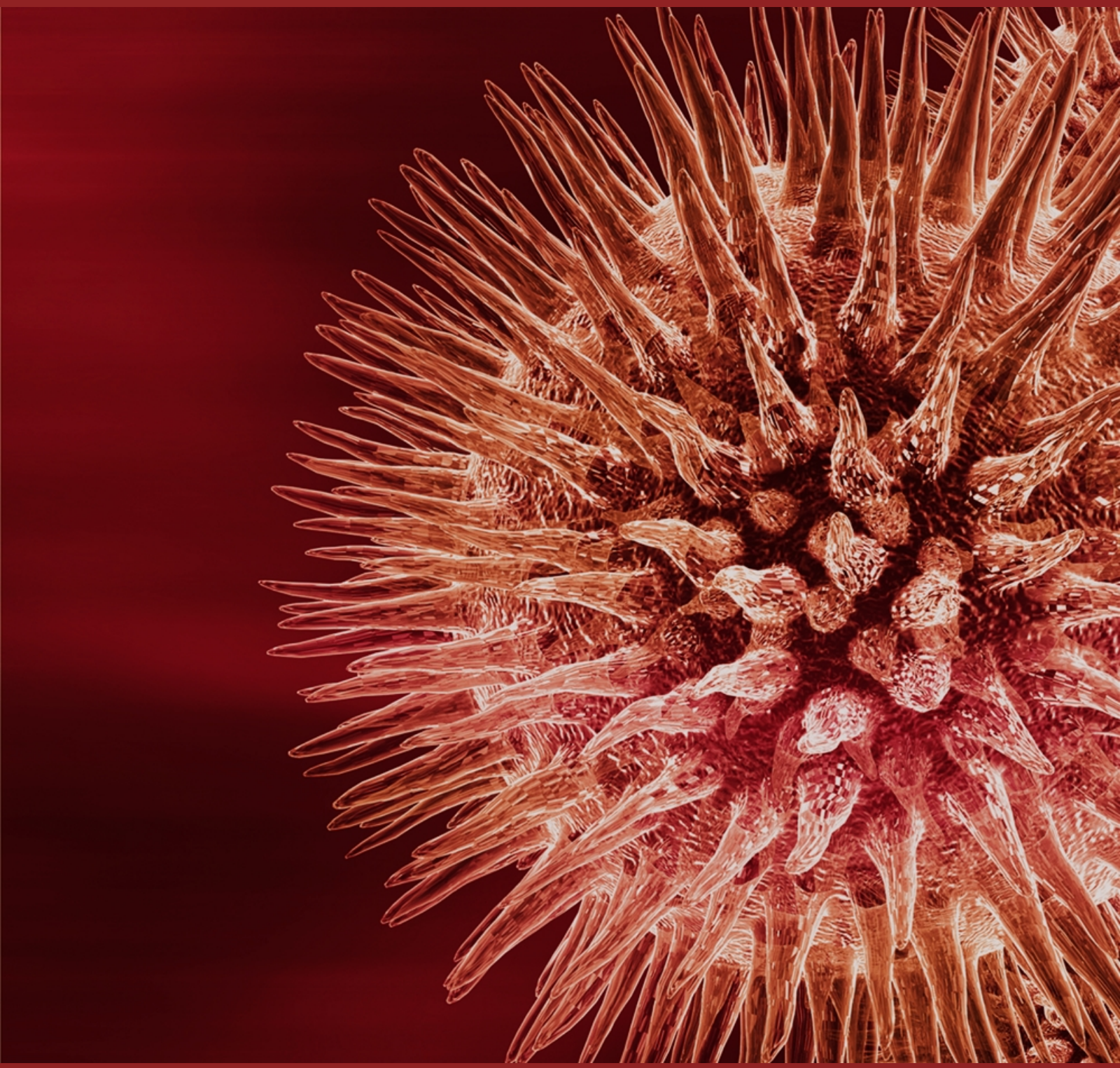
Journal of Biomedicine and Biotechnology

Proteomics in Health and Disease—Part II

Guest Editors: George L. Wright Jr and O. John Semmes



Journal of Biomedicine and Biotechnology

Proteomics in Health and Disease—Part II

Journal of Biomedicine and Biotechnology

Proteomics in Health and Disease—Part II

Guest Editors: George L. Wright Jr and O. John Semmes

Contents

Journal of Biomedicine and Biotechnology • 2003:5 (2003) 257–266 • PII. S1110724303209220 • http://jbb.hindawi.com

REVIEW ARTICLE

Protein and Chemical Microarrays—Powerful Tools for Proteomics

Qingchai Xu and Kit S. Lam*

Division of Hematology and Oncology, Department of Internal Medicine, UC Davis Cancer Center, University of California, Davis, 4501 X Street, Sacramento, CA 95817, USA

Received 11 September 2002; accepted 10 December 2002

In the last few years, protein and chemical microarrays have emerged as two important tools in the field of proteomics. Specific proteins, antibodies, small molecule compounds, peptides, and carbohydrates can now be immobilized on solid surfaces to form high-density microarrays. Depending on their chemical nature, immobilization of these molecules on solid support is accomplished by in situ synthesis, nonspecific adsorption, specific binding, nonspecific chemical ligation, or chemoselective ligation. These arrays of molecules can then be probed with complex analytes such as serum, total cell extracts, and whole blood. Interactions between the analytes and the immobilized array of molecules are evaluated with a number of different detection systems. In this paper, various components, methods, and applications of the protein and chemical microarray systems are reviewed.

INTRODUCTION

Chemical array is a form of the combinatorial library method that was first described by Geysen et al in 1984 [1]. Peptides were synthesized on polyethylene pins in a 96-well footprint and used for B-cell epitope mapping. Enzyme-linked immunosorbent assays were used for such analysis. In 1991, Foder et al of Affymax, Inc (Palto Alto, Calif) [2] reported the use of photolithography in conjunction with light-directed peptide synthesis to generate 1024 peptide on a 1.6-cm^2 glass surface and the use of fluorescent microscopy to analyze interactions between the peptides and fluorescent-labeled antibodies. In 1992, Frank described the synthesis of a peptide array as spots on paper [3]. These three techniques all use parallel synthesis methods to generate arrays of peptides that are spatially separable and addressable. The identity of each peptide spot is known prior to any biological assay. In 1991, we reported the use of the highly efficient "split-mix" synthesis method to generate "one-bead one-compound" (OBOC) combinatorial peptide library (millions of peptides), in which each 80-μm bead displayed only one peptide entity [4]. This spatially separable but nonaddressable peptide microarray or library was screened with an enzyme-linked colorimetric assay, and individual color beads were then physically isolated for microsequencing. These early studies paved the road for the microarray field. In 1994, scientists from Affymetrix (Santa Clara, Calif), a spin-off company from Affymax, reported the use of photolithography/photochemistry approaches to synthesize the first DNA chip [5]. Microarrays of oligonucleotides were synthesized in situ on glass surface, and fluorescent-labeled octanucleotide probes were used to identify the complementary oligonucleotides on this 256 array. In 1995, Brown et al introduced a different form of a DNA microarray chip by using high-speed robot to spot array of different cDNAs on the glass surface. Fluorescent-labeled cDNA derived from mRNA of whole cell lysate was then used to probe the DNA microarray, allowing determination of expression levels of thousands of genes simultaneously [6, 7]. Several automatic arrayers have since become commercially available. The successful application of DNA microarrays to gene expression analysis [8, 9, 10], genetic diagnosis [11], and drug target identification [12], and the rapid development of the proteomics field have propelled many to think about the use of protein or chemical microarray as an efficient tool to evaluate the function of complex protein mixtures. In the last few years, several groups independently developed different chemical and protein microarray methods and applied them to study various biological and chemical problems [13, 14]. In this review, we will give an overview in the field and highlight recent developments.

TYPES OF MICROARRAY

Microarray, sometimes referred to as a chip or arrayed library, can be classified into two general types: biochips (biomicroarray) and chemical microarray. Biochips are usually generated from biochemical or biological components, such as, protein (including enzymes and antibodies), DNA, cell [15], and tissue [16]. Chemical microarrays consist of arrays of organic compounds including small organic molecules, peptides, and sugars. Based

on how chemical microarrays are constructed, they can also be categorized as in situ synthesis array and spotting array. The chemistry of the in situ synthesis approach is more limited, particularly when photochemical reaction is a required synthetic step. As a result, only oligomeric molecules such as oligonucleotides or peptides are used in the in situ synthesis array. A spotting array refers to an array of compounds that are presynthesized and directly transferred and immobilized on a solid surface. This approach is more versatile and can be applied to generate a microarray of almost any molecules.

IN SITU SYNTHESIS MICROARRAY APPROACH

Two typical in situ synthesis approaches are SPOT-synthesis [3] and light-directed parallel synthesis [2]. The SPOT-synthesis method involves parallel peptide synthesis on membrane or paper. In this method, a small volume of solutions containing Fmoc-amino acids plus coupling reagents is dispensed onto the designated spot on the membrane. After the coupling reaction is complete, the whole membrane is washed and the N-terminal protecting group is deprotected prior to the next coupling cycle. Limited by possibility of contamination by reagents from adjacent spots, the distance between each spot cannot be too small. This results in a rather low-density array (eg, 25 spots/cm^2). Production of peptide arrays with this technology has been reviewed recently [17]. Automated SPOT-synthesizers, such as the Auto-Spot Robot ASP222 (Intavis Inc, Germany), are now commercially available.

The light-directed in situ synthesis was initially developed for peptide synthesis [2] but has now been widely used for the synthesis of DNA microarrays. The commercially available Affymetrix chips are prepared by this approach. Figure 1 depicts the deprotection chemistries of light-directed parallel synthesis. The original method uses amino acids with a photolabile protecting group (eg nitroveratryloxycarbonyl, NVOC) as building blocks, and photolithographic method with appropriate masks is used to spatially deprotect the N-terminal protecting group during peptide synthesis. This approach was later adapted to oligonucleotide synthesis [5, 6]. In 1996, McGall et al reported a new photolithographic masking method that involves the use of a polymeric photoresistant film to construct a pattern onto the glass surface [18]. The patterned photoresistant film is used to mask selected regions of the substrate from exposure to standard chemical reagents during synthesis. The main advantage of this approach is that the resolution is superior, and therefore the density of the microarray can increase significantly.

Instead of using photolithographic masks, Singh-Gasson et al [19], reported a maskless light-directed array synthesizer, which utilizes a digital micromirror device consisting of a 600×800 array of micromirrors to form virtual masks. With 1 : 1 imaging, the synthesizer can produce 480,000 pixels of synthetic oligomers in a 10×14 mm area. A similar maskless microarray setup was also reported by Gao et al to synthesize oligonucleotide [20] and peptide [21] microarrays. However, this method is unique in that it uses light to generate acid in situ to deprotect the acid labile protecting groups of the growing chain (4,4′-dimethoxytrityl (DMT) for nucleotides or Boc for amino acids, Figure 1) instead of using photolabile protecting groups. As a result, commercially available standard building blocks for both peptide and oligonucleotide synthesis can be used with this method.

SPOTTING MICROARRAY APPROACH

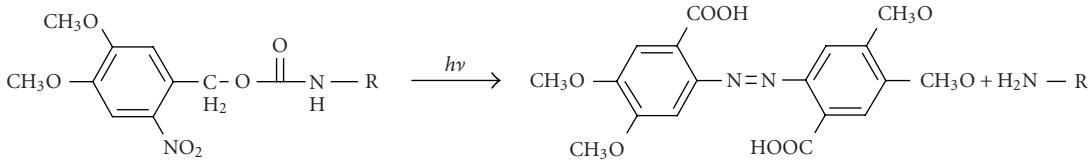
As previously indicated, in situ synthesis methods are mainly used to generate arrays of oligomeric compounds such as oligonucleotide and peptides. In contrast, the spotting microarray approach, in conjunction with a proper immobilization method, can generate many copies of the same chip more efficiently since the compounds need to be synthesized only once. Although macromolecules such as proteins, DNA, and larger peptides can easily be adsorbed onto solid surface through noncovalent interactions, immobilization of small organic molecules, short peptides, or simple sugars will require covalent attachment, preferably through chemoselective site-specific ligation.

IMMOBILIZATION VIA PHYSICAL ADSORPTION

The first and simplest type of immobilization is through surface adsorption. This approach is particularly useful for proteins, and it has been used in standard ELISA, dot blot, and Western blot for many years. The commonly used solid supports are hydrophobic plastics such as polystyrene. Protein microarrays have also been generated on an aminosilane cationic surface [22], nitrocellulose membrane [23], and Hybond ECL membrane [24, 25]. Through adsorption, Wang et al have prepared carbohydrate arrays on nitrocellulose bonded slides [26]. They discovered that the efficiency of adsorption of dextran to nitrocellulose increases with size. The 2000-kd dextran bound significantly stronger than the 20-kd dextran. Wang et al [27] have used HydroGel-coated slides for arraying 43 monoclonal antibodies (mAb) against cytokines and chemokines, and used a fluorescence-based multiplexed immunoassays to quantitate the level of these proteins. We spotted peptide-protein and peptide-agarose conjugates on polystyrene slide as a way to display peptides (Xu and Lam et al, unpublished data, May 2003). Fang et al [28] have developed a membrane protein microarray by spotting cell membrane preparations containing G-protein-coupled receptors on γ-aminopropylsilane-derivatized glass or gold-coated glass surface.

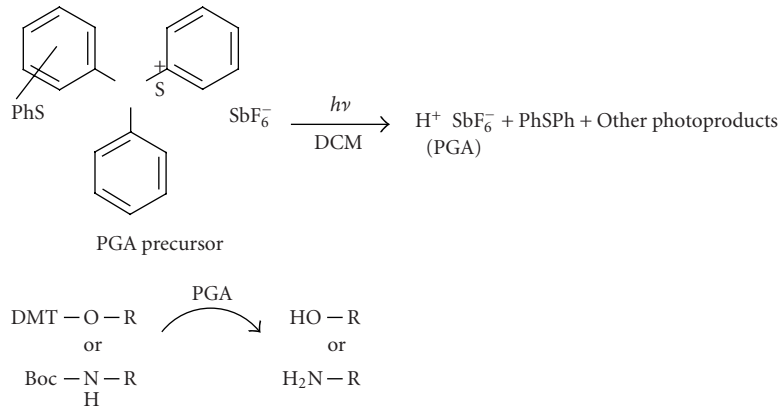
IMMOBILIZATION VIA SPECIFIC SURFACE INTERACTION

In addition to immobilization via nonspecific physical adsorption, molecules can be tagged and immobilized



6-nitroveratryloxycarbonyl (NVOC)

(a) The photolabile protecting group (NVOC) is removed by light illumination.



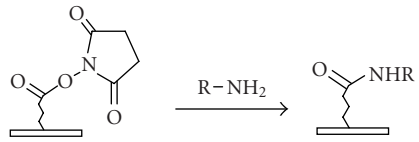
(b) Acid labile protecting groups (DMT and Boc) are removed by a photogenerated acid (PGA).

Figure 1. Deprotection chemistry in two approaches of light-directed parallel synthesis.

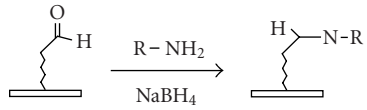
through specific noncovalent interactions between the tag and the already immobilized capturing molecule for the tag. A typical example is the biotin-streptavidin system for immobilizing biotinylated proteins onto streptavidin-coated surfaces [29]. Likewise, a small molecule can also be biotinylated and printed onto a surface that has been precoated with a monolayer of streptavidin or neutravidin. Neutravidin is sometimes preferred as it has less nonspecific interaction with other proteins. We were able to successfully print a microarray of biotinylated synthetic peptides onto a neutravidin-coated polystyrene microscope slide using a Wittech 03 arrayer (Wittech, Inc, Taiwan) [30]. Lesaicherre et al have recently reported [31] the derivatization of glass slides with avidin for immobilization of biotinylated peptides. The poly-His-Ni^{2+} system has also been reported for protein microarray [32]. In this method, proteins containing a poly-histidine tag were printed onto Ni^{2+}-chelating surfaces. Another approach is to use anti-GST antibody or glutathione-coated slides to capture a series of GST-fusion proteins. Boronic acid groups are known to form very stable complex with some moieties such as cis or coaxial 1,2-diol. Immobilized phenylboronic acids have been used in chromatography [33] and protein immobilization [34]. Similar strategies can, in principle, be applied to protein or chemical microarrays as well.

IMMOBILIZATION VIA COVALENT ATTACHMENT

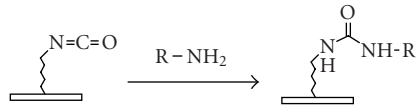
Although nonspecific physical adsorption has been used successfully for generating a microarray of macromolecules, this approach is less useful for the preparation of small molecule or small peptide microarrays. These small molecules can be conjugated to a tag which in turn binds to the immobilized capturing agent (see above). Alternatively, they can be immobilized via covalent attachment to a functional group on the solid surface. Figures 2 and 3 summarize some of the common chemistries used in generating microarrays by covalent attachment. Chemical modification of the solid surface is necessary to create functional groups for covalent immobilization and to achieve homogeneous immobilization. Commercially available aldehyde-derivatized glass slides that have been used for DNA immobilization can also be used for protein microarrays [35]. The aldehyde groups on the glass surface react with primary amines on the protein to form Schiff's base linkages. BSA is used to block the remaining unreacted aldehyde groups or other nonspecific binding sites. Zhu et al [32, 36] have described the use of a 3-glycidoxypropyltrimethoxysilane (GPTS) to activate polydimethylsiloxane (PDMS) on the slide surface prior to protein immobilization. Lin et al have reported [37] the printing of protein microarrays on an aminopropyltrimethoxysilane surface activated with



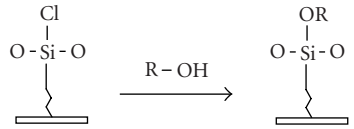
(a) NHS/NH_2 [37, 38].



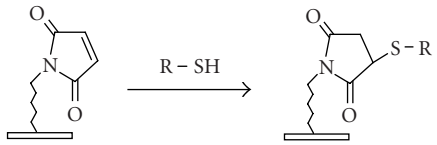
(b) Aldehyde/NH_2 [35].



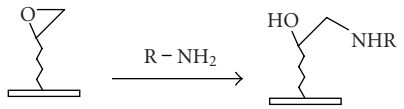
(c) Isocyanate/NH_2 [38].



(d) Chlorinated glass/OH [40].

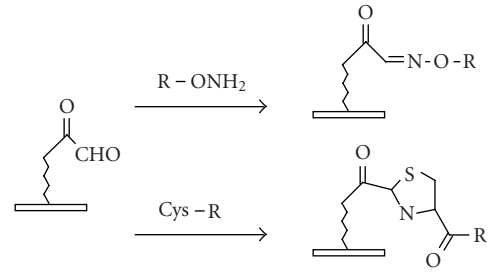


(e) Maleimide/thiol [39].

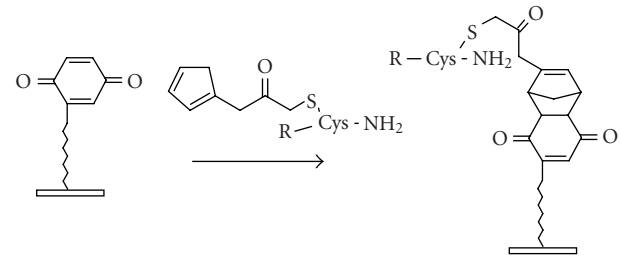


(f) Glycidoxy/NH_2 [36].

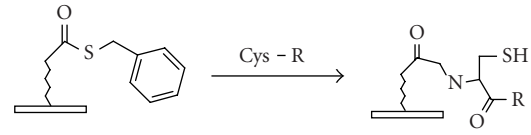
Figure 2. Chemistries of covalent immobilization (nonselective ligation).



(a) Glyoxylyl/NH_2O, N-terminal Cys [42, 43].



(b) Quinine/cyclodiene [44].



(c) Thioester/N-terminal Cys [31].

Figure 3. Chemistries of covalent immobilization (chemoselective ligation).

bis-sulfosuccinimidyl suberate. Furthermore, Benters et al [38] have demonstrated the use of succinimidyl ester- or isocyanate-functionalized dendrimer on a solid surface for nucleic acid and protein microarrays.

Immobilization of small molecules or short peptides often requires covalent linkage of the compounds onto the solid support. Michael addition has been used by Schreiber's group to ligate thiol-containing compounds to maleimide-derivatized glass slides to form a microarray of small molecules [39]. They have also described [40] the covalent attachment of alcohol-containing small molecule compounds onto chlorinated glass slides. In covalent immobilization of molecules, the site of ligation is often important, especially for molecules containing multiple biologically important functional groups. For example, if a peptide requires its free amino and sulfhydryl group to be biologically active, ligation with standard aldehyde, N-hydroxyl succimide, isocyanate, or maleimide chemistries will not work. Carlson and Beal [41] have observed that the attachment point of immobilized peptide-acridine conjugates significantly affects RNA binding. In our group, we have applied the highly selective chemoselective ligation reaction to covalently attach peptides or small molecules onto the glyoxylyl-derivatized glass slide via oxime bond or thiazolidine ring formation [42]. With this chemoselective ligation method, peptides or small molecules are immobilized on the solid support at a predetermined site (aminoxyl or 1,2 aminothiol functional groups), and other α-amino groups, ε-amino groups, or isolated thiol groups present in the molecules will

not react. This eliminates the risk of inactivating the compounds during immobilization. We have recently reported a new strategy to significantly improve the loading of glyoxylyl groups onto glass surfaces by using acrylic acid as a starting material [43]. Further improvement of loading can be accomplished by adding hydrophilic linkers and bifurcating amino acids such as lysine between the glyoxylyl group and the glass surface (Xu and Lam, May 2002). Lesaicherre et al [31] have recently reported the derivatization of glass slides with thioester for chemoselective ligation of peptides with an N-terminal cysteine (1,2 aminothiol group) to form an amide bond. Houseman et al [44] have described a peptide chip prepared by the Diels-Alder-mediated immobilization of peptides on the quinone-functionalized surfaces. In this method, a cyclodiene moiety is incorporated into peptides for the purpose of immobilization. As an alternative to immobilizing peptide-protein conjugates onto plastic slides via physical adsorption, we have used a chemically derivatized protein scaffold (eg, glyoxylyl-functionalized BSA) to first coat the polystyrene slide via nonspecific adsorption. Peptides or small molecules are then printed onto the functionalized protein-coated slides. After incubation, these compounds are immobilized onto the coating surface via a site-specific ligation reaction. In some applications, immobilization of the ligand onto the glass surface via a long hydrophilic linker or a protein may be beneficial for biological activity. For example, we have found that a long linker is needed for phosphorylation of a peptide substrate by $p60^{c\text{-}src}$ protein tyrosine kinase [42].

CD, MICROFLUIDICS, MULTIPLEX-BEADS, AND FIBER-OPTIC MICROARRAYS

Kido et al [45] have developed a compact disc-based microarray system for immunoassays, which uses a piezoelectric inkjet applicator to generate a high-density antigen microarray (75-μm spot) on a polycarbonate disc. Competitive inhibition immunoassays with fluorescent antibodies are performed on the disc, and the final readout is accomplished with a commercially available fluorescence scanner.

Microfluidics is another format of microarray that is based on microchannel systems. Shi et al [46] have developed a radial capillary array electrophoresis microplate system, which consists of 96 separate microfabricated separation channels that connect a central common anode reservoir to 96 injectors at the perimeter of the 10-cm diameter wafer. A laser-excited rotary confocal scanner with four-color detection channels is used for detection. A high-quality restriction map of the 96 samples can be obtained in less than 120 seconds with this device.

Han et al have described [47] the development of quantum-dot-tagged microbeads for multiplexed optical coding of biomolecules. This group immobilizes different peptides, proteins, or nucleic acids on various samples of microbeads into which different-size hydrophobic quantum dots are embedded. These quantum-dot-tagged microbeads are used in suspension in conjunction with flow cytometry to track the identity of the immobilized molecules when mixed with the analytes. These investigators have used DNA hybridization studies to demonstrate that the coding and target signals can be read at a single bead level simultaneously. Similar technology using traditional fluorophores to encode microbeads has been commercialized by Luminex (Austin, Tex).

Walt has pioneered the fiber-optic microarray biosensor technology, which is commercially known as Illumina BeadArray [48]. It involves the patterning of an array of wells at the tip of an optical imaging fiber bundle. The depth of each well can be controlled by etching time and the concentration of hydrofluoric acid used. The fiber-optic tip is then dipped into a suspension of a library of OBOC expressing microbeads (3 micron diameter). The chemicals on the microbeads can be synthesized by "split-mix" synthesis or through parallel synthesis. Alternatively, proteins can be adsorbed on the beads in parallel. Through surface tension, each of the microwells is filled with just one microbead. This random microbead array is decoded [49] for a specific application, prior to analysis of unknown samples. This technology has been applied for the detection of a variety of analytes ranging from DNA to organic vapors [50, 51, 52]. One major advantage of this method is that the packing density of the array is extremely high with center-to-center distance of each bead at 5 μm. However, the necessary decoding step is a major drawback.

PRINTING METHODS AND INSTRUMENTATION

Photolithography, mechanical microspotting, and inkjet application are the three general methods for microarray printing. The mask and maskless photolithography methods [3, 19, 20, 21] that involve in situ light-directed synthesis of oligomers can generate high-quality, high-density microarrays. The mechanical spotting (or surface contact) technique uses a pin to transfer liquid samples to the slides by direct surface contact. The pin-ring technique (eg, GSM 417 arrayer, Affymetrix) and the stamp microcontact technique (eg, SpotBot Personal Microarrayer, TeleChem International, Inc, Sunnyvale, Calif) allow for fairly reproducibly sized spots. The inkjet method employs an electric current to dispense the liquid sample onto the solid support. This noncontact piezoelectric technique accurately and rapidly dispenses monodisperse droplets. Moerman et al [53] have reported the use of electrospraying in a stable cone-jet mode to generate a highly reproducible spot of biological material 130–350 μm in diameter, and as small as 50 pL. More recently, Avseenko et al [54] have used the electrospray method to deposit dry protein onto a dextran-grafted surface, followed by incubation in a 100% humidity chamber. This method was able to generate a $0.6 \times 0.6\,\text{mm}^2$ array with each spot 30–40 μm in diameter and consisting of 28 different protein antigens and allergens. Ringeisen et al [55] have utilized a laser transfer technique, which allows

accurate deposition of picoliter volumes of active proteins onto standard microarray substrates.

MICROARRAY APPLICATION

Table 1 summarizes the various aspects of the microarray technology: attachment chemistries, immobilized molecules, analytes, and detection methods. In most cases, the analytes that have been used to probe the microarray are complex mixtures of molecules: for example, mRNA or cDNA preparations, total cell extracts or intact cells, complex protein mixtures, body fluids such as serum or whole blood, fermentation broths, environmental samples, or microorganisms. In some cases, a pure protein, such as a protein kinase or a protease, is used to probe the peptide microarray to determine the substrate profile of these enzymes. Protein microarrays are frequently used to study different protein functions such as protein-protein interactions. Zhu et al [32] have identified many new calmodulin- and phospholipid-interacting proteins through assaying a microarray of 5800 yeast proteins using known proteins and phospholipids as probes. Similar technique was also used to evaluate protein substrate profile of 119 yeast protein kinases [36]. Sreekumar et al [56] have applied protein microarrays to discover novel radiation-regulated proteins. Huang et al [25] have used antibody arrays to quantitate a large number of different cytokines from sera and culture media. Knezevic et al [57] have reported antibody microarrays to analyze protein expression in cancer tissue of the oral cavity. Ziauddin and Sabatini have reported the use of plasmid DNA microarray to transfect cells in situ to form a new cell microarray with newly expressed proteins encoded by the plasmid DNA [15]. The use of protein microarrays for serum marker detection and discovery using prostate cancer as a model disease has recently been reviewed [58]. Mezzasoma et al [59] have reported the use of microbial antigen arrays to detect serum antibodies against the ToRCH antigens in a panel of characterized human sera. Antibody microarrays have been used for proteomic profiling of the cancer microenvironment [57]. Paweletz et al [60] have reported the development of the so-called reverse-phase protein microarrays to quantitate proteins derived from microdissected tissues. They have immobilized serial dilution of total cell lysates from microdissected tissues on nitrocellulose-coated glass slides and have used different enzyme-linked antibodies to probe specific proteins or phosphoproteins.

Park and Clark [61] have described a sol-gel-encapsulated enzyme array to screen biocatalytic activity or enzyme inhibition. Rakow and Suslick [62] have developed a colorimetric sensor array for detection of volatile chemicals at a concentration below two parts per million. A chemical microarray technique has also been applied to monitor chemical reactions by determining the enantiomeric excess of thousands of samples [63].

All the assay methods developed for on-bead screening of OBOC combinatorial libraries [4, 64] are applicable to chemical microarrays. In our laboratory, we use the OBOC combinatorial library method (100,000 to a few million different compounds per library) to identify ligands or substrates that are of biological interest. The positive hits are then resynthesized and spotted in a microarray format, in multiple replicate sets for subsequent probing with a number of different analytes and under different conditions. In principal, this stepwise approach will enable us to focus our attention on a finite number of ligands that can subsequently be characterized and developed into a diagnostic chip. For example, we have used whole cell binding assay to screen random peptide libraries for leukemia cell binding ligands. The positive ligands are then resynthesized, immobilized on plastic or glass slides, and these peptide microarrays can be used to probe whole blood derived from patients with leukemia [65]. Our ultimate goal is to develop microarrays of cancer targeting peptides that can be probed, allowing physicians to rapidly identify the therapeutic peptide cocktail effective for a specific patient. Similarly, we have used an on-bead functional assay to screen OBOC combinatorial libraries for protein kinase substrates [66]. Unique peptide substrates for a number of protein kinases have been identified using this approach [67, 68, 69, 70]. Immobilization of these peptides with a long hydrophilic linker to form a peptide microarray, in principle, will enable the development of protein kinase substrate chips [42]. These chips can be used to profile the protein kinase activities of whole cell extracts derived from biopsy samples of cancer patients. Other groups have also reported the use of peptide array to profile protein kinase activities [31, 44, 71].

METHODS OF DETECTION

Standard immunodetection techniques, such as enzyme-linked colorimetric, fluorescent, FRET, chemiluminescence, or luminescence methods, are useful to analyze chemical microarrays. As mentioned above, arrayed peptides or protein substrates can be phosphorylated by protein kinases in the presence of [γ^{33}P]ATP, followed by detection with autoradiography or by a phosphor imager [42, 66]. Cell adhesion assays can be performed on a microarray and detected by microscopy using cell staining [42, 65]. One very useful detection method that does not need reporter systems or tags is surface plasmon resonance (SPR) spectroscopy [72]. SPR applies unlabeled probes (eg, antibodies, proteins, and drug candidates) to the surface of a gold-coated glass chip where testing molecules are immobilized. The chip is scanned from below by a light beam. The beam is reflected back by the gold layer, and the angle of reflection varies according to the mass of the molecules attached. Compounds captured by the immobilized array can therefore be located and quantitated by measuring this angle on each array spot or every pixel corresponding to the entire surface of the

Table 1. Summary of microarray methods and detection techniques.

Attachment chemistries
– In situ synthesis via covalent bond: spot synthesis [3, 17]; light-directed parallel synthesis [2, 19, 20, 21].
– Nonspecific adsorption: polystyrene or polymer-coated surface [23, 24, 25, 26]; glass surface, and cationic surface amino groups [22].
– Nonspecific covalent attachment via activated surfaces: aldehyde [35]; succinimidyl ester [37, 38]; isocyanate [38]; glycidoxy [36]; chlorine [40].
– Chemoselective ligation via activated surfaces: glyoxylyl [42, 43]; quinone [44]; thioester [31]; maleimide [39].
Arrayed molecules: mode of immobilization
– Direct link with solid support: DNAs [6, 7]; target proteins [32, 59]; antibodies [6, 19, 25, 20, 57]; alcohol-containing small molecules [40]; carbohydrates [26]; thiolated small molecules [39].
– Indirect link with solid support via immobilized capturing molecules: anti-tag Ab/tagged-target proteins, streptavidin/biotin-target proteins, streptavidin/biotin-carbohydrates, streptavidin/biotin-peptides [31, 65, 42]; streptavidin/biotin-organic molecules [42]; glutathione/GST-fusion proteins, Ni/His tag proteins [32]; gold/thiolated DNAs, peptides, and proteins.
Analytes
mRNAs, cDNAs, total cell extract [57]; protein mixture, body fluid, serum [25, 58, 59]; environmental sample [58]; pure enzyme, microorganisms, and intact cells [42, 65].
Detection methods
Fluorescence, fluorescent-quenching, chemiluminescence, luminescence, FRET, color-dye, enzyme-linked, radiolabel: Phosphoryl imager or autoradiogram [42]; scintillation proximity [71, 72]; protease activity, protein kinase activity [31, 42, 44, 66, 71]; plasmon resonance spectroscopy [70]; whole cell [42, 61]; and cell function [42].

slide. Furthermore, this technology may also provide a measure of the affinity as well as the "on" and "off" rate of binding between the capturing agents and the captured molecules. Morozov et al [73] have described the use of a charge-coupled device to quantitatively detect isotope-labeled ligands bound to a protein microarray. Although it has not been applied in microarray detection, in principle, a homogenous assay such as scintillation proximity assay [74, 75] can also be used. In this case the analyte has to be radiolabeled (eg, by ^{35}S), and the surface of the slide coated with a layer of scintillant.

PERSPECTIVES AND CHALLENGES

Microarray technologies enable the evaluation of thousands to tens of thousands of molecular interactions simultaneously in a high-throughput manner. Microarrays have made significant impact on biology, medicine, drug discovery, and many other related fields and are considered indispensable in genomic and proteomic research pursuits. In the field of drug discovery, microarray techniques can be utilized to identify drug targets that are unique to a disease. Chemical microarray, a form of combinatorial libraries, can also be used for lead identification, as well as optimization of these leads. The impact of microarray on medicine in the future will be significant. Just to mention a few, genetic diseases will be routinely diagnosed and confirmed by gene chips; hundreds to thousands of blood tests will be performed simultaneously on a chip using a few drops of blood from the patient; nucleic acids or proteins derived from a cancer specimen will be analyzed on a chip and a correct diagnosis will be made immediately and, based on the analysis, target-specific anticancer drugs will be prescribed by the physician. In this era of bioterrorism, the development of a chip capable of detecting a multitude of biological or chemical agents in the environment will be of great interest to the law enforcement agencies.

Challenges in this field include the development of novel material or surfaces with minimal nonspecific binding of biological molecule and yet allow for specific ligation of the testing molecule on the surface. Better, more site-specific ligation chemistries for immobilization of synthetic ligands or proteins to the solid support must also be developed. Although it has been reported that compounds derived from individual beads of the OBOC library can be recovered and immobilized on glass surface to form multiple replicates of chemical microarray [76, 77], consistent recovery of enough material from every bead and efficient ligation of the minute amount of material to the glass surface in a site-specific manner remains a big challenge.

Homogenous assays are popular in drug screen but wide application of these detection approaches to microarrays needs to be developed. In principle, the use of plasmon resonance spectroscopy to analyze the entire

microarray for real-time association and dissociation among every spot is feasible, but successful development and commercialization of such an instrument remains to be developed. Mass spectrometry, in principle, can be used to (a) identify the molecular masses of all the captured molecules that bind to each of the microarray spot without the requirement of labeling or amplification [78] and (b) determine the amino acid sequence of some of the peptide fragments obtained from the captured proteins. However, to achieve these goals, a more sensitive mass spectrometer with special labeling techniques and sampling devices will be needed. Microarray technologies have already proven to be invaluable in the field of genomics and proteomics and solutions to these challenges will undoubtedly facilitate the development of clinically useful diagnostic chips in the foreseeable future.

ACKNOWLEDGMENTS

We would like to thank Amanda Enstrom for editorial assistance. This work was supported by Grants NIH CA78868, NIH CA78909, and NIH CA86364.

REFERENCES

[1] Geysen HM, Meloen RH, Barteling SJ. Use of peptide synthesis to probe viral antigens for epitopes to a resolution of a single amino acid. *Proc Natl Acad Sci USA*. 1984;81(13):3998–4002.

[2] Fodor SP, Read JL, Pirrung MC, Stryer L, Lu AT, Solas D. Light-directed, spatially addressable parallel chemical synthesis. *Science*. 1991;251(4995):767–773.

[3] Frank R. Spot-synthesis: An easy technique for the positionally addressable, parallel chemical synthesis on a membrane support. *Tetrahedron*. 1992; 48(42):9217–9232.

[4] Lam KS, Salmon SE, Hersh EM, Hruby VJ, Kazmierski WM, Knapp RJ. A new type of synthetic peptide library for identifying ligand-binding activity. *Nature*. 1991;354(6348):82–84.

[5] Pease AC, Solas D, Sullivan EJ, Cronin MT, Holmes CP, Fodor SPA. Light-generated oligonucleotide arrays for rapid DNA sequence analysis. *Proc Natl Acad Sci USA*. 1994;91(11):5022–5026.

[6] Schena M, Shalon D, Davis RW, Brown PO. Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. 1995;270(5235):467–470.

[7] Schena M, Shalon D, Heller R, Chai A, Brown PO, Davis RW. Parallel human genome analysis: microarray-based expression monitoring of 1000 genes. *Proc Natl Acad Sci USA*. 1996;93(20):10614–10619.

[8] Lockhart DJ, Winzeler EA. Genomics, gene expression and DNA arrays. *Nature*. 2000;405(6788):827–836.

[9] Kurella M, Hsiao LL, Yoshida T, et al. DNA microarray analysis of complex biologic processes. *J Am Soc Nephrol*. 2001;12(5):1072–1078.

[10] Cuzin M. DNA chips: a new tool for genetic analysis and diagnostics. *Transfus Clin Biol*. 2001;8(3):291–296.

[11] Maughan NJ, Lewis FA, Smith V. An introduction to arrays. *J Pathol*. 2001;195(1):3–6.

[12] Marton MJ, DeRisi JL, Bennett HA, et al. Drug target validation and identification of secondary drug target effects using DNA microarrays. *Nat Med*. 1998;4(11):1293–1301.

[13] Lam KS, Renil M. From combinatorial chemistry to chemical microarray. *Curr Opin Chem Biol*. 2002; 6(3):353–358.

[14] Khandurina J, Guttman A. Microchip-based high-throughput screening analysis of combinatorial libraries. *Curr Opin Chem Biol*. 2002;6:359–366.

[15] Ziauddin J, Sabatini DM. Microarrays of cells expressing defined cDNAs. *Nature*. 2001;411(6833): 107–110.

[16] Moch H, Kononen T, Kallioniemi OP, Sauter G. Tissue microarrays: what will they bring to molecular and anatomic pathology? *Adv Anat Pathol*. 2001; 8(1):14–20.

[17] Reineke U, Volkmer-Engert R, Schneider-Mergener J. Applications of peptide arrays prepared by the SPOT-technology. *Curr Opin Biotechnol*. 2001; 12(1):59–64.

[18] McGall G, Labadie J, Brock P, Wallraff G, Nguyen T, Hinsberg W. Light-directed synthesis of high-density oligonucleotide arrays using semiconductor photoresists. *Proc Natl Acad Sci USA*. 1996;93(24): 13555–13560.

[19] Singh-Gasson S, Green RD, Yue Y. et al. Maskless fabrication of light-directed oligonucleotide microarrays using a digital micromirror array. *Nat Biotechnol*. 1999;17(10):974–978.

[20] LeProust E, Pellois JP, Yu P, et al. Digital light-directed synthesis. A microarray platform that permits rapid reaction optimization on a combinatorial basis. *J Comb Chem*. 2000;2(4):349–354.

[21] Pellois JP, Wang W, Gao X. Peptide synthesis based on t-Boc chemistry and solution photogenerated acids. *J Comb Chem*. 2000;2(4):355–360.

[22] Martin BD, Gaber BP, Patterson CH, Turner DC. Direct protein microarray fabrication using a hydrogel stamper. *Langmuir*. 1998;14:3971–3975.

[23] Ge H. UPA, a universal protein array system for quantitative detection of protein-protein, protein-DNA, protein-RNA and protein-ligand interactions. *Nucleic Acids Res*. 2000;28(2):e3.

[24] Huang R-P. Detection of multiple proteins in an antibody-based protein microarray system. *J Immunol Methods*. 2001;255(1-2):1–13.

[25] Huang R-P, Huang R, Fan Y, Lin Y. Simultaneous detection of multiple cytokines from conditioned

media and patient's sera by an antibody-based protein array system. *Anal Biochem*. 2001;294(1):55–62.
[26] Wang D, Liu S, Trummer BJ, Deng C, Wang A. Carbohydrate microarrays for the recognition of cross-reactive molecular markers of microbes and host cells. *Nat Biotechnol*. 2002;20(3):275–281.
[27] Wang CC, Huang R-P, Sommer M, et al. Array-based multiplexed screening and quantitation of human cytokines and chemokines. *J Proteome Res*. 2002;1(4):337–343.
[28] Fang Y, Frutos AG, Lahiri J. Membrane protein microarrays. *J Am Chem Soc*. 2002;124(11):2394–2395.
[29] Ruiz-Taylor LA, Martin TL, Zaugg FG, et al. Monolayers of derivatized poly(L-lysine)-grafted poly(ethylene glycol) on metal oxides as a class of biomolecular interfaces. *Proc Natl Acad Sci USA*. 2001;98(3):852–857.
[30] Aina OH, Sroka TC, Chen ML, Lam KS. Therapeutic cancer targeting peptides. *Biopolymers*. 2002; 66(3):184–199.
[31] Lesaicherre ML, Uttamchandani M, Chen GYJ, Yao SQ. Developing site-specific immobilization strategies of peptides in a microarray. *Bioorg Med Chem Lett*. 2002;12(16):2079–2083.
[32] Zhu H, Bilgin M, Bangham R, et al. Global analysis of protein activities using proteome chips. *Science*. 2001;293(5537):2101–2105.
[33] Singhal RP, DeSilva SSM. Boronate affinity chromatography. *Adv Chromatogr*. 1992;31:293–335.
[34] Stolowitz ML, Ahlem C, Hughes KA, et al. Phenylboronic acid-salicylhydroxamic acid bioconjugates. 1. A novel boronic acid complex for protein immobilization. *Bioconjug Chem*. 2001;12(2):229–239.
[35] MacBeath G, Schreiber SL. Printing proteins as microarrays for high-throughput function determination. *Science*. 2000;289(5485):1760–1763.
[36] Zhu H, Klemic JF, Chang S, et al. Analysis of yeast protein kinases using protein chips. *Nat Genet*. 2000;26(3):283–289.
[37] Lin SC, Tseng FG, Huang HM, Huang CY, Chieng CC. Microsized 2D protein arrays immobilized by micro-stamps and micro-wells for disease diagnosis and drug screening. *Fresenius J Anal Chem*. 2001;371(2):202–208.
[38] Benters R, Niemeyer CM, Wohrle D. Dendrimer-activated solid supports for nucleic acid and protein microarrays. *Chembiochem*. 2001;2(9):686–694.
[39] MacBeath G, Koehler AN, Schreiber SL. Printing small molecules as microarrays and detecting protein-ligand interactions en masse. *J Am Chem Soc*. 1999;121:7967–7968.
[40] Hergenrother PJ, Depew KM, Schreiber SL. Small-molecule microarrays: Covalent attachment and screening of alcohol containing small molecules on glass slides. *J. Am. Chem. Soc*. 2000;122:7849–7850.
[41] Carlson CB, Beal PA. Point of attachment and sequence of immobilized peptide-acridine conjugates control affinity for nucleic acids. *J Am Chem Soc*. 2002;124(29):8510–8511.
[42] Falsey JR, Renil M, Park S, Li S, Lam KS. Peptide and small molecule microarray for high throughput cell adhesion and functional assays. *Bioconjug Chem*. 2001;12(3):346–353.
[43] Xu Q, Lam KS. An efficient approach to prepare glyoxylyl functionality on solid-support. *Tetrahedron Letters*. 2002;43(25):4435–4437.
[44] Houseman BT, Huh JH, Kron SJ, Mrksich M. Peptide chips for the quantitative evaluation of protein kinase activity. *Nat Biotechnol*. 2002;20(3):270–274.
[45] Kido H, Maquieira A, Hammock BD. Disc-based immunoassay microarrays. *Anal. Chim. Acta*. 2000;411(1-2):1–11.
[46] Shi Y, Simpson PC, Scherer JR, et al. Radial capillary array electrophoresis microplate and scanner for high-performance nucleic acid analysis. *Anal Chem*. 1999;71(23):5354–5361.
[47] Han M, Gao X, Su JZ, Nie S. Quantum-dot-tagged microbeads for multiplexed optical coding of biomolecules. *Nat Biotechnol*. 2001;19(7):631–635.
[48] Walt DR. Techview: molecular biology. Bead-based fiber-optic arrays. *Science*. 2000;287(5452):451–452.
[49] Albert KJ, Gill DS, Pearce TC, Walt DR. Automatic decoding of sensor types within randomly ordered, high-density optical sensor arrays. *Anal Bioanal Chem*. 2002;373(8):792–802.
[50] Albert KJ, Walt DR. High-speed fluorescence detection of explosives-like vapors. *Anal Chem*. 2000;72(9):1947–1955.
[51] Ferguson JA, Steemers FJ, Walt DR. High-density fiber-optic DNA random microsphere array. *Anal Chem*. 2000;72(22):5618–5624.
[52] Stitzel SE, Cowen LJ, Albert KJ, Walt DR. Array-to-array transfer of an artificial nose classifier. *Anal Chem*. 2001;73(21):5266–5271.
[53] Moerman R, Frank J, Marijnissen JCM, Schalkhammer TGM, van Dedem GWK. Miniaturized electrospraying as a technique for the production of microarrays of reproducible micrometer-sized protein spots. *Anal Chem*. 2001;73(10):2183–2189.
[54] Avseenko NV, Morozova TY, Ataullakhanov FI, Morozov VN. Immunoassay with multicomponent protein microarrays fabricated by electrospray deposition. *Anal Chem*. 2002;74(5):927–933.
[55] Ringeisen BR, Wu PK, Kim H, et al. Picoliter-scale protein microarrays by laser direct write. *Biotechnol Prog*. 2002;18(5):1126–1129.
[56] Sreekumar A, Nyati MK, Varambally S, et al. Profiling of cancer cells using protein microarrays: discovery of novel radiation-regulated proteins. *Cancer Res*. 2001;61(20):7585–7593.
[57] Knezevic V, Leethanakul C, Bichsel VE, et al. Proteomic profiling of the cancer microenvironment by antibody arrays. *Proteomics*. 2001;1(10):1271–1278.
[58] Miller JC, Butler EB, Teh BS, Haab BB. The application of protein microarrays to serum diagnostics: prostate cancer as a test case. *Dis Markers*. 2001; 17(4):225–234.

[59] Mezzasoma L, Bacarese-Hamilton T, Di Cristina M, Rossi R, Bistoni F, Crisanti A. Antigen microarrays for serodiagnosis of infectious diseases. *Clin Chem*. 2002;48(1):121–130.
[60] Paweletz CP, Charboneau L, Bichsel VE, et al. Reverse phase protein microarrays which capture disease progression show activation of pro-survival pathways at the cancer invasion front. *Oncogene*. 2001;20(16):1981–1989.
[61] Park CB, Clark DS. Sol-gel encapsulated enzyme arrays for high-throughput screening of biocatalytic activity. *Biotechnol Bioeng*. 2002;78(2):229–235.
[62] Rakow NA, Suslick KS. A colorimetric sensor array for odour visualization. *Nature*. 2000;406(6797):710–713.
[63] Korbel GA, Lalic G, Shair MD. Reaction microarrays: a method for rapidly determining the enantiomeric excess of thousands of samples. *J Am Chem Soc*. 2001;123(2):361–362.
[64] Lam KS, Lebl M, Krchnak V. The "one-bead-one-compound" combinatorial library method. *Chem Rev*. 1997;97(2):411–448.
[65] Healey BG, Walt DR. Fast temporal response fiber-optic chemical sensors based on the photodeposition of micrometer-scale polymer arrays. *Anal Chem*. 1997;69(11):2213–2216.
[66] Wu J, Ma QN, Lam KS. Identifying substrate motifs of protein kinases by a random library approach. *Biochemistry*. 1994;33(49):14825–14833.
[67] Lam KS, Wu J, Lou Q. Identification and characterization of a novel synthetic peptide substrate specific for src-family protein tyrosine kinases. *Int J Pept Protein Res*. 1995;45(6):587–592.
[68] Lou Q, Leftwich ME, Lam KS. Identification of GIYWHHY as a novel peptide substrate for human $p60^{c\text{-}src}$ protein tyrosine kinase. *Bioorg Med Chem*. 1996;4(5):677–682.
[69] Wu JJ, Phan H, Lam KS. Comparison of the intrinsic kinase activity and substrate specificity of c-Abl and Bcr-Abl. *Bioorg Med Chem Lett*. 1998;8(17):2279–2284.
[70] Wu JJ, Afar DE, Phan H, Witte ON, Lam KS. Recognition of multiple substrate motifs by the c-Abl protein tyrosine kinase. *Comb Chem High Throughput Screen*. 2002;5(1):83–91.
[71] Lesaicherre ML, Uttamchandani M, Chen GY, Yao SQ. Antibody-based fluorescence detection of kinase activity on a peptide array. *Bioorg Med Chem Lett*. 2002;12(16):2085–2088.
[72] Rich RL, Myszka DG. Advances in surface plasmon resonance biosensor analysis. *Curr Opin Biotechnol*. 2000;11(1):54–61.
[73] Morozov VN, Gavryushkin AV, Deev AA. Direct detection of isotopically labeled metabolites bound to a protein microarray using a charge-coupled device. *J Biochem Biophys Methods*. 2002;51(1):57–67.
[74] Cook ND. Scintillation proximity assays: a versatile high throughput screening technology. *Drug Discov Today*. 1996;1(7):287–294.
[75] Picardo M, Hughes KT. Scintillation proximity assays. In: *High Throughput Screening*. New York, NY: Marcel Dekker; 1997:307–316.
[76] Blackwell HE, Pérez L, Stavenger RA, et al. A one-bead, one-stock solution approach to chemical genetics: part 1. *Chem Biol*. 2001;8(12):1167–1182.
[77] Clemons PA, Koehler AN, Wagner BK, et al. A one-bead, one-stock solution approach to chemical genetics: part 2. *Chem Biol*. 2001;8(12):1183–1195.
[78] Fung ET, Thulasiraman V, Weinberger SR, Dalmasso EA. Protein biochips for differential profiling. *Curr Opin Biotechnol*. 2001;12(1):65–69.

* Corresponding author.
E-mail: kit.lam@ucdmc.ucdavis.edu
Fax: +1 916 734 7946; Tel: +1 916 734 8012

Journal of Biomedicine and Biotechnology • 2003:5 (2003) 267–290 • PII. S1110724303209232 • http://jbb.hindawi.com

REVIEW ARTICLE

Proteomics in Vaccinology and Immunobiology: An Informatics Perspective of the Immunone

Irini A. Doytchinova, Paul Taylor, and Darren R. Flower*

Edward Jenner Institute for Vaccine Research, High Street, Compton, Berkshire, RG20 7NN, UK

Received 5 July 2002; accepted 18 December 2002

The postgenomic era, as manifest, inter alia, by proteomics, offers unparalleled opportunities for the efficient discovery of safe, efficacious, and novel subunit vaccines targeting a tranche of modern major diseases. A negative corollary of this opportunity is the risk of becoming overwhelmed by this embarrassment of riches. Informatics techniques, working to address issues of both data management and through prediction to shortcut the experimental process, can be of enormous benefit in leveraging the proteomic revolution. In this disquisition, we evaluate proteomic approaches to the discovery of subunit vaccines, focussing on viral, bacterial, fungal, and parasite systems. We also adumbrate the impact that proteomic analysis of host-pathogen interactions can have. Finally, we review relevant methods to the prediction of immunome, with special emphasis on quantitative methods, and the subcellular localization of proteins within bacteria.

INTRODUCTION

Genomics has changed the world. Or at least, it has changed the intellectual landscape of the biosciences: its implications suggest that we should be able to gain access to information about biological function at a rate, and on a scale, previously beyond our wildest expectations. As ever, our hopes and dreams are yet to be fulfilled. What we can conceive of still far exceeds what can actually be done at the laboratory bench. Experimental science is playing catch up, developing so-called postgenomic strategies that seek to exploit the opportunities created by the information explosion implicit within genomics. Biology is at risk of being overcome by a bewildering deluge of new data on a hitherto unknown scale and of a hitherto unknown complexity. This is clearly both a blessing and a curse; the trick is to tease out useful information from the data with the hope that this will, in its turn, yield first knowledge, and then, ultimately, true understanding and the ability to efficiently manipulate biological systems.

Postgenomic approaches are legion. They include genomic sequencing, transcriptomics, proteomics, and the analysis of protein-protein interactions, as well as applied techniques, such as the high-throughput screening (HTS) for drug candidates, and integrated informatic strategies, including structure-function prediction. The key underlying factor here is parallelization: the ability to address specific questions not on an individual basis, through complex, intricate experiments, but en masse through elegantly conceived procedures that examine not a single biological object but hundreds, thousands, even hundreds of thousands. This is the area of functional genomics. Functional genomics relies implicitly on high-throughput techniques for measuring the mRNA (the transcriptome), protein (the proteome), and metabolite (the metabolome) components of cells, tissues, organs, and whole organisms.

We pause, momentarily, to examine some definitions. The word orismology, which, in English, dates to 1816, is the science of making and defining terms, especially scientific and technical ones. The need for constant and reliable definitions of terms in science is clear but is seldom realized. Orismology, the science of defining technical terms, seeks to address issues such as these. However, as in many areas of life, that which is apparently rigorous is often anything but risible. The meanings of words drift and vary with time and take on new (sub)discipline-dependent meanings. While this is natural and unavoidable, it can, at times, become obfuscatingly discombobulating. We have seen that with the rise of -ologies (psychology, sociology, even lipocalinology [1]) and, more recently, with the explosions of -omes and -omics. The now famous, or perhaps infamous, website http://www.genomicglossaries.com/content/omes.asp lists literally hundreds of different -omes and -omics. From the genome, countless other -omes have arisen, and within this -omic revolution there are many many conflicting definitions, some are useful, some are not. For the sake of completeness, we list a few of the most useful and the most germane to our present discussions.

The genome is the DNA sequence of an organism. The number of sequenced genomes is now large and ever increasing. In the space of a few years the sequencing of a genome has gone from a transcendent achievement capable of stopping the scientific world in its tracks to the

almost mundane, worthy of only a minor mention in a second line journal. In future times, genomic sequencing may simply become a workaday laboratory technique. Within a decade it may become the stuff of postgraduate students' theses; undergraduates might need to sequence a dozen to complete their final year projects.

Transcriptome is the complement of messenger RNAs (mRNAs) transcribed from a genome. This is a dynamic set of proteins, unlike the genome, which is constantly changing with time in response to the conditions experienced by the cell, hence the development of transcriptomics: the analysis, typically using MicroArrays, of mRNA expression profiles.

Proteome is the protein complement of a cell corresponding to the genome and transcriptome described above. Proteomics is the science that has developed to study the proteome. The proteome is, like the transcriptome, highly dynamic. Conceptually, the proteome is biology in a way that neither the genome nor transcriptome could ever be. Proteins make nature function. Genes, as nucleic acid memes (if one is to believe Richard Dawkins and his ilk), are the essence of inheritance, but it is only through the medium of the protein world, that they are able to propagate themselves.

Metabolome is the complement of all low molecular weight molecules present in a cell. As before, the state of the metabolome is highly dependent on the particular physiological or developmental state or the environmental challenge of the cell. We can usefully distinguish between primary and secondary metabolites. Primary metabolites are the intermediates (ATP, amino acids, membrane phospholipids, etc) in the key metabolic pathways of the cell. Secondary metabolites, at least in the context of microbial natural products, are compounds with no explicit role in the internal metabolic economy of the microbe that biosynthesized them. One argument predicates their existence within an evolutionary rationale: secondary metabolites enhance the survival of their producer organisms by binding specifically to macromolecular receptors in competing organisms with a concomitant physiological action. As a consequence of this intrinsic capacity for interaction with biological receptors, made manifest in their size and complexity, natural products will be generally predisposed to form macromolecular complexes. On this basis, and within a drug discovery context, one might expect that natural products would possess a high hit rate when screened and a good chance of high initial activity and selectivity.

Focussing on the size and complexity of the proteome, we continue by briefly looking at the vexed question of gene number. Pre-genome estimates of the size of the human genome have been revised down from an initial "best-guess" figure in excess of 100,000. As this review is being written, estimates of gene number are converging from a preliminary postgenomic estimate of 30,000–40,000 to a more realistic 65,000–70,000. This may also prove to be an underestimate. The proteome is, however, much larger, principally through the existence of splice variants [2], but also due to the existence of protein splicing elements (inteins) which catalyze their own excision from flanking amino acid sequences (exteins) thus creating new proteins in which the exteins are linked directly by a peptide bond [3]. Other mechanisms include posttranslational modifications, cleavage of precursors, and other types of proteolytic activation. Some estimates place the estimated number of proteins encoded by the human genomes to be two to three orders of magnitude higher than the number of genes. In certain senses at least, the proteome is, as we have said, also much more dynamic than the genome; it varies according to the cell type and the functional state of the cell. In addition, the proteome shows characteristic perturbations in response to disease and external stimuli. Proteomics, as a scientific discipline is relatively new, but is based upon rather older techniques, combining sophisticated analytical methods, such as 2*D* electrophoresis and mass spectrometry (MS), with bioinformatics. Thus proteomics is the study of gene expression at a functional level.

Returning once more to definitions, a comprehensive description of the proteome provides not only a catalogue of all proteins encoded by the genome but also data on protein expression under defined conditions, the occurrence of posttranslational modifications and, importantly, the distribution of specific proteins within the cell [4]. A forerunner to the current proteome paradigm was the concept adumbrated by Anderson and Anderson [5]: the "human protein index." They wished to characterize all the proteins expressed by a cell using high-resolution two-dimensional electrophoresis (2DE). They thought that the human protein index would prove useful in clinical chemistry, pathology, and toxicology. In its proteomic form, this conceit has proved all too true.

The biome, and hence biomics, is an overall term encompassing all of these definitions and including informatic approaches as well. An oft-neglected part of the biome is the immunome: the set of antigenic peptides, or possibly immunogenic proteins, within a microorganism, be that virus, bacteria, fungus, or parasite [6, 7]. There are alternative definitions of the immunome that also include immunological receptors and accessory molecules, but in what follows we will restrict discussion to this initial definition. It is also possible to talk of the self-immunome, the set of potentially antigenic self-peptides. This is clearly important within the context of, for example, cancer (the cancer-immunone) and autoimmunity (the autoimmunome), which affect about 30% and 3% of the global population, respectively.

Many -omes are virtual, rather than literal, biological entities. For example, the recently christened chemome, or chemizome, may be defined as the set of all artificially created or natural products that interact with biological targets in the organism. In practice, this set is not bounded. It is not possible to ever derive or find all the molecules that are encompassed by this definition. In contrast, the immunone, at least for a particular

pathogen, can be realized only in the context of a particular, defined host.

The nature of the immunome is clearly dependent upon the host as much as it is on what we will, for convenience, call the pathogen. This is implicit in the term antigenic or immunogenic. A peptide is not antigenic if the immune system does not respond to it. A good example of this is the major histocompatibility complex (MHC) restriction of T-cell responses. A particular MHC allele will have a peptide specificity that may, or may not, overlap, with other expressed alleles, but the total specificity of all individual alleles will not cover the whole possible sequence space of peptides. Thus peptides that do not bind to any of an individual's allelic MHC variants cannot be antigenic within a cellular context. The ability to define the specificity of different MHCs computationally, which we may call in silico immunomics or in silico immunological proteomics for want of a more succinct term, is an important, but eminently realizable goal of immuninformatics, the application of informatics techniques to immunological macromolecules, a newly emergent subdiscipline within bioinformatics. We will return to this key topic later.

From the perspective of human disease, a proper understanding of the immune system is vital. Indeed, the immune system has evolved to combat the threat of infectious disease. Disease is, arguably, the most significant cause of death worldwide, but it is also the greatest source of preventable human mortality, in that, and in contrast to other causes of death, it can be attacked systematically through the use of biological and chemical entities, such as vaccines and drugs, and through the efforts of surgeons and physicians, and through improvements in public health, drinking water, and sanitation. Although it may be argued quite cogently that the greatest benefit to man has come through improved public health, it is clear that drugs and vaccines have made a large contribution. In contrast, other than by dispensing of drugs and other therapies, the contribution made to public wellbeing by trained medics, though more direct, is also relatively small.

Immunology is also pivotal in other areas of human disease. Cancer is often a prey to immunological mechanisms, and the augmentation of the immune response to carcinomas and cancer antigens is a vital area for future development. Likewise, the inappropriate response of the immune system to self-proteins, as manifest in allergy and, more importantly, in autoimmune diseases, is an area where immunotherapy and immunomodulators can be effective. The discovery and development of vaccines is an important component of publically funded healthcare programs throughout the developed and underdeveloped worlds. Most Western countries have a well developed or long standing centres devoted to its study. The Edward Jenner Institute for Vaccine Research is the United Kingdom's contribution to this worldwide movement.

From a wealth creation viewpoint, rather than from a purely humanitarian one, the world human vaccine market is currently only in the region of $5 billion. One must put this figure against the total worldwide annual sales for all human therapeutic drugs of about $350 billion and an annual global investment in R&D of around $30 billion. To put these large numbers into context, this $350-billion figure is comparable to the yearly gross national product of Taiwan, the Netherlands, or Los Angeles County. However, sales in the vaccine market are increasing at around 12% per annum compared to a yearly rate of about 5% for drugs. Likewise, increased concern by consumers regarding both chemical-free food and environmental and animal welfare has led to an increased interest in vaccines within the farm livestock and companion animal health markets, which worth $18 billion and $3 billion respectively [8]. In the aftermath of AIDS, antibiotic resistance, and the threat from bioterrorism, interest in vaccines has increased dramatically from, say, 10–15 years ago when the vaccine industry was floundering. In 1990, there were about 10 companies in the area worldwide, compared today to over 100, although the majority of current vaccine production is still in the hands of only four big players. The design of therapeutic vaccines (pharmaccines) is, then, an active area of research. Novel ways to rationalize and accelerate vaccine discovery are desperately needed however. Advances in molecular biology and computer science are now accelerating candidate vaccine antigens discovery rates.

To bring this introduction full circle, proteomics is poised to make a significant contribution to the elaboration of the immunome, and thus vaccinology. Proteomics is a pivotal discipline, or more accurately disciplines, within functional genomics. It is an umbrella term for the large-scale analysis of proteins. In fact, proteomics encompasses many different methods seeking to identify the protein complement of a cell or tissue at a given time. These include comparing apparent differences between treated and untreated or between normal and diseased samples, the determination of posttranslational modification (the most common of which are glycosylation and phosphorylation), and the large-scale identification of protein-protein interactions. We will begin with a review of experimental approaches to proteomic vaccinology and then we will present an analysis of computational approaches to vaccinology.

PROTEOMICS IN VACCINOLOGY

The discovery of vaccination is generally attributed to Edward Jenner (1749–1823). However, at the beginning of the 18th century, inoculation against smallpox had been brought to England by Lady Mary Wortley Montagu (1689–1762). Lady Mary, who is, perhaps, better known to history as a poet and witty correspondent, was born in London, the eldest child of Evelyn Pierrepont, Earl of Kingston. In 1716, after the accession of the first Hanoverian monarch George I (1660–1727) on the death of the last Stuart monarch Queen Anne (1665–1714), Lady Mary's husband was appointed

Ambassador to Turkey. The Wortley Montagu's long and dangerous transcontinental journey, which was undertaken in the dead of winter was considered something of an achievement at the time. Constantinople was full of wonders which Lady Mary, unlike so many European wives, set out to explore and understand, immersing herself in all Turkish things, even learning the language. She visited the zenanas, meeting the upper class women secluded there, whom she came to admire, and absorbed Turkish customs. Her record of her experiences, *Turkish Embassy Letters*, is a primary source for historians of this period.

The Wortley Montagu's visit occurred during the reign of the sophisticated, cultured, and tulip-obsessed ottoman sultan Ahmed III (1667–1736). His reign marked something of a renaissance for the Ottoman Empire after its relative decline during the 17th century. Influenced by his son-in-law, or damut, vizier Ibrahim Pasha Kulliyesi, Ahmed III increasingly looked to the West, creating the first fire brigade and printing presses in Constantinople and also establishing the Empire's first foreign embassies. Ahmed III reigned from 1703 to 1730, the so-called Tulip Era, or lale devri, a period of rare hedonistic extravagance centering on the sultan's love for the tulip.

Wortley was recalled due to a change in English relations with Turkey, and the family appeared in London in the fall of 1718. Lady Mary discovered that the Turks inoculated healthy children with a weakened strain of smallpox in order to confer immunity from the more virulent strains of the disease, and determined to bring the practice to England. Lady Mary had her own son and daughter inoculated against smallpox, which had killed her brother and left her scarred by her 1715 bout, and thus introduced the custom to the nobility. However, Lady Mary struggled to interest the English medical establishment in inoculation. Their main objection seems to have been to being told by a woman what it was their business to know. While it has become fashionable among feminist revisionists to credit Lady Mary, rather than Jenner, with the discovery of vaccination, this is hardly accurate. While it is important to recognize her contribution, it is important as well to recall that protective immunity has been recognized for several millenia at least; in 430 BC Thucydides, principal historian of the Peloponnesian War, noted that during an Athenian plague only those who had recovered from the plague were able to nurse the sick without themselves falling ill. During the 15th century, both the Chinese and Turks deliberately induced immunity by inhaling dried crusts from smallpox pustules or by inserting the crusts into cuts in the skin.

After a period of first training in London and then working for a time as an army surgeon, Jenner, a native of Gloucestershire, spent his entire career working in the county as a country doctor. Jenner had noted that milkmaids who had contracted cowpox, a related virus, seemed to be immune to smallpox. On 14th May 1796, he introduced the fluid from a cowpox pustule he used to build protective immunity against smallpox in his gardener's 8-year old son. Jenner then infected him with smallpox. The boy did not become ill. Later, Louis Pasteur (1822–1895) adopted "Vaccination," the word Jenner had invented for his treatment (from the Latin *vacca*, a cow), for immunization against any disease. Pasteur also made important empirical advances in vaccination, discovering that chickens injected with attenuated fowl cholera bacteria survived an infection with the virulent form. Later, Pasteur immunized sheep with attenuated anthrax bacillus and challenged them with virulent anthrax and showed that the attenuated anthrax protected the sheep from disease, and in 1885 Louis Pasteur saved the life of a boy bitten by a rabid dog by administering a rabies vaccine he had created. It is now generally accepted that mass vaccination, taking account, as it does, of the principal of herd immunity, is one of the most effective prophylactic approaches to the treatment, or rather, pretreatment, of infectious disease.

However, vaccination has, until relatively recently, been a highly empirical science, relying of poorly understood, nonmechanistic approaches to the development of new vaccines. As a consequence of this, relatively few effective vaccines have been developed and deployed during most of the two centuries that have elapsed since Jenner's work. This has been prompted by, amongst other things, worries over the emergence of antibiotic resistance and, latterly, bioterrorism.

Vaccinology is slowly evolving into immunovaccinology, a discipline that uses the rapid advances in immunological understanding extant within the last few decades to effect a paradigm shift in thinking within the discipline. Reverse immunogenetic approaches offer the tantalizing prospect of short cutting the process of vaccine discovery and also producing safer and more effective vaccines. Postgenomic approaches, of which proteomics is amongst the most prominent, are another broad tranche of techniques which offers much in this context.

Antigenicity or immunogenicity manifests itself within both humoral immunology (mediated primarily through the binding of whole antigens by antibodies) and cellular immunology (mediated by binding of proteolytically cleaved peptides). In the main, we will concentrate our attention on that part of the adaptive immune response that is mediated by T cells. Within the context of cellular immunology, the immunogenicity of peptides strongly depends on their ability to bind to MHC and to be recognized subsequently by T-cell receptors (TCR). Traditionally, T-cell epitopes, the small peptide fragments of whole proteins that cellular immunity recognizes, have been identified by examining the responses of T cells to sets of overlapping peptides generated from target antigens. This is adequate, if labour intensive, for the study of a single, small protein, but the experimental overhead becomes prohibitive for the study of proteomes from large viruses, bacteria, or parasites, which may contain thousands, if not tens of thousands, of gene products.

The application of proteomics, perhaps in combination with transcriptomic approaches, together with

bioinformatics, should allow us to reduce the virtual set of open reading frames (ORFs) apparent within a genome. This set might number a few hundred for viruses, a few thousands for bacteria, or a few tens of thousands for parasitic microorganisms. Leverage of these technologies could reduce this to a manageably short list of candidate vaccines, perhaps numbering no more than a few dozens. Such candidates would then require channeling through a set of subsequent processes including recombinant expression, purification, and testing for immunogenicity and protective efficacy [9].

Hitherto, proteomics has been seen as a primarily analytical science, which combines multidimensional polyacrylamide gel electrophoretic techniques with sensitive biological MS, supported by rapidly growing protein and DNA databases, to effect the high-throughput identification of protein populations from different cell types or cells experiencing different environmental conditions. As we have said, the unambiguously identification of a protein is a prerequisite to their full functional investigation. This identification is usually effected through matrix-assisted laser desorption/ionization mass spectrometry (MALDI-MS), which is one of the current analytical methods for linking sequence databases to gel-separated proteins. There are at least two main MALDI-MS identification methods: peptide mass fingerprinting (PMF) and post-source decay analysis. PMF identifies proteins by comparison of experimentally and theoretically derived profiles of proteolytically digested peptides. Because both experimental data and sequence databases are limited, there usually remains some ambiguity with regard to posttranslational modification(s) and intrinsic sequence variation. Moreover, the role of electroblotting and Edman N-terminal sequencing as tools in protein identification should not be overlooked. As proteins derived from the same gene may be largely identical, and might differ only in limited yet functionally important details, the identification of proteins must not only pinpoint numerous proteins *en masse* but also differentiate between close relatives.

But obviously proteomics is more than a few techniques, however sophisticated. Indeed, it is a cohesive and overarching intellectual environment, replete with ideas, many now beginning to yield advanced, if less established, techniques. The cutting edge of proteomics has much to offer.

Other techniques, such as the yeast two-hybrid system [10], also cower under the proteomic umbrella, but are less relevant to vaccine discovery, and so are excluded from our discussion. Perhaps, the most exciting array of emergent proteomic techniques are the so-called protein arrays [11], where recombinant proteins can be arrayed to study protein-ligand and protein-protein interactions. Based on the rationale that altered abundance or a change in structure of proteins can lead to disease, and although protein arrays are currently more expensive and more technically difficult to produce than nucleotide arrays, protein and antibody arrays are now generating considerable excitement. Alternatively, arrays of protein-specific antibodies can quantitate protein levels, analogous to the detection of mRNA by microarrays [12]. It is to be hoped that as protein arrays become more sophisticated, they will impact on infectious disease research by profiling sera and body fluids to discover prognostic and diagnostic markers of particular infections.

The identification of antigenic or immunogenic proteins as putative whole protein subunit vaccines is a key goal of immunovaccinology. It offers the hope of eliciting significant responses from both humoral and cellular immune systems, far exceeding the efficacy of peptide vaccines, while avoiding potential toxicity problems associated with whole microbe vaccines [13]. Before we continue, however, we must raise a minor caveat: proteomics is, after all, only one part of a much larger postgenomic initiative. While our current focus will be on the role of proteomics in immunovaccinology, it is as well to note that it is a complement, rather than a replacement, for other like-minded technologies, such as genomics and transcriptomics. In order to maximize the potential didactic benefit of reading this review, it should be read in conjunction with other papers which cover other postgenomic techniques as applied to close related areas [12, 13, 14, 15, 16, 17, 18]. In what follows, we will concentrate, on viral, bacterial, fungal, and parasite proteomics, as well as host-pathogen interactions, largely, but not exclusively, within the context of vaccinology. However, we specifically omit discussion of autoimmune disease, cancer, and cancer antigen proteomics. These are primarily host-only proteomics that is clearly beyond our current scope.

Viral proteomics

As we will see in later sections, proteomics of bacterial systems is now well advanced, as cancer proteomics is, a subject we will adumbrate but not describe in great detail. By contrast, direct analysis of viruses has been rather limited. Mass spectrometry has long been used to increase our understanding of structure and function in viral proteins: being used to identify posttranslational modifications and mutants, and characterize individual capsid proteins. For example, VP6, the major structural protein of rotavirus, which makes up its inner capsid, has been studied recently using MALDI-TOF and electrospray ionization mass spectrometry [19]. Emslie et al were able to differentiate serovars of the virus and identify a number of posttranslational modifications, including the N-terminal acetylated methionine and deamidated ASP107. Other mass-based approaches, combined with time-resolved proteolysis (mass mapping), have revealed the dynamic nature of viral particles in solution [20]. More recently, Yao et al have used a novel isotope labelling approach, based on the differential incorporation of ^{18}O, to investigate differences between the proteomes of two serotypes (Ad5 and Ad2) of adenovirus [21].

Most other proteomic studies have examined host-virus interactions. Currently, our understanding of the

effects of virus infection on the proteomes of infected cells is poor. Toda et al profiled proliferative B-lymphoblastoid cell lines infected with Epstein-Barr virus using proteomic techniques and identified a spot, corresponding to the 16-kd protein phosphoprotein stathmin, that decreased significantly in immortalized cells [22]. Diaz and coworkers [23, 24] examined ribosomal modifications induced by herpes simplex virus type 1. Comparison of the highly basic ribosomal protein maps from infected and noninfected cells indicated that virus infection induces unusual phosphorylation of proteins of the small ribosomal subunit, including S2 and S3a, and the large subunits, including protein L30. Their most significant observation was the permanent phosphorylation of ribosomal protein S6, which plays an important role in controlling the translation of mRNAs that code for components of the translation apparatus.

In a proteomic analogue of transcriptomic analysis, Rodriguez et al [25] used electrophoresis to examine protein expression patterns in Vero infected with African swine fever virus (ASFV) attenuated strain BA71V and porcine alveolar macrophages cells treated with the ASFV virulent strain E70. The resultant data sets, for noninfected cells, included 177 basic and 818 acidic polypeptides from the macrophage and 1,127 acidic and 271 basic polypeptides from the Vero cell. Comparison of infected and noninfected proteomes indicated that ASFV infection shuts off protein synthesis for 65% of cellular proteins, while a small number of proteins—28 proteins (macrophages) and 48 proteins (Vero cells)—show a greater than 2-fold increase in expression.

Bacterial proteomics

In the last decade, rapid advancements in sequencing technology have lead to the completion of a whole tranche of bacterial genomes. Two main routes to bacterial genomics have been followed. The first was contingent upon the generation of a physical map using cloned genomic fragments in a phage or plasmid library, with the individual cloned fragments then being sequenced and aligned to the physical map. The genome sequence of *Escherichia coli* was determined in this way [26]. In the second, essentially random fragments of the genome were cloned in plasmid and phage libraries, with the inserts' terminal sequences then determined and the sequenced fragments assembled into the complete genome sequence. This methodology has been used to determine the genome sequences of many other bacteria including *Haemophilus influenzae* [27], *Mycoplasma genitalium* [28], *Methanococcus jannaschii* [29], and *Helicobacter pylori* [30]. The rest, of course, is history. Presently, the number of completed or partially completed sequencing projects is in the region of hundreds, rather than tens, of genomes [31].

Once determined, analysis of genome sequences using gene prediction programs has identified large numbers of ORFs, many were previously unknown. While, it proved possible to assign functions to proteins encoded by the majority of ORFs on the basis of their homology to extant sequences, a significant number of ORFs show no obvious similarity to genes of known function. As we have said, this has led to the development of many postgenomic strategies, such as proteomics, which seek to determine function. Bacteria have special features, generally lacking in other organisms, for proteomic analysis, that result from the abundance of information on their genomes, their low levels of functional redundancy, their relative simplicity of gene regulation, and their experimental tractability.

Within the context of vaccinology, one of the key goals of postgenomic research is to determine differences between two related microbes, or, more generally, cells, or between the same microbe or cell under different growth conditions. Proteomics approaches to this problem have been applied, with particular success, to bacteria. This work includes the determination of the proteomes for several bacterial species: *Salmonella typhimurium* [32], *Bacillus subtilis* [33], and *Mycoplasma pneumoniae* [34].

In another study, *Chlamydia pneumoniae*, an obligate intracellular human pathogen that causes acute and chronic respiratory tract diseases, was cultured in Hep-2 cells and proteins from its infectious elementary bodies were separated by two-dimensional gel electrophoresis [35]. Two hundred sixty-three protein spots were extracted in the pH range 3–11, these corresponded to 167 genes (about 15% of the genome) were identified. The proteins identified included 31 hypothetical proteins including several involved in the type III secretion apparatus, an important mediator of virulence amongst intracellular bacteria, and others involved in energy metabolism. In a related study, global gene expression in *Chlamydia trachomatis* serovars A, D, and L2, each is responsible for a different chlamydial disease, was investigated using proteomics [36]. Seven hundred protein spots were detected, from which 250 proteins, deriving from 144 genes, were identified, again from the elementary body. As well as again identifying proteins associated with the type III secretion system, 25 hypothetical ORFs and 5 polymorphic membrane proteins were also identified. Correlating protein expression with type of serovar suggests ways of tailoring the identification of specific antigens to particular disease states.

In another study on different, but closely related, bacteria, Piechaczek et al examined uropathogenic *E coli* strain 536 and some of its mutants [37]. Differences in proteins expressed by wild-type *E coli* as well as mutants 536delta102, 536-21, and 536R3, which differ in the presence or absence of different pathogenicity islands with their genome, were examined using two-dimensional polyacrylamide gel electrophoresis and MALDI-TOF mass spectrometry. The presence of 39 intracellular proteins with markedly different expression in the different strains was determined, of which 34 could be identified using MALDI-TOF-MS. Comparison of the different derivatives indicated that proteomics was an efficient approach to studying global gene expression and that the expression of various proteins including those

encoded by many housekeeping genes is affected by the presence of different pathogenicity islands. Malhotra et al analyzed two strains, PAO1 and PD0300, of *Pseudomonas aeruginosa* to determine proteins that are differentially expressed as a consequence of mucoid conversion, a process implicated in chronic pulmonary infections in cystic fibrosis [38]. Using proteomic methods, they identified 6 proteins more abundant in mucoid strain PD0300, including 2 implicated in alginate biosynthesis (AlgA and AlgD), porin F, and DsbA (a disulfide bond isomerase).

We will now shift our emphasis and restrict our focus to two particular pathogens: *Mycobacterium tuberculosis* and *H pylori*. This pair of pathogens is chosen not only as a demonstration of what has been done but also as an example of what might easily be achieved for other bacterial pathogens.

Every day, over 5700 people will die from tuberculosis (TB), a chronic bacterial infection. It causes greater morbidity than any other infectious disease and is the only such disease to be declared a "global emergency" by the World Health Organization, yet it is over 40 years since a novel anti-TB drug was introduced. The intracellular pathogen *M tuberculosis*, the causative agent of TB, infects about one-third of the world's population, around 1.7 billion people. Although most infected people do not develop active TB, over 8 million people do develop the disease annually. The rapid spread of AIDS, especially in developing countries, has contributed to the recent sudden escalation in TB cases. This problem is exacerbated by the increased spread of antibiotic- or multidrug-resistant strains of *M tuberculosis*.

One approach to targeting TB is the development of novel antibiotics. For example, in the genomic era, a tranche of new drug targets, including mycobacterial cell wall components, which are vital for bacterial viability, and the metabolic pathways that biosynthesize them, have become available. Vaccines are another important research avenue. Only a few years ago, it was generally accepted that clinical trials of TB vaccines would not occur for at least a decade, yet the first trials are now beginning.

A number of studies, building on early work [39, 40], have begun to build a picture of the TB proteome, and how pathogenic and nonpathogenic strains of TB differ. For example, proteomic approaches can identify novel genes not apparent from automated gene hunting within genome sequences, as has been found for TB [41], where the existence of six ORFs was shown by electrophoresis and MS.

In a ground breaking study, proteomics was used to compare the proteome of two nonvirulent vaccine strains of attenuated *Mycobacterium bovis* Bacillus Calmette-Guerin (BCG) with two virulent strains of *M tuberculosis*. *M tuberculosis* usually resides within the host macrophage, but its mechanisms of survival are poorly understood. Whatever evidence exists suggests that *M bovis* BCG is both a deletion and regulatory mutant, yet retains the ability to live within the macrophage and is immunoprotective, albeit at a relative low efficacy. This leads to the identification of around 25 different proteins, which are either differentially expressed or modified, from a set of 2600 resolved protein spots out of the 3924 ORFs identified in the TB genome [42]. In a more recent study, the same group has identified a number of putative virulence factors and diagnostic markers of TB as well as interesting candidates for vaccination against tuberculosis [43]. About 1800 distinct protein spots were identified by electrophoresis, of which 56 spots were unique to virulent strains and 40 spots to the attenuated strains. Twelve spots specific for *M tuberculosis* were identified as proteins previously shown to be missing from *M bovis* BCG, while 20 *M tuberculosis*-specific spots were identified as genes not previously thought to be deleted in *M bovis* BCG.

Some of these differences seen in this last experiment may reflect differences in environment-dependent expression rather than differences between the complete proteome. In order to investigate this, a number of workers have examined the proteome of *M tuberculosis* and BCG under different conditions. In an early study, Wong et al [44] used proteomics to examine the effect of high and low extracellular iron concentration on the expression of genes in *M tuberculosis*. The expression of 15 proteins was induced, and the expression of 12 proteins was decreased under low-iron conditions. Mass spectrometry identified 10 proteins including fur and aconitase proteins, both of which are regulated by iron in certain bacterial systems. More recently, Monahan et al [45] have tried to define differences in gene expression during the interaction of BCG with macrophage cell line THP-1. They found that BCG resident within macrophages express different proteins than those expressed during growth in culture or under conditions of heat shock. In particular, they identified six abundant proteins with increased macrophage expression: Rv2623, InhA, GroEL-1, GroEL-2, alpha-crystalline, and elongation factor Tu. In a related study, Betts et al [46] have examined a laboratory model of the latent or "persistent" form of TB that may mimic its nongrowing, drug-resistant persistence in vivo. By using microarray and proteome analysis, they investigated the response of a nutrient-starved *M tuberculosis* and identified a number of interesting target proteins. In an earlier study, Betts and coworkers analyzed the recent clinical isolate CDC1551 *M tuberculosis* with laboratory strain H37Rv, which has been subject to in vitro passage, using standard proteomic techniques [47]. Although the two strains demonstrate different in vivo and in vitro phenotypes, visualization of 1750 protein spots indicated that their protein profiles were very similar. Of the 17 protein spot differences, 7 were unique to CDC 1551, 3 to H37Rv, and 2 showed increased expression in H37Rv.

Identification of proteins by a strategy that targets the differences between *M tuberculosis* and BCG, as well as strains grown under different conditions, will help elucidate the molecular basis of attenuation and the vaccine potential of BCG, as well as identifying TB-specific

antigens, virulence factors, and diagnostic biomarkers that can distinguish vaccination by BCG from infection with *M tuberculosis*. Identification of potential subunit vaccines is greatly facilitated using this spot-the-difference technique or alternative proteomic approaches which focus on the identification of secreted proteins. In either case, it is often necessary to undertake old-style serial experiments where a set of potential antigens is expressed by hand and evaluated as a source of B-cell or T-cell epitopes. The work of Covert et al indicates a rapid, parallel, and facile postgenomic approach to this problem using proteomics to elucidate immunodominant T-cell antigens of pathogenic bacteria [48]. Subcellular protein fractions from *M tuberculosis* were resolved into 355 and 299 fractions of filtrate and cytosolic proteins. The reactions of splenocytes from C57Bl/6 mice infected with *M tuberculosis* were used to analyze dominant T-cell responses from these fractions, leading to the identification of 38 immunodominant fractions and 30 corresponding individual proteins. Many of these were previously known antigens, but 17 were novel T-cell antigens.

We now turn to a discussion of *H pylori*. The human stomach, on the basis of its low pH, has long been considered as an extremely hostile environment for the growth of bacteria. However, this view has changed dramatically with the discovery of the spiral microaerophilic bacterium *H pylori* from the human gastric mucosa. A report by Langenberg et al [49] began to unravel the mechanism of pathogenicity demonstrated by *H pylori*, by observing that it could produce large amounts of the virulence factor urease, thus explaining urease activity observed earlier in the mammalian stomach. This understanding, combined with evidence that *H pylori* causes chronic and acute gastritis, initiated interest into the prevalence and incidence of this bacterial infection. Epidemiological studies are consistent with the view that *H pylori* causes gastric infection in half the human population worldwide and over 80% of populations from developing countries. The prevalence of *H pylori* in gastric ulcer disease is greater than 90% and curing infection results in a cure for the gastric ulcer.

The definition of *H pylori* surface proteins is of particular importance in vaccine discovery. Two-dimensional electrophoresis combined with antibody detection and N-terminal sequencing was used to detect *H pylori* antigens [50, 51, 52]. Jungblut et al [53] studied *H pylori* whole cell proteins extensively by 2DE and 152 proteins were identified by MS. A single patient's serum was used to determine antibody reactivity. A small number of antigenic proteins were identified, leading the authors to suggest that several antigens may be minor components in whole cell lysates and therefore beyond detection in the absence of enrichment. Sample fractionation and enrichment of proteins using a chromatographic step prior to electrophoresis improves the identification of proteins at a low expression level. It may also improve the ratio of immunogenic versus nonimmunogenic proteins in a complex antigen preparation. In some studies of *H pylori* proteins, large pH gradients were used [51, 52, 53, 54] and basic proteins, common in *H pylori*, may have been poorly resolved. Isoelectric focusing using a more appropriate pH gradient allows greater resolution of proteins, and by 2DE immunoblotting it is possible to identify specific antigenic proteins as well as evaluate complex antigens. More precise identification of such immunogens will be necessary, in order to produce recombinant proteins, using either advanced MS-MS sequencing or more classical N-terminal microsequencing.

In a study designed to directly address the direct identification of vaccine targets, Chakravarti et al analyzed the *H pylori* genome [55] using both proteomic and genomic approaches. Two different approaches were taken for the identification of a set of potential candidate vaccines. In the first, proteins were identified from outer membrane preparations using proteomic technologies. An outer membrane fraction, purified from disrupted cells, was treated with Triton X-100, centrifuged, treated with detergent, centrifuged again, and then separated by 1D SDS-PAGE. Those proteins are reacting against monoclonal antibodies and are identified by mass fingerprinting. In the second approach, outer membrane proteins were separated by 2DE and transferred to PVDF membrane. Spots were trypsin-digested, and extracted peptides were analyzed by MALDI-TOF-MS. In a complementary study, Haas et al [56] compared the reactivity of sera from *H pylori*-infected patients, a control group with non- *H pylori* gastric illness, and patients with gastric cancer to electrophoretically separated proteins from *H pylori* strain HP 26695. Three hundred ten proteins were recognized by *H pylori*-positive sera. Notable amongst these were serine protease HtrA (HP1019), Cag3 (HP0522), and the predicted coding region HP0231. In an interesting variant study, McAtee et al examined protein differences between bacterial lysates from *H pylori* strain 26695, which is resistant to metronidazole (MTZ) due to a mutation in nitroreductases gene, rdxA, grown in the presence and absence of small quantity of MTZ [57]. The expression of a number of proteins decreased by twofold or more during growth with MTZ, yet the levels of various isoforms of alkylhydroperoxide reductase (AHP) (encoded by gene ahpC HP1563 and linked to oxygen toxicity resistance) increased.

Fungal, parasite, and cancer proteomics

In the following section, we briefly adumbrate several areas in eukaryotic proteome research. Two are emerging areas, fungal and parasite proteomics, while the third is relatively well developed, cancer proteomics. In examining the last of this triumvirate, it is difficult to disentangle it from host proteomics, which is clearly beyond the scope of this review. As a consequence, we will touch on the subject only briefly.

Currently, and in contrast to the application of genomic technologies, fungal proteomics is a ripe area for exploitation. Our present understanding of fungal virulence factors is somewhat limited and largely confined

to fungi-plant interactions [58]. They may be classified as

(1) toxins and enzymes that degrade host defenses. These can be readily assessed via biochemical assays and were amongst the first virulence factors identified;

(2) elicitors that induce host defenses;

(3) transporters and signal transduction components that protect the fungus from host responses;

(4) signal transduction proteins that aid sensing of the host environment;

(5) penetration effectors, such as melanin or hydrophobins.

The group of fungal virulence factors is still small and is obviously incomplete given the complex lifestyle of pathogenic fungi [59, 60, 61]. Thus the aggressive use of proteomic methods, in conjunction with genome-wide comparisons coupled with transcriptomic expression profiling, will have much to contribute to studies of fungal pathogenesis.

A recent study by Lim et al [62] will illuminate what is possible. Two hundred twenty proteins associated with the cell envelope were extracted from active and quiescent mycelia of *Trichoderma reesei*. Of these, 56 spots were examined by MS and 20 spots were identified as known proteins on the basis of sequence, indicating that most fungal cell wall proteins are novel. Identified proteins included translation elongation factor beta, diphosphate kinase, disulfide isomerase, outer membrane porin, transaldolase, vacuolar protease A, enolase, and glyceraldehyde-3-phosphate dehydrogenase. However, the most abundant protein in active and quiescent mycelia was HEX1. This is the major protein in Woronin bodies which are only found in filamentous fungi. Future studies will identify genes that specifically determine fungal lifestyle and genes that distinguish between filamentous and single-cell growth. It will also allow genes and pathways involved in pathogenicity to be identified, leading to the identification of further virulence factors, and thus further candidate fungal vaccines.

Parasitic infections are a very common cause of serious disease, particularly in third world countries and amongst domesticated animal populations, engendering a greatly enhanced interest in developing prophylactic vaccines against them [63, 64]. Human vaccines against malaria and other parasites have not been overly successful. However, vaccines able to control the major parasites of livestock have proved more useful [8, 65], particularly those directed against major nematode and trematode infections. Apart from attenuated-live vaccines for the control of avian coccidiosis, toxplasmosis in sheep and anaplasmosis in cattle, vaccines have been developed against *Haemonchus contortus*, the pathogenic nematode of sheep and goats, and *Fasciola hepatica*, the liver fluke of sheep and cattle; Bm86 vaccine against *Boophilus microplus*; 45w and EG95 recombinant proteins against *Taenia ovis* and *Echinococcus granulosis*; and broad-spectrum gastrointestinal worm vaccines against *Ostertagia* and *Trichostrongylus* species. Vaccines in development include the cathepsin L vaccines against the liver fluke *F hepatica*, and the H11 vaccine against *H contortus*.

Jefferies et al [66] analyzed the excretory-secretory proteins from *F hepatica* using proteomics, identifying a number of proteins including cathepsin L proteases and other enzymes involved in protection from the host immune responses as part of a reactive oxygen detoxification system: superoxide dismutase, thioredoxin peroxidase, and glutathione S-transferases. Interestingly, host superoxide dismutase was the only such protein identified on the gel.

By comparison, molecular vaccines against protozoans are proving considerable more elusive in both animals and humans. This is no where more apparent than in the case of malaria. This disease, caused, in its most severe form, by the protozoan parasite *Plasmodium falciparum*, has plagued humanity throughout recorded history and results in the death of over 2 million people per year. Other parasitic diseases, such as leishmaniasis and schistosomiasis, are also important diseases in developing countries. Leishmaniasis, in its cutaneous (CL), mucocutaneous (MCL), and visceral (VL) forms, affects directly about 2 million people per year, with about 350 million individuals at risk worldwide. The 35-Mb genome of *Leishmania*, which should be sequenced late in 2002, contains about 8500 genes that will translate into more than 10000 proteins. Of all vaccines against human parasitic disease, those targeting malaria, leishmaniasis, and schistosomiasis are in the most advanced stages of development. However, despite the remarkable progress made in identifying protective antigens, at present there are no generally accepted vaccines against parasitic diseases. Vaccines for malaria and leishmaniasis have been taken to clinical trials while vaccines for schistosomiasis are in phase I/II trials. The control of leishmaniasis remains a problem and no vaccines exist for the VL, CL, or MCL forms of the disease.

While postgenomic approaches are being pursued actively for *Leishmania* [67], which combine MicroArray transcriptomics with random vaccine screening using cDNA libraries, relatively little has been done within the proteomic arena. Thiel and Bruchhaus [68] have analyzed the expression specific differences between the proteomes characterizing the promastigotes and amastigotes forms of *Leishmania* and also the transition between them. They mapped the *Leishmania donovani* proteome during distinct metamorphic stages, identifying stage-specific proteins and regulons, using isoelectric focusing compatible protocol. Around 400 proteins could be visualized and a significant decrease in protein synthesis during differentiation from promastigotes to amastigotes could be observed.

Toxoplasma gondii is another protozoan parasite that has been investigated using proteomic technology. There are two forms of *T gondii* associated with human hosts. The rapidly growing tachyzoites give rise to acute illness

and the slowly dividing encysted bradyzoites can remain dormant within tissues for a lifetime. During infection, conversion occurs between the rapidly dividing tachyzoite stage (responsible for acute toxoplasmosis) and the much more slowly replicating bradyzoite, a process central to both pathogenesis and longevity of infection. Proteomics has helped identify several proteins specific to these different stages.

Cohen et al [69] analyzed proteins expressed during the tachyzoite stage of *T gondii* and separated over 1000 proteins in the pH ranges 4–7 and 6–11. Because the genome was not available in full, they were obliged to combine their proteomic approaches with searches of EST databases in order to identify proteins less equivocally. Many protein spots were encoded by the same gene, indicating that posttranslational modification and alternative splicing are common features of gene expression in *T gondii*. In a similar study, Dlugonska et al [70] analyzed a lysate of the tachyzoite stage of *T gondii* and separated 224 proteins. They could identify 14 proteins using mass fingerprinting including the excretory dense granule proteins GRA1-GRA8, S16/acid phosphatase, nucleoside triphosphate hydrolase, and the H4 protein, and two secreted antigens p36 and p40 were identified.

Proteomic analysis of host-pathogen interactions

The story of the proteomic analysis of host-pathogen interactions is the story of a series of dichotomies, which is to say that we can partition the subject into a bifurcating series of binary divisions. One division is between the nature of target cells (antigen presenting cells versus T cells), another is between the nature of stimulation used to engender changes in gene expression within target cells (bacterial or viral infection versus isolated immunomodulators, such as LPS). In the following section we will briefly review a number of studies addressing these issues. Each highlights a different theme or aspect relevant to the development of proteomic immunovaccinology. We begin with alterations apparent in gene expression within a small number of bacterial systems.

Fletcher et al investigated the effect of environmental factors on the expression and release of secreted or surface proteins, containing many virulence factors, from *Actinobacillus actinomycetemcomitans*, a bacteria implicated in periodontal diseases, where gum inflammation is associated with bone loss and gum recession leading to the formation of a so-called periodontal pocket [71]. Differences in expression of many proteins, including glycolytic enzyme triose phosphate isomerase, were observed for bacteria grown under varied conditions (anaerobic versus aerobic growth, biofilm versus planktonic growth, under iron depletion, or in the presence or absence of serum or blood), indicating its adaptability to changes within the periodontal microenvironment. Monahan et al analyzed changes in protein expression in attenuated vaccine strain *M bovis* BCG induced by host macrophage phagocytosis [72]. They used proteomics to show that BCG phagocytosed by the human macrophage cell line THP-1 expresses proteins not seen during heat shock or growth in culture media, and were able to identify six proteins showing increased expression: 16 kd alpha-crystalline (HspX), GroEL-1 and GroEL-2, a 31.7-kd hypothetical protein (Rv2623), InhA, and elongation factor Tu (Tuf).

We now turn to proteomic changes in antigen presenting cells and begin with the inverse experiment to that performed by Betts. Ragno et al combined transcriptomic and proteomic methods to evaluate changes in gene and protein expression in the leukaemic macrophage cell line THP-1 after infection with TB [73]. Initially, microarrays of 375 immunologically implicated human genes identified a set of early upregulated proteins that not unsurprisingly included a range of chemokines and cytokines, as well as other cell surface molecules. It was more difficult to detect changes using proteomics, although human IL-1beta and superoxide dismutase were shown to have increased expression after infection, and, in contrast, the heat-shock protein hsp27 was downregulated. In a similar study, Kovarova et al analyzed phagosome extracts from macrophages derived from host organisms resistant or susceptible to infection by *Francisella tularensis* LVS (live vaccine strain) [74]. They identified several proteins upregulated in susceptible macrophages including host proteins mitochondrial ATP synthase beta-chain and NADH-ubiquinone oxidoreductase as well as the bacterial 60-kd chaperonin GroEL and a hypothetical 23-kd protein, whose expression level correlate with susceptibility and *F tularensis* LVS pathogenicity. Pizarro-Cerda et al examined the molecular components that facilitate cellular uptake of *Listeria monocytogenes* into the phagosome in the human epithelial cell line LoVo using proteomics [75]. Their results confirmed literature precedents, with the exception of MSF, a member of the septin family of GTPases, which forms filaments that colocalize with the actin cytoskeleton in quiescent cells.

Moving now from cells presenting antigen to cells mediate immune recognition, we focus now on T cells. Truffa-Bachi et al used protemics to analyze the changes contingent on the removal of *Concanavalin* A or *Cyclosporin* A from cultures of activated murine T cells [76]. They found that a large number of proteins were strongly upregulated and downregulated after the immunosuppressive drugs were removed, indicating that T cells were programmed by *Cyclosporin* A to change expression levels without reactivation. In the context of developing a proteomic database of helper T cells, Nyman et al activated $CD4^+$ T cells with anti-$CD3^+$ anti-CD28 antibodies and visualized 2000 spots with autoradiography and 1500 spots using silver staining and identified 91 proteins using mass fingerprinting [77]. By using proteomics, Fratelli et al sought to identify T-cell proteins that undergo glutathionylation, the formation of mixed disulfides between glutathione and other proteins, under conditions of oxidative stress [78]. They observed several proteins not previously known to be glutathionylated, including enzymes, such as enolase (which is inhibited by glutathionylation); redox enzymes, such as peroxiredoxin 1 or cytochrome

c oxidase; cytoskeletal proteins, such as vimentin, profilin, and actin; cyclophilin (which is not inhibited by glutathionylation); stress proteins, such as HSP60 and HSP70; and a number of miscellaneous proteins, such as galectin and fatty acid binding protein. The authors felt that their results supported the view that glutathionylation is a common global mechanism for the regulation of protein function.

INFORMATICS OF THE IMMUNOME

Informatic support for proteomics is now well established, and it would be futile to reiterate the content of many useful reviews on the subject (see [79, 80, 81, 82] and references therein). Equally software for the analysis and exploration of proteomics is now well developed and widely distributed. Indeed, online databases of proteomes or collections of proteomes have now proliferated. However, the informatic analysis of the immunome is currently less well developed. In many ways the informatic analysis of the imunome is the complement of the experimental analyzes described above. The immunome, the complement of short immunogenic peptides derived, by the complex, poorly understood molecular machinery of the immune system, from the proteome of some microbe is itself a subset of the peptidome. The peptidome is the set of all peptides, as opposed to proteins generated by the cell. It is composed of both genomic peptides, with a specific function, such as hormones or neuropeptides, and cleavage products generated by proteases. In some respects, it lies somewhere between the proteome and metabolome of small biosynthesized molecules and is highly compartmentalized within the cell. Bioscience is only now beginning to explore the peptidome. Because experimental methods do not address either the peptidome or immunome, informatic prediction has much to contribute here.

Approaches to predicting the immunone

A specialized type of immune cell mediates cellular immunity, the T cell, which constantly patrols the body searching out proteins that originate from a pathogenic organism, be that virus, bacterium, fungus, or parasite. The surface of T cells is, unsurprisingly, enriched in TCRs, which function by binding MHCs expressed on the surfaces of other cells. These proteins bind small peptide fragments derived from both host and pathogen proteins. It is the recognition of such complexes that lies at the heart of the cellular immune response. These short peptides are referred to as epitopes. The overall process leading to the presentation of antigen-derived epitopes on the surface of cells is a complicated, and not yet fully understood, process. There are many alternative processing pathways, but we will confine our attention to the two major types: class I and class II.

Class I MHCs are expressed by almost all cells in the body. They are recognized by T cells whose surfaces are rich in CD8 coreceptor protein. Class II MHCs are only expressed on the so-called "professional antigen presenting cells" and are recognized by T cells whose surfaces are rich in CD4 coreceptors. Class I peptides are ultimately derived from intracellular proteins, such as viruses. These proteins are targeted to the proteasome, which cuts them into short peptides of 8 to 11 amino acids in length. These peptides are then bound by the transmembrane peptide transporter TAP, which translocates them from the cell cytoplasm to the endoplasmic reticulum where they are bound by MHCs. Theoretical analyzes of proteasomal cleavage patterns have been conducted by several groups [83, 84], leading in turn to a number of prediction methods [85], some of which are available via the Internet [86, 87]. The amount of data studied remains relatively small, and the predictive power possessed by these different methods has yet to be evaluated objectively. Nonetheless, they represent useful contributions and important starting points for future study. Likewise, studies have also been conducted on the peptide substrate specificities of the TAP transporter [88], leading to the development of predictive models [89] for the determination of peptides that bind to TAP. Together, studies on proteasomal cleavage and TAP transport represent a good first attempt to produce useful, predictive tools for the processing aspect of class I restricted epitope presentation.

For class II, receptor-mediated ingestion of extracellular protein derived from a pathogen is targeted to an endosomal compartment, where the proteins are cleaved by cathepsins, a particular class of protease, to produce slightly longer peptides of 15–20 amino acids. Class II MHCs then bind these peptides. The peptide specificity of protein cleavage by cathepsins has also been investigated and simple cleavage motifs are well known [90]. However, more precise investigations are required before accurate predictive methods can be realized. The first attempts to computerize the identification of MHC binding peptides led to the development of motifs characterizing the peptide specificity of different MHC alleles. Such motifs—a concept with wide popularity amongst immunologists—characterize a short peptide in terms of dominant anchor positions with a strong preference for certain amino acids. Probably the first proper attempt to analyze MHC binding in terms of specific allele-dependant sequence motifs was undertaken by Sette et al [91]. They defined motifs for the mouse alleles I-Ad and I-Ed after measuring affinity for a large set of synthetic peptides originating from eukaryotic and prokaryotic organisms, as well as viruses; in addition they also assayed a set of overlapping peptides encompassing the entire staphylococcal nuclease molecule. Sette et al quote prediction rates at the 75% level for these two alleles. A large number of succeeding papers, both from this group and others, have extended this approach to many other human and mouse alleles.

As we have said, these motifs are usually expressed in terms of anchor residues: the presence of certain amino acids at particular positions that are thought to be essential for binding. For example, human class I allele

HLA-A*0201, probably the best studied of all alleles, has anchor residues at peptide positions P2 and P9 for a nine amino acid peptide. At P2, acceptable amino acids would be L and M, and at the P9 anchor position would be amino acids V and L. Secondary anchors, residues that are favourable, but not essential for binding, can also be present. Moreover, sequencing of peptides, that are known to bind, show preferences for particular amino acids at particular positions, although whether this represents anything other than the inherent bias in protein sequences is seldom addressed. The method is admirably simple: it is easy to implement either by eye or more systematically using a computer to scan through protein sequences.

However, there are many problems with the motif approach. Although it is possible to score the relative contributions of primary and secondary anchors to produce a rough and ready measure of binding affinity [92, 93], the most significant problem with the motif approach is that it is, fundamentally, a deterministic method. A peptide is either a binder or is not a binder. Even a brief reading of the immunological literature shows that matches to motifs produce many false positives, and are, in all probability, producing an equal number of false negatives, though peptides predicted to be nonbinders are seldom screened.

While useful in themselves, binding motifs are, as we have said, very simplistic. They are not quantitative and their over-reliance on anchor positions can lead to unacceptable levels of false positives and false negatives. Alternative approaches abound and have different strengths and different weaknesses. The strategy adopted by many workers is to use data from binding experiments to generate matrices able to predict MHC binding. For want of a better term, we refer to these approaches as experimental matrix methods, as most such methods use their own measured data and relatively uncomplicated statistical treatments to produce their predictive models.

A step forward from deterministic motifs came with the work of Kenneth Parker [94]. This method, which is based on regression analysis, gives quantitative predictions in terms of half-lives for the dissociation of β_2-microglobulin from the MHC complex. It is founded on a series of important observations about peptide binding to MHC molecules [95, 96, 97, 98, 99, 100, 101, 102] and has been used in a number of applications [103, 104]. Moreover, apart from its intrinsic utility, one of the other important contributions of this approach is that it was the first to be made available online (http://bimas.dcrt.nih.gov/molbio/hla_bind/). This method, often referred to as BIMAS, or occasionally, COMBIFORM, by immunologists, is, for this reason, widely used. Other empirical methods include EpiMatrix and EpiMer developed by DeGroot and coworkers and TEPITOPE developed by Hammer and colleagues.

A number of groups have used techniques from artificial intelligence research, such as artificial neural networks (ANNs) and hidden Markov models (HMMs), to tackle the problem of predicting peptide-MHC affinity. ANNs and HMMs are, for slightly different applications, the particular favourites when bioinformaticians look for tools to build predictive models. However, the development of ANNs is often complicated by several adjustable factors whose optimal values are seldom known initially. These can include, inter alia, the initial distribution of weights between neurons, the number of hidden neurons, the gradient of the neuron activation function, and the training tolerance. Other than chance effects, neural networks have, in their application, suffered from three kinds of limiting factors: overfitting, overtraining (or memorization), and interpretation. As new, more sophisticated neural network methods have been developed and statistics has been applied to their use, overfitting and overtraining have been largely overcome. Interpretation, however, remains an intractable problem; few, if any, can easily visualize or interpret the very complex weighting schemes used by neural networks.

Notwithstanding these potential problems, many workers have adopted an ANN strategy in seeking to solve the prediction of peptide-MHC binding. Bisset and Fierz [105] were amongst the first to use ANN in this context. They trained an ANN to relate binding to the class II allele HLA-DR1 to peptide structure and reported a correlation coefficient of 0.17 with a statistical significance of $P = .0001$. Amongst the best known names of those interested in the area of MHC binding prediction is Vladimir Brusic. Over many years, he and his coworkers have developed a range of artificial intelligence techniques, including, inter alia, ANN, HMMs, and evolutionary algorithms, aimed at solving problems of this kind [106, 107, 108, 109]. His work contains models of both class I and class II MHC alleles, as well as the TAP transporter [88, 89], and within the context of his own classification scheme [110], his models seem highly predictive.

A quite different approach to obtaining predictions of peptide is that MHC binding is based on atomistic molecular dynamic simulations. It attempts to calculate the free energy of binding for a given molecular system, which is closely related to experimentally observable quantities such as equilibrium constants or IC_{50}s. It has the advantage that, in principal, there is no reliance on known binding data, as it attempts the de novo prediction of all relevant parameters given certain knowledge of the system. Essentially, all what is required is the experimentally determined structure, or a convincing homology model, of an MHC peptide complex.

DeLisi and coworkers were among the first to apply molecular dynamics to peptide, MHC binding, and have, subsequently, developed a series of different methods [111, 112, 113]. Part of this work has concentrated on accurate docking using molecular dynamics and another part on determining free energies from peptide MHC complexes. Didier Rognan has, over a long period, also made important contributions to this area [114, 115, 116].

In his work, dynamic properties of the solvated protein-peptide complexes, such as atomic fluctuations, solvent accessible surface areas, and hydrogen bonding patterns correlated well with available binding data. He has been able to discriminate between binders that remain tightly anchored to the MHC molecule and nonbinders that are significantly weaker. Other work in this area has come from two directions. The first direction is interested in using the methodology to analyze and predict features of peptide-MHC complexes. These methods have looked at both class I [117, 118] and class II [119]. The second direction is more interested in developing novel aspects of molecular dynamics (MD) methodology, including both simulation methodology [120] and solvation [121], and using the MHC peptide systems as a convenient example of binary molecular complex.

Quantitative approaches to predicting the immunone

In this section, we review quantitative approaches to the developing field of computational immunovaccinology. This includes our own contribution, including a discussion of our newly released JenPep database and two powerful new techniques for T-cell epitope prediction. The first is a 2D quantitative structure-activity relationships, or 2D-QSAR, approach which we have christened the "additive" method [122]. The other is a 3D-QSAR approach, based on comparative molecular similarity indices analysis (CoMSIA) [123, 124]. The methods were prototyped using the common class I allele, HLA-A*0201, for which numerous binding data is available.

Virtual screening

A methodology closely related to MD, both being based, to a large degree, on molecular mechanics force fields, or, at least, drawing on analogies from pairwise atomistic potential energy functions, is a set of techniques grouped loosely under the name of "virtual screening." There are two principal types of virtual screening methodology that have, thus far, been applied to the prediction of MHC binding. One derives from computational chemistry and the other from structural bioinformatics and the development of tools for fold prediction. Virtual screening is an expression derived from pharmaceutical research that is the use of predicted ligand-receptor interactions to rank or filter molecules as an alternative to high-throughput screening. Approaches to virtual screening cover a spectrum of methods which vary in complexity from molecular descriptors and QSAR variables, through simple scoring functions (such as Ludi, FlexX, Gold, or Dock), potentials of mean force (PMF) (such as Bleep), force field methods, QM/MM and linear response methods, to free energy perturbations. In this transition from, say, atom counts through to full molecular dynamics, we see a tremendous increase in required computer time. Virtual screening can be seen as seeking a pragmatic solution to the "accuracy gained" versus "time taken" equation. The point at which one stops on this spectrum is contingent upon the system being evaluated, the number of peptides being evaluated, and the computing resources available.

Didier Rognan has developed a virtual screening method called FRESNO and applied this algorithm, which relies on a simple physicochemical model of host-guest interaction, to the prediction of peptide binding to MHCs [125]. This model was trained on a combination of data and experimentally derived 3D structures from the alleles HLA-A*0201 and H-2Kk. He found that lipophilic interactions contributed the most to HLA-A*0201-peptide interactions, whereas H-bonding predominated in H-2Kk recognition. Cross-validated models were afterward used to predict the binding affinity of a test set of 26 peptides to HLA-A*0204 (an allele closely related to HLA-A0201) and of a series of 16 peptides to H-2Kk. He concluded from their initial study that their scoring function was able to predict, with reasonable accuracy, binding free energies from 3D models. In a more comparative study [126], Rognan and colleagues found that for predicting the binding affinity of 26 peptides to the class I MHC molecule HLA-B*2705, FRESNO outperformed six other available methods (Chemscore, Dock, FlexX, Gold, Pmf, and Score).

Turning now to bioinformatic-based approaches, others are using amino acid pair potentials, initially developed to predict the fold of a protein, to identify those peptides which will bind well to an MHC. Margalit and colleagues have proposed a number of virtual screening methodologies [127, 128], each is of increasing complexity. They used amino acid pair potentials, originally developed by Miyazawa and Jernigan [129], to evaluate the interprotein contact complementarity between peptide sequences and MHC binding site residues. They presented an analysis of peptide binding to four MHC alleles (HLA-A2, HLA-A68, HLA-B27, and H-2Kb), and were successful in predicting peptide binding to MHC molecules with hydrophobic binding pockets but not when MHC molecules with charged or hydrophilic pockets were investigated. Again focussing on class I alleles, a more recent study from this group [130] used an updated set of statistical pairwise potentials. These were developed from the Miyazawa and Jernigan potential by Betancourt and Thirumalai [131] and described the hydrophilic interactions more appropriately. This enables more accurate modelling of the threading of the candidate peptide sequence.

Because of the relative celerity of virtual screening methods compared with MD methods and its ability to tackle MHC alleles for which no known binding data is available, this method has considerable potential. While both MD and related methods hold out the greatest hope for such true de novo predictions of MHC binding, their present success rate is very much lower than that of data driven models.

Positional scanning peptide libraries

An alternative strategy is the use of positional scanning peptide libraries (PSPLs) to generate such matrices. A number of such studies have been conducted. Some are aimed at investigating the problem of MHC-peptide interaction [132, 133, 134], while others concern themselves with evaluating how variations in peptide sequence contribute to TCR recognition and T-cell activation [135, 136]. One of the most recent of these is also one of the most promising; Udaka et al [137] have used PSPLs to investigate the influence of positional sequence variation on binding to the mouse class I alleles Kb, Db, and Ld. From their analysis, a program that could score MHC-peptide interaction was developed and used to predict the experimental binding of an independent test set. Their results showed a good linear correlation but with substantial deviation. About 80% of peptides could be predicted within a log unit.

QSAR approaches

JenPep

Version 1.0 of JenPep [138] is composed of three subdatabases: (i) a compilation of quantitative affinity measures for 6000 peptides which bind class I and class II MHC; (ii) a compendium of 2300 dominant and subdominant T-cell epitopes; and (iii) a set of quantitative data for 400 peptide binding to the TAP peptide transporter. The database, and an HTML interface for searching, is freely available via the Internet. It can be found at http://www.jenner.ac.uk/JenPep. JenPep contains binding data on a wide variety of different MHC alleles; for class I MHC molecules, JenPep has data for 68 different restriction alleles with more than 50 genotype variations. For class II MHC molecules, there are over 40 restriction alleles with 52 genotype designations. Peptide lengths for class I MHC molecules are in the range of 7–16 residues and for class II MHC molecules are in the range of 9–35 residues. Measures of binding affinity include radiolabelled and fluorescent IC_{50} values, BL_{50}, and half-lives. JenPep is the first database in immunology to concentrate on quantitative measurements, complementing existing systems. This compilation of binding data underlies our attempts to derive statistically sound QSAR tools for the accurate prediction of peptide binding to immunological molecules.

A 2D-QSAR method for binding affinity prediction

We have developed predictive techniques based on the so-called additivity concept, whereby each substituent makes an additive and constant contribution to the biological activity regardless of variation in the rest of the molecule. The IBS hypothesis, developed by Parker [94], is the immunological analogue of this idea. We extended this concept by adding additional terms that account for near neighbour side-chain interactions [122]. The binding affinity of a peptide will depend on contributions from each amino acid as well as interactions between adjacent and every second side-chain:

$$\text{binding affinity} = \text{const} + \sum_{i=1}^{9} P_i + \sum_{i=1}^{8} P_i P_{i+1} + \sum_{i=1}^{7} P_i P_{i+2}, \quad (1)$$

where the const accounts, at least nominally, for the peptide backbone contribution, $\sum_{i=1}^{9} P_i$ is the sum of amino acids contributions at each position, $\sum_{i=1}^{8} P_i P_{i+1}$ is the sum of adjacent peptide side-chain interactions, and $\sum_{i=1}^{7} P_i P_{i+2}$ is the sum of every second side-chain interactions.

Four hundred twenty IC_{50} values for 340 nonamer peptides were used in the development of the additive method. The peptide sequences and their binding affinities to the HLA-A*0201 molecule were extracted from the JenPep database. More than one IC_{50} value was found for some of the peptides. As is common practice amongst QSAR practitioners, IC_{50} values were converted to P-units (negative decimal logarithm).

A program was developed to transform the nine amino acid peptide sequence into a row of a table. A term is equal to 1 when a certain amino acid at a certain position or a certain interaction exists, and equal to 0 when they are absent. Thus a matrix of 420 rows and 6120 columns was generated. One hundred eighty columns account for the contributions of the amino acids (20 amino acids $\times$ 9 positions), 3200 for the adjacent side-chains, or 1–2 interactions ($20 \times 20 \times 8$), and 2800 for the 1–3 side-chain interactions ($20 \times 20 \times 7$). To reduce the number of columns, the program omits columns that contain only zeros. The final matrix consists of 420 rows and 2158 columns.

As the columns are more numerous than the rows, the equations were solved using partial least square (PLS) method. The predictive power was assessed by the cross-validated q^2 (as generated by "leave-one-out" cross-validation [LOO-CV]), standard error of predictions (SEP), and residuals between the experimental and predicted by LOO-CV PIC_{50} values. A mean |residual| value and standard deviation for the set were also calculated. The non-cross-validated model was assessed by multiple linear regression (MLR) parameters: explained variance (r^2), standard error of estimate (SEE), and F ratio.

The final equation derived by the additive method consists of 1815 terms including the constant. It contains the contributions of the amino acids and the contributions of the significant side-chain interactions. There were 172 very well-predicted (residuals $\leq$ |0.5| log unit) peptides (50.5%), 128 well-predicted (|0.5| $\leq$ residuals $\leq$ |1.0| log unit) peptides (37.5%), and only 41 poorly predicted (residuals $>$ |1.0| log unit) peptides (12.0%).

A 3D-QSAR method for binding affinity prediction

One of the most reliable methods for investigating the structure-activity trends within sets of biological

molecules is 3D-QSAR. The explanatory power of 3D-QSAR methods is considerable, manifests not only in their ability to accurately predict binding affinities, but also in their capacity to display advantageous and disadvantageous interaction potential mapped onto the structures of molecules being investigated. We have applied the 3D-QSAR method (CoMSIA) to gain an understanding of the relationship between physicochemical properties (steric bulk, electrostatic potential, local hydrophobicity, hydrogen-bond donor, and hydrogen-bond acceptor abilities) and the affinities of peptides that bind to the MHC molecule HLA-A*0201 [123, 124].

Two hundred sixty-six nonamer peptides are included in the CoMSIA study. Their IC_{50} values were collected from the JenPep database and converted to *P*-units. All molecular modelling and QSAR calculations were performed using the sybyl 6.7 molecular modelling software. The X-ray structure of the nonameric viral peptide TLTSCNTSV was used as a starting conformation. The structures of the remaining peptides were built to this conformation. The partial atomic charges used in CoMSIA were computed using the AM1 semiempirical method, as available in MOPAC.

Five types of similarity index (steric, electrostatic, hydrophobic, and hydrogen-bond donor and acceptor) were calculated, using a common probe atom with 1 Å radius, charge +1, hydrophobicity +1, hydrogen-bond donor and acceptor properties +1. SEP, q^2, and residuals assessed the predictive power of the final model. The initial CV model had low q^2 and r^2 values. This result was not surprising, given the great diversity of peptides collected from a variety of sources. One hundred fifty-one were very well predicted, 83 were well-predicted peptides, and only 32 peptides were poorly predicted. The mean |residual| was 0.553. The model was improved by excluding a limited number of poorly predicted peptides in a stepwise manner, beginning with the peptide with the highest residual. The final CV model had significantly higher parameter values: $q^2 = 0.683$ at 7 components and $r^2 = 0.891$. This model was used to predict the binding affinities of the excluded peptides. The predictions were better for both the group of very well-predicted peptides and the group of poorly predicted peptides.

Comparison of the two methods in the context of peptide structure

It has long been known that all nine side-chains of the bound peptide contact the HLA-A*0201 molecule and influence the energetics of binding. The antigen-binding groove has a 30-Å long surface accessible to a solvent probe. There are six pockets in the surface denoted by A through F. Some of them are nonpolar and can form hydrophobic contacts, while others contain polar atoms and can make hydrogen bonds with the side-chains. As statistical approaches, the additive method and CoMSIA seek to correlate relative differences in discriminating molecular descriptor values to a dependent property (eg, the binding affinity). In that respect, CoMSIA is a method able to map similarities or dissimilarities between molecules. The additive method is able to quantify the contributions made to the binding affinity by each amino acid, at each position, and by the interactions between them. Comparing the results of the additive method and CoMSIA, we have found a remarkable degree of congruence.

Positions within the peptide are defined as P1 to P9. CoMSIA suggests that hydrophobic steric bulk with negative potential is well tolerated at P1. Topologically, P1 corresponds to pocket A. The most suitable amino acids for this position seem to be Phe and Tyr. According to the additive method, Tyr is the favourite amino acid for P1. Phe and Lys also make positive contributions, while Arg, His, and Thr are not preferred. The steric map at P2 indicates that long side-chains such as Leu, Ile, and Met are well tolerated here. The additive method distinguishes two favourite amino acids for this position (Met and Leu). Ala, Cys, Gly, and Thr make negative contributions.

Hydrophobic volume with negative potential is preferred at P3. The side-chains of the amino acids at this position fall into pocket D. The hydrogen bonding ability map indicates that amino acids able to form hydrogen bonds will also be well accepted here. Tyr and Trp have the greatest positive contributions for this position, but Leu and Phe are also well accepted. Glu, Cys, His, Pro, and Ser negatively contribute. Short hydrophilic amino acids able to form hydrogen bonds are well tolerated at P4. Ser or Thr would be well tolerated here. According to the additive method, there is no favourite amino acid at P4. Gly, Pro, Ser, and Thr are well accepted here while Ile, Phe, Cys, and Met make negative contributions.

The maps indicate that amino acids with hydrophobic, branched or aromatic side-chains ending with small hydrophilic groups are well tolerated at P5. Favourite amino acids for P5 are Phe and Tyr. His, Leu, and Trp also positively contribute, while Arg should be avoided at this position. Amino acids with long hydrophobic side-chains are preferred at P6. Hydrogen-bond ability is an additional priority. Ile, Leu, Thr, and Tyr are well accepted here. Ala, Arg, Asp, Gln, His, and Lys negatively contribute. This side-chain falls into pocket C. This pocket is predominantly polar, which explains the acceptance of the hydrophilic Thr and Tyr, but it cannot explain the preference for the hydrophobic Ile and Leu. Short side-chains are favoured sterically at P7. The side-chain at P7 falls into pocket E. Pro is the favourite amino acid for this position according to the additive method, although His also makes a good contribution. Asn, Arg, Gln, Gly, Ser, and Thr are deleterious.

The side-chain at P8 should be short, with a hydrophobic core and an end capable of forming hydrogen bonds. Gln, Phe, Pro, and Ser are all well accepted here. The presence of Asp, Ile, His, Met, or Val is deleterious. Amino acids with hydrophobic, short side-chains are required for P9. Val is the favourite amino acid here. Interestingly, a small hydrophilic area, carrying negative potential, appears near P9, which is due to the Thr introduced

here by the intermediate binder MLQDMAILT and the high binder YMLDLQPET. However, according to the additive method, Ser and Thr should be avoided.

Predicting subcellular location

There are obviously many other aspects to computational vaccine design other than the prediction of potential epitopes. Many of them are as yet only poorly developed. While we have seen that T-cell epitope prediction is now well developed, at least to the stage where it is beginning to become useful, the prediction of immunogenicity, particularly for subunit vaccines, which necessarily involves a deeper understanding of host responses, remains primitive. The prediction of antibody- or B-cell-mediated antigenicity is at an even more primitive stage [139, 140]. This relies on concepts of some antiquity [141, 142, 143] and quite simplistic software [144, 145]. However, some other techniques complementary to the prediction of host responses, such as the prediction of the subcellular location of potential antigen proteins, have reached a greater level of maturity. A prevailing hypothesis, amongst many, involved directly in the hunt for protective antigens is a belief that the majority of such immunogens will be secreted proteins. Proteomics can help in the systematic search for secreted proteins [146, 147]. This is also an area where computational techniques can produce direct results.

Consider a microbial genome or, more specifically, a bacterial genome. The total protein complement—say a few thousand gene products—is distributed between the inner and outer compartments of the bacteria. Some will reside in the cytoplasm, some will find their way to the periplasmic space, at least in Gram-negative bacteria, and others will be secreted from the cell. Some proteins will become integral membrane proteins located in the inner or outer membranes and some will become lipoproteins. An ability to predict these locations would be a great benefit when choosing which proteins to investigate as candidate vaccines; a secreted protein, for example, can be regarded, at least naively, to be a more likely target than, say, a cytoplasmic enzyme. A number of bioinformatic methods have been developed which address the prediction of subcellular location which has proved to be more complex than was originally envisaged.

In 1982, a strong link between amino acid composition (eg, Leu and Trp favoured, Pro disfavoured [148]) and cellular location was identified [149], but as the number of available protein structures increased, this relationship has become more blurred [150]. Despite the ambiguous relationship between amino acid composition and subcellular localization, many methods of increasing sophistication have been created that exploit this connection [151, 152, 153]. Nakashima and Nishikawa [154] describe a method where the average amino acid composition for a number of proteins, whose subcellular localization is known, was calculated. From these simply obtained results, trends in amino acid composition were observed such as intracellular proteins relatively rich in aliphatic residues. Just using basic rules like these, they were able to correctly identify 78% of the test set as being either intracellular or extracellular.

This idea was developed further by Andrade et al [155] who hypothesized that throughout evolution, each subcellular location has maintained a characteristic physiochemical environment. The proteins in each location would have adapted to the environment and therefore each location would have proteins with signature structural characteristics. These characteristics are more likely to manifest at the surface (which is exposed to the environment) and therefore the surface residue composition is likely to give a very strong identification of the subcellular location. This method predicted 77% of protein locations accurately. Although amino acid composition is correlated with subcellular location, the former cannot be exclusively defined by the later. Neural networks have also been applied to this problem [156] and are the basis of the NNPSL web-based server. This provided an accuracy of 81% for prokaryotic prediction but only 66% for eukaryotic. This seems likely to be due to the persistent neural network shortcoming of overfitting to training data especially when the variables are complex.

The majority of methods for predicting localization are based on protein sorting signals [157]. These signals are normally represented as a short sequence with variable levels of conservation. Many are represented as well-defined motifs while others show vague sequence features that are undetectable by simple homology searching [158]. The most obvious protein sorting signal to investigate is the signal peptide. Looking at a simple bacterial model, if a protein has a signal peptide but no transmembrane domain, then it will be excreted through the inner membrane. If a protein with a signal peptide has a transmembrane domain, then it will become inserted into the membrane [159]. All signal peptides have a 3-region structure, the amino (N), the hydrophobic (H), and the carboxy (C) with a weak consensus pattern specifying the cleavage site [160]. Signal peptides are divided into classes on the basis of variation of structure of the N, H, and C regions, structure of the cleavage site, and different propensities for amino acids [161].

Many approaches have been taken to try and predict subcellular location from signal peptides and cleavage variations. The different amino acid propensities of N, H, and C regions for different classes can be identified by multivariate analysis of the individual amino acids [162]. A wide range of characteristics of amino acid properties has been determined, and the similarities/dissimilarities in the property profiles for different signal peptide classes were compared. Initially this method was applied just to *E coli* with some success but later expansion to Gram-positive bacteria was less successful and varied greatly from species to species [163]. Though, there were some factors such as charge, length, side-chain hydrophobicity, and volume that proved reasonably reliable factors that could be used as part of possible new techniques. The prediction of cleavage sites and

inference of subcellular location has proved more fruitful than amino acid composition-based methods, with prediction as high as 96% [164, 165].

DISCUSSION

Prophylactic vaccination has made an essential contribution to the improvement of human health over the 20th century. However, we still lack efficient vaccines against major human diseases such as malaria or tuberculosis. Historically, at least in the area of parasite vaccinology, as in many other areas of the subject, one of the greatest problems has been the scarcity of relevant material and our concomitant inability to generate purified vaccine candidates using conventional protein chemistry. Proteomic and other postgenomic and molecular biology approaches, through the preparation of cDNA expression libraries, are now proving central to the identification of immunogenic proteins.

Preceding sections have addressed two approaches to the identification of the immunome: experimental proteomic analysis of microbial proteins and mechanistic informatic prediction. However, the present review is by no means exhaustive, nor does it pretend to be. We have not specifically addressed other important uses for proteomics within vaccine research, such as the systematic discovery of adjuvants and diagnostic and prognostic biomarkers. Rather, we have to suggest how computational strategies and experimental proteomic approaches are highly complimentary to the aim of identifying the immunome. In particular, proteomics will prove crucial in the correct identification of appropriate posttranslational modification and conformation, upon which the immunogenicity of many vaccines will depend. Various informatics strategies hold out the hope that they will be able to short cut some of the more intractable experimental procedures by quickly prioritizing candidate genes.

As we have shown, experimental proteomics can identify proteins that represent potential candidate vaccines. It can achieve this either by identification of highly expressed genes or proteins secreted from the cell. The discovery of potential virulence factors or antigens is achieved by comparing the proteomes of virulent and avirulent microbes, or microbes grown under different conditions, or changes apparent upon infection, or by identifying proteins that are coregulated with already known virulence genes. For example, identification of proteins by such strategies may help elucidate the molecular basis of the attenuation of BCG, and may provide antigens that distinguish infection with *M tuberculosis* from vaccination with BCG. Proteomics can also help trace out how pathogenic bacteria cope with the challenges imposed on them by therapy or host responses to infection. Generally, however, proteomics will only form part of large postgenomic strategies, incorporating many other techniques. Appropriate use of this technology should allow us to reduce the large number of protein products within the proteome down a much more manageable short list of candidate vaccines, perhaps numbering no more than a few dozens. Such candidates would then require subsequent channeling through recombinant expression, purification, and testing for immunogenicity and protective efficacy [55]. For example, the electroeluting of single protein spots, and the subsequent testing of eluted protein against an APC-T-cell clone system, for immunogenicity, is an interesting combination postgenomic approach which addresses the concept of whole protein antigenicity. Epitomics, the postgenomic identification of epitopes, is also an area falling under the proteomics revolution. Mass spectrometry is now being used routinely to sequence peptides eluted from MHC molecules [166, 167, 168].

The experimental and informatic techniques described above address the determination of immunogenicity, albeit parenthetically. Immunogenicity is one of the most widely used terms within immunobiology. Simply, immunogenicity is that property of a chemical moiety—be that protein, lipid, carbohydrate, or some combination thereof—that allows it to induce a significant response of the immune system. An exact definition might not be possible to formulate, being dependent on context. Put simply, a protein which is highly immunogenic within one species, within one population, or within one particular individual within a population is not necessarily immunogenic within another species, population, or individual. Immunogenicity is not the same as protective immunity although it is bound up with it, particularly from an immunovaccinology perspective. Protective immunity is, essentially, an enhanced immunity to reinfection, or to a first infection in the case of a successful vaccine. It is composed of an augmentation of preformed immune reactants, such as antigen-specific antibodies, and the formation of long lasting immune memory, which is mediated by memory B cells and memory T cells. Immunogenicity, per se, is an obvious requirement for protective immunity, yet while it is necessary, it is also clearly not sufficient. There are other factors—probably many other factors—as yet unknown, that mediate protection.

Although we cannot easily define immunogenicity, nonetheless, a fundamental understanding of immunological mechanisms operating, within this context, at the molecular level underlies most modern attempts to design vaccines rationally. The newly emergent discipline of immunovaccinology is bound up with the development of immunobiology as a postgenomic science. The sequences of genomes from both microbial pathogens and vertebrate hosts are now available, and the power of parallel approaches such as transcriptomics and proteomics is now being felt in the search for new vaccines. However, the manifestation of immunology at the whole animal level is an exceedingly complex phenomenon. It is only by investigating each of its individual stages, at the level of interacting molecules and cells, and in a physicochemical manner, that we can hope to formulate ways of modelling and manipulating the process effectively.

REFERENCES

[1] Paine K, Flower DR. The lipocalin website. *Biochim Biophys Acta*. 2000;1482(1-2):351–352.

[2] Ji H, Zhou Q, Wen F, Xia H, Lu X, Li Y. AsMamDB: an alternative splice database of mammals. *Nucleic Acids Res*. 2001;29(1):260–263.

[3] Perler FB. InBase: the Intein Database. *Nucleic Acids Res*. 2002;30(1):383–384.

[4] Wasinger VC, Cordwell SJ, Cerpa-Poljak A, et al. Progress with gene-product mapping of the mollicutes: *Mycoplasma genitalium*. *Electrophoresis*. 1995;16(7):1090–1094.

[5] Anderson NG, Anderson L. The human protein index. *Clin Chem*. 1982;28(4 pt 2):739–748.

[6] Holtappels R, Grzimek NK, Thomas D, Reddehase MJ. Early gene m18, a novel player in the immune response to murine cytomegalovirus. *J Gen Virol*. 2002;83(pt 2):311–316.

[7] Holtappels R, Thomas D, Podlech J, Reddehase MJ. Two antigenic peptides from genes m123 and m164 of murine cytomegalovirus quantitatively dominate CD8 T-cell memory in the H-2d haplotype. *J Virol*. 2002;76(1):151–164.

[8] Dalton JP, Mulcahy G. Parasