_x_at	C8orf30A	0.76873
203028_s_at	CYBA	1.2656	202182_at	KAT2A	0.73306	1570410_at	CYGB	0.77039
229348_at	UBIAD1	1.26796	40359_at	RASSF7	0.73899	48580_at	CXXC1	0.77492
227811_at	FGD3	1.27493	222674_at	C9orf114	0.74328	40359_at	RASSF7	0.77801
206138_s_at	PI4KB	1.2802	1555866_a_at	HEXDC	0.74495	36545_s_at	SFI1	0.78158
36545_s_at	SFI1	1.28182	43977_at	TMEM161A	0.74878	229348_at	UBIAD1	0.79097
1570410_at	CYGB	1.2902	221629_x_at	C8orf30A	0.75543	43977_at	TMEM161A	0.79156
211512_s_at	OGFR	1.29953	208779_x_at	DDR1	0.75607	213681_at	CYHR1	0.79345
203419_at	MLL4	1.35051	213681_at	CYHR1	0.76656	221629_x_at	C8orf30A	0.80694
210749_x_at	DDR1	1.56079	210749_x_at	DDR1	0.90993	210749_x_at	DDR1	1

regulatory mechanisms and thus offer insights into various transcriptional and biological processes.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Luman Wang and Qiaochu Mo are equally contributors.

Acknowledgments

This study was supported by National Nature Science Foundation of China (no. 31470675) and Special Fund for Forest Scientific Research in Public Welfare of China (no. 201404102).

References

[1] T. Barrett, D. B. Troup, S. E. Wilhite et al., "NCBI GEO: archive for high-throughput functional genomic data," *Nucleic Acids Research*, vol. 37, no. 1, pp. D885–D890, 2009.

[2] A. Brazma, H. Parkinson, U. Sarkans et al., "ArrayExpress—a public repository for microarray gene expression data at the EBI," *Nucleic Acids Research*, vol. 31, no. 1, pp. 68–71, 2003.

[3] A. Al-Qahtani, M. Al-Anazi, A. A. Abdo et al., "Correlation between genetic variations and serum level of interleukin 28B with virus genotypes and disease progression in chronic hepatitis C virus infection," *Journal of Immunology Research*, vol. 2015, Article ID 768470, 10 pages, 2015.

[4] N. Nagi-Miura, D. Okuzaki, K. Torigata et al., "CAWS administration increases the expression of interferon γ and complement factors that lead to severe vasculitis in DBA/2 mice," *BMC Immunology*, vol. 14, article 44, 2013.

[5] W. C. Yim, Y. Yu, K. Song, C. S. Jang, and B.-M. Lee, "PLANEX: the plant co-expression database," *BMC Plant Biology*, vol. 13, 83, 2013.

[6] T. Obayashi, K. Kinoshita, K. Nakai et al., "ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in *Arabidopsis*," *Nucleic Acids Research*, vol. 35, no. 1, pp. D863–D869, 2007.

[7] Y. Ogata, H. Suzuki, N. Sakurai, and D. Shibata, "CoP: a database for characterizing co-expressed gene modules with biological information in plants," *Bioinformatics*, vol. 26, no. 9, pp. 1267–1268, 2010.

[8] Z. Fei, J.-G. Joung, X. Tang et al., "Tomato functional genomics database: a comprehensive resource and analysis package for tomato functional genomics," *Nucleic Acids Research*, vol. 39, no. 1, pp. D1156–D1163, 2011.

[9] M. Lescot, P. Déhais, G. Thijs et al., "PlantCARE, a database of plant cis-acting regulatory elements and a portal to tools for in silico analysis of promoter sequences," *Nucleic Acids Research*, vol. 30, no. 1, pp. 325–327, 2002.

[10] T. Obayashi, S. Hayashi, M. Shibaoka, M. Saeki, H. Ohta, and K. Kinoshita, "COXPRESdb: a database of coexpressed gene networks in mammals," *Nucleic Acids Research*, vol. 36, no. 1, pp. D77–D82, 2008.

[11] S. van Dam, T. Craig, and J. P. de Magalhães, "GeneFriends: a human RNA-seq-based gene and transcript co-expression database," *Nucleic Acids Research*, vol. 43, no. 1, pp. D1124–D1132, 2015.

[12] I. Michalopoulos, G. A. Pavlopoulos, A. Malatras et al., "Human gene correlation analysis (HGCA): a tool for the identification of transcriptionally co-expressed genes," *BMC Research Notes*, vol. 5, article 265, 2012.

[13] P. Wang, H. Qi, S. Song et al., "ImmuCo: a database of gene co-expression in immune cells," *Nucleic Acids Research*, vol. 43, no. 1, pp. D1133–D1139, 2015.

[14] N. Gupta and S. Aggarwal, "MIB: using mutual information for biclustering gene expression data," *Pattern Recognition*, vol. 43, no. 8, pp. 2692–2697, 2010.

[15] I. Priness, O. Maimon, and I. Ben-Gal, "Evaluation of gene-expression clustering via mutual information distance measure," *BMC Bioinformatics*, vol. 8, article 111, 2007.

[16] R. Steuer, J. Kurths, C. O. Daub, J. Weise, and J. Selbig, "The mutual information: detecting and evaluating dependencies between variables," *Bioinformatics*, vol. 18, supplement 2, pp. S231–S240, 2002.

[17] 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, article R80, 2004.

[18] P. E. Meyer, F. Lafitte, and G. Bontempi, "minet: A R/bioconductor package for inferring large transcriptional networks using mutual information," *BMC Bioinformatics*, vol. 9, article 461, 2008.

[19] 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.

[20] A. J. Butte and I. S. Kohane, "Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements," *Pacific Symposium on Biocomputing*, vol. 5, pp. 418–429, 2000.

[21] J. Wang, B. Chen, Y. Wang et al., "Reconstructing regulatory networks from the dynamic plasticity of gene expression by mutual information," *Nucleic Acids Research*, vol. 41, no. 8, article e97, 2013.

[22] T. Barrett, S. E. Wilhite, P. Ledoux et al., "NCBI GEO: archive for functional genomics data sets—update," *Nucleic Acids Research*, vol. 41, no. 1, pp. D991–D995, 2013.

[23] D. Sean and P. S. Meltzer, "GEOquery: a bridge between the Gene Expression Omnibus (GEO) and BioConductor," *Bioinformatics*, vol. 23, no. 14, pp. 1846–1847, 2007.

[24] C. L. Wilson and C. J. Miller, "Simpleaffy: a BioConductor package for Affymetrix quality control and data analysis," *Bioinformatics*, vol. 21, no. 18, pp. 3683–3685, 2005.

[25] S. D. Pepper, E. K. Saunders, L. E. Edwards, C. L. Wilson, and C. J. Miller, "The utility of MAS5 expression summary and detection call algorithms," *BMC Bioinformatics*, vol. 8, article 273, 2007.

[26] L. Gautier, L. Cope, B. M. Bolstad, and R. A. Irizarry, "Affy—analysis of Affymetrix GeneChip data at the probe level," *Bioinformatics*, vol. 20, no. 3, pp. 307–315, 2004.

[27] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, no. 3, pp. 379–423, 1948.

[28] A. F. Villaverde, J. Ross, F. Morán, and J. R. Banga, "MIDER: network inference with mutual information distance and entropy reduction," *PLoS ONE*, vol. 9, no. 5, Article ID e96732, 2014.

[29] F. M. Giorgi, G. Lopez, J. H. Woo, B. Bisikirska, A. Califano, and M. Bansal, "Inferring protein modulation from gene expression data using conditional mutual information," *PLoS ONE*, vol. 9, no. 10, Article ID e109569, 2014.

[30] T. Obayashi and K. Kinoshita, "COXPRESdb: a database to compare gene coexpression in seven model animals," *Nucleic Acids Research*, vol. 39, no. 1, pp. D1016–D1022, 2011.

Hindawi Publishing Corporation
Journal of Immunology Research
Volume 2015, Article ID 401956, 14 pages
http://dx.doi.org/10.1155/2015/401956

Research Article

The Role of Aggregates of Therapeutic Protein Products in Immunogenicity: An Evaluation by Mathematical Modeling

Liusong Yin,[1] Xiaoying Chen,[2] Abhinav Tiwari,[2] Paolo Vicini,[3] and Timothy P. Hickling[1]

[1]*Pharmacokinetics, Dynamics and Metabolism-New Biological Entities, Pfizer, Andover, MA 01810, USA*
[2]*Pharmacokinetics, Dynamics and Metabolism-New Biological Entities, Pfizer, Cambridge, MA 02138, USA*
[3]*Pharmacokinetics, Dynamics and Metabolism-New Biological Entities, Pfizer, San Diego, CA 92121, USA*

Correspondence should be addressed to Timothy P. Hickling; timothy.hickling@pfizer.com

Received 31 July 2015; Accepted 7 October 2015

Academic Editor: Marzio Pennisi

Copyright © 2015 Liusong Yin et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Therapeutic protein products (TPP) have been widely used to treat a variety of human diseases, including cancer, hemophilia, and autoimmune diseases. However, TPP can induce unwanted immune responses that can impact both drug efficacy and patient safety. The presence of aggregates is of particular concern as they have been implicated in inducing both T cell-independent and T cell-dependent immune responses. We used mathematical modeling to evaluate several mechanisms through which aggregates of TPP could contribute to the development of immunogenicity. Modeling interactions between aggregates and B cell receptors demonstrated that aggregates are unlikely to induce T cell-independent immune responses by cross-linking B cell receptors because the amount of signal transducing complex that can form under physiologically relevant conditions is limited. We systematically evaluate the role of aggregates in inducing T cell-dependent immune responses using a recently developed multiscale mechanistic mathematical model. Our analysis indicates that aggregates could contribute to T cell-dependent immune response by inducing high affinity epitopes which may not be present in the nonaggregated TPP and/or by enhancing danger signals to break tolerance. In summary, our computational analysis is suggestive of novel insights into the mechanisms underlying aggregate-induced immunogenicity, which could be used to develop mitigation strategies.

1. Introduction

Therapeutic protein products (TPP) from nonhuman, humanized, and human origins include monoclonal antibodies (mAbs), Fc fusion proteins, blood factors, hormones, cytokines, chemokines, and engineered protein scaffolds [1]. They have been widely used to treat a variety of human diseases, including cancer, anemia, hemophilia, rheumatoid arthritis, multiple sclerosis, and inflammatory bowel diseases [1, 2]. Their large success is mainly due to increased target specificity, decreased intrinsic toxicity, and longer half-lives compared with small molecule drugs [3]. These advantages have led to the expansion of TPP in the drug market, with annual revenues of over 100 billion US dollars [1, 2]. However, unwanted immune responses against TPP, such as generation of anti-drug antibodies (ADA), have raised concerns on both drug efficacy and patient safety [4–8]. The effect of ADA on clinical outcomes ranges from no obvious impact to severe loss of efficacy and adverse effects such as infusion reactions [7]. The mechanisms leading to the generation of immunogenicity are yet to be established, but several risk factors have been proposed [9–12], which can be classified as follows: (i) patient-related: genetic background, immunological status, and prior exposure [10], (ii) treatment-related: route, dose, and frequency of administration [7, 13], and (iii) product-related: drug origins, characteristics such as protein structures and aggregates, and formulations [10].

Among these risk factors, aggregates of TPP are of particular concern due to their potential role in inducing both T cell-independent and T cell-dependent immune responses [14–17] (Figure 1). It has been previously found that aggregated recombinant human interferon alpha2b generated by thermal stress, low pH, or oxidization stress is more immunogenic in mice compared with nonaggregated product [18–20]. High immunogenicity in mice has also been observed for aggregates of other TPP, such as human mAbs [21–23],

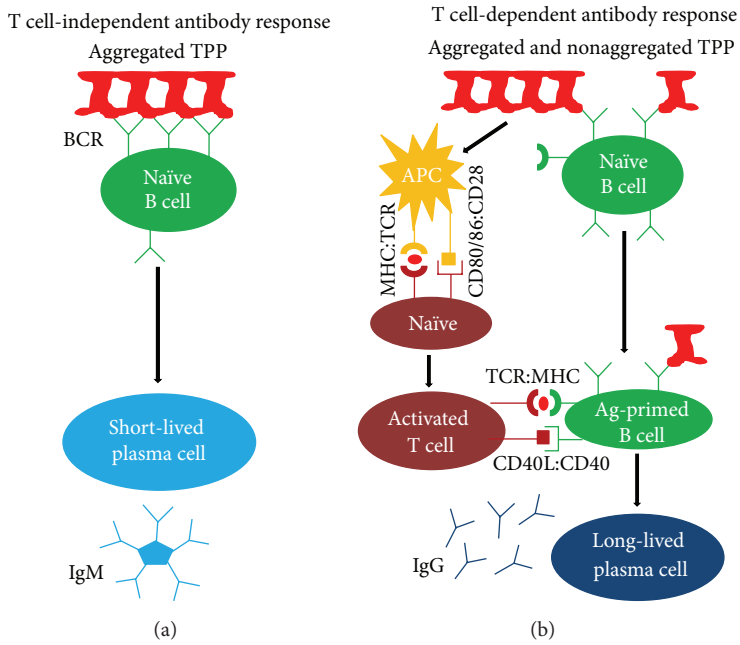


Figure 1: Schematic overview of aggregate-induced T cell-independent and T cell-dependent antibody responses. (a) In the T cell-independent pathway aggregates of TPP cross-link BCRs and activate B cells, which differentiate into short-lived plasma cells that generate antigen-specific IgM pentamers. (b) In the T cell-dependent pathway both aggregated and nonaggregated TPP can be captured by B cells or by APC which present TPP-derived epitopes to activate T cells, which in turn activate antigen-primed B cells. The activated B cells differentiate into long-lived plasma cells that generate isotype-switched IgG.

human epoetin alfa [24], human factor VIII [25, 26], human interferon beta [27], and murine growth hormone [28]. In the clinic, the different ADA incidence rates for several recombinant human interferon beta drugs have been attributed to the differences in aggregation levels [29]. However, the detailed mechanism by which aggregates increase immunogenicity, especially in humans, is yet to be established. For example, it is unknown whether aggregates increase immunogenicity through a T cell-dependent or T cell-independent pathway; and which processes of ADA production could be altered by aggregates is also unknown. In the case of TPP, immunogenicity could be induced through both T cell-dependent and T cell-independent pathways [9, 12]. In the T cell-dependent pathway, antigenic peptides derived from TPP could be presented by major histocompatibility complex class II molecules (MHC II) on antigen-presenting cells (APC) that have been matured by danger signal to stimulate antigen-specific CD4+ T cells. Activated CD4+ T cells would then stimulate antigen-specific B cells that will be responsible for the production of ADA, which are usually affinity matured IgG. It has been found that, in comparison with the nonaggregated form, aggregated mAb results in an increase in the amount of total peptides and the number of epitopes eluted from MHC II [30]. This suggests that aggregates may increase immunogenicity by enhancing antigen processing and presentation in the T cell-dependent pathway. Aggregates could also contribute to T cell-dependent immunogenicity by increasing the danger signal for dendritic cell maturation. Consistent with this, a recent study suggested that aggregated mAb induces significantly higher dendritic cell maturation compared with unstressed mAb [30]. Lastly, aggregates could form repetitively arranged B cell epitopes in a paracrystalline manner to cross-link B cell receptors (BCRs), which in turn will activate antigen-specific B cells to generate ADA, mostly IgM, via the T cell-independent pathway [14]. However, the scarcity of clinical data and the difficulty to isolate the impact of aggregates from other immunogenicity risk factors are major impediments to understand the mechanisms of aggregate-induced immunogenicity.

Mathematical modeling offers the advantage of fast and cost-effective assessment and so it can be used in complement with experimental analysis to study immune responses [31–34]. It also provides quantitative means to dissect each component of a complex response for a deeper understanding of the mechanisms underlying aggregate-induced immunogenicity. Multiple mechanistic mathematical models have been previously developed to study immune responses against various pathogens. For example, antigen processing and presentation by APC and the activation of T helper cells by interactions between T cell receptors and MHC II-peptide complexes have been modeled and the simulation results agree with a variety of experimental data [35]. A mathematical model was also developed for predicting the clonal selection of B cells and antibody production by plasma cells [36]. The role of activation threshold and infections in the dynamics of autoimmune diseases has been studied mathematically as well [37, 38]. Mathematical models have been proposed and experimentally validated for T cell-dependent antibody responses to a wide range of antigens, including *Haemophilus influenzae* type b, hepatitis B virus, cancer antigens, and influenza A virus [39–43]. The T cell-independent activation of B cells by multivalent

hapten-polymer has been modeled, where fitting to experimental data revealed that a minimum number of BCRs, in the range of 7 to 15, need to be cross-linked by a single multivalent ligand to stimulate a B cell [44, 45]. With regards to TPP-induced immunogenicity, several pharmacokinetics (PK) models have been developed to study the impact of ADA on mAb therapy [32]. For example, by incorporating ADA-drug interactions into empirical PK modeling, we developed a PK/ADA model to quantitatively assess the extent and timing of ADA generation, affinity maturation, and ADA-mediated TPP elimination [46]. More recently, we built a mechanistic, multiscale mathematical model of TPP-induced immunogenicity, recapitulating the key processes underlying T cell-dependent generation of ADA, such as antigen presentation, activation of immune cells, and production of ADA as well as *in vivo* disposition of ADA and TPP [47, 48]. This system-level model consists of a subcellular module for antigen presentation, a cellular module for immune system activation and antibody production, and a whole-body module for drug disposition. The model is able to reproduce key immunological phenomena such as antibody affinity maturation and enhanced secondary response [47, 48]. More importantly, a case study on immune response against adalimumab (a fully human anti-TNF alpha IgG1 mAb) showed reasonable agreement between model simulations and experimental observations [47, 48]. Owing to its flexibility and comprehensiveness this system-level model provides us with an ideal platform to probe mechanisms through which aggregates could generate immunogenicity.

In this study, we evaluate whether aggregates could induce T cell-independent or T cell-dependent immune response. In the former case, we model the interactions between multivalent aggregates and BCRs and examine the formation of signal-transducing complex (STC) under physiologically relevant conditions. For the latter case, we use our previously developed system-level model to investigate the impact of antigen processing and presentation, number and affinity of epitopes, and danger signal on ADA production due to aggregates.

2. Materials and Methods

2.1. Aggregates in the T Cell-Independent Pathway: Interactions between Multivalent Aggregates and BCRs. An aggregate (Ag_a) is assumed to be a homogeneous product formed by the combination of n monomers, which gives it a valency of n. The binding of Ag_a to BCR is assumed to be sequential (see Figure 2(a) for an example with $n = 4$) and can be represented by the following second-order reactions:

$$\begin{aligned}
&\mathrm{Ag}_a + \mathrm{BCR} \underset{k_{-1}}{\overset{k_1}{\rightleftharpoons}} \mathrm{Ag}_a\mathrm{BCR}_1 \\
&\mathrm{Ag}_a\mathrm{BCR}_1 + \mathrm{BCR} \underset{k_{-2}}{\overset{k_2}{\rightleftharpoons}} \mathrm{Ag}_a\mathrm{BCR}_2 \\
&\qquad\vdots \\
&\mathrm{Ag}_a\mathrm{BCR}_{n-1} + \mathrm{BCR} \underset{k_{-n}}{\overset{k_n}{\rightleftharpoons}} \mathrm{Ag}_a\mathrm{BCR}_n,
\end{aligned} \tag{1}$$

where k_i and k_{-i} are the ith reaction's binding and dissociation rates, respectively, and $\mathrm{Ag}_a\mathrm{BCR}_i$ is the complex formed by binding of Ag_a to i BCRs. It is assumed that a BCR could bind to any free site on Ag_a and dissociate from any bound site on $\mathrm{Ag}_a\mathrm{BCR}_i$. The above reactions can be described by the following ordinary differential equations that govern the time evolution of $\mathrm{Ag}_a\mathrm{BCR}_i$, Ag_a, and BCR:

$$\begin{aligned}
\frac{d\mathrm{Ag}_a\mathrm{BCR}_1}{dt} &= n \cdot k_1 \cdot \mathrm{BCR} \cdot \mathrm{Ag}_a + 2 \cdot k_{-2} \cdot \mathrm{Ag}_a\mathrm{BCR}_2 \\
&\quad - \left[k_{-1} + (n-1) \cdot k_2 \cdot \mathrm{BCR}\right] \cdot \mathrm{Ag}_a\mathrm{BCR}_1 \\
\frac{d\mathrm{Ag}_a\mathrm{BCR}_i}{dt} &= (n-i+1) \cdot k_i \cdot \mathrm{BCR} \cdot \mathrm{Ag}_a\mathrm{BCR}_{i-1} + (i \\
&\quad + 1) \cdot k_{-(i+1)} \cdot \mathrm{Ag}_a\mathrm{BCR}_{i+1} - \left[ik_{-i} + (n-i) \cdot k_{(i+1)}\right. \\
&\quad \left. \cdot \mathrm{BCR}\right] \cdot \mathrm{Ag}_a\mathrm{BCR}_i, \quad 1 \le i \le n-1 \\
\frac{d\mathrm{Ag}_a\mathrm{BCR}_n}{dt} &= k_n \cdot \mathrm{BCR} \cdot \mathrm{Ag}_a\mathrm{BCR}_{n-1} - n \cdot k_{-n} \\
&\quad \cdot \mathrm{Ag}_a\mathrm{BCR}_n \\
\frac{d\mathrm{Ag}_a}{dt} &= -k_1 \cdot n \cdot \mathrm{BCR} \cdot \mathrm{Ag}_a + k_{-1} \cdot \mathrm{Ag}_a\mathrm{BCR}_1 \\
\frac{d\mathrm{BCR}}{dt} &= -k_1 \cdot n \cdot \mathrm{BCR} \cdot \mathrm{Ag}_a + k_{-1} \cdot \mathrm{Ag}_a\mathrm{BCR}_1 \\
&\quad - \sum_{j=2}^{n} \left(k_j \cdot (n-j+1) \cdot \mathrm{BCR} \cdot \mathrm{Ag}_a\mathrm{BCR}_{j-1} + k_{-j} \cdot j\right. \\
&\quad \left. \cdot \mathrm{Ag}_a\mathrm{BCR}_j\right).
\end{aligned} \tag{2}$$

We selected three (low, medium, and high) physiologically relevant levels for input parameters association constant ($K_a = k_1/k_{-1}$) and initial Ag_a concentration ($[\mathrm{Ag}_a{}^0]$). $[\mathrm{Ag}_a{}^0]$ is Ag_a concentration at $t = 0$, as an initial condition for ordinary differential equations, which is estimated using the following equation:

$$\left[\mathrm{Ag}_a{}^0\right] = [\mathrm{Ag}] \cdot \frac{p}{n}, \tag{3}$$

where [Ag] is the total TPP concentration, p is the aggregation percentage in TPP, and n is the valency of aggregates. [Ag] ranges from 500 to 10^5 pM based on 30 μg dose of interferon beta 1b and 40 mg dose of anti-TNF mAb adalimumab, respectively [29, 47–49]; p spans from 2 to 15% based on a previous report on the characterization and quantitation of aggregates in recombinant human interferon beta drug products [29]; and n varies from 10 to 100 based on the sizes of nonaggregated and aggregated TPP [18, 23, 29, 50, 51]. Taken together, the low and high levels of $\mathrm{Ag}_a{}^0$ are 0.1 and 1500 pM, respectively. The association constant K_a has been previously reported to be 10^{-7} pM^{-1} for antibodies with low intrinsic affinities and 10^{-3} pM^{-1} for affinity matured antibodies, and hence these were selected as low and high levels [52]. The middle levels for total $\mathrm{Ag}_a{}^0$ (12 pM) and K_a (10^{-5} pM^{-1}) are the geometric means of corresponding low and high levels.

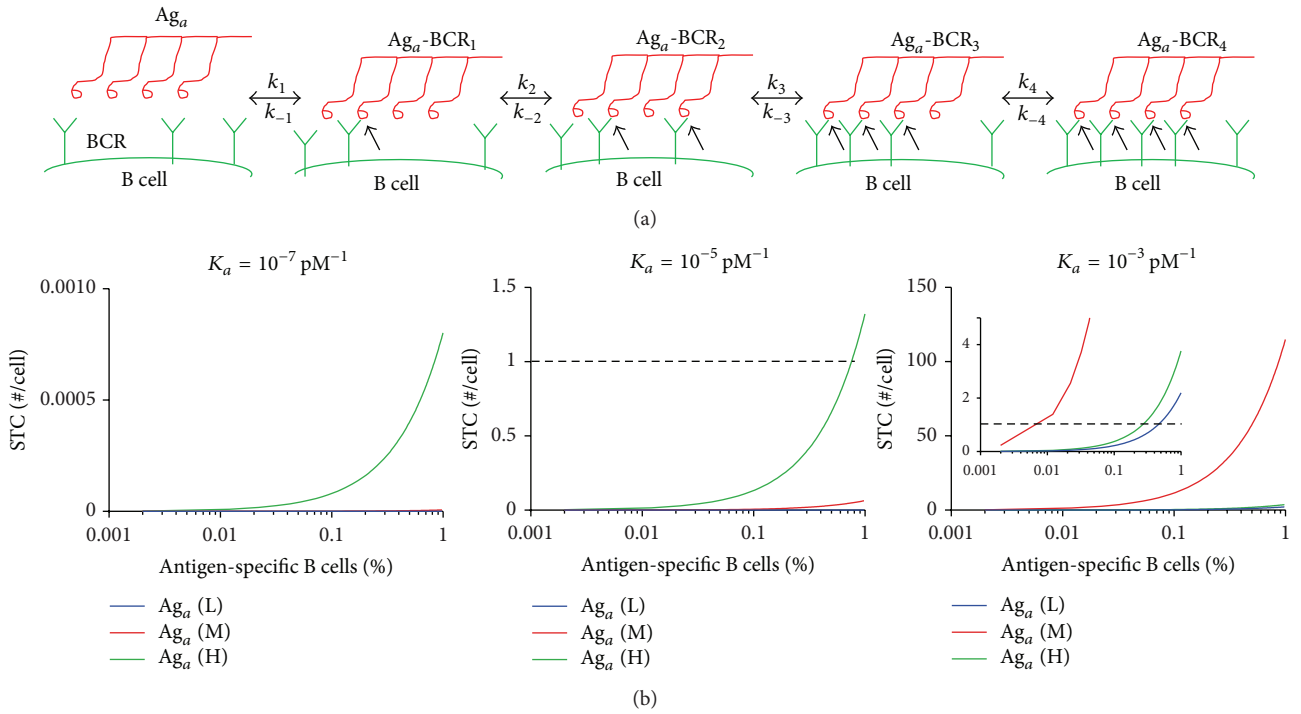


Figure 2: Significant number of STC per cell only forms under limited conditions. (a) Schematic representation of a tetravalent aggregate (Ag_a) binding to BCRs to form Ag_aBCR_i, where i denotes the number of BCRs bound to Ag_a. Black arrow points out the binding of Ag_a to a BCR. Each binding step i is governed by its binding (k_i) and dissociation (k_{-i}) rates. (b) Simulated levels of STC formed per cell are plotted against percentage of antigen-specific B cells under low (L, 0.1 pM), medium (M, 12 pM), and high (H, 1500 pM) levels of total Ag_a, for binding affinity $K_a = 10^{-7}$ pM^{-1} (left panel), 10^{-5} pM^{-1} (middle panel), and 10^{-3} pM^{-1} (right panel). Inset in the right panel is the zoomed-in version of the plot. STC per B cell is defined as the number of aggregates that cross-link a minimum number (s) of BCRs. Here $s = 2$ and valency $n = 100$. The horizontal dashed line denotes one STC.

The rate of binding of an antigen to its corresponding BCR, k_i, is relatively constant [52, 53], so we fixed it to 8.64×10^{-3} pM^{-1} day^{-1}. By contrast, the rate of dissociation (k_{-i}) is expected to increase with i because the resistance of Ag_a against torsion and bending grows due to the steric hindrance from progressive binding of BCRs [45]. For simplicity we assume that k_{-i} decreases exponentially with i and the base for exponential decay is 0.5 as previously identified while modeling interactions between multivalent hapten-polymer and BCRs [45]. The initial BCR concentration is the product of number of BCRs per cell, B cell concentration, and percentage of antigen-specific B cells. The number of BCRs per cell and B cell concentration have been previously reported as ~10^5 and ~10^8 L^{-1}, respectively [41, 44, 45, 47, 48]. Studies on the percentage of antigen-specific B cells are limited, but it has been reported to be <0.002% for vaccinia virus [54] and <1% for individual antigens [55]. The above estimates were used to define the input range of BCR concentration at $t = 0$ as an initial condition for the ordinary differential equations in the simulation.

In the model, the STC is the number of Ag_a that cross-links at least s BCRs as defined in [44, 45]:

$$\text{STC} = \sum_{s}^{n} \text{Ag}_a\text{BCR}_s. \quad (4)$$

The model was simulated using the ordinary differential equation solver *ode15s* in MATLAB (The MathWorks, Inc., Natick, MA).

2.2. Aggregates in the T Cell-Dependent Pathway: Impact on Antigen Processing and Presentation and Danger Signal. For this analysis, we use our previously developed mechanistic, multiscale mathematical model for T cell-dependent ADA production [47, 48]. In this system-level model aggregates could contribute to increased ADA production by enhancing either the antigen processing and presentation or the danger signal for dendritic cell maturation (denoted by red arrows in Figure 3). We simulate the impact of aggregates by increasing (i) the rate of internalization of TPP into the endosome, (ii) the rate of degradation/processing of TPP into antigenic peptides, (iii) the number of epitopes generated, (iv) the affinity of epitopes to MHC II, and (v) the level of danger signal. Subsequently, for each of these conditions, we examine the endosomal levels of aggregates and epitope, the number of MHC II-peptide complexes on APC, and the levels of ADA production. To simulate B cell clonal selection and antibody affinity maturation, B cells and ADA are divided into 17 subgroups based on the binding affinity to antigen [36, 47, 48]. In our analysis, we define ADA production as the sum of the 17 subgroups.

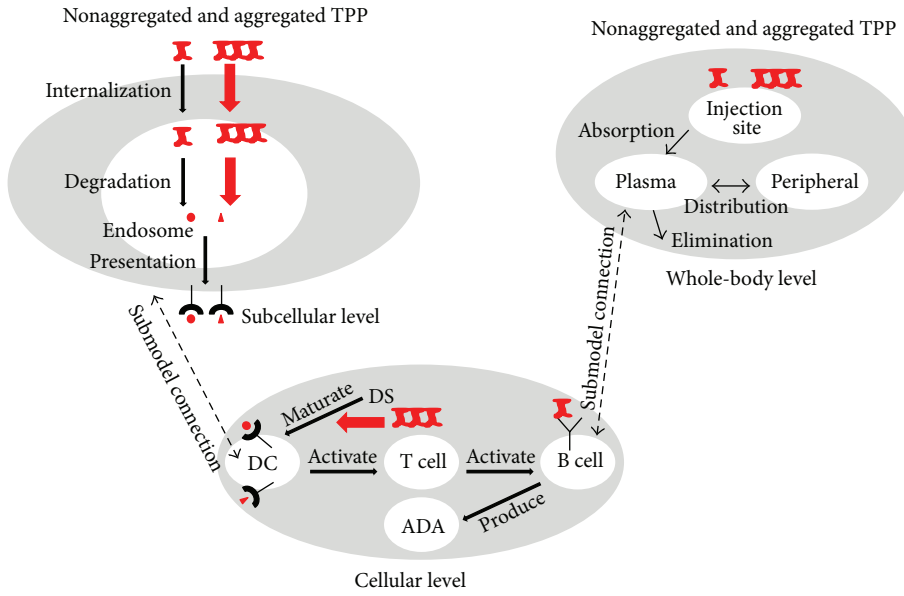


FIGURE 3: Schematic highlighting of the potential role of aggregates in T cell-dependent ADA production. A recapitulation of our system-level model for T cell-dependent ADA production [47, 48]. At the subcellular level, TPP are internalized into endosome of APC, such as dendritic cells (DC), and then degraded into antigenic peptides. Epitopes derived from TPP could be loaded onto MHC II and presented on the surface of APC. Aggregates could contribute to enhanced ADA production by having increased internalization or degradation rate or number and affinity of epitopes generated (indicated by thick red arrows). At the cellular level, danger signal (DS) maturated DC activate T cells which in turn activate B cells to generate ADA. Aggregates could enhance the DS to maturate DC (see red arrow). At the whole-body level, aggregated and nonaggregated TPP are absorbed from the injection site into plasma and will be distributed into periphery, eliminated, or captured by B cells through BCR binding.

3. Results

3.1. Aggregates Are Unlikely to Induce T Cell-Independent Immune Response because the Number of STC Formed Is Limited. To evaluate whether aggregates could induce T cell-independent antibody responses through BCR cross-linking, we examine the number of STC formed per B cell for different parameter combinations (see Section 2 for details). The model output for interactions between aggregates and BCR is the STC formed per B cell, which was previously defined as the number of Ag_a which cross-links a minimum number of BCRs [44, 45]. It has been reported that a multivalent ligand stimulates B cell activation only if it cross-links a minimum number (s) of BCRs, which is usually between 7 and 15 [44, 45]. We calculated the number of STC for s = 2, 5, and 10 under different total Ag_a, K_a, and BCR levels. Surprisingly, our computer simulation analysis showed that if s = 10 or 5, no more than one STC per cell could be observed under physiological levels of total Ag_a, BCR, and K_a (data not shown). Even if s is lowered to 2, more than one STC per cell can form only under limited conditions, when the sensitive parameters are near the upper limits of the physiologically plausible ranges (Figure 2(b)). In the case of $K_a = 10^{-7}$ pM^{-1}, no more than one STC could form (Figure 2(b), left panel). For $K_a = 10^{-5}$ pM^{-1}, more than one STC could form at high levels of total Ag_a (1.5×10^{-3} pM) but only near the upper limit of antigen-specific B cells percentage (1%) (Figure 2(b), middle panel). Finally, when $K_a = 10^{-3}$ pM^{-1}, more than one STC could form at all total Ag_a levels but only with antigen-specific B cell percentage >0.006% (Figure 2(b), right panel). These results from our computer simulation showed that STC per cell is very sensitive to K_a and total concentrations of Ag_a and BCRs (but not to binding rate k_i, data not shown). Overall, this analysis suggests that aggregates are unlikely to induce T cell-independent activation of B cells and consequent ADA production under physiologically plausible conditions. Therefore, aggregates may only contribute to ADA production through a T cell-dependent pathway, which we explore next.

3.2. Aggregates Could Enhance ADA Production by Increasing the Danger Signal to Maturate Dendritic Cells. To evaluate the T cell-dependent effect of aggregates on ADA production, we modulated those parameters in our system-level immunogenicity model [47, 48] that may be impacted by aggregation. This model consists of a subcellular module for antigen presentation in APC, a cellular module for immune cell activation and ADA production, and a whole-body module for drug and ADA disposition (Figure 3). Aggregates have been previously shown to increase danger signal for dendritic cell maturation and T cell activation [12, 22, 30, 56]. Specifically, aggregated mAb upregulated the dendritic cell maturation marker CD83 and CD4+ T cell costimulatory molecules CD80 and CD86 as well as cytokines produced by CD4+ T cells, such as IL-2 and IL-10 [30, 56]. Due to the complexity of dendritic cell maturation by danger signal and

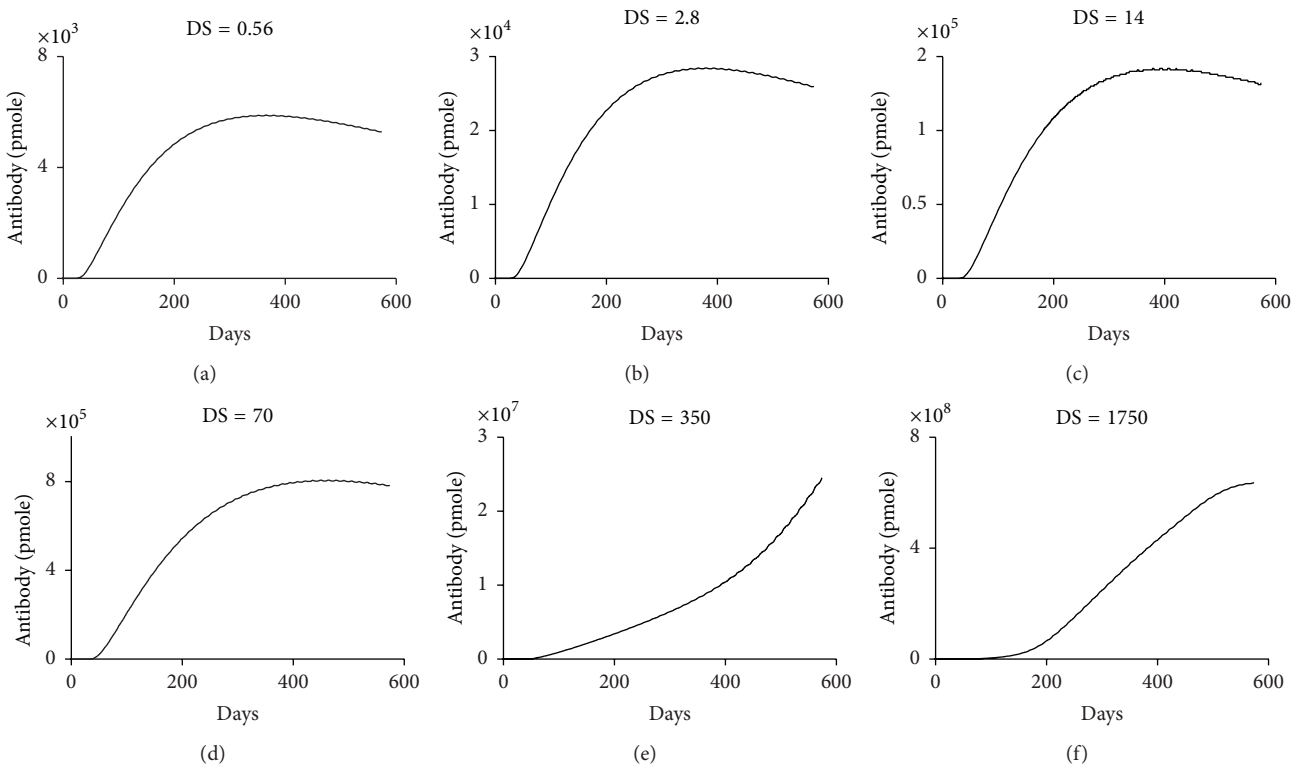


Figure 4: Aggregates could contribute to ADA production by increasing danger signal to maturate dendritic cells. ((a)–(f)) Simulated ADA production is shown at various levels of danger signal (DS) which is modeled as the amount of LPS in ng. Remaining parameter values are the same as in the original simulation for nonaggregated adalimumab [47, 48]. DS = 350 ng LPS shows the original simulation for nonaggregated adalimumab [47, 48]. ADA production is shown as the sum of the 17 subgroups. Dose = 40 mg administered biweekly.

the unavailability of many parameters associated with this process, it is simply modeled as being driven by endotoxin lipopolysaccharide (LPS) [47, 48]. LPS is widely used in immunological studies for dendritic cell maturation [57–61] and is present in many TPP [62]. The cytokine profiles induced by LPS and aggregates of mAb are very similar [22, 63]. Using our system-level model, we previously simulated ADA production induced by adalimumab, a fully anti-TNF alpha IgG1 mAb used to treat various inflammatory and autoimmune diseases, with a danger signal of 350 ng LPS [47] (Figure 4(e)). If aggregates increase the danger signal by 5-fold, ADA production is increased by 20-fold (Figure 4(f)). We also simulated ADA production for low danger signal levels (Figures 4(a)–4(d)) as the actual amount induced by nonaggregated TPP is unknown. In essence, ADA production depends on the level of danger signal (Figures 4(a)–4(f)). Therefore, our simulations suggest aggregates could enhance ADA production by increasing danger signal to enhance maturation of dendritic cells and subsequently activate T cells.

3.3. Aggregates Could Not Enhance ADA Production by Increasing Antigen Processing and Presentation If High Affinity Epitopes Are Already Present in Nonaggregated TPP. Antigen processing and presentation are the key events in T cell-dependent immunogenicity of TPP [12]. Previous studies have demonstrated that aggregation enhances antigen's uptake, processing, and presentation by APC [12, 22, 30, 56, 64]. More recently, a study showed that aggregated mAb could directly increase the total number of different peptides and the number of epitopes presented by MHC II compared with nonaggregated mAb [30]. To evaluate whether aggregation-enhanced antigen processing and presentation could increase ADA production, we simulated these effects of aggregates in our model by changing its internalization or degradation rate or the number and affinity of epitopes generated and assessing their impact on final ADA production.

We previously simulated ADA production induced by adalimumab with an internalization rate of 14.4 day^{-1} (IR_0), a degradation rate of 17.28 day^{-1} (DR_0), and two predicted adalimumab epitopes with high binding affinities of 123 and 85 nM to common MHC II allele DRB1*04:01 [47]. To model the aggregates' effect on antigen processing, we increased either internalization or degradation rate by 16.6-fold based on a previous study which reported that aggregated mAb resulted in a 16.6-fold increase in total peptides associated with MHC II [30] and then assessed the levels of endosomal aggregates and epitopes, MHC II-peptide complexes on cell surface, and ADA production. As expected, conditional on the parameters and structure of the model simulation, increasing internalization rate by 16.6-fold resulted in a similar fold increase in aggregates internalized into endosome and epitopes generated by its degradation (Figures 5(a)-5(b) and 5(e)-5(f)). Increasing degradation rate by 16.6-fold resulted

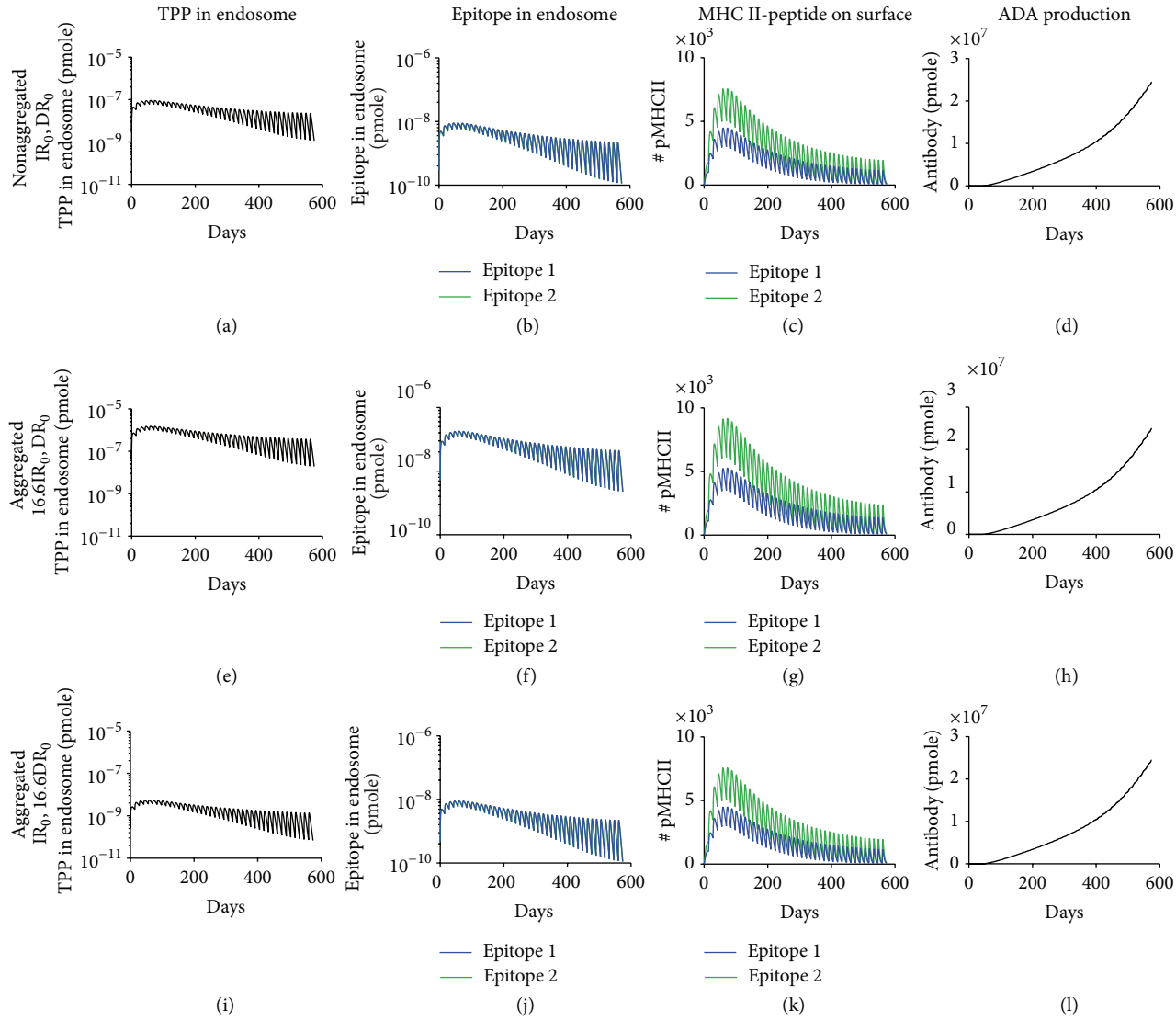


Figure 5: Aggregates could not enhance ADA production through faster antigen internalization or degradation if high affinity epitopes are already present in nonaggregated TPP. Simulated levels of nonaggregated and aggregated TPP in endosome, epitopes in endosome, MHC II-peptide complex on cell surface, and ADA production are shown for ((a)–(d)) original internalization ($IR_0 = 14.4\ day^{-1}$) and degradation ($DR_0 = 17.28\ day^{-1}$) rate for nonaggregated adalimumab [47, 48], ((e)–(h)) $16.6IR_0$ and DR_0 for hypothetical aggregated form, and ((i)–(l)) IR_0 and $16.6DR_0$ for hypothetical aggregated form. ADA production has the same definition and dose has the same value as in Figure 4.

in the same fold decrease in endosomal aggregates, but the levels of epitopes were unchanged, which suggested that epitope generation was limited by the amount of aggregates internalized and not by the degradation rate (Figures 5(a)-5(b) and 5(i)-5(j)). Moreover, increasing internalization or degradation rate by 16.6-fold did not significantly change the number of MHC II-peptide complexes presented on the surface of APC (Figures 5(c), 5(g), and 5(k)). Aggregates could also impact the FcR binding and potentially affect the antigen uptake [44]. We therefore evaluated a larger range of internalization and degradation rate. Our conclusions were unaffected by larger increases (200-fold) in internalization or degradation rate (data not shown). Consistent with MHC II-peptide complex presentation levels, increasing internalization or degradation rate by 16.6-fold had little impact on final ADA production (Figures 5(d), 5(h), and 5(l)). We next modeled the effect of aggregates on the number of epitopes presented. As expected, including aggregate-induced generation of new epitopes led to the surface presentation of corresponding MHC II-peptide complexes whose levels depend on the binding affinity of epitope to MHC II (Figures 6(a)–6(c), 6(e)–6(g), and 6(i)–6(k)). Surprisingly, if two high affinity epitopes are already present, then the inclusion of new epitopes did not increase ADA production (Figures 6(d), 6(h), and 6(l)). Taken together, these analyses suggest that aggregate-induced high antigen processing and presentation

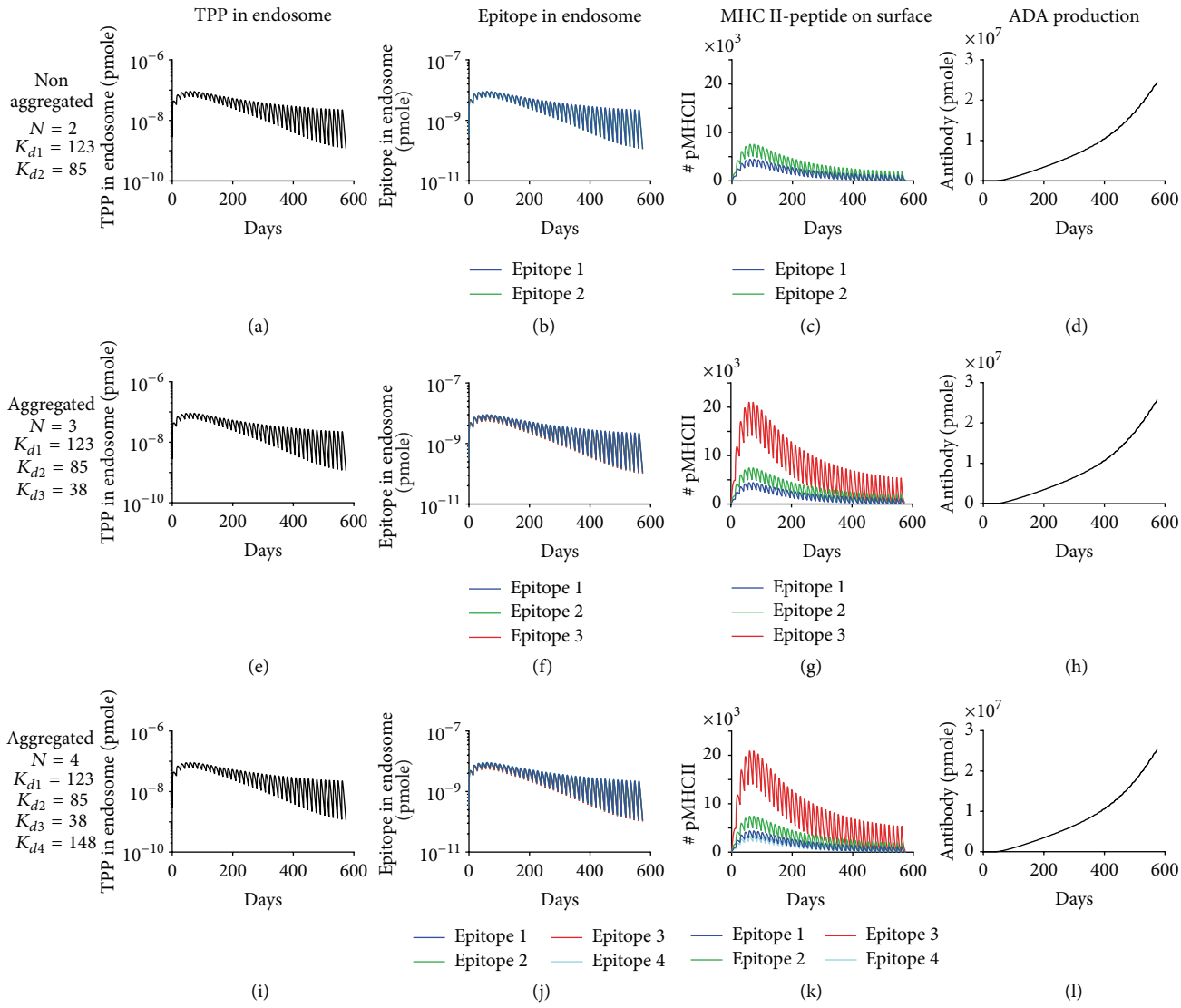


Figure 6: Aggregates could not enhance ADA production by increasing number of epitopes if high affinity epitopes are already present in nonaggregated TPP. Simulated levels of TPP in endosome, epitope in endosome, MHC II-peptide complex on cell surface, and ADA production are shown for ((a)–(d)) original two predicted epitopes for nonaggregated adalimumab [47, 48], ((e)–(h)) three epitopes for hypothetical aggregated form, and ((i)–(l)) four epitopes for hypothetical aggregated form. The predicted dissociation constant (K_d, unit: nM) for binding of each epitope to MHC II is indicated. ADA production has the same definition and dose has the same value as in Figure 4.

cannot enhance ADA production if high affinity epitopes are already present.

3.4. Aggregates Could Enhance ADA Production by Inducing the Presentation of Epitopes with Higher Affinities than Those from Nonaggregated TPP. MHC II-restricted epitopes are generated with μM to nM affinity range [65]. We next evaluated whether aggregate-induced high antigen processing and presentation could increase immunogenicity when nonaggregated TPP present low affinity (μM range) epitopes. We started with 40 mg dose of nonaggregated TPP administered biweekly and two epitopes with K_d of 1230 and 850 nM representing low affinity epitopes of μM range [65, 66] and monitored the number of MHC II-peptide complexes on surface of APC and ADA production (Figures 7(a)–7(d)). We next increased the internalization rate by 16.6-fold to mimic the effect of aggregates and again saw no increase in antigen presentation and ADA production (Figures 7(e)–7(h)). Notably, when aggregates induced the presentation of a high affinity epitope (K_d = 38 nM), ADA production increased by >4-fold (Figure 7(l)) due to enhanced antigen presentation (Figures 7(i)–7(k)). We further evaluated the effect of aggregate-induced high affinity epitopes on ADA production under different dose levels, all of which demonstrated that induction of a high affinity epitope could significantly increase ADA production (compare top and bottom rows in Figure 8), whereas increase in internalization rate had no effect (compare top and middle rows in Figure 8).

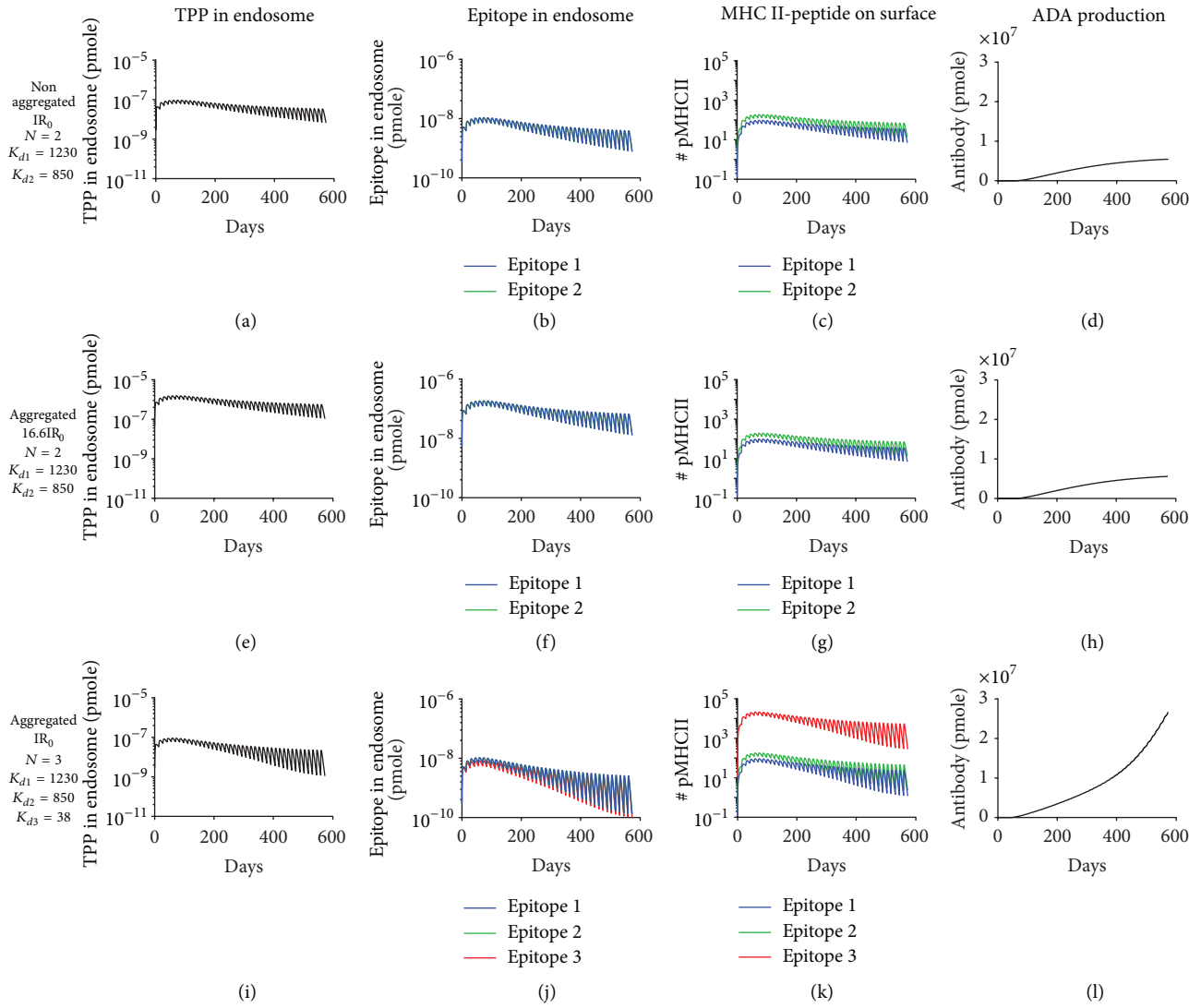


FIGURE 7: Aggregation could contribute to ADA production by inducing the presentation of high affinity epitopes that may not be present in nonaggregated TPP. Simulated levels of TPP in endosome, epitope in endosome, MHC II-peptide complex on cell surface, and ADA production are shown for ((a)–(d)) original internalization rate (IR_0) and two low affinity epitopes for hypothetical nonaggregated TPP, ((e)–(h)) 16.6IR_0 and two low affinity epitopes for hypothetical aggregated form, and ((i)–(l)) IR_0 and inclusion of a high affinity third epitope for hypothetical aggregated form. The predicted dissociation constant (K_d, unit: nM) for binding of each epitope to MHC II is indicated. ADA production has the same definition and dose and IR_0 have the same values as in Figure 4.

These computational modeling results indicate that aggregates could contribute to ADA production by inducing the presentation of high affinity epitopes that may not be present in nonaggregated TPP.

4. Discussion

In this study, we used mathematical modeling to comprehensively evaluate mechanisms through which aggregates of TPP could contribute to immunogenicity. By modeling the interactions between aggregates and BCRs, we find that aggregates are unlikely to induce T cell-independent antibody responses through BCR cross-linking due to the limited number of STC that could form under physiologically plausible conditions. Thereafter, using our previously developed multiscale, mechanistic mathematical model for the T cell-dependent induction of ADA by TPP, we systematically evaluated the potential roles of aggregates in ADA generation by dissecting the individual steps leading to it. Our analyses indicate that aggregates could contribute to immunogenicity by increasing the danger signal to maturate dendritic cells and activate T cells and/or by inducing the presentation of high affinity epitopes that may not be present in nonaggregated TPP.

TPP could aggregate during manufacturing, storage, handling, or delivery to patients due to agitation, light exposure, temperature elevation, oxidation, pH change, and leaching [12, 17, 23, 24, 29, 30, 56, 67]. Aggregation has been proposed

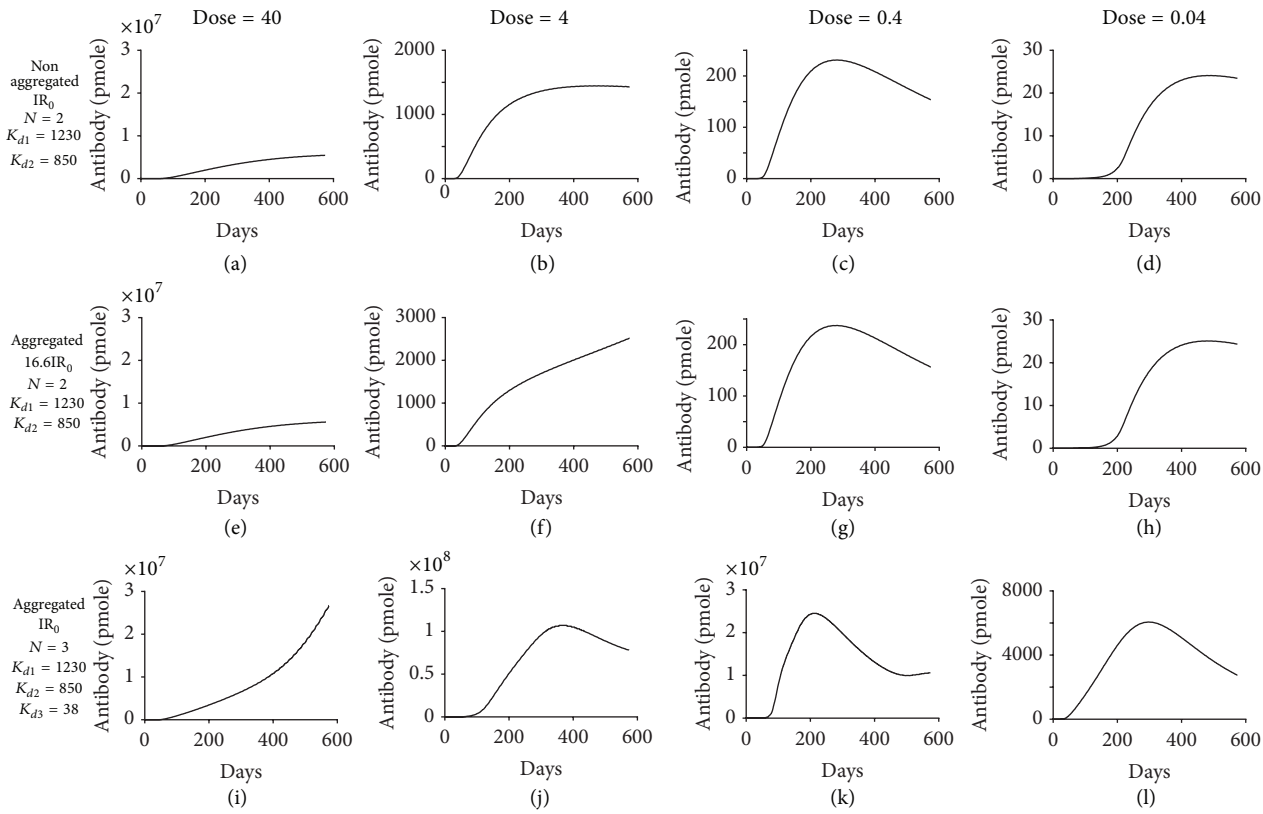


Figure 8: Aggregation could contribute to immunogenicity by inducing the presentation of high affinity epitopes that may not be present in nonaggregated TPP under a variety of drug doses. Simulated ADA production is shown for the same conditions as in Figure 7 for biweekly administered dose of 40 mg ((a), (e), and (i)), 4 mg ((b), (f), and (j)), 0.4 mg ((c), (g), and (k)), and 0.04 mg ((d), (h), and (l)). ADA production has the same definition as in Figure 4.

as a strong risk factor for TPP-induced immunogenicity due to its potential role in both T cell-independent and T cell-dependent antibody responses [10, 12, 14, 16, 17]. Several previous studies in mice have demonstrated that for different TPP aggregates induced a stronger ADA production compared with nonaggregated forms [18–21, 25, 27, 28]. However, the mechanisms underlying aggregate-induced ADA production are not clear. A recent study in mice transgenic for human IgG demonstrated that only light-induced oligomers of IgG induced an immune response, which was ablated by the depletion of CD4+ cells [66]. The data from this mouse model are in agreement with the mathematical model in which aggregates induce immune responses in a T cell-dependent manner.

Repetitively arranged epitopes in a paracrystalline structure of viral particles could cross-link BCRs to induce T cell-independent IgM or in some cases IgG3 responses [68–72]. It is expected that aggregates of TPP, potentially resembling the structure of highly repetitive epitopes, could induce T cell-independent antibody responses in a similar way [12, 16, 17]. The model does not directly consider the nature of a polyclonal B cell response, but it is consistent with that. Specifically, multiple epitopes from the aggregates being bound by the BCR can be represented by the differential binding rate constants in the model, and the different number of B cell epitopes on aggregates can be captured by the complex forming between aggregates and various number of BCRs. Surprisingly, by modeling the interactions between aggregates and BCRs, we find that aggregates are unlikely to induce T cell-independent antibody responses because only a few STC can form under physiologically plausible conditions for antigen-specific B cells, antigen dose, and binding affinity (Figure 2(b), left and center panels). This is consistent with previous studies in mice that showed no significant T cell-independent IgG3 antibody response against aggregated recombinant murine growth hormone [28] or anti-TNFα murine mAb [23], although IgM production was not evaluated in either case. However, it should be noted that, under conditions of high binding affinity and BCR concentration and appropriate antigen concentrations, significant number of STC could form, with a potential to induce T cell-independent antibody response (Figure 2(b), right panel). High BCR concentration can be achieved through high percentage of antigen-specific B cells for particular TPP or through B cell proliferation due to lowering of activation threshold by cytokines [73], second messenger diacylglycerol [74], costimulatory signal [75], or Bruton's tyrosine kinase [76]. Appropriate antigen concentration can result from specific dosing strategies. Therefore, particular attention should be given while administering TPP to patients in those

conditions. Future experiments directly investigating the downstream signaling of BCR cross-linking in the presence of aggregated and nonaggregated TPP and studies evaluating whether T cell-independent IgM is induced in response to aggregates can further elucidate the role of aggregates in T cell-independent ADA production.

T cell-dependent ADA production is thought to be the major pathway through which TPP induce immunogenicity as in the case of IgG1 and IgG4 generated against anti-TNFα mAb to treat a variety of inflammatory and autoimmune diseases [7]. Antigen processing and presentation by professional APC, such as dendritic cells, macrophages, and B cells, play a key role in T cell-dependent antibody responses [12]. It has been shown that aggregates could enhance antigen uptake by APC thereby increasing peptides associated with MHC II and could induce dendritic cell maturation and T cell activation [22, 30, 56]. However, human data directly ascribing ADA levels to aggregates are still lacking. In this study, we systematically evaluated whether aggregate-enhanced antigen processing and presentation could increase ADA production. Our computer simulation suggests that the amount of antigenic peptides in endosome is limited by antigen internalization rate, not degradation rate, and the number of MHC II-peptide complexes presented on cell surface is mainly restricted by the binding affinity of epitopes. Our modeling analyses indicate that induction of high affinity epitopes by aggregates that may not be present in nonaggregated TPP and increased danger signal by aggregates to maturate dendritic cells could result in increased ADA production (Figures 4–8). A specifically designed experimental study that examines the binding affinities of peptides to MHC II derived from dendritic cells treated with aggregated or nonaggregated TPP would verify whether aggregates can induce the presentation of high affinity epitopes not present in nonaggregated TPP. In this work, we modeled aggregates-induced danger signal as LPS. However, it should be noted that aggregates have the potential to bind to a variety of pattern recognition receptors as well as FcR. Therefore the kinetics, activation thresholds, and receptors engaged by aggregates are more diverse and complicated than those of LPS and need further investigation.

This work improves our understanding of aggregate-induced immunogenicity and could be utilized to develop prediction and mitigation strategies. Overall, this modeling study suggests that aggregates could enhance immunogenicity; therefore enough attention should be given to reduce aggregation during manufacturing, storage, handling, and administration. In particular, potential high affinity CD4+ T cell epitopes are of great concern because their presentation in nonaggregated TPP will result in high levels of immunogenicity regardless of aggregation. On the other hand, even if they are not presented in nonaggregated TPP, an aggregation-induced presentation will also result in enhanced immunogenicity. Thus, efforts should be made towards experimental identification or *in silico* prediction of high affinity epitopes during immunogenicity assessment, and potential high affinity epitopes should be avoided while designing novel TPP as they carry a strong risk for ADA generation.

Our recently developed mechanistic system-level mathematical model for ADA production is a useful tool to evaluate human immunogenicity against TPP as it incorporates protein-specific antigenic properties and host-specific immunological characteristics, although further experimental validation is needed to increase confidence in ADA predictions [47, 48]. Multiple product- and patient-related risk factors have been proposed to impact immunogenicity of TPP [7, 8, 10–14, 77, 78]. As confidence in its properties increases, this system-level model could potentially be used to design new hypotheses and study other risk factors besides aggregation. For example, though the model is developed for healthy subjects, it can be easily modified to account for the effect of different disease statuses. For example, the profile of ADA generation observed in autoimmune patients [79, 80] can be simulated by including either a lower activation threshold for immune cells [37, 38] or preexisting immunity against TPP [79, 80]. Also, peptide editor HLA-DM plays a key role in MHC II antigen presentation and CD4+ T cell epitope selection by favoring the presentation of peptides with higher kinetic stabilities [65, 81–84]. To evaluate the effect of HLA-DM-mediated epitope selection on ADA production, it could be included in the subcellular module of antigen processing and presentation to select the epitopes presented based on peptide susceptibility to HLA-DM-mediated peptide exchange [84]. Other ADA production impact factors that could be evaluated by this system-level model include time delays between administration, immune cell activation and migration from tissue to lymphoid compartments [42], contraction of effector B cells and T cells [85, 86], effect of immunomodulators through comedication [87], and different antibody isotypes generated by short- and long-lived plasma cells [42, 88, 89]. Therefore, this model could generate new hypotheses about immunogenicity and could be used with experiments to decipher the mechanisms underlying immunogenicity of TPP and develop corresponding mitigation strategies.

5. Conclusion

In summary, our computational analyses suggest that aggregates are unlikely to induce T cell-independent antibody responses through BCR cross-linking due to limited formation of STC under physiologically plausible conditions. In contrast, aggregates could contribute to immunogenicity via the T cell-dependent pathway by inducing the presentation of high affinity epitopes that may not be present in nonaggregated TPP and/or by enhancing danger signal to maturate dendritic cells and activate T cells. This study provides novel insights into how aggregates could contribute to overall immunogenicity and suggests novel mechanistic hypotheses eventually suitable for experimental testing.

Disclosure

Paolo Vicini is currently working at Clinical Pharmacology and DMPK, MedImmune, Cambridge CB21 6GH, UK.

Conflict of Interests

All authors are current or former employees of Pfizer Inc.

Authors' Contribution

Xiaoying Chen and Abhinav Tiwari contributed to the work equally.

Acknowledgments

This work was supported by a Pfizer Worldwide Research and Development Postdoctoral Fellowship.

References

[1] D. S. Dimitrov, "Therapeutic proteins," in *Therapeutic Proteins*, vol. 899 of *Methods in Molecular Biology*, pp. 1–26, Humana Press, 2012.

[2] G. Walsh, "Biopharmaceutical benchmarks 2014," *Nature Biotechnology*, vol. 32, no. 10, pp. 992–1000, 2014.

[3] V. Brinks, D. Weinbuch, M. Baker et al., "Preclinical models used for immunogenicity prediction of therapeutic proteins," *Pharmaceutical Research*, vol. 30, no. 7, pp. 1719–1728, 2013.

[4] R. T. Purcell and R. F. Lockey, "Immunologic responses to therapeutic biologic agents," *Journal of Investigational Allergology & Clinical Immunology*, vol. 18, no. 5, pp. 335–342, 2008.

[5] V. Jawa, L. P. Cousens, M. Awwad, E. Wakshull, H. Kropshofer, and A. S. De Groot, "T-cell dependent immunogenicity of protein therapeutics: preclinical assessment and mitigation," *Clinical Immunology*, vol. 149, no. 3, pp. 534–555, 2013.

[6] J. R. Maneiro, E. Salgado, and J. J. Gomez-Reino, "Immunogenicity of monoclonal antibodies against tumor necrosis factor used in chronic immune-mediated inflammatory conditions: systematic review and meta-analysis," *JAMA Internal Medicine*, vol. 173, no. 15, pp. 1416–1428, 2013.

[7] P. A. van Schouwenburg, T. Rispens, and G. J. Wolbink, "Immunogenicity of anti-TNF biologic therapies for rheumatoid arthritis," *Nature Reviews Rheumatology*, vol. 9, no. 3, pp. 164–172, 2013.

[8] G. Shankar, S. Arkin, L. Cocea et al., "Assessment and reporting of the clinical immunogenicity of therapeutic proteins and peptides—harmonized terminology and tactical recommendations," *The AAPS Journal*, vol. 16, no. 4, pp. 658–673, 2014.

[9] A. S. De Groot and D. W. Scott, "Immunogenicity of protein therapeutics," *Trends in Immunology*, vol. 28, no. 11, pp. 482–490, 2007.

[10] S. K. Singh, "Impact of product-related factors on immunogenicity of biotherapeutics," *Journal of Pharmaceutical Sciences*, vol. 100, no. 2, pp. 354–387, 2011.

[11] C. Krieckaert, T. Rispens, and G. Wolbink, "Immunogenicity of biological therapeutics: from assay to patient," *Current Opinion in Rheumatology*, vol. 24, no. 3, pp. 306–311, 2012.

[12] S. Sethu, K. Govindappa, M. Alhaidari, M. Pirmohamed, K. Park, and J. Sathish, "Immunogenicity to biologics: mechanisms, prediction and reduction," *Archivum Immunologiae et Therapiae Experimentalis*, vol. 60, no. 5, pp. 331–344, 2012.

[13] A. C. Moss, V. Brinks, and J. F. Carpenter, "Review article: immunogenicity of anti-TNF biologics in IBD—the role of patient, product and prescriber factors," *Alimentary Pharmacology & Therapeutics*, vol. 38, no. 10, pp. 1188–1197, 2013.

[14] S. Kumar, S. K. Singh, X. Wang, B. Rup, and D. Gill, "Coupling of aggregation and immunogenicity in biotherapeutics: T- and B-cell immune epitopes may contain aggregation-prone regions," *Pharmaceutical Research*, vol. 28, no. 5, pp. 949–961, 2011.

[15] S. Kumar, M. A. Mitchell, B. Rup, and S. K. Singh, "Relationship between potential aggregation-prone regions and HLA-DR-binding T-cell immune epitopes: implications for rational design of novel and follow-on therapeutic antibodies," *Journal of Pharmaceutical Sciences*, vol. 101, no. 8, pp. 2686–2701, 2012.

[16] K. D. Ratanji, J. P. Derrick, R. J. Dearman, and I. Kimber, "Immunogenicity of therapeutic proteins: influence of aggregation," *Journal of Immunotoxicology*, vol. 11, no. 2, pp. 99–109, 2014.

[17] M. Sauerborn, V. Brinks, W. Jiskoot, and H. Schellekens, "Immunological mechanism underlying the immune response to recombinant human protein therapeutics," *Trends in Pharmacological Sciences*, vol. 31, no. 2, pp. 53–59, 2010.

[18] S. Hermeling, L. Aranha, J. M. A. Damen et al., "Structural characterization and immunogenicity in wild-type and immune tolerant mice of degraded recombinant human interferon alpha2b," *Pharmaceutical Research*, vol. 22, no. 12, pp. 1997–2006, 2005.

[19] S. Hermeling, H. Schellekens, C. Maas, M. F. B. G. Gebbink, D. J. A. Crommelin, and W. Jiskoot, "Antibody response to aggregated human interferon alpha2b in wild-type and transgenic immune tolerant mice depends on type and level of aggregation," *Journal of Pharmaceutical Sciences*, vol. 95, no. 5, pp. 1084–1096, 2006.

[20] P. Human, H. Ilsley, C. Roberson et al., "Assessment of the immunogenicity of mechanically induced interferon aggregates in a transgenic mouse model," *Journal of Pharmaceutical Sciences*, vol. 104, no. 2, pp. 722–730, 2015.

[21] V. Filipe, W. Jiskoot, A. H. Basmeleh, A. Halim, H. Schellekens, and V. Brinks, "Immunogenicity of different stressed IgG monoclonal antibody formulations in immune tolerant transgenic mice," *mAbs*, vol. 4, no. 6, pp. 740–752, 2012.

[22] M. K. Joubert, M. Hokom, C. Eakin et al., "Highly aggregated antibody therapeutics can enhance the in vitro innate and late-stage T-cell immune responses," *The Journal of Biological Chemistry*, vol. 287, no. 30, pp. 25266–25279, 2012.

[23] A. J. Freitag, M. Shomali, S. Michalakis et al., "Investigation of the immunogenicity of different types of aggregates of a murine monoclonal antibody in mice," *Pharmaceutical Research*, vol. 32, no. 2, pp. 430–444, 2015.

[24] A. Seidl, O. Hainzl, M. Richter et al., "Tungsten-induced denaturation and aggregation of epoetin alfa during primary packaging as a cause of immunogenicity," *Pharmaceutical Research*, vol. 29, no. 6, pp. 1454–1467, 2012.

[25] V. S. Purohit, C. R. Middaugh, and S. V. Balasubramanian, "Influence of aggregation on immunogenicity of recombinant human factor VIII in hemophilia A mice," *Journal of Pharmaceutical Sciences*, vol. 95, no. 2, pp. 358–371, 2006.

[26] D. S. Pisal, M. P. Kosloski, C. R. Middaugh, R. B. Bankert, and S. V. Balu-Iyer, "Native-like aggregates of factor VIII are immunogenic in von Willebrand factor deficient and hemophilia a mice," *Journal of Pharmaceutical Sciences*, vol. 101, no. 6, pp. 2055–2065, 2012.

[27] M. M. C. van Beers, M. Sauerborn, F. Gilli, V. Brinks, H. Schellekens, and W. Jiskoot, "Aggregated recombinant human interferon beta induces antibodies but no memory in immune-tolerant transgenic mice," *Pharmaceutical Research*, vol. 27, no. 9, pp. 1812–1824, 2010.

[28] A. H. Fradkin, J. F. Carpenter, and T. W. Randolph, "Glass particles as an adjuvant: a model for adverse immunogenicity of therapeutic proteins," *Journal of Pharmaceutical Sciences*, vol. 100, no. 11, pp. 4953–4964, 2011.

[29] J. G. Barnard, K. Babcock, and J. F. Carpenter, "Characterization and quantitation of aggregates and particles in interferon-β products: potential links between product quality attributes and immunogenicity," *Journal of Pharmaceutical Sciences*, vol. 102, no. 3, pp. 915–928, 2013.

[30] V. Rombach-Riegraf, A. C. Karle, B. Wolf et al., "Aggregation of human recombinant monoclonal antibodies influences the capacity of dendritic cells to stimulate adaptive T-cell responses in vitro," *PLoS ONE*, vol. 9, no. 1, Article ID e86322, 2014.

[31] C. Lundegaard, O. Lund, C. Keşmir, S. Brunak, and M. Nielsen, "Modeling the adaptive immune system: predictions and simulations," *Bioinformatics*, vol. 23, no. 24, pp. 3265–3275, 2007.

[32] J. D. Gómez-Mantilla, I. F. Trocóniz, Z. Parra-Guillén, and M. J. Garrido, "Review on modeling anti-antibody responses to monoclonal antibodies," *Journal of Pharmacokinetics and Pharmacodynamics*, vol. 41, no. 5, pp. 523–536, 2014.

[33] T. P. Hickling, X. Chen, P. Vicini, and S. Nayak, "A review of quantitative modeling of B cell responses to antigenic challenge," *Journal of Pharmacokinetics and Pharmacodynamics*, vol. 41, no. 5, pp. 445–459, 2014.

[34] S. Palsson, T. P. Hickling, E. L. Bradshaw-Pierce et al., "The development of a fully-integrated immune response model (FIRM) simulator of the immune response through integration of multiple subset models," *BMC Systems Biology*, vol. 7, article 95, 2013.

[35] N. G. B. Agrawal and J. J. Linderman, "Mathematical modeling of helper T lymphocyte/antigen-presenting cell interactions: analysis of methods for modifying antigen processing and presentation," *Journal of Theoretical Biology*, vol. 182, no. 4, pp. 487–504, 1996.

[36] G. I. Bell, "Mathematical model of clonal selection and antibody production," *Journal of Theoretical Biology*, vol. 29, no. 2, pp. 191–232, 1970.

[37] K. B. Blyuss and L. B. Nicholson, "The role of tunable activation thresholds in the dynamics of autoimmunity," *Journal of Theoretical Biology*, vol. 308, pp. 45–55, 2012.

[38] K. B. Blyuss and L. B. Nicholson, "Understanding the roles of activation threshold and infections in the dynamics of autoimmune disease," *Journal of Theoretical Biology*, vol. 375, pp. 13–20, 2015.

[39] M. Oprea and A. S. Perelson, "Exploring the mechanisms of primary antibody responses to T cell-dependent antigens," *Journal of Theoretical Biology*, vol. 181, no. 3, pp. 215–236, 1996.

[40] A. Rundell, R. DeCarlo, H. HogenEsch, and P. Doerschuk, "The humoral immune response to *Haemophilus influenzae* type b: a mathematical model based on T-zone and germinal center B-cell dynamics," *Journal of Theoretical Biology*, vol. 194, no. 3, pp. 341–381, 1998.

[41] F. Castiglione, F. Toschi, M. Bernaschi et al., "Computational modeling of the immune response to tumor antigens," *Journal of Theoretical Biology*, vol. 237, no. 4, pp. 390–400, 2005.

[42] H. Y. Lee, D. J. Topham, S. Y. Park et al., "Simulation and prediction of the adaptive immune response to influenza A virus infection," *Journal of Virology*, vol. 83, no. 14, pp. 7151–7165, 2009.

[43] S. M. Ciupe, R. M. Ribeiro, and A. S. Perelson, "Antibody responses during hepatitis B viral infection," *PLoS Computational Biology*, vol. 10, no. 7, Article ID e1003730, 2014.

[44] B. Sulzer and A. S. Perelson, "Equilibrium binding of multivalent ligands to cells: effects of cell and receptor density," *Mathematical Biosciences*, vol. 135, no. 2, pp. 147–185, 1996.

[45] B. Sulzer and A. S. Perelson, "Immunons revisited: binding of multivalent antigens to b cells," *Molecular Immunology*, vol. 34, no. 1, pp. 63–74, 1997.

[46] X. Chen, T. Hickling, E. Kraynov, B. Kuang, C. Parng, and P. Vicini, "A mathematical model of the effect of immunogenicity on therapeutic protein pharmacokinetics," *The AAPS Journal*, vol. 15, no. 4, pp. 1141–1154, 2013.

[47] X. Chen, T. P. Hickling, and P. Vicini, "A mechanistic, multiscale mathematical model of immunogenicity for therapeutic proteins: part 2—model applications," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 3, no. 9, article e134, 10 pages, 2014.

[48] X. Chen, T. P. Hickling, and P. Vicini, "A mechanistic, multiscale mathematical model of immunogenicity for therapeutic proteins: part 1—theoretical model," *CPT: Pharmacometrics & Systems Pharmacology*, vol. 3, no. 9, pp. 1–9, 2014.

[49] S. E. Grossberg, J. Oger, L. D. Grossberg, A. Gehchan, and J. P. Klein, "Frequency and magnitude of interferon beta neutralizing antibodies in the evaluation of interferon beta immunogenicity in patients with multiple sclerosis," *Journal of Interferon & Cytokine Research*, vol. 31, no. 3, pp. 337–344, 2011.

[50] L. O. Narhi, J. Schmit, K. Bechtold-Peters, and D. Sharma, "Classification of protein aggregates," *Journal of Pharmaceutical Sciences*, vol. 101, no. 2, pp. 493–498, 2012.

[51] M. Reth, "Matching cellular dimensions with molecular sizes," *Nature Immunology*, vol. 14, no. 8, pp. 765–767, 2013.

[52] J. Foote and H. N. Eisen, "Kinetic and affinity limits on antibodies produced during immune responses," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 92, no. 5, pp. 1254–1256, 1995.

[53] J. Foote and C. Milstein, "Kinetic maturation of an immune response," *Nature*, vol. 352, no. 6335, pp. 530–532, 1991.

[54] S. Crotty, P. Felgner, H. Davies, J. Glidewell, L. Villarreal, and R. Ahmed, "Cutting edge: long-term B cell memory in humans after smallpox vaccination," *The Journal of Immunology*, vol. 171, no. 10, pp. 4969–4973, 2003.

[55] A. P. Kodituwakku, C. Jessup, H. Zola, and D. M. Roberton, "Isolation of antigen-specific B cells," *Immunology and Cell Biology*, vol. 81, no. 3, pp. 163–170, 2003.

[56] M. Ahmadi, C. J. Bryson, E. A. Cloake et al., "Small amounts of sub-visible aggregates enhance the immunogenic potential of monoclonal antibody therapeutics," *Pharmaceutical Research*, vol. 32, no. 4, pp. 1383–1394, 2015.

[57] C. Buelens, V. Verhasselt, D. De Groote, K. Thielemans, M. Goldman, and F. Willems, "Human dendritic cell responses to lipopolysaccharide and CD40 ligation are differentially regulated by interleukin-10," *European Journal of Immunology*, vol. 27, no. 8, pp. 1848–1852, 1997.

[58] F. Granucci, E. Ferrero, M. Foti, D. Aggujaro, K. Vettoretto, and P. Ricciardi-Castagnoli, "Early events in dendritic cell maturation induced by LPS," *Microbes and Infection*, vol. 1, no. 13, pp. 1079–1084, 1999.

[59] R. M. Cisco, Z. Abdel-Wahab, J. Dannull et al., "Induction of human dendritic cell maturation using transfection with RNA encoding a dominant positive toll-like receptor 4," *Journal of Immunology*, vol. 172, no. 11, pp. 7162–7168, 2004.

[60] K. Kawamura, K. Iyonaga, H. Ichiyasu, J. Nagano, M. Suga, and Y. Sasaki, "Differentiation, maturation, and survival of dendritic cells by osteopontin regulation," *Clinical and Diagnostic Laboratory Immunology*, vol. 12, no. 1, pp. 206–212, 2005.

[61] Y.-C. Wang, X.-B. Hu, F. He et al., "Lipopolysaccharide-induced maturation of bone marrow-derived dendritic cells is regulated by notch signaling through the up-regulation of CXCR4," *The Journal of Biological Chemistry*, vol. 284, no. 23, pp. 15993–16003, 2009.

[62] D. Verthelyi and V. Wang, "Trace levels of innate immune response modulating impurities (IIRMIs) synergize to break tolerance to therapeutic proteins," *PLoS ONE*, vol. 5, no. 12, Article ID e15252, 2010.

[63] M. Rossol, H. Heine, U. Meusch et al., "LPS-induced cytokine production in human monocytes and macrophages," *Critical Reviews in Immunology*, vol. 31, no. 5, pp. 379–446, 2011.

[64] J. C. Jones, E. W. Settles, C. R. Brandt, and S. Schultz-Cherry, "Virus aggregating peptide enhances the cell-mediated response to influenza virus vaccine," *Vaccine*, vol. 29, no. 44, pp. 7696–7703, 2011.

[65] L. Yin, J. M. Calvo-Calle, O. Dominguez-Amorocho, and L. J. Ster, "HLA-DM constrains epitope selection in the human CD4 T cell response to vaccinia virus by favoring the presentation of peptides with longer HLA-DM-mediated half-lives," *Journal of Immunology*, vol. 189, no. 8, pp. 3983–3994, 2012.

[66] J. Bessa, S. Boeckle, H. Beck et al., "The immunogenicity of antibody aggregates in a novel transgenic mouse model," *Pharmaceutical Research*, vol. 32, no. 7, pp. 2344–2359, 2015.

[67] W. Wang, "Protein aggregation and its inhibition in biopharmaceutics," *International Journal of Pharmaceutics*, vol. 289, no. 1-2, pp. 1–30, 2005.

[68] M. F. Bachmann and R. M. Zinkernagel, "Neutralizing antiviral B cell responses," *Annual Review of Immunology*, vol. 15, pp. 235–270, 1997.

[69] E. Szomolanyi-Tsuda and R. M. Welsh, "T-cell-independent antiviral antibody responses," *Current Opinion in Immunology*, vol. 10, no. 4, pp. 431–435, 1998.

[70] C. Babin, N. Majeau, and D. Leclerc, "Engineering of papaya mosaic virus (PapMV) nanoparticles with a CTL epitope derived from influenza NP," *Journal of Nanobiotechnology*, vol. 11, article 10, 2013.

[71] C. M. Snapper, T. M. McIntyre, R. Mandler et al., "Induction of IgG3 secretion by interferon γ: a model for T cell-independent class switching in response to T cell-independent type 2 antigens," *The Journal of Experimental Medicine*, vol. 175, no. 5, pp. 1367–1371, 1992.

[72] T. Fehr, M. F. Bachmann, E. Bucher et al., "Role of repetitive antigen patterns for induction of antibodies against antibodies," *The Journal of Experimental Medicine*, vol. 185, no. 10, pp. 1785–1792, 1997.

[73] P. K. A. Mongini, P. F. Highet, and J. K. Inman, "Human B cell activation: effect of T cell cytokines on the physicochemical binding requirements for achieving cell cycle progression via the membrane IgM signaling pathway," *Journal of Immunology*, vol. 155, no. 7, pp. 3385–3400, 1995.

[74] M. L. Wheeler, M. B. Dong, R. Brink, X.-P. Zhong, and A. L. DeFranco, "Diacylglycerol kinase zeta limits B cell antigen receptor-dependent activation of ERK Signaling to inhibit early antibody responses," *Science Signaling*, vol. 6, no. 297, article ra91, 2013.

[75] G. G. B. Klaus, M. Holman, C. Johnson-Léger, J. R. Christenson, and M. R. Kehry, "Interaction of B cells with activated T cells reduces the threshold for CD40-mediated B cell activation," *International Immunology*, vol. 11, no. 1, pp. 71–79, 1999.

[76] L. P. Kil, M. J. W. de Bruijn, M. van Nimwegen et al., "Btk levels set the threshold for B-cell activation and negative selection of autoreactive B cells in mice," *Blood*, vol. 119, no. 16, pp. 3744–3756, 2012.

[77] D. W. Scott and A. S. De Groot, "Can we prevent immunogenicity of human protein drugs?" *Annals of the Rheumatic Diseases*, vol. 69, supplement 1, pp. i72–i76, 2010.

[78] V. Brinks, W. Jiskoot, and H. Schellekens, "Immunogenicity of therapeutic proteins: the use of animal models," *Pharmaceutical Research*, vol. 28, no. 10, pp. 2379–2385, 2011.

[79] L. Xue, M. Fiscella, M. Rajadhyaksha et al., "Pre-existing biotherapeutic-reactive antibodies: survey results within the american association of pharmaceutical scientists," *The AAPS Journal*, vol. 15, no. 3, pp. 852–855, 2013.

[80] L. Xue and B. Rup, "Evaluation of pre-existing antibody presence as a risk factor for posttreatment anti-drug antibody induction: analysis of human clinical study data for multiple biotherapeutics," *The AAPS Journal*, vol. 15, no. 3, pp. 893–896, 2013.

[81] V. S. Sloan, P. Cameron, G. Porter et al., "Mediation by HLA-DM of dissociation of peptides from HLA-DR," *Nature*, vol. 375, no. 6534, pp. 802–806, 1995.

[82] N. K. Nanda and A. J. Sant, "DM determines the cryptic and immunodominant fate of T cell epitopes," *The Journal of Experimental Medicine*, vol. 192, no. 6, pp. 781–788, 2000.

[83] A. J. Sant, F. A. Chaves, S. A. Jenks et al., "The relationship between immunodominance, DM editing, and the kinetic stability of MHC class II:peptide complexes," *Immunological Reviews*, vol. 207, pp. 261–278, 2005.

[84] L. Yin, P. Trenh, A. Guce et al., "Susceptibility to HLA-DM protein is determined by a dynamic conformation of major histocompatibility complex class II molecule bound with peptide," *The Journal of Biological Chemistry*, vol. 289, no. 34, pp. 23449–23464, 2014.

[85] K. R. Garrod, H. D. Moreau, Z. Garcia et al., "Dissecting T cell contraction in vivo using a genetically encoded reporter of apoptosis," *Cell Reports*, vol. 2, no. 5, pp. 1438–1447, 2012.

[86] H. Lee, S. Haque, J. Nieto et al., "A p53 axis regulates B cell receptor-triggered, innate immune system-driven B cell clonal expansion," *Journal of Immunology*, vol. 188, no. 12, pp. 6093–6108, 2012.

[87] C. L. M. Krieckaert, G. M. Bartelds, W. F. Lems, and G. J. Wolbink, "The effect of immunomodulators on the immunogenicity of TNF-blocking therapeutic monoclonal antibodies: a review," *Arthritis Research & Therapy*, vol. 12, no. 5, article 217, 2010.

[88] A. Radbruch, G. Muehlinghaus, E. O. Luger et al., "Competence and competition: the challenge of becoming a long-lived plasma cell," *Nature Reviews Immunology*, vol. 6, no. 10, pp. 741–750, 2006.

[89] C. H. Rozanski, R. Arens, L. M. Carlson et al., "Sustained antibody responses depend on CD28 function in bone marrow-resident plasma cells," *The Journal of Experimental Medicine*, vol. 208, no. 7, pp. 1435–1446, 2011.

Hindawi Publishing Corporation
Journal of Immunology Research
Volume 2015, Article ID 737168, 10 pages
http://dx.doi.org/10.1155/2015/737168

Review Article

Computational Modelling Approaches on Epigenetic Factors in Neurodegenerative and Autoimmune Diseases and Their Mechanistic Analysis

Afroza Khanam Irin,[1,2] Alpha Tom Kodamullil,[1,2] Michaela Gündel,[1,2] and Martin Hofmann-Apitius[1,2]

[1]*Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing, 53754 Sankt Augustin, Germany*
[2]*Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Dahlmannstrasse 2, 53113 Bonn, Germany*

Correspondence should be addressed to Martin Hofmann-Apitius; martin.hofmann-apitius@scai.fraunhofer.de

Received 31 July 2015; Revised 19 October 2015; Accepted 20 October 2015

Academic Editor: Francesco Pappalardo

Copyright © 2015 Afroza Khanam Irin et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Neurodegenerative as well as autoimmune diseases have unclear aetiologies, but an increasing number of evidences report for a combination of genetic and epigenetic alterations that predispose for the development of disease. This review examines the major milestones in epigenetics research in the context of diseases and various computational approaches developed in the last decades to unravel new epigenetic modifications. However, there are limited studies that systematically link genetic and epigenetic alterations of DNA to the aetiology of diseases. In this work, we demonstrate how disease-related epigenetic knowledge can be systematically captured and integrated with heterogeneous information into a functional context using Biological Expression Language (BEL). This novel methodology, based on BEL, enables us to integrate epigenetic modifications such as DNA methylation or acetylation of histones into a specific disease network. As an example, we depict the integration of epigenetic and genetic factors in a functional context specific to Parkinson's disease (PD) and Multiple Sclerosis (MS).

1. Introduction

In the 19th century, Gregor Mendel defined the mechanism of inheritance patterns, which laid the ground for genetics in modern biology. However, Mendel's theories could explain neither how different individuals in a population are genetically similar but exhibit different phenotypes, nor how identical twins are prone to different diseases. Recent studies confirmed that copy number variations, single nucleotide polymorphism, or any heritable changes in the DNA sequence could be a plausible additional explanation for Mendel's observation. In 1942, Waddington used the term *epigenotype* as a name for the study of causal mechanisms through which genes exhibit phenotypic effects and their adaptive interaction with the environment [1]. These epigenetic causal mechanisms involve histone modifications, DNA methylation, and abnormal RNA regulation, which can alter normal biological processes by heritable silencing of genes, although they do not cause any nucleotide sequence changes in chromosomal components [2]. Gill published the first paper describing epigenetic mechanism in drosophila egg promorphology [3]. In 1971, Tsanev and Sendov proposed the role of epigenetics in neoplastic transformation and the process of carcinogenesis [4]. Holliday reviewed the methylation of cytosine in DNA and how they are consistent to the levels of gene expression in higher organisms like human, mouse, and hamster [5]. He also illustrated that epigenetic effects are closely linked to aging such that decrease in methylation correlates with lifespan. It has later been demonstrated that epigenetic modifications are tissue-specific phenomena that can have dramatic effects on the silencing, the increase, or the reduction of the expression of genes in a given tissue. Song et al. observed variations of the methylation status in different developmental stages [6]. Additionally, Chen and

Zhang showed the risk of neonatal mortality due to maternal vascular underperfusion, which is a result of epigenetic modifications in several genes during pregnancy [7].

Several studies illustrate how nutrition and environmental factors influence epigenetic modifications. A study based on an African-American cohort demonstrated that epigenetic factors like psychological stress and social context are related to inflammation in coronary heart disease and stroke [8]. In the progression of type-2 diabetes mellitus (T2DM), Prattichizzo et al. [9] reviewed interactions between epigenetic (DNA methylation, posttranslational histone modifications, and miRNA regulation) and environmental factors (lifestyle and mainly dietary habits). Duru et al. proposed several dietary chemoprevention agents—such as Retinoids/Vitamin A, Resveratrol, EGCG/Green Tea, and Vitamin D—which act on miRNA-signalling pathways to be novel therapeutics in breast cancer [10].

It is noteworthy that environmental exposures during early stage of life can also induce persistent alterations in the epigenome, which may lead to an increased risk of disease later in life. Reviews by Van Dijk et al. and Cordero et al. investigated different epigenomics patterns in obesity during early and later stage of life [11, 12]. They elucidated the role of dietary supplements and environmental conditions on epigenetic mechanisms during the pregnancy period, which lead to the risk of obesity in offspring.

2. Epigenetics in Neurodegenerative and Autoimmune Disease

With the rising momentum of biomedical science, several studies on neurodegenerative diseases (NDDs) not only showed environmental influences on molecular and cellular changes [13, 14] but also established possible relationships between genes and the environment [15]. The major mechanisms for epigenetic alterations found in these diseases include DNA methylation, histone tail modifications, chromatin remodelling, and mechanisms regulated by small RNA molecules [16–18]. Epigenetics in neurodegenerative and autoimmune diseases are of current interest to many researchers and more recently several studies have shed light on the role of epigenetic alterations in autoimmune diseases and NDDs.

Ravaglia et al. discussed the association of folate and Vitamin B12 levels in nutritional diet with the prevalence of NDD [19]. An experiment performed on aged monkeys showed epigenetic changes in APP expression and amyloid beta level due to lead (Pb) exposure [20]. Another study by Baccarelli and Bollati explained how air-pollutants (black carbon, benzene) and toxic chemicals (arsenic, nickel, and diethylstilbestrol) alter gene expression accompanied by epigenetics changes [21]. This paper reviewed all possible metals and chemicals; those are responsible for up- or downregulation of disease specific gene such as BDNF.

Since NDDs are prevalent in the aged population, experiments conducted on NDD patients have revealed how environmental factors such as age, lifestyle, diet, and level of education influence the development of diseases and also highlighted the crosstalk of environmental factors with genes [22]. HDAC gene expression has been shown to be downregulated by Kaliman et al. due to moderate physical activities, which in turn reduce the expression of proinflammatory genes in NDDs [23]. Other than physical exercise, Nicolia et al. reviewed the role of environmental factors such as stressors (physical and behavioral), pesticides, and mental exercise causing DNA methylation in age-related diseases, specifically in AD [24]. The authors suggested that longer lifespan increases the risk of environment-induced epigenetic changes. In a detailed study [25] of epigenetics in AD, decreased DNA methylation was observed in the temporal neocortex of monozygotic AD twins. Manipulation of histone tail acetylation with HDAC inhibitors also has been investigated in several animal models of AD [26]. Martí et al. have explained a set of deregulated miRNAs that participate in altered gene expression in neurodegeneration, especially in Huntington's disease [27].

A hypothesis, namely, "hapten hypothesis," was introduced by Mintzer et al. in 2009, which describes that drugs like Penicillin and Clozapine play the role as haptens to produce antibodies against neutrophils in case of autoimmune diseases, such as Systemic Lupus Erythematosus (SLE) [28]. Uhlig et al. mentioned smoking as risk factor in addition to age and gender in another systemic autoimmune disorder, that is, Rheumatoid Arthritis (RA) [29]. Similarly, ultraviolet radiation also alters the immune mechanisms that may result in Lupus Erythematosus (LE) [30]. From the above discussion it is evident that epigenetic factors play a significant role in the context of NDD and autoimmune disease.

Although there is growing interest in epigenetics of NDDs and autoimmune diseases, only a few studies have been performed specifically on PD and MS. In fact, only a very limited number of studies deal with the functional consequences of epigenetic modifications and perturbed mechanisms leading to a particular phenotype. A systematic comparison of the number of epigenetic studies in AD, PD, and MS in the last years is shown in Figure 1(a). The graph shows that the number of scientific publications on epigenetics in PD and MS is significantly lower than the number of papers on epigenetics in AD. Figure 1(b) represents the overall trend in epigenetic studies; it becomes obvious that AD, PD, and MS represent only a minority fraction of the literature on epigenetics mechanisms, in particular when compared with the predominant indication areas arthritis, cancer, and diabetes.

3. Computational Modelling of Epigenetic Factors in a Functional Context

To represent, manipulate, and visualize large amounts of biological data from different sources, computational modelling has become an intuitive approach. Artyomov et al. proposed an "epigenetic and genetic regulatory network" that describes how transcription factors affect cellular differentiation by reprogramming embryonic cells [31]. Irrespective of any specific disease context, a computational micromodel for epigenetic mechanisms was developed by Raghavan et al.,

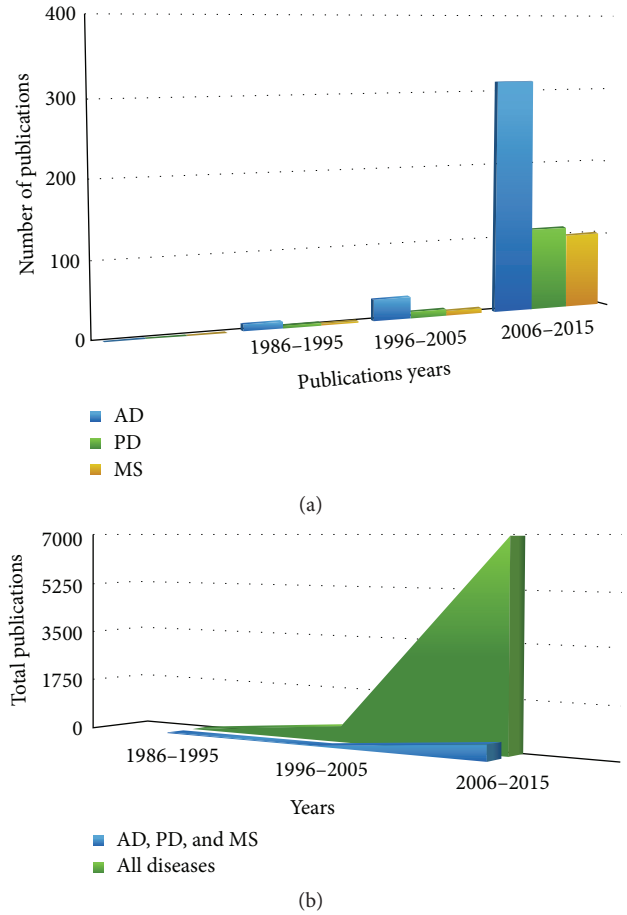


Figure 1: (a) Statistics over scientific publications around epigenetics related neurodegenerative (AD and PD), autoimmune diseases and other diseases using PubMed with queries (("Parkinson's disease") AND epigenetics), (("Alzheimer's disease") AND epigenetics), and (("Multiple Sclerosis") AND epigenetics), last accessed on 7/20/2015. In (a), blue, green, and orange coloured bars represent the total number of publications, for AD, PD, and MS, respectively. (b) This figure illustrates the trend of research on other diseases around epigenetics compared to NDD (AD and PD) and autoimmune (MS) disease, where green coloured portion representing the studies on all sorts of diseases and blue portion covers only AD, PD, and MS related researches.

demonstrating the interaction of histone modifications with DNA methylation and transcription process [32]. The model was able to identify the transcription rate when the level of DNA methylation is known.

From high throughput gene expression data of 12 human cell lines, a model integrating transcriptomic data and histone modification has been developed, called Epigenetic Regulatory Network [33], which identifies the main contributing epigenetic factors among different cell types. To facilitate the systematic integration of High Throughput Sequencing (HTS) epigenetic data, Althammer et al. have described a new computational framework. This workflow was inspired by machine learning algorithms and can be used to find alterations of epigenetic states between two given cell types [34]. Artificial Epigenetic Regulatory Network (AERN) proposed by Turner et al. has included DNA methylation and chromatin modification as the epigenetic elements in addition to genetic factors. They showed an example of how disease specific genes can be allocated in the network according to environmental changes and how gene expression regulation can be analysed within the network [35]. In a recent review paper [36], Hidden Markov Models (HMM) have been used to handle the complexity of epigenetic mechanisms, especially different patterns of DNA methylation. For autoimmune diseases, Farh et al. developed an algorithm, named "Probabilistic Identification of Causal SNPs (PICS)," which was able to find out the possibility of SNPs to be causal variants in immune cell enhancers when epigenetic modifications on that chromatin site are known [37].

Although there are algorithms that identify epigenetic modifications, there are no previous evidences describing the interpretation of functional consequences of epigenetic modifications in disease mechanisms. Here, we propose a computer-readable modelling strategy that is competent of fusing knowledge and data based information, which is capable of explaining the functional consequences of epigenetic modification in a mechanistic fashion. In this paper, we introduce the Biological Expression Language (BEL; http://www.openbel.org/) that is the main base of building models for epigenetics analysis of PD and MS.

BEL integrates literature-derived "cause and effect" relationships into network models, which can be subjected to causal analysis and used for mechanism-based hypothesis generation [38]. The semantic triple-based modelling language used here enables the application of Reverse Causal Reasoning (RCR) algorithms, which support the identification of mechanistic hypotheses from the corresponding causal network. The RCR methodology allows for investigating to what extent a knowledge-based set of triples is supported by omics data (e.g., gene expression data); the method is therefore suited for inference based on qualitatively significant data [39]. To enable a quantitative assessment and to perform comparative mechanistic analysis, another algorithm is integrated in the BEL framework: the Network Perturbation Amplitude (NPA) method. Although it uses the same network structure like RCR, its main purpose is to estimate the activity changes of a specific biological process when a pathophysiology state is compared to a nonperturbed condition [40].

Until now, BEL based network modelling approaches have been used in various applications such as early patient stratification, biomarker identification [41], and personalized drug discovery [42] in the context of cancer research by different groups. Our objective behind this computational modelling approach aims at harvesting relevant scientific knowledge from unstructured text and to systematically understand the functional impact of epigenetic modification in the context of PD and MS using BEL.

4. Role of Epigenetics in Parkinson's Disease Using BEL Models

PD is characterized by a loss of midbrain dopaminergic neurons leading to motor abnormalities and autonomic

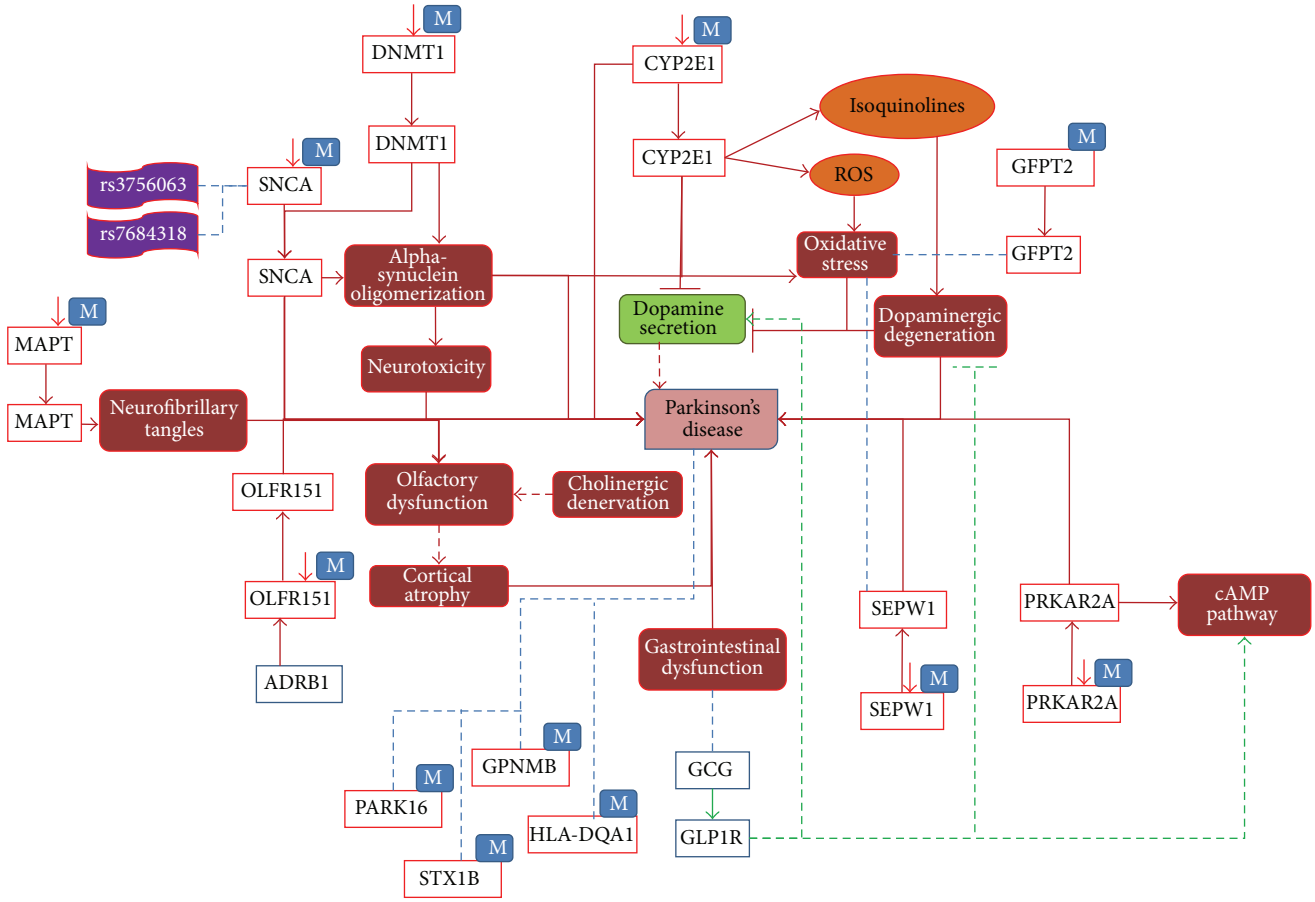


Figure 2: The role of epigenetics modification; hypomethylation around certain genes in PD. In this figure, red lines indicate the disease state interactions and green lines show normal state. Blue lines show the association between entities with unknown direction. Dotted lines are the interpretation, which needs to be further analysed. "M" associated with a gene entity denotes a methylation process and down-arrows besides represent decreased methylation.

dysfunctions [43]. Genes such as *SNCA*, *parkin*, *PINK1*, and *FBX07* have been identified to be responsible for pathophysiological mechanisms like mitochondrial damage, repair, and oxidative stress [17]. There are evidences suggesting that the above-mentioned key genes are epigenetically modified under disease conditions. For example, studies in familial as well as sporadic PD patients suggested that demethylation of the *SNCA* gene stimulates its upregulation [17, 44, 45]. Increasing amounts of *CYP2E1* have been found to promote the formation of toxic metabolites, which further degenerate the dopaminergic neurons [46]. Abnormal epigenetic modifications involved in the pathogenesis of PD have been studied by Feng et al.; in that study, detailed insights on DNA methylation and histone acetylation mechanisms and their association with the disease are reported [47].

To construct an epigenetics model for PD, we have made use of SCAIView (http://bishop.scai.fraunhofer.de/scaiview/), a literature mining environment to extract all relevant articles using the query ([*MeSH Disease*: "*Parkinson Disease*"]) *AND* ([*Parkinson Ontology*: "*Epigenetics*"]). Based on this literature mining approach, we have manually selected 78 articles, which were found to contain relevant information about PD epigenetics. The content of these publications was subsequently encoded in BEL. The model consists of 235 nodes and 407 edges representing 339 BEL statements. The nodes contain 67 proteins/genes, 21 biological processes, 6 SNPs, 3 complexes, 24 chemical entities, 26 miRNAs, and 88 other nodes representing translocation, degradation, and association functions.

As shown in Figure 2, seven representative genes, namely, *SNCA*, *MAPT*, *DNMT1*, *CYP2E1*, *OLFR151*, *PRKAR2A*, and *SEPW1*, were reported to be hypomethylated under disease conditions. In these cases, hypomethylation causes overexpression of genes that perturb normal biological processes. Increased expression of *SNCA* and *DNMT1* caused by decreased methylation of these genes results in alpha-synuclein oligomerization, which in turn causes neurotoxicity in PD [48]. Along with that, two SNPs, rs3756063 and rs7684318, were associated with hypomethylation of SNCA in PD patients. Similarly, the *CYP2E1* gene was detected to be upregulated due to (i) hypomethylation, (ii) release of isoquinolines, and (iii) Reactive Oxygen Species (ROS), which lead to dopaminergic degeneration and oxidative stress, respectively [49]. Increased neurofibrillary tangles in

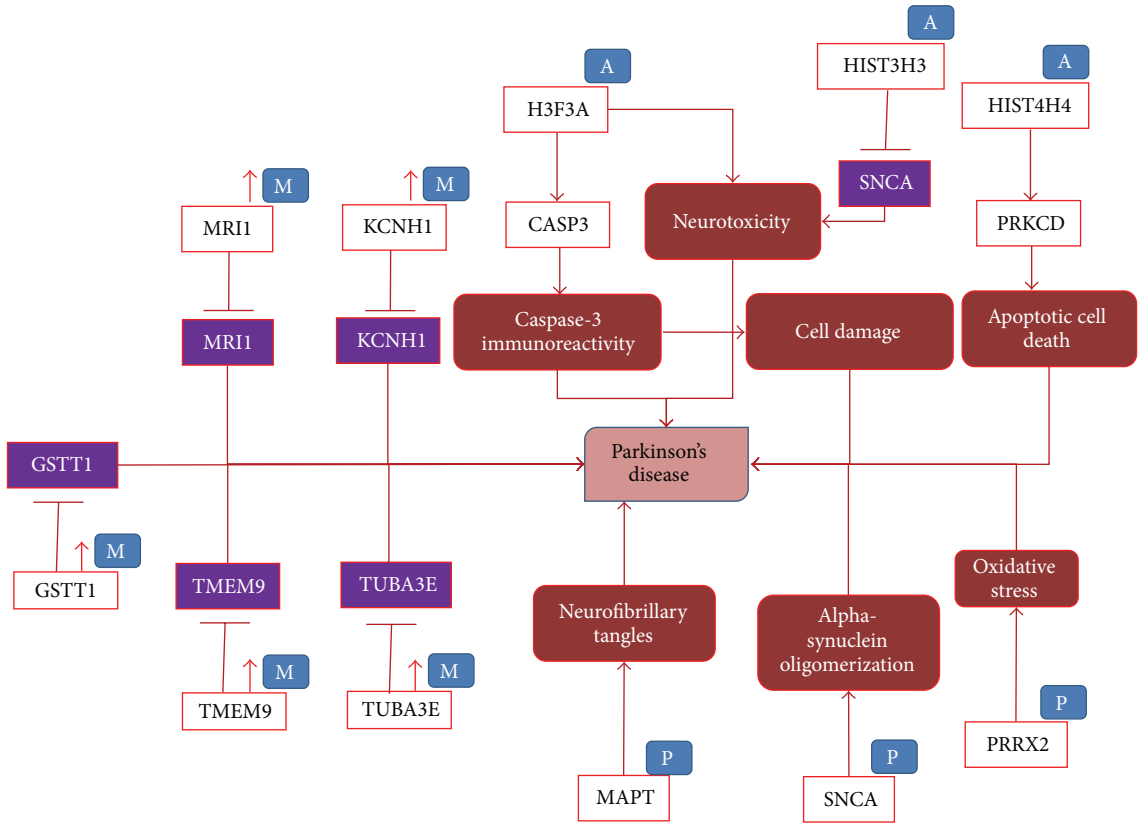


FIGURE 3: The role of epigenetics modification, hypermethylation, phosphorylation, and acetylation around certain genes in PD. In this figure also red lines indicate the disease state interactions. "M" associated with a gene entity denotes a methylation process and up-arrows besides represent an increase of methylation. "P" and "A" represent the phosphorylation and acetylation processes, respectively. Genes in purple boxes denote lower expression of genes.

PD have been reported to be linked with high expression of *MAPT* gene, as a consequence of reduced methylation [50]. Furthermore, *ADRB1* induced the hypomethylation of the *OLFR151* gene [51]. As a result, overexpression of *OLFR151* leads to olfactory dysfunction and cortical atrophy, which are early symptoms of PD [52].

GWAS and epigenomic studies suggest that *SEPW1* and *PRKAR2A* were overexpressed due to hypomethylation in PD patients [53]. However, there is lack of well-established knowledge about the functional role of *SEPW1* and *PRKAR2A* in the context of PD. We identified only one study that reports the association of *SEPW1* with PD brains [53]. Similarly, we did not find any direct biological consequences of *PRKAR2A* to play a role in the disease state. We employed a dedicated data mining approach in our model and identified the association of *PRKAR2A* with the cAMP pathway. It has been found that cAMP signal transduction pathway is stimulated by *GCG* (glucagon) [54] and its receptor *GLP1R*, which is secreted by the gastrointestinal mucosa [55]. *GLP1R* is also known to play a role in dopamine secretion and inhibiting dopaminergic degeneration [56]. Therefore we speculate that gastrointestinal dysfunction (an early symptom of PD) may result in a perturbation of the cAMP pathway and that this could be a possible mechanistic link to hypomethylation of *PRKAR2A* in PD. In addition to the above-mentioned hypomethylated genes, five more methylated genes were identified in the PD context, namely, *GFPT2*, *GPNMB*, *PARK16*, *STX1B*, and *HLA-DQA1*, where only *GFPT2* was inferred to be associated with oxidative stress [57]. These examples demonstrate that even though the analysis of high throughput data like GWAS or epigenetic studies do predict many disease-associated risk genes, no further research has been carried out to understand the functional impact of these genes.

In addition to Figure 2, we represent in our modelling approach three more highly relevant epigenetics modifications, namely, hypermethylation, phosphorylation, and acetylation (Figure 3). Five genes, *GSTT1*, *MRI1*, *KCNH1*, *TMEM9*, and *TUBA3E*, were reported to be significantly hypermethylated resulting in low expression of genes [58]. However, there were no studies describing the functional role of these genes in the PD context. In case of acetylation modification, *H3F3A*, *HIST3H3*, and *HIST4H4* were shown to be acetylated under disease conditions. Acetylated *H3F3A* increases *CASP3* activity and thereby may cause cell damage [59]. Acetylation in *HIST3H3* decreases the expression of *SNCA* leading to neurotoxicity [60], whereas *HIST4H4* acetylation induces the activity of *PRKCD*, which promotes

apoptotic cell death [59]. Phosphorylation of *MAPT*, *SNCA*, and *PRRX2* causes deposition of neurofibrillary tangles, alpha-synuclein oligomerization, and oxidative stress, respectively, in PD [50, 61].

The enlisted microRNAs in Table 1 were suggested to regulate the epigenetic modification in disease state of Parkinson. These microRNAs bind to their target and downregulate or upregulate their expression in diseased condition. For instance, *MIR34C* induces the expression of the *PARK7* gene, which in turn causes oxidative stress in PD. Some microRNAs function together (i.e., *MIR34B* and *MIR34C*) while others target individually specific genes such as *PARK7*, *PARK2*, and *TP53* to cause dysregulation in target genes, which may contribute to the disease aetiology [62].

Table 1: Role of microRNAs in PD epigenetics. 26 microRNAs have been identified that have been reported to control PD pathways. Positive and negative correlations of these microRNAs with PD mean if they are inducing or inhibiting the disease state, respectively. Also, we have enlisted the target genes for retrieved microRNAs.

Role of microRNAs in PD epigenetics		
MicroRNA	Relation to PD	Target
MIR133B	Negative correlation	*PITX3*
MIR1	Negative correlation	*TPPP, BDNF*
MIR29A	Negative correlation	—
MIR221	Negative correlation	—
MIR222	Negative correlation	—
MIR223	Negative correlation	—
MIR224	Negative correlation	—
MIR30A	Positive correlation	*SLC6A3, FGF20, GRIN1, GRIA1*
MIR16-2	Positive correlation	*FGF20*
Mir26a-2	Associated	*Gria1, Tyr*
MIR886	Positive correlation	—
MIR133B	Negative correlation	—
MIR433	Negative correlation	*FGF20*
MIR7-1	Negative correlation	—
MIR7-2	Negative correlation	—
MIR-7	Positive correlation	*SNCA*
MIR34B	Positive correlation	*PARK7, PARK2, TP53*
MIR34C	Positive correlation	*PARK7, PARK2, TP53*
MIR219A1	Negative correlation	*GRIN1, CD164*
MIR219A2	Negative correlation	*GRIN1, CD164*
MIR124-1	Positive correlation	*PPP1R13L*
Mir219a-1	Negative correlation	*Grin1*
Mir219a-2	Negative correlation	*Grin1*
Mir124a-1	Negative correlation	—
Mir124a-2	Negative correlation	—
Mir124a-3	Negative correlation	—

5. Role of Epigenetics in Multiple Sclerosis Using BEL Models

Multiple Sclerosis, a complex autoimmune disease of the central nervous system, is characterized by inflammation, demyelination, and destruction of the axons in the central nervous system [63]. Although the aetiology is not known, there is accumulating evidence that, in a cohort with genetic predisposition, environmental factors may play a key role in the development of the disease [64]. Epigenetic studies of this autoimmune disease have shown that disorders of epigenetic processes may influence chromosomal stability and gene expression, resulting in complicated syndromes [65, 66]. In a more detailed study, increased immunoreactivity for acetylated histone H3 in oligodendrocytes was found in a subset of MS samples [67]. Various microRNAs have been shown to differentially express in MS samples; particularly *MIR223* was found to be upregulated in MS patients compared to healthy controls [68]. Major epigenetic mechanisms involved in MS have been listed in a current review article [69], for example, DNA methylation, histone citrullination, and histone acetylation.

Similar to the approach taken with the PD model, we have started with a systematic literature analysis using SCAIView. We extracted information from all articles that could be retrieved with the query ([*MeSH Disease*: "*Multiple Sclerosis*"]) *AND* ([*Multiple Sclerosis Ontology*: "*Epigenetics*"]). An overall number of 75 highly relevant articles were used to build the BEL model for MS epigenetics. From this corpus of relevant literature, we have extracted 339 BEL statements to develop a network comprising 215 nodes and 536 edges. The nodes consist of 69 proteins/genes, 43 biological processes, 8 complexes, 18 chemical entities, 38 miRNAs, 8 protein families, and 31 other entities representing translocation, degradation, and association functions.

Most frequent epigenetic factors affecting MS were found to be miRNA regulation, histone citrullination, and lifestyle factors. We found 24 miRNAs that positively regulate the pathogenesis of MS and *miR23B*, *miR487B*, *miR184,* and *miR656* seem to be less expressed in the diseased context [70]. Apart from these, many epigenetics modifications like acetylation and citrullination were found in cytokines (*IFNG*, *TNF*) [71], chemokines (*CCR5*, *CCL5*, *CXCR3*, *CXCL10*, *CXCL8*, and *CXCR6*) [72], neurotrophic factors (*BDNF*, *NTF3*) [73], surface antigens (*CD8A*, *CD8B*) [74], and other genes like *GFAP*, *MBP*, *SNORD24*, and *NOTCH4*. In addition, dietary factors such as Vitamin D, intake of fruit juice, fruit/vegetables, cereal, bread, grains, and fish products reduce the risk of MS whereas intake of high energy and animal food such as fat, pork, hot dogs, and sweets increase risk of the disease (Figure 4).

6. Discussion

Epigenetics is a major mechanism that accommodates gene-expression changes in response to gene-environment interactions. In the last few decades, it has been shown that epigenetic factors play an important role in neurodegenerative as well as in autoimmune diseases. Even though there are strategies to identify new epigenetic modifications, there are very few studies, which link these alterations in DNA to the aetiology of the disease. Given the complexity and the wide variety of entities like epigenetic modifications and genetic variants, which perturb normal biological processes,

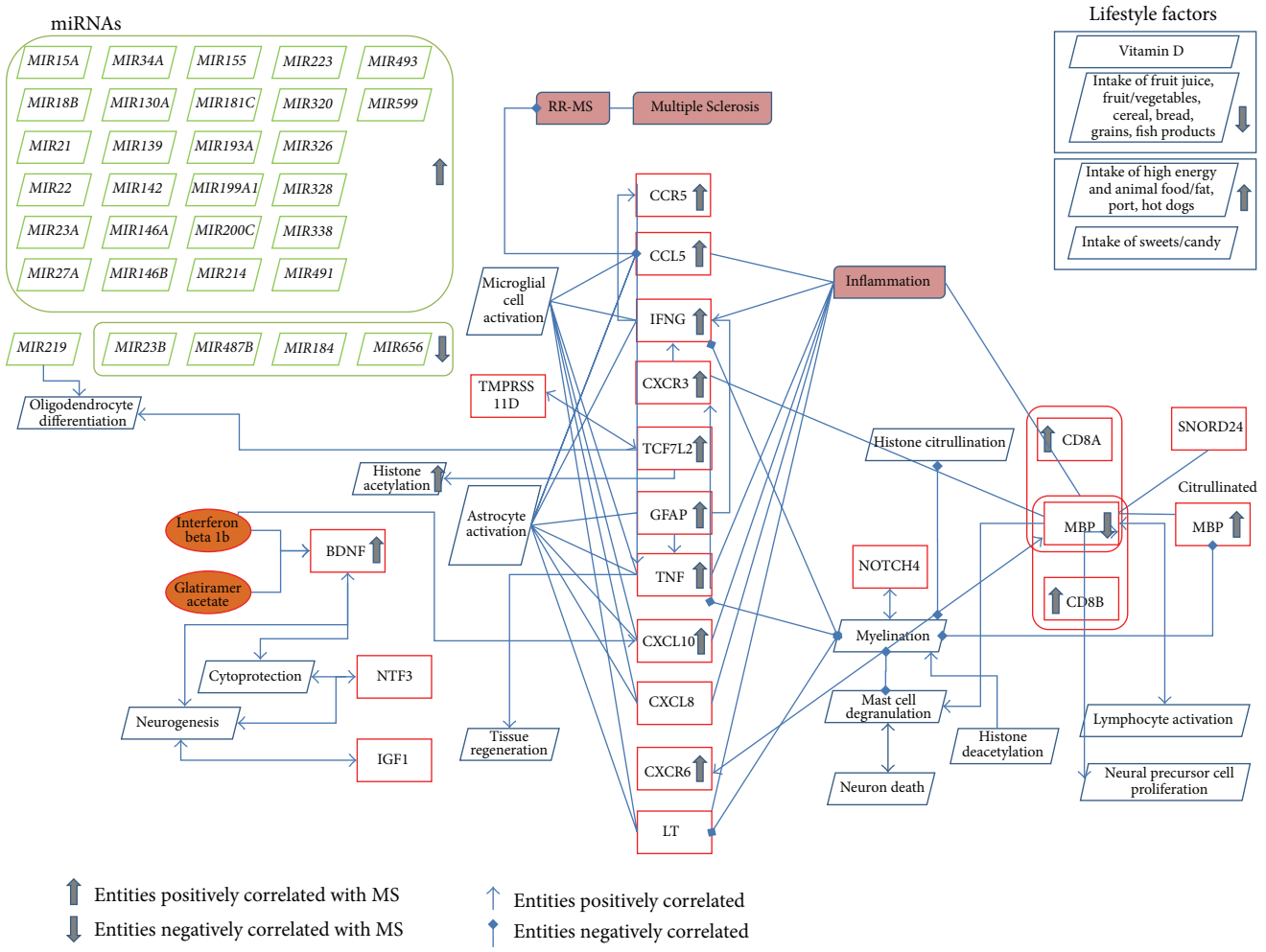


FIGURE 4: Model around epigenetic factors in Multiple Sclerosis. The figure shows the different epigenetic factors which regulate MS pathology including miRNAs represented in green colour boxes and genes in red colour boxes, chemicals in orange colour, and bioprocess in blue rhombus shape.

we need new strategies to integrate data driven and knowledge driven approaches to unravel the mechanisms behind these complex diseases. We demonstrated that it is possible to collectively capture disease-related, epigenetic knowledge and integrate it into a functional context using the modelling language BEL. An adaptation of the BEL syntax enables us to integrate epigenetic modification information like methylation (hypo and hyper), acetylation, phosphorylation, and miRNAs regulation into a specific disease network. In addition to these mechanisms, we have also included the role of many environmental factors such as food habit and obesity to the model which are responsible for the epigenetic modifications.

Although fewer studies related to PD and MS around epigenetics have been published until now, we tried to integrate all available knowledge from the scientific literature. In the case of PD, the main genes which are epigenetically regulated through methylation are *SNCA*, *PARK6*, *CYP2E1*, *PINK1*, *BDNF*, *FGF*, *MAPT*, *MTHFR*, *OLFR 151*, *PARK16*, *PARK2*, *PARK7*, *TPPP*, *PDE4D*, and *METRNL*. Also we have found acetylation in *H3F3A*, *HIST3H3*, and *HIST4H4* genes and phosphorylation in *MAPT*, *SNCA*, and *PRRX2* genes as major epigenetic modifications in PD along with miRNA regulation. Similarly for MS, we have found several citrullinated or acetylated cytokines, chemokines, transcription factors, neurotrophins, and many dietary factors, which can influence disease processes.

Some of the genes identified are well studied, but for others still an in-depth analysis is needed. Since there are no studies published on these novel candidates derived from data driven approaches, we were not able to link the functional impact of epigenetic modifications to the disease aetiology. For instance, there are about 30 GWAS studies associating the *PARK16* gene with PD, but no detailed information about the functional context of *PARK16* in the pathophysiology of PD exists in the literature. We observe a clear bias towards well-known candidate genes like *SNCA* for PD and *MBP* for MS; in order to overcome this bias, dedicated effort towards investigating the role of the new candidate genes and related bioprocesses is required.

Although BEL has the capability to integrate different biological entities and modifications at the levels of proteins, the current version of BEL is not efficient in representing epigenetic modifications at gene level, so that it is not yet possible to reason over epigenetic effects automatically (e.g., using RCR). It is obvious that we need to extend the syntax of the modelling language in order to formally represent this type of variation and develop algorithms that assess the functional impact based on biological network models.

Abbreviations

BEL: Biological Expression Language
PD: Parkinson's disease
MS: Multiple Sclerosis
AD: Alzheimer's disease
NDDs: Neurodegenerative diseases
T2DM: Type-2 diabetes mellitus
DNA: Deoxyribonucleic acid
RNA: Ribonucleic acid
HDAC: Histone deacetylase
APP: Amyloid Precursor Protein
SLE: Systemic Lupus Erythematosus
RA: Rheumatoid Arthritis
HTS: High throughput sequencing
AGRN: Artificial Gene Regulatory Network
AERN: Artificial Epigenetic Regulatory Network
HMM: Hidden Markov Model
SNP: Single nucleotide polymorphism
ROS: Reactive Oxygen Species
GWAS: Genome-Wide Association Study
cAMP: Cyclic adenosine $3'$,$5'$-monophosphate
miRNA: MicroRNA.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors wish to thank Professor Dr. Ullrich Wüllner from Universitätsklinikum Bonn for fruitful discussions and active support for this work. Furthermore, the authors acknowledge the financial support from the B-IT foundation that sponsors part of the academic work in their department. Finally, the authors would like to acknowledge the strong motivation that came from their involvement in the Neuroallianz project, a project funded by the German Ministry of Research and Science (BMBF). Part of the research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under Grant Agreement no. 115568 (project AETIONOMY), resources of which are composed of financial contribution from the European Union's Seventh Framework Program (FP7/2017-2013) and European Federation of Pharmaceutical Industries and Associations companies in kind contribution.

References

[1] C. H. Waddington, "The epigenotype. 1942," *International Journal of Epidemiology*, vol. 41, no. 1, pp. 10–13, 2012.

[2] G. Egger, G. Liang, A. Aparicio, and P. A. Jones, "Epigenetics in human disease and prospects for epigenetic therapy," *Nature*, vol. 429, no. 6990, pp. 457–463, 2004.

[3] K. S. Gill, "Epigenetics of the promorphology of the egg in *Drosophila Melanogaster*," *Journal of Experimental Zoology*, vol. 155, no. 1, pp. 91–104, 1964.

[4] R. Tsanev and B. Sendov, "An epigenetic mechanism for carcinogenesis," *Zeitschrift für Krebsforschung und Klinische Onkologie*, vol. 76, no. 4, pp. 299–319, 1971.

[5] R. Holliday, "The inheritance of epigenetic defects," *Science*, vol. 238, no. 4824, pp. 163–170, 1987.

[6] F. Song, S. Mahmood, S. Ghosh et al., "Tissue specific differentially methylated regions (TDMR): changes in DNA methylation during development," *Genomics*, vol. 93, no. 2, pp. 130–139, 2009.

[7] M. Chen and L. Zhang, "Epigenetic mechanisms in developmental programming of adult disease," *Drug Discovery Today*, vol. 16, no. 23-24, pp. 1007–1018, 2011.

[8] K. L. Saban, H. L. Mathews, H. A. de Von, and L. W. Janusek, "Epigenetics and social context: implications for disparity in cardiovascular disease," *Aging and Disease*, vol. 5, no. 5, pp. 346–355, 2014.

[9] F. Prattichizzo, A. Giuliani, A. Ceka et al., "Epigenetic mechanisms of endothelial dysfunction in type 2 diabetes," *Clinical Epigenetics*, vol. 7, article 56, 2015.

[10] N. Duru, R. Gernapudi, G. Eades, R. Eckert, and Q. Zhou, "Epigenetic regulation of miRNAs and breast cancer stem cells," *Current Pharmacology Reports*, vol. 1, no. 3, pp. 161–169, 2015.

[11] S. J. Van Dijk, P. L. Molloy, H. Varinli et al., "Epigenetics and human obesity," *International Journal of Obesity*, vol. 39, no. 1, pp. 85–97, 2015.

[12] P. Cordero, J. Li, and J. A. Oben, "Epigenetics of obesity: beyond the genome sequence," *Current Opinion in Clinical Nutrition & Metabolic Care*, vol. 18, no. 4, pp. 361–366, 2015.

[13] T. Palomo, T. Archer, R. J. Beninger, and R. M. Kostrzewa, "Gene-environment interplay in neurogenesis and neurodegeneration," *Neurotoxicity Research*, vol. 6, no. 6, pp. 415–434, 2004.

[14] T. L. Spires and A. J. Hannan, "Nature, nurture and neurology: gene-environment interactions in neurodegenerative disease. FEBS Anniversary Prize Lecture delivered on 27 June 2004 at the 29th FEBS Congress in Warsaw," *FEBS Journal*, vol. 272, no. 10, pp. 2347–2361, 2005.

[15] F. Coppedè, M. Mancuso, G. Siciliano, L. Migliore, and L. Murri, "Genes and the environment in neurodegeneration," *Bioscience Reports*, vol. 26, no. 5, pp. 341–367, 2006.

[16] S. C. Marques, C. R. Oliveira, C. M. Pereira, and T. F. Outeiro, "Epigenetics in neurodegeneration: a new layer of complexity," *Progress in Neuro-Psychopharmacology & Biological Psychiatry*, vol. 35, no. 2, pp. 348–355, 2011.

[17] F. Coppedè, "Genetics and epigenetics of Parkinson's disease," *The Scientific World Journal*, vol. 2012, Article ID 489830, 12 pages, 2012.

[18] J. M. Greer and P. A. McCombe, "The role of epigenetic mechanisms and processes in autoimmune disorders," *Biologics*, vol. 6, pp. 307–327, 2012.

[19] G. Ravaglia, P. Forti, F. Maioli et al., "Homocysteine and folate as risk factors for dementia and Alzheimer disease," *The American Journal of Clinical Nutrition*, vol. 82, no. 3, pp. 636–643, 2005.

[20] J. Wu, M. R. Basha, B. Brock et al., "Alzheimer's disease (AD)-like pathology in aged monkeys after infantile exposure to environmental metal lead (Pb): evidence for a developmental origin and environmental link for AD," *The Journal of Neuroscience*, vol. 28, no. 1, pp. 3–9, 2008.

[21] A. Baccarelli and V. Bollati, "Epigenetics and environmental chemicals," *Current Opinion in Pediatrics*, vol. 21, no. 2, pp. 243–251, 2009.

[22] I. A. Qureshi and M. F. Mehler, "Advances in epigenetics and epigenomics for neurodegenerative diseases," *Current Neurology and Neuroscience Reports*, vol. 11, no. 5, pp. 464–473, 2011.

[23] P. Kaliman, M. J. Álvarez-López, M. Cosín-Tomás, M. A. Rosenkranz, A. Lutz, and R. J. Davidson, "Rapid changes in histone deacetylases and inflammatory gene expression in expert meditators," *Psychoneuroendocrinology*, vol. 40, no. 1, pp. 96–107, 2014.

[24] V. Nicolia, M. Lucarelli, and A. Fuso, "Environment, epigenetics and neurodegeneration: focus on nutrition in Alzheimer's disease," *Experimental Gerontology*, vol. 68, pp. 8–12, 2015.

[25] D. Mastroeni, A. Grover, E. Delvaux, C. Whiteside, P. D. Coleman, and J. Rogers, "Epigenetic changes in Alzheimer's disease: decrements in DNA methylation," *Neurobiology of Aging*, vol. 31, no. 12, pp. 2025–2037, 2010.

[26] Y. I. Francis, M. Fà, H. Ashraf et al., "Dysregulation of histone acetylation in the APP/PS1 mouse model of Alzheimer's disease," *Journal of Alzheimer's Disease*, vol. 18, no. 1, pp. 131–139, 2009.

[27] E. Martí, L. Pantano, M. Bañez-Coronel et al., "A myriad of miRNA variants in control and Huntington's disease brain regions detected by massively parallel sequencing," *Nucleic Acids Research*, vol. 38, no. 20, pp. 7219–7235, 2010.

[28] D. M. Mintzer, S. N. Billet, and L. Chmielewski, "Drug-induced hematologic syndromes," *Advances in Hematology*, vol. 2009, Article ID 495863, 11 pages, 2009.

[29] T. Uhlig, K. B. Hagen, and T. K. Kvien, "Current tobacco smoking, formal education, and the risk of rheumatoid arthritis," *Journal of Rheumatology*, vol. 26, no. 1, pp. 47–54, 1999.

[30] A. Kuhn and S. Beissert, "Photosensitivity in lupus erythematosus," *Autoimmunity*, vol. 38, no. 7, pp. 519–529, 2005.

[31] M. N. Artyomov, A. Meissner, and A. K. Chakraborty, "A model for genetic and epigenetic regulatory networks identifies rare pathways for transcription factor induced pluripotency," *PLoS Computational Biology*, vol. 6, no. 5, Article ID e1000785, 2010.

[32] K. Raghavan, H. J. Ruskin, D. Perrin, F. Goasmat, and J. Burns, "Computational micromodel for epigenetic mechanisms," *PLoS ONE*, vol. 5, no. 11, Article ID e14031, 2010.

[33] L. Y. Wang, P. Wang, M. J. Li et al., "EpiRegNet: constructing epigenetic regulatory network from high throughput gene expression data for humans," *Epigenetics*, vol. 6, no. 12, pp. 1505–1512, 2011.

[34] S. Althammer, A. Pagès, and E. Eyras, "Predictive models of gene regulation from high-throughput epigenomics data," *Comparative and Functional Genomics*, vol. 2012, Article ID 284786, 13 pages, 2012.

[35] A. P. Turner, M. A. Lones, L. A. Fuente, S. Stepney, L. S. D. Caves, and A. M. Tyrrell, "The incorporation of epigenetics in artificial gene regulatory networks," *BioSystems*, vol. 112, no. 2, pp. 56–62, 2013.

[36] K.-E. Lee and H.-S. Park, "A review of three different studies on hidden markov models for epigenetic problems: a computational perspective," *Genomics & Informatics*, vol. 12, no. 4, pp. 145–150, 2014.

[37] K. K.-H. Farh, A. Marson, J. Zhu et al., "Genetic and epigenetic fine mapping of causal autoimmune disease variants," *Nature*, vol. 518, no. 7539, pp. 337–343, 2015.

[38] A. T. Kodamullil, E. Younesi, M. Naz, S. Bagewadi, and M. Hofmann-Apitius, "Computable cause-and-effect models of healthy and Alzheimer's disease states and their mechanistic differential analysis," *Alzheimer's & Dementia*, 2015.

[39] N. L. Catlett, A. J. Bargnesi, S. Ungerer et al., "Reverse causal reasoning: applying qualitative causal knowledge to the interpretation of high-throughput data," *BMC Bioinformatics*, vol. 14, article 340, 2013.

[40] F. Martin, T. M. Thomson, A. Sewer et al., "Assessment of network perturbation amplitudes by applying high-throughput data to causal biological networks," *BMC Systems Biology*, vol. 6, article 54, 2012.

[41] D. Laifenfeld, D. A. Drubin, N. L. Catlett et al., "Early patient stratification and predictive biomarkers in drug discovery and development: a case study of ulcerative colitis anti-TNF therapy," *Advances in Experimental Medicine and Biology*, vol. 736, pp. 645–653, 2012.

[42] D. A. Fryburg, D. H. Song, and D. De Graaf, "Early patient stratification is critical to enable effective and personalised drug discovery and development," *Drug Discovery World*, vol. 12, no. 3, pp. 47–56, 2011.

[43] B. Thomas and M. F. Beal, "Molecular insights into Parkinson's disease," *F1000 Medicine Reports*, vol. 3, article 7, 2011.

[44] A. Jowaed, I. Schmitt, O. Kaut, and U. Wüllner, "Methylation regulates alpha-synuclein expression and is decreased in Parkinson's disease patients' brains," *The Journal of Neuroscience*, vol. 30, no. 18, pp. 6355–6359, 2010.

[45] L. Matsumoto, H. Takuma, A. Tamaoka et al., "CpG demethylation enhances alpha-synuclein expression and affects the pathogenesis of Parkinson's disease," *PLoS ONE*, vol. 5, no. 11, Article ID e15522, 2010.

[46] A. G. Riedl, P. M. Watts, P. Jenner, and C. D. Marsden, "P450 enzymes and Parkinson's disease: the story so far," *Movement Disorders*, vol. 13, no. 2, pp. 212–220, 1998.

[47] Y. Feng, J. Jankovic, and Y.-C. Wu, "Epigenetic mechanisms in Parkinson's disease," *Journal of the Neurological Sciences*, vol. 349, no. 1-2, pp. 3–9, 2015.

[48] O. W. Wan and K. K. K. Chung, "The role of alpha-synuclein oligomerization and aggregation in cellular and animal models of Parkinson's disease," *PLoS ONE*, vol. 7, no. 6, Article ID e38545, 2012.

[49] O. Kaut, I. Schmitt, and U. Wüllner, "Genome-scale methylation analysis of Parkinson's disease patients' brains reveals DNA hypomethylation and increased mRNA expression of cytochrome P450 2E1," *Neurogenetics*, vol. 13, no. 1, pp. 87–91, 2012.

[50] M. J. Devine and P. A. Lewis, "Emerging pathways in genetic Parkinson's disease: tangles, Lewy bodies and LRRK2," *The FEBS Journal*, vol. 275, no. 23, pp. 5748–5757, 2008.

[51] C. Hague, M. A. Uberti, Z. Chen et al., "Olfactory receptor surface expression is driven by association with the beta2-adrenergic receptor," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 101, no. 37, pp. 13672–13676, 2004.

[52] E.-Y. Lee, P. J. Eslinger, G. Du, L. Kong, M. M. Lewis, and X. Huang, "Olfactory-related cortical atrophy is associated with olfactory dysfunction in Parkinson's disease," *Movement Disorders*, vol. 29, no. 9, pp. 1205–1208, 2014.

[53] P. Desplats, B. Spencer, E. Coffee et al., "Alpha-synuclein sequesters Dnmt1 from the nucleus: a novel mechanism for epigenetic alterations in Lewy body diseases," *The Journal of Biological Chemistry*, vol. 286, no. 11, pp. 9031–9037, 2011.

[54] P. Viitala, K. Posti, A. Lindfors, O. Pelkonen, and H. Raunio, "cAMP mediated upregulation of CYP2A5 in mouse hepatocytes," *Biochemical and Biophysical Research Communications*, vol. 280, no. 3, pp. 761–767, 2001.

[55] Y. Fujii, N. Osaki, T. Hase, and A. Shimotoyodome, "Ingestion of coffee polyphenols increases postprandial release of the active glucagon-like peptide-1 (GLP-1(7–36)) amide in C57BL/6J mice," *Journal of Nutritional Science*, vol. 4, article e9, 9 pages, 2015.

[56] Y. Li, T. Perry, M. S. Kindy et al., "GLP-1 receptor stimulation preserves primary cortical and dopaminergic neurons in cellular and rodent models of stroke and Parkinsonism," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 106, no. 4, pp. 1285–1290, 2009.

[57] P. Prasad, A. K. Tiwari, K. M. P. Kumar et al., "Association analysis of ADPRT1, AKR1B1, RAGE, GFPT2 and PAI-1 gene polymorphisms with chronic renal insufficiency among Asian Indians with type-2 diabetes," *BMC Medical Genetics*, vol. 11, article 52, 2010.

[58] E. Masliah, W. Dumaop, D. Galasko, and P. Desplats, "Distinctive patterns of DNA methylation associated with Parkinson disease: identification of concordant epigenetic changes in brain and peripheral blood leukocytes," *Epigenetics*, vol. 8, no. 10, pp. 1030–1038, 2013.

[59] C. Song, A. Kanthasamy, V. Anantharam, F. Sun, and A. G. Kanthasamy, "Environmental neurotoxic pesticide increases histone acetylation to promote apoptosis in dopaminergic neuronal cells: relevance to epigenetic mechanisms of neurodegeneration," *Molecular Pharmacology*, vol. 77, no. 4, pp. 621–632, 2010.

[60] I. F. Harrison and D. T. Dexter, "Epigenetic targeting of histone deacetylase: therapeutic potential in Parkinson's disease?" *Pharmacology and Therapeutics*, vol. 140, no. 1, pp. 34–52, 2013.

[61] D. Qu, J. Rashidian, M. P. Mount et al., "Role of Cdk5-Mediated Phosphorylation of Prx2 in MPTP Toxicity and Parkinson's Disease," *Neuron*, vol. 55, no. 1, pp. 37–52, 2007.

[62] E. Miñones-Moyano, S. Porta, G. Escaramís et al., "MicroRNA profiling of Parkinson's disease brains identifies early downregulation of miR-34b/c which modulate mitochondrial function," *Human Molecular Genetics*, vol. 20, no. 15, pp. 3067–3078, 2011.

[63] D. A. Umphred, *Neurological Rehabilitation*, Mosby, St. Louis, Mo, USA, 2001.

[64] M. Iridoy Zulet, L. Pulido Fontes, T. Ayuso Blanco, F. Lacruz Bescos, and M. Mendioroz Iriarte, "Epigenetic changes in neurology: DNA methylation in multiple sclerosis," *Neurología*, 2015.

[65] S. Ruhrmann, P. Stridh, L. Kular, and M. Jagodic, "Genomic imprinting: a missing piece of the multiple sclerosis puzzle?" *The International Journal of Biochemistry & Cell Biology*, vol. 67, pp. 49–57, 2015.

[66] Z. Zhang and R. Zhang, "Epigenetics in autoimmune diseases: pathogenesis and prospects for therapy," *Autoimmunity Reviews*, vol. 14, no. 10, pp. 854–863, 2015.

[67] X. Pedre, F. Mastronardi, W. Bruck, G. López-Rodas, T. Kuhlmann, and P. Casaccia, "Changed histone acetylation patterns in normal-appearing white matter and early multiple sclerosis lesions," *The Journal of Neuroscience*, vol. 31, no. 9, pp. 3435–3445, 2011.

[68] A. Keller, P. Leidinger, J. Lange et al., "Multiple sclerosis: microRNA expression profiles accurately differentiate patients with relapsing-remitting disease from healthy controls," *PLoS ONE*, vol. 4, no. 10, Article ID e7440, 2009.

[69] C. İ. Küçükali, M. Kürtüncü, A. Çoban, M. Çebi, and E. Tüzün, "Epigenetics of multiple sclerosis: an updated review," *NeuroMolecular Medicine*, vol. 17, no. 2, pp. 83–96, 2015.

[70] A. Junker, M. Krumbholz, S. Eisele et al., "MicroRNA profiling of multiple sclerosis lesions identifies modulators of the regulatory protein CD47," *Brain*, vol. 132, no. 12, pp. 3342–3352, 2009.

[71] M. Sospedra and R. Martin, "Immunology of multiple sclerosis," *Annual Review of Immunology*, vol. 23, pp. 683–747, 2005.

[72] T. L. Sørensen, M. Tani, J. Jensen et al., "Expression of specific chemokines and chemokine receptors in the central nervous system of multiple sclerosis patients," *Journal of Clinical Investigation*, vol. 103, no. 6, pp. 807–815, 1999.

[73] R. Gandhi, A. Laroni, and H. L. Weiner, "Role of the innate immune system in the pathogenesis of multiple sclerosis," *Journal of Neuroimmunology*, vol. 221, no. 1-2, pp. 7–14, 2010.

[74] Y. C. Q. Zang, S. Li, V. M. Rivera et al., "Increased $CD8^+$ cytotoxic T cell responses to myelin basic protein in multiple sclerosis," *Journal of Immunology*, vol. 172, no. 8, pp. 5120–5127, 2004.

Hindawi Publishing Corporation
Journal of Immunology Research
Volume 2015, Article ID 173593, 20 pages
http://dx.doi.org/10.1155/2015/173593

Research Article

Geometry Dynamics of α-Helices in Different Class I Major Histocompatibility Complexes

Reiner Ribarics,[1] Michael Kenn,[1] Rudolf Karch,[1] Nevena Ilieva,[2] and Wolfgang Schreiner[1]

[1] *Section of Biosimulation and Bioinformatics, Center for Medical Statistics, Informatics and Intelligent Systems (CeMSIIS), Medical University of Vienna, Spitalgasse 23, 1090 Vienna, Austria*
[2] *Institute of Information and Communication Technologies (IICT), Bulgarian Academy of Sciences, Acad. G. Bonchev Street, Block 25A, 1113 Sofia, Bulgaria*

Correspondence should be addressed to Wolfgang Schreiner; wolfgang.schreiner@meduniwien.ac.at

Received 27 July 2015; Revised 30 September 2015; Accepted 30 September 2015

Academic Editor: Francesco Pappalardo

Copyright © 2015 Reiner Ribarics et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

MHC α-helices form the antigen-binding cleft and are of particular interest for immunological reactions. To monitor these helices in molecular dynamics simulations, we applied a parsimonious fragment-fitting method to trace the axes of the α-helices. Each resulting axis was fitted by polynomials in a least-squares sense and the curvature integral was computed. To find the appropriate polynomial degree, the method was tested on two artificially modelled helices, one performing a bending movement and another a hinge movement. We found that second-order polynomials retrieve predefined parameters of helical motion with minimal relative error. From MD simulations we selected those parts of α-helices that were stable and also close to the TCR/MHC interface. We monitored the curvature integral, generated a ruled surface between the two MHC α-helices, and computed interhelical area and surface torsion, as they changed over time. We found that MHC α-helices undergo rapid but small changes in conformation. The curvature integral of helices proved to be a sensitive measure, which was closely related to changes in shape over time as confirmed by RMSD analysis. We speculate that small changes in the conformation of individual MHC α-helices are part of the intrinsic dynamics induced by engagement with the TCR.

1. Introduction

T cells play a major role in both innate immunity and adaptive immunity. Their surface-bound T cell receptors (TCRs) recognise antigens and thereby detect hazardous organisms or changes inside cells. TCRs recognize short peptide fragments (p) that are bound to major histocompatibility complexes (MHC). Different interactions of TCR/peptide-MHC (TCR/pMHC) are believed to be the basis for distinctive stimuli that lead to and trigger different fates of the T cell, for example, T cell development, thymic selection, lineage commitment, differentiation into effector cells, or memory T cell responses to foreign antigens [1].

MHCs are surface-bound proteins and their role is to present peptide fragments to TCRs, be they self- or alloantigens. To achieve this, MHCs have a cleft that is able to bind peptide fragments. This cleft comprises two α-helices and five subjacent lateral β-strands forming a sheet or floor of the cleft. TCRs engage peptide MHCs in a diagonal arrangement that seems to be a common mode of interaction across TCRs [2, 3]. The two MHC α-helices interact with the TCR complementarity determining regions (CDR) 1 and 2, while the MHC-bound peptide interacts with CDR3, although CDR1 has also been shown to interact with the terminal parts of the peptide. Most of the sequence variability of TCRs is found within CDRs; these regions are also referred to as hypervariable regions. The two MHC α-helices are of particular interest as they represent stable secondary structural domains interacting with TCRs.

Adhesion and signaling proteins together with a set of TCR/pMHC complexes form a junction between T cell and an antigen-presenting cell that is referred to as an immunological synapse. It serves as a platform for assembly and segregation of signaling complexes, which cooperatively

decide the outcome of T cell activation and effector function. New findings show that this supramolecular complex forms too late to be relevant for initial TCR signaling that happens within seconds after pMHC engagement, with the TCR initiating a tyrosine phosphorylation cascade [4]. Such an early signal may be sufficient to trigger effector function like killer T cell cytolysis of target cells. Stimulation of proliferation, however, requires engagement and signaling for many minutes or even hours [5–8]. It is important to note that $\alpha\beta$-TCR itself has no signaling motif but associates with homo- and heterodimeric cluster of differentiation 3 (CD3) subunits, $\zeta\zeta$, $\varepsilon\delta$, and ε in a noncovalent way. These subunits contain immunoreceptor tyrosine-based activation motifs (ITAMs) that can be phosphorylated and initiate downstream signaling of TCR activation.

Protein structures as found in protein databases (e.g., Protein Data Bank (PDB)) show a static image. In contrast to that, in molecular dynamics (MD) simulations, the protein explores many conformations and allows one to capture its dynamics. Computational immunoinformatics is a well-established, rapidly evolving field [9]. In previous papers [10, 11] we presented a first evaluation of three TCR/pMHC systems that differ only slightly in the MHC amino acid sequence. Macdonald et al. [12] determined binding characteristics and immunogenicity of MHC alleles HLA-B*44:02, HLA-B*44:03, and HLA-B*44:05 in complex with the ABCD3 self-peptide (EEYLQAFTY) and LC13 TCR. HLA-B*44:02 and B*44:05 trigger an immune response when bound to the LC13 TCR, while HLA-B*44:02 does not despite extensive amino acid sequence homology. This renders these HLA alleles an interesting set to study TCR allorecognition. Macdonald et al. [12] determined the X-ray structure of the ternary complex HLA-B*44:05, ABCD3 nonapeptide, and LC13 TCR that is accessible on the protein database (PDB, http://www.pdb.org/, PDB-ID: 3KPS). Due to extensive sequence identity we were able to use in silico site-directed mutagenesis to obtain 3D structures of the missing TCR/pMHC complexes. MD simulations in the nanosecond range could possibly show short-lived changes in dynamic behaviour, conformation, propagation of forces, or early activation signals [13, 14].

The aim of the present paper is to put forward adequate tools for identifying and monitoring of conformational changes with possible functional relevance in MHC α-helices, and in particular to monitor geometric characteristics of the MHC peptide-binding groove. Here, we present (i) a robust and parsimonious method to find an approximation of a protein's α-helical axis, (ii) an evaluation of polynomial degree adequacy in describing bending or hinge movements of particular α-helical axes, and (iii) application of polynomial fitting of α-helical axis to monitor α-helical conformations along MD simulations in different TCR/pMHC immunological complexes.

Our previous studies established the use of polynomials to model α-helices in MHC molecules and monitor their dynamic behaviour [15]. To mathematically describe the structural dynamics of MHC α-helices at the TCR/pMHC binding interface we first identified those helical regions which were stable and therein those which are close to the protein-protein interface. Then we extracted the α-helical axis and finally determined a polynomial that approximates this axis in a least-squares sense.

Table 1: Molecular dynamics simulations.

Number	Molecular system (TCR/peptide/MHC)	Simulation length
1	LC13 TCR/ABCD3/HLA-B*44:02 (B4402)	250 ns
2	LC13 TCR/ABCD3/HLA-B*44:03 (B4403)	250 ns
3	LC13 TCR/ABCD3/HLA-B*44:05 (B4405)	250 ns

Various methods to extract a helix axis have been developed [16], including rotational fitting, using C_α atoms as control points of B-splines [17] or fitting to a helix. We used a fragment-fitting method, based on previous work [16], to locate the axis of α-helices as follows: an ideal, linear α-helix fragment comprising four C_α atoms is superimposed on successive pieces of MHC α-helices in a least-squares sense. Along the axis of the fitted helical fragment, we adopt points as estimates of the MHC α-helix axis and fit a polynomial through these points. From the polynomial, geometric parameters can be derived to monitor conformational changes. Polynomials fitted to the α-helical axis can, in principle, be of any degree. However, polynomials of higher order tend to oscillate, adding noise to geometrical quantities computed thereof. We therefore evaluated polynomials of different degrees for their ability to reproduce bending and hinge motions of an α-helix with minimum relative error. Between two adjacent α-helices, as found in MHC proteins, the polynomials serve to span a ruled surface. This interhelical surface lends itself to derive several quantitative characteristics of shape: (a) total area [18, 19], (b) a profile of interhelical distances along the binding cleft, and (c) heuristic "centre line of the cleft" which may be constructed, along which the surface torsion, that is, a twist or screw of the interhelical surface, can be computed. The latter characterizes the positions and bending of helices relative to each other and defines the geometrical shape of the peptide-binding cleft that is ligated to the TCR. Dynamics in the shape of the protein-protein interface might modulate the TCR/pMHC binding affinity.

2. Methods

2.1. Homology Modelling of TCR/pMHC Complexes. Three TCR/pMHC systems listed in Table 1 were simulated. The X-ray structure of TCR/pMHC B4405 (number 3 in Table 1) was taken from the PDB (PDB-ID: 3KPS). Structures B4402 and B4403 were engineered by means of in silico site-directed mutagenesis [20] using B4405 as a structural template. We introduced

(i) mutation Y116D to the MHC molecule to get LC13/ABCD3/HLA-B*44:02 complex (B4402),

(ii) mutations Y116D and D156L to the MHC molecule to get LC13/ABCD3/HLA-B*44:03 complex (B4403).

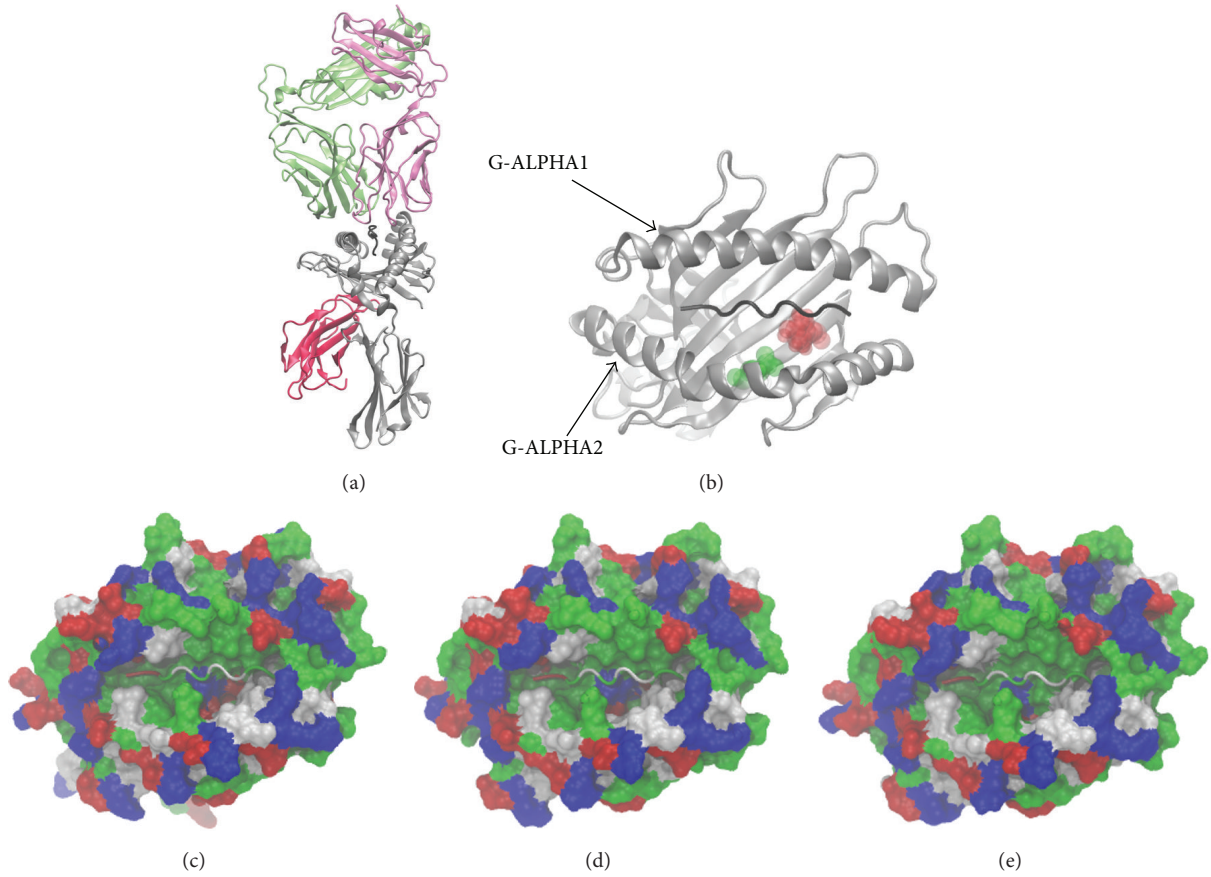


Figure 1: TCR/pMHC. The three HLA molecules studied in this paper are closely related and differ only at amino acid position 116 and/or 156. The X-ray structure of HLA-B*44:05 in complex with LC13 TCR and ABCD3 peptide (b) is available from http://www.pdb.org (PDB-ID: 3KPS) and was used as a template to model similar systems containing HLA-B*44:03 and HLA-B*44:02. HLA-B*44:05 $\xrightarrow{\text{Y116D}}$ HLA-B*44:02 $\xrightarrow{\text{D156L}}$ HLA-B*44:03. (a) Cartoon representation of TCR/pMHC system. The TCR comprises two chains (lime and pink). Each chain is made up of a constant domain and a variable domain. The constant domain faces the membrane. The CDR loops 1–3 are highly polymorphic regions that interact with the MHC. Beta-sheets are the main secondary structural element of the TCR. MHC (grey) class I comprises alpha-helices and beta-sheets. Alpha-helices G-ALPHA1 and G-ALPHA2 together with the underlying beta-sheets comprise the peptide-binding pocket and present digested peptide fragments on the cell surface. (b) Cartoon representation of MHC class I. HLA molecule (grey), peptide (black), tyrosine at position 116 (red), and aspartic acid at position 156 (green). (c, d, and e) Surface representation of MHC binding grooves of B4402 (c), B4403 (d), and B4405 (e). Nonpolar residues (white), basic residues (blue), acidic residues (red), and polar residues (green). The ABCD3 peptide is embraced in the peptide-binding groove displayed in cartoon representation. Helix G-ALPHA2 is dominated by alternating acidic and basic residues. The Y116D mutation introduces a negatively charged residue (compare panel (c) with (b): a red spot appears at the right-hand side of the peptide). The D156L mutation substitutes a charged residue with an apolar residue. Structures are taken from the first frame of MD simulations.

See Figure 1(b) for a 3D representation of amino acid positions Y116 and D156. 3D structures were edited and mutations introduced with the Swiss PDB Viewer. The program automatically browses a rotamer library and selects an amino acid rotamer minimizing a scoring function in order to fit the new amino acid in its surrounding and avoid steric clashes with other atoms. All MHC molecules simulated have an amino acid sequence identity of more than 99% and stay in very similar 3D fold. MHC molecules appeared stable during all our MD simulations as seen in RMSD plots in Figure 15.

The full amino acid sequence of HLA-B alleles is accessible from the HLA library (https://www.ebi.ac.uk/ipd/imgt/hla/) and a description of its topology is accessible from UniProt (http://www.uniprot.org/uniprot/Q95365). Not surprisingly, a sequence comparison showed that a transmembrane helix (24 amino acids) and cytoplasmic tail (30 amino acids) are missing from the MHC X-ray structure as plasma membrane structures and flexible protein parts are hard to determine using X-ray crystallography. The LC13 TCR is also missing its transmembrane helix.

2.2. Molecular Dynamics Simulations. GROMACS 4.0.7 was used for molecular dynamics simulation. First, water molecules were added to the protein structure, immersing it in an artificial water bath of rectangular form and allowing a minimum distance of 2 nm between the protein and

the box boundaries. Second, water molecules were replaced by sodium and chloride ions to yield a salt concentration of 0.15 mol/L and neutralize the protein net charge. Third, the energy of the solvated system was minimized using a steepest descent method and then the system was gradually heated up to 310 K during 100 ps position restraint MD simulation. Finally, MD production runs were done with the LINCS constraint algorithm acting on all bonds using the Gromos 53A6 force field [21]. Hydrogen and fast improper dihedral motions were removed, allowing for an integration step of 5 fs. Van der Waals and Coulomb interactions were computed using a cut-off of 1.4 nm. Long-range Coulomb interactions were computed by Ewald summation. Velocity rescale temperature coupling was set to 310 K and Berendsen isotropic pressure coupling was set to 1 bar. The selection of the force field and MD parameters for pMHCs were evaluated by Omasits et al. [22] and set accordingly.

2.3. Finding Dynamically Stable α-Helices at the Protein-Protein Interface. As outlined in the introduction, the MHC protein comprises α-helices and beta-sheets. We are interested in monitoring C_α atoms that form stable α-helices and are in close contact with the TCR. From MD simulation data we calculated the following:

(i) The relative presence of α-helical structure in the protein complexes over the simulation time: we used the DSSP algorithm [23] as implemented in GROMACS to identify secondary structural elements of the protein complex over simulation time. α-helices were considered stable if the ratio

$$\frac{t_{\alpha\text{-helix}}}{t_{\text{sim}}} \geq 0.5 \tag{1}$$

with $t_{\alpha\text{-helix}}$ being the time the main chain (backbone atoms plus carbonyl oxygen atoms) meets the DSSP criterion for an α-helix and t_{sim} being the total simulation time. The cut-off value of 0.5 is justified by the distinctly bimodal distribution (see Figure 11(a)).

(ii) The relative presence of close contacts at the protein-protein interface: C_α atoms that are less than 1.4 nm apart (i.e., the cut-off for electrostatic interactions in our MD simulations) are defined as being in close contact. Atom-atom contacts between TCR and MHC are defined stable if the ratio

$$\frac{t_{\text{contact}}}{t_{\text{sim}}} \geq 0.5 \tag{2}$$

t_{contact} is the time during which two atoms are in close contact, and t_{sim} is the simulation time. Again, the cut-off value of 0.5 is justified by the distinctly bimodal distribution (see Figure 11(b)). To get the residue-wise relative contact time, we averaged the atom-wise relative contact time (defined in (2)) over all atoms per residue.

The resulting sets of amino acid residues defined by procedures (i) and (ii) were intersected (workflow, see Figure 2; results, see Figure 3) before applying further methods described in Sections 2.4 and 2.5. Dynamically stable helices (green atomic surface in Figure 3) as defined by procedure (i) overlap well with the atoms in close contact of the TCR (red atomic surface in Figure 3) especially with the cut-off set to 1.4 nm. Note that at the end of the MHC's peptide-binding pocket both G-ALPHA1 and G-ALPHA2 exhibit a kink followed by a short α-helix. These short α-helices are present in the crystal structure, but we found them being unstable during MD simulations as they fold and unfold. We do not consider these helices in our analysis, as they are not in close contact with the TCR.

2.4. Approximating the Axis of an α-Helix. In order to mathematically describe and quantify α-helical geometry and movements, polynomials $\mathbf{c}(u)$ of degree m were fitted to the α-helices, where $\mathbf{c}$ represents a vector of 3D coordinates and u is the curve parameter. Prior to fitting we extracted C_α atom coordinates of those amino acids, which fulfil the criterion of stable α-helices and close contacts as described in Section 2.3.

According to the structural definition of α-helices [24] a model of one ideal α-helical turn, that is, a fragment consisting of four C_α atoms, is constructed, with its axis coinciding with x-axis. The coordinates of its kth C_α were assigned as

$$\begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} p \cdot (k-1) \\ r \cdot \cos\left(\varphi \cdot (k-1)\right) \\ r \cdot \sin\left(\varphi \cdot (k-1)\right) \end{pmatrix} \quad k = 1, \ldots, 4 \tag{3}$$

with pitch $p = 0.15$ nm (advance from one C_α to the next), helix radius $r = 0.23$ nm, and $\varphi = 100 \cdot \pi/180$. Along the axis of this helical fragment we consider three axis points, the *initial* $(0, 0, 0)$, the *intermediate* $(1.5 \cdot p, 0, 0)$, and the *final* $(3.0 \cdot p, 0, 0)$.

Out of an α-helix with N C_α atoms we pick moving groups of four successive C_α atoms each, to which we fit the fragment model defined above in a least-squares sense [25]. Along x-axis of the fitted fragment model we adopt points as estimates of the axis of the α-helix, see Figure 4. From the very first fitted fragment (C_α atoms $1, \ldots, 4$) we adopt two points as points of the helix: the transformed initial axis point as $\mathbf{a}_1$ and the transformed *intermediate* point as $\mathbf{a}_2$. From fragments fitted subsequently we adopt only the respective intermediate points, and from the last fragment (fitted to C_α atoms $N-3, \ldots, N$), we again adopt 2 points, its "intermediate" as $\mathbf{a}_{N-2}$ and its final point as $\mathbf{a}_{N-1}$. Thus, $N-1$ points $(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{N-1})$ represent the axis of the α-helix.

2.5. Fitting the Axis of an α-Helix. The points $(\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_{N-1})$ representing the helical axis are fitted by polynomials $\mathbf{c}(u)$ of degree m in a least-squares sense. Separate regression functions f_x, f_y, and f_z are computed for x-, y-, and z-coordinates:

$$\mathbf{c}(u) = \begin{pmatrix} f_x(u) \\ f_y(u) \\ f_z(u) \end{pmatrix}; \tag{4}$$

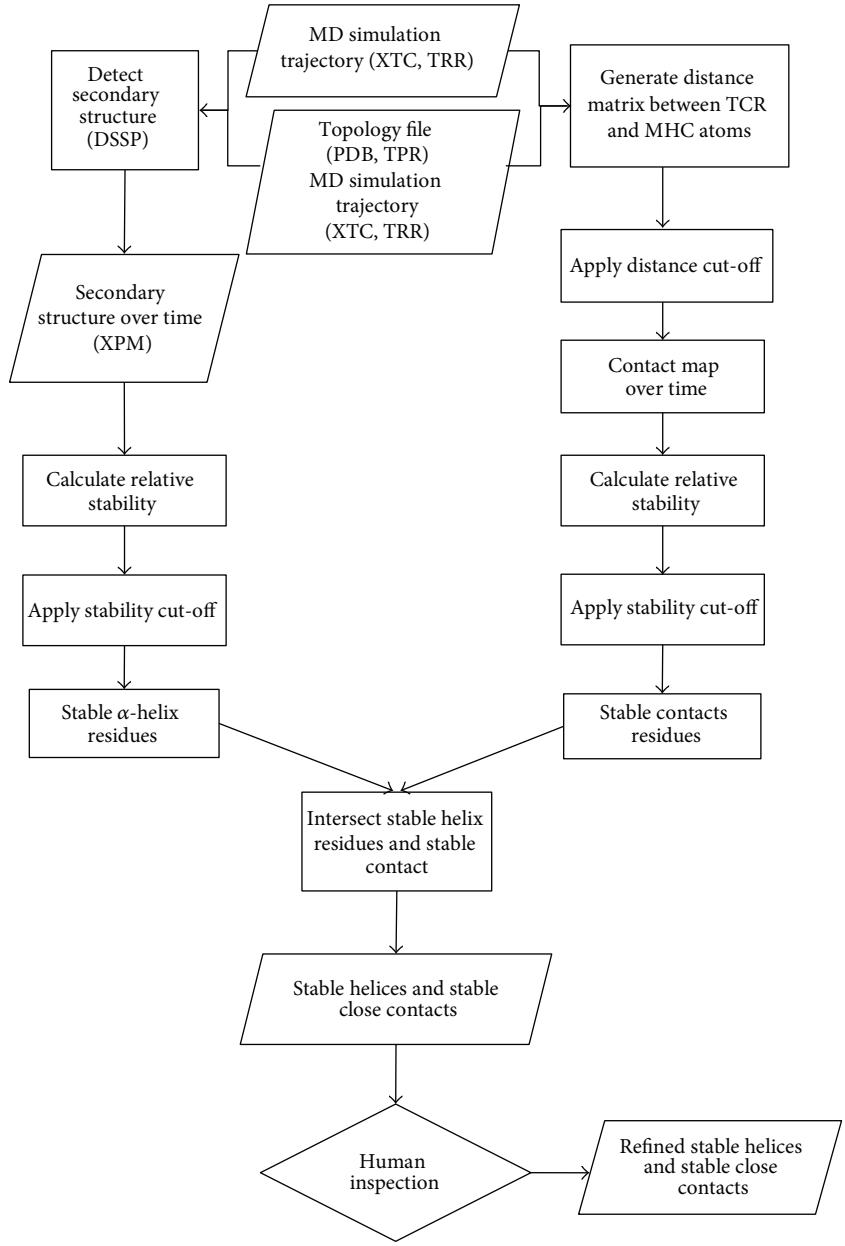


Figure 2: Workflow for selecting stable α-helical contact residues. Starting from MD simulation data we calculated the relative presence of α-helical structures using the DSSP algorithm (left path) and the relative presence of close contacts over the simulation time (right path). The resulting sets of amino acids are intersected as to yield one list of amino acids that fulfil both criteria: (i) being located within a certain distance to the TCR for more than half of the simulation time and (ii) being part of an α-helix for more than half of the simulation time. The process results in a list of amino acids that are stable α-helices and stable close contacts. The authors inspected the list in order to rule out the fact that only parts of α-helices were selected. Subjecting only parts of a helix to the fragment-fitting method would result in the calculation of a meaningless helical axis.

for example, for x-coordinate,

$$\begin{aligned}\mathbf{c}_x(u) &= f_x(u) \\ &= p_{x,m}u^m + p_{x,m-1}u^{m-1} + \cdots + p_{x,1}u + p_0.\end{aligned} \tag{5}$$

The total number of parameters for this model is $N_{\text{parameter}} = 3 \cdot (m + 1)$. When fitting the curve, parameter u is evaluated at discrete values $u = 1, 2.5, 3.5, \ldots, N - 1.5, N$ of pitch, corresponding to the positions of estimated points along the helical axis. After the regression has been performed, $\mathbf{c}(u)$ in (5) may be evaluated for arbitrary, continuous values $1 \leq u \leq N$, yielding a continuous representation of the helical axis.

Both G-ALPHA1 and G-ALPHA2 helices were modelled in the same way, yielding models $\mathbf{c}_1(u)$ and $\mathbf{c}_2(u)$ with equal polynomial degrees k. It is well known that (with equidistant data points) interpolating polynomials of too high a degree

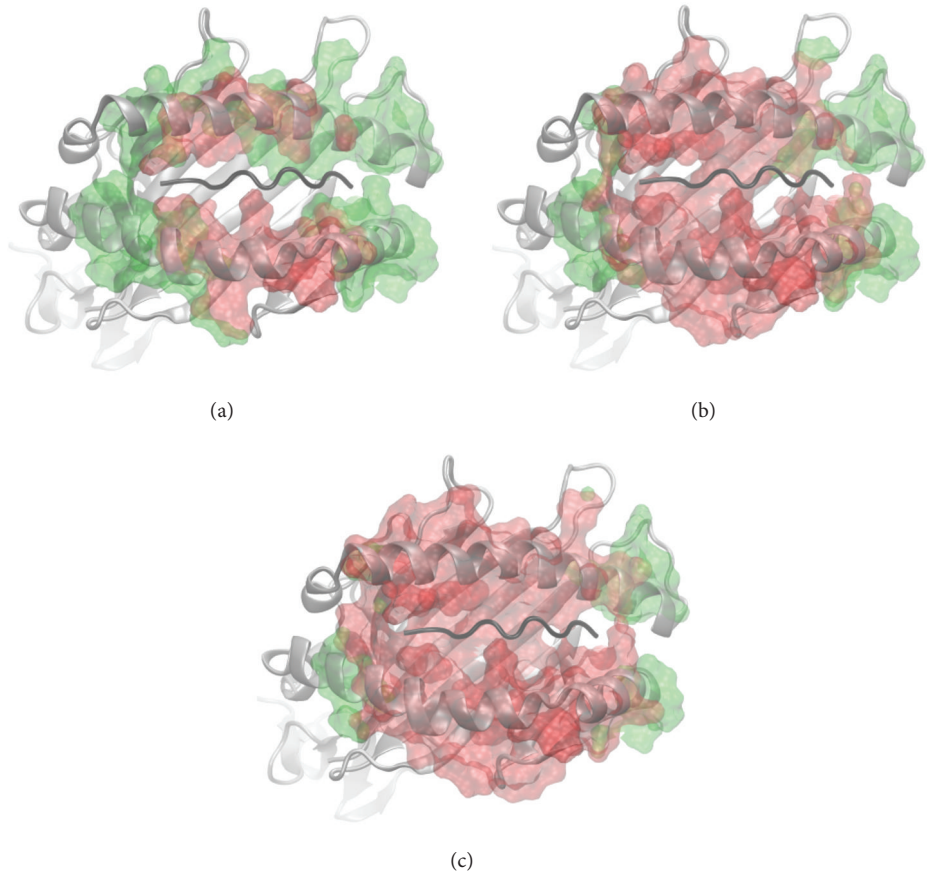


Figure 3: Helical residues and close TCR contacts. The atomic surface in green represents stable α-helical amino acids of the MHC that were identified by the method described in Section 2.3. The atomic surface in red represents amino acids in close proximity of the TCR with varying distance cut-offs: (a) 0.8 nm, (b) 1.2 nm, and (c) 1.4 nm. Some parts of the MHC α-helices G-ALPHA1 and G-ALPHA2 comprise the protein-protein interface between TCR and pMHC and, as expected, atomic surfaces in green and red overlap. However, not all parts of the α-helices belong to the close contacts even when the cut-off is set at 1.4 nm. For calculation of the helical axis we skipped the parts that are not overlapping and not directly interacting with the TCR. Visualization was done with VMD [43]. The contact map was calculated between C_α atoms to determine which atoms are in close contacts.

may exhibit severe oscillations near the ends of the interpolation interval [26]. (Interpolating polynomials actually pass through all data points.) This is also true for approximating polynomials [27]. (Approximating polynomials need not pass through data points but rather approximate the shape of their functional dependence in a least-square sense.) We therefore kept the polynomial degrees as low as possible.

2.6. Geometric Quantities

2.6.1. Interhelical Distance and Area of Interhelical Surface. For each polynomial, the curve parameter ranges within $1 \leq u \leq N$, with possibly different values (N_1, N_2) for each helix. We consider $L = \min(N_1, N_2)$ equidistant values of u, yielding L reference points on each helix model. Connecting corresponding reference points by straight lines yields rulings, $\mathbf{X}(u) = \mathbf{c}_2(u) - \mathbf{c}_1(u)$, which span a ruled surface (see Figure 5). (The rule for defining "corresponding" points has to be adopted in a reasonable way but finally remains to some degree arbitrary.) From rulings we estimate distances $|\mathbf{X}(u)|$ between polynomials across the cleft. The resulting polygon mesh was triangulated and interhelical area A was calculated, as previously outlined [18, 28]. Each of these quantities may be monitored over time, for example, $A(t)$; see Figure 9. Respective graphs provide well-defined estimates of changes in width of the intrahelical gap (binding cleft) as a function of both helical position and time. Likewise, median, quartiles, and extreme values of interhelical distances for each u and of A can be obtained.

2.6.2. Torsion of Interhelical Surface. The ruled surface between both polynomial helix models may bend and wind in various ways. Describing all aspects would call for a comprehensive mathematical treatment in terms of differential geometry, from which we refrain. Instead, we restrict ourselves to describe something like the "torsion of the interhelical surface" in a simple, intuitive way; see Figure 5. (In everyday terminology "torsion" is often called "winding." It may apply to (curved) lines as well as to surfaces in 3D space. We use both terms in parallel with equal meaning.)

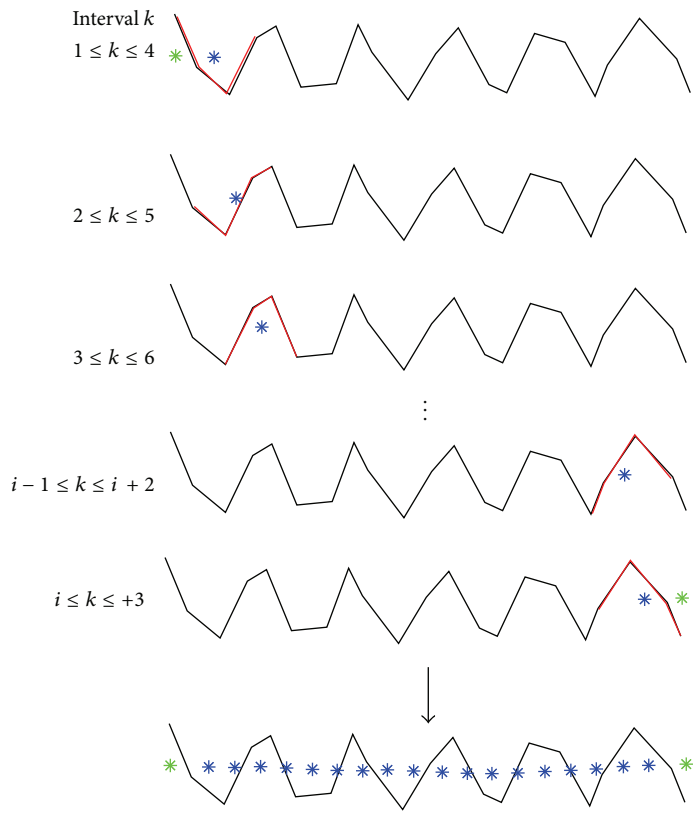
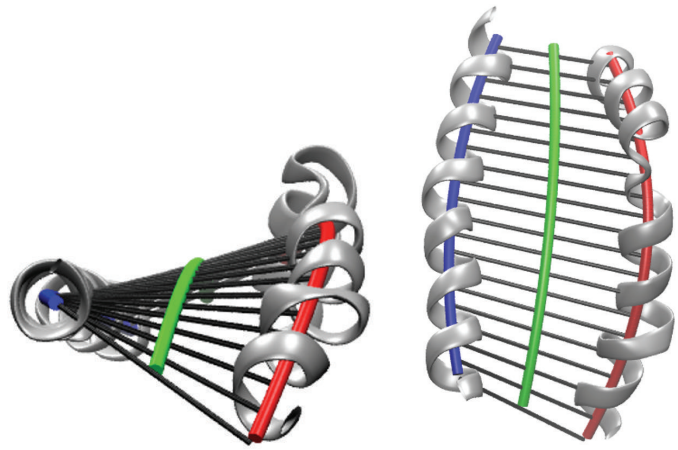


FIGURE 4: Visualization of the fragment-fitting process. An ideal helical fragment of four C_α atoms (red) is superimposed on successive pieces of an α-helix (grey) in a least-squares sense. Along the axis of the fitted helical fragment we adopt points (blue) as estimates of points on the axis of the MHC α-helix. From the very first and last superimposition we adopt one extra point each (green). The sequence of blue and green points (shown at the bottom of the figure) represents an estimate of the axis of the α-helix. Subsequently, a polynomial is fitted to these points in a least-squares sense, yielding a simple model of the α-helix from which geometric quantities can be derived.



(a) (b)

FIGURE 5: Surface torsion of the interhelical surface. Front view (left) and top-down view (right) on MHC α-helices G-ALPHA1 and G-ALPHA2 whose axes are modelled by second-degree polynomials (blue and red). Lines (in grey, called "rulings") between the polynomials span a ruled surface. Taking the mean coordinates of the lines coloured in blue and red results in the centre line (green). When moving over the surface along the centre line we see that rulings change direction, which can be quantified by a parameter called "surface torsion." Surface torsion describes the extent and orientation of twist of a surface along a given line, which is the centre line in our case. The surface torsion describes important aspects of the relative orientation of the two helix axes towards each other.

To this end we note that the observer may slide across a surface along various paths and may, in general, observe different values of surface torsion along each of the paths. We notice that torsion in strict mathematical terms is a local characteristic of the surface, and even more, it depends upon the path one takes to inspect the surface. For the sake of simplicity we deliberately adopt the centre line between both helix models:

$$\mathbf{m}(u) = \frac{1}{2}\left[\mathbf{c}_1(u) + \mathbf{c}_2(u)\right], \tag{6}$$

and let $\mathbf{t}(u)$ be the unit tangent vector of $\mathbf{m}(u)$. Of note, the centre line intersects with each of the rulings. While proceeding along the centre line we inspect the directional change of rulings. This provides us with a particular characterization of the deformation of the interhelical surface. To quantify this intuitive description, we introduce a parameter "surface torsion." It is based on the change in direction between successive rulings [29] and is determined by the derivative $\mathbf{X}'(u) = \mathbf{c}_1'(u) + \mathbf{c}_2'(u)$. (In strict mathematical terms (of differential geometry) this quantity is called "parameter of distribution," which we think is a very nonintuitive label, prone to be mixed up with probability distributions. Hence, we prefer the more intuitive term "surface torsion" to quantify the winding of a surface while proceeding along a given path (in our case the centre line), not that surface torsion is generally different in different directions, even in the very same point of a surface.) What we call *surface torsion* is given by [29]

$$\tau(u) = \frac{\left(\mathbf{X} \times \mathbf{X}'\right) \cdot \mathbf{t}}{|\mathbf{X}|^2}. \tag{7}$$

The integral value

$$T = \int_{\text{path } \mathbf{m}(u)} \tau(u)\, du \tag{8}$$

quantifies the relative tilt of the surface between start and endpoint of the path. Note that $\tau(u) > 0$ indicates right-handed surface torsion, $\tau(u) < 0$ indicates left-handed surface torsion, and $\tau(u) = 0$ indicates that the surface at this point is developable and rulings are parallel. ("Developable" means that a sheet of paper could be bended to exactly match the surface; the surface could be "flattened.") In summary, the definition of the intrahelical surface depends on a reasonable selection of rulings, $\mathbf{X}$, and a path, $\mathbf{m}$, which are both arbitrary to some extent. Despite these shortcomings in a strict mathematical sense, the concept of torsion, as introduced here, mirrors some essential features in describing the interplay between the shape of α-helices and their relative orientation in forming the MHC binding cleft.

2.6.3. Curvature of Helices. Derivatives of each polynomial helix model can be obtained analytically for each value of the parameter u, yielding the vectors

$$\mathbf{c}'(u) = \frac{d\mathbf{c}}{du}, \qquad \mathbf{c}''(u) = \frac{d^2\mathbf{c}}{du^2}. \tag{9}$$

Curvature κ is a scalar quantity being per definition positive in Euclidian 3D space and is obtained via [29]:

$$\kappa(u) = \frac{\left|\mathbf{c}'(u) \times \mathbf{c}''(u)\right|}{\left|\mathbf{c}'(u)\right|^3}. \tag{10}$$

2.6.4. Construction of a Curved Helix Model and Helix Hinge Model. So far we have described how an α-helical axis is reconstructed by our newly proposed fragment-fitting method and fitted by a polynomial of a certain degree. From this polynomial, we calculate the curvature integral as described in Section 2.6.3. We wanted to find an optimal degree for these polynomials in order to retrieve conformational deformations with minimal errors. To do so we modelled two different motions such that MHC α-helices could perform: a bending motion and a hinge motion (see Figure 6).

To model a bending motion we created a curved helix backbone model with known curvature, as described in detail in Appendix A. Following polynomial representation of the backbone, the curvature integral can be obtained analytically, yielding the reference, $\int \kappa_{\text{bending}}^{\text{reference}} du$ (see (A.2) in the appendix).

To test our method, the curved helix model was subjected to the fragment-fitting method and the resulting helix axis was fitted by a polynomial in a least-squares sense as described in Sections 2.4 and 2.5. The polynomial was evaluated at 100 equidistant points and curvature and the curvature integral were calculated numerically, yielding $\int \kappa_{\text{bending}}^{\text{detector}} du$ (see (10)). As a quality criterion for regaining the correct curvature integral we used the relative error

$$\eta_{\text{bending}} = \frac{\int \kappa_{\text{bending}}^{\text{detector}} du}{\int \kappa_{\text{bending}}^{\text{reference}} du} - 1 \tag{11}$$

and evaluated it for polynomial degrees 1 to 8. The results are depicted in Figure 6(b).

To model a hinge motion, we constructed an ideal, linear α-helix comprising 31 C_α atoms. Subsequently, the helix was split into two parts: one C_α atom was selected and the remaining part of the helix rotated around the selected C_α atom to model a hinge motion (see Figure 6(c)). The position to split the helix (number of C_α selected) was varied from C_α atom 5 to 15. The aim was to examine α-helix hinges with two legs unequal in length. From this series of α-helical models (different positions of the hinge and varying hinge angles, α_{hinge}) we then reconstructed the α-helical axis using the fragment-fitting method. A polynomial of kth order was fitted to the traced α-helical axis in a least-squares sense and the curvature integral calculated numerically, yielding $\int \kappa_{\text{hinge}}^{\text{detector}} du$. The relative error

$$\eta_{\text{hinge}} = \frac{\int \kappa_{\text{hinge}}^{\text{detector}} du}{\alpha_{\text{hinge}}} - 1 \tag{12}$$

was obtained for polynomials of degrees 1 to 8, with $0 \leq \alpha_{\text{hinge}} \leq \pi/2$ being the angle between the two parts of the helix.

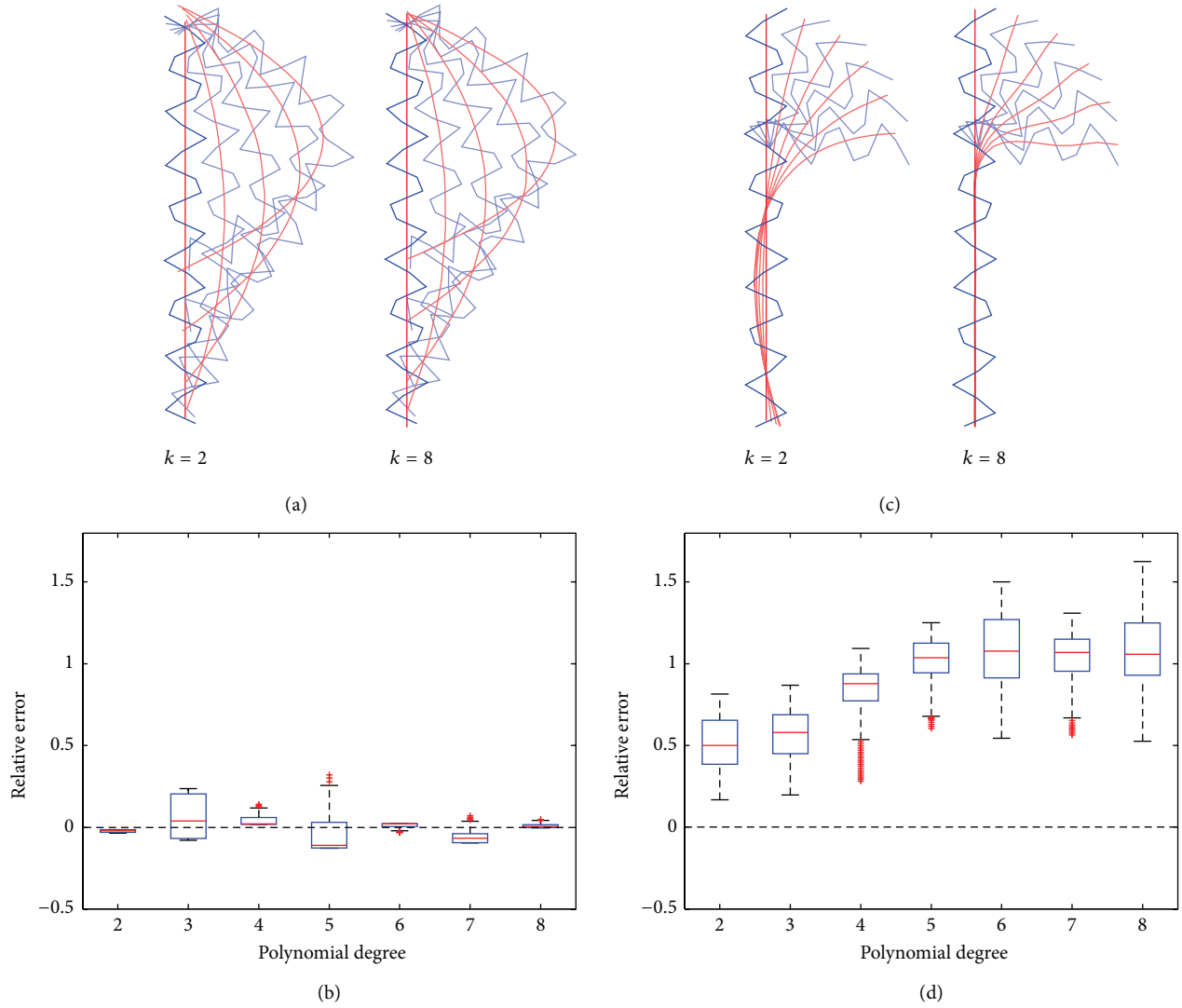


FIGURE 6: Capturing helix motions with polynomials of different degrees. Panel (a) is an illustration of a helix bending movement. We created a model of an ideal linear helix (blue) comprising only C_α atoms whose axis is gradually bent by a mathematically well-defined function. From this function we can easily derive the curvature ($\int \kappa_{\text{bending}}^{\text{reference}} du$) and compare it to the curvature we measure (detect) from the polynomial fitted to the helical axis ($\int \kappa_{\text{bending}}^{\text{detector}} du$) that was calculated by the fragment-fitting method. From this comparison we derive the relative error; see (11). Panel (b) shows the relative errors of bending motion for polynomial degrees 1 to 8. Panel (c) is an illustration of a helix hinge movement. We created a model of an ideal linear helix (blue) comprising only C_α atoms and split it into two parts. One part was 10-atom long and the other part was 20-atom long. Then one part was rotated around a pivotal point as to simulate a hinge movement. The curvature integral of the helical axis was compared to the hinge angle by calculating the relative error. Panel (d) shows the relative errors of hinge motion for polynomials degrees 1 to 8. Polynomials of second degree were found to reproduce the bending and hinge angles with minimal relative errors.

3. Results

3.1. Finding the Optimal Polynomial Degree for Monitoring Helix Motions. To monitor deformations of α-helices in MD simulations we propose an approximation of the α-helical axis using a fragment-fitting method (Section 2.4), fitting the resulting axis by a polynomial in least-squares sense (Section 2.5), and derive geometric quantities from it (Section 2.6). To find an optimal degree of polynomials we applied our method by applying it to modelled α-helical bending and hinge motions (Figures 6(a) and 6(b)) and evaluated the polynomial degree that retrieves a certain α-helical motion with minimal relative error. For the hinge motion we noticed that relative errors increase with polynomials of higher order. For the bending motion, choosing sixth- or eighth-degree polynomials increases the degree of freedom while at the same time failing to add adequate improvement in relative error. Generally, relative errors in the detection of hinge motions are larger than those for bending motions. The second-order polynomials reproduced the measured quantities of bending and hinge motions with low relative error and hence were adopted for monitoring

Table 2: Median RMSD values of case studies.

	Orange Median RMSD	Red Median RMSD	Red versus orange Median RMSD
B4402 G-ALPHA2	0.0870 nm	0.0682 nm	0.1350 nm
B4403 G-ALPHA1	0.0789 nm	0.0671 nm	0.1170 nm
B4405 G-ALPHA1	0.0592 nm	0.0556 nm	0.0794 nm

We compared helix conformations between different phases of the simulation, that is, before and after an inflection point or continous decrease/increase in the curvature integral (see Figure 7, upper row). Median values of the RMSD distribution are shown in this table.

MHC α-helices in all subsequent analyses. We admit that second-order polynomials may fail to model α-helical axes in full detail, but it seems appropriate for capturing trends in helical motions and shape during an MD simulation with minimal relative error.

We applied the geometric analysis to model MHC α-helices of TCR/pMHC complexes B4402, B4403, and B4405. By inspecting the curvature integral for MHC α-helices, G-ALPHA1 and G-ALPHA2, three interesting cases could be identified showing changes along the 250 ns MD simulation (Figure 7, panels in upper row). Phases during which the curvature integral changes either abruptly ((b) and (c)) or gradually (a) are highlighted in orange and red. Helix conformations ("bundles") corresponding to these phases are shown in panels in the middle row. To check if changes in the curvature integral actually indicate a conformational change we calculated RMSD matrices of conformations within and between orange and red helix bundles (panels in lower row). In all three cases, RMSD between phases is distinctly larger (shifted to the right) as compared to RMSD within phases. A shift towards higher RMSD values indicates a conformational change and is seen in all three cases (see Table 2 for median RMSD values). These case studies demonstrate that changes in conformation of α-helices during MD simulations can be monitored using second-degree polynomials fitted to helical axes.

3.2. Geometric Quantities Characterizing the Shape of MHC α-Helices. The MHC peptide-binding groove comprises two α-helices that interact with the TCR. Using second-order polynomials, we computed (i) the integral of the curvature of individual MHC α-helices, (ii) the area of interhelical surface, and (iii) the surface torsion along the imaginary centre line derived from both polynomials that model MHC α-helices for single time steps in MD simulations. Items (ii) and (iii) are geometric properties of the ruled surface. They are used to quantify the geometric relation between the two helices, for example, their relative orientation, and thus capture important aspects of the geometry of the peptide-binding groove, as illustrated in Figure 5.

Curvature is a local feature of a curve. Considering the curvature integral we obtain a measure of the overall bending of the whole curve [30]. Curvatures of polynomials modelling single helices were integrated and monitored over time as seen in Figure 8. G-ALPHA1 of B4403 and B4405 each undergoes abrupt fluctuations in helical conformation, reflected in the curvature integral, but is stable in the phases in-between. The curvature integral for G-ALPHA1 of B4402 shows a gradually increasing trend. We notice that the curvature integral for G-ALPHA2 is generally higher than that for G-ALPHA1, reflecting the kink near its N-terminal end. The two helical parts of G-ALPHA2 form a hinge. G-ALPHA2 of B4403 and B4405 shows only minor fluctuations and no abrupt changes in the curvature integral indicating that the hinge angle stays stable. For B4402, the hinge angle decreases in the first half of the MD simulation and remains stable thereafter.

The area of interhelical surface depends on the relative location of both helices. It changes, as the helices drift apart or elongate. It also changes when both helices bend in opposite directions as these amount to a distension of the surface. Whenever helices bend in similar ways in the same direction, interhelical area will be relatively unaffected. Depending on the complexity in shape of the helical axis, second-order polynomials might not adapt to the axis' path accurately. Figure 9 shows an increasing trend of interhelical area for complexes B4403 and B4405, while a declining trend of interhelical area is seen for B4402. These changes occur as the helices of the peptide-binding groove move closer together or further apart, respectively. This is also reflected in the distance of α-helical centres of masses (data not shown). However, inspecting Figure 5 clearly demonstrates that the actual shape of an interhelical surface cannot be characterized by a single quantity such as the area of interhelical surface.

A more elaborate descriptor of the shape of a surface is the surface torsion. This parameter describes the change in angle between subsequent tangent planes along a given path, in our case the centre line. High values of surface torsion, regardless of being positive or negative, describe a rapid change in the angle. A pair of helices, each being somehow deformed and both being in varying positions to each other, gives rise to an interhelical surface with a plethora of possible shapes. For describing these shapes geometrically, surface torsion is an important concept, lending itself to quantify the twist in the surface along a prescribed path.

For B4402, surface torsion along the centre line (see Figure 5(a)) is left-handed most of the time, reaching a minimum (ca. −2, data not shown) at half of its length. Positive values of surface torsion (ca. 0.2, data not shown) are seen only in rare cases and near both surface termini. The ruled surface of the peptide-binding cleft at the N-terminus of G-ALPHA1 appears to be more stable for B4405 and B4403 than for B4402. Similar trends are seen in the surface torsion integral, see Figure 10. It is stable for B4405, but trends are observed for the other two complexes: B4402 shows a drop in the surface torsion integral during the first 60 ns followed by a gradual rise. B4403 is rather stable during the first 60–70 ns and then shows a gradual decrease in surface torsion.

We also looked at the shape complementarity (S_C) (Lawrence and Colman introduced a method to measure how well the surfaces of two proteins at their interface match [31]. The parameter output by this method is called

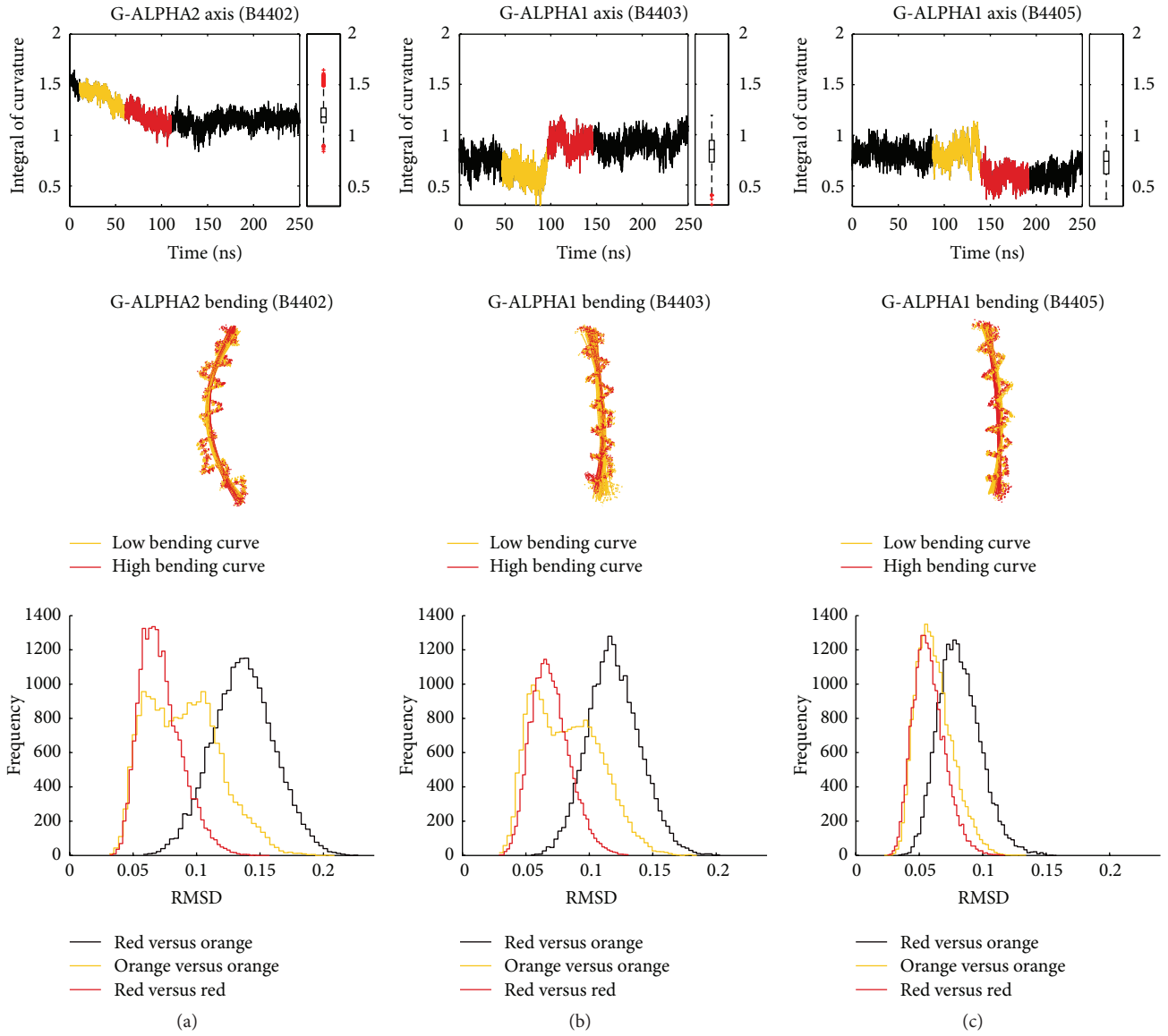


FIGURE 7: Curvature integral and helix conformations. Three case studies showing the time course of the curvature integral for G-ALPHA1 (b and c) and G-ALPHA2 (a). Upper row shows curvature integral along the 250 ns MD simulation. Phases of increasing or decreasing trends are highlighted in orange and red. Middle row shows 3D conformations of MHC α-helices (C_α atoms) and polynomials derived by fragment-fitting. Colours correspond to the highlighted phases in the upper row. We see that the red bundle of conformations differs from the orange bundle, especially near the ends. Lower row shows that RMSD matrices between helix conformations were calculated and frequencies of RMSD values plotted. Red and orange lines represent frequency distribution of RMSD values between configurations within red and orange phases, respectively. Black lines represent RMSD distributions between configurations in the red and configurations in the orange phases. The difference between lower RMSD within phases and higher RMSD between phases confirms a conformational change between these phases.

shape complementarity (S_C).) of the protein-protein interface surface and backbone C_α RMSD. We found S_C to be stable for all three TCR/pMHC systems (see Figure 14). As S_C is unaffected by deformations of the binding cleft (as reflected by surface torsion, see above) we conclude that TCRs follow the conformational changes of the MHC surface. RMSDs of B4402 and B4403 reach a low plateau after a few nanoseconds and remain stable thereafter. RMSD of B4405 shows an increasing trend over 250 ns simulation time, rising from approximately 0.3 nm to 0.6 nm; see Figure 15.

4. Discussion

The mechanism of TCR activation is still controversially debated and several models have been proposed [32–38]. A recent work of Dustin and Depoil [34] summarizes new insights into the function of the T cell synapse. The authors grouped T cell synapse into three interactive layers including interactions of receptors, a signaling layer, and a cytoskeleton layer, all contributing to TCR activation, regulation, and fine-tuning of signaling and responses. Conformational changes

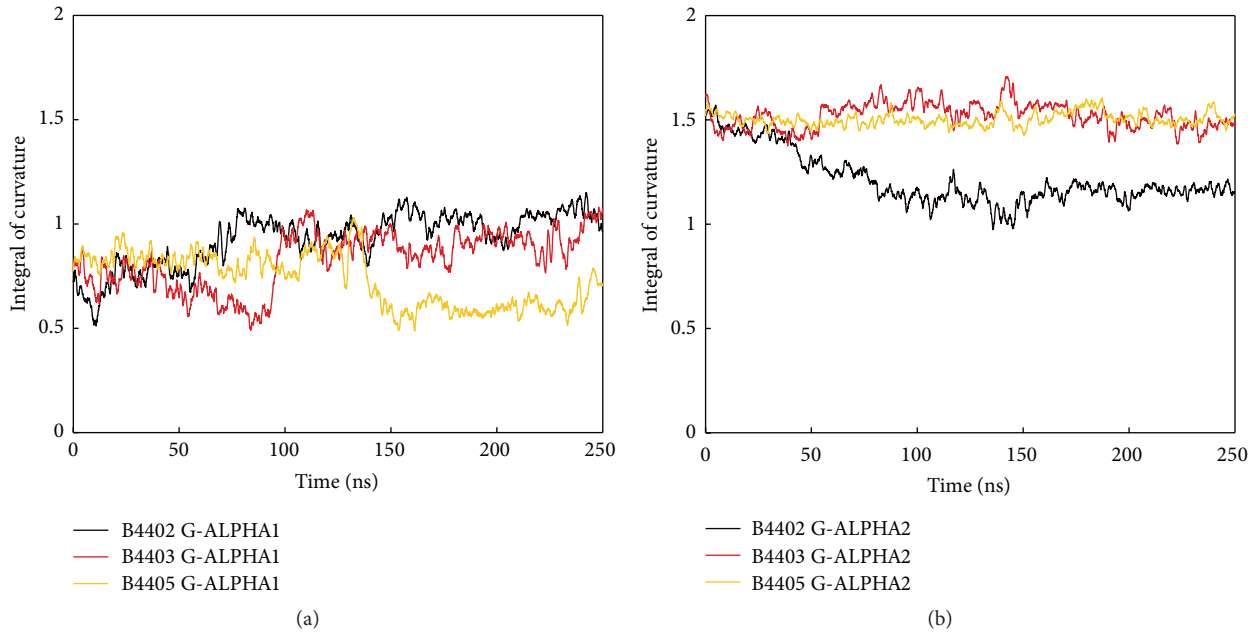


FIGURE 8: Curvature integral of MHC α-helices. Moving average of the curvature integral of MHC helices G-ALPHA1 (a) and G-ALPHA2 (b) along 250 ns MD simulations of TCR/pMHC systems B4402, B4403, and B4405. Curvature of G-ALPHA2 is higher compared to G-ALPHA1 due to its kink. G-ALPHA1 shows greater fluctuations but is comparable between all three TCR/pMHC complexes. G-ALPHA2 of B4402 shows a decreasing trend, consistently reflected in the area of the ruled surface spanned between both MHC α-helices (see Figure 9).

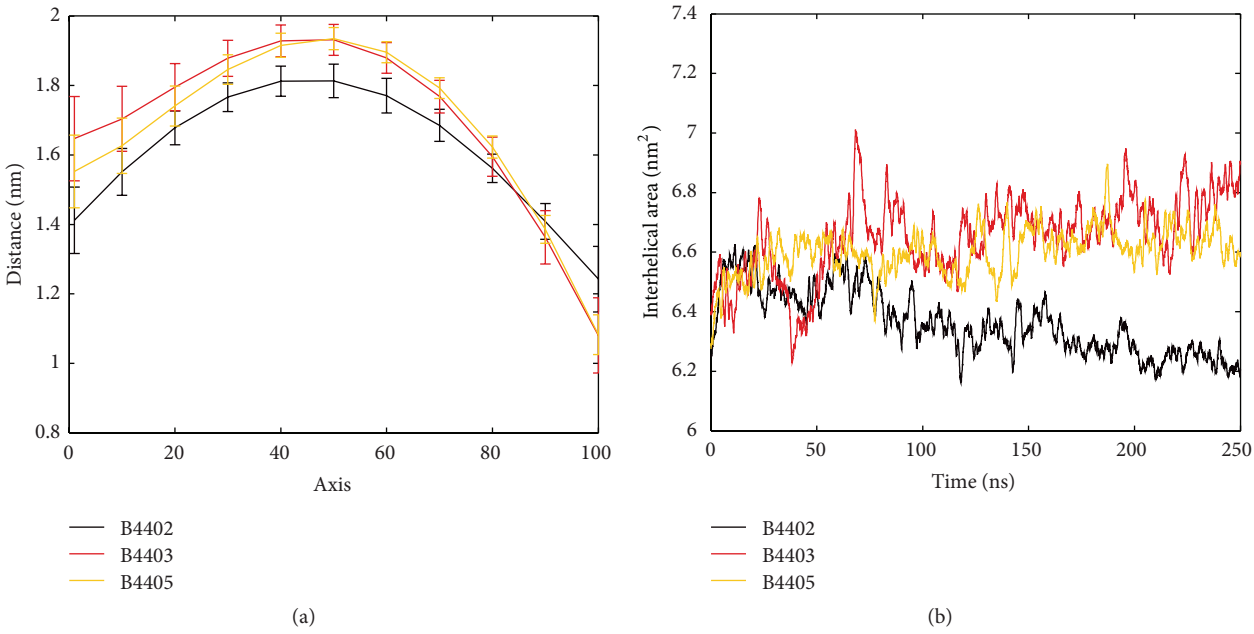


FIGURE 9: Interhelical distance and area of interhelical surface of MHC α-helices. (a) Mean distances between the two MHC α-helices as measured at 11 different points along the helices. x-axis describes the running parameter of the helices with each helical axis divided into 100 equidistant points. The orientation of the running parameters of both helices is from N-terminus to C-terminus of G-ALPHA1. Distances are measured between corresponding points on each helical axis of G-ALPHA1 and G-ALPHA2. The standard deviation of the mean is shown in the error bars. This distance plot describes the shape and size of the peptide-binding pocket. B4403 and B4405 show a very similar pocket shape. (b) The two MHC α-helices span a ruled surface. Moving average of interhelical area along the MD simulation is shown. The magnitudes of interhelical area of B4403 (nonimmunogenic) and B4405 (immunogenic) are similar and slightly increasing, while B4402 shows a declining trend. The curvature integral (Figure 8) for individual helices shows a concomitant bending and relaxing, explaining the shrinkage of interhelical area.

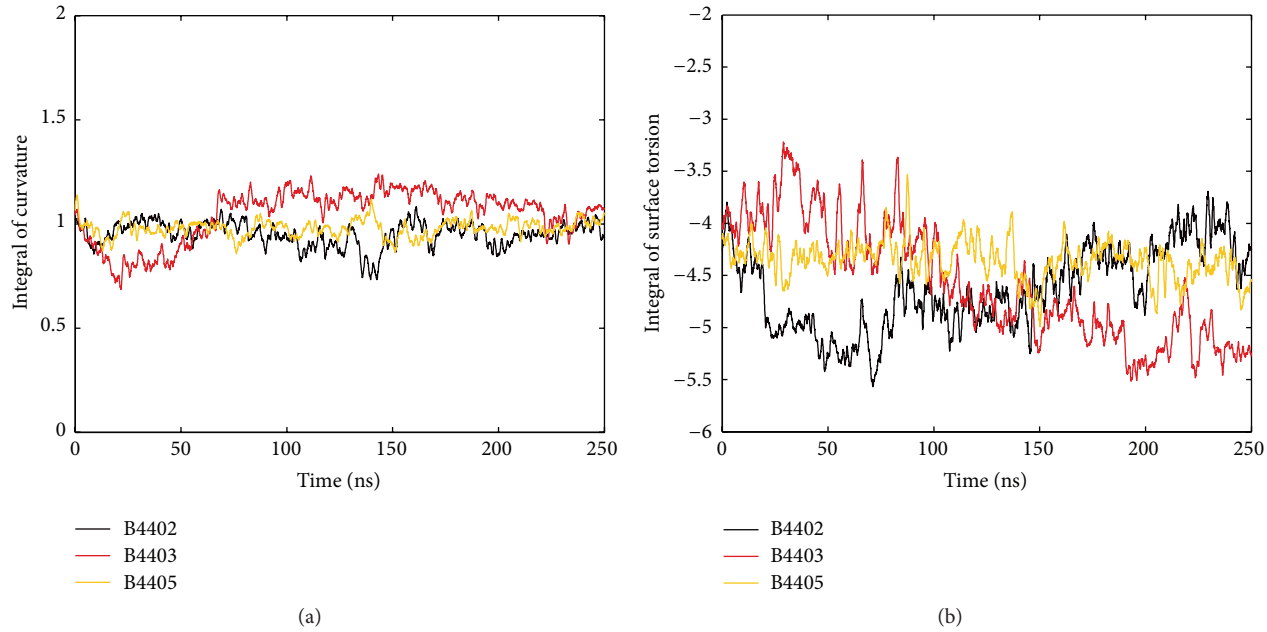


Figure 10: Curvature integral and surface torsion integral of the centre line. (a) The centre line is an imaginary line combining polynomials fitted to axis of G-ALPHA1 and G-ALPHA2 (see Figure 5, green line). The curvature integral of the centre line is stable along the MD simulation for all three TCR/pMHC complexes. (b) The surface torsion integral along the centre line between two polynomials that approximate MHC G-ALPHA1 and G-ALPHA2 is stable for B4405. The other two complexes differ: B4402 shows a drop in the surface torsion integral during the first 60 ns and rises afterwards. B4403 shows a generally decreasing trend.

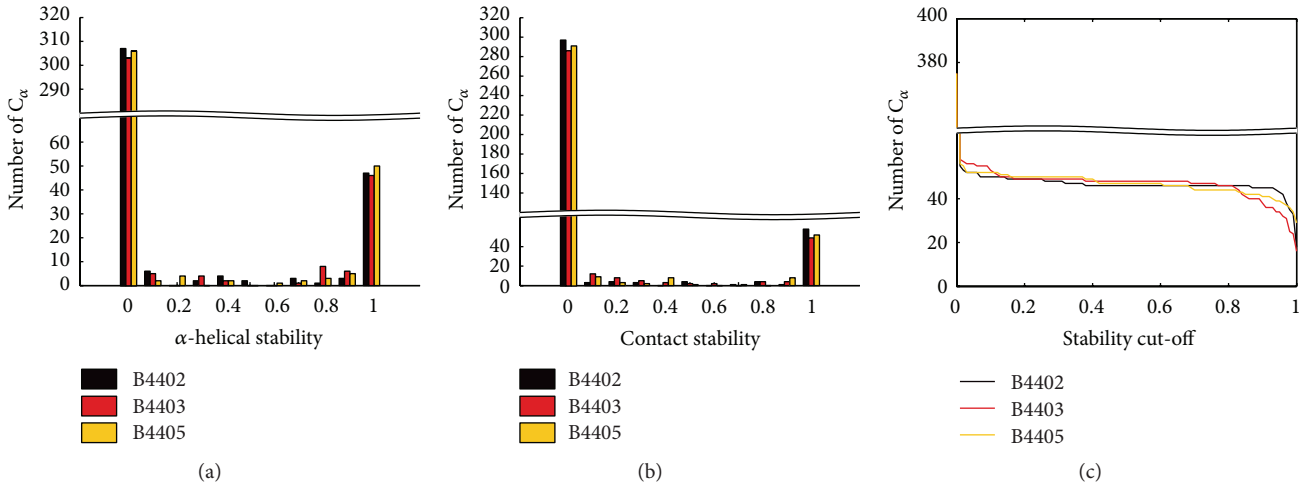


Figure 11: Stability of MHC α-helices and protein-protein contacts. (a) Histogram of amino acids that form α-helices in the MHC and β-2-microglobulin protein complex. α-helical stability of 0 means that a given C_α atom is never part of an α-helix during the MD simulation. On the contrary, an α-helical stability of 1 means that this C_α atom is part of an α-helix in every time step of the MD simulation and thus is part of a very stable α-helix. The histogram shows a distinctly bimodal distribution. (b) Histogram of C_α atoms forming stable close contacts (atoms being less than 1.4 nm apart) at the protein-protein interface. The distribution is also distinctly bimodal. Contact stability of 0 means that a C_α atom never forms a close contact during the MD simulation. A contact stability of 1 means that this C_α atom forms very stable contacts throughout the MD simulation. (c) The number of stable residues on y-axis is calculated by intersecting both sets of stable helix C_α atoms and stable close contacts C_α atoms. In Section 2.3, we claim that, due to the distinctly bimodal distributions, neither stable α-helices nor the number of close contacts is insensitive to the choice of the cut-off. The resulting number of stable residues will roughly stay constant for a wide range of cut-off values (from 0.2 to 0.8), therefore justifying our choice of 0.5 as the stability cut-off.

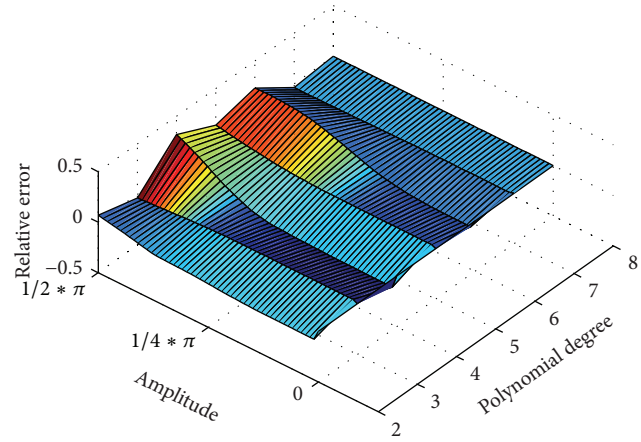


Figure 12: Curvature integral as a function of helical bending. In order to model a bending motion, an ideal linear helix comprising 31 C_α atoms is distorted by bending its imaginary, linear axis along a cosine function. The curvature of the imaginary axis is known and its integral serves as the reference for the amount of bending. We compare this reference curvature integral with that derived via our fragment-fitting method by calculating the relative error. An ideal method would show a very close to linear correlation between the reference and the measured value. Polynomials of third and fifth degree show the highest relative error, especially for large magnitudes of bending. Sixth-order polynomials or polynomials of higher order look quite promising regarding relative error. Polynomials of higher order were ruled out because of overfitting and the fact that spurious terminal oscillations might occur. Second-order polynomials show a well behaved, close to linear dependence and were therefore adopted to model α-helices of MHCs.

of the TCR complex have been demonstrated to be relevant for signaling of the TCR [39]. Association of CD3 proteins to the TCR/pMHC complex is necessary to transmit the activation signal to intracellular signaling molecules [40]. Evidence suggests that TCR conformational changes are required for full activation, but there are certain signaling pathways that can also be activated in the absence of conformational changes [41]. The three TCR/pMHC complexes analysed in this work differ only by one or two amino acids in the MHC molecule. HLA-B*44:05 (B4405) and HLA-B*44:02 (B4402) MHC types trigger LC13 TCR activation in the presence of the ABCD3 self-peptide. Surprisingly HLA-B*44:03 (B4403) does not trigger TCR activation.

To characterise the dynamics of the MHC antigen binding cleft, we applied (similar to the methods reported by Christopher et al. [16]) a fragment-fitting method to model stable α-helical regions that are in close proximity ($\leq$1.4 nm) to the TCR and monitored their geometric parameters. However, it is known that geometric quantities derived from polynomial approximations may vary substantially depending on the polynomial degree chosen [10]. To select an appropriate polynomial degree, we tested the ability of polynomials with different degrees to retrieve predefined parameters of helical motions and deformation. In a simplified model, we tested the ability of the fragment-fitting method to reproduce the curvature integral of helical bending and hinge motions. We found that second-order polynomials are best suited to model these α-helical motions with low relative error. The curvature integral derived from polynomials of individual α-helices can be related to conformational changes of α-helices. Between the two MHC α-helices a ruled surface can be spanned, of which we computed the area as an estimate for the size of the peptide-binding cleft. We also calculated the surface torsion along an imaginary centre line characterizing the orientation of both α-helices relative to each other. We applied this method to MD simulations of three TCR/pMHC complexes. However, we were not able to find correlations between immunogenicity and certain patterns in the α-helical movements.

4.1. Limitations. A limitation of the current analysis is that the ruled surface between the MHC α-helices that we use to model the MHC surface presented to the TCR does not consider the shape and dynamics of the peptide that lies between the two helices. We cannot assume that phase space has been sampled comprehensibly for these large molecules. Stepwise fluctuations in measured variables are visible (see Figure 7(b), upper row, and Figure 15). Also, even with highly optimized simulation performance of 15 ns/day on 1024 computing cores of the IBM BlueGene, statistics to discriminate between different simulated systems is not feasible. Enhanced sampling techniques and adequate collective variables might be useful to identify adequate collective variables for such systems. Interpretation of single simulations should therefore be done with caution. Signaling may involve a series of other proteins of the immunological synapse and interactions that are not considered in these simplified TCR/pMHC models. Interactions between TCR and pMHC take place between two cells that are in close contact to each other. It has been shown that plasma membrane lipids affect the activity of signaling networks [42] and some models of TCR/pMHC interaction propose involvement of the plasma membrane [33].

4.2. Conclusion. In this work, changes of MHC shape and their dynamics were quantified. We applied the quantification method to three large TCR/pMHC complexes, due to their size being accessible by MD simulation studies only since recently. We saw that MHC α-helices undergo rapid changes

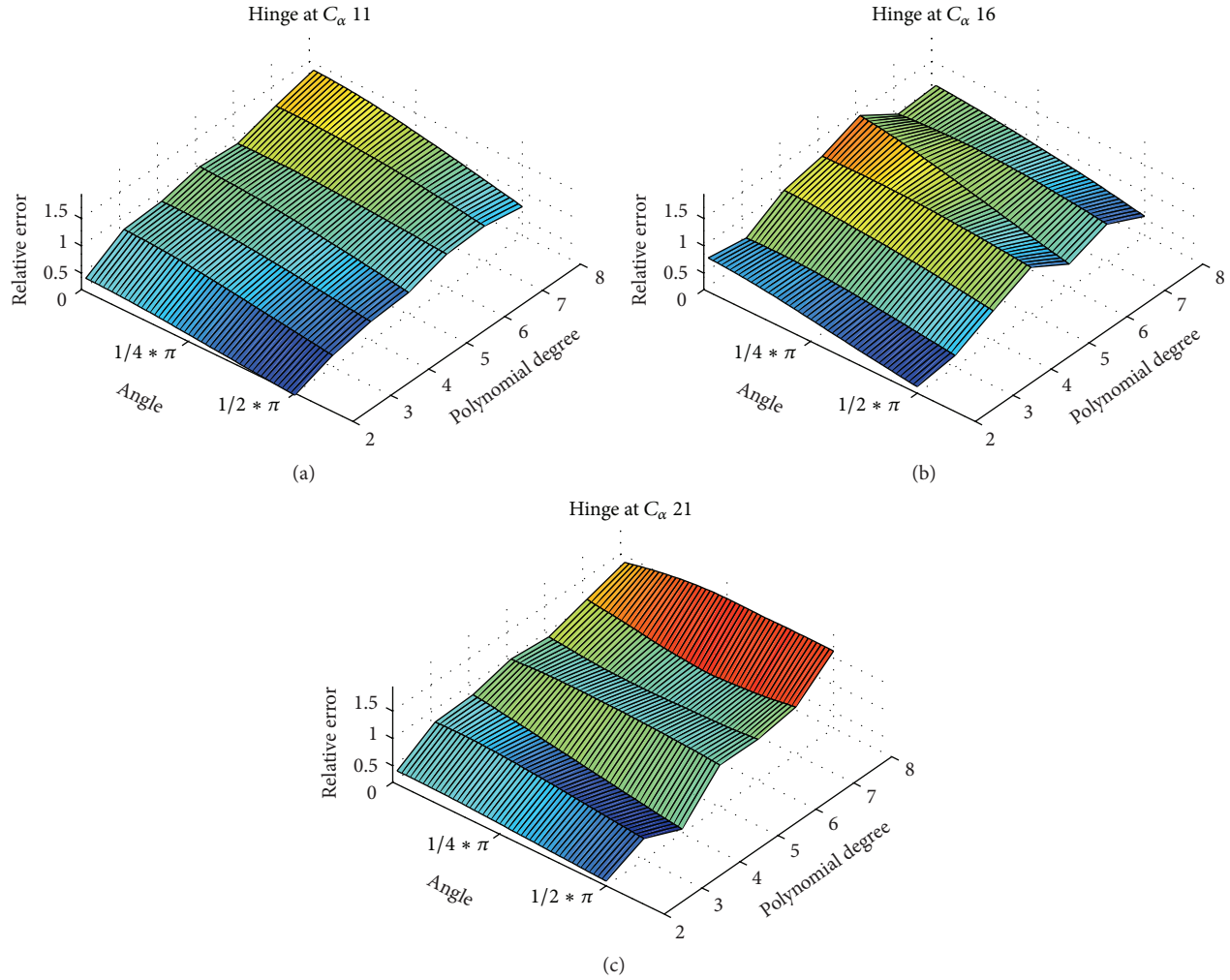


FIGURE 13: Curvature integral as a function of hinge movement. The relative error in retrieving the correct hinge angle is plotted against helix hinge angle and polynomial degree. To model the hinge motion, a kink of varying angle was introduced to an ideal linear helix comprising only C_α atoms (31 atoms). Images (a), (b), and (c) show the same data for different positions of the kink in the helix. We refer to the kink angle as the signal we want to measure. We compared the signal to the curvature integral of the polynomial fitted to the helical axis by calculating the relative error. An ideal method would show a linear correlation between signal and the measured value. We see that polynomials of higher order show a higher relative error and overestimate the magnitude of the kink. We also see that the position of the kink modulates the relative error. Second-order polynomials have a nearly linear dependency and were therefore adopted to model α-helices of MHCs.

in conformation by either bending motion or hinge motion. Surface torsion used for characterizing the MHC surface presented to the TCR is stable in B4405, which is the most immunogenic complex. We speculate that rapid changes in helical conformation are part of the intrinsic dynamics of MHCs when engaging with TCRs.

Though we were not able to find a clear correlation between immunogenicity and certain patterns in the α-helical movements, we could demonstrate that single amino acid polymorphisms in the MHC seem to have a subtle influence on the helixes' shape dynamics and that it would be interesting to apply the same method in the case of peptide polymorphism.

In summary, the present work demonstrates the feasibility and reliability of deriving shape parameters from simulation data. Next, the influence of the detected small conformational changes on the microscopic dynamics will be investigated to clarify their relation to the biological functions of the complexes of interest. Conclusions regarding functional differences between TCR/pMHC systems characterized by a single-residue polymorphism certainly require advanced sampling techniques in order to sample the conformational phase space appropriately for molecules of this size. Future studies might investigate if the small conformational changes in MHC α-helices transmit forces to the TCR.

Appendices

A. Construction of a Curved Helix Model

The backbone is modelled along a cosine curve. We compute equidistant points along the curve and create C_α atoms

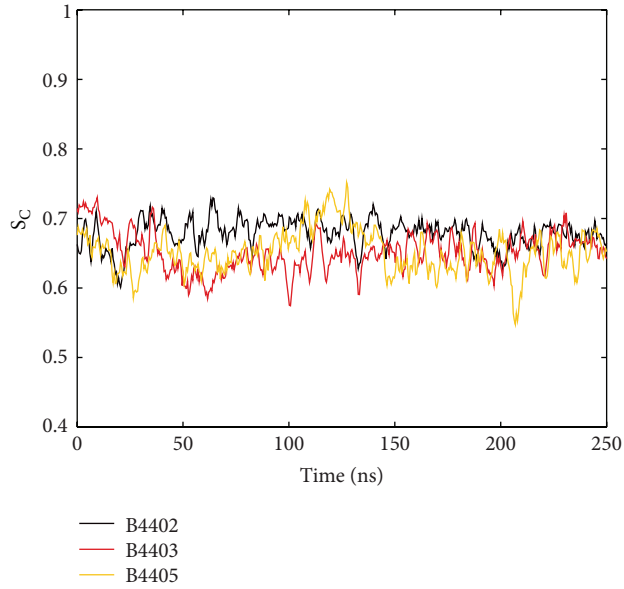


FIGURE 14: Shape complementarity of TCR/pMHC interface. Lawrence and Colman [31] introduced shape complementarity statistics comparing the surface normal alignment on dots from molecular protein-protein surfaces generated according to Connolly [44]. S_C is a measure of how good two protein surfaces fit together. It assumes values between 0 and 1, with 1 indicating a perfect fit. S_C is stable and similar for all three TCR/pMHC complexes along 250 ns MD simulations.

with constant normal distances to the backbone. Angles between successive C_α atoms were set to $\beta = 100^\circ$. Formally, coordinates along the axis of the helix are given by

$$\mathbf{x} = \begin{cases} x = t \\ y = a\cos t \\ z = 0, \end{cases} \tag{A.1}$$

where a represents the maximum elongation (amplitude) of the curved helix, compared to $a = 0$, corresponding to a straight model. Increasing a in a stepwise fashion generates models of increasing curvature.

For the curvature we obtain

$$\kappa(t) = \frac{a\sin t}{\left(1 + a^2\sin^2 t\right)^{3/2}}, \qquad \kappa(0) = a. \tag{A.2}$$

In order to keep the length of backbones constant, the limits for parameter t have to be adjusted appropriately; $t \in [-b, b]$. In fact $b = b(a)$ is chosen to make the arc length equal to 1:

$$\int_{-b}^{b} \sqrt{1 + a^2\sin^2 t}\, dt = 1. \tag{A.3}$$

By varying a within $0 \le a \le \pi/2$ different models were created. Curvature decreases with decreasing a; in particular $a = 0$ corresponds to a straight line without curvature. In order to find the appropriate values for $b = b(a)$ with $a > 0$, an elliptical integral (see (A.3)) has to be solved. To find N equidistant points along the backbone $\mathbf{x}_n = (x_n, y_n, z_n) = (x(t_n), y(t_n), z(t_n))$, for $n = 1, \ldots, N$, we proceed, similar to the normalization of the arc length, via numerically solving the elliptical integral:

$$\int_{-b}^{b} \sqrt{1 + a^2\sin^2 t}\, dt = \frac{n-1}{N-1}. \tag{A.4}$$

In the next step, C_α atoms are positioned in distance r to the backbone and successively rotated by $\beta = 100^\circ$. The tangent vector of unit length at position $\mathbf{x}_n$ is given by

$$\mathbf{r}_n = \frac{1}{\sqrt{1 + a^2\sin^2 t_n}} \begin{bmatrix} 1 \\ -a\sin t_n \\ 0 \end{bmatrix}. \tag{A.5}$$

Next, the radius r of α-helical turns around the axis is set in proper relation 0.23 nm/0.15 nm to the total length of the helix:

$$r = \frac{0.23}{0.15}\frac{1}{N-1} \tag{A.6}$$

and the vector $\mathbf{s}_n$ to be rotated is given by

$$\mathbf{s}_n = \frac{r}{\sqrt{1 + a^2\sin^2 t_n}} \begin{bmatrix} -a\sin t_n \\ -1 \\ 0 \end{bmatrix}. \tag{A.7}$$

The point $\mathbf{x}_n + \mathbf{s}_n$ is rotated around the axis $\mathbf{r}_n$ by angles $\beta_n = \beta_0 + (n-1)\beta$ to create coordinates of successive C_α atoms ($n = 1, \ldots, N$). Finally, all coordinates are scaled via $\mathbf{x} \to \mathbf{x}\cdot N\cdot 0.15$ nm to arrive at proper dimensions. Thus, we obtain a chain of C_α atoms rotated in a clockwise manner to represent

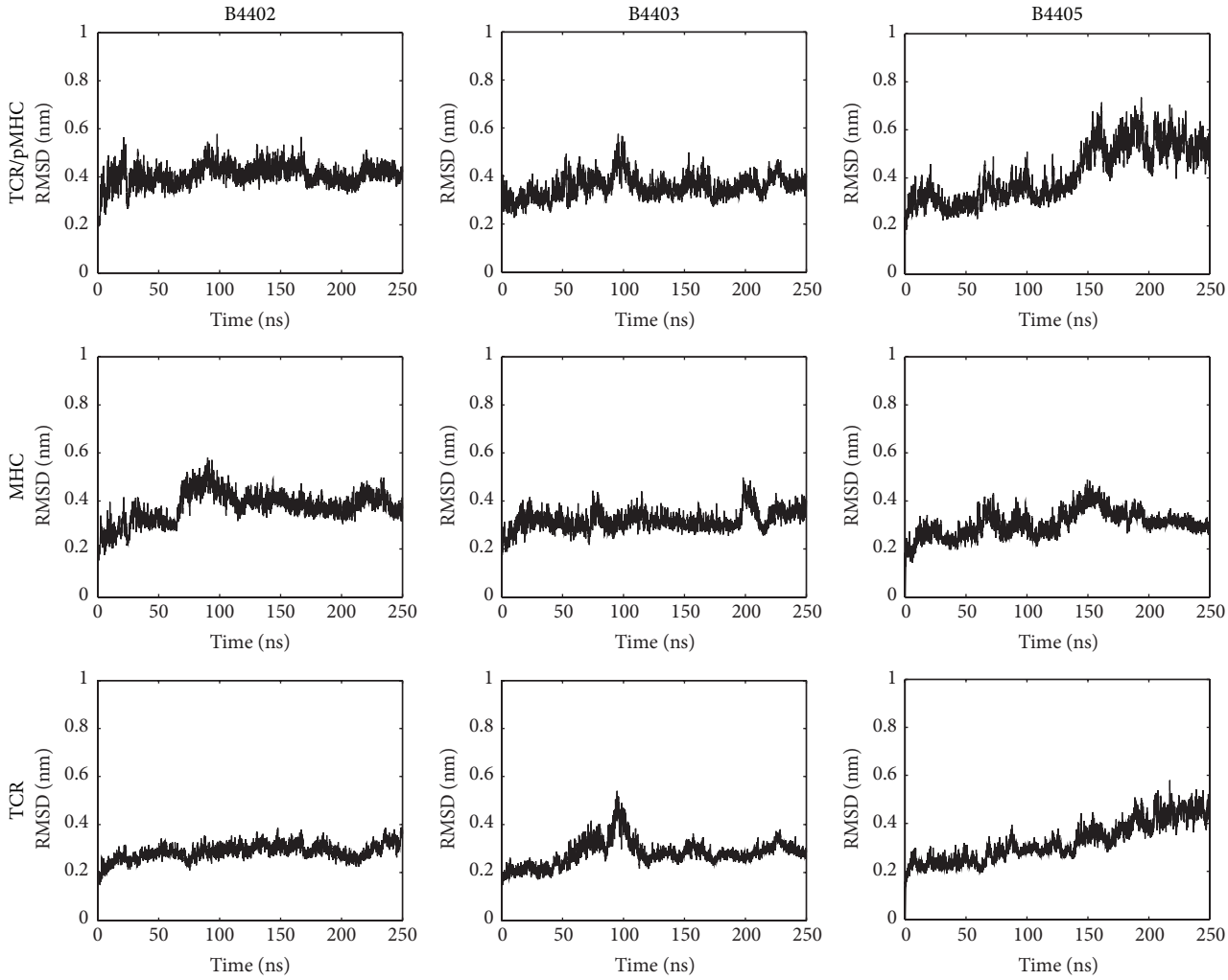


Figure 15: Root mean square deviations. Root mean square deviations of TCR/pMHC systems B4402, B4403, and B4405. Superposition of successive frames was done with respect to protein C_α backbone of the first frame of the simulation (nonprogressive fitting) and RMSD was calculated between protein C_α backbones. Row TCR/pMHC shows that the whole protein system, TCR/pMHC, was fitted to itself and RMSD calculated for the whole protein. Row MHC shows that TCR was fitted to itself and RMSD calculated for TCR. Row TCR shows that MHC was fitted to itself and RMSD calculated for MHC. Results for B4405 indicate that 250 ns of simulation time does not suffice to sample the whole phase space, which is a common finding for such large proteins. RMSD time courses for B4402 and B4403 do not explicitly indicate nonstationary behaviour. They indeed show slower and smaller growth of RMSD over time than does B4405, also indicating their stability as a molecule (despite two point mutations introduced). As noted before, the present work intends to delineate techniques for modelling geometries of MHC components and does not aim at statistical comparisons between the motions of different HLA alleles. Ergodicity is hence not a vital issue; see the discussion in Schreiner et al. [45].

a curved right-handed helix along a predefined, cosinusodial backbone.

The curved helix model has a few limitations. (i) Due to the curvature of the cosine arc and the construction of C_α atoms with normal distances, the typical distance between successive C_α atoms is not maintained. (ii) We refrained from constructing the full backbone also including nitrogen, oxygen, and hydrogen atoms and restricted the model to C_α carbon atoms only. This seems justifiable, however, since the fragment-fitting algorithm also takes into account C_α atoms only as does the DSSP algorithm and the calculation of close contacts. Therefore, modelling the full backbone would not add more information to the model. (iii) The cosine arc is planar and hence cannot incorporate any torsion within the curved helix model.

B. Additional Data

See Figures 11–16.

Conflict of Interests

Reiner Ribarics is an employee and stockholder of Gilead Sciences. The authors declare that there is no conflict of interests regarding the publication of this paper.

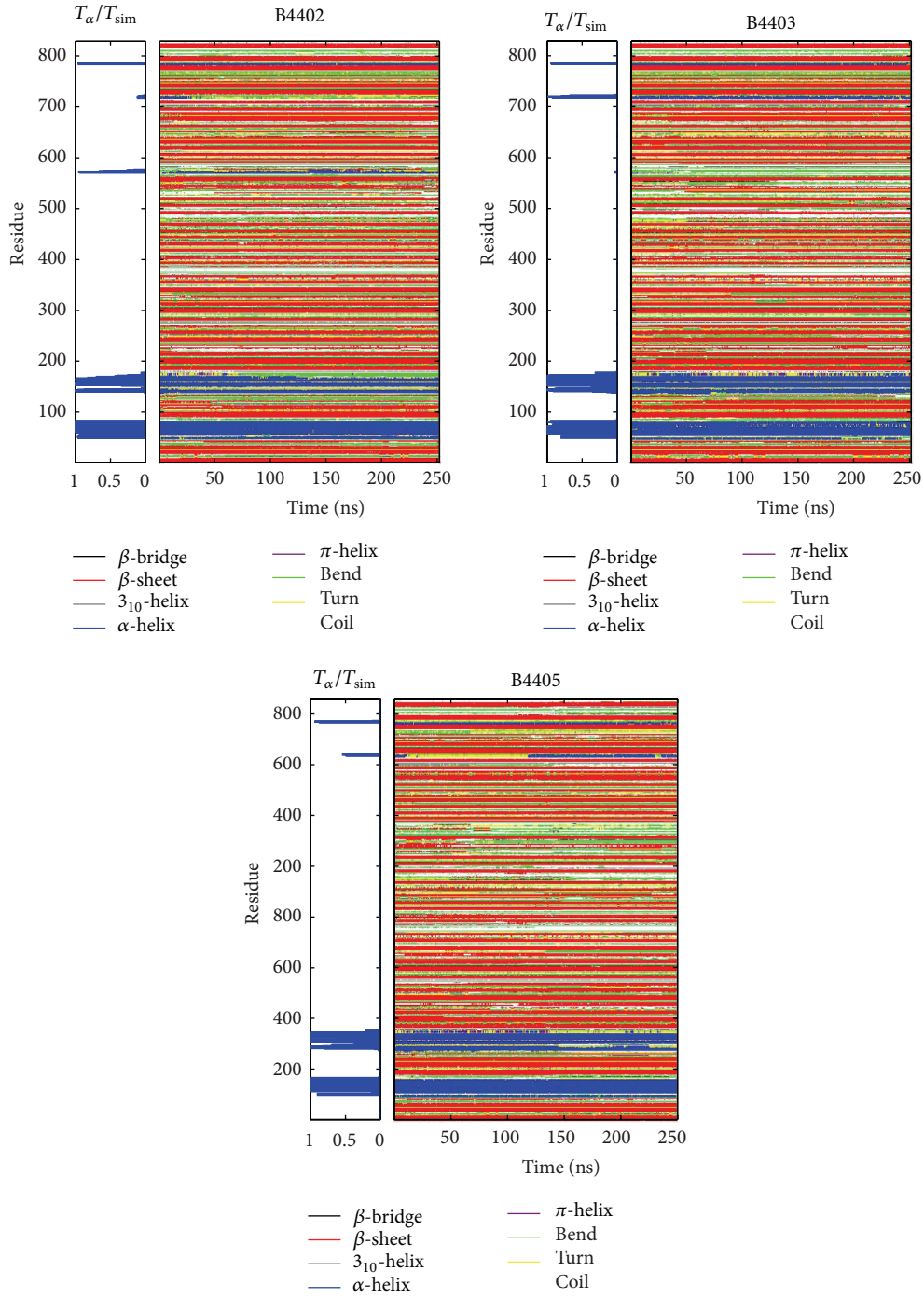


Figure 16: Dynamics of secondary structure. The TCR/pMHC systems B4402, B4403, and B4405 comprise 829 amino acids. The following list shows which residues belong to which protein: MHC, residues 1–276; β-2-microglobulin, residues 277–375; ABCD3 peptide, residues 376–384; TCR α, residues 384–584; TCR β, residues 585–825. The graph on the right-hand side displays the structural behaviour of amino acid residues along the simulation time. Different secondary structural elements are assigned different colours as shown in the legend. Secondary structures are stable along the 250 ns MD simulation for all three TCR/pMHC systems. The graph on the left-hand side displays the relative simulation time that an amino acid residue is part of an α-helix. Extended and stable α-helices in these TCR/pMHC systems are only present in the MHC molecule.

Acknowledgments

The MD trajectories used in the present work were generated on the IBM BlueGene/P computer facility at Bulgarian National Centre for Supercomputing Applications (NCSA, http://www.scc.acad.bg/). The work was supported in part by BSF and OeAD under Grants nos. DNTS-A 01-2/2013 and WTA-BG 06/2013.

References

[1] O. Acuto, V. Di Bartolo, and F. Michel, "Tailoring T-cell receptor signals by proximal negative feedback mechanisms," *Nature Reviews Immunology*, vol. 8, no. 9, pp. 699–712, 2008.

[2] D. N. Garboczi, P. Ghosh, U. Utz, Q. R. Fan, W. E. Biddison, and D. C. Wiley, "Structure of the complex between human T-cell receptor, viral peptide and HLA-A2," *Nature*, vol. 384, no. 6605, pp. 134–141, 1996.

[3] Y.-H. Ding, K. J. Smith, D. N. Garboczi, U. Utz, W. E. Biddison, and D. C. Wiley, "Two human T cell receptors bind in a similar diagonal mode to the HLA- A2/Tax peptide complex using different TCR amino acids," *Immunity*, vol. 8, no. 4, pp. 403–411, 1998.

[4] J. K. Burkhardt, "Seeing is believing: sorting out signaling events at the immunological synapse," *The Journal of Immunology*, vol. 194, no. 9, pp. 4059–4060, 2015.

[5] L. E. Samelson, M. D. Patel, A. M. Weissman, J. B. Harford, and R. D. Klausner, "Antigen activation of murine T cells induces tyrosine phosphorylation of a polypeptide associated with the T cell antigen receptor," *Cell*, vol. 46, no. 7, pp. 1083–1090, 1986.

[6] G. R. Crabtree, "Contingent genetic regulatory events in T lymphocyte activation," *Science*, vol. 243, no. 4889, pp. 355–361, 1989.

[7] L. A. Timmerman, N. A. Clipstone, S. N. Ho, J. P. Northrop, and G. R. Crabtree, "Rapid shuttling of NF-AT in discrimination of Ca^{2+} signals and immunosuppression," *Nature*, vol. 383, no. 6603, pp. 837–840, 1996.

[8] A. Weiss, R. Shields, M. Newton, B. Manger, and J. Imboden, "Ligand-receptor interactions required for commitment to the activation of the interleukin 2 gene," *Journal of Immunology*, vol. 138, no. 7, pp. 2169–2176, 1987.

[9] F. Pappalardo, V. Brusic, F. Castiglione, and C. Schönbach, "Computational and bioinformatics techniques for immunology," *BioMed Research International*, vol. 2014, Article ID 263189, 2 pages, 2014.

[10] R. Ribarics, R. Karch, N. Ilieva, and W. Schreiner, "Geometric analysis of alloreactive HLA α-helices," *BioMed Research International*, vol. 2014, Article ID 943186, 8 pages, 2014.

[11] M. Kenn, R. Ribarics, N. Ilieva, and W. Schreiner, "Finding semirigid domains in biomolecules by clustering pair-distance variations," *BioMed Research International*, vol. 2014, Article ID 731325, 13 pages, 2014.

[12] W. A. Macdonald, Z. Chen, S. Gras et al., "T cell allorecognition via molecular mimicry," *Immunity*, vol. 31, no. 6, pp. 897–908, 2009.

[13] W. Stacklies, M. C. Vega, M. Wilmanns, and F. Gräter, "Mechanical network in titin immunoglobulin from force distribution analysis," *PLoS Computational Biology*, vol. 5, no. 3, Article ID e1000306, 2009.

[14] M. Karplus and J. Kuriyan, "Molecular dynamics and protein function," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 19, pp. 6679–6685, 2005.

[15] B. Hischenhuber, F. Frommlet, W. Schreiner, and B. Knapp, "MH2c: characterization of major histocompatibility α-helices—an information criterion approach," *Computer Physics Communications*, vol. 183, no. 7, pp. 1481–1490, 2012.

[16] J. A. Christopher, R. Swanson, and T. O. Baldwin, "Algorithms for finding the axis of a helix: fast rotational and parametric least-squares methods," *Computers and Chemistry*, vol. 20, no. 3, pp. 339–345, 1996.

[17] J. A. Lopera, J. N. Sturgis, and J.-P. Duneau, "Ptuba: a tool for the visualization of helix surfaces in proteins," *Journal of Molecular Graphics and Modelling*, vol. 23, no. 4, pp. 305–315, 2005.

[18] B. Hischenhuber, H. Havlicek, J. Todoric, S. Höllrigl-Binder, W. Schreiner, and B. Knapp, "Differential geometric analysis of alterations in MH α-helices," *Journal of Computational Chemistry*, vol. 34, no. 21, pp. 1862–1879, 2013.

[19] B. Hischenhuber, H. Havlicek, J. Todoric, S. Höllrigl-Binder, W. Schreiner, and B. Knapp, "Corrigendum: differential geometric analysis of alterations in MH alpha-helices," *Journal of Computational Chemistry*, vol. 34, no. 32, p. 2834, 2013.

[20] P. K. Warme, F. A. Momany, S. V. Rumball, R. W. Tuttle, and H. A. Scheraga, "Computation of structures of homologous proteins. α-lactalbumin from lysozyme," *Biochemistry*, vol. 13, no. 4, pp. 768–782, 1974.

[21] C. Oostenbrink, A. Villa, A. E. Mark, and W. F. Van Gunsteren, "A biomolecular force field based on the free enthalpy of hydration and solvation: the GROMOS force-field parameter sets 53A5 and 53A6," *Journal of Computational Chemistry*, vol. 25, no. 13, pp. 1656–1676, 2004.

[22] U. Omasits, B. Knapp, M. Neumann et al., "Analysis of key parameters for molecular dynamics of pMHC molecules," *Molecular Simulation*, vol. 34, no. 8, pp. 781–793, 2008.

[23] W. Kabsch and C. Sander, "Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features," *Biopolymers*, vol. 22, no. 12, pp. 2577–2637, 1983.

[24] L. Pauling, R. B. Corey, and H. R. Branson, "The structure of proteins: two hydrogen-bonded helical configurations of the polypeptide chain," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 37, no. 4, pp. 205–211, 1951.

[25] W. Kabsch, "A solution for the best rotation to relate two sets of vectors," *Acta Crystallographica A*, vol. 32, no. 5, pp. 922–923, 1976.

[26] C. Runge, "Über empirische Funktionen und die Interpolation zwischen äquidistanten Ordinaten," *Zeitschrift für Mathematik und Physik*, vol. 46, no. 1, pp. 224–243, 1976.

[27] J. P. Boyd and F. Xu, "Divergence (Runge Phenomenon) for least-squares polynomial approximation on an equispaced grid and Mock-Chebyshev subset interpolation," *Applied Mathematics and Computation*, vol. 210, no. 1, pp. 158–168, 2009.

[28] M. Peternell, H. Pottmann, and B. Ravani, "On the computational geometry of ruled surfaces," *CAD Computer Aided Design*, vol. 31, no. 1, pp. 17–32, 1999.

[29] W. Kühnel, *Differentialgeometrie Kurven—Flächen—Mannigfaltigkeiten*, Vieweg+Teubner, Wiesbaden, Germany, 2008.

[30] P. W. Verbeek and L. J. van Vliet, "Curvature and bending energy in digitized 2D and 3D images," in *Proceedings of the 8th Scandinavian Conference on Image Analysis*, pp. 1403–1410, Tromsø, Norway, May 1993.

[31] M. C. Lawrence and P. M. Colman, "Shape complementarity at protein/protein interfaces," *Journal of Molecular Biology*, vol. 234, no. 4, pp. 946–950, 1993.

[32] D. Aivazian and L. J. Stern, "Phosphorylation of T cell receptor zeta is regulated by a lipid dependent folding transition," *Nature Structural Biology*, vol. 7, no. 11, pp. 1023–1026, 2000.

[33] K. Choudhuri and P. A. van der Merwe, "Molecular mechanisms involved in T cell receptor triggering," *Seminars in Immunology*, vol. 19, no. 4, pp. 255–261, 2007.

[34] M. L. Dustin and D. Depoil, "New insights into the T cell synapse from single molecule techniques," *Nature Reviews Immunology*, vol. 11, no. 10, pp. 672–684, 2011.

[35] S. T. Kim, K. Takeuchi, Z.-Y. J. Sun et al., "The $\alpha\beta$ T cell receptor is an anisotropic mechanosensor," *Journal of Biological Chemistry*, vol. 284, no. 45, pp. 31028–31037, 2009.

[36] C. Xu, E. Gagnon, M. E. Call et al., "Regulation of T cell receptor activation by dynamic membrane binding of the CD3epsilon cytoplasmic tyrosine-based motif," *Cell*, vol. 135, no. 4, pp. 702–713, 2008.

[37] M. S. Kuhns, A. T. Girvin, L. O. Klein et al., "Evidence for a functional sidedness to the $\alpha\beta$TCR," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 11, pp. 5094–5099, 2010.

[38] H. Zhang, S.-P. Cordoba, O. Dushek, and P. A. van der Merwe, "Basic residues in the T-cell receptor zeta cytoplasmic domain mediate membrane association and modulate signaling," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 48, pp. 19323–19328, 2011.

[39] N. Martinez-Martin, R. M. Risueno, A. Morreale et al., "Cooperativity between T cell receptor complexes revealed by conformational mutants of CD3ε," *Science Signaling*, vol. 2, no. 83, article ra43, 2009.

[40] G. Ryan, "T cell signalling: CD3 conformation is crucial for signalling," *Nature Reviews Immunology*, vol. 10, no. 1, p. 7, 2010.

[41] R. Blanco, A. Borroto, W. Schamel, P. Pereira, and B. Alarcon, "Conformational changes in the T cell receptor differentially determine T cell subset development in mice," *Science Signaling*, vol. 7, no. 354, Article ID ra115, 2014.

[42] A. A. Petruk, S. Varriale, M. R. Coscia, L. Mazzarella, A. Merlino, and U. Oreste, "The structure of the CD3 $\zeta\zeta$ transmembrane dimer in POPC and raft-like lipid bilayer: a molecular dynamics study," *Biochimica et Biophysica Acta (BBA)—Biomembranes*, vol. 1828, no. 11, pp. 2637–2645, 2013.

[43] W. Humphrey, A. Dalke, and K. Schulten, "VMD: visual molecular dynamics," *Journal of Molecular Graphics*, vol. 14, no. 1, pp. 33–38, 1996.

[44] M. L. Connolly, "Analytical molecular surface calculation," *Journal of Applied Crystallography*, vol. 16, no. 5, pp. 548–558, 1983.

[45] W. Schreiner, R. Karch, B. Knapp, and N. Ilieva, "Relaxation estimation of RMSD in molecular dynamics immunosimulations," *Computational and Mathematical Methods in Medicine*, vol. 2012, Article ID 173521, 9 pages, 2012.

Hindawi Publishing Corporation
Journal of Immunology Research
Volume 2015, Article ID 739706, 10 pages
http://dx.doi.org/10.1155/2015/739706

Review Article

Understanding Experimental LCMV Infection of Mice: The Role of Mathematical Models

Gennady Bocharov,[1] Jordi Argilaguet,[2] and Andreas Meyerhans[2,3]

[1]*Institute of Numerical Mathematics, Russian Academy of Sciences, Gubkina Street 8, Moscow 119333, Russia*
[2]*Infection Biology Laboratory, Department of Experimental and Health Sciences (DCEXS), Universitat Pompeu Fabra, Doctor Aiguader 88, 08003 Barcelona, Spain*
[3]*Institució Catalana de Recerca i Estudis Avançats (ICREA), Passeig Lluis Companys 23, 08010 Barcelona, Spain*

Correspondence should be addressed to Gennady Bocharov; gbocharov@gmail.com

Received 30 July 2015; Accepted 27 September 2015

Academic Editor: Francesco Pappalardo

Copyright © 2015 Gennady Bocharov et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Virus infections represent complex biological systems governed by multiple-level regulatory processes of virus replication and host immune responses. Understanding of the infection means an ability to predict the systems behaviour under various conditions. Such predictions can only rely upon quantitative mathematical models. The model formulations should be tightly linked to a fundamental step called "coordinatization" (Hermann Weyl), that is, the definition of observables, parameters, and structures that enable the link with a biological phenotype. In this review, we analyse the mathematical modelling approaches to LCMV infection in mice that resulted in quantification of some fundamental parameters of the CTL-mediated virus control including the rates of T cell turnover, infected target cell elimination, and precursor frequencies. We show how the modelling approaches can be implemented to address diverse aspects of immune system functioning under normal conditions and in response to LCMV and, importantly, make quantitative predictions of the outcomes of immune system perturbations. This may highlight the notion that data-driven applications of meaningful mathematical models in infection biology remain a challenge.

1. Introduction

One of the best-studied model systems of viral infections is that of the lymphocytic choriomeningitis virus (LCMV) in mice (Figure 1) [1–3]. LCMV is an RNA virus of Arenaviridae that is noncytopathic in vivo. Thus, the virus itself does not cause direct damage to cells and tissues. This feature enables relating any damage that appears in the course of an infection to host responses against the virus. Another important feature of the LCMV model system is the existence of several well-characterized viral strains that differ in their replicative capacity, host range (cell tropism and mouse strain), and experimental routes of infection (intracranial versus intraperitoneal (i.p.) or intravenous (i.v.)) and thus show different infection outcomes. This enables directly linking easily measurable viral dynamic properties to pathogenic consequences and studying the fundamental issue of chronic infections.

With the use of the LCMV infection model system, a large number of conceptual discoveries in immunology have been made, of which we cite here just a few. First, back in 1974/75, Zinkernagel and Doherty demonstrated that cytotoxic T lymphocytes (CTLs) recognize foreign antigens only in the context of proteins of the major histocompatibility complex (MHC) [4, 5]. For this finding of "MHC restriction," they were awarded the Nobel Prize in 1996. Second, with the help of knockout mice, the mechanism of CTL-mediated destruction of LCMV-infected target cells in vivo could be directly linked to perforin, a pore-forming protein contained in granules of this cell type [6, 7]. Third, fundamental properties of "memory" of the adaptive immune response have been studied. For example, a quantitative understanding on the number of epitope-specific precursor T cells, their expansion and contraction in the course of an acute LCMV strain Armstrong infection, and their maintenance was established [8–10] (for details, see below).

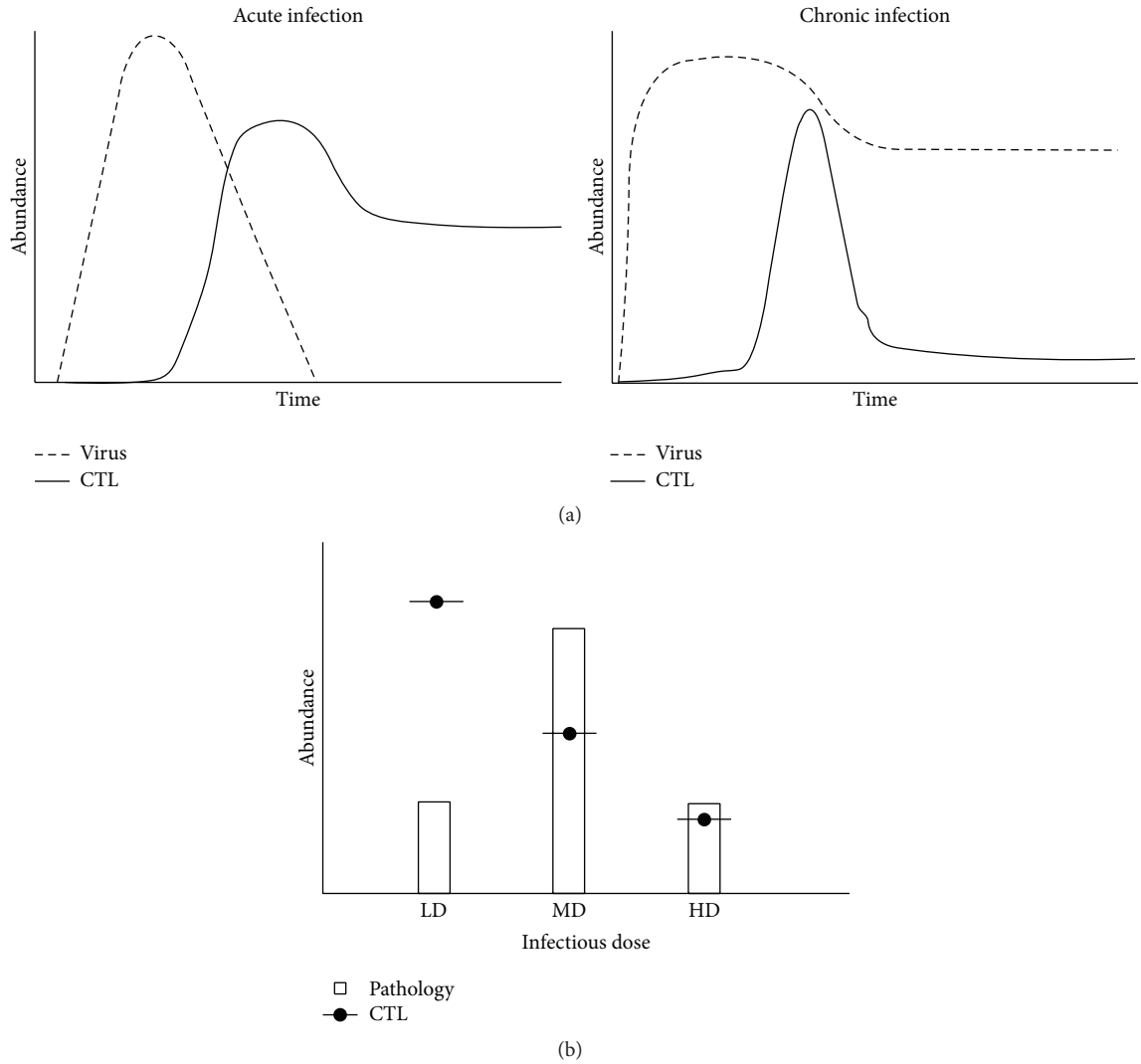


Figure 1: (a) Schematic view of acute and chronic LCMV infections including virus and cytotoxic T lymphocyte (CTL) dynamics. (b) Representation of the CTL-induced immunopathology dependence on the initial viral infectious dose at day 13 after infection (adapted from Cornberg et al. [62]). LD: low-dose infection; MD: medium-dose infection; HD: high-dose infection.

Further studies defined the requirements for CTL memory to prevent the establishment of a persistent LCMV infection [11]. Fourth, NK cells of the innate immune response have been recognized as an important regulator of the helper T cell support for antiviral CTL [12]. Fifth, a critical role of organized secondary lymphoid organs in the induction of naive T and B cells and subsequent virus control was established [13]. Sixth, the concept of immunopathology, that is, the damage of tissues and organs due to the antiviral immune response rather than the infecting virus itself, was established. Mediators of immunopathology include CTL, macrophages, neutrophils, and interferons [14–16]. Seventh, based on the amino acid similarities between viral antigens and host proteins, the so-called molecular mimicry, viral infections can trigger autoimmunity and influence the course of subsequent infections with other viral pathogens [17–19]. Eighth, important observations towards an acute versus a persistent LCMV infection outcome were made [20–23]. Which infection fate is followed depends on the infecting viral dose and the infecting viral strain and thus can be easily directed experimentally. LCMV persistence is associated with CTL exhaustion, a reversible, nonfunctional state of CTL. As CTL exhaustion seems to be a physiological consequence of persistent antigen exposure and has been observed both in persistent human viral infections and in cancers, the LCMV system is highly attractive to understand infection fate regulation in general terms. As CTL exhaustion can be reversed by antibodies against PD1 or PD-L1 that block the negative signalling pathway, novel immunotherapeutic modalities arose which show exciting promises as antiviral and anticancer therapies [24–26]. Several clinical studies in this direction have been initiated and are ongoing.

The LCMV infection model system offers sufficient experimental data to develop mathematical models in a problem-oriented manner. Indeed, although only around 20 mathematical modelling studies have been published today

addressing specific aspects of LCMV infection, they are more instructive than studies of other infections including HIV [27]. Indeed, the modelling studies of LCMV resulted in experimentally testable predictions concerning the mechanisms of the infection control, for example, (i) protective numbers of the initial CTL precursors to protect from a chronic LCMV infection outcome, (ii) minimal number of antigen presenting DCs in spleen for robust induction of CTL responses, and (iii) the effect of virus growth rate on the magnitude of the clonal expansion of CTLs, to name just the major of them. In this review, we summarize how the modelling approaches were tailored to address diverse aspects of immune system functioning under normal conditions and in response to LCMV and, importantly, make quantitative predictions of the outcomes of immune system perturbations.

2. Mathematical Models of LCMV Infection

Mathematical models developed for the analysis of experimental LCMV infection in mice enabled quantifying some fundamental parameters of the virus-host interaction including the numbers and turnover rates of immune cells and the growth rates of viruses. A maximum likelihood approach to nonlinear model parameter estimation utilizing precise and comprehensive data sets proved to be instrumental. The estimates of the fundamental parameters of T cell response to LCMV infection are summarized in Table 1.

2.1. Basic Numbers for Induction of the Clonal Expansion: Precursor CTLs and DCs. The initial frequency of virus antigen-specific T lymphocytes is a primary control parameter determining the speed and the magnitude of the antiviral immune response. Although the total number of lymphocytes in mice is high, that is, about 10^9 [28], the fraction of the cells specific for a given viral antigen is very low being of the order of 10^{-5}–10^{-6}. As the division time of the lymphocytes is rather slow (~12 hours) compared to the replication rate of LCMV, the precise quantitation of the precursor T cell number is important in predicting the time needed for their clonal expansion above the threshold required to eliminate the virus from an infected animal. Early estimates of the number of LCMV-specific CTL precursor cells in spleen of naive C57BL/6 mice suggested that it is less than 1 in 10^5 [29]. The first quantitative mathematical model of the LCMV infection provided the best-fit estimate of the number of LCMV Docile-specific precursor CTLs of about 27 cells per spleen [30]. A similar parameter estimation from experimental LCMV WE infections led to the value of about 110 naive virus-specific CTLs per spleen [31]. Taking into account the fact that the splenic population of lymphocytes is about 5% of the total lymphocyte number, the extrapolation from spleen to the whole mouse suggests that about 540 to 2200 precursor CTLs are specific for LCMV. This estimate is between the values obtained by later experimental examination of the precursor CTLs specific for the H-2D^b-restricted GP33-41 (GP33) epitope of LCMV quantitated from in vivo competition assay [8] and those specific for an entire virus as quantitated via in vivo limiting-dilution assay [32]. The estimated numbers for naive C57BL/6 mice are from 100 to 200 GP33-specific CTLs to about 6,761 LCMV-specific CD8$^+$ T cells, respectively. A finer quantitative dissection of CD4$^+$ and CD8$^+$ T cell responses to infection of C57BL/6 mice with LCMV Armstrong by considering the epitope-specific clones was done using the exponential growth and contraction model in [33]. The data-driven parameter estimation suggested the following population sizes for T cells starting the proliferation (i.e., for the lower bound on the precursor number) with their respective 95% confidence intervals ($CI_{95\%}$). For CD8$^+$ T lymphocytes, the per-spleen estimates are 12 cells (3, 33) for GP33, 7 cells (1, 21) for NP396, 6 cells (0.4, 40) for GP118, 5 cells (1, 14) for GP276, 29 cells (9, 82) for NP205, and 165 cells (34, 519) for GP92 and in total ~224 cells per spleen. For CD4$^+$ T lymphocytes, the estimates are 22 cells (19, 27) for GP61 and 56 cells (46, 74) for NP309 and in total 78 precursor cells. The above estimates were obtained by fitting the model to epitope-specific T cell data quantitated by intracellular cytokine staining. Similar parameter estimation from the data on Ag-specific CD8$^+$ T cells in spleen measured by MHC tetramer staining resulted in the following numbers [34]: 8 cells (7, 11) for GP33, 5 cells (4, 6) for NP396, 2 cells (1.5, 3) for GP34, and 5 cells (4, 6) for GP276 and in total 20 cells.

The activation of T cells requires MHC-restricted antigen presentation by professional antigen presenting cells. The assessment of the threshold number of the dendritic cells for the induction of robust CD8$^+$ T cell responses in secondary lymphoid organs was made using data-driven mathematical modelling. The study by Ludewig et al. [35] examined the impact of the dendritic cells (DCs) number on the induction of the CTL response. C57BL/6 mice were adoptively transferred by intravenous injection with 2×10^4, 2×10^5, and 2×10^6 GP33-presenting DCs from transgenic mice ubiquitously expressing the LCMV glycoprotein peptide GP33 (H8-mice). The population dynamics of activated GP33-specific CTL (H2-D^b/GP33-tetramer-binding, CD8$^+$CD62L$^-$) and quiescent "memory" CTL (CD62L$^+$) in blood, spleen, and liver were followed. For the data analysis, a three-compartment delay differential equation model was formulated which considered the population dynamics of DCs and CTLs. The maximum likelihood approach to the model calibration was used to estimate the relevant parameter of the cell circulation and interaction. The model predicted that the threshold number of DCs in the spleen for induction of half maximal proliferation of CTLs is about 212 cells with the corresponding 95% uncertainty interval (75, 1200). A later independent study of the minimum number of DCs required to initiate a T cell response arrived at similar numbers [36]. The analysis combined the experimental assessment of the T cell and antigen-bearing DC encounters in popliteal lymph nodes (LN) with intravital 2-photon and confocal images and flow cytometry examination of the phosphorylation following the footpad injection of DCs and i.v. injection of CD4$^+$ T cells and Dby peptides. The developed computational model of T cell DC encounter described the Brownian motion of moving T cell and static DCs in a spherical volume approximating the

Table 1: Fundamental parameters of T cell response to LCMV infection and the relevant LCMV epitopes.

Parameter (mouse strains)	Estimated parameter: value, range ($CI_{95\%}$), union of ranges	References
Number of single LCMV-epitope-specific precursor $CD8^+$ T cells (C57BL/6 mice)	100–200 (per mouse): GP33 5–165 (per spleen): GP33, GP92, GP118, GP276, NP205, and NP396 2–8 (per spleen): GP33, GP34, GP276, and NP396	[8, 32–34]
Number of single LCMV-epitope-specific precursor $CD4^+$ T cells (C57BL/6 mice)	22–56 (per spleen): GP61, NP309	[33]
Total number of the precursor $CD8^+$ T cells specific for an entire virus (C57BL/6 mice)	27–110 (per spleen): WE and Docile 6,761 (per mouse): GP33 + GP118 + GP276 + NP396 + NP205 224 (per spleen): GP33 + GP118 + GP276 + NP396 + NP205 20 (per spleen): GP33 + GP34 + GP276 + NP396	[30–34]
Total number of the precursor $CD4^+$ T cells specific for an entire virus (C57BL/6 mice)	78 (per spleen): GP61 + NP309	[33]
Number of dendritic cells required for induction of CD8 T cell clonal expansion (C57BL/6 mice)	212, $CI_{95\%} = (75, 1200)$ (per spleen): H8-mice DCs	[35]
Doubling time of LCMV-epitope-specific $CD8^+$ T cells during clonal expansion phase (C57BL/6 mice, BALB/c mice)	7.5–16.7 (hours): GP33, GP92, GP118, GP276, NP205, and NP396 5.5–7.6 (hours): GP238, NP118	[33, 37]
Doubling time of LCMV-epitope-specific $CD4^+$ T cells during clonal expansion phase (C57BL/6 mice)	10.5–17.3 (hours): GP61, NP309	[33]
Half-lives of epitope-specific $CD8^+$ T cells during contraction phase (BALB/c mice)	19.6–87.6 (hours): GP238, NP118	[37]
Half-lives of infected target cells killed by epitope-specific $CD8^+$ T cells (C57BL/6 mice)	1.4 (hours) at day 8, 2.9 (hours) at day 30, and 8.9 (hours) at day 300: H8-spleen cells 0.11–0.46 (hours) at day 8 with effector frequency 0.05 per spleen: GP276, NP396 0.048–0.16 (hours) at day 8: GP276, NP396 0.11–0.24 (hours) for acute infection (day 8), 0.27–0.37 (hours) for memory phase (day 42), and 0.31–0.44 (hours) for chronic infection (day 42): GP33	[38–41]
Threshold frequency of CTLs in spleen at which the infected cells elimination rate is half-maximum (C57BL/6 mice, infection with LCMV Docile)	0.004–0.023 for acute (day 8) and chronic (day 42) infection, 0.007–0.088 for memory phase of infection (day 42): GP33	[41]
Protective number of memory CTLs against infection (C57BL/6 mice, infection with LCMV Armstrong)	1.3×10^5 cells per spleen for GP276, NP396	[43]
Protective number of naive precursor CTLs against chronic infection (C57BL/6 mice)	10^5 cells per spleen for infection with 10^5 pfu LCMV Docile (cells from TCR318 mice)	[11]
Dependence of CTL clonal expansion on virus growth rate (C57BL/6 mice)	Bell-shaped; both slow and fast replicating virus strains can induce weak $CD8^+$ T cell clonal expansion (GP33)	[47]

LN. The model allowed one to calculate the probability of a T cell to interact with antigen-bearing DC within 24 hours, which is 0.58 for 100 DCs and increases up to 0.99 for 10^3 DCs, respectively.

2.2. T Cell Proliferation. Following the stimulation with LCMV antigens, the specific T lymphocytes enter the expansion phase and after reaching a peak of expansion the population starts to decline during the contraction phase. Although the general scales of the clonal expansion and contraction can be assessed directly from experiments with LCMV infection, for example, [21], the mathematical modelling in conjunction with the data on kinetics of the LCMV-specific T cell responses provided a fine kinetic characterization of the proliferation and death rates of the epitope-specific $CD4^+$ and $CD8^+$ T cells for both primary and chronic infection phases [33, 34, 37]. The data on LCMV-Armstrong i.p. infection of BALB/c mice were inverted into the estimates of the net doubling time of NP118 and GP283-specific $CD8^+$ T cell and their half-life for the expansion and contraction phases of the primary immune response, respectively [37]. A piecewise linear system of ordinary differential equations was used to describe the population dynamics of naive, activated, and memory phenotype $CD8^+$ T cells. A simplifying assumption was used saying that the viral load just switches the proliferation of the T cells between the full and zero modes in a time-dependent manner. The viral kinetics invariant estimates of the proliferation and death rates during the

expansion and contraction phases, respectively, are presented as the doubling time and half-lives with the respective 95% uncertainty intervals. The best-fit doubling times are 5.7 (hours) (5.5, 6.2) for the NP118 epitope and 6.4 (hours) (5.5, 7.6) for the GP283 epitope. The best-fit half-lives are 32.6 (hours) (26, 39.6) for the NP118 epitope and 46.2 (hours) (19.8, 87.6) for the GP283 epitope.

A similar analysis has been used to assess differences in proliferation rates between $CD4^+$ and $CD8^+$ T cells during the clonal expansion phase [33]. The epitope-specific $CD4^+$ and $CD8^+$ T cells from spleens of C57BL/6 mice after intraperitoneal infection with 10^5 pfu of LCMV Armstrong were used. Although it was pointed out that multiple mathematical formulations can be developed to describe the data, the information-theoretic criteria were not applied to rank the models. The best-fit estimates with the respective 95% confidence intervals obtained using a model similar to that of De Boer et al. [37] and extended to consider a biphasic contraction phase suggested the following doubling times for $CD4^+$ and $CD8^+$ T cells: for GP61-specific (immunodominant epitope) and NP309-specific (subdominant epitope) $CD4^+$ T cells, the values were 11.3 (hours) (10.5, 12.1) and 15 (hours) (13.3, 17.3), respectively. The doubling times of $CD8^+$ T cell were shorter with dominance ranking of GP33 > NP396 > GP118 > GP276 > NP205 > GP92 and were as follows: 8.8 (hours) (8, 9.6), 8.7 (hours) (7.8, 9.5), 8.9 (hours) (7.5, 10.4), 8.9 (hours) (8.1, 9.8), 10.9 (hours) (9.8, 12.1), and 14.7 (hours) (12.5, 16.7), respectively. The immunodominant epitopes exhibited faster proliferation rates.

2.3. Target Cell Elimination. It is not only the number of virus-specific CTLs that is important for the elimination of infection but also the efficacy of target cell elimination in vivo. The elimination rates of transferred cells expressing LCMV antigens into immune mice at the peak of an acute response and during the memory phase were first estimated using a simple exponential decay model in [38]. The best-fit estimates of the half-lives of the target cells at day 8 after infection were 1.4 hours and increased to 2.9 hours 30 days and finally to 8.9 hours 300 days after infection. The estimates of the elimination rates of the LCMV epitopes expressing cells by virus-specific CTLs in vivo were a subject of a number of follow-up studies in which more complex mathematical models were used.

Regoes et al. considered the migration of donor cells from blood to spleen [39]. The killing of the antigen-pulsed target cells by CTL in immune mice was assumed to follow a mass action law. The data used for the estimation of the target cell elimination rate were derived from experiments on i.p. infection of C57BL/6 mice with 2×10^5 pfu LCMV Armstrong. The number of virus-specific CTLs at day 8 after infection was assumed to be 5×10^6 per gram of spleen. The half-life of the PKH26-stained target cells due to the killing by NP396-epitope-specific CTL was estimated to be 0.17 (hours) with the 95% confidence interval (0.11, 0.33), whereas for the subdominant epitope the values were 0.33 (hours) (0.24, 0.46).

In the follow-up study [40] based on the same data, the data fitting procedure was refined to reduce the variability between animals (splenocyte numbers, magnitude of the CTL response to LCMV infection, and inocula size). This was done by (i) pairing the estimates of unpulsed and pulsed target cell frequencies in each animal and (ii) splitting the parameter estimation procedure into two stages, the estimation of the transfer rate of target cells from blood to spleen and the estimation of the target cell killing rate. This procedure led to about 3-fold increase of the best-fit estimate of the killing rate at the peak (day 8) of the acute infection.

The target cell elimination kinetics depend on many factors and processes including the migration of the cells into the spleen, the decay of the epitope from target cells, the number of antigen-specific CTLs, the functional status of CTL (i.e., effector state or exhaustion state), the load of LCMV peptides on target cells, and the parameterization of the target cell-CTL interaction. The last two issues have been systematically examined by Garcia et al. [41] following a model analysis oriented data generation approach. The in vivo killing assay was conducted in C57BL/6 mice acutely (200 pfu) or chronically (10^6 pfu) infected with LCMV Docile. Six different peptide (GP33) loads spanning four orders of magnitude were used to pulse the adoptively transferred homozygous splenocytes. It was established that the ability of CTL to recognize and kill infected target cells depends on the number of peptide-MHC complexes presented on the cell surface. In addition, a saturation effect in target cell killing rates for high CTL numbers was suggested by the analysis of the quality of the data fitting. From the modelling perspective, a more accurate mathematical description of the target cell killing kinetics in relation to the peptide load (λ) and the abundance of CTL in the spleen (C) was proposed:

$$f \propto \frac{k_{\max}\lambda}{\lambda_{0.5} + \lambda} \times \frac{C}{C_{0.5} + C}. \quad (1)$$

The parameters characterizing the maximum killing rate, half-maximum peptide density, and half-maximum CTL frequency ($k_{\max}, \lambda_{0.5}$, and $C_{0.5}$) were estimated from the data. The minimal half-lives of infected cells (defined by the maximum killing rate in the respective groups) in face of their elimination by the epitope-specific CTLs were estimated to be 0.17 (hours), $CI_{95\%}$ = (0.11, 0.24), for acute infection, 0.32 (hours), $CI_{95\%}$ = (0.27, 0.37), for the memory phase, and, surprisingly, 0.38 (hours), $CI_{95\%}$ = (0.31, 0.44), for the chronic infection phase. The value of $\lambda_{0.5}$ characterizing the sensitivity of CTL to the peptide frequency of the presented epitopes on the target cells was estimated to be the highest in acute infection and similar for the chronic infection and the memory infection phase. Furthermore, the limits of validity of the mass action law in the description of the target cell elimination, as generally used in data analysis, were examined. The values of the CTL abundance in spleen were estimated for acute infection, memory infection phase, and chronic infection to be 0.042, 0.031, and 0.006, respectively. Earlier computational studies based on a cellular automata model predicted that, above CTL frequencies of 0.03, saturation effects of target cell elimination have to be taken into account [42]. The estimated values of the threshold

density of CTL $C_{0.5}$ at which the elimination rate is half-maximum are 0.013 with $CI_{95\%}$ = (0.004, 0.023) for acute and chronic infection but increase 4-fold to 0.051, $CI_{95\%}$ = (0.007, 0.088), for the memory phase of infection. Overall, the study led to a number of novel insights into the mechanics of target cell elimination: (i) there is no evidence of an increased recognition sensitivity of memory CTL compared to acute or chronic CTLs; (ii) the killing ability of CTL in chronic LCMV infection is at least as strong as during acute infection.

3. Acute and Chronic LCMV Infection

There are relatively few mathematical models of LCMV infection in which the population dynamics of viruses and immune responses as shown in Figure 1 were formulated [30, 43–45]. The validity of the law of mass action in the description of target cell elimination by $CD8^+$ T cells was shown [43]. The mathematical model formulated with ODE described the growth and elimination of the virus population by CTLs. The dynamics of the adoptively transferred LCMV peptide-loaded target cells were described analytically following the model by Ganusov and De Boer [46]. An exponential growth model was used to describe the expansion of CTLs. The combined equations allowed the authors to estimate the critical number of CTLs at which the virus growth can be prevented from the start of infection. For LCMV Armstrong with the exponential growth rate assumed to be 5 per day, the protective number of memory CTLs was around 1.3×10^5 cells.

One of the first quantitative models of LCMV infection was developed by Bocharov [30] to describe the population dynamics of virus, precursor CTL, and effector CTL using delay differential equations. The data in low-, intermediate-, and high-dose infection of C57BL/6 mice with LCMV Docile [21] were used for model calibration. The model was used to predict the effect of the variation in the number of precursor CTLs on the outcome of LCMV infection. The model-generated predictions were tested experimentally by Ehl et al. [11] and the following conclusions have been made: (i) a minimal threshold number of about 25–50 naive LCMV-specific CTL precursors (CTLp) are necessary for control of infections in the range of 1–10^4 pfu; (ii) with 10-fold higher doses, a 100-fold increase in CTLp is required to restore virus control; (iii) in high-dose infection (above 10^6 pfu), elevations in CTLp were found to be detrimental as they changed the outcome of infection from harmless virus persistence to lethal immunopathology. Overall, above a critical threshold, the time when effector function is reached by CTLs is more important than the initial number of virus-specific CTL precursors.

The mathematical model developed by Bocharov [30] was further used to predict the impact of the virus replication kinetics on the magnitude of the CTL response in acute LCMV infection. The experimental analysis of the clonal expansion of CTLs in C57BL/6 mice to LCMV strains (Armstrong, WE-Armstrong, WE, Traub, and Docile) differing in their replication rate [47] confirmed that there is a bell-shaped relationship between the LCMV growth rate and the peak CTL response. It was shown that both slow and fast replicating LCMV strains produce weaker CTL responses. A mechanism of virus persistence by sneaking through immune surveillance due to slow replication kinetics was hypothesized and its relevance for HBV and HCV infections was shown. The "underwhelming" infection mechanism (supplementing the "overwhelming" infection [21]) fits the concept of the sensitivity of immune responses to perturbations [48].

Infection of mice with certain strains of LCMV can result in the development of lifelong virus persistence. The role of various host and viral parameters in the development of chronic LCMV infection has been examined experimentally. One of the fundamental features of the establishment of LCMV persistence was associated with the exhaustion of antiviral cytotoxic effector T cells after their early and complete induction [21]. The phenomenon of exhaustion was defined as complete disappearance of CTL activity and the clonal deletion of virus-specific CTLs. The exhaustion of antiviral CTL responses was a stepwise process observed in an overwhelming infection with LCMV Docile or LCMV Clone 13. Following the initial activation, LCMV-specific T cells become anergic for 3 to 5 days and then disappear because of activation-induced cell death (apoptosis). (Of note, the observed lack of T cell functionality was in time of the described experiments termed "anergy"; however, this functional state of T cells was subsequently studied in more detail and shown to be a nonresponsive state after continuous antigen exposure that is now termed "exhausted"; for a detailed discussion, see Wherry and Kurachi [49].) The phenomenology of conventional and exhaustive CTL responses was quantitatively described in the mathematical model by Bocharov [30]. The single characteristic that appeared to be sufficient to control conventional versus exhaustive responses of CTLs was the cumulative viral load since the beginning of the infection $W(t) = \int_0^t V(s)ds$. The increase of W above a certain threshold value in conjunction with the high viral load in the host for about 5 days results in the shift of the infection phenotype from an "acute with recovery" to a chronic infection. The model allowed estimating the fraction of virus population homing to spleen (V_{spleen}) as a saturating function of the inoculum size (IS):

$$V_{\text{Spleen}} = \frac{0.37 \times \text{IS}}{\left(1 + \text{IS}/\left(0.84 \times 10^5\right)\right)}. \quad (2)$$

The model was used to generate biologically relevant predictions amenable to experimental testing: (i) the impact of the precursor CTL number on the dynamics of LCMV infection and (ii) the effect of the virus growth rate on CTL expansion. A bifurcation analysis of the model was used to specify the parametric conditions of low level LCMV persistence after an acute infection [50]. In addition, extensions of the model were used to theoretically examine the efficacy of protection and immunopathology by effector memory versus naive CTLs against intravenous or peripheral infections [31], the role of antigen-specific versus bystander stimulation for persistence, and the structure of CTL memory in LCMV infection [51].

Models of LCMV infections of mice are an example of rival approaches to the description of the exhaustion phenomenon. While the Bocharov models assumed CTL

anergy and apoptosis, two other models used nonoverlapping assumptions: (1) the virus infecting APCs and $CD4^+$ T cells that are later killed by CTLs, thus negatively affecting the clonal expansion loop [45], or (2) the direct competition between the innate immunity and CTLs that is mediated by direct elimination of CTLs by innate immunity and indirect inhibition of CTLs by elimination of the antigen [44]. It is interesting to note that very recent studies demonstrated a direct and an indirect contribution of innate NK cells to T cell exhaustion during primary and chronic infection phases [12, 52–54]. A model refinement based on these new findings seems worthwhile.

The LCMV polymerase is error prone. Escape mutations in the viral envelope of LCMV have been considered as another factor in the establishment of chronic infections when acute CTL responses fail to eliminate the virus. Initially, the physical deletion of CTLs was attributed to the exhaustion phenomenon [21]. Recent studies have suggested that the exhaustion phenotype results from a gradual process and is associated with long-term persistence of CTLs with a reduced functionality, that is, the presence of mono- or bifunctional CTL populations [55]. The overall approach was model driven and focussed on the analysis of the protective efficacy of individual CTL specificities in chronic infection of C57BL/6 mice coinfected with wild type LCMV Cl13 and mutant viruses (CTL epitope mutants: GP33, NP396, and GP276). The mathematical model for the population dynamics of the mutant and wild type (WT) viruses was used to quantify the epitope-specific CTL selection pressure in chronically infected mice. The kinetics of the selection pressure exhibited extensive diversity between individual mice. However, the CTL selection pressure was not lost during the chronic phase of the infection. The early onset of the CTL pressure on the GP276 and GP33 epitopes was documented with the action of GP276 during the first 50 days of persistent infection and the biphasic model of GP33-specific selection pressure. The NP396-specific CTLs were documented to become relevant for selection in a later stage between days 30 and 80 after infection. Interestingly, lack of correlation between the GP276-specific CD8 T cell frequencies in peripheral blood and epitope-specific CTL pressure was observed. Thus, the conventional approach to study the immune correlates of CTL efficacy in terms of avidity, proliferative capacity, perforin expression, resistance to immune regulation, and so forth by monitoring the cells in peripheral blood seems to be not informative enough. Because of the skewing of virus-specific CTLs to LCMV-infected tissues, it is necessary to assess the cell functionality in a tissue-related manner. To overcome the above limitations, novel methodological approaches based upon the analysis of gene expression profiles in conjunction with markers of CTL efficacy, that is, global transcriptome examinations of infected organs, are needed. These issues obviously represent a challenge to mathematicians in terms of the computational tools that are needed for big-data- and multiscale modelling of LCMV-host interaction dynamics quantitated by using modern high-throughput experimental technologies.

4. Model Ranking and Selection

The interaction between a virus and the immune system can be described by multiple mechanisms using various types of modeling formalism (differential equations, cellular automata, and lumped versus spatial considerations). Furthermore, immunological and mathematical knowledge enter models a priori in the form of simplifying assumptions. However, a major limitation is that these assumptions about biological processes are often incompletely understood and the consequences of the necessary simplifications are therefore difficult to predict. Modeling the LCMV infection of mice is an example of the above dilemma.

Three essentially differing mathematical models of the virus-immune response dynamics were developed to explain the phenomenon of CTL exhaustion, reflecting the fact that translation of an immunological phenomenon into a mathematical structure is a nonunique procedure [30, 44, 45]. A computationally intense analysis of various model formulations (30 variants) for the regulation of immune responses in LCMV infection was presented by Rouzine et al. [56]. The ranking was based upon the mean square deviation and the analysis of the Akaike criterion for model and experimental data on LCMV infection in BALB/c and 129/ScEv mice with 4 different strains (Armstrong, Docile, Clone 13, and Aggressive). The best-fit model of the CTL regulation is characterized by a linear control function of the activation process by APCs.

In the process of model development and calibration, some principles have to be considered [57]. A basic requirement is a computational methodology for discriminating between rival models that are constructed from observed data. If there are a number of candidate models, the task is not simply to identify the one with the smallest least squares deviation function but to consider the principle of parsimony for the maximum use of information implicit in the data. If one has confidence in the forms of nested models, one criterion by which to rank them may be the size of the objective function. Model discrimination for nested models is based upon standard hypothesis tests such as the F-test. Following an information-theoretic approach to model building one quantitates the information lost when the model is used to approximate the reality or "full truth." The ranking methodology is that associated with minimum information loss. The latter expression is taken here in terms of the Kullback-Leibler information-theoretic measure of the "distance between" two probabilistic models. It provides a basis for deriving "information-theoretic" criteria such as the Akaike, Schwarz, and Takeuchi indices [58]. The minimal value of the Akaike index suggests that the preferred model ensures a balance between overabundance of parameters (overfitting the data) and sparsity of parameters (underfitting the data). The minimum description length (MDL) provides a selection method that is sensitive to a model's functional form and favors the model that permits the greatest compression of data in its description [59]. Though well established, only three modelling studies of the LCMV infection utilized the information-theoretic approach to model ranking [40, 41, 56, 60].

The mathematical models for the virus-CTL interaction in acute LCMV infections (see Figure 1) can be defined within a set of two- or three-dimensional ordinary differential equations (ODEs) or delay differential equations (DDEs) representing the dynamics of the virus and that of virus-specific CTL (activated and memory cells) populations. It is remarkable that the best-approximating model according to the Akaike criterion for the typical data set of LCMV-CTL population dynamics in primary infections appears to be the one which was introduced elsewhere in an ad hoc manner [33, 37]. Specifically, the parsimonious form of the proliferation term implies that the CTL response to a low-dose LCMV infection is a process regulated by the virus load in an "on" (full activation) and "off" (no activation at all) way.

5. Conclusions

In conclusion, the mathematical modelling approaches to LCMV infection in mice provided rate estimates for T cell turnover, infected target cell elimination, and precursor frequencies but also shed new light on quantitative relationships or "the numbers game" between the virus and the host [28]. This represents the level of virus and cell population dynamics. However, in the immune system, complexity exists on additional levels including the single cell level with tunable responsiveness and the level of the complete host with its anatomical context [48]. With the growing body of high-throughput data from various perturbation experiments at the molecular, the cellular, and the tissue level, the field will be dependent on the ability to develop innovative computational methodologies for data assimilation, analysis, and predictions. A necessary prerequisite is a tight collaboration and mutual understanding between mathematicians and immunologists. There are richness of opportunities and myriads of challenges [61].

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Authors' Contribution

Sections 2, 3, and 4 were done by Gennady Bocharov and Andreas Meyerhans under the support of the Russian Science Foundation (Grant no. 15-11-00029). Sections 1 and 5 were done by Andreas Meyerhans and Jordi Argilaguet under the support of the Spanish Ministry of Economy and Competitiveness and FEDER (Grant no. SAF2013-46077-R).

Acknowledgments

The authors are supported by Grants nos. 15-11-00029 (to Gennady Bocharov and in part Andreas Meyerhans) from the Russian Science Foundation and SAF2013-46077-R (to Andreas Meyerhans and Jordi Argilaguet) from the Spanish Ministry of Economy and Competitiveness and FEDER.

References

[1] P. Klenerman and R. M. Zinkernagel, "What can we learn about human immunodeficiency virus infection from a study of lymphocytic choriomeningitis virus?" *Immunological Reviews*, vol. 159, pp. 5–16, 1997.

[2] X. Zhou, S. Ramachandran, M. Mann, and D. L. Popkin, "Role of lymphocytic choriomeningitis virus (LCMV) in understanding viral immunology: past, present and future," *Viruses*, vol. 4, no. 11, pp. 2650–2669, 2012.

[3] R. M. Zinkernagel, "Lymphocytic choriomeningitis virus and immunology," *Current Topics in Microbiology and Immunology*, vol. 263, pp. 1–5, 2002.

[4] P. C. Doherty and R. M. Zinkernagel, "H 2 compatibility is required for T cell mediated lysis of target cells infected with lymphocytic choriomeningitis virus," *The Journal of Experimental Medicine*, vol. 141, no. 2, pp. 502–507, 1975.

[5] R. M. Zinkernagel and P. C. Doherty, "Restriction of in vitro T cell-mediated cytotoxicity in lymphocytic choriomeningitis within a syngeneic or semiallogeneic system," *Nature*, vol. 248, no. 5450, pp. 701–702, 1974.

[6] D. Kägi, B. Ledermann, K. Bürki et al., "Cytotoxicity mediated by T cells and natural killer cells is greatly impaired in perforin-deficient mice," *Nature*, vol. 369, no. 6475, pp. 31–37, 1994.

[7] D. Masson and J. Tschopp, "Isolation of a lytic, pore-forming protein (perforin) from cytolytic T-lymphocytes," *The Journal of Biological Chemistry*, vol. 260, no. 16, pp. 9069–9072, 1985.

[8] J. N. Blattman, R. Antia, D. J. D. Sourdive et al., "Estimating the precursor frequency of naive antigen-specific CD8 T cells," *Journal of Experimental Medicine*, vol. 195, no. 5, pp. 657–664, 2002.

[9] D. Homann, L. Teyton, and M. B. A. Oldstone, "Differential regulation of antiviral T-cell immunity results in stable $CD8^+$ but declining $CD4^+$ T-cell memory," *Nature Medicine*, vol. 7, no. 8, pp. 913–919, 2001.

[10] S. M. Kaech, E. J. Wherry, and R. Ahmed, "Effector and memory T-cell differentiation: implications for vaccine development," *Nature Reviews Immunology*, vol. 2, no. 4, pp. 251–262, 2002.

[11] S. Ehl, P. Klenerman, R. M. Zinkernagel, and G. Bocharov, "The impact of variation in the number of $CD8^+$T-cell precursors on the outcome of virus infection," *Cellular Immunology*, vol. 189, no. 1, pp. 67–73, 1998.

[12] S. N. Waggoner, M. Cornberg, L. K. Selin, and R. M. Welsh, "Natural killer cells act as rheostats modulating antiviral T cells," *Nature*, vol. 481, no. 7381, pp. 394–398, 2012.

[13] U. Karrer, A. Althage, B. Odermatt et al., "On the key role of secondary lymphoid organs in antiviral immune responses studied in alymphoplastic (*aly/aly*) and spleenless ($Hox11^{-/-}$) mutant mice," *Journal of Experimental Medicine*, vol. 185, no. 12, pp. 2157–2170, 1997.

[14] G. A. Cole, N. Nathanson, and R. A. Prendergast, "Requirement for theta-bearing cells in lymphocytic choriomeningitis virus-induced central nervous system disease," *Nature*, vol. 238, no. 5363, pp. 335–337, 1972.

[15] J. V. Kim, S. S. Kang, M. L. Dustin, and D. B. McGavern, "Myelomonocytic cell recruitment causes fatal CNS vascular injury during acute viral meningitis," *Nature*, vol. 457, no. 7226, pp. 191–195, 2009.

[16] Y. Riviere, I. Cresser, J. C. Guillon, and M. G. Tovey, "Inhibition by anti-interferon serum of lymphocytic choriomeningitis virus disease in suckling mice," *Proceedings of the National Academy*

of Sciences of the United States of America, vol. 74, no. 5, pp. 2135–2139, 1977.

[17] H. D. Chen, A. E. Fraire, I. Joris, M. A. Brehm, R. M. Welsh, and L. K. Selin, "Memory $CD8^+$ T cells in heterologous antiviral immunity and immunopathology in the lung," *Nature Immunology*, vol. 2, no. 11, pp. 1067–1076, 2001.

[18] P. S. Ohashi, S. Gehen, K. Buerki et al., "Ablation of 'tolerance' and induction of diabetes by virus infection in viral antigen transgenic mice," *Cell*, vol. 65, no. 2, pp. 305–317, 1991.

[19] M. B. A. Oldstone, M. Nerenberg, P. Southern, J. Price, and H. Lewicki, "Virus infection triggers insulin-dependent diabetes mellitus in a transgenic model: role of anti-self (virus) immune response," *Cell*, vol. 65, no. 2, pp. 319–331, 1991.

[20] D. L. Barber, E. J. Wherry, D. Masopust et al., "Restoring function in exhausted CD8 T cells during chronic viral infection," *Nature*, vol. 439, no. 7077, pp. 682–687, 2006.

[21] D. Moskophidis, F. Lechner, H. Pircher, and R. M. Zinkernagel, "Virus persistence in acutely infected immunocompetent mice by exhaustion of antiviral cytotoxic effector T cells," *Nature*, vol. 362, no. 6422, pp. 758–761, 1993.

[22] E. J. Wherry, S.-J. Ha, S. M. Kaech et al., "Molecular signature of $CD8^+$ T cell exhaustion during chronic viral infection," *Immunity*, vol. 27, no. 4, pp. 670–684, 2007.

[23] A. J. Zajac, J. N. Blattman, K. Murali-Krishna et al., "Viral immune evasion due to persistence of activated T cells without effector function," *The Journal of Experimental Medicine*, vol. 188, no. 12, pp. 2205–2213, 1998.

[24] O. Leavy, "Tumour immunology: a triple blow for cancer," *Nature Reviews Immunology*, vol. 15, no. 5, pp. 265–265, 2015.

[25] L. Trautmann, L. Janbazian, N. Chomont et al., "Upregulation of PD-1 expression on HIV-specific $CD8^+$ T cells leads to reversible immune dysfunction," *Nature Medicine*, vol. 12, no. 10, pp. 1198–1202, 2006.

[26] V. Velu, K. Titanji, B. Zhu et al., "Enhancing SIV-specific immunity in vivo by PD-1 blockade," *Nature*, vol. 458, no. 7235, pp. 206–210, 2009.

[27] G. Bocharov, V. Chereshnev, I. Gainova et al., "Human immunodeficiency virus infection: from biological observations to mechanistic mathematical modelling," *Mathematical Modelling of Natural Phenomena*, vol. 7, no. 5, pp. 78–104, 2012.

[28] R. M. Zinkernagel, H. Hengartner, and L. Stitz, "On the role of viruses in the evolution of immune responses," *British Medical Bulletin*, vol. 41, no. 1, pp. 92–97, 1985.

[29] U. Assmann-Wischer, D. Moskophidis, M. M. Simon, and F. Lehmann-Grube, "Numbers of cytolytic T lymphocytes (CTL) and CTL precursor cells in spleens of mice acutely infected with lymphocytic choriomeningitis virus," *Medical Microbiology and Immunology*, vol. 175, no. 2-3, pp. 141–143, 1986.

[30] G. A. Bocharov, "Modelling the dynamics of LCMV infection in mice: conventional and exhaustive CTL responses," *Journal of Theoretical Biology*, vol. 192, no. 3, pp. 283–308, 1998.

[31] G. Bocharov, P. Klenerman, and S. Ehl, "Modelling the dynamics of LCMV infection in mice: II. Compartmental structure and immunopathology," *Journal of Theoretical Biology*, vol. 221, no. 3, pp. 349–378, 2003.

[32] M. O. Seedhom, E. R. Jellison, K. A. Daniels, and R. M. Welsh, "High frequencies of virus-specific $CD8^+$ T-cell precursors," *Journal of Virology*, vol. 83, no. 24, pp. 12907–12916, 2009.

[33] R. J. De Boer, D. Homann, and A. S. Perelson, "Different dynamics of $CD4^+$ and $CD8^+$ T cell responses during and after acute lymphocytic choriomeningitis virus infection," *Journal of Immunology*, vol. 171, no. 8, pp. 3928–3935, 2003.

[34] C. L. Althaus, V. V. Ganusov, and R. J. De Boer, "Dynamics of CD^{8+} T cell responses during acute and chronic lymphocytic choriomeningitis virus infection," *Journal of Immunology*, vol. 179, no. 5, pp. 2944–2951, 2007.

[35] B. Ludewig, P. Krebs, T. Junt et al., "Determining control parameters for dendritic cell-cytotoxic T lymphocyte interaction," *European Journal of Immunology*, vol. 34, no. 9, pp. 2407–2418, 2004.

[36] S. Celli, M. Day, A. J. Müller, C. Molina-Paris, G. Lythe, and P. Bousso, "How many dendritic cells are required to initiate a T-cell response?" *Blood*, vol. 120, no. 19, pp. 3945–3948, 2012.

[37] R. J. De Boer, M. Oprea, R. Antia, K. Murali-Krishna, R. Ahmed, and A. S. Perelson, "Recruitment times, proliferation, and apoptosis rates during the $CD8^+$ T-cell response to lymphocytic choriomeningitis virus," *Journal of Virology*, vol. 75, no. 22, pp. 10663–10669, 2001.

[38] W. Barchet, S. Oehen, P. Klenerman et al., "Direct quantitation of rapid elimination of viral antigen-positive lymphocytes by antiviral $CD8^+$ T cells in vivo," *European Journal of Immunology*, vol. 30, no. 5, pp. 1356–1363, 2000.

[39] R. R. Regoes, D. L. Barber, R. Ahmed, and R. Antia, "Estimation of the rate of killing by cytotoxic T lymphocytes in vivo," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 104, no. 5, pp. 1599–1603, 2007.

[40] A. Yates, F. Graw, D. L. Barber, R. Ahmed, R. R. Regoes, and R. Antia, "Revisiting estimates of CTL killing rates in vivo," *PLoS ONE*, vol. 2, no. 12, Article ID e1301, 2007.

[41] V. Garcia, K. Richter, F. Graw, A. Oxenius, and R. R. Regoes, "Estimating the in vivo killing efficacy of cytotoxic T lymphocytes across different Peptide-MHC complex densities," *PLoS Computational Biology*, vol. 11, no. 5, Article ID e1004178, 2015.

[42] F. Graw and R. R. Regoes, "Investigating CTL mediated killing with a 3D cellular automaton," *PLoS Computational Biology*, vol. 5, no. 8, Article ID e1000466, 2009.

[43] V. V. Ganusov, D. L. Barber, and R. J. De Boer, "Killing of targets by CD8+ T cells in the mouse spleen follows the law of mass action," *PLoS ONE*, vol. 6, no. 1, Article ID e15959, 2011.

[44] C. Keşmir and R. J. De Boer, "Clonal exhaustion as a result of immune deviation," *Bulletin of Mathematical Biology*, vol. 65, no. 3, pp. 359–374, 2003.

[45] D. Wodarz, P. Klenerman, and M. A. Nowak, "Dynamics of cytotoxic T-lymphocyte exhaustion," *Proceedings of the Royal Society B: Biological Sciences*, vol. 265, no. 1392, pp. 191–203, 1998.

[46] V. V. Ganusov and R. J. De Boer, "Estimating in vivo death rates of targets due to CD8 T-cell-mediated killing," *Journal of Virology*, vol. 82, no. 23, pp. 11749–11757, 2008.

[47] G. Bocharov, B. Ludewig, A. Bertoletti et al., "Underwhelming the immune response: effect of slow virus growth on $CD8^+$-T-lymphocyte responses," *Journal of Virology*, vol. 78, no. 5, pp. 2247–2254, 2004.

[48] Z. Grossman and W. E. Paul, "Dynamic tuning of lymphocytes: physiological basis, mechanisms, and function," *Annual Review of Immunology*, vol. 33, no. 1, pp. 677–713, 2015.

[49] E. J. Wherry and M. Kurachi, "Molecular and cellular insights into T cell exhaustion," *Nature Reviews Immunology*, vol. 15, no. 8, pp. 486–499, 2015.

[50] T. Luzyanina, K. Engelborghs, S. Ehl, P. Klenerman, and G. Bocharov, "Low level viral persistence after infection with LCMV: a quantitative insight through numerical bifurcation analysis," *Mathematical Biosciences*, vol. 173, no. 1, pp. 1–23, 2001.

[51] G. Bocharov, P. Klenerman, and S. Ehl, "Predicting the dynamics of antiviral cytotoxic T-cell memory in response to different stimuli: cell population structure and protective function," *Immunology and Cell Biology*, vol. 79, no. 1, pp. 74–86, 2001.

[52] P. A. Lang, K. S. Lang, H. C. Xu et al., "Natural killer cell activation enhances immune pathology and promotes chronic infection by limiting $CD8^+$ T-cell immunity," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 4, pp. 1210–1215, 2012.

[53] K. D. Cook and J. K. Whitmire, "The depletion of NK cells prevents T cell exhaustion to efficiently control disseminating virus infection," *Journal of Immunology*, vol. 190, no. 2, pp. 641–649, 2013.

[54] S. N. Waggoner, K. A. Daniels, and R. M. Welsh, "Therapeutic depletion of natural killer cells controls persistent infection," *Journal of Virology*, vol. 88, no. 4, pp. 1953–1960, 2014.

[55] S. Johnson, A. Bergthaler, F. Graw et al., "Protective efficacy of individual CD8+ T cell specificities in chronic viral infection," *The Journal of Immunology*, vol. 194, no. 4, pp. 1755–1762, 2015.

[56] I. M. Rouzine, K. Murali-Krishna, and R. Ahmed, "Generals die in friendly fire, or modeling immune response to HIV," *Journal of Computational and Applied Mathematics*, vol. 184, no. 1, pp. 258–274, 2005.

[57] S. M. Andrew, C. T. H. Baker, and G. A. Bocharov, "Rival approaches to mathematical modelling in immunology," *Journal of Computational and Applied Mathematics*, vol. 205, no. 2, pp. 669–686, 2007.

[58] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference—A Practical Information-Theoretic Approach*, Springer, New York, NY, USA, 2nd edition, 2002.

[59] P. D. Grunwald, I. J. Myung, and M. A. Pitt, *Advances in Minimum Description Length Theory and Applications*, MIT Press, Cambridge, Mass, USA, 2005.

[60] C. T. H. Baker, G. A. Bocharov, J. M. Ford et al., "Computational approaches to parameter estimation and model selection in immunology," *Journal of Computational and Applied Mathematics*, vol. 184, no. 1, pp. 50–76, 2005.

[61] W. E. Paul, "The immune system—complexity exemplified," *Mathematical Modelling of Natural Phenomena*, vol. 7, no. 5, pp. 4–6, 2012.

[62] M. Cornberg, L. L. Kenney, A. T. Chen et al., "Clonal exhaustion as a mechanism to protect against severe immunopathology and death from an overwhelming CD8 T cell response," *Frontiers in Immunology*, vol. 4, article 475, 2013.

Hindawi Publishing Corporation
Journal of Immunology Research
Volume 2015, Article ID 156241, 17 pages
http://dx.doi.org/10.1155/2015/156241

Review Article

Structural and Computational Biology in the Design of Immunogenic Vaccine Antigens

Lassi Liljeroos, Enrico Malito, Ilaria Ferlenghi, and Matthew James Bottomley

Novartis Vaccines & Diagnostics S.r.l. (a GSK Company), Via Fiorentina 1, 53100 Siena, Italy

Correspondence should be addressed to Lassi Liljeroos; lassi.j.liljeroos@gsk.com

Received 29 May 2015; Accepted 2 August 2015

Academic Editor: Guanglan Zhang

Copyright © 2015 Lassi Liljeroos et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Vaccination is historically one of the most important medical interventions for the prevention of infectious disease. Previously, vaccines were typically made of rather crude mixtures of inactivated or attenuated causative agents. However, over the last 10–20 years, several important technological and computational advances have enabled major progress in the discovery and design of potently immunogenic recombinant protein vaccine antigens. Here we discuss three key breakthrough approaches that have potentiated structural and computational vaccine design. Firstly, genomic sciences gave birth to the field of reverse vaccinology, which has enabled the rapid computational identification of potential vaccine antigens. Secondly, major advances in structural biology, experimental epitope mapping, and computational epitope prediction have yielded molecular insights into the immunogenic determinants defining protective antigens, enabling their rational optimization. Thirdly, and most recently, computational approaches have been used to convert this wealth of structural and immunological information into the design of improved vaccine antigens. This review aims to illustrate the growing power of combining sequencing, structural and computational approaches, and we discuss how this may drive the design of novel immunogens suitable for future vaccines urgently needed to increase the global prevention of infectious disease.

1. Introduction

Vaccines are one of the most successful medical interventions in human history and estimated to prevent more than 2.5 million deaths every year [1, 2]. In essence, vaccination is about convincing the immune system to treat a noninfectious artificially introduced substance as an invading pathogen and to raise an immune response that would protect the vaccine from future infection. Vaccination ideally induces an immune response equal to or better than that caused by natural infection. As a result, long-term immunity against a pathogen can be obtained that prevents the individual from disease as well as from transmitting the pathogen thus contributing to the herd protection of the whole society. The history of vaccination is considered to have started with Edward Jenner's experiments in 1796 showing that vaccination with pus from milk maids' blisters caused by cowpox protected humans against smallpox. Since then, the science of vaccines has come a long way, from using inactivated pathogens or toxins and attenuated live pathogens to recombinant subunit and glycoconjugate vaccines, and most recently towards structurally designed epitope-focused vaccines [3].

In its simplest form, effective vaccination can be achieved with inactivated or attenuated pathogens. This has been and still remains the best available solution against many diseases such as measles, mumps, and varicella. Such vaccines have resulted in complete or almost complete eradication of devastating diseases like smallpox and polio. Despite the proven effectiveness in many cases, inactivated pathogens do not always generate adequate protection and attenuated pathogens have safety concerns caused by possible reverse mutations. Further, the logistics of immunizing people in developing countries with live attenuated vaccines is problematic due to the often strict requirements of an uninterrupted cold chain to keep the pathogens alive. Live attenuated vaccines also pose an increased risk for immunocompromised subjects that may not be able to respond adequately to limit the infection.

In the 1970s, glycoconjugate and recombinant subunit vaccines revolutionized the field allowing the development

of safer and more effective vaccines. Glycoconjugate vaccines superseded the previous capsular polysaccharide vaccines, enabling efficient T-cell activation required for long-term immunity. The general mechanism of how the carrier protein helps in T-cell engagement is still, however, unconfirmed [4, 5]. Glycoconjugate vaccines have proven successful against a number of bacterial pathogens such as *Haemophilus influenzae* type B, *Neisseria meningitidis* serotypes A, C, W-135, and Y, and *Streptococcus pneumoniae* [6]. All of the current glycoconjugate vaccines target bacterial pathogens, but the technology is also potentially suitable against viruses like HIV, since their main antigens are highly glycosylated and some of the broadly neutralizing antibodies (bnAbs) have been shown to bind to glycan moieties on the envelope attachment and fusion protein (Env) surface [7].

The advent of recombinant DNA technology in the late 1970s was quickly adopted in the vaccines field. The new techniques enabled heterologous large-scale production of single proteins from pathogens and their modification in order to optimize proteins for vaccine use (e.g., by detoxification of undesirable catalytic activity). Recombinant subunit vaccines initially proved their usefulness against viruses like hepatitis B virus [8, 9] and human papilloma viruses [10] but have since also been used in bacterial vaccines. An example of this is the recently approved 4-component vaccine against *Neisseria meningitidis* serogroup B (MenB) (Bexsero) that is composed of three recombinant proteins and an outer membrane vesicle preparation from the bacteria. This MenB vaccine is also the first vaccine approved for human use for which the starting point of development relied on genomic data and bioinformatics to select the initial pool of antigen candidates by reverse vaccinology (RV) [11].

The great developments in the speed of DNA sequencing and the associated computational methods have enabled large-scale antigen mining by RV and it has already been used for several pathogens [12–17], mainly bacteria, but recently also for herpes simplex virus [18] to find surface expressed or secreted antigen candidates. Initially, candidates are often found to be suboptimal in terms of stability, safety, immunogenicity, or generating broad protection against all strains of the pathogen. Structural vaccinology (SV) is a rational approach that can be used to address these issues. Major aims in SV are the identification of protective B-cell epitopes on the antigens and optimizing the antigens in terms of stability, epitope presentation, ease of production and safety. SV is a symbiosis between experimental methods like X-ray crystallography, electron microscopy and mass spectrometry, and computational methods like structural modeling, computational scaffold design and epitope prediction. Recent breakthrough examples in the fields of respiratory syncytial virus (RSV) [19], human immunodeficiency virus 1 (HIV-1) [20], MenB [21] and group B streptococcus (GBS) [22] indicate that SV has the potential to become another revolutionary step in vaccine development given that many of the important infectious diseases currently not preventable by vaccines are not amenable to traditional approaches. In this review we discuss the experimental and computational aspects of three key modules of a modern vaccine development pipeline: reverse vaccinology, epitope characterization, and structure-based antigen optimization and design. Of the myriad of experimental methods, algorithms, and software developed for these approaches, we highlight the ones we consider of highest practical relevance for vaccine development. We summarize our view of the whole vaccine development process for current and near future vaccines in Figure 1.

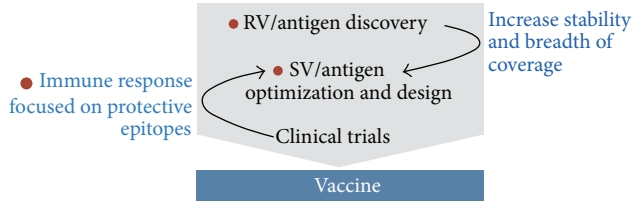


Figure 1: A simplified pipeline for a vaccine development project. Contributions from modern, mainly SV approaches are indicated in blue, while red dots indicate steps aided or made possible by computational methods.

2. Genomic Era, Next-Generation Sequencing and Reverse Vaccinology

Traditionally, vaccine antigen candidates for subunit vaccines have been selected based on experimental data on function, abundance, and immunogenicity. These methods may have overlooked many potentially excellent candidates present with high abundance under the natural conditions during colonization and infection but which may have been present only in lower amounts on the pathogen surface during experimental characterization of antigen expression under laboratory conditions [23]. Initially with shotgun sequencing approach [24] and more recently with next-generation sequencing [25] allowing rapid determination of sequences of whole genomes, vaccine candidate discovery especially on bacterial pathogens has shifted more towards computational prediction of suitable vaccine antigen candidates from genomic sequences using RV. To date, a number of different methodologies have been developed for high-throughput genomic sequencing of which the most commonly used are the sequencing by synthesis (Illumina), ion semiconductor detection (Life Technologies), pyrosequencing (Roche Diagnostics), and more recently single-molecule real time sequencing (PacBio) [26]. Common to most NGS technologies, the output is a set of millions or even billions of short (50–700 bases) sequence reads accompanied by a base-call quality metric. Thus, a major part of any NGS project is assembling the short reads together in a precise and reliable manner. Assembling raw sequencing data of a whole genome is still not trivial and every sequencing technology has its typical types of reads characteristics and sequencing errors that have to be accounted for. A detailed review of assembly methods is beyond the scope of this paper, but we suggest looking into other reviews for a description of the current state of the art [27, 28].

With the number of sequenced bacterial genomes already in the tens of thousands and availability of multiple, for some

bacteria >100, complete genome sequences, it has become possible to use a core set of genes shared by all strains in the RV computational analysis [16, 38]. Alternatively, the availability of multiple genome sequences from one species enables a comparison between the genomes of pathogenic and nonpathogenic strains, which can reveal genes important for pathogenesis that can often be good vaccine antigen candidates [39]. The first step of any RV approach is to predict all ORFs from the genomic sequence and pass them individually through several computational selection filters. In the classical RV approach the features selected are based on the assumption that, in order to be available for interactions with protective antibodies, the antigen has to be either surface-associated or secreted [11]. Features that are typically computationally analyzed include transmembrane domains, leader peptides, homology to known surface proteins, lipoprotein signatures, outer membrane anchoring motifs, and host cell binding motifs such as RGD. The main software that was used in the first RV projects and has retained its popularity is PSORT, now in its third generation for prokaryotic sequences [34]. Since the initial version of PSORT, a number of other software packages for protein cellular localization prediction in bacteria have been published, most using support vector machines on experimental datasets to train the software for prediction [40–44].

Since PSORT remains one of the most widely used subcellular localization prediction software packages in RV, we briefly describe its underlying principles here. PSORT comes with its own database PSORTdb [45] composed of several thousand proteins with experimentally verified subcellular localization that it uses as a reference set for queries. PSORT is a modular program analyzing several features known to be relevant for protein localization like sequence homology to proteins with known localization, signal peptides, amino acid composition, and motifs. In its latest versions PSORTb 2.0 and PSORTb 3.0 the software has also been trained against an extended PSORTdb dataset using an n-peptide composition-based support vector machine, a kernel learning algorithm to improve the percentage of proteins for which a prediction is reported. PSORTb 3.0 covers all prokaryotes, including archea and bacteria with atypical cell wall or membrane topologies, and was reported to predict the subcellular localization with over 95% precision and with a recall of over 90% for both Gram negative and positive test datasets [34].

In general, based on subcellular location only, a large fraction (typically around 30%) of the whole proteome gets selected. In the first applications of RV on MenB an astonishing number of 350 candidates and for GBS 312 candidates were expressed and tested in mice to find promising vaccine antigens [11, 16]. Carrying such a large panel of proteins through the whole workflow of cloning, expression, and purification and above all animal experiments is impractical, which has led to the development of narrowing methods combining computational and experimental, mainly mass spectrometric (MS), methods [46–48]. Identification of an antigen candidate by MS provides proof of expression and can confirm surface localization. Expression can, however, vary in different culture conditions and what is observed *in vitro* may not always be representative of *in vivo* conditions. Several purely computational methods, summarized in Table 1, have also been specifically developed for RV purposes. NERVE added, on top of PSORT subcellular localization analysis, exclusion of multipass membrane proteins and human homologues and positive selection for adhesin-like features, that is, proteins likely to have host cell adhesion functions. Raising antibodies against adhesins can have an inhibitory effect on colonization and infection [33]. Vaxign further added MHC I and MHC II epitope prediction. Jenner-Predict, while using some of the same filtering criteria as NERVE and Vaxign, put more weight on the known host-pathogen interaction domains on proteins [32]. Vacceed framework extended the vaccine antigen prediction also to eukaryotes [35], which are typically much more challenging targets due to their complexity compared to prokaryotes. Also approaches based on existing experimental data on features of known protective antigens have been used in vaccine antigen prediction. VaxiJen uses an alignment-free approach and is based on statistical methods using auto-cross covariance transformation of protein sequences into uniform vectors of principal amino acid properties [37]. For whole proteomes, VaxiJen was, however, reported to identify a set of proteins almost as large as traditional RV approaches [49]. To increase selectivity, another method based on identification of structural and functional features in known bacterial protective antigens and using this data to discover new protective antigens was developed [49]. This method relies on databases and tools such as Pfam and SMART to find features correlated with protectivity of antigens and searches for these features within the protein coding sequences of a particular genome. The program is designed not to take into account any localization signals, enabling recognition of intracellular antigens, the majority of which are T-cell antigens. This "protectome analysis" method was reported to be more selective than VaxiJen leaving only a few dozen candidate antigens identified from a whole bacterial proteome to test experimentally, while still being able to select all the antigens in the MenB (Bexsero) and *Bordetella pertussis* (Daptacel) vaccines.

Once a panel of candidate antigens has been selected, they need to be tested in preclinical animal models. The challenge here is that for many pathogens the correlates of protection are not clear. In other words, the animal experiments may not be reliable indicators of protection in humans, and following wrong or too few metrics, like immunogenicity only, can misguide antigen selection. Nevertheless, the candidate antigens have to fulfill at least three general key features: they have to be immunogenic, they have to be conserved and expressed in natural infection, and they have to be safely tolerated. Since the majority of vaccine-induced protection is generally based on antibodies, the antigen candidate has to elicit measurable, preferentially high antibody titers. Also, if a protein is toxic to experimental animals, the risks of use in humans are too high to consider it as a vaccine antigen as such. In some cases, like that of MenB, at the time of antigen selection it was known that, for protection, antibodies that have complement-mediated bactericidal activity are required, which simplified the selection process [50].

Table 1: Software for antigen discovery.

Software	Vaccine candidate prediction targets			Main selection criteria	Multipass membrane	Host-pathogen	MHC
	Prokaryotes	Archaea	Eukaryotes	Subcellular localization	protein/human homolog detection	interactions	epitopes
Jenner-Predict [32]	+	−	−	+	+	+	−
NERVE [33]	+	−	−	+[3]	+	+	−
PSORTb 3.0 [34]	+	+	−	+[2]	−	−	−
Vacceed [35]	−	−	+	+	+	−	+
Vaxign [36]	+	−	−	+[2]	+	+	+
VaxiJen v2.0[1] [37]	+	−	−	−	−	−	−

[1]VaxiJen uses an alignment-independent approach making it not directly comparable to the other software in the antigen selection criteria.
[2]Using PSORTb 2.0 for subcellular localization prediction.
[3]Using PSORTb 3.0 for subcellular localization prediction.

With many viruses, there are few antigen candidates to choose from and reverse vaccinology is not required. However, surface-exposed viral antigens often exhibit high degrees of sequence variability and conformational heterogeneity, which can make it difficult to produce stable immunogenic conformations and find epitopes conserved across different strains of the viruses. Independently of how one arrives at the selected antigen(s), the next step in understanding the molecular basis of protection and optimizing the antigen is to determine its structure and find the epitopes where protective antibodies bind.

3. Structural Characterization of Antigens and Antigen-Antibody Interactions

From the initial 7 entries deposited in the Protein Data Bank (PDB) when it was established in 1971, the total number of macromolecular structures deposited is currently >100,000, a vast collection of data achieved due to many technological innovations and advances in experimental methods of structure determination. The structures of most vaccine antigens and their epitopes can nowadays be determined and used for advanced, rational, structure-based vaccine design [51]. Knowing the structures of antigens that are candidates to become vaccines enables rational design to fine tune their presentation to the immune system or to facilitate their manufacturing. At the same time, structures of antigen-antibody complexes provide useful information to understand the molecular nature of host-pathogen interactions and of pathogen- or vaccine-induced antibody responses. This information can in turn help to elucidate the effects of immunization and also provides useful knowledge for the development of general principles and computational tools for *in silico* predictions of protective epitopes [52]. Systematic structure-based approaches applied to vaccine research can potentially save time and resources and can also aid the development of vaccines for difficult antigen targets that resist other traditional approaches [53].

3.1. Epitope Mapping and Discovery. Knowing the exact regions of an antigen that are recognized and bound by antibodies provides essential information for antigen engineering and can be used to guide vaccine design and optimization. The experimental methods necessary to obtain such information are collectively called "epitope mapping," and their important role in the early stages of vaccine design has been recognized for several years [54]. Importantly, the postgenomic era and the rapidly increasing number of available structures of antigens and antigen-antibody complexes now open new possibilities for the development of novel tools that can reliably predict protective epitopes. Using either the sequence or the structure of antigens that are target vaccine candidates, these methods could drastically reduce experimental efforts in discovering epitopes needed for the design of vaccines. Below, we will first review recent applications of epitope mapping methods with a focus on the high-resolution mapping by X-ray crystallography and the emerging potential of cryo-EM, and we will then review the current status of computational methods for the prediction and design of B-cell epitopes. Reviews of computational methods for the prediction of T-cell epitopes have been provided previously [55, 56].

3.1.1. Experimental Epitope Mapping. An epitope can be defined as the collection of atoms that directly contact or bind an antibody. Electrostatic attractions, water-mediated hydrogen-bond networks, and long-range forces such as ionic and hydrophobic interactions contribute to the overall binding affinity between an antigen and an antibody [57–60]. In addition, it has been recently suggested that allosteric effects between constant and variable regions of antibodies play a role on antigen affinity or specificity [58, 61]. B-cell epitopes can be either linear (continuous or sequential) or conformational (discontinuous). While linear epitopes are short peptides made of a contiguous amino acid sequence fragment of a protein, conformational epitopes are composed of noncontiguous amino acids (in primary sequence) that are brought into close proximity within the folded 3-dimensional protein structure. It is estimated that most B-cell epitopes (up to 90%) are conformational [62].

An empirical approach to epitope mapping is still essential in order to generate reliable information about antibody-antigen interactions, and many experimental methods of epitope mapping are available. Among those methods that do not require knowledge of the tertiary structure, of either the antigen or the antibody, are (i) the use of synthetic libraries of peptide fragments to scan their binding by an antibody (Pepscan); (ii) the use of bacteriophages (or other organisms) displaying libraries of peptides on their surface to study their binding by antibodies (Phage Display); (iii) various strategies of mass spectrometry (MS) such as epitope extraction, excision, differential chemical modification, and more recently hydrogen-deuterium exchange (H/DX-MS); (iv) solution NMR epitope mapping [63, 64]. In contrast with the information provided by the mainly sequence-based methods cited above, X-ray crystallography delivers a clear visual definition and an atomic-level description of the epitope and paratope atoms forming the antigen-antibody interface. To date there are over 100 nonredundant antibody-antigen (i.e., Fab-protein) complex structures deposited in the PDB, a number that is likely to grow further due to the increasing ease in obtaining and producing human Fabs [65].

An example of epitope mapping approaches has recently been published, showing the importance of employing different methods when a high-resolution picture of the epitope-paratope interface is not available [66]. After failing to obtain crystals of the complex between the programmed death ligand 1 protein (PD-L1) and a monoclonal antibody that targets PD-L1 binding to its receptor, the programmed death protein 1 (PD-1), Hao et al. used limited proteolysis, HDX-MS, mutational studies, and surface plasmon resonance (SPR) to reveal and characterize the epitope.

Another example of interdisciplinary approaches to epitope mapping that instead illustrates the limitations of some methods has been reported for the MenB factor H binding protein (fHbp) and its interaction with a monoclonal antibody (mAb 12C1) [67]. Here, the crystal structure of

the complex fHbp-Fab 12C1 was solved at 1.8 Å resolution, revealing high-resolution details on an extensive epitope-paratope interface involving the variable heavy (VH) and variable light (VL) chains of mAb 12C1 and both the N- and C-terminal domains of fHbp. This interface involves 23 fHbp and 33 Fab residues and generates buried surfaces of ~1000 Å on fHbp and ~880 Å on Fab 12C1. In addition to the crystal structure of the complex, also HDX-MS revealed a large, discontinuous, conformational epitope, though with broader boundaries. Instead, Pepscan and Phage Display identified partial epitopes only and, as expected, of exclusively linear nature. However, these partially mapped and linear regions were also part of the main epitope as identified both by the crystal structure and by HDX-MS [67].

Epitope mapping by X-ray crystallography is probably one of the most powerful and important applications of structural biology in the field of vaccines research. But it is important to recognize that X-ray crystallography cannot be considered a universal solution to epitope mapping, as there are also limitations such as (i) the generation of crystals typically requiring large amounts of sample material, (ii) the lack of certainty that any protein, or antigen-antibody complex, will produce high-quality crystals, and (iii) the unpredictable timelines for the crystallization and structure determination processes that while in the most favourable cases can be very short (days-weeks) or indeed may even never be achieved. As an alternative, for small antigens (<30 kda), solution NMR can also be a rapid and accurate method for epitope mapping and is particularly useful if HSQC peak assignment for the antigen is already available. Examples of NMR epitope mapping include work on MenB and gonococcus fHbp and the DIII domain of dengue virus E protein [68–70].

An emerging method in macromolecular structure determination is cryoelectron microscopy (cryo-EM). Due to the impressive recent progress mainly in electron detectors and software algorithms, it is now possible to determine protein cryo-EM structures to quasiatomic resolution using single-particle methods in 3D reconstruction [71–74]. The great attraction of single-particle cryo-EM compared to X-ray crystallography is that it requires only micrograms of sample and crystallization is not required. Moreover, since images of individual molecules are obtained, computational classification methods can be used to reveal multiple conformational states. Obtaining high-resolution reconstructions (<4 Å) is however greatly facilitated by having a rigid, homogenous complex, preferably several hundreds of kilodaltons in molecular weight. Yet, even low-resolution EM maps can be useful in providing information on the overall architecture of a protein or a protein complex and intermediate-resolution EM maps can already offer insights into the arrangement of domains and localization of functional sites on the macromolecules. Docking of X-ray structures of individual subcomponents in EM maps can be done with high precision, thus increasing the apparent resolvability of the results and making it possible to extract atomic details from the maps. Flexible fitting methodologies can be used to further improve the fitting of X-ray structures into the EM density maps [75]. Recent examples, where X-ray structures have been used to help detailed interpretation of cryo-EM maps, include work on alternative function-related conformational states of a complex, a protein in complex with a cofactor and an antigen-antibody complex [76–78].

Several recent publications illustrate the usefulness of cryo-EM for epitope mapping. Aiyegbo and coworkers described a hybrid method approach for epitope mapping, based on single-particle cryo-EM and enhanced H/DX-MS to determine the location and mode of binding of RV6-25 Fab directed against the VP6 epitope of human *Rotavirus* [79]. Interestingly, the structure of the RV6-25 Fab attached to the double-layered particle (DLP) complex determined by cryo-EM indicated a rather complex binding pattern that revealed differences in accessibility of the VP6 epitope depending on its position in the type I, II, or III channels (located at the icosahedral 5-fold and 3-fold axes of which the former serve as egress points of nascent viral mRNA during viral transcription) (see Figure 2 in [79]). These variations in the accessibility of the RV VP6 capsid layer led to position-specific differences in occupancy for binding of the RV6-25 Fab. A second innovative publication by Bannwarth and collaborators described a new structural approach to characterize in 3D the poliovirus type I epitopes in virus-antibody immune complexes [80]. Briefly, the inactivated polio vaccine (IPV) contains serotypes 1, 2, and 3 of poliovirus. All three serotypes share the D-antigen, which induces protective antibodies. The antigenic structure of PVs is composed by at least four different antigenic sites; thus the D-antigen content results in the combined activity of multiple epitopes. Characterization of the epitopes recognized by the different mAbs was fundamental to map the entire virus surface and ensure the presence of epitopes able to induce neutralizing antibodies. In their new approach the authors describe how combination of single-particle cryo-EM with X-ray crystallographic data allowed the identification of the antigenic sites for these mAbs. The generation and comparison of five different 3D EM maps generated from five different specific Fab-virus and one mAb-virus complex allowed the identification of exposed amino acid residues and finally the mAb-antigen sites. This new approach can be used to map the whole "epitopic" viral surface and provide a comprehensive picture of main epitopes on the surface.

In addition to the examples described above, cryo-EM has successfully been used to map epitopes on several icosahedral viruses, generally difficult to crystallize due to their large size [81, 82]. Cryo-EM also allows characterization of antigen-antibody complexes *in situ*, for example, on enveloped virus surface when tomographic reconstruction methods are used. Such approaches have allowed characterization of influenza HA in complex with a stem-directed mAb showing that the stem is accessible for mAb binding even though the virion surface is densely packed with HA and NA spikes [83].

Fabs have also been used to increase the size of the protein of interest for cryo-EM imaging and reconstruction purposes [84]. Since small (<100 kDa) proteins have up to now been difficult to reconstruct due to lack of evident features and consequent failure in particle alignment, Fabs with their well-known overall structure can be used to aid alignment and validation. Not only does this enable the structure

determination of the protein of interest, but simultaneously produce a structure of the antigen-antibody complex. Thus, antigen-Fab complexes are in fact easier to reconstruct than small antigens alone, which extends the capabilities of cryo-EM to studies of smaller antigens. In a recent example, a Fab from HIV bnAb PGV04 was used to help reconstructing an Env-Fab complex to 5.8 Å resolution [77].

Although X-ray crystallography is still the leading method for high-resolution antigen-antibody interaction characterization, single-particle cryo-EM holds the promise to become complementary in cases where crystallography is not possible or feasible. Furthermore, since cryo-EM can deal with heterogeneous samples, it may in the future become possible to characterize multiple complexes (e.g., from polyclonal sera) within the same sample, an important aspect for throughput and completeness in characterizing the full repertoire of antigen-antibody interactions, rather than just a few mAb-antigen interactions.

3.1.2. Computational Methods for Epitope Prediction and Design. Epitope mapping studies performed over the last few decades have provided a wealth of information on epitope-paratope interfaces that have allowed a certain sophistication in the definition of an epitope [85]. Some common themes of antigen-antibody interactions are starting to emerge, with potential benefits for the development of computational methods for the reliable identification of B-cell epitopes. Also, the constant evolution and refinement of computational methods for protein folding and design, driven by the growth of available protein structures in the PDB, may now further aid in elucidating the molecular bases of antigen-antibody interactions.

Starting in the early 1980s, several sequence- and structure-based B-cell epitope prediction methods have been developed, as extensively reviewed elsewhere [52, 86]. However, developing robust computational methods for epitope prediction has proven to be a very difficult task, and their predictive performance remains far from ideal. Among the possible reasons for the limitations of current epitope prediction tools are the belief that we still possess somewhat weak or wrong hypotheses on the true nature of B-cell epitopes and on the structural bases for the ability of protective antigens to elicit functional antibodies [57, 87]. For example, the evidence that any residue of an antigen may become an epitope under certain circumstances poses a serious complication for most of the prediction methods [88]. However, most of the currently available prediction algorithms try to discriminate between epitopic and nonepitopic antigen surface residues [58]. Another obstacle to obtaining reliable predictions is the lack of large robust benchmark datasets and standard data formats, for which the community has proposed several solutions [88]. In support of the need of more robust benchmarks, the inclusion of additional biological information has been shown to greatly enhance prediction performances [89].

The most common computational methods of epitope prediction are sequence-based, and they specialize on predicting linear or continuous B-cell epitopes. These methods initially utilized sequence profiling by use of amino acid scales, where hydrophobicity, flexibility, solvent accessibility, or other physicochemical properties scales assign a propensity value to each amino acid, thus measuring their tendency to be part of a B-cell epitope [90]. These methods were later exhaustively reviewed and questioned, showing how almost 500 propensity scales performed only slightly better than random and thus leading to the conclusion that they cannot yet be used to predict epitope location reliably [91]. The use of machine-learning, knowledge-based methods was subsequently introduced in order to increase accuracy and reliability of the predictions.

Several lines of evidence indicate that most epitopes are conformational [62], and prediction of discontinuous or conformational epitopes presents further challenges as they require as a prerequisite the antigen structure. Indeed, all discontinuous epitope prediction methods currently available require the 3D structures of the antigen, which may in some cases be very difficult or impossible to obtain [92]. Despite continuous incremental advances in computational tools for the *in silico* prediction of 3D protein structures [93], the prediction of discontinuous epitopes still preferentially requires the experimental 3D structure to increase reliability. Some of the earlier approaches to conformational epitopes prediction used correlations of known epitopes with crystallographic temperature factors, protrusions from the protein globular surface, solvent accessibility, and flexibility. Also, both protein-protein binding site prediction tools [94] and docking algorithms [95] were introduced and tested for epitope prediction.

It has been recently estimated that the accuracy of continuous B-cell epitope predictions methods can reach 60–66% [52]. Importantly, recent computational and experimental validation of both continuous and discontinuous epitope predictions made with the most well established methods currently available show that they all still perform rather poorly [96].

More recent developments of computational methods for epitope prediction include the methods of electrostatic desolvation profiles (EDP) [97] and matrix of local coupling energies (MLCE) [98–100]. Both of these methods aim to elucidate the physicochemical determinants of antibody recognition by an antigen, starting from the structure of the antigen and the *in silico* analyses of its surface properties. This in turn allows making hypotheses on the optimal interface formation for protein-protein complexes in general (EDP) and for antigen-antibody complexes (MLCE). In particular, the MLCE algorithm takes into account both dynamic and energetic properties of a protein surface, looking for sites that because of their intrinsic low-intensity energetic couplings with the rest of the protein will likely undergo conformational changes, as well as mutations with minimal energetic expense, which are both desirable properties for antigenic epitopes and will influence the way an antigen-antibody complex forms. The MLCE does not require previous knowledge on antibody binding or the structure of an antigen-antibody complex, and as such it can be applied to any isolated protein antigen. Once those regions that are minimally coupled to the rest of the protein or the antigen, thus likely involved in antibody recognition, are localized, it is possible to introduce mutations that will increase the affinity for the antibody

and at the same time will not affect the antigen's overall stability. Or it is possible to engineer stable regions of the antigen as more stable or dominant conformation of the one found in the original structural background of the entire antigen. Successful applications of the EDP and MLCE methods have been recently reported, which helped elucidate antigenic regions or immunogenic epitopes of several vaccine target candidates from *Burkholderia pseudomallei* (namely, BPSL1050 [101], the oligopeptide-binding protein A (OppABp) [102], the flagellar hook-associated protein (FlgKBp) [103], and the acute phase antigen BPSL2765 [104]).

A novel antibody-dependent prediction method has been recently introduced that instead of classifying Ag residues a priori as epitopic or nonepitopic will predict the potential match between a given Ab and a given epitope [105]. In addition to the antigen structure, this method also utilizes antibody sequences or structures and thus it promises to bring a new paradigm in epitope prediction by trying to predict which region of an antigen will bind a specific antibody or group of related antibodies, rather than any generic antibody. This concept is similar to the one used for T-cell epitopes prediction and specifically in the assumption of the specific major histocompatibility complex molecule presenting the epitope, rather than the antigen [55]. Combined with the growing field of immunoglobulin repertoire sequencing, this new approach promises to significantly increase the accuracy of B-cell epitope prediction methods. Also, by using antibodies structural or sequence information, this method might help focus the search for epitopes on certain antigenic regions and thus overcome the limitation of the antigen surface as a potentially continuous landscape of epitopes [58]. For example, groups of clonally related antibodies will often bind to the same, or similar, antigenic sites. By determining the high-resolution structures of these antibody-antigen complexes and by analyzing their interaction, a clearer picture of the molecular bases for recognition and binding may emerge, to be then exploited to improve prediction performances [106]. A recent study on the D8 protein of the vaccinia virus (VACV), which is a target of neutralizing Abs elicited by the smallpox vaccine, shows how this computational prediction method performs better than the state-of-the-art prediction methods available [107]. Importantly, in this same study it was shown how a significant increase of the prediction performance can be obtained when combining the antibody-specific predictions with relevant experimental data.

The sections above have described how novel candidate antigens can be discovered initially by sequencing and analyzing the genome of a pathogen and how structural biology—enriched by the combined addition of immunological and epitope mapping data—can yield highly detailed information on the most immunogenic and protective regions of such antigens. The following sections aim to illustrate how this information can drive the design of improved vaccine antigens. In general, antigen optimization strategies can be divided into two main branches, one branch aiming for better antigenic properties in terms of presentation of epitopes that elicit neutralizing or protective antibodies broadly reactive against antigens from multiple strains of the target pathogen; the other branch—equally important—aiming for structural stabilization, homogeneity, ease of production, and safety of the antigen.

4. Overcoming the Challenge of Antigen Sequence Diversity

Many pathogens manage to escape the host immune system by encoding surface-exposed antigens that exhibit high amino acid sequence diversity. The extent of this diversity can range considerably, depending on the pathogen. The "changing face" presented by a variable pathogen represents a challenge for the host immune system and consequently for the successful design of vaccines with broad coverage. One example where this issue has been tackled, with promising preclinical results, regards the antigen fHbp of *Neisseria meningitidis*.

FHbp is a highly immunogenic 28 kDa lipoprotein present on the surface of the majority of meningococcal strains. It was identified both by the computational reverse vaccinology strategy and by more traditional membrane-fractionation methods and is an effective antigen included in two recently licensed vaccines that protect against MenB (Bexsero, Trumenba), as reviewed recently [108]. There are now over 800 unique fHbp peptide sequences deposited in the public *Neisseria* database (http://pubmlst.org/neisseria/fHbp/) [109], hundreds of which are from MenB strains. The latter has implications when considering the design of a protein-based vaccine against MenB; that is, in contrast with the highly successful glycoconjugate vaccines that efficiently target the capsular polysaccharides of serogroup MenA, C, W, and Y [110], a protein-based vaccine should elicit diverse cross-reactive antibodies affording coverage of as many strains as possible, where protein antigens may display extreme sequence diversity. Computational and immunological analyses of these fHbp proteins allowed their grouping into three major sequence variants, which have as little as ~65% sequence identity between variant groups (but typically >90% identity within variant groups) which explains why the different variants are immunologically distinct [111]. In short, it appears that each meningococcal strain can use one of three immunologically different fHbp molecules in order to achieve factor H binding and thus downregulation of the host alternative complement pathway to promote its survival in the blood [112]. This antigenic variability promotes evasion of anti-fHbp directed immune responses, because the different fHbps are not broadly recognized by the antibodies previously induced by antigens of other variant groups.

In order to overcome the meningococcal fHbp antigen sequence diversity, attempts were made to generate a single fHbp antigen capable of eliciting cross-reactive antibodies sufficient to enable bactericidal activity against all meningococcal strains. The three-dimensional (3D) structures of fHbp determined by NMR spectroscopy and later also by X-ray crystallography revealed a molecule composed of two similarly sized domains: an N-terminal taco-shaped beta-barrel and a C-terminal beta-barrel [113–115]. From this structural starting point, Scarselli et al. designed chimeric fHbp molecules, using variant 1.1 fHbp as a scaffold to display surface patches representing epitopes from fHbp variants 2

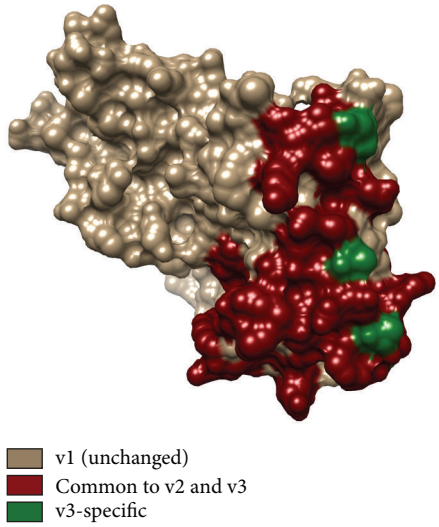


Figure 2: Broad coverage chimeric fHbp generated by rational design as described in [21]. The surface of variant 1 fHbp used as a scaffold is shown in brown. The engineered area carrying heterologous epitopes is colored in dark red (residues in common between variant 2 (V2) and variant 3 (V3)) and in green (variant 3-specific residues).

and 3 (Figure 2) [21]. The size and location of the grafted surface patches were selected and designed on the basis of two key analyses. Firstly, computational analyses of nonredundant antibody-antigen complex structures indicated that the typical epitope-paratope interface involves from 600 to 2000 $Å^2$ surface area on each molecule [85, 116], thus suggesting the need to genetically engineer relatively large new surface patches. Indeed, single amino acids grafted from v2 or v3 were insufficient to elicit cross-reactive antigenicity. Secondly, epitope mapping studies performed using sera obtained from mice immunized with fHbp molecules of each variant group (v1, v2, and v3) suggested that the most immunogenic and protective epitopes of v1, v2, and v3 lay in nonoverlapping regions of the structure—with the C-terminal domain of fHbp harboring most of the v2 and v3 epitopes. Thus, combined sequence- and structure-based inputs led to the computational design of numerous partially overlapping patches each containing a sufficient number of new surface-exposed residues that could potentially form at least one v2- and/or v3-specific conformational epitope on the v1 scaffold. Over 50 mutants were prepared and used to immunize mice. The resulting sera were tested in serum bactericidal assays for their ability to induce complement-mediated killing of MenB strains displaying fHbp molecules of v1, v2, or v3 sequence types. This approach led to the successful design of a novel antigen able to induce broadly protective bactericidal responses against MenB [21].

A somewhat similar approach was recently used by Rippa et al. to graft fHbp v1-specific epitopes onto the gonococcal orthologue of fHbp (Ghfp) [68], which was previously reported to induce strongly bactericidal antibodies against fHbp v2 and v3, but only to a limited extent against fHbp v1 strains of MenB [117]. Immunization of mice with this Ghfp scaffold chimera displaying v1 epitopes was shown to induce bactericidal antibodies against all three fHbp variants, thus demonstrating that combining epitopes from different yet closely related species can be a viable strategy to generate broadly protective antigens [68]. To our knowledge, these two fHbp-centric studies currently represent the only demonstration that the antigenic sequence diversity of a pathogen can be overcome by computational and structure-based design. Clearly, a prerequisite and potential obstacle is detailed knowledge of the 3D antigen structure, though if a reliable template is available it may be possible to start with a homology model generated computationally, for example, using tools, such as I-Tasser or Rosetta, which have performed strongly during recent CASP tests [93]. We anticipate that these proofs of principle will pave the way for similar epitope grafting strategies targeting the variable antigens of alternative pathogens where strain variation causes incomplete protection upon immunization with antigens from one strain only.

5. Enhancing Antigen Homogeneity and Stability

A second area where structural and computational biology have been combined synergistically is in the rational optimization of an antigen that presented the confounding issue of conformational heterogeneity. This approach was illustrated by research performed to identify an effective vaccine antigen to protect against respiratory syncytial virus (RSV). RSV is an important unmet medical need; it is the main viral cause of severe respiratory tract disease in children worldwide, being responsible for approximately 6% of infant deaths [118, 119] and also affecting immunocompromised adults and the elderly [120]. The fusion glycoprotein F is an obvious candidate vaccine antigen, and indeed is the target of a licensed therapeutic mAb (palivizumab, or Synagis) [121]. The F protein is a well-conserved trimeric surface antigen of 150 kDa, but vaccine development was hampered by its conformational variability, typical of viral fusion glycoproteins that undergo large structural changes from prefusion to postfusion states when mediating membrane fusion [122]. In short, while prefusion F might conceptually be the preferable antigen due to its exposure on the virion, it is only metastable as a recombinant protein and converts into a postfusion conformation. In contrast, a simple postfusion antigen was unsuitable, due to its tendency to aggregate, caused by an exposed hydrophobic fusion peptide segment [123].

A promising postfusion F candidate antigen for RSV was rationally designed by removal of the hydrophobic fusion peptide, the transmembrane segment and the cytosolic region. The resulting antigen was readily produced, nonaggregating, homogeneous, highly thermostable and presented the key neutralizing epitopes recognized by palivizumab. Moreover, this postfusion F construct raised high titers of neutralizing antibodies in rodent models of RSV infection, suggesting that it is a promising antigen for clinical trials to protect against RSV [124].

Subsequently, it was reported that much of the neutralizing activity present in human sera after infection targeted the prefusion F state [125]. Design of a stable prefusion F construct was not feasible based on previously known structures but became possible following cocrystallization and structure determination of F in complex with the prefusion-specific human Fab D25 [126]. This antibody-antigen complex provided a platform for the computer-assisted design of point mutations to stabilize the protein in the conformation captured by the antibody. Analysis of Cβ–Cβ bond distances enabled design of a number of Cys substitutions that would introduce disulfide bonds that might covalently lock the F protein in the prefusion conformation (with the most successful pair of mutations being S155C and S290C). In addition, structural analysis revealed a number of sites where amino acid substitutions could enhance protein stability by filling only partially occupied cavities and thus increase hydrophobic packing interactions, in particular the cavity-filling mutation S190F. The designed antigen was further stabilized in its trimeric state by a C-terminal foldon domain, added specifically to ensure stable trimerization. This novel prefusion F candidate was subsequently shown to induce high titers of neutralizing antibodies in rodent and nonhuman primate models [19]. Collectively, the various studies performed to generate both pre- and postfusion F antigen candidates demonstrate how structural studies, computational analyses, and modeling can guide site-directed mutagenesis to generate novel antigens for consideration in RSV vaccine trials.

The HIV-1 Env surface glycoprotein has also been a target for extensive research in the structural vaccinology field. Until recently, the structure of the native prefusion trimer had remained elusive. With the help of an engineered disulfide bond between the GP120 and GP41 subunits and an additional stabilizing mutation required to keep the GP41 in its prefusion conformation, a stable BG505 SOSIP.664 construct was obtained and crystallized and its structure was solved alone and in complex with bnAbs PGT122 and 35O22 [127–129]. Immunization of rabbits and macaques with SOSIP.664 constructs induced autologous neutralizing antibodies, a highly promising sign as the major hurdle in HIV vaccine development has been the inability to induce germline B-cells to mature and mutate to secrete the required bnAbs [130]. Recent breakthrough research indicates that a successful strategy for HIV immunization is likely to be composed of several temporally separate injections of which the first ones contain antigens capable of efficiently stimulating the rare B-cell precursors and subsequent injections containing native-like Env that stimulate the already activated B-cell populations to undergo further somatic mutation thus evolving to bind the native Env on the virion surface [131, 132]. Such a germline-targeting immunogen could be the minimal engineered outer domain (eOD) assembled on nanoparticles developed by Jardine et al. [132] and a native-like Env construct could be that recently developed through structure-based design on the SOSIP.664 background by Do Kwon et al. [20]. Structure-based optimization of a vaccine antigen has also been used for *Borrelia burgdorferi* outer surface protein A [133]. The authors used NMR epitope mapping to reveal that protective epitopes were located exclusively on a C-terminal globular domain. Based on this knowledge, a truncated version of the molecule was designed and produced but found to be unstable and to induce poorer protection in mice compared to the wild-type protein. Replacing some of the charged residues within the core of the domain by hydrophobic residues improved the stability of the construct to levels similar to the wild-type, and likely as a consequence of the increased stability, the construct was found to be equally good as the wild-type in eliciting protective immunity in mice.

6. Optimizing Epitope Presentation

Vaccinating with native antigens is not always optimal and engineered constructs containing only the protective epitopes may perform better in eliciting the optimal immune response. An important example is provided by the influenza haemagglutinin (HA) where immunization with the native protein in seasonal influenza vaccines drives a response mainly directed to the highly variable head region of this antigen resulting in the need to develop a new vaccine almost every year to fight the strains prevalent during a given year [134]. A growing body of evidence suggests that the much less variable stalk region of HA contains neutralizing epitopes, and is therefore a rational point of focus in developing HA antigens [134]. The challenge here is that, in order to direct the immune response to the stalk, the interfering effects of the variable immunodominant head have to be circumvented. The most common approach has been to attempt to make constructs containing only the stalk region. These constructs have had the tendency to be poorly producible in soluble form and to adopt the post-fusion conformation, where neutralizing epitopes are not retained. Recently, however constructs faithfully reproducing the pre-fusion stalk and capable of inducing bnAbs have been reported [135, 136]. Mallajosyula et al. used a computational minimalization approach to design fragment constructs that, basing on interaction network analysis, contained only the residues essential to faithfully reproduce the epitopes. Hydrophobic residues outside the epitope were mutated to prevent aggregation and the fragments connected by flexible linkers and trimerization enhanced by isoleuzine zippers or foldon domains. The obtained antigens were able to elicit bnAbs and confer robust protection against lethal, heterologous viral challenge in mice [136]. Another approach shown to improve the elicited antibody titers to HA, including bnAbs to the stalk, is to express the full-length protein on the surface of ferritin nanoparticles [137]. While none of the HA antigens reported so far can be considered as a universal influenza antigen, the promise is that through further design and engineering an antigen capable of inducing neutralizing antibodies to most if not all influenza strains can be developed.

The examples described above demonstrate how knowledge of an antigen structure and its protective epitopes can be combined to generate novel vaccine candidates with improved characteristics based on closely related predefined scaffolds. However, it is also conceivable to identify protective epitopes that can be targeted by neutralizing antibodies and mount them as conformationally relevant fragments on

non-related scaffold structures, thus enabling new degrees of freedom in antigen design that may overcome issues related to intrinsically problematic behavior of the full-length antigen. For example, as described above, the native RSV F antigen displays properties unsuitable for development as a vaccine antigen (e.g., meta-stability, or tendency for aggregation). Therefore, alternative methods were sought to enable design of novel RSV F conformational epitope presentation strategies, since the known epitope did not elicit neutralizing antibodies when used as a peptide immunogen [138], likely to due to lack of appropriate conformation of the unconstrained peptide. Briefly, the neutralizing epitope targeted by the therapeutic mAb (palivizumab) was characterized by determination of its crystal structure in complex with the Motavizumab Fab—an affinity-enhanced derivative of palivizumab with a picomolar dissociation constant (K_D) for the same epitope in full-length F [29]. The complex structure revealed a highly complementary paratope/epitope interface, with the epitope fragment of 24 residues in a helix-turn-helix conformation making many contacts between the two helices and the Fab. Initial attempts were made to computationally screen all known structures in the Protein Data Bank (PDB) that might be able to host this F-derived helix-turn-helix motif in a conformationally faithful manner, thus enabling epitope presentation on a heterologous scaffold. Indeed, a subset of structures was identified and subsequent grafting of the F epitope into three of these structures was attempted. One of these epitope scaffolds bound Motavizumab with reasonable kinetics (although with an affinity considerably lower than the native F protein). However, when tested in mice, although this immunogen did induce antibodies able to recognize the F antigen, the immune sera lacked RSV neutralizing activity [30]. There was not a clear explanation for the apparent inability to elicit a protective response, which may be linked to the lower affinity observed for Motavizumab binding or to insufficient epitope mimicry or because additional epitopes outside the helix-turn-turn motif are required.

To further develop the epitope scaffold strategy, Correia et al. devised new computational methods to design *de novo* scaffold proteins more ideally suited to display and accurately mimic the RSV neutralizing F epitope [31]. Their method, termed "Fold From Loops (FFL)," has four main steps: (i) selection of the functional motif and target topology to host the motif, (ii) *ab initio* folding to identify suitable main chain structures, (iii) iterations to select the most compatible low-energy side chain solutions, and (iv) automated and human-guided fine tuning to select the best structural candidates. In the specific test case, the latter step involved manual replacement of surface-exposed residues outside the epitope with those from the scaffold template protein and the computational selection of large hydrophobic residues to be inserted within the buried protein core. Importantly, leading designs were biophysically and structurally characterized, and at least eight constructs displayed key signs of being soluble and monomeric, with correct folding and high thermostability ($T_m > 75^{\circ}$C); several also bound to Motavizumab with high affinity (K_D 6–94 pm), suggesting their faithful reproduction of the neutralizing epitope on these scaffolds (K_D for wild-type F glycoprotein was 35 pm), confirmed by crystal structure determinations, and thus representing a major improvement on their previous efforts. Ultimately, the epitope scaffold designs were tested in mice and nonhuman primate animal models. Macaques produced robust binding responses against the autologous antigen scaffolds and RSV F protein, and neutralizing activity was detected in sera in up to 12 of 16 animals. Notably, some of the animals had neutralization titers comparable to those induced by natural human infection. We illustrate the development pathway from the F prefusion Motavizumab epitope to the latest protection-inducing scaffolds of Correia et al. in Figure 3. To summarize, this new structural and computational approach enabled generation of novel epitope scaffolds with robust antigenic properties and presented the F epitope in the desired conformation, as confirmed experimentally by structure determination alone or in complex with Fabs and by immune recognition using sera from RSV-seropositive humans. These studies, which included preclinical experiments in nonhuman primates, ultimately provided a proof of principle for the design of epitope-focusing scaffolds that can successfully elicit neutralizing antibodies against a desired protective epitope. Clearly, this approach could be applied to the design of antigens against a variety of pathogens and could potentially be further developed by the incorporation of multiple epitopes per scaffold, thus increasing breadth of protection elicited by the antigen. Indeed, in the search for potent antigens to protect against HIV, a few promising studies have been performed using scaffolds to stably display portions of gp120 or gp41 [139–141]. Ultimately, the structure-based computational design of epitope scaffolds appears to be a versatile and high-precision approach to vaccine discovery that holds great promise.

7. Conclusion

Structural and computational biology have become important in designing vaccines against diseases unamenable to traditional empirical vaccine development strategies. Computational antigen selection tools are now sophisticated enough to allow a relatively straightforward selection of a limited number of vaccine antigen candidates from whole genomes as a starting point for vaccine development.

However, bioinformatics predictions can fail to correctly identify the posttranslational modifications such as glycosylation, phosphorylation, and molecular rearrangement following proteolytic cleavage that can change the structure and potentially the antigenic properties of bacterial antigens. Integration with proteomics can represent a valid strategy to refine the antigen characterization as well as provide useful insights on abundance and subcellular localization of bacterial antigens [46, 142, 143].

With the increasing speed of X-ray crystallographic structure determination and the promise of cryo-EM in rapid high-resolution structure determination, the number of antigen-antibody complex structures in the PDB is expected to rise at an increasing speed. A significant contribution to this will likely be provided by the fast characterization and cloning of antibodies enabled by B-cell sequencing.

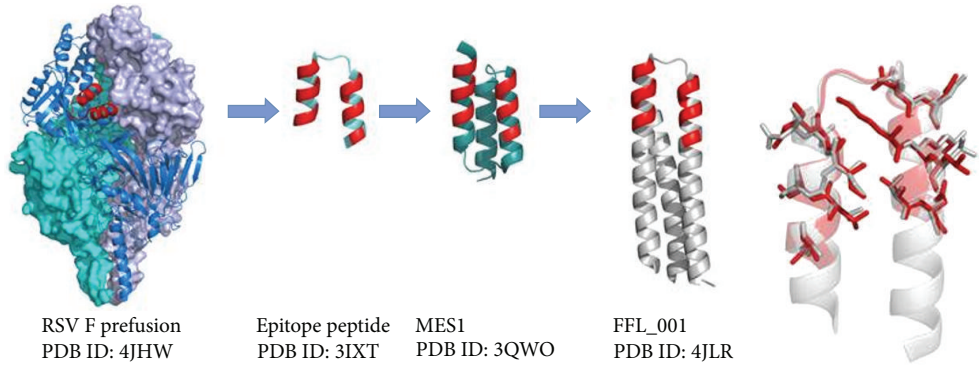


Figure 3: Development of a respiratory syncytial virus (RSV) F scaffold antigen. (a) RSV F Motavizumab epitope is conserved in both pre- and postfusion conformations and a peptide epitope in complex with Motavizumab was shown to have a similar conformation as the epitope in RSV F (PDB ID: 2IXT) [29]. This information was used to design scaffolds presenting the epitope of which the best, MES1 (PDB ID: 3QWO), bound Motavizumab with high affinity but failed to induce protection in mice upon immunization [30] FFL_001 was the first example of a computationally designed scaffold that when used in immunization of macaques induced protection against the virus (PDB ID: 4JLR) [31]. FFL_001 was obtained through a computational *de novo* epitope scaffold design approach called "Fold From Loops" aimed at faithfully reproducing the epitope with strong immunogenic properties. The epitope on RSV F and the residues responsible for interaction with Motavizumab on the antigen constructs are shown in red. (b) Overlay of the RSV F epitope from prefusion F and the epitope region from FFL_001 illustrates the faithful reproduction of the epitope in FFL_001. Epitope residues important for Motavizumab binding from RSV F are shown in red and the corresponding residues from FFL_001 in gray.

The availability of more experimental structures will also help in developing more reliable computational tools for epitope prediction as well as in designing scaffolds for epitope presentation; that is, the more we know, the more we can predict.

Recent developments in epitope-oriented scaffold-based antigen design show great promise but still require additional successful examples to become the norm. A burning question, especially in the case of HIV, is that to what extent the epitope information of bnAbs is useful for vaccination purposes since the germline B-cell receptors that need to first recognize the antigen are very diverse from the bnAbs after somatic hypermutation.

We expect that structural optimization in terms of thermostability, conformational heterogeneity, and safety are likely to show up as the first examples of SV in the pool of vaccines for approved use in human. Indeed, it will be very interesting to see how many of the promising preclinical candidates perform in clinical trials and which will get the privilege to lead the way for other SV-based vaccines of the future.

Conflict of Interests

All authors are employees of Novartis Vaccines & Diagnostics S.r.l. (a GSK company). Bexsero is a product of GlaxoSmithKline.

Acknowledgments

The authors thank Maria Scarselli for critical reading of the paper and constructive comments and Alessandro Muzzi for helpful discussion. Lassi Liljeroos is funded by a European Union FP7 Framework Programme FP7-PEOPLE-2013-IEF Grant (623168).

References

[1] E. De Gregorio and R. Rappuoli, "From empiricism to rational design: a personal perspective of the evolution of vaccine development," *Nature Reviews Immunology*, vol. 14, no. 7, pp. 505–514, 2014.

[2] WHO, *Global Vaccine Action Plan 2011–2020*, WHO, 2013.

[3] R. Cozzi, M. Scarselli, and I. Ferlenghi, "Structural vaccinology: a three-dimensional view for vaccine development," *Current Topics in Medicinal Chemistry*, vol. 13, no. 20, pp. 2629–2637, 2013.

[4] R. Rappuoli and E. De Gregorio, "A sweet T cell response," *Nature Medicine*, vol. 17, no. 12, pp. 1551–1552, 2011.

[5] F. Y. Avci, X. Li, M. Tsuji, and D. L. Kasper, "A mechanism for glycoconjugate vaccine activation of the adaptive immune system and its implications for vaccine design," *Nature Medicine*, vol. 17, no. 12, pp. 1602–1609, 2011.

[6] D. Pace, "Glycoconjugate vaccines," *Expert Opinion on Biological Therapy*, vol. 13, no. 1, pp. 11–33, 2013.

[7] S. Horiya, I. S. MacPherson, and I. J. Krauss, "Recent strategies targeting HIV glycans in vaccine design," *Nature Chemical Biology*, vol. 10, no. 12, pp. 990–999, 2014.

[8] E. Dahl-Hansen, J. Siebke Chr., S. S. Froland, and M. Degre, "Immunogenicity of yeast-derived hepatitis B vaccine from two different producers," *Epidemiology and Infection*, vol. 104, no. 1, pp. 143–149, 1990.

[9] P. Valenzuela, A. Medina, W. J. Rutter, G. Ammerer, and B. D. Hall, "Synthesis and assembly of hepatitis B virus surface antigen particles in yeast," *Nature*, vol. 298, no. 5872, pp. 347–350, 1982.

[10] R. Kirnbauer, F. Booy, N. Cheng, D. R. Lowy, and J. T. Schiller, "Papillomavirus L1 major capsid protein self-assembles into virus-like particles that are highly immunogenic," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 89, no. 24, pp. 12180–12184, 1992.

[11] M. Pizza, V. Scarlato, V. Masignani, and et al, "Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing," *Science*, vol. 287, no. 5459, pp. 1816–1820, 2000.

[12] 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.

[13] A. Naz, F. M. Awan, A. Obaid et al., "Identification of putative vaccine candidates against Helicobacter pylori exploiting exoproteome and secretome: a reverse vaccinology based approach," *Infection, Genetics and Evolution*, vol. 32, pp. 280–291, 2015.

[14] M. H. Chiang, W. C. Sung, S. P. Lien et al., "Identification of novel vaccine candidates against *Acinetobacter baumannii* using reverse vaccinology," *Human Vaccines & Immunotherapeutics*, vol. 11, no. 4, pp. 1065–1073, 2015.

[15] S. Talukdar, S. Zutshi, K. S. Prashanth, K. K. Saikia, and P. Kumar, "Identification of potential vaccine candidates against *Streptococcus pneumoniae* by reverse vaccinology approach," *Applied Biochemistry and Biotechnology*, vol. 172, no. 6, pp. 3026–3041, 2014.

[16] D. Maione, I. Margarit, C. D. Rinaudo et al., "Identification of a universal Group B streptococcus vaccine by multiple genome screen," *Science*, vol. 309, no. 5731, pp. 148–150, 2005.

[17] 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.

[18] Z. Xiang and Y. He, "Genome-wide prediction of vaccine targets for human herpes simplex viruses using Vaxign reverse vaccinology," *BMC Bioinformatics*, vol. 14, supplement 4, article S2, 2013.

[19] J. S. McLellan, M. Chen, M. G. Joyce, and et al, "Structure-based design of a fusion glycoprotein vaccine for respiratory syncytial virus," *Science*, vol. 342, no. 6158, pp. 529–598, 2013.

[20] Y. Do Kwon, M. Pancera, P. Acharya et al., "Crystal structure, conformational fixation and entry-related interactions of mature ligand-free HIV-1 Env," *Nature Structural & Molecular Biology*, vol. 22, no. 7, pp. 522–531, 2015.

[21] M. Scarselli, B. Aricò, B. Brunelli et al., "Rational design of a meningococcal antigen inducing broad protective immunity," *Science Translational Medicine*, vol. 3, no. 91, 2011.

[22] A. Nuccitelli, R. Cozzi, L. J. Gourlay et al., "Structure-based approach to rationally design a chimeric protein for an effective vaccine against Group B *Streptococcus* infections," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 25, pp. 10278–10283, 2011.

[23] L. Fagnocchi, A. Biolchi, F. Ferlicca et al., "Transcriptional regulation of the nada gene in *Neisseria meningitidis* impacts the prediction of coverage of a multicomponent meningococcal serogroup b vaccine," *Infection and Immunity*, vol. 81, no. 2, pp. 560–569, 2013.

[24] E. D. Green, "Strategies for the systematic sequencing of complex genomes," *Nature Reviews Genetics*, vol. 2, no. 8, pp. 573–583, 2001.

[25] J. Shendure and H. Ji, "Next-generation DNA sequencing," *Nature Biotechnology*, vol. 26, no. 10, pp. 1135–1145, 2008.

[26] H. P. J. Buermans and J. T. den Dunnen, "Next generation sequencing technology: advances and applications," *Biochimica et Biophysica Acta—Molecular Basis of Disease*, vol. 1842, no. 10, pp. 1932–1941, 2014.

[27] S. El-Metwally, T. Hamza, M. Zakaria, and M. Helmy, "Next-generation sequence assembly: four stages of data processing and computational challenges," *PLoS Computational Biology*, vol. 9, no. 12, Article ID e1003345, 2013.

[28] M. Howison, F. Zapata, and C. W. Dunn, "Toward a statistically explicit understanding of de novo sequence assembly," *Bioinformatics*, vol. 29, no. 23, pp. 2959–2963, 2013.

[29] J. S. McLellan, M. Chen, A. Kim, Y. Yang, B. S. Graham, and P. D. Kwong, "Structural basis of respiratory syncytial virus neutralization by motavizumab," *Nature Structural and Molecular Biology*, vol. 17, no. 2, pp. 248–250, 2010.

[30] J. S. McLellan, B. E. Correia, M. Chen et al., "Design and characterization of epitope-scaffold immunogens that present the motavizumab epitope from respiratory syncytial virus," *Journal of Molecular Biology*, vol. 409, no. 5, pp. 853–866, 2011.

[31] B. E. Correia, J. T. Bates, R. J. Loomis et al., "Proof of principle for epitope-focused vaccine design," *Nature*, vol. 507, no. 7491, pp. 201–206, 2014.

[32] V. Jaiswal, S. K. Chanumolu, A. Gupta, R. S. Chauhan, and C. Rout, "Jenner-predict server: prediction of protein vaccine candidates (PVCs) in bacteria based on host-pathogen interactions," *BMC Bioinformatics*, vol. 14, no. 1, article 211, 2013.

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

[34] N. Y. Yu, J. R. Wagner, M. R. Laird et al., "PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes," *Bioinformatics*, vol. 26, no. 13, pp. 1608–1615, 2010.

[35] S. J. Goodswen, P. J. Kennedy, and J. T. Ellis, "Vacceed: a high-throughput in silico vaccine candidate discovery pipeline for eukaryotic pathogens based on reverse vaccinology," *Bioinformatics*, vol. 30, no. 16, pp. 2381–2383, 2014.

[36] 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, 15 pages, 2010.

[37] I. A. Doytchinova and D. R. Flower, "VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines," *BMC Bioinformatics*, vol. 8, article 4, 2007.

[38] G. Vernikos, D. Medini, D. R. Riley, and H. Tettelin, "Ten years of pan-genome analyses," *Current Opinion in Microbiology*, vol. 23, pp. 148–154, 2015.

[39] D. G. Moriel, I. Bertoldi, A. Spagnuolo et al., "Identification of protective and broadly conserved vaccine antigens from the genome of extraintestinal pathogenic *Escherichia coli*," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 20, pp. 9072–9077, 2010.

[40] J.-M. Chang, E. C.-Y. Su, A. Lo, H.-S. Chiu, T.-Y. Sung, and W.-L. Hsu, "PSLDoc: protein subcellular localization prediction based on gapped-dipeptides and probabilistic latent semantic analysis," *Proteins: Structure, Function, and Bioinformatics*, vol. 72, no. 2, pp. 693–710, 2008.

[41] S. Matsuda, J.-P. Vert, H. Saigo, N. Ueda, H. Toh, and T. Akutsu, "A novel representation of protein sequences for prediction of subcellular location using support vector machines," *Protein Science*, vol. 14, no. 11, pp. 2804–2813, 2005.

[42] E. C.-Y. Su, H.-S. Chiu, A. Lo, J.-K. Hwang, T.-Y. Sung, and W.-L. Hsu, "Protein subcellular localization prediction based on compartment-specific features and structure conservation," *BMC Bioinformatics*, vol. 8, article 330, 2007.

[43] C.-S. Yu, Y.-C. Chen, C.-H. Lu, and J.-K. Hwang, "Prediction of protein subcellular localization," *Proteins*, vol. 64, no. 3, pp. 643–651, 2006.

[44] M. Zhou, J. Boekhorst, C. Francke, and R. J. Siezen, "LocateP: genome-scale subcellular-location predictor for bacterial proteins," *BMC Bioinformatics*, vol. 9, article 173, 2008.

[45] N. Y. Yu, M. R. Laird, C. Spencer, and F. S. L. Brinkman, "PSORTdb—an expanded, auto-updated, user-friendly protein subcellular localization database for Bacteria and Archaea," *Nucleic Acids Research*, vol. 39, no. 1, pp. D241–D244, 2011.

[46] G. Bensi, M. Mora, G. Tuscano et al., "Multi high-throughput approach for highly selective identification of vaccine candidates: the Group A *Streptococcus* case," *Molecular & Cellular Proteomics*, vol. 11, no. 6, 2012.

[47] F. Doro, S. Liberatori, M. J. Rodríguez-Ortega et al., "Surfome analysis as a fast track to vaccine discovery: identification of a novel protective antigen for Group B *Streptococcus* hypervirulent strain COH1," *Molecular & Cellular Proteomics*, vol. 8, no. 7, pp. 1728–1737, 2009.

[48] M. J. Rodríguez-Ortega, N. Norais, G. Bensi et al., "Characterization and identification of vaccine candidate proteins through analysis of the group A *Streptococcus* surface proteome," *Nature Biotechnology*, vol. 24, no. 2, pp. 191–197, 2006.

[49] E. Altindis, R. Cozzi, B. Di Palo et al., "Protectome analysis: a new selective bioinformatics tool for bacterial vaccine candidate discovery," *Molecular and Cellular Proteomics*, vol. 14, no. 2, pp. 418–429, 2015.

[50] M. C. J. Maiden, J. A. Bygraves, E. Feil et al., "Multilocus sequence typing: a portable approach to the identification of clones within populations of pathogenic microorganisms," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 95, no. 6, pp. 3140–3145, 1998.

[51] P. R. Dormitzer, J. B. Ulmer, and R. Rappuoli, "Structure-based antigen design: a strategy for next generation vaccines," *Trends in Biotechnology*, vol. 26, no. 12, pp. 659–667, 2008.

[52] J. V. Ponomarenko and M. H. Van Regenmortel, "B cell epitope prediction," in *Structural Bioinformatics*, John Wiley & Sons, 2nd edition, 2009.

[53] D. W. Kulp and W. R. Schief, "Advances in structure-based vaccine design," *Current Opinion in Virology*, vol. 3, no. 3, pp. 322–331, 2013.

[54] J. M. Gershoni, A. Roitburd-Berman, D. D. Siman-Tov, N. T. Freund, and Y. Weiss, "Epitope mapping: the first step in developing epitope-based vaccines," *BioDrugs*, vol. 21, no. 3, pp. 145–156, 2007.

[55] D. V. Desai and U. Kulkarni-Kale, "T-cell epitope prediction methods: an overview," *Methods in Molecular Biology*, vol. 1184, pp. 333–364, 2014.

[56] C. Lundegaard, I. Hoof, O. Lund, and M. Nielsen, "State of the art and challenges in sequence based T-cell epitope prediction," *Immunome Research*, vol. 6, supplement 2, article S3, 2010.

[57] M. H. V. Van Regenmortel, "What is a B-cell epitope?" in *Methods in Molecular Biology, Epitope Mapping Protocols*, pp. 3–20, Humana Press, 2nd edition, 2009.

[58] I. Sela-Culang, V. Kunik, and Y. Ofran, "The structural basis of antibody-antigen recognition," *Frontiers in Immunology*, vol. 4, article 302, 2013.

[59] P. Ball, "Water as an active constituent in cell biology," *Chemical Reviews*, vol. 108, no. 1, pp. 74–108, 2008.

[60] C. J. Van Oss, "Hydrophobic, hydrophilic and other interactions in epitope-paratope binding," *Molecular Immunology*, vol. 32, no. 3, pp. 199–211, 1995.

[61] I. Sela-Culang, S. Alon, and Y. Ofran, "A systematic comparison of free and bound antibodies reveals binding-related conformational changes," *The Journal of Immunology*, vol. 189, no. 10, pp. 4890–4899, 2012.

[62] G. Walter, "Production and use of antibodies against synthetic peptides," *Journal of Immunological Methods*, vol. 88, no. 2, pp. 149–161, 1986.

[63] R. C. Ladner, "Mapping the epitopes of antibodies," *Biotechnology and Genetic Engineering Reviews*, vol. 24, pp. 1–30, 2007.

[64] M. Bardelli, E. Livoti, L. Simonelli et al., "Epitope mapping by solution NMR spectroscopy," *Journal of Molecular Recognition*, vol. 28, no. 6, pp. 393–400, 2015.

[65] Y. Zhao, L. Gutshall, H. Jiang et al., "Two routes for production and purification of Fab fragments in biopharmaceutical discovery research: papain digestion of mAb and transient expression in mammalian cells," *Protein Expression and Purification*, vol. 67, no. 2, pp. 182–189, 2009.

[66] G. Hao, J. S. Wesolowski, X. Jiang, S. Lauder, and V. D. Sood, "Epitope characterization of an anti-PD-L1 antibody using orthogonal approaches," *Journal of Molecular Recognition*, vol. 28, no. 4, pp. 269–276, 2015.

[67] E. Malito, A. Faleri, P. L. Surdo et al., "Defining a protective epitope on factor H binding protein, a key meningococcal virulence factor and vaccine antigen," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 9, pp. 3304–3309, 2013.

[68] V. Rippa, L. Santini, P. Lo Surdo et al., "Molecular engineering of Ghfp, the gonococcal orthologue of *Neisseria meningitidis* factor H binding protein," *Clinical and Vaccine Immunology*, vol. 22, no. 7, pp. 769–777, 2015.

[69] M. Scarselli, F. Cantini, L. Santini et al., "Epitope mapping of a bactericidal monoclonal antibody against the factor H binding protein of *Neisseria meningitidis*," *Journal of Molecular Biology*, vol. 386, no. 1, pp. 97–108, 2009.

[70] L. Simonelli, M. Pedotti, M. Beltramello et al., "Rational engineering of a human anti-dengue antibody through experimentally validated computational docking," *PLoS ONE*, vol. 8, no. 2, Article ID e55561, 2013.

[71] S. Asano, Y. Fukuda, F. Beck et al., "Proteasomes. A molecular census of 26S proteasomes in intact neurons," *Science*, vol. 347, no. 6220, pp. 439–442, 2015.

[72] A. Bartesaghi, A. Merk, S. Banerjee et al., "2.2 Å resolution cryo-EM structure of β-galactosidase in complex with a cell-permeant inhibitor," *Science*, vol. 348, no. 6239, pp. 1147–1151, 2015.

[73] Y. Cheng, "Single-particle Cryo-EM at crystallographic resolution," *Cell*, vol. 161, no. 3, pp. 450–457, 2015.

[74] K. Nagayama and R. Danev, "Phase-plate electron microscopy: a novel imaging tool to reveal close-to-life nano-structures," *Biophysical Reviews*, vol. 1, no. 1, pp. 37–42, 2009.

[75] E. Villa and K. Lasker, "Finding the right fit: chiseling structures out of cryo-electron microscopy maps," *Current Opinion in Structural Biology*, vol. 25, pp. 118–125, 2014.

[76] C. Bebeacua, A. Förster, C. McKeown, H. H. Meyer, X. Zhang, and P. S. Freemont, "Distinct conformations of the protein complex p97-Ufd1-Npl4 revealed by electron cryomicroscopy,"

Proceedings of the National Academy of Sciences of the United States of America, vol. 109, no. 4, pp. 1098–1103, 2012.

[77] D. Lyumkis, J.-P. Julien, N. De Val et al., "Cryo-EM structure of a fully glycosylated soluble cleaved HIV-1 envelope trimer," *Science*, vol. 342, no. 6165, pp. 1484–1490, 2013.

[78] D. R. Southworth and D. A. Agard, "Client-loading conformation of the Hsp90 molecular chaperone revealed in the cryo-EM structure of the human Hsp90:Hop complex," *Molecular Cell*, vol. 42, no. 6, pp. 771–781, 2011.

[79] M. S. Aiyegbo, I. M. Eli, B. W. Spiller et al., "Differential accessibility of a rotavirus VP6 epitope in trimers comprising type I, II, or III channels as revealed by binding of a human rotavirus VP6-specific antibody," *Journal of Virology*, vol. 88, no. 1, pp. 469–476, 2014.

[80] L. Bannwarth, Y. Girerd-Chambaz, A. A. Arteni et al., "Structural studies of virus-antibody immune complexes (poliovirus type I): characterization of the epitopes in 3D," *Molecular Immunology*, vol. 63, no. 2, pp. 279–286, 2015.

[81] G. Fibriansah, J. L. Tan, S. A. Smith et al., "A highly potent human antibody neutralizes dengue virus serotype 3 by binding across three surface proteins," *Nature Communications*, vol. 6, p. 6341, 2015.

[82] H. Lee, S. A. Brendle, S. M. Bywaters et al., "A cryo-electron microscopy study identifies the complete H16.V5 epitope and reveals global conformational changes initiated by binding of the neutralizing antibody fragment," *Journal of Virology*, vol. 89, no. 2, pp. 1428–1438, 2015.

[83] A. K. Harris, J. R. Meyerson, Y. Matsuoka et al., "Structure and accessibility of HA trimers on intact 2009 H1N1 pandemic influenza virus to stem region-specific neutralizing antibodies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 12, pp. 4592–4597, 2013.

[84] S. Wu, A. Avila-Sakar, J. Kim et al., "Fabs enable single particle cryoEM studies of small proteins," *Structure*, vol. 20, no. 4, pp. 582–592, 2012.

[85] J. V. Kringelum, M. Nielsen, S. B. Padkjær, and O. Lund, "Structural analysis of B-cell epitopes in antibody: protein complexes," *Molecular Immunology*, vol. 53, no. 1-2, pp. 24–34, 2013.

[86] Y. El-Manzalawy and V. Honavar, "Recent advances in B-cell epitope prediction methods," *Immunome Research*, vol. 6, supplement 2, p. S2, 2010.

[87] L. J. Gourlay, G. Colombo, M. Soriani, G. Grandi, X. Daura, and M. Bolognesi, "Why is a protective antigen protective?" *Human Vaccines*, vol. 5, no. 12, 2009.

[88] 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.

[89] J. V. Kringelum, C. Lundegaard, O. Lund, and M. Nielsen, "Reliable B cell epitope predictions: impacts of method development and improved benchmarking," *PLoS Computational Biology*, vol. 8, no. 12, Article ID e1002829, 2012.

[90] M. H. 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.

[91] 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.

[92] B. Yao, D. Zheng, S. Liang, and C. Zhang, "Conformational B-cell epitope prediction on antigen protein structures: a review of current algorithms and comparison with common binding site prediction methods," *PLoS ONE*, vol. 8, no. 4, Article ID e62249, 2013.

[93] Y. J. Huang, B. Mao, J. M. Aramini, and G. T. Montelione, "Assessment of template-based protein structure predictions in CASP10," *Proteins*, vol. 82, supplement 2, pp. 43–56, 2014.

[94] S. Jones and J. M. Thornton, "Prediction of protein-protein interaction sites using patch analysis," *Journal of Molecular Biology*, vol. 272, no. 1, pp. 133–143, 1997.

[95] I. Halperin, B. Ma, H. Wolfson, and R. Nussinov, "Principles of docking: an overview of search algorithms and a guide to scoring functions," *Proteins*, vol. 47, no. 4, pp. 409–443, 2002.

[96] E. S. Bergmann-Leitner, S. Chaudhury, N. J. Steers et al., "Computational and experimental validation of B and T-cell epitopes of the in vivo immune response to a novel malarial antigen," *PLoS ONE*, vol. 8, no. 8, Article ID e71610, 2013.

[97] S. Fiorucci and M. Zacharias, "Prediction of protein-protein interaction sites using electrostatic desolvation profiles," *Biophysical Journal*, vol. 98, no. 9, pp. 1921–1930, 2010.

[98] C. Peri, P. Gagni, F. Combi et al., "Rational epitope design for protein targeting," *ACS Chemical Biology*, vol. 8, no. 2, pp. 397–404, 2012.

[99] G. Scarabelli, G. Morra, and G. Colombo, "Predicting interaction sites from the energetics of isolated proteins: a new approach to epitope mapping," *Biophysical Journal*, vol. 98, no. 9, pp. 1966–1975, 2010.

[100] E. Malito and R. Rappuoli, "Finding epitopes with computers," *Chemistry and Biology*, vol. 20, no. 10, pp. 1205–1206, 2013.

[101] D. Gaudesi, C. Peri, G. Quilici et al., "Structure-based design of a B cell antigen from *B. pseudomallei*," *ACS Chemical Biology*, vol. 10, no. 3, pp. 803–812, 2015.

[102] P. Lassaux, C. Peri, M. Ferrer-Navarro et al., "A structure-based strategy for epitope discovery in *Burkholderia pseudomallei* OppA antigen," *Structure*, vol. 21, no. 1, pp. 167–175, 2013.

[103] L. J. Gourlay, R. J. Thomas, C. Peri et al., "From crystal structure to in silico epitope discovery in the *Burkholderia pseudomallei* flagellar hook-associated protein FlgK," *FEBS Journal*, vol. 282, no. 7, pp. 1319–1333, 2015.

[104] L. J. Gourlay, C. Peri, M. Ferrer-Navarro et al., "Exploiting the burkholderia pseudomallei acute phase antigen BPSL2765 for structure-based epitope discovery/design in structural vaccinology," *Chemistry and Biology*, vol. 20, no. 9, pp. 1147–1156, 2013.

[105] I. Sela-Culang, Y. Ofran, and B. Peters, "Antibody specific epitope prediction—emergence of a new paradigm," *Current Opinion in Virology*, vol. 11, pp. 98–102, 2015.

[106] I. Sela-Culang, S. Ashkenazi, B. Peters, and Y. Ofran, "PEASE: predicting B-cell epitopes utilizing antibody sequence," *Bioinformatics*, vol. 31, no. 8, pp. 1313–1315, 2015.

[107] I. Sela-Culang, M. R.-E. Benhnia, M. H. Matho et al., "Using a combined computational-experimental approach to predict antibody-specific B cell epitopes," *Structure*, vol. 22, no. 4, pp. 646–657, 2014.

[108] K. L. Seib, M. Scarselli, M. Comanducci, D. Toneatto, and V. Masignani, "Neisseria meningitidis factor H-binding protein fHbp: a key virulence factor and vaccine antigen," *Expert Review of Vaccines*, vol. 14, no. 6, pp. 841–859, 2015.

[109] K. A. Jolley and M. C. J. Maiden, "BIGSdb: scalable analysis of bacterial genome variation at the population level," *BMC Bioinformatics*, vol. 11, article 595, 2010.

[110] D. Pace, "Quadrivalent meningococcal ACYW-135 glycoconjugate vaccine for broader protection from infancy," *Expert Review of Vaccines*, vol. 8, no. 5, pp. 529–542, 2009.

[111] V. Masignani, M. Comanducci, M. M. Giuliani et al., "Vaccination against *Neisseria meningitidis* using three variants of the lipoprotein GNA1870," *Journal of Experimental Medicine*, vol. 197, no. 6, pp. 789–799, 2003.

[112] G. Madico, J. A. Welsch, L. A. Lewis et al., "The meningococcal vaccine candidate GNA1870 binds the complement regulatory protein factor H and enhances serum resistance," *Journal of Immunology*, vol. 177, no. 1, pp. 501–510, 2006.

[113] F. Cantini, D. Veggi, S. Dragonetti et al., "Solution structure of the factor H-binding protein, a survival factor and protective antigen of Neisseria meningitidis," *The Journal of Biological Chemistry*, vol. 284, no. 14, pp. 9022–9026, 2009.

[114] L. Cendron, D. Veggi, E. Girardi, and G. Zanotti, "Structure of the uncomplexed Neisseria meningitidis factor H-binding protein fHbp (rLP2086)," *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, vol. 67, part 5, pp. 531–535, 2011.

[115] A. Mascioni, B. E. Bentley, R. Camarda et al., "Structural basis for the immunogenic properties of the meningococcal vaccine candidate LP2086," *The Journal of Biological Chemistry*, vol. 284, no. 13, pp. 8738–8746, 2009.

[116] N. D. Rubinstein, I. Mayrose, D. Halperin, D. Yekutieli, J. M. Gershoni, and T. Pupko, "Computational characterization of B-cell epitopes," *Molecular Immunology*, vol. 45, no. 12, pp. 3477–3489, 2008.

[117] I. Jongerius, H. Lavender, L. Tan et al., "Distinct binding and immunogenic properties of the gonococcal homologue of meningococcal factor h binding protein," *PLoS Pathogens*, vol. 9, no. 8, Article ID e1003528, 2013.

[118] R. Lozano, M. Naghavi, K. Foreman et al., "Global and regional mortality from 235 causes of death for 20 age groups in 1990 and 2010: a systematic analysis for the Global Burden of Disease Study 2010," *The Lancet*, vol. 380, no. 9859, pp. 2095–2128, 2012.

[119] H. Nair, D. J. Nokes, B. D. Gessner et al., "Global burden of acute lower respiratory infections due to respiratory syncytial virus in young children: a systematic review and meta-analysis," *The Lancet*, vol. 375, no. 9725, pp. 1545–1555, 2010.

[120] A. R. Falsey, P. A. Hennessey, M. A. Formica, C. Cox, and E. E. Walsh, "Respiratory syncytial virus infection in elderly and high-risk adults," *The New England Journal of Medicine*, vol. 352, no. 17, pp. 1749–1759, 2005.

[121] E. M. Connor, "Palivizumab, a humanized respiratory syncytial virus monoclonal antibody, reduces hospitalization from respiratory syncytial virus infection in high-risk infants," *Pediatrics*, vol. 102, no. 3, pp. 531–537, 1998.

[122] J. S. McLellan, W. C. Ray, and M. E. Peeples, "Structure and function of respiratory syncytial virus surface glycoproteins," *Current Topics in Microbiology and Immunology*, vol. 372, pp. 83–104, 2013.

[123] M. B. Ruiz-Argüello, L. González-Reyes, L. J. Calder et al., "Effect of proteolytic processing at two distinct sites on shape and aggregation of an anchorless fusion protein of human respiratory syncytial virus and fate of the intervening segment," *Virology*, vol. 298, no. 2, pp. 317–326, 2002.

[124] K. A. Swanson, E. C. Settembre, C. A. Shaw et al., "Structural basis for immunization with postfusion respiratory syncytial virus fusion F glycoprotein (RSV F) to elicit high neutralizing antibody titers," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 23, pp. 9619–9624, 2011.

[125] M. Magro, V. Mas, K. Chappell et al., "Neutralizing antibodies against the preactive form of respiratory syncytial virus fusion protein offer unique possibilities for clinical intervention," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 109, no. 8, pp. 3089–3094, 2012.

[126] J. S. McLellan, M. Chen, S. Leung et al., "Structure of RSV fusion glycoprotein trimer bound to a prefusion-specific neutralizing antibody," *Science*, vol. 340, no. 6136, pp. 1113–1117, 2013.

[127] J.-P. Julien, A. Cupo, D. Sok et al., "Crystal structure of a soluble cleaved HIV-1 envelope trimer," *Science*, vol. 342, no. 6165, pp. 1477–1483, 2013.

[128] M. Pancera, T. Zhou, A. Druz et al., "Structure and immune recognition of trimeric pre-fusion HIV-1 Env," *Nature*, vol. 514, no. 7523, pp. 455–461, 2014.

[129] R. W. Sanders, R. Derking, A. Cupo et al., "A next-generation cleaved, soluble HIV-1 Env trimer, BG505 SOSIP.664 gp140, expresses multiple epitopes for broadly neutralizing but not non-neutralizing antibodies," *PLoS Pathogens*, vol. 9, no. 9, Article ID e1003618, 2013.

[130] R. W. Sanders, M. J. van Gils, R. Derking et al., "HIV-1 VACCINES. HIV-1 neutralizing antibodies induced by native-like envelope trimers," *Science*, vol. 349, no. 6244, 2015.

[131] P. Dosenovic, L. von Boehmer, A. Escolano et al., "Immunization for HIV-1 broadly neutralizing antibodies in human Ig knockin mice," *Cell*, vol. 161, no. 7, pp. 1505–1515, 2015.

[132] J. Jardine, J.-P. Julien, S. Menis et al., "Rational HIV immunogen design to target specific germline B cell receptors," *Science*, vol. 340, no. 6133, pp. 711–716, 2013.

[133] S. Koide, X. Yang, X. Huang, J. J. Dunn, and B. J. Luft, "Structure-based design of a second-generation Lyme disease vaccine based on a C-terminal fragment of *Borrelia burgdorferi* OspA," *Journal of Molecular Biology*, vol. 350, no. 2, pp. 290–299, 2005.

[134] F. Krammer, P. Palese, and J. Steel, "Advances in universal influenza virus vaccine design and antibody mediated therapies based on conserved regions of the hemagglutinin," *Current Topics in Microbiology and Immunology*, vol. 386, pp. 301–321, 2015.

[135] G. Bommakanti, X. Lu, M. P. Citron et al., "Design of *Escherichia coli*-expressed stalk domain immunogens of H1N1 hemagglutinin that protect mice from lethal challenge," *Journal of Virology*, vol. 86, no. 24, pp. 13434–13444, 2012.

[136] V. V. A. Mallajosyula, M. Citron, F. Ferrara et al., "Influenza hemagglutinin stem-fragment immunogen elicits broadly neutralizing antibodies and confers heterologous protection," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 111, no. 25, pp. E2514–E2523, 2014.

[137] M. Kanekiyo, C.-J. Wei, H. M. Yassine et al., "Self-assembling influenza nanoparticle vaccines elicit broadly neutralizing H1N1 antibodies," *Nature*, vol. 499, no. 7456, pp. 102–106, 2013.

[138] J. A. Lopez, L. D. Andreu, C. Carreno, P. Whyte, G. Taylor, and J. A. Melero, "Conformational constraints of conserved neutralizing epitopes from a major antigenic area of human respiratory syncytial virus fusion glycoprotein," *Journal of General Virology*, vol. 74, no. 12, pp. 2567–2577, 1993.

[139] M. L. Azoitei, B. E. Correia, Y.-E. A. Ban et al., "Computation-guided backbone grafting of a discontinuous motif onto a protein scaffold," *Science*, vol. 334, no. 6054, pp. 373–376, 2011.

[140] B. E. Correia, Y.-E. A. Ban, M. A. Holmes et al., "Computational design of epitope-scaffolds allows induction of antibodies specific for a poorly immunogenic HIV vaccine epitope," *Structure*, vol. 18, no. 9, pp. 1116–1126, 2010.

[141] G. Ofek, F. J. Guenaga, W. R. Schief et al., "Elicitation of structure-specific antibodies by epitope scaffolds," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 107, no. 42, pp. 17880–17887, 2010.

[142] A. Lopez-Campistrous, P. Semchuk, L. Burke et al., "Localization, annotation, and comparison of the *Escherichia coli* K-12 proteome under two states of growth," *Molecular & Cellular Proteomics*, vol. 4, no. 8, pp. 1205–1209, 2005.

[143] 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.

Hindawi Publishing Corporation
Journal of Immunology Research
Volume 2015, Article ID 323506, 9 pages
http://dx.doi.org/10.1155/2015/323506

Research Article

Utilities for High-Throughput Analysis of B-Cell Clonal Lineages

William D. Lees and Adrian J. Shepherd

Institute of Structural and Molecular Biology, Birkbeck College, University of London, Malet Street, London WC1E 7HX, UK

Correspondence should be addressed to William D. Lees; william@lees.org.uk

Received 1 June 2015; Accepted 12 July 2015

Academic Editor: Guanglan Zhang

Copyright © 2015 W. D. Lees and A. J. Shepherd. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

There are at present few tools available to assist with the determination and analysis of B-cell lineage trees from next-generation sequencing data. Here we present two utilities that support automated large-scale analysis and the creation of publication-quality results. The tools are available on the web and are also available for download so that they can be integrated into an automated pipeline. Critically, and in contrast to previously published tools, these utilities can be used with any suitable phylogenetic inference method and with any antibody germline library and hence are species-independent.

1. Introduction

Today it is possible to perform high-throughput sequencing of antibody repertoires at a depth that enables the molecular response to a pathogen to be examined [1, 2]. A key focus is on the identification of clonal lineages of B-cells undergoing the process of somatic hypermutation in germinal centres and the maturation pathways by which these lineages develop over time. It is anticipated that a greater understanding of development pathways will facilitate effective vaccine design for challenging targets such as HIV [3], as well as supporting research into autoimmune disease [4] and immune reactions to therapeutic agents.

B-cell receptor variable regions, which contain the hypervariable complementary-determining regions (CDRs), are encoded by cellular DNA which is transformed in the developing cell through a process of somatic recombination known as junction rearrangement. In the light chain, this process involves the rearrangement of two gene segments, V and J, while, in the heavy chain, three segments, V, D, and J, are rearranged [5, 6]. One source of antibody diversity arises from the selection of V(D)J gene segments from the germline, which contains multiple segments and alleles at different genetic loci, while further diversity arises from the rearrangement process itself, in which gene segments are truncated and additional nucleotides inserted. In a process usually requiring T-cell activation, naive B-cells having affinity to an encountered antigen proliferate and are subjected to somatic hypermutation, in which additional mutations are introduced into the variable region of descendent cells, and mutated descendants binding with higher affinity to the target antigen are selected [7, 8]. The large number of germline gene segments, and the stochastic nature of the gene rearrangement process, makes it unlikely that two cells will develop identical arrangements: the arrangement locus, or junction, shared by all descendants, therefore acts as a unique fingerprint that can be used to trace clonally related sequences through this process of affinity maturation, although the additional mutations generated by the process of somatic hypermutation introduce uncertainty [9].

A number of tools have been developed to identify the germline gene segments and junction rearrangement underlying a particular sequence. IMGT [10], in particular, is widely used for large-scale analysis of NGS-derived repertoires. While only available as an online service, it is capable of analysing sets of up to 500,000 sequences at a time. It is supported by (and can only be used in conjunction with) a curated set of antibody germline libraries, covering a number of commonly used experimental species. In NGS studies, clonally related families are typically identified from the output of such tools by collecting sequences that share descent from the same V and J germline segments and have high junction sequence identity at the nucleotide or amino acid level [11, 12]. D germline ancestry is generally not considered, as the junction D-segment is often < 10 nt and the germline can be difficult to identify categorically.

There are few tools available for the analysis of clonally related lineages, and the majority of studies published to date rely on in-house software. ClonalRelate [13] enables the identification of clonally related families based on junction analysis results from iHMMune-align [14], but the two tools are limited to heavy chain sequences. Vidjil [15] provides an innovative junction analysis that can be used as a prescreening step but does not provide definitive germline attribution. ARPP [16] uses sophisticated phylogenetic techniques to reconstruct a B-cell lineage from a set of clonally related sequences but is restricted to human sequences, employing a germline library that is integrated into the program. IgTree [17] develops lineage trees using a novel algorithm as opposed to traditional phylogenetic methods and is distributed under a restricted licence. Our aim with this toolset is to provide open-source tools that can be combined with any available methods for junction analysis and inference of descent, without constraints in terms of germline usage or species. We foresee them being used both as part of a high-throughput pipeline and in the preparation of accurate and informative figures for publication.

In developing an automated pipeline for large-scale analysis of clonally related lineages, we identified two use cases which were not addressed by available tooling and which were time-consuming (and potentially error-prone) to carry out by hand, even on a small scale. The first is the inference of a germline sequence from a junction analysis, either for rooting a phylogenetic tree or for determining the most likely germline CDR configuration corresponding to an isolated sequence. This requires accurate alignment of the germline V(D)J sequences and appropriate handling of the intervening N- and P-sequences, where, in some cases, one may wish to leave the inference to phylogenetic analysis, while in other cases one may wish to make the best inference possible from available sequence data, possibly taking account of information from a number of related sequences. The second use case concerns the inference of ancestral intermediates from a phylogenetic tree and subsequent reporting. Here a number of packages are available for ancestral reconstruction, for example, in PHYLIP [18], PAML [19], and HyPhy [20]. While these tools provide substantial value in an analysis, their direct use imposes constraints on sequence identifier names which are frequently incompatible with those encountered in real-life examples and do not support the direct generation of phylogenetic trees and other reports which embody standard numbering schemes such as those used by IMGT [21] or by the Protein Data Bank [22] or otherwise embody understanding of CDR locations. Compiling reports on clonal lineages using these tools is therefore likely to require input and output file reformatting, followed by manual cross-referencing and labelling of position identifiers and CDR locations.

Here we present two tools to assist with these use cases. The tools are species-independent in that they can be used with any desired germline library and are available both as online services and as open-source code for integration into a local pipeline. Well-established open-source packages are leveraged for phylogenetic analysis, sequence manipulation, and results presentation. The first tool, *RevertToGermline*, uses a simple technique to infer the ancestral sequence from which a clonally related sequence is derived. The second, *AnnotateTree*, takes a phylogenetic tree rooted on this ancestral sequence and provides annotated trees and alignments showing intermediate sequences and amino acid transitions, based on inferred ancestral states.

2. Materials and Methods

2.1. Algorithms and Functionality. RevertToGermline takes as input a junction analysis of a variable region sequence in which the V(D)J germline gene segments are identified and in which the sequences are divided into regions associated with gene segments and with the intervening spacer regions. In an IMGT analysis, this information is encapsulated in the "nt sequences" section of the analysis, and the tool takes its input in the IMGT format. Use of other junction analysis tools is possible, provided that the output is converted to the simple comma- or tab-separated formats used by IMGT: examples are provided in Supplementary Information available online at http://dx.doi.org/10.1155/2015/323506. The output of RevertToGermline is a sequence in which the V(D)J segments are reverted to germline, while the spacer regions are preserved. A germline library (again in FASTA format, with sequence identifiers in IMGT format) is used to obtain germline segments.

Although they are uncommon, in-frame insertions and deletions can arise in the V-region, when compared to germline. The target V-region is therefore aligned against the germline V-gene at the amino acid level. If the target contains an insertion, the inserted codon is inserted into the derived germline at the same point. If the target contains a deletion, the equivalent codon is deleted from the derived germline. This corresponds to a hypothesis that such insertions and deletions most likely have occurred at the time of junction rearrangement and should therefore be included in the derived germline.

RevertToGermline provides three analysis options, allowing for use in a variety of circumstances (Figure 1). In the first, the germline V-gene is mapped against the input sequence, and remaining nucleotides are gapped. The germline V-gene is trimmed to occupy just that region of the sequence that, according to the junction analysis, is derived from the V-gene in the input sequence. This option provides a convenient root for a phylogenetic tree but yields little information on the likely junction residues of the germline B-cell. The second option maps germline V(D)J sequences in the same manner, putting gaps in just those locations that junction analysis has determined are filled by intervening N and P nucleotides, while the third carries through the N and P nucleotides as well, meaning that there are no gaps in the output sequence. A final option directs RevertToGermline to construct a consensus germline, from germlines inferred for a set of input sequences. Here, in addition to any gaps implied by the above analysis, positions will be gapped if the consensus value is observed in less than 70% of output sequences.

RevertToGermline emits nucleotide sequences that cover whole codons and are aligned on a codon boundary, removing stray nucleotides at the 5′ and 3′ ends. A nucleotide gap in

	V-region	N1	D-region	N2	J-region
Original sequence	...tgtgcga	aagatctgggagaaagggaaaatgaagagtgggcgtcggat	tattacgattttgggagagatta	ccctggccaagacccacggggcgtggttggaagtattgacact	tggggc...
Germline-V	...tgtgcg-	----------	----------	----------	------...
Germline-VDJ	...tgtgcg-	----------	tattacgattttggagtggt--	----------	tggggc...
Germline-full	...tgtgcga	aagatctgggagaaagggaaaatgaagagtgggcgtcggat	tattacgattttggagtggtta	ccctggccaagacccacggggcgtggttggaagtattgacact	tggggc...

Figure 1: RevertToGermline analysis options. The junction decomposition of a representative heavy chain sequence is shown alongside the three inference options available from RevertToGermline. In the first (V), whole codons in the V-region are reverted to the inferred germline, and other regions are gapped out. In the second (VDJ), whole codons in the D- and J-regions are also reverted to their inferred germlines. In the third (full) option, remaining nucleotides are carried into the inferred germline from the original sequence.

any codon position will be extended to cover the entire codon. These steps allow the output to be consumed without further processing by AnnotateTree and other protein-oriented tools. RevertToGermline can conveniently be run against all sequences in a clonal family. Substantial deviations from consensus, as indicated by gaps in the consensus sequence, may indicate a need to apply stricter criteria when identifying the clonal family members, or postrecombinatorial revision [23]. However in our experience the inferred germlines based on V(D)J sequences provide a useful first approximation to the germline, allowing rapid analysis of multiple clonal families and giving a first indication of changes from germline based on the output from an automated pipeline. In many cases, the germline will not be correctly inferable directly from available sequences, but once the universal common ancestor (UCA) of all sequences is available from ancestral reconstruction, anomalies between the UCA and the inferred germline can be investigated.

Current germline library coverage of allelic variants is known to be incomplete [24]. To assist with the identification of variant alleles unrecorded in the germline library, RevertToGermline will optionally report on the presence of "mutated" positions, insertions, and deletions that are observed in all sequences sharing an imputed V germline ancestor, the implication being that these sequences may have descended from a different V germline not present in the germline library. The analysis is conducted for each germline for which a threshold number of sequences are present in the sample: the threshold is user-configurable in the command-line script and set to 20 in the online service.

Having established an initial germline with RevertToGermline, a rooted phylogenetic tree can be inferred using one of the many established packages such as PHYLIP [18] or IQ-TREE [25]. AnnotateTree uses the resulting tree, and the set of clonally related sequences, to perform the following analyses.

2.1.1. Ancestry Reconstruction. Ancestral sequences are inferred by a maximum likelihood method, using PHYLIP's dnaml [18]. AnnotateTree manages the creation of input files for dnaml and presents the results in a convenient form for the user. As dnaml restricts the format and length of sequence names, the names used in the user's input files are mapped to names acceptable to dnaml during processing and mapped back to the user-provided values in output results. The full dnaml report is available for review. An annotated tree, showing the position of each inferred sequence, is produced, together with an amino acid alignment of submitted and inferred sequences. Further trees are produced showing the total number of amino acid changes along each branch and showing the position of inferred intermediate nodes. Output trees are provided as rendered graphics, and also in Newick format (Figure 2). Nucleotide and amino acid sequences are provided in FASTA format for further analysis. dnaml default settings are used, but in the downloadable software the input parameters are exposed in the file dnaml.ctl and can be changed if required.

2.1.2. Tree Annotation. Amino acid substitutions, determined from ancestry reconstruction, are added to the input tree as node labels. The resulting tree is provided as a rendered graphic (in SVG and PNG formats), and also in Newick format, in which annotations are present as node names.

2.1.3. Position Numbering. The position identifiers of amino acids (as used in the alignment and in the labelling of substitutions) can be flexibly defined by the user. The scheme supports both the PDB-style scheme (e.g., 99, 99A, 99B, and 100) and the scheme used by IMGT [21] (e.g., 111, 111.1, 112.1, and 112), in which it will be noted that insertions can precede their ordinal. Deletions are supported in both schemes. To define the scheme for an alignment, the user specifies the following:

(i) The position identifier of the first residue in the sequence.

(ii) The position identifiers of any deletions.

(iii) The position identifiers of any insertions occurring *before* the ordinal position identifier (112.1 in the above examples).

(iv) The position identifiers of any insertions occurring *after* the ordinal position identifier (99A, 99B, and 111.1).

2.1.4. CDR Analysis. If, additionally, the locations of the CDRs are specified, each amino acid location within the CDRs is categorised as follows:

(i) Conserved to germline: the same residue is present at that location in all sequences, including the germline (the first sequence in the submitted sequence file is taken to be the germline).

(ii) Common to trunk: the same residue is present at that location in all sequences except the germline.

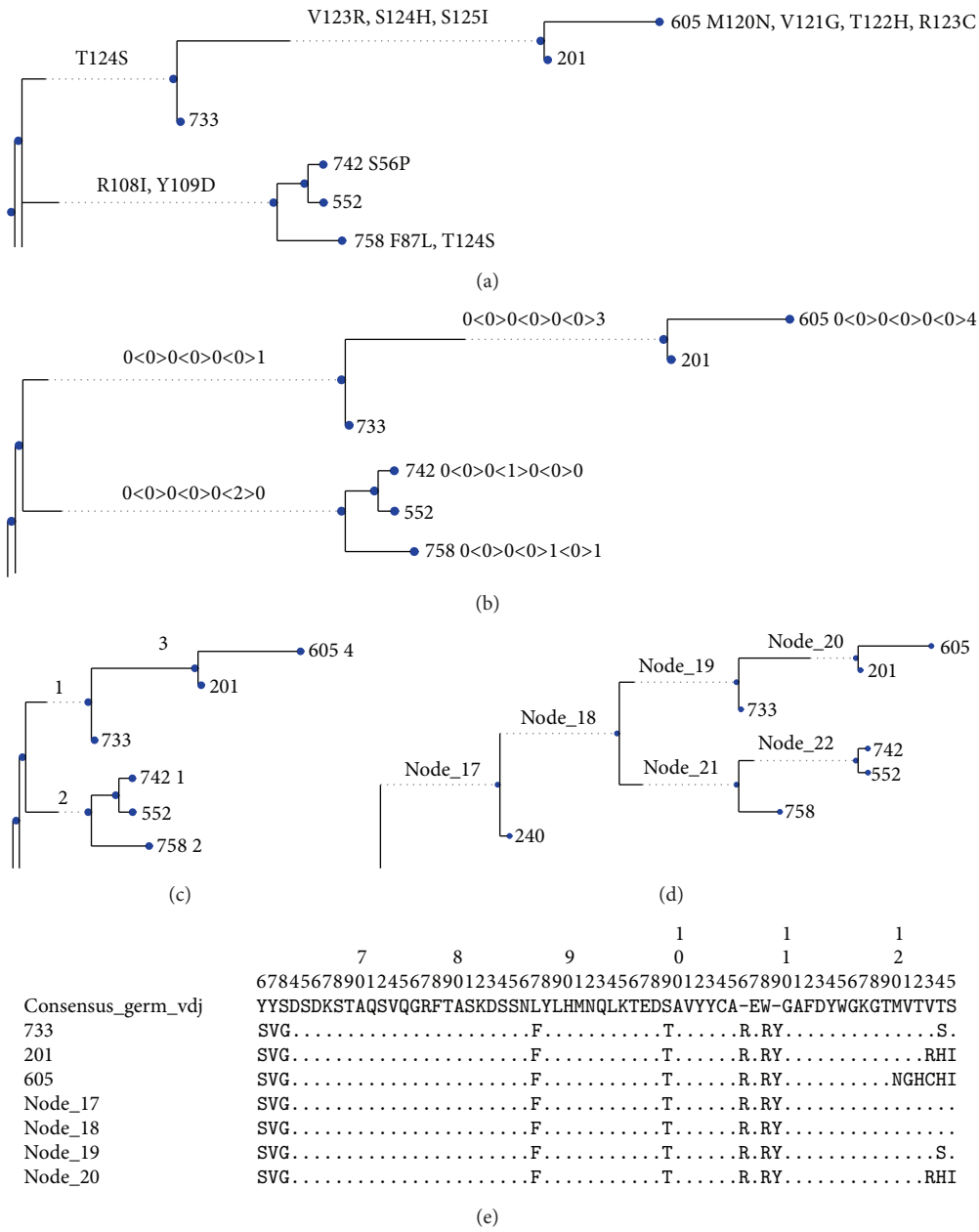


Figure 2: Sample output from AnnotateTree. The samples are taken from the zebrafish repertoire described in more detail in Results and Discussion. (a) Annotation with AA substitutions, using IMGT standard numbering. (b) Summary annotation showing a count of the number of AA substitutions in each FR and CDR using the convention FR1<CDR1>FR2<CDR2>FR3<CDR3>FR4. (c) Annotation showing the total number of AA substitutions on each branch. (d) Tree showing the location of each inferred ancestral node, for cross-reference against AA and nt alignments. (e) Extract from the AA alignment, showing selected nodes from the above trees.

(iii) Variation across samples: differing residues are present at that location across the clonal family.

A further tree is also produced, showing a count of the number of amino acid changes in each CDR and framing region, along each branch of the tree.

2.2. Case Studies. Three case studies, drawing on previously published sequence sets, are used to demonstrate results from the tools. The sequence sets and full results are provided as Supplementary Information.

2.3. PW99 Sequence Set. The PW99 dataset is provided as part of the source distribution of ClonalRelate [13]. Sequences with indeterminate nucleotides were removed. 5′ sequence start positions were aligned by hand and the 3′ end was trimmed to eliminate gaps. This provided 84 distinct sequences with length 375 nucleotides for further analysis.

The sequences were analysed by IMGT to determine CDR positions, and the presence of a uniform junction rearrangement was confirmed by review of the IMGT junction analysis. The V-, D-, and J-segments of all sequences were reverted to germline using RevertToGermline, and the consensus of these sequences was used as the root. The phylogenetic tree was inferred by IQ-TREE v1.2.2 using the K3Pu + G4 substitution model, which was determined to be optimal by the software. AnnotateTree was used to derive the annotated tree and ancestral sequences.

2.4. Zebrafish Repertoire Sequence Set. An NGS repertoire derived from a sampled zebrafish at each of 5 timepoints was downloaded from https://sites.google.com/site/zebrafishdev/files [26] (in each case the sampled fish labelled "A" was chosen). The sequence set consisted of 22,798 annotated heavy chain reads of 224 nt length, spanning the V-D-J junction. The sequences were analysed by IMGT High V-Quest using the IMGT zebrafish germline library. Clonally related families of productive sequences were determined by clustering the junction nucleotide sequences using CD-HIT [27] with parameters that required identical length and >80% sequence identity. This yielded a total of 381 clusters with two or more members, of which the largest had 127 distinct junction sequences. The full database of 22,798 reads was queried for junction sequences from that cluster that occurred at the 2-week timepoint. 84 matching reads were extracted, deduplicated, and trimmed at the 5′ and 3′ ends using HyPhy [20], yielding 59 distinct sequences of length 208 nt. The IMGT junction analysis of this set was reviewed, and 15 sequences which did not match the consensus V-gene IGHV9-2*01 or the consensus J-gene IGHJ2-1*0 were removed, as was one further unproductive sequence. The V-, D-, and J-segments of remaining sequences were reverted to germline using RevertToGermline, and the consensus of these reversions was used as the root. The phylogenetic tree was inferred by IQ-TREE v1.2.2 using the K2P + G4 substitution model, which was determined to be optimal by the software. AnnotateTree was used to derive the annotated tree and ancestral sequences.

2.5. HIV-Neutralizing Antibodies Sequence Set. The heavy chain sequence set and inferred tree were downloaded in Nexus format from the Supplementary Information of the original study [28] and converted to separate FASTA and Newick format files using HyPhy [20] before processing with AnnotateTree. As insertions are developed in the course of the lineage and gaps are not properly handled by PHYLIP dnaml, the UCA inferred by dnaml was compared against the nearest sequence in the phylogenetic tree (038-234314) and matching gaps were created to replace ancestral nucleotides incorrectly inferred by dnaml.

3. Results and Discussion

3.1. VH4-34 Lineage in the Human Tonsil. A characterisation of VH4-34-encoded antibodies isolated from tonsils of healthy human subjects has been previously described [29], and the PW99 dataset, consisting of 99 sequences isolated from a single sample and known to be derived from the same V-D-J rearrangement, has been used in a previous analysis of clonal diversity [13]. A phylogenetic tree inferred by IQ-TREE and annotated by AnnotateTree (Figure 3 and Supplementary Information) shows broad development from the germline, with the absence of CDR-based mutations in the trunk and relatively short development pathways suggesting a repertoire formed by primary rearrangement.

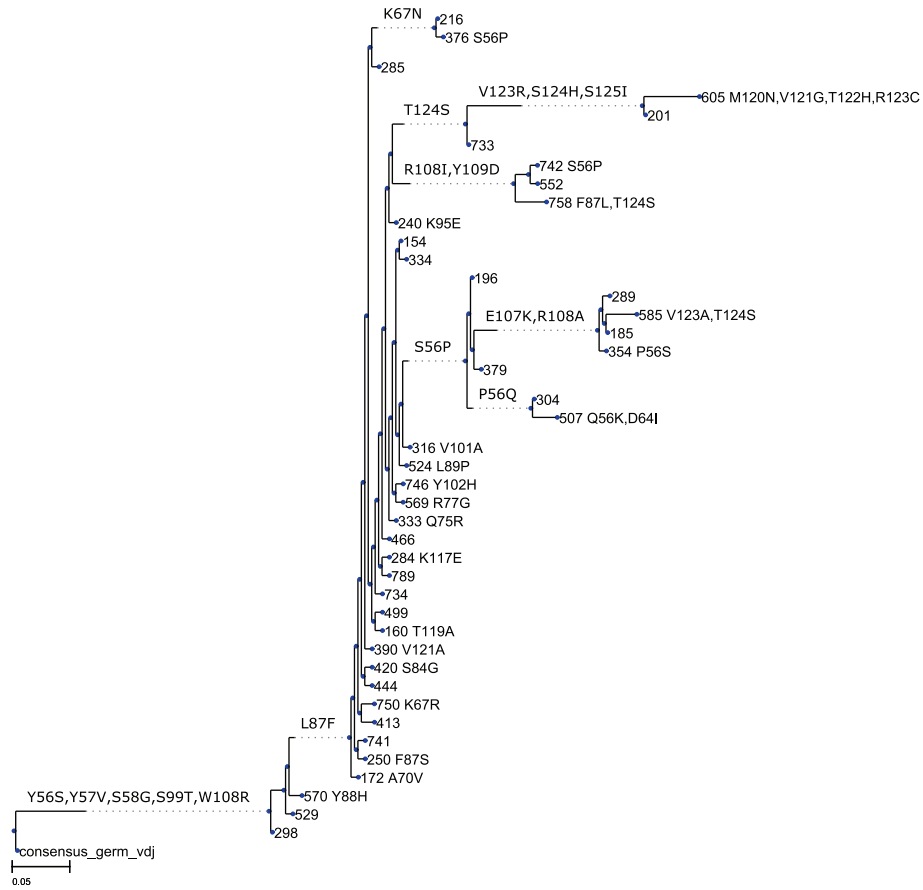
3.2. B-Cell Heavy Chain Development in the Juvenile Zebrafish. In a contrasting study, we isolated and analysed a clonal lineage of 43 partial V-gene sequences isolated from a 2-week-old zebrafish [26]. The pattern is again one of broad development (Figure 3 and Supplementary Information) but the presence of conserved mutations in the trunk is suggestive of more focussed development. The development of substitutions in the framing regions, particularly FR4, compared to the CDRs, is notable.

3.3. Developmental Pathway of HIV-Neutralizing Antibodies. The lineage of neutralizing antibody CAP256-VRC26, which binds to the variable regions 1 and 2 of the HIV-1 envelope, has been described [28]. We analysed the heavy chain lineage sequence set, consisting of 692 sequences, using the inferred tree provided by the authors, which is rooted on the IGHV3-30*18 germline (Figure 3 and Supplementary Information). This sequence set contains samples from 8 timepoints over a 4-year period. A number of insertions can be seen in the amino acid alignment, and it will be noted that insertions are not annotated on the output tree. This is because dnaml treats sequence gaps as unknown nucleotides and therefore does not represent them in intermediate sequences [30]. The UCA inferred by dnaml was corrected to restore the gaps observed in the phylogenetically closest sequence (see Section 2.5). V, D, and J segments were then reverted to germline using RevertToGermline, and the resulting output was observed to be in entire agreement with the UCA, indicating that V, D, and J segments in the UCA represent original germline values and showing that the lineage can be traced back close to the germline, although prior unobserved changes in the N-regions cannot be ruled out.

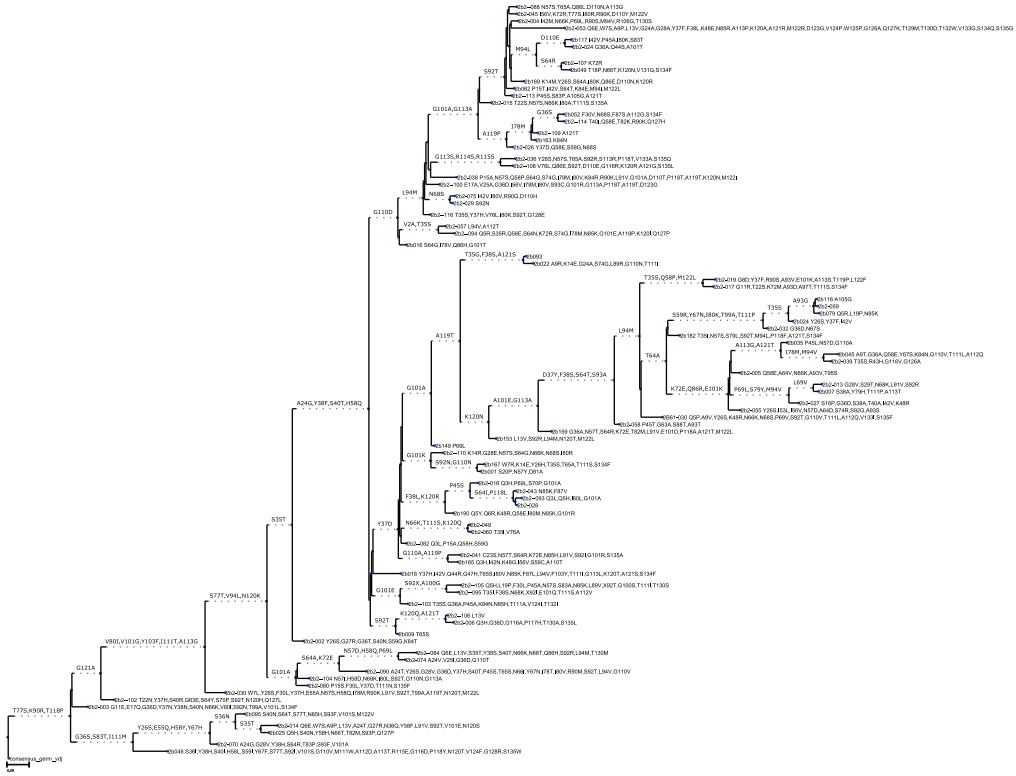
3.4. Availability. Both tools are available online at http://cimm.ismb.lon.ac.uk/pat. The tools are written in Python 2.7 with BioPython [31] and the ETE Toolkit [32]: source code and installation instructions for command-line scripts may be downloaded from the above location.

4. Conclusion

The tools described in this paper were developed to meet the needs of our own work but we are making them publicly available as a small contribution towards the important goal of developing an accepted standard for the analysis of antibody repertoires. Although this work is primarily directed at the analysis of B-cell clonal families, AnnotateTree

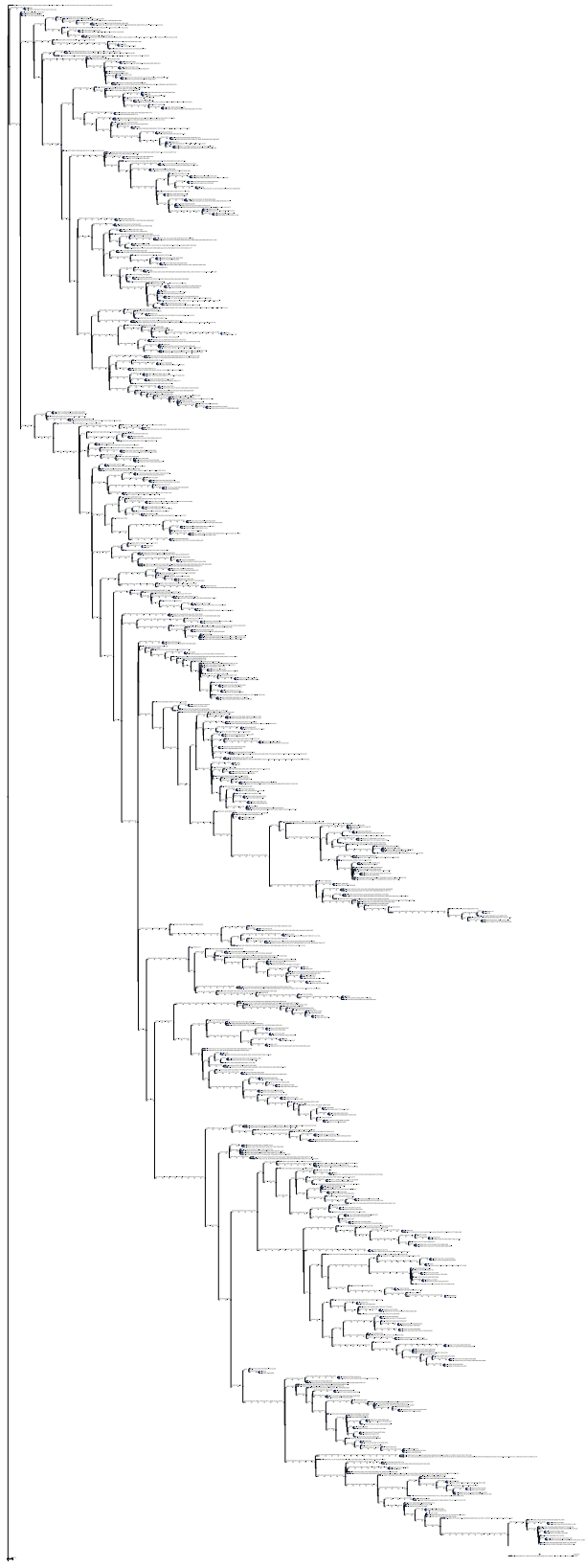


(a)



(b)

Figure 3: Continued.



(c)

Figure 3: Phylogenetic trees for case studies discussed in this paper. (a) VH4-34 lineage in the human tonsil (84 sequences); (b) heavy chain clonal family from a 2-week-old zebrafish (43 sequences); (c) developmental lineage of the HIV bnAb CAP256-VRC26 over 8 timepoints (692 sequences). Larger copies of these trees, plus other output from AnnotateTree, can be found in the Supplementary Information.

may be useful for the creation of alignments, annotations, and ancestral reconstructions of other sequences.

Given the ever-increasing volumes of data, and the increasingly widespread use of NGS by experts in other fields, we feel that it is important to create tools that can be accessed conveniently online for casual use but also installed locally as part of a high-throughput automated pipeline, avoiding the need to work around the volume limitations or queue sizes of online shared services. Given the rapid development of knowledge in the field, toolsets should be as open as possible so that germline libraries and third-party components can be readily updated. Our utilities embody these principles.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

References

[1] G. Georgiou, G. C. Ippolito, J. Beausang, C. E. Busse, H. Wardemann, and S. R. Quake, "The promise and challenge of high-throughput sequencing of the antibody repertoire," *Nature Biotechnology*, vol. 32, no. 2, pp. 158–168, 2014.

[2] N. Fischer, "Sequencing antibody repertoires," *mAbs*, vol. 3, no. 1, pp. 17–20, 2011.

[3] P. D. Kwong, J. R. Mascola, and G. J. Nabel, "Broadly neutralizing antibodies and the search for an HIV-1 vaccine: the end of the beginning," *Nature Reviews Immunology*, vol. 13, no. 9, pp. 693–701, 2013.

[4] R. Mehr, M. Sternberg-Simon, M. Michaeli, and Y. Pickman, "Models and methods for analysis of lymphocyte repertoire generation, development, selection and evolution," *Immunology Letters*, vol. 148, no. 1, pp. 11–22, 2012.

[5] D. G. Schatz and Y. Ji, "Recombination centres and the orchestration of V(D)J recombination," *Nature Reviews Immunology*, vol. 11, no. 4, pp. 251–263, 2011.

[6] D. G. Schatz, "V(D)J recombination," *Immunological Reviews*, vol. 200, no. 1, pp. 5–11, 2004.

[7] K. M. Murphy, *Janeway's Immunobiology*, Garland Science, 8th edition, 2012.

[8] K. Rajewsky, "Clonal selection and learning in the antibody system," *Nature*, vol. 381, no. 6585, pp. 751–758, 1996.

[9] S. D. Boyd, E. L. Marshall, J. D. Merker et al., "Measurement and clinical monitoring of human lymphocyte clonality by massively parallel V-D-J pyrosequencing," *Science Translational Medicine*, vol. 1, no. 12, Article ID 12ra23, 2009.

[10] E. Alamyar, P. Duroux, M.-P. Lefranc, and V. Giudicelli, "IMGT tools for the nucleotide analysis of immunoglobulin (IG) and t cell receptor (TR) V-(D)-J repertoires, polymorphisms, and IG mutations: IMGT/V-QUEST and IMGT/HighV-QUEST for NGS," *Methods in Molecular Biology*, vol. 882, pp. 569–604, 2012.

[11] Y. Wine, D. R. Boutz, J. J. Lavinder et al., "Molecular deconvolution of the monoclonal antibodies that comprise the polyclonal serum response," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 110, no. 8, pp. 2993–2998, 2013.

[12] Y.-C. Wu, D. Kipling, and D. K. Dunn-Walters, "Age-related changes in human peripheral blood IGH repertoire following vaccination," *Frontiers in Immunology*, vol. 3, article 193, 2012.

[13] Z. Chen, A. M. Collins, Y. Wang, and B. A. Gaëta, "Clustering-based identification of clonally-related immunoglobulin gene sequence sets," *Immunome Research*, vol. 6, supplement 1, article S4, 2010.

[14] B. A. Gaëta, H. R. Malming, K. J. L. Jackson, M. E. Bain, P. Wilson, and A. M. Collins, "iHMMune-align: hidden Markov model-based alignment and identification of germline genes in rearranged immunoglobulin gene sequences," *Bioinformatics*, vol. 23, no. 13, pp. 1580–1587, 2007.

[15] M. Giraud, M. Salson, M. Duez et al., "Fast multiclonal clusterization of V(D)J recombinations from high-throughput sequencing," *BMC Genomics*, vol. 15, no. 1, article 409, 2014.

[16] T. B. Kepler, "Reconstructing a B-cell clonal lineage. I. Statistical inference of unobserved ancestors," *F1000Research*, vol. 2, article 103, 2013.

[17] M. Barak, N. S. Zuckerman, H. Edelman, R. Unger, and R. Mehr, "IgTree: creating immunoglobulin variable region gene lineage trees," *Journal of Immunological Methods*, vol. 338, no. 1-2, pp. 67–74, 2008.

[18] J. Felsenstein and G. A. Churchill, "A Hidden Markov Model approach to variation among sites in rate of evolution," *Molecular Biology and Evolution*, vol. 13, no. 1, pp. 93–104, 1996.

[19] Z. Yang, "PAML: a program package for phylogenetic analysis by maximum likelihood," *Computer Applications in the Biosciences*, vol. 13, no. 5, pp. 555–556, 1997.

[20] S. L. K. Pond, S. D. W. Frost, and S. V. Muse, "HyPhy: hypothesis testing using phylogenies," *Bioinformatics*, vol. 21, no. 5, pp. 676–679, 2005.

[21] M.-P. Lefranc, C. Pommié, M. Ruiz et al., "IMGT unique numbering for immunoglobulin and T cell receptor variable domains and Ig superfamily V-like domains," *Developmental and Comparative Immunology*, vol. 27, no. 1, pp. 55–77, 2003.

[22] H. M. Berman, T. Battistuz, T. N. Bhat et al., "The protein data bank," *Acta Crystallographica D: Biological Crystallography*, vol. 58, no. 6, pp. 899–907, 2002.

[23] P. C. Wilson, K. Wilson, Y.-J. Liu, J. Banchereau, V. Pascual, and J. D. Capra, "Receptor revision of immunoglobulin heavy chain variable region genes in normal human B lymphocytes," *The Journal of Experimental Medicine*, vol. 191, no. 11, pp. 1881–1894, 2000.

[24] D. Gadala-Maria, G. Yaari, M. Uduman, and S. H. Kleinstein, "Automated analysis of high-throughput B-cell sequencing data reveals a high frequency of novel immunoglobulin V gene segment alleles," *Proceedings of the National Academy of Sciences*, vol. 112, no. 8, pp. E862–E870, 2015.

[25] B. Q. Minh, M. A. T. Nguyen, and A. von Haeseler, "Ultrafast approximation for phylogenetic bootstrap," *Molecular Biology and Evolution*, vol. 30, no. 5, pp. 1188–1195, 2013.

[26] N. Jiang, J. A. Weinstein, L. Penland, R. A. White, D. S. Fisherd, and S. R. Quake, "Determinism and stochasticity during maturation of the zebrafish antibody repertoire," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 108, no. 13, pp. 5348–5353, 2011.

[27] L. Fu, B. Niu, Z. Zhu, S. Wu, and W. Li, "CD-HIT: accelerated for clustering the next-generation sequencing data," *Bioinformatics*, vol. 28, no. 23, pp. 3150–3152, 2012.

[28] N. A. Doria-Rose, C. A. Schramm, J. Gorman et al., "Developmental pathway for potent V1V2-directed HIV-neutralizing antibodies," *Nature*, vol. 509, no. 7498, pp. 55–62, 2014.

[29] N.-Y. Zheng, K. Wilson, X. Wang et al., "Human immunoglobulin selection associated with class switch and possible tolerogenic origins for C delta class-switched B cells," *The Journal of Clinical Investigation*, vol. 113, pp. 1188–1201, 2004.

[30] *Frequently Asked Questions: What Do I Do about Deletions and Insertions in My Sequences?*, 2015, http://evolution.genetics.washington.edu/phylip/faq.html#indels.

[31] P. J. A. Cock, T. Antao, J. T. Chang et al., "Biopython: freely available Python tools for computational molecular biology and bioinformatics," *Bioinformatics*, vol. 25, no. 11, pp. 1422–1423, 2009.

[32] J. Huerta-Cepas, J. Dopazo, and T. Gabaldón, "ETE: a python environment for tree exploration," *BMC Bioinformatics*, vol. 11, article 24, 2010.

Hindawi Publishing Corporation
Journal of Immunology Research
Volume 2015, Article ID 380975, 11 pages
http://dx.doi.org/10.1155/2015/380975

Research Article

FluKB: A Knowledge-Based System for Influenza Vaccine Target Discovery and Analysis of the Immunological Properties of Influenza Viruses

Christian Simon,[1,2] Ulrich J. Kudahl,[1,3] Jing Sun,[3] Lars Rønn Olsen,[3,4] Guang Lan Zhang,[3,5,6] Ellis L. Reinherz,[3,6,7] and Vladimir Brusic[3,5,6]

[1]*Center for Biological Sequence Analysis, Department of Systems Biology, Technical University of Denmark, 2800 Lyngby, Denmark*
[2]*Department of Disease Systems Biology, Novo Nordisk Foundation Center for Protein Research, University of Copenhagen, 2200 Copenhagen, Denmark*
[3]*Cancer Vaccine Center, Dana-Farber Cancer Institute, Harvard Medical School, Boston, MA 02115, USA*
[4]*Bioinformatics Centre, Department of Biology, University of Copenhagen, 1017 Copenhagen, Denmark*
[5]*Department of Computer Science, Metropolitan College, Boston University, Boston, MA 02215, USA*
[6]*Department of Medicine, Harvard Medical School, Boston, MA 02115, USA*
[7]*Laboratory of Immunobiology, Department of Medical Oncology, Dana-Farber Cancer Institute, Boston, MA 02115, USA*

Correspondence should be addressed to Vladimir Brusic; vladimir.brusic@nu.edu.kz

Received 16 January 2015; Accepted 12 March 2015

Academic Editor: Peirong Jiao

Copyright © 2015 Christian Simon et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

FluKB is a knowledge-based system focusing on data and analytical tools for influenza vaccine discovery. The main goal of FluKB is to provide access to curated influenza sequence and epitope data and enhance the analysis of influenza sequence diversity and the analysis of targets of immune responses. FluKB consists of more than 400,000 influenza protein sequences, known epitope data (357 verified T-cell epitopes, 685 HLA binders, and 16 naturally processed MHC ligands), and a collection of 28 influenza antibodies and their structurally defined B-cell epitopes. FluKB was built using a modular framework allowing the implementation of analytical workflows and includes standard search tools, such as keyword search and sequence similarity queries, as well as advanced tools for the analysis of sequence variability. The advanced analytical tools for vaccine discovery include visual mapping of T- and B-cell vaccine targets and assessment of neutralizing antibody coverage. FluKB supports the discovery of vaccine targets and the analysis of viral diversity and its implications for vaccine discovery as well as potential T-cell breadth and antibody cross neutralization involving multiple strains. FluKB is representation of a new generation of databases that integrates data, analytical tools, and analytical workflows that enable comprehensive analysis and automatic generation of analysis reports.

1. Introduction

An estimated 250,000–500,000 people die from seasonal influenza infection each year. The economic impact of influenza is immense due to the large number of lost working hours, hospitalizations, further medical complications, and treatment costs. Although vaccines against influenza exist, the rapid mutation of influenza virus calls for constant surveillance and annual vaccine reformulation [1]. A huge body of sequence data, annotations, and knowledge is available in the literature, online resources, and biological databases such as GenBank [2], UniProt [3], Protein Data Bank [4], EpiFlu Database [5], OpenFlu Database [6], Influenza Research Database (IRD) [7], and the Immune Epitope Database (IEDB) [8]. However, the underlying mechanisms of host/pathogen interaction are still not completely understood. The lack of a "universal" or broadly neutralizing influenza vaccine can be attributed to, among other factors, combinatorial complexity of the host immune system and the highly variable nature of viral antigens leading to

immune escape of the emerging influenza variants [9, 10]. One approach, in an attempt to overcome challenges of immune escape, is to raise a T-cell response against class I or class II epitopes conserved among viral strains [11, 12]. Public databases represent valuable resource for the study and development of broadly protective T-cell vaccines, but our ability to analyze these data falls behind the pace of data accumulation.

Numerous computational analysis tools that are useful for vaccine target discovery are available. They include keyword and text search tools, sequence comparison tools such as the BLAST algorithm [13] or multiple sequence alignment tools such as MAFFT [14], MUSCLE [15], and the Clustal [16], 3D structure visualization tools [17, 18], HLA binding prediction algorithms [19–21], and conservation analysis tools [22, 23], among others. The application of these tools in discrete steps can yield valuable information; however the extraction of higher-level knowledge requires integrating data from multiple databases and employing various analytical tools to answer specific questions. For example, when a new infectious influenza strain emerges (such as H9N7 avian flu [24] or a new seasonal flu) it is desirable to rapidly investigate its similarities and dissimilarities with known sequences, its epidemic or pandemic potential in humans, how different it is from the past vaccine strains, and its T- and B-cell epitopes from previously circulating strains and estimate its immune escape potential. Additionally, for new pandemic strains (such as 2009 swine flu [25]) it is desirable to establish origin and identify strains that are useful vaccine candidates. Well-defined workflows enable rapid extraction of such knowledge and automated generation of reports that contain such information, for which knowledge-based systems have previously been utilized [26, 27]. The need for integration and advanced analysis of available data is rapidly increasing. The integration of multistep analysis of multidimensional data for vaccine analysis and discovery requires the automation of analytical workflows [28].

FluKB is a knowledge-based system that integrates multiple types of influenza data and analytical tools into such workflows to support vaccine target discovery. The datasets in FluKB consist of curated, enriched, and standardized protein sequence data, immunological data from multiple data sources, and a set of modular analysis tools. The analysis tools infrastructure comprises a library of individual tools along with standard (applicable to multiple pathogens) and specific influenza vaccine target discovery workflows. Furthermore, we developed a standardized nomenclature to enable and speed up data mining using automated workflows. FluKB has a user-friendly web-based interface to access the data, tools, predefined workflows, and workflow reports. The overall architecture of FluKB is shown in Figure 1.

2. Materials and Methods

2.1. KB-Builder. FluKB was implemented using the KB-builder framework [29]. Briefly, KB-builder consists of seven major functional modules that enable automated data extraction from multiple sources, data cleaning, import to a central repository, integration of basic analysis tools, integration of advanced analysis tools, workflow definition, and update and maintenance. The KB-builder framework enabled setting up a web-accessible knowledge base and the analysis workflows. A workflow takes the user request, performs complex analyses which combines specific data and analytical tools, and feeds the results into subsequent analyses to produce a comprehensive report. The web-accessible interface uses a set of graphical user interface forms. These interfaces access search routines, analytical tools, and workflows that use a combination of PHP, Perl, Common Gateway Interface (CGI), and C background software. Development of KB-build and FluKB was carried out in CentOS 5.11 Linux environment. The web server used is Apache HTTP server 2.2.3 so the access to the web server is per default parallelized for each user. The Linux server is a 16-core server with 32 GB ram and should be able to handle the traffic to the website.

2.2. Data Sources for FluKB Data Repository. The data repository within FluKB contains four types of data: protein sequence data, HLA-related data (T-cell epitopes and HLA ligands), 3D crystal structures, and neutralizing antibody-related data. Protein sequences available from Influenza A, B, and, C viruses were collected from IRD [7] and GenBank [2], while HLA-related data were collected from IEDB [8]. The complex structures of neutralizing antibodies against influenza virus hemagglutinin (HA) were collected from the Protein Databank (PDB) [4]. The binding and neutralization assays of each neutralizing antibody were collected from primary literature, as described in [30].

2.3. Data Collection, Cleaning, and Enrichment

2.3.1. Protein Data Entries and Their Updates. The following sequence data and annotations were extracted from the IRD: protein sequence, GenBank identifiers (GI and Accession Numbers), UniProt ID, and nomenclature (type, host, location, ID, year, and subtype) when available. The metadata included a vocabulary that comprises the instances of the following terms: protein names, host names, geographic locations, years, and subtypes. The vocabulary was generated from the collected entries and enriched using primary literature. The vocabulary comprised correct terms as well as variants of the terms, and erroneous terms. During updates, each new entry is checked against the list of correct and erroneous terms. If all terms within the entry are matched to the correct terms, the entry is automatically annotated and converted into the FluKB format. If an erroneous term is identified, the curator is alerted and the correct term is proposed for the entry. The curator then approves the change or manually corrects the entry and updates the vocabulary, when needed. Each new term identified in the update dataset is inspected by the curator and added to the vocabulary. The vocabulary iteratively increases in size with each update and less than 20 new terms are usually identified for each update. This approach enables data curation that is of extremely high quality and can be performed very fast (Figure 2 upper path). We converted existing nomenclatures, whenever possible, into standardized formats. To verify the protein annotations of newly added entries, protein

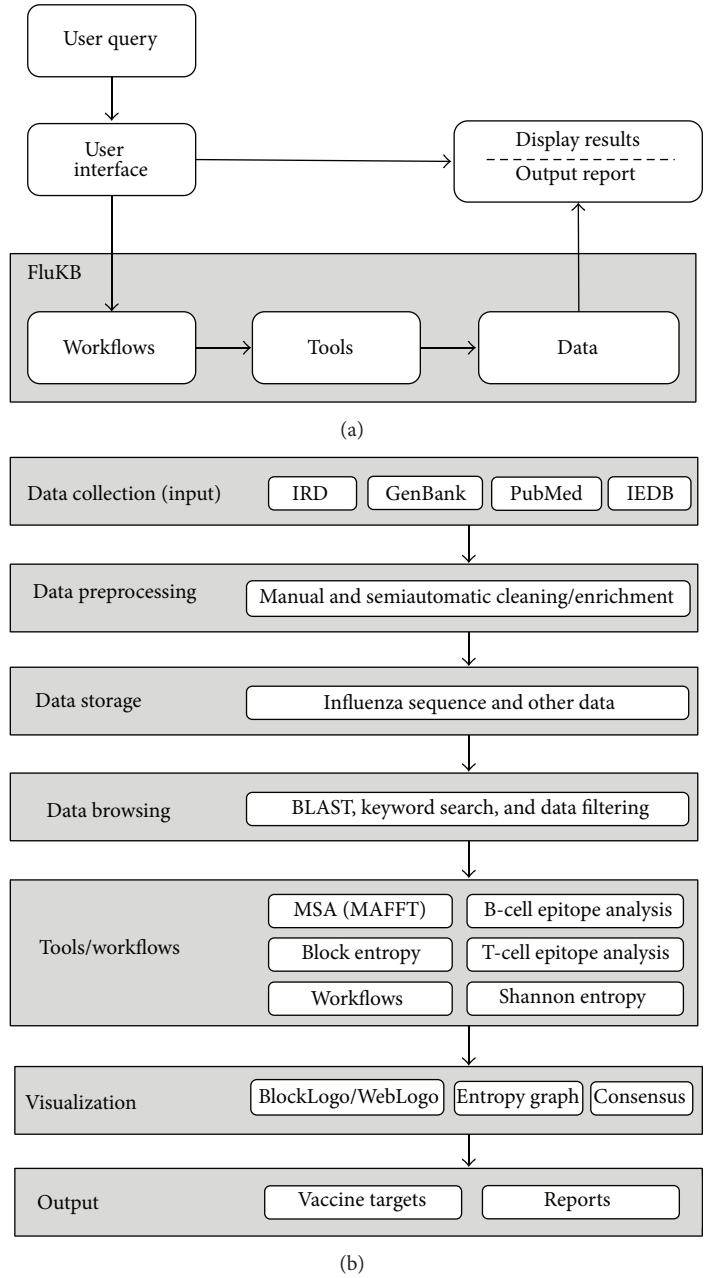


Figure 1: Overview of the architecture of FluKB. (a) Users can access FluKB through an interactive user interface where they can select specific data and tools or deploy a predefined workflow. (b) Top to bottom: data are collected, cleaned, and enriched. Higher-level knowledge extraction is enabled by utilization of tools assembled into workflows.

assignment of the strain proteomes was done by aligning the entry proteome sequence to a representative influenza reference sequences using the BLAST algorithm. We selected 10 proteins from UniProt that have detailed annotations as reference sequences for each protein of the two influenza types (A and B). The reference sequences are shown in Supplemental Table S1 (see Supplementary Material available online at http://dx.doi.org/10.1155/2015/380975).

2.3.2. Immunological Data. The immunological data extracted from IEDB include T-cell epitope sequence, epitope type (HLA binding, naturally processed, or T-cell epitope), PubMed references, experimental methods and results, and HLA-allele restriction (Figure 2 lower path). FluKB includes entries from IEDB that bind HLA class I or II allele with at least one positive result, or those reported as T-cell epitopes.

2.3.3. Neutralizing Antibody Data. For each neutralizing antibody, the following descriptive information is provided in the FluKB: isolation information (from human samples or from antibody phage-display libraries) from the primary

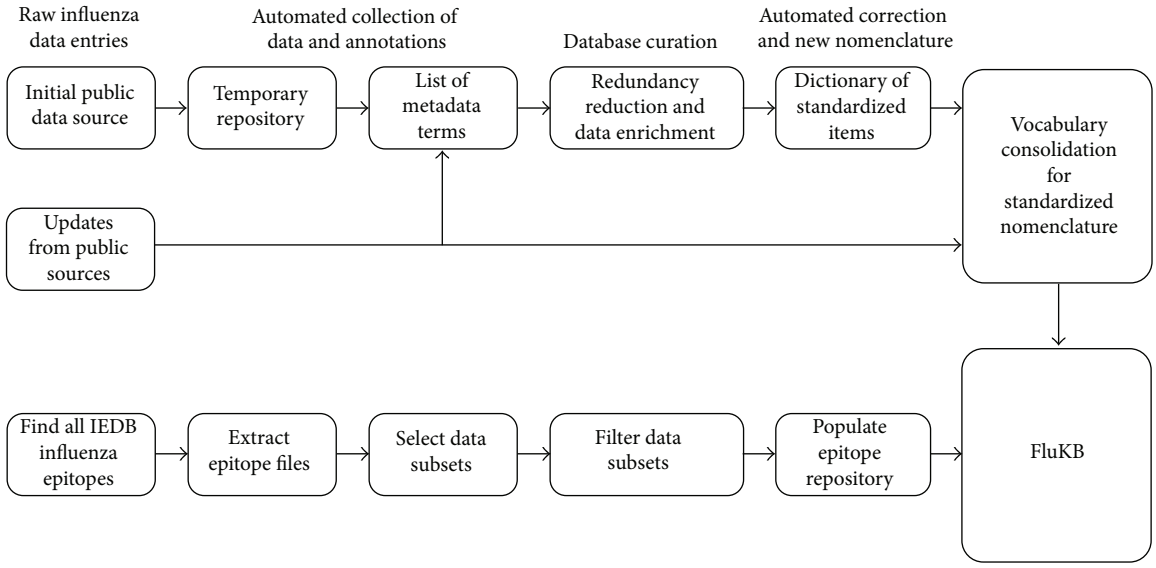


Figure 2: Semiautomated generation and updating of protein sequence repository of FluKB (upper path) and data extraction and repository creation of the influenza specific epitopes from IEDB (lower path).

literature, the corresponding crystal structure of antibody-HA complex from the PDB database, the B-cell epitope sequence variants detected from experimentally validated strains, and computationally defined B-cell epitopes on HA protein displayed as both sequence and 3D structure with Jmol [31].

2.4. Influenza Nomenclature Standardization and Definition of Data Mining Keys. Each influenza strain sample is annotated using the nomenclature originally proposed by the WHO [32] providing a shorthand description of influenza virus strains. However, the lack of a standardized vocabulary has made these nomenclatures error prone. For example, the nomenclature of an influenza isolate of type A, subtype H2N3, isolated from a duck in Heinersdorf in 1986, is written, using the original nomenclature as

A/Peking duck/Heinersdorf/648-4/1986(H2N3)

The lack of standardization of nomenclature has led to inconsistent nomenclature and incomplete metadata, thus increasing the difficulty in extraction of specific data subsets for analysis. The nomenclature in the Heinersdorf 1986 strain 648-4 has two issues. First the organism name term is erroneous ("Peking duck" is a traditional Chinese dish, while "Pekin duck" is the organism). Second, it is unclear where Heinersdorf is located. To ensure a complete access control over the sequence data within FluKB each entry was given a standardized data mining key. This key converts the nonstandardized nomenclature fields into a fully standardized format. The key is represented in FluKB as a standardized FASTA header that provides a condensed and detailed summary of the sample's information. The FluKB data mining key for the Heinersdorf sample HA protein is

>FLU0175850|type:A|host:/Mallard;Anas platyrhynchos;8839|location:Heinersdorf;Berlin-Germany;DE-BE|isolate:648-4|year:1986|subtype:H2N3|protein:HA|seqtype:complete|variant:1|vaccine:_|genbank: CY117179|geneinfo:386644010|uniprot:_|original: A/Peking duck/Heinersdorf/648-4/1986|key:yes

This search key compresses detailed sequence annotations into the FASTA format enabling easy combination of the results of sequence comparison with the analysis of annotations. Standardized formats for host and geographic location enable proper grouping and mapping of results. For host species, the NCBI-Taxonomy IDs [33] and Bird Life International taxonomy (http://www.birdlife.org/datazone/info/taxonomy) names were used as standard terms, including the NCBI taxonomy number. For the geographical locations, two-letter ISO codes for the countries and provinces were used (ISO-3166, 2012) [33]. This allows for each of the host species and each geographic term to be described in nonambiguous terms. Examples of corrected ambiguities are shown in Table 1. The FASTA format is easily understandable because of the descriptive nature of the fields in the FASTA header. Finally we use the term *key:yes* for all entries that could be fully annotated allowing them to be utilized within the analysis framework of FluKB, while those that could not be fully annotated are assigned with *key:no* and can only be found by a search.

2.5. Implementation of Analytical Tools. A set of analytical and visualization tools have been integrated within the FluKB. These tools include a selection of keyword searching tools: MAFFT [14] for multiple sequence alignment (MSA) and BLAST [13] for sequence similarity search. Specialized

Table 1: All the alternative instances of the host Mallard in all of the entries in the knowledge base of FluKB. The standardized name for the search key is *Mallard*; *Anas platyrhynchos*; *8839*.

Alternative names for Mallard	Number of times present in dataset	Status
Anas domesticus	1	Ambiguous
Anasplatyrhynchos	9	Error
Anasplatyrhynchos	39	Correct
Domestic duck	58	Variant
Domestic Mallard	11	Variant
Feral duck	1	Ambiguous
Khaki campbell duck	12	Variant
Mallard	25,844	Correct
Mallard duck	1,172	Redundant
Pekin duck	137	Variant
Peking duck	92	Error
Sentinel duck	13	Error
Wdk	11	Ambiguous (abbreviation)
Wild duck	952	Ambiguous
Total	28,352	

tools for the analysis of variability include sequence conservation metrics and their visualization using block entropy analysis [34]. The T-cell epitope prediction tools for HLA Class I and Class II have been integrated within FluKB for vaccine-related analyses. WebLogo [22] and BlockLogo [23] tools are used for visualization of results.

2.5.1. Sequence Conservation Metrics. FluKB enables conservation analysis of single positions within protein sequences, of linear blocks of amino acids extracted from multiple sequence alignments (MSA) of proteins using block entropy [34]. In addition, virtual peptides can be constructed from discontinuous epitopes within MSA and can be analyzed using block entropy, enabling the variability analysis of B-cell epitopes [30]. All these calculations are based on the calculation of Shannon entropy [35].

2.5.2. Visualization of Sequence Variability. WebLogo enables fast and easy interpretation of the position specific variability in an alignment. It displays amino acids by their corresponding one-letter code on a graph where all the amino acids present in one position are stacked on top of each other, and the frequency in the position is then based on their individual height in the graph. BlockLogo enables variability analysis of peptides (either linear peptides or discontinuous strings of amino acids), rather than single residues. The combinatorial number of possible blocks that can be created from a WebLogo can quickly become very large because of variation in individual positions. BlockLogo only shows the exact peptides that are most frequent for the investigated positions.

2.5.3. Prediction of MHC Class I and II Binders. For class I HLA binding prediction, the NetMHCpan 2.8a algorithm [19] was implemented. The alleles available for predictions are HLA-A∗02:01, -A∗03:01, -A∗11:01, -A∗2402, -B∗07:02, -B∗08:01, and -B∗15:01, since prediction accuracies for these alleles are relatively high [36]. Additionally, we predict binders to HLA-A∗01:01 since this allele is of very high frequency in the human population, although these predictions are of slightly lower accuracy than the seven benchmarked alleles. Collectively these alleles cover approximately 82% of human population [37]. For class II MHC predictions we used NetMHCIIpan 3.0 [21], with which users can predict binders to HLA-DRB1∗01:01, ∗03:01, ∗04:01, ∗07:01, ∗11:01, ∗13:01, and ∗15:01, as these have been benchmarked and validated for high accuracy [21]. These alleles cover approximately 40% of the human population [37].

2.5.4. Analysis of Sequence Similarity and Geographical Mapping (Strain Mapper). The standardized data mining key in FluKB enables sequence similarity searching and the display of a sequence's origin on the world map. For this purpose, we developed the specialized tool, Strain Mapper. The query sequence is entered in the search window and optionally the maximal number of amino acid mismatches can be selected. We used Google Earth software (http://www.google.com/earth/) for geographic map display. It is combined with sequence similarity analysis to allow user to select a query sequence and find its closest matches in FluKB and visualize them on the world map. This feature provides visualization of epidemiological information useful for evaluating the spread of a given virus and closely related variants.

3. FluKB Database

3.1. Data Repository. The FluKB sequence repository as of December 2014 contains 402,306 sequence entries from 75,426 unique strains of influenza. There are 67,907 type A, 7,028 type B, 194 type C, and 297 unknown type sequence entries. Out of 330,435 full-length sequences and 71,501 fragments, 370 protein sequences failed to align well to any of the template alignments during the annotation process. Each sequence entry contains information about location, host, and time of isolation, as well as a standardized nomenclature for identification of strains. Each entry contains a protein sequence with standardized, curated, and enriched annotations (Table 2).

The epitope repository contains a total of 357 unique T-cell epitopes (194 class I and 163 class II) and 685 unique HLA binders (572 class I and 113 class II). Each record describes the type of epitope (T-cell epitope, naturally processed, or HLA binder), epitope sequence type (only exact epitopes are included in the repository), experimental method used for validation, HLA-restriction, and literature references.

Twenty-eight neutralizing antibodies against influenza virus have crystal structures of HA/antibody complexes available in PDB. The functional data and neutralizing specificity of these antibodies were collected from published articles. Twenty of these antibodies target the globular head of the HA

Table 2: The data fields in each protein entry.

Field title	Field content
CVC accession	Accession number unique to FluKB
NCBI accession	Accession number unique to NCBI
GenBank GI	Accession number unique to GenBank
Type	The influenza virus type
Subtype	The serotype
Host	Host of collection
Country	Location of collection
Year	Time of collection
Isolate	Isolate name
Vaccine strain	Years the strain has been used as a vaccine strain
Original nomenclature	The original nomenclature from the raw data
Protein name	The protein name, based on template BLAST
Sequence type	Full or fragment of the protein sequence
Predict HLA binders	Predict HLA binders to the sequence
Blast sequence	Blast the sequence for similar sequences of FluKB
Sequence	The amino acid sequence of the entry
Epitopes in sequence	IEDB epitopes found in the protein sequence

protein, and the binding sites of the remaining eight antibodies are located on HA stem region. All of these antibodies were classified as broadly neutralizing (cross neutralization within subtype or across subtypes) and strain-specific antibodies.

We plan to have yearly updates of FluKB moving forward as new data and tools become available.

3.2. Data Cleaning, Quality Control, and Enrichment. The sequence data collected from IRD were subject to extensive cleaning, quality control (QC), and enrichment of annotations. We found that 142,232 (38.25%) of the 402.306 entries contained at least one type of error, ambiguity, or missing data. Most errors were in the geographic location fields where 72,340 (17.9%) had an error and 6,821 (12.1%) had missing information in the entry (see Table 3). In the initial screen of the data we found 2,977 entries that did not conform to nomenclature standard, including 305 entries that lacked information about host species, 867 entries that lacked separation fields within the nomenclature, and 1,805 other deficiencies. These entries were manually corrected and their nomenclatures were updated.

Furthermore, abbreviations, alternative, and misspelled names constituted the largest proportion of errors and were present in more than 10,000 entries. All name-related errors were corrected by the dictionary consolidation using the dictionary of standardized metadata terms. An example of the redundancy is shown in Table 1 where the host "Mallard" is found in 16,457 of the FluKB entries described by 14 different terms. In total, 96.41% of errors of various types described above were corrected and 469 standardized forms of missing data (such as location and host species) were added by manually searching the original literature.

Table 3: The types of errors found for the geographical location in the metadata of the entry.

Type of error	Number
Case errors	26,335
Redundant information	14,165
Alternative name and abbreviations	11,208
Misspellings	2027
Alternative spellings	9,112
True location could not be determined	9493
Total	72,340

Our effort in the data cleaning and enrichment stage focused on minimizing errors and maximizing data completeness to enhance knowledge extraction for discovery of potential vaccine targets in influenza, as well as genetic and epidemiological modeling of viral strains. Because of the system of reference sequences, templates, and reference MSA implemented in FluKB, we expect that the majority of future entries will be automatically corrected, if they contain errors and redundancies already encountered by the system. Any new errors will be subject to manual curating and updating of dictionaries.

3.3. Standardized Nomenclature. To enable automated data mining and workflows, we created data mining keys from the original nomenclature of influenza viruses with standardized terms. The data mining keys utilized NCBIs taxID database for host species [32] and the ISO codes for geographical location (ISO-3166, 2012) [33]. A total of 398,078 sequences (98.95%) were assigned the new developed nomenclature, while 4,228 (1.05%) could not be assigned. The original standard nomenclature is included in the data mining key as a reference for additional literature searches and text mining of article databases [37]. Data mining for vaccine targets often requires the analysis of subsets of data, for example, patient data such as specific HLA profile, age group, phenotypes, or other factors. Similarly, epidemiological modelling may need analysis of sequences from certain hosts, for example, specific migrating birds, or limited to geographical locations. The host, time, and location of collection are key information that help determine the spread of specific influenza strains and are central for better understanding of influenza outbreaks [38]. The data mining keys enable such analyses by having a standardized nomenclature, which pattern recognition algorithms can utilize as labels. Entries without the data mining key are made unavailable to the analysis on FluKB as inclusion of entries that lack data could affect the reliability and the outcome of the results. The data mining keys are furthermore nomenclature crucial for the automation of computational analyses; standardization of nomenclature fields allows the computer to interpret the data automatically, which previously was limited. For instance, the taxonomy ID of hosts enables host specific

Table 4: The sources of the integrated tools in FluKB and URL for their stand-alone versions web services.

Tool	URL	Reference
BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi	[11]
MAFFT MSA	http://mafft.cbrc.jp/alignment/software/	[12]
NetMHCpan 2.8a	http://www.cbs.dtu.dk/services/NetMHCpan/	[15]
NetMHCIIpan 3.0	http://www.cbs.dtu.dk/services/NetMHCIIpan/	[19]
Block entropy	http://research4.dfci.harvard.edu/cvc/flukb/HTML/blockentropy.php	ISO-3166
BlockLogo	http://mafft.cbrc.jp/alignment/software/	[21]

analyses that can potentially reveal features important for interspecies transmission of influenza. Proper organization of data allows for grouping of data by ancestral species and the variability can be followed over time. Furthermore the ISO codes for the geographical location by country and provinces enhance analyses in, for instance, epidemiological studies where an increased resolution in terms of actual spread can be analysed. This information can be used for the analysis of changes in T-cell and B-cell epitopes.

4. FluKB Tools

4.1. Database Searching and Querying. In FluKB, two search strategies can be deployed for sequence search: annotation-based or epitope-based. The first is a keyword search that enables the user to extract the data of FluKB into specific subsets and the second is a sequence similarity search by BLAST [13]. These search types are vital for the following data analysis as they enable the user to select the needed datasets based on specific scientific questions.

4.1.1. Keyword Search. The user can query the sequence entries for information such as the type, protein, subtype, year range, country, province, host, original nomenclature, and sequence type (fragment or full protein) by keyword search. The sequence entry database is indexed in order to decrease the retrieval time. An example of entry page retrieved by ID FLU0306481 or Strain Name A/Guangdong/1/2013, protein HA, is shown in Figure 3.

4.1.2. Sequence Similarity Search. FluKB has an indexed database generated from sequence entries that can be searched for sequence similarity using BLAST algorithm. The standard parameters are used: E value (10), word size (≥2), substitution matrix (BLOSOM62), gap cost (11) and extension (1), size of the result list (500), and pairwise list (250). Besides sequence search, the FluKB entries can be searched for T-cell epitopes and B-cell epitopes.

4.1.3. Stand-Alone Tools. FluKB can be queried using a selection of analytical tools under the tab "Tools." Sequence alignment by MSA can be performed under the "Sequence alignment" tab by entering either a list of sequence IDs in the query window or a selection of subsets of proteins by name, influenza type, subtype, range of years of identification, country, province, host, or complete sequences/fragments. The protein subsets selection window (Supplemental Figure S1)

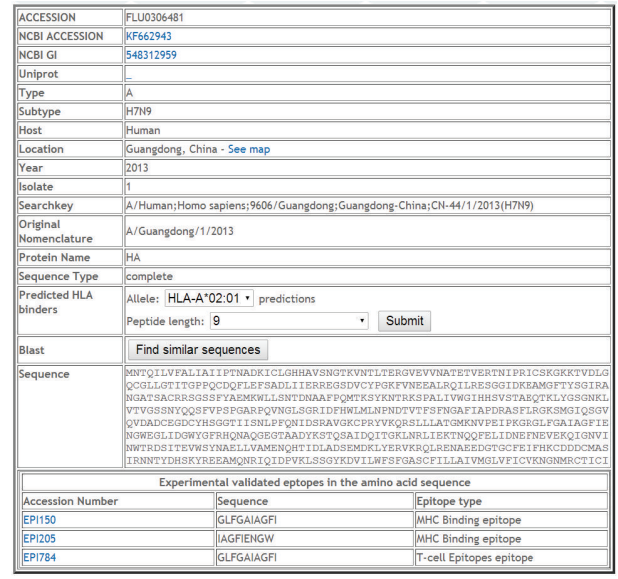
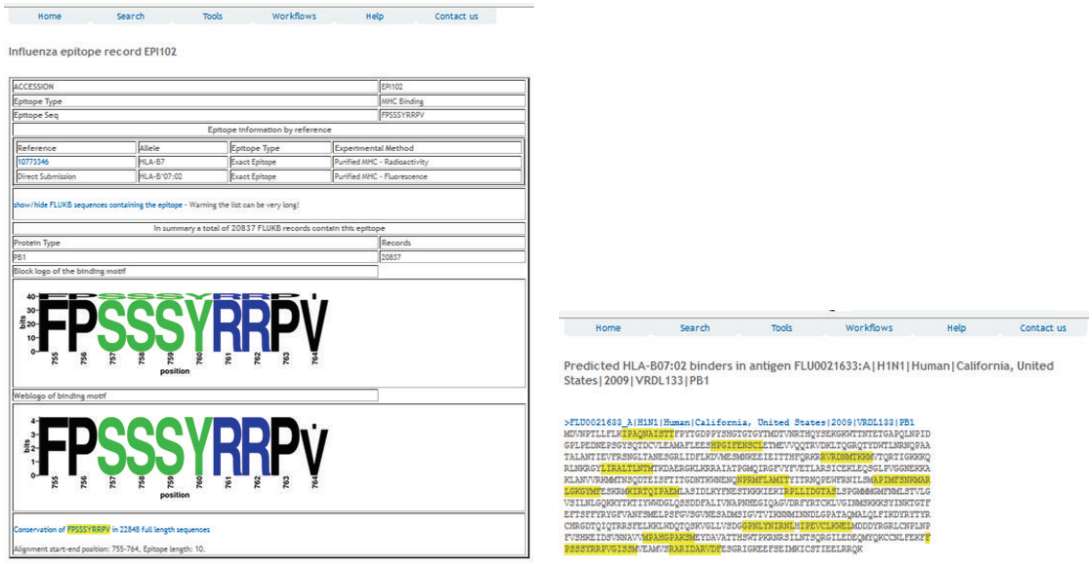


Figure 3: Record FLU0306481, Protein HA representing the 2013 H7N9 outbreak of bird flu in China.

can be used for sequence variability analysis and block entropy calculation under appropriate tabs. Epitope block entropy (T-cell epitopes) uses protein subsets selection window that enables the input of epitope sequence. The strain mapping tool is described in Section 2.5.4.

4.2. Variability Analysis. The analysis of variability of viral sequences is important for understanding the emergence of new strains, immune escape, changes in pathogenicity, the extent of spread of viral strains, and vaccine design. The variability analysis can be performed interactively, but the variability analysis tools are also integrated in the T-cell and B-cell mapping tools and relevant workflows. The interactive variability analysis can be performed using individual sequences or sets of sequences. The tools used in variability analysis are shown in Table 4. The main tools for the analysis of variability are BLAST search that can be accessed from individual entries (Figure 3, "find similar sequences") or from the "Sequence alignment" under the "Tools" tab. The MSA can be performed from the results of BLAST search by selection of sequences and clicking the "Align them…" key. The positions of variability within the MSA results are color-coded for better visual inspection and each sequence is hyperlinked to its record (Supplemental



(a) (b)

Figure 4: T-cell epitope EPI102 and the T-cell epitope analysis from the record FLU0021633. (a) Entry EPI150 with the graphical display of present T-cell epitope variants. (b) Graphical display of predicted T-cell epitopes (HLA-B∗07:02, length 10, IC_{50} < 1000 nM).

Figure S2). The "Sequence variability analysis" tool plots entropy (red curve) and the percentage of sequences (blue curve) containing the consensus amino acid at all positions along with the consensus sequence (Supplemental Figure S3). Further visualization can be achieved by "Block entropy" calculation, which visualizes the conservation of peptides of lengths appropriate for immune recognition, rather than individual residues [34]. The "Epitope block entropy" calculation displays variability of a specific epitope across the selected subset of sequences.

4.3. T-Cell Epitope Search. T-cell epitope analysis can be performed directly from the protein entry. For example, three T-cell epitope entries are displayed on the record entry page (Figure 3). The epitope entry page for EPI150 is shown in Figure 4(a). In addition, the prediction of T-cell epitopes can be performed by selecting allele and peptide length in the "Predicted HLA Binders" field followed by the "Submit" action. The visual display of experimentally verified T-cell epitopes is shown in Figure 4(b).

T-cell epitope search can also be initiated from the "Search" tab, where epitope can be searched by the sequence. The results will appear as a list of epitopes along with their binding or T-cell restriction specificities. The list is hyperlinked to the epitope record, an example of which is shown in Figure 4(a). Finally, T-cell epitope search can be performed using the workflow titled "Vaccine targets" under the Workflows tab. After selection of input parameters, for example, Allele "HLA-A∗0201," Protein "HA," Influenza type "A," subtype "H7 N9," year(s) "2013 2014," Affinity threshold "500 nM," and Conservation Threshold "95%" (and remaining values "default"), 98 sequences will be selected and the report will be generated.

4.4. B-Cell Epitope Search. B-cell epitope analysis can be performed from the Search tab by selecting "Antibodies list." By the end of May, 2014, 28 antibodies and their detailed neutralizing and structural information have been deposited in FluKB. All of these antibodies are neutralizing antibodies against hemagglutinin protein on influenza virus. These antibodies are listed on the webpage, while their respective B-cell epitopes can be displayed on three interactive structures: X-31 strain-specific antibodies on HA structure from 1KEN, broadly neutralizing antibodies on HA structure from 1EO8, and influenza B virus antibodies on HA structure from 4FQK. This feature enables visual comparison of antibody-specific B-cell epitopes.

For each neutralizing antibody, the isolation information, structure information, and computational identified B-cell epitope information can be accessed. Also, the neutralized motifs and escape motifs extracted from experimentally validated strains from the primary literature are presented as well [30]. In addition, two workflows have been implemented for further analysis: the neutralization coverage estimation and B-cell epitope mapping (Supplemental Figure S4). The neutralized/escape coverage by a specific existing neutralizing antibody is calculated for the complete population of influenza strains. The strain population coverage by a neutralizing antibody can be assessed within any selected subset of influenza strains, such as year range, specific subtype, and geographic coverage. The B-cell epitope mapping is performed by submitting a query hemagglutinin sequence. Cross neutralization coverage of a known neutralizing antibody can be estimated based on sequence comparison to the known neutralizing epitopes. A discontinuous peptide is extracted based on epitope positions determined from crystal structures. This tentative discontinuous peptide is then compared

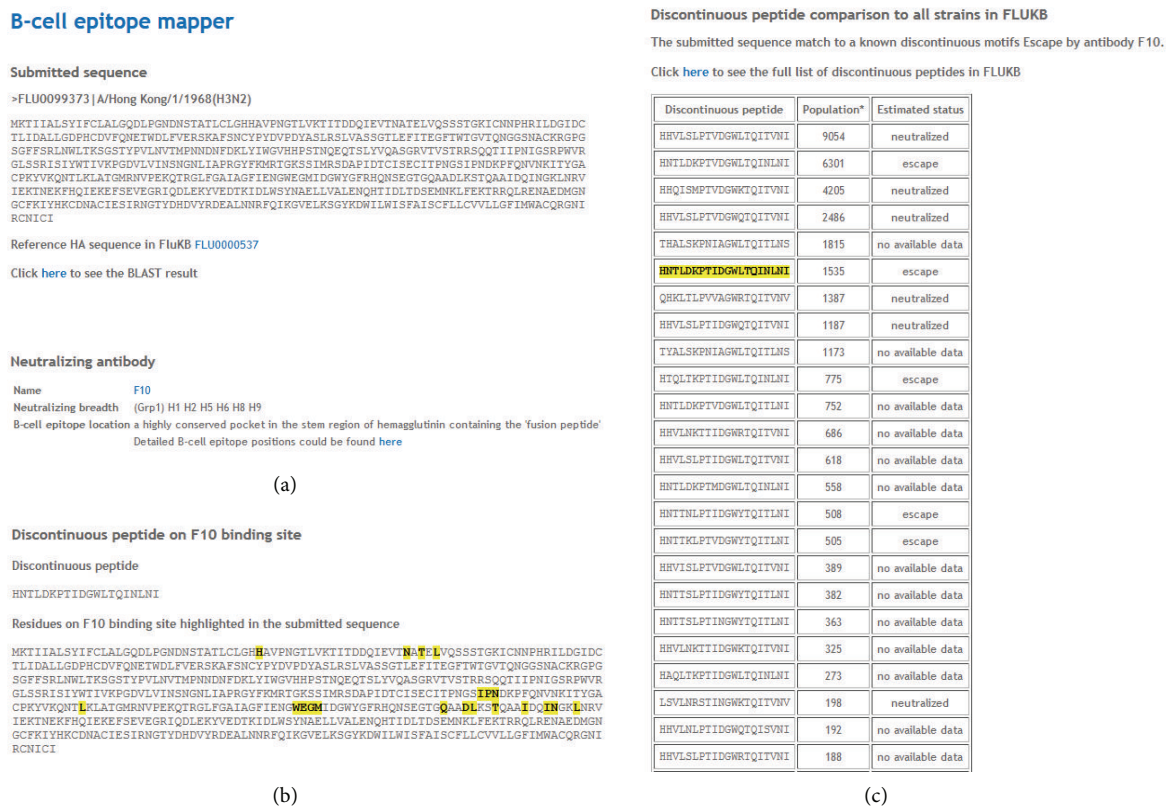


Figure 5: The results of B-cell epitope analysis II. Analysis of entry FLU0099373 for broadly neutralizing antibody F10 interaction. (a) The BLAST result of the query sequence to the sequence with highest identity in FLUKB, (b) the discontinuous peptide extracted from the query sequence and respective residue positions highlighted in full sequence, and (c) the summary information of neutralizing antibody. The discontinuous peptide from the query sequence is compared to all discontinuous peptides generated from FLUKB (with their neutralizing status listed). The identical sequence is highlighted in yellow with estimated status of neutralization given.

to the B-cell epitopes of experimentally validated strains. An example of B-cell epitope analysis is shown in Figure 5.

FluKB offers the capability to address complex questions relating to sequence variability on very specific subsets, identification of potential T-cell epitopes, and selection and combination of these epitopes into polyvalent vaccine constructs. The modular structure of the workflow renders FluKB highly flexible. The tools and data can be reorganized and more tools can be created to answer additional questions, for example, relating to epidemiological modeling and analysis of cross protective potential of neutralizing antibodies. The overall architecture can be viewed in Supplemental Figure S5.

5. Conclusion

Publically available influenza data are a valuable resource for computational analyses with applications in vaccine design. Similarly, existing bioinformatics tools provide the means for extraction of information and new knowledge. However, to utilize the full potential of these resources, data preprocessing must be performed and analytical tools must be carefully combined into well-defined workflows. These workflows allow users to ask specific questions (scientific, technical, and clinical) and provide means for systematic data analysis. These workflows can automatically generate comprehensive analysis reports. The infrastructure of data and tools is the backbone of FluKB and similar knowledge-based systems [26, 27].

Despite many years of research and available vaccines, influenza remains a major public health burden and a threat of a major new pandemic. Multiple data sources provide information on protein and nucleotide sequences and immune epitopes in influenza [2, 5–8, 39, 40]. They represent well-maintained catalogues of influenza sequences and annotations, along with a selection of basic search tools. They focus mainly on providing access to data, extraction, and simple analyses. The FluKB was developed focusing on a different purpose, the facilitation of data mining for influenza vaccinology and immunology of influenza infection. The FluKB has very clean and standardized data, integrating information on antigen sequences, and immunological epitopes. The set of integrated analysis tools and workflows are designed to aid rational vaccine design. This includes the discovery of vaccine targets, assessment of variability, and in-depth analysis of immune epitope. FluKB is a unique data mining system for largely automated knowledge discovery from the ever-increasing body of influenza data with applications in both T-cell and B-cell immunology and vaccinology.

Systematic discovery of influenza vaccine targets requires highly accurate, up-to-date, and standardized data of influenza antigens and immune epitopes. The sequence and epitope data available through publications, various reports, and databases vary in quality, granularity, and data formats. The extraction of knowledge and discovery of vaccine targets from diverse and scattered data sources are a challenging and time-consuming task. FluKB integrates the content and the analytical tools in a unified system that enables the automation of complex queries and discovery. FluKB is a contribution to the long-standing quest for universal influenza vaccines [41, 42] by allowing a large-scale analysis on a large collection of annotated influenza sequences. FluKB is publicly available at http://research4.dfci.harvard.edu/cvc/flukb/.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgment

This work was supported by the NIH Grant U01 AI90043.

References

[1] "Prevention and control of seasonal influenza with vaccines: recommendations of the advisory committee on immunization practices—United States, 2013-2014.," *Morbidity and Mortality Weekly Report*, vol. 62, no. 7, pp. 1–43, 2013.

[2] D. A. Benson, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers, "GenBank," *Nucleic Acids Research*, vol. 38, no. 1, pp. D46–D51, 2009.

[3] The UniProt Consortium, "Reorganizing the protein space at the Universal Protein Resource (UniProt)," *Nucleic Acids Research*, vol. 40, no. 1, pp. D71–D75, 2012.

[4] H. M. Berman, J. Westbrook, Z. Feng et al., "The protein data bank," *Nucleic Acids Research*, vol. 28, no. 1, pp. 235–242, 2000.

[5] The GISAID Epiflu database.

[6] R. Liechti, A. Gleizes, D. Kuznetsov et al., "OpenFluDB, a database for human and animal influenza virus," *Database*, vol. 2010, Article ID baq004, 2010.

[7] R. B. Squires, J. Noronha, V. Hunt et al., "Influenza research database: an integrated bioinformatics resource for influenza research and surveillance," *Influenza and other Respiratory Viruses*, vol. 6, no. 6, pp. 404–416, 2012.

[8] R. Vita, L. Zarebski, J. A. Greenbaum et al., "The immune epitope database 2.0," *Nucleic Acids Research*, vol. 38, no. 1, pp. D854–D862, 2010.

[9] A. T. Heiny, O. Miotto, K. N. Srinivasan et al., "Evolutionarily conserved protein sequences of influenza a viruses, avian and human, as vaccine targets," *PLoS ONE*, vol. 2, no. 11, Article ID e1190, 2007.

[10] I. Sitaras, D. Kalthoff, M. Beer, B. Peeters, and M. C. M. de Jong, "Immune escape mutants of highly pathogenic avian influenza H5N1 selected using polyclonal sera: identification of key amino acids in the HA protein," *PLoS ONE*, vol. 9, no. 2, Article ID e84628, 2014.

[11] Y. Furuya, J. Chan, M. Regner et al., "Cytotoxic T cells are the predominant players providing cross-protective immunity induced by γ-irradiated influenza A viruses," *Journal of Virology*, vol. 84, no. 9, pp. 4212–4221, 2010.

[12] D. B. Keskin, B. B. Reinhold, G. L. Zhang, A. R. Ivanov, B. L. Karger, and E. L. Reinherz, "Physical detection of influenza A epitopes identifies a stealth subset on human lung epithelium evading natural CD8 immunity," *Proceedings of the National Academy of Sciences*, vol. 112, no. 7, pp. 2151–2156, 2015.

[13] 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.

[14] K. Katoh, K. Misawa, K.-I. Kuma, and T. Miyata, "MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform," *Nucleic Acids Research*, vol. 30, no. 14, pp. 3059–3066, 2002.

[15] R. C. Edgar, "MUSCLE: multiple sequence alignment with high accuracy and high throughput," *Nucleic Acids Research*, vol. 32, no. 5, pp. 1792–1797, 2004.

[16] M. A. Larkin, G. Blackshields, N. P. Brown et al., "Clustal W and Clustal X version 2.0," *Bioinformatics*, vol. 23, no. 21, pp. 2947–2948, 2007.

[17] The PyMOL Molecular Graphics System.

[18] Jmol: an open-source Java viewer for chemical structures in 3D.

[19] C. Lundegaard, K. Lamberth, M. Harndahl, S. Buus, O. Lund, and M. Nielsen, "NetMHC-3.0: accurate web accessible predictions of human, mouse and monkey MHC class I affinities for peptides of length 8-11," *Nucleic Acids Research*, vol. 36, pp. W509–W512, 2008.

[20] M. Nielsen, S. Justesen, O. Lund, C. Lundegaard, and S. Buus, "NetMHCIIpan-2.0—improved pan-specific HLA-DR predictions using a novel concurrent alignment and weight optimization training procedure," *Immunome Research*, vol. 6, no. 1, article 9, 2010.

[21] 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 296, 2009.

[22] G. E. Crooks, G. Hon, J.-M. Chandonia, and S. E. Brenner, "WebLogo: a sequence logo generator," *Genome Research*, vol. 14, no. 6, pp. 1188–1190, 2004.

[23] L. R. Olsen, U. J. Kudahl, C. Simon et al., "BlockLogo: visualization of peptide and sequence motif conservation," *Journal of Immunological Methods*, vol. 400-401, no. 1, pp. 37–44, 2013.

[24] T. Kageyama, S. Fujisaki, E. Takashita et al., "Genetic analysis of novel avian A(H7N9) influenza viruses isolated from patients in China, February to April 2013," *Eurosurveillance*, vol. 18, no. 15, Article ID 20453, 2013.

[25] J. S. M. Peiris, L. L. M. Poon, and Y. Guan, "Emergence of a novel swine-origin influenza A virus (S-OIV) H1N1 virus in humans," *Journal of Clinical Virology*, vol. 45, no. 3, pp. 169–173, 2009.

[26] L. R. Olsen, G. L. Zhang, E. L. Reinherz, and V. Brusic, "FLAVIdB: a data mining system for knowledge discovery in flaviviruses with direct applications in immunology and vaccinology," *Immunome Research*, vol. 7, no. 3, pp. 1–9, 2011.

[27] G. L. Zhang, A. B. Riemer, D. B. Keskin, L. Chitkushev, E. L. Reinherz, and V. Brusic, "HPVdb: a data mining system for knowledge discovery in human papillomavirus with applications in T cell immunology and vaccinology," *Database*, vol. 2014, Article ID bau031, 2014.

[28] J. Söllner, A. Heinzel, G. Summer et al., "Concept and application of a computational vaccinology workflow," *Immunome Research*, vol. 6, supplement 2, article S7, 2010.

[29] G. L. Zhang, L. Chitkushev, L. R. Olsen, U. J. Kudahl, C. Simon, and V. Brusic, "Streamlining the development of immunological knowledge bases," in *Genomics Drug Discovery*, M. Sakharkar, Ed., 2014.

[30] J. Sun, U. J. Kudahl, C. Simon, Z. Cao, E. L. Reinherz, and V. Brusic, "Large-scale analysis of B-cell epitopes on influenza virus hemagglutinin—implications for cross-reactivity of neutralizing antibodies," *Frontiers in Immunology*, vol. 5, article 58, 2014.

[31] R. M. Hanson, "Jmol-a paradigm shift in crystallographic visualization," *Journal of Applied Crystallography*, vol. 43, no. 5, pp. 1250–1260, 2010.

[32] WHO, "A revision of the system of nomenclature for influenza viruses: a WHO memorandum," *Bulletin of the World Health Organization*, vol. 58, pp. 585–591, 1980.

[33] S. Federhen, "The NCBI Taxonomy database," *Nucleic Acids Research*, vol. 40, no. 1, pp. D136–D143, 2012.

[34] L. R. Olsen, G. L. Zhang, D. B. Keskin, E. L. Reinherz, and V. Brusic, "Conservation analysis of dengue virust-cell epitope-based vaccine candidates using peptide block entropy," *Frontiers in Immunology*, vol. 2, article 69, 2011.

[35] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, vol. 27, pp. 379–423, 1948.

[36] 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.

[37] L. Gragert, A. Madbouly, J. Freeman, and M. Maiers, "Six-locus high resolution HLA haplotype frequencies derived from mixed-resolution DNA typing for the entire US donor registry," *Human Immunology*, vol. 74, no. 10, pp. 1313–1320, 2013.

[38] D. Onozuka and A. Hagihara, "Spatial and temporal dynamics of influenza outbreaks," *Epidemiology*, vol. 19, no. 6, pp. 824–828, 2008.

[39] U. Consortium, "Ongoing and future developments at the Universal Protein Resource," *Nucleic Acids Research*, vol. 39, pp. D214–D219, 2010.

[40] P. W. Rose, C. Bi, W. F. Bluhm et al., "The RCSB Protein Data Bank: new resources for research and education," *Nucleic Acids Research*, vol. 41, no. 1, pp. D475–D482, 2013.

[41] A. J. McMichael, F. M. Gotch, G. R. Noble, and P. A. S. Beare, "Cytotoxic T-cell immunity to influenza," *The New England Journal of Medicine*, vol. 309, no. 1, pp. 13–17, 1983.

[42] J. Sui, W. C. Hwang, S. Perez et al., "Structural and functional bases for broad-spectrum neutralization of avian and human influenza A viruses," *Nature Structural and Molecular Biology*, vol. 16, no. 3, pp. 265–273, 2009.

Hindawi Publishing Corporation
Journal of Immunology Research
Volume 2015, Article ID 573165, 9 pages
http://dx.doi.org/10.1155/2015/573165

Research Article

Automated Classification of Circulating Tumor Cells and the Impact of Interobsever Variability on Classifier Training and Performance

Carl-Magnus Svensson,[1] Ron Hübler,[1,2] and Marc Thilo Figge[1,2]

[1] *Applied Systems Biology, Leibniz Institute for Natural Product Research and Infection Biology–Hans-Knöll-Institute (HKI), Beutenbergstraße 11a, 07745 Jena, Germany*
[2] *Friedrich Schiller University Jena, Fürstengraben 1, 07743 Jena, Germany*

Correspondence should be addressed to Marc Thilo Figge; thilo.figge@hki-jena.de

Received 27 August 2015; Accepted 15 September 2015

Academic Editor: Francesco Pappalardo

Copyright © 2015 Carl-Magnus Svensson et al. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Application of personalized medicine requires integration of different data to determine each patient's unique clinical constitution. The automated analysis of medical data is a growing field where different machine learning techniques are used to minimize the time-consuming task of manual analysis. The evaluation, and often training, of automated classifiers requires manually labelled data as ground truth. In many cases such labelling is not perfect, either because of the data being ambiguous even for a trained expert or because of mistakes. Here we investigated the interobserver variability of image data comprising fluorescently stained circulating tumor cells and its effect on the performance of two automated classifiers, a random forest and a support vector machine. We found that uncertainty in annotation between observers limited the performance of the automated classifiers, especially when it was included in the test set on which classifier performance was measured. The random forest classifier turned out to be resilient to uncertainty in the training data while the support vector machine's performance is highly dependent on the amount of uncertainty in the training data. We finally introduced the consensus data set as a possible solution for evaluation of automated classifiers that minimizes the penalty of interobserver variability.

1. Introduction

The identification and enumeration of circulating tumor cells is an important tool for evaluation of the disease progression in especially breast cancer [1–3] and is also under consideration as a diagnostic tool in various other types including lung and colorectal cancer [4–6]. The type of CTCs found also serves as a potential marker for changes in the chemotherapy resistance of a cancer [7]. The extreme rarity of CTCs in patient blood, typically one CTC per 10^8 blood cells [8], makes both collection and detection of these cells extremely challenging. The collection of CTCs from peripheral blood is in a majority of studies done by antiepithelial-cell-adhesion-molecule (EpCAM) antibody-coated isolation systems [5, 9], but also other types of immunomagnetic devices [10, 11], density gradient centrifugation [12], and membrane filtration [13] are used for CTC enrichment. The detection of CTCs after collection is done by immunocytological staining or polymerase chain reaction (PCR) [14]. In the case of immunocytological staining the standard method of CTC enumeration is manual counting either at the microscope or from microscopy images [15, 16]. However, progress was lately made in using machine learning techniques for the detection of CTCs from fluorescence microscopy images [17, 18]. In these studies, as well as in any study applying classifiers to data, manual labelling was used for validation and also for training using (semi)supervised training regiments. The use of computational methods, in this case machine vision, makes the screening of the vast amounts of data that is readily available today quicker and more efficient. Instead of having

a highly trained expert performing the time-consuming task of looking at numerous images, this can be done by the computer. Even if the computer is not able to completely take over the manual analysis it can at least screen the image data for regions of interest and provide a second opinion in difficult cases.

This paper builds on the previous result in enumerating CTCs in images using image analysis techniques combined with support vector machines (SVMs) and naïve Bayesian classifiers (NBCs) [18]. Data for the study was collected with a *functionalized and structured medical wire* (FSMW) [19] that is a CE-certified medical device for the isolation of CTCs. Human carcinoma cells expresses the epithelial cell adhesion molecule (EpCAM) on their surface while this molecule is absent from the surface of haematological cells [20–22]. The FSMW is functionalized with anti-EpCAM antibodies and was inserted into the cubital vein of a patient through a standard 20 G intravenous cannula, where it was left for 30 minutes collecting CTCs from the blood that flows past [19]. After cell collection the FSMW was fluorescently stained and microscopy images were made in which we aim to enumerate CTCs. Ideally only CTCs should adhere to the FSMW but because of the many blood cells compared to CTCs, even the unlikely event of catching a blood cell occurs regularly. The first step in the analysis was to identify regions of interest (ROIs) which are candidates as CTCs but may in fact also be a blood cell, some kind of debris or a staining artifact. In the previous study we concluded that both SVMs and NBCs achieved an accuracy of CTC detection in the range of 85–90% after ROIs were identified [18]. In that study, the annotation used for evaluation of classifier performance and training of the classifiers were based on the manual classification of the ROIs by one author (CMS).

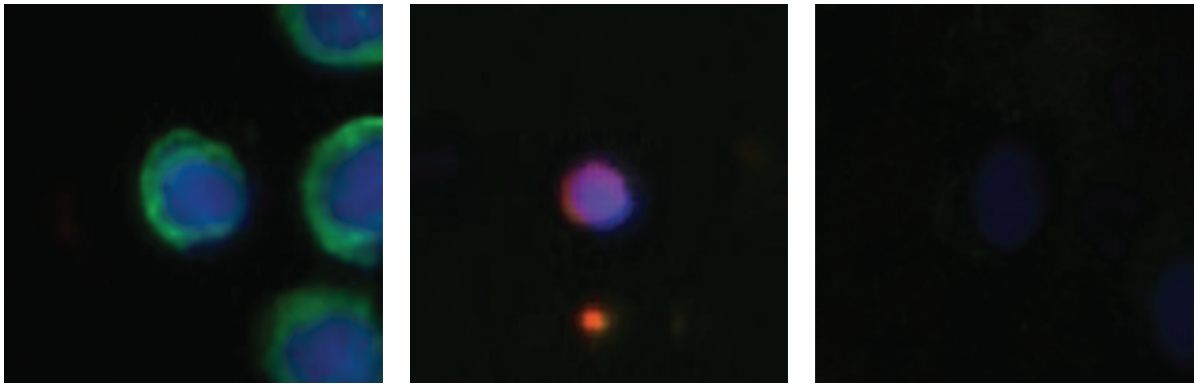
The use of different machine learning and machine vision techniques is an active research field with the aim of making disease diagnosis more accurate and efficient [23]. Especially in the diagnosis and treatment evaluation of different cancer types, including but not limited to prostate [24, 25] and colorectal cancer [26], automated algorithms are used. However, interobserver variability is a known issue in diagnostics of different cancer types and a disagreement of more than 15% is not uncommon when multiple observers, normally all trained experts, are interpreting patient image data of different types [27–29]. When training and evaluating an automated classifier the labels provided by observers are of great importance as any inconsistencies will affect the performance of the classifier. In this study, we investigated how uncertainty in annotation, so called label noise, affects the performance of automated classification using a random forest (RF) and a SVM and relate that to the performance of earlier studies [17, 18]. Interobserver variability for disease progression using CTCs is reported to be as low as 1% but is then related to the question if the patient has more than 4 CTCs per 7.5 mL blood [27]. When considering the manual classification of images of possible CTCs, Scholtens et al. presented that observers disagree on approximately 15% of the data points [17]. To investigate how this variability affects the estimated performance of the classifiers, we in this study carefully identified possible label noise through analysis of the manual annotation. Moreover, a consensus annotation was identified and training and testing of the classifiers with a controlled amount of label noise in both training and test sets was evaluated.

2. Materials and Methods

2.1. Image Data. The data set used for this study was the same as used in our earlier publication [18], where CTCs were captured using the FSMW both *in vivo* and *in vitro* [19]. After collection the FSMW was fluorescently stained for cell nuclei (blue), EpCAM, or cytokeratins (green) and counterstained for CD45 (red) in order to differentiate between CTCs and blood cells that may have attached to the wire. Images were taken using a 10x ocular and 10x, 20x, or 40x objective resulting in 1.0 μm^2, 0.5 μm^2, or 0.25 μm^2 pixel resolution of the images. CTCs are those cells that exhibit nuclear dye (blue) colocalized with the antibodies against cytokeratin and/or EpCAM (both green); see Figure 1(a). ROIs, for example, objects that may be CTCs and most likely at least some type of cell, were identified based on the blue signal that indicates positive staining of a cell nucleus. For full details of the collection, staining, imaging, and ROI identification we refer the reader to earlier publications using this data set [18, 19]. The data points used for CTC classification were obtained by cutting out an image with area 100×100 pixels around the center of each identified ROI resulting in 617 data points from 61 original microscopy images.

2.2. Manual Annotation. Manual annotation was needed for both training and evaluation of the classifiers as well as for the determination of interobserver variability. The observers were instructed to determine if the most central object in each image cutout was a CTC or not. For an example of multiple objects occurring in the same cutout see Figure 1(a). According to guidelines used in earlier studies [18, 19], observers were instructed to count the object as a CTC if the nuclei (blue staining) were intact and the object showed positive staining for EpCAM or cytokeratin (green staining). The blue and green staining had to be distinguishable from each other; for example, the nuclei and the EpCAM staining should be structured. While it was required that the nuclei should be intact, it was allowed for CTCs to have irregular shapes or be clustered. Any object that showed positive CD45 staining (red) was not to be counted as a CTC; see Figure 1(b). All $N_{\text{obs}} = 11$ observers (with 5/6 male/female) had normal or corrected to normal eyesight and no one had any known issues with color vision. Cutouts were presented on individual laptops in one session to avoid different light conditions and without any time restrictions. The order of the data points was random and the observers were instructed not to confer.

2.3. Data Preprocessing and Automatic Classification. As the cutouts have been taken at different magnifications we first normalized the image matrix to cover a region of the size of 2500 μm^2 around the center of each image cutout. This means that the cutouts had 100×100, 71×71, or 50×50 pixels, depending on if they were from an image taken with a 40x,



(a) (b) (c)

Figure 1: Examples of image cutouts classified with regard to the most central object being a CTC or not. (a) All eleven human observers agree that this is a CTC. (b) All eleven observers agree that this is not a CTC. (c) Six observers classify that this is a CTC while the other five say that it is not a CTC.

20x, or 10x ocular. We then applied a Gauss convolution filter with standard deviation $\sigma = 1$ pixel to the cutouts to reduce the effects of high frequency noise. The classifiers we used, SVM and RF, both required inputs with fixed dimensions and therefore all cutouts were downsampled to 50×50 pixels using the raster package in R (https://cran.r-project.org/package=raster). The color space of the cutouts were red-green-blue (RGB) when read but for the classification we transformed them to hue-saturation-value (HSV) using the grDevices package (https://stat.ethz.ch/R-manual/). This was done as HSV has a natural division between color dimensions (H and S) and intensity (V), which is not present in RGB space. In the HSV space we dropped the V dimension as preliminary tests revealed that this factor is not decisive in the classification of cutouts containing a CTC or not. The image matrix was then vectorized so that each cutout is then represented by an array with 5000 entries with the hue and saturation values of the cutout. For the rest of this paper any reference to automated classification of a data point or cutout will mean that this vector containing the hue and saturation values of a cutout was presented to the classifier.

The automated classifiers were implemented in R using the h2o interface (https://cran.r-project.org/package=h2o) for the RF and the kernlab package (https://cran.r-project.org/package=kernlab) for the SVM. The RF was an implementation of the Breiman forest [30] consisting of 500 trees. The SVM with radial basis function (RBF) kernel [31] had the parameters $C = 2$ and $\gamma = 0.005$, where C is the soft margin penalty and γ the inverted radius of the RBF. Parameters, number of trees as well as C and γ for the SVM, were optimized to give the highest accuracy possible on a subset of the data.

To get the classifier responses to the data, all data sets, that is, both the entire set of 617 cutouts and subsets that will be described, were divided into randomized folds. Training of the classifier was then performed on a number of folds and testing was then done on one or more folds that were not used for training. This was done iteratively with new folds chosen for training and testing until all data points were classified. For each subset of data the number of folds and how they were used for training and testing are described in the text where appropriate.

3. Results

3.1. Interobserver Variability Reveals Differences in Bias and Large Degree of Uncertainty. In the $N = 617$ data points the observers found on average 300 CTCs with the median being 318, but the number varied largely as can be seen in Table 1. The lowest number of CTCs was found by the observer MTF with 221 CTCs and the largest number was 354 CTCs observed by ST. The largest interobserver distance in an ordered list was between MB (223) and MP (281) with 48 CTCs, while the second largest distance was 17 between JP (330) and CMS (347). The initial conclusion is therefore that two observers, MTF and MB, had a much more conservative opinion on what was to be considered a CTC than the other observers, thereby minimizing the risk for false positive CTC annotation. The rest of the observers have a range of detected CTCs that corresponds to approximately 10% of the total number of data points presented.

We define the agreement between observers A and B as

$$P_{\text{Agr}}(A, B) = 1 - \frac{1}{N}\sum_{i=1}^{N} \text{abs}\left(A_i - B_i\right), \quad (1)$$

where $A_i, B_i \in [0, 1]$ indicates the annotation of image cutout i as CTC (1) or no CTC (0) by the respective observer. If two observers agreed on all data points their agreement is one, whereas total disagreement gives the value zero. In Figure 2 a heat map showing the agreement between all observers is shown. It is worth noticing that the maximal agreement was 0.91 and that the average (median) agreement was 0.85 (0.87). This average agreement can be compared to the study by Scholtens et al. [17] that also had an interobserver agreement of 0.85, in that case across five observers. It should, however, be noted that the classification task in their study was not binary but objects were classified into one of five classes

Table 1: The number of CTCs identified in the data set by each observer.

Observer	SD	MP	TL	ST	MB	BH	MTF	CMS	SB	JP	RH
Number of found CTCs (N_{CTC}), all data, $N = 617$	307	281	303	354	223	327	221	347	318	330	294
Number of found CTCs (N_{CTC}), consensus, $N = 502$	244	244	248	258	206	257	210	260	255	250	245

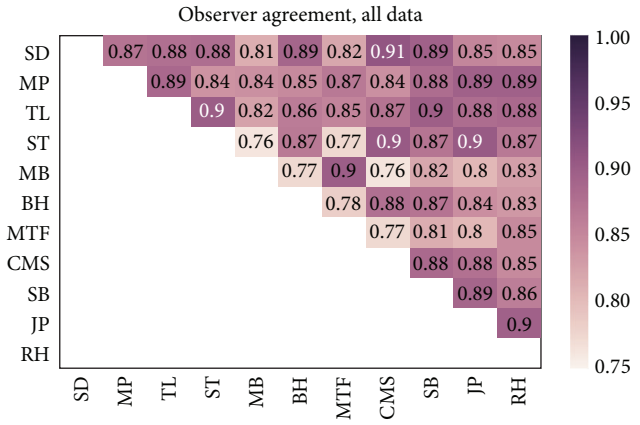


Figure 2: The agreement, P_{Agr}, between observers across all $N = 617$ cutouts.

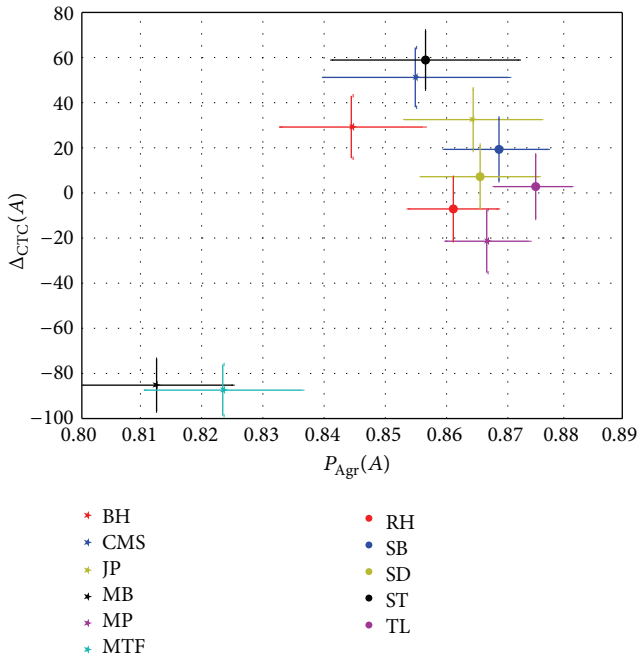


Figure 3: The average agreement for the observers plotted against the difference in number of found CTCs between the observers. Bars indicate the standard errors around the means.

dividing the data into different types of CTCs and other objects including leukocytes. On the other hand, all observers in their study were referred to as experts, whereas in the present study the observers comprise experts as well as non-experts that were asked to identify CTC for the first time according to the criteria described in Section 2.

In Figure 3 we present the average agreement per observer against the average difference in identified CTCs between one specific observer and all other observers. The average agreement between observer A and the others is defined as

$$P_{\text{Agr}}(A) = \frac{1}{N_{\text{obs}} - 1} \sum_{B \in (\text{observers} \neq A)} P_{\text{Agr}}(A, B) \quad (2)$$

and the mean difference in the number of CTCs found for observer A against all other observers is given by

$$\Delta_{\text{CTC}}(A) = \frac{1}{N_{\text{obs}} - 1} \sum_{B \in (\text{observers} \neq A)} N_{\text{CTC}}(A) - N_{\text{CTC}}(B). \quad (3)$$

It can be seen from the clustering in Figure 3 that the two observers avoiding false positives in their indication of CTCs are isolated from the rest of the observers in both dimensions. Even though all participants were given identical instructions, both written and orally, these two observers made a different interpretation on how to annotate the data compared to the other nine observers. While the majority of observers tried to make a guess on cases where they were unsure, the observers MTF and MB always went for no CTC when unsure. For the two observers which avoided false positive CTC annotation the difference in the number of found CTCs seems to be the underlying reason for the low agreement with the other observers. The difference or similarity in number of identified CTCs however does not uniquely predict observer agreement. As an example we consider the two observers SD and TL, who have indicated 307 and 303 CTCs, respectively, that have an agreement of 0.88. On the other hand the observers MP and JP had an agreement of 0.89 although JP identified 49 more CTCs than MP. This emphasizes the need for a multidimensional and a multiobserver analysis regarding the interobserver agreement, rather than just looking at pairwise agreement and averages to identify observers with different biases.

The average agreement between any pair of observers was 85%; that is, $P(A = B) = 0.85$, and the assumption that the probability of agreement would be equal for each image cutout can be inserted into the Bernoulli distribution

$$P(A = B) = \binom{2}{2} p^2 = 0.85, \quad (4)$$

resulting in $p = \sqrt{0.85} \approx 0.92$. Based on this value, we estimate that all eleven observers should agree in $617 \cdot p^{11} \approx 259$ of the cases. In our dataset, all eleven observers agreed on 365 data points and we refer to these data points as the total consensus data set. This implies in turn an average pairwise agreement of $P(A = B) = 0.91$, which is significantly different ($p < 10^{-13}$, Student's t-test) from the measured agreements in Figure 2. From these considerations we can draw the conclusion that the probability for disagreement is not the same for all image cutouts. To exemplify this, we in Figure 1 show a cutout for which all observers agreed of having a CTC (a) and one for which all agreed that there is no CTC (b). In the first case the conditions for a CTC are clearly fulfilled with the strong green staining and the clear integrity of the nucleus shown by the blue staining. The flanking objects were apparently not disturbing the observers. In Figure 1(b) the red staining identifies the object as a blood cell and all observers agreed that this is not a CTC. The third case, Figure 1(c), shows an example where the decision was split six versus five. The staining intensity in this cutout is lower than for the other cutouts and it is therefore hard to verify the integrity of the nucleus. Furthermore, it is difficult to determine if the green staining is structured enough for a positive CTC classification. It is also quite possible that some observers did not see the green staining at all due to the low color intensity.

3.2. Interobserver Agreement Does Not on Average Exceed 93% for Consensus Data. The requirement that all observers should agree may be unnecessarily harsh as we may then discard data that a single observer made a mistake on. In studies where observers are not well supervised and possibly anonymous, as in the case of citizen science projects [32, 33], a single observer that misunderstands the task (or for some reason willingly gives false annotations) can severely damage the integrity of the data set. To determine how many observers we require to vote either CTC or no CTC, we defined a consensus limit, c, for which we say that consensus was reached. As the decision between CTC or no CTC is binary, we required that for the $N_{\text{obs}} = 11$ observers

$$P_{\text{bin}} = \sum_{i=1}^{N_{\text{obs}}} \binom{N_{\text{obs}}}{c} \left(\frac{1}{2}\right)^{N_{\text{obs}}} < 0.05; \tag{5}$$

that is, the probability that c observers by chance annotated the cutout as containing a CTC or not should be less than 5%. In our case this means that $c = 9$ observers had to agree that the cutout does or does not contain a CTC for consensus to be reached and in our data set consensus was reached in 502 of the 617 cutouts (81%). For the consensus data set the interobserver agreement was naturally higher with mean (median) of 0.93 (0.95). In Figure 4 the agreement between observers for the consensus data set is shown as a heat map.

In the case of consensus data points, the observers that avoided false positive CTC annotation again had considerably lower number of CTCs than the other nine observers; see Table 1. Excluding the two observers with the no false positive bias (MB and MTF), the other nine observers are identified between 244 and 260 CTCs which gave a variation of around 3% of the total number of cutouts presented.

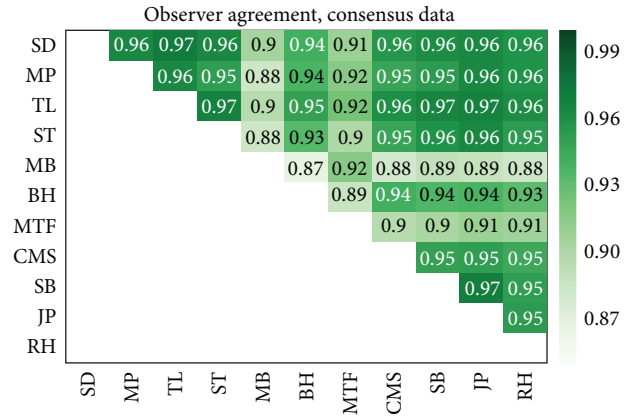


Figure 4: The agreement between observers for the 502 data points of the consensus data set.

Given the distinctly different number of CTCs (see Table 1) identified by observers MTF and MB and their deviation from consensus (see Figure 4) the hypothesis that these two observers had a different bias than the others is further validated. However, instead of discarding the two observers as outliers, we decided that it may be rather interesting to see how annotations that arise from different biases affect the training of automated classifiers. In a setting where fewer observers are used it may not be possible to identify such differences in bias and it is also not sure that the differences in bias is restricted to a clear minority of observers.

3.3. Performance of Automated Classification Strongly Affected by Annotation Ambiguities. When evaluating automated classifiers different performance measures are used to show their agreement with an annotation considered to be ground truth. We have so far demonstrated that for certain data sets the annotation can vary strongly depending on the observer performing the annotation. The performance measures we use to evaluate the automated classifiers are defined with the help of correctly identified CTCs (TP), falsely identified CTCs (FP), objects correctly identified as not CTCs (TN), and CTCs that were not identified as such (FN). Our performance measures are then defined as accuracy Acc:

$$\text{Acc} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FP} + \text{TN} + \text{FN}}, \tag{6}$$

precision Pre:

$$\text{Pre} = \frac{\text{TP}}{\text{TP} + \text{FP}}, \tag{7}$$

and recall Rec:

$$\text{Rec} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{8}$$

Here, accuracy quantifies the fraction of correctly classified data points relative to all data points, whereas a high precision

Table 2: The performance of the classifiers when trained on two folds from probGT and GT in different combinations. It is cyclically tested on one GT fold that was not used for training.

	Training on two GT folds	Training on one GT fold and one probGT fold	Training on two probGT folds	Entire data set (across observers)
RF	Acc: 0.98 ± 0.00	Acc: 0.96 ± 0.01	Acc: 0.94 ± 0.02	Acc: 0.86 ± 0.04
	Pre: 0.98 ± 0.00	Pre: 0.94 ± 0.02	Pre: 0.89 ± 0.05	Pre: 0.83 ± 0.06
	Rec: 0.98 ± 0.00	Rec: 0.98 ± 0.01	Rec: 0.97 ± 0.03	Rec: 0.88 ± 0.08
SVM	Acc: 0.96 ± 0.00	Acc: 0.92 ± 0.01	Acc: 0.81 ± 0.04	Acc: 0.86 ± 0.03
	Pre: 0.95 ± 0.00	Pre: 0.88 ± 0.02	Pre: 0.75 ± 0.08	Pre: 0.85 ± 0.05
	Rec: 0.96 ± 0.00	Rec: 0.95 ± 0.01	Rec: 0.84 ± 0.07	Rec: 0.85 ± 0.08

(recall) indicates a low number of falsely identified CTCs (missed CTCs).

In our earlier study [18], a support vector machine (SVM) achieved accuracy Acc = 0.89, precision Pre = 0.87, and recall Rec = 0.93 on the data set used here, given an annotation of data points performed by one observer (CMS). In the same study, a naïve Bayesian classifier (NBC) was trained without the use of labels, also known as unsupervised learning, which achieved accuracy Acc = 0.87, precision Pre = 0.85, and recall Rec = 0.92.

Our results from the interobserver variability study indicate that a different observer might have annotated the data quite differently. We divided the data set into five randomized folds, without any regard to whether the observers agreed on data points and train a random forest (RF) and a SVM on 3 of those and test on 1 fold. We got the average performance measures Acc = 0.86 ± 0.04, Pre = 0.83 ± 0.06, and Rec = 0.88 ± 0.08 across observers for the RF and the performance measures Acc = 0.86 ± 0.03, Pre = 0.85 ± 0.05, and Rec = 0.85 ± 0.08 for the SVM (see Table 2). Thus, the performance of the SVM and NBC in our previous study [18] was within one standard deviation of the numbers found here, for both the RF and the SVM. It should be noted that besides different implementations of the classifiers and the fact that only one annotation was used in [18], different features were also used. In our previous study, the features used were one-dimensional color histograms while in the present study the hue and saturation channels of HSV images were used. Taken together, we have used three automated classifiers (one RF, two SVM implementations, and one NBC) that performed almost equal on the data set. To add to this, the average interobserver variability was conspicuously close the accuracy of the classifiers, strongly suggesting that the performance of the classifiers was strongly influenced by annotation ambiguities.

To examine if and how differences in annotation affected the classifiers' performance, we split the data set into the total consensus data set that can be considered ground truth (GT) with 365 data points and a part with probabilistic annotation (probGT) containing the remaining 252 data points. From probGT different annotations can be generated by assigning the label for each data point from a randomly chosen observer. On average 81 data points will change label between two probabilistic annotations. For classifier evaluation, GT was in turn split into three folds and the probGT into two folds, giving in total five folds with approximately the same number of cutouts. To get prediction by the classifiers we trained on two folds and tested on a third fold. The test fold was always one of the GT folds as we were here trying to separate the effects of uncertain labels in the test set from uncertainty in training labels. Averages and standard deviations were obtained by 50 repetitions of the training and testing across the folds with new annotations drawn for probGT between each repetition. When training on only GT folds, which do not change any labels between repetitions, we repeated the procedure 10 times to check if any randomness originated in the training of the classifiers.

In Table 2 the performances of the classifiers are listed as we introduced different amounts of uncertainty in the training data. If training and testing were done only on the GT part of the data, the RF achieved performance measures Acc = 0.98 ± 0.00, Pre = 0.98 ± 0.00, and Rec = 0.98 ± 0.00, whereas the SVM achieved performance measures Acc = 0.96 ± 0.00, Pre = 0.95 ± 0.00, and Rec = 0.96 ± 0.00. The standard deviations were less than 1% confirming that both classifiers were stable between training runs and any deviations of this magnitude would originate from annotation changes in the probGT folds. The RF performances did vary in the order of 0.1%, which is due to the probabilistic build of the forest. Compared with the values achieved on the full data set this was a clear improvement when we tested and trained on noise-free data.

If we, instead of training only on GT, took one fold from GT and one from probGT and then tested on one GT fold the RF achieved performance measures Acc = 0.96 ± 0.01, Pre = 0.94 ± 0.02, and Rec = 0.98 ± 0.01 and the SVM achieved performance measures Acc = 0.92 ± 0.01, Pre = 0.88 ± 0.02, and Rec = 0.95 ± 0.01. This means that the label noise during training generally decreased the performance with a stronger performance reduction for the SVM than for the RF. The performances of both classifiers were still better than that recorded on the entire data set where testing was done against partly probabilistic annotation. Even when we trained the RF on the two probGT folds, which we know has a high degree of label noise and it can be assumed that the data in probGT is of a lower quality than in GT, its performance measures were Acc = 0.94 ± 0.02, Pre = 0.89 ± 0.05, and Rec = 0.97 ± 0.03. An example of what we refer to as low quality data is the low

color intensity cutout shown in Figure 1(c). For this setting the performance of the SVM clearly dropped to Acc = 0.81 ± 0.04, Pre = 0.75 ± 0.08, and Rec = 0.84 ± 0.07.

This nicely illustrates that the RF is more robust when faced with label noise than many other classifiers, as was shown in the comparison between RFs and decision trees by Breiman [30]. While the SVM performed well in the pure GT case, its performance dropped more rapidly than the RF when uncertainty was introduced. When the training data contained at least 50% certain cases the SVM still performed better than it did on the entire data set, but when only probGT was used the SVM dropped to considerably lower levels. For the RF the performance level seen for the whole data set was mainly because the classifier is tested on unreliable annotation; that is, the training on unreliable labels did have an effect but that is fairly mild in comparison.

3.4. Consensus Data Provides a Base for Classifier Evaluation. To at least partly solve the issue of uncertain annotation affecting the performance of the automated classifiers we evaluated the classifiers against the consensus data set. As discussed earlier, it is reasonable that the consensus data set is defined as cutouts for which at least nine out of eleven observers agree with each other, because in this case the probability for random annotation of cutouts as containing a CTC or not is less than 5%. In the case of five observers it would be required that all five observers agree in order to satisfy this condition. Thus, the consensus limit varies with the number of observers. When training and evaluating the classifiers against the consensus data set we split the data set of 502 consensus data points into four folds, trained on three of them and tested on the fourth.

In Figure 5 the performances of the manual observers, RF and SVM versus the consensus labeling, are plotted. The performances of the RF and the SVM were close to each other. The SVM had a bit better precision, whereas the RF had a somewhat better recall. The performance measures for the RF were Acc = 0.94, Pre = 0.96, and Rec = 0.93, whereas the SVM had performance measures Acc = 0.94, Pre = 0.95, and Rec = 0.94. In comparison with our earlier study [18], we found an increase of the accuracy by approximately 5%, a precision increase by around 9%, while recall remained unchanged. Given the uncertainty in annotation that has been demonstrated in this study these values are much more representative performance measures for the task of automated classification of fluorescently stained CTCs. The majority of observers had better performances than the automated classifiers, but it should be noted that each observer had a vote when determining the consensus, whereas the RF and SVM did not. It should also be noted that none of the observers reached perfect performance in any of the measures. Hence, there exists not a subset of observers that could have served as a substitute for the consensus annotation.

In summary, the use of a consensus data set for training and evaluation of automated classifiers turned out to be a good option for evaluation of automated classification. In combination with the resilience of the RF to label noise

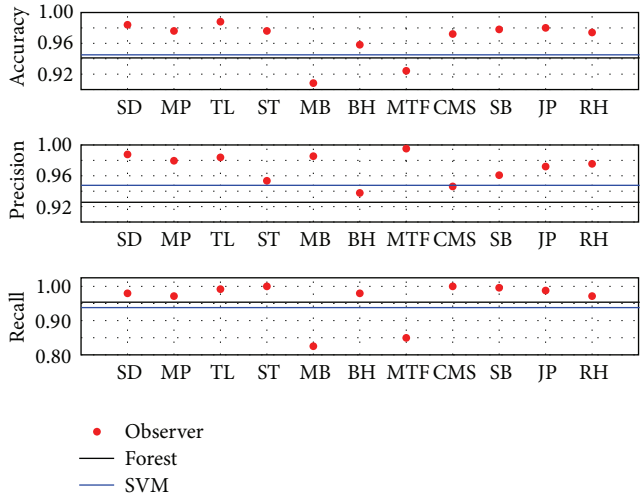


FIGURE 5: The accuracy, precision, and recall of observers (dots), RF (black line), and SVM (blue line) with the consensus annotation of the 502 cutouts for which consensus could be reached.

in the training data, it seems that especially the test set has to be carefully chosen to give a correct evaluation of how well the classifier performs. The issue remains to find a good consensus data set as the manual annotation of data is hard to come by and time-consuming for the observers. This is especially the case when the annotation requires expert knowledge and experience in interpreting, for example, radiology images [28, 29].

4. Conclusions

As the use of computational methods is growing in cell biology, both for classification and modeling of biological systems [17, 18, 34, 35], we have in this paper investigated the effect of label noise caused by uncertain or faulty annotation on the performance of automated classification tasks. In total, eleven observers were asked to manually classify 617 image cutouts that may or may not contain CTCs. The rarity of CTCs in patient blood [7] can easily inflate the accuracy of any classifier due to the many true negatives (TN) that are normally present. The cutouts we used here were identified using a morphological classifier among approximately 35000 foreground objects found during initial image segmentation. The morphological classifier was designed for high recall so that very few CTCs were overlooked at the initial stage, for full details of the procedure see Svensson et al. [18]. For the 617 cutouts, it was revealed that observers agreed with probability 85% whether a CTC was present or not. This degree of agreement is comparable to the uncertainty often seen in manual assessment of medical image data [19, 23–29]. When only considering cutouts on which all observers agreed, the classifiers RF and SVM reached performance measures above 95% (see Table 2). This is considerably higher than the previously reported performances when attempting to automatically classify images of CTCs using RF and SVM [17, 18]. The RF turned out to be quite resilient to noise in the training data, even when using only uncertain data

points in the course of training it performed better on the total consensus test set than classifiers in previous studies (see Table 2). The SVM was more sensitive to label noise in the training data and actually performed worse than it did when the whole data set was used for training and testing. These findings are in line with the findings of Breiman [30] that RFs are stable with regard to noise, although in that study RFs were only compared with decision trees. Going beyond that study, here we have demonstrated that they are also more stable than SVMs with a radial basis function (RBF) kernel. To test classifiers on data which is with a high probability incorrectly annotated or for which it cannot be uniquely decided on the actual class, as is the case for probGT, is of disadvantage for classifiers that cannot be corrected by machine learning algorithms. Any performance improvement above the uncertainty in annotation will be a type of overfitting and even if the achieved performance measures seem impressive the algorithm will most likely not perform well on other data sets. On the other hand, if the test data suffers from label noise it is of great importance to take this into consideration when evaluation any automated classifier.

Two very pressing questions remain to be investigated: (i) how to determine what is good data to use for training and testing the classifiers and (ii) how to detect and treat data that may occur in a clinical setting that is not appropriate for classification using the automated classification. Regarding the first question, we have shown that the creation of a consensus data set is a valid approach, but this normally requires a considerable effort from many observers to make the consensus statistically sound. In many cases these observer also have to be experts, for example, trained physicians that may not be very motivated to annotate data for machine learning algorithms rather than dealing with patients. It can be imagined that machine vision could step in to provide additional observers supplementing human observers. In this case care must be taken that the automated classification can be interpreted as an independent observer that is not getting slaved by human observers. This study suggests that RFs may be a strong candidate for this issue, because we have shown that noise in training data does not strongly affect the RF's performance on a total consensus test set. Another possibility is to use generative models which can be trained without labels [18] and which are therefore independent of the performance of the human observers. For the second question the ideal solution would be if the CTC imaging procedure would be (close to) perfect. In the case of CTC collection using FSMW, as done for the present data set, the cylindrical or spiral shape of the wire presents a considerable imaging challenge to get the entire surface in focus [18]. Even assuming a close-to-perfect data collection technique, it can be expected that clinical use will regularly produce data of a type that was not seen in training of the classifiers. A human observer could in such cases easily conclude that this is an uncertain case, whereas an SVM or RF will be forced to make a decision by design. In the machine learning literature there are methods for outlier detection and these may have to be implemented and developed to handle this task [36, 37]. For outlier detection to be efficient in this classification task, a further subgrouping of objects would probably be needed as the class representing objects that are not CTCs is a very inhomogeneous group of objects.

Instead of simply enumerating CTCs, as done here, it is desirable to determine subgroups within the CTC population, for example, to distinguish between apoptotic and viable CTCs [7, 17]. In order to do this, new sets of features may have to be identified that complement or even replace the color content of the cutouts. Examples of possible features would be further morphological quantities and Fourier-ring descriptors [38]. To apply machine learning to the subgrouping task would require more data than used here and a more rigorous manual classification performed by experts. As Scholtens et al. [17] demonstrated, we would in that case still be faced with a considerable interobserver variability that would require a handling along the lines presented in this study.

Conflict of Interests

The authors declare that there is no conflict of interests regarding the publication of this paper.

Acknowledgments

The authors would like to thank GILUPI GmbH for providing the microscopic images and all eleven observers for the time spent on annotating the 617 cutouts.

References

[1] M. Cristofanilli, D. F. Hayes, G. T. Budd et al., "Circulating tumor cells: a novel prognostic factor for newly diagnosed metastatic breast cancer," *Journal of Clinical Oncology*, vol. 23, no. 7, pp. 1420–1430, 2005.

[2] D. F. Hayes, M. Cristofanilli, G. T. Budd et al., "Circulating tumor cells at each follow-up time point during therapy of metastatic breast cancer patients predict progression-free and overall survival," *Clinical Cancer Research*, vol. 12, no. 14, pp. 4218–4224, 2006.

[3] L. Zhang, S. Riethdorf, G. Wu et al., "Meta-analysis of the prognostic value of circulating tumor cells in breast cancer," *Clinical Cancer Research*, vol. 18, no. 20, pp. 5701–5710, 2012.

[4] A. Rolle, R. Günzel, U. Pachmann, B. Willen, K. Höffken, and K. Pachmann, "Increase in number of circulating disseminated epithelial cells after surgery for non-small cell lung cancer monitored by MAINTRAC is a predictor for relapse: a preliminary report," *World Journal of Surgical Oncology*, vol. 3, article 18, 2005.

[5] S. J. Cohen, C. J. A. Punt, N. Iannotti et al., "Relationship of circulating tumor cells to tumor response, progression-free survival, and overall survival in patients with metastatic colorectal cancer," *Journal of Clinical Oncology*, vol. 26, no. 19, pp. 3213–3221, 2008.

[6] C. Alix-Panabières and K. Pantel, "Challenges in circulating tumour cell research," *Nature Reviews Cancer*, vol. 14, no. 9, pp. 623–631, 2014.

[7] J. Nieva, M. Wendel, M. S. Luttgen et al., "High-definition imaging of circulating tumor cells and associated cellular events

in non-small cell lung cancer patients: a longitudinal analysis," *Physical Biology*, vol. 9, no. 1, Article ID 016004, 2012.

[8] A. G. J. Tibbe, M. C. Miller, and L. W. M. M. Terstappen, "Statistical considerations for enumeration of circulating tumor cells," *Cytometry Part A*, vol. 71, no. 3, pp. 154–162, 2007.

[9] X. Zheng, L. S.-L. Cheung, J. A. Schroeder, L. Jiang, and Y. Zohar, "A high-performance microsystem for isolating circulating tumor cells," *Lab on a Chip*, vol. 11, no. 19, pp. 3269–3276, 2011.

[10] P. R. C. Gascoyne, J. Noshari, T. J. Anderson, and F. F. Becker, "Isolation of rare cells from cell mixtures by dielectrophoresis," *Electrophoresis*, vol. 30, no. 8, pp. 1388–1398, 2009.

[11] W. Sheng, T. Chen, R. Kamath, X. Xiong, W. Tan, and Z. H. Fan, "Aptamer-enabled efficient isolation of cancer cells from whole blood using a microfluidic device," *Analytical Chemistry*, vol. 84, no. 9, pp. 4199–4206, 2012.

[12] G. Chausovsky, M. Luchansky, A. Figer et al., "Expression of cytokeratin 20 in the blood of patients with disseminated carcinoma of the pancreas, colon, stomach, and lung," *Cancer*, vol. 86, no. 11, pp. 2398–2405, 1999.

[13] A. A. S. Bhagat, H. W. Hou, L. D. Li, C. T. Lim, and J. Han, "Pinched flow coupled shear-modulated inertial microfluidics for high-throughput rare blood cell separation," *Lab on a Chip*, vol. 11, pp. 1870–1878, 2011.

[14] K. Tjensvoll, O. Nordgård, and R. Smaaland, "Circulating tumor cells in pancreatic cancer patients: methods of detection and clinical implications," *International Journal of Cancer*, vol. 134, no. 1, pp. 1–8, 2014.

[15] J. den Toonder, "Circulating tumor cells: the grand challenge," *Lab on a Chip*, vol. 11, no. 3, pp. 375–377, 2011.

[16] L. Yu, S. R. Ng, Y. Xu, H. Dong, Y. J. Wang, and C. M. Li, "Advances of lab-on-a-chip in isolation, detection and post-processing of circulating tumour cells," *Lab on a Chip*, vol. 13, no. 16, pp. 3163–3182, 2013.

[17] T. M. Scholtens, F. Schreuder, S. T. Ligthart, J. F. Swennenhuis, J. Greve, and L. W. M. M. Terstappen, "Automated identification of circulating tumor cells by image cytometry," *Cytometry Part A*, vol. 81, no. 2, pp. 138–148, 2012.

[18] C.-M. Svensson, S. Krusekopf, J. Lücke, and M. T. Figge, "Automated detection of circulating tumor cells with naive Bayesian classifiers," *Cytometry Part A*, vol. 85, no. 6, pp. 501–511, 2014.

[19] N. Saucedo-Zeni, S. Mewes, R. Niestroj et al., "A novel method for the *in vivo* isolation of circulating tumor cells from peripheral blood of cancer patients using a functionalized and structured medical wire," *International Journal of Oncology*, vol. 41, no. 4, pp. 1241–1250, 2012.

[20] M. Balzar, M. J. Winter, C. J. de Boer, and S. V. Litvinov, "The biology of the 17-1A antigen (Ep-CAM)," *Journal of Molecular Medicine*, vol. 77, no. 10, pp. 699–712, 1999.

[21] P. T. H. Went, A. Lugli, S. Meier et al., "Frequent EpCAM protein expression in human carcinomas," *Human Pathology*, vol. 35, no. 1, pp. 122–128, 2004.

[22] R. J. Amato, V. Melnikova, Y. Zhang et al., "Epithelial cell adhesion molecule-positive circulating tumor cells as predictive biomarker in patients with prostate cancer," *Urology*, vol. 81, no. 6, pp. 1303–1307, 2013.

[23] K. Doi, "Current status and future potential of computer-aided diagnosis in medical imaging," *The British Journal of Radiology*, vol. 78, supplement 1, pp. S3–S19, 2005.

[24] X. Hu, H. Cammann, H.-A. Meyer, K. Miller, K. Jung, and C. Stephan, "Artificial neural networks and prostate cancer-tools for diagnosis and management," *Nature Reviews Urology*, vol. 10, no. 3, pp. 174–182, 2013.

[25] S. Wang, K. Burtt, B. Turkbey, P. Choyke, and R. M. Summers, "Computer aided-diagnosis of prostate cancer on multiparametric MRI: a technical review of current research," *BioMed Research International*, vol. 2014, Article ID 789561, 11 pages, 2014.

[26] Y. Kominami, S. Yoshida, S. Tanaka et al., "Computer-aided diagnosis of colorectal polyp histology by using a real-time image recognition system and narrow-band imaging magnifying colonoscopy," *Gastrointestinal Endoscopy*, 2015.

[27] G. T. Budd, M. Cristofanilli, M. J. Ellis et al., "Circulating tumor cells versus imaging—predicting overall survival in metastatic breast cancer," *Clinical Cancer Research*, vol. 12, article 6403, 2006.

[28] D. S. Gierada, T. K. Pilgram, M. Ford et al., "Lung cancer: interobserver agreement on interpretation of pulmonary findings at low-dose CT screening," *Radiology*, vol. 246, no. 1, pp. 265–272, 2008.

[29] B. G. Muller, J. H. Shih, S. Sankineni et al., "Prostate cancer: interobserver agreement and accuracy with the revised prostate imaging reporting and data system at multiparametric MR imaging," *Radiology*, 2015.

[30] L. Breiman, "Random forests," *Machine Learning*, vol. 45, no. 1, pp. 5–32, 2001.

[31] M. D. Buhmann, *Radial Basis Functions: Theory and Implementations*, vol. 12 of *Cambridge Monographs on Applied and Computational Mathematics*, Cambridge University Press, Cambridge, UK, 2003.

[32] R. Bonney, C. B. Cooper, J. Dickinson et al., "Citizen science: a developing tool for expanding science knowledge and scientific literacy," *BioScience*, vol. 59, no. 11, pp. 977–984, 2009.

[33] D. Müllensiefen, B. Gingras, J. Musil, and L. Stewart, "The musicality of non-musicians: an index for assessing musical sophistication in the general population," *PLoS ONE*, vol. 9, no. 6, Article ID e101091, 2014.

[34] K. Hünniger, K. Bieber, R. Martin et al., "A second stimulus required for enhanced antifungal activity of human neutrophils in blood is provided by anaphylatoxin C5a," *Journal of Immunology*, vol. 194, no. 3, pp. 1199–1210, 2015.

[35] S. Durmus, T. Çakir, A. Özgür, and R. Guthke, "A review on computational systems biology of pathogen-host interactions," *Frontiers in Microbiology*, vol. 6, article 235, 2015.

[36] M. Kemmler, E. Rodner, E.-S. Wacker, and J. Denzler, "One-class classification with Gaussian processes," *Pattern Recognition*, vol. 46, no. 12, pp. 3507–3518, 2013.

[37] Y. Zhang, N. Meratnia, and P. J. M. Havinga, "Distributed online outlier detection in wireless sensor networks using ellipsoidal support vector machine," *Ad Hoc Networks*, vol. 11, no. 3, pp. 1062–1074, 2013.

[38] T. Emerson, M. Kirby, K. Bethel et al., "Fourier-ring descriptor to characterize rare circulating cells from images generated using immunofluorescence microscopy," *Computerized Medical Imaging and Graphics*, vol. 40, pp. 70–87, 2015.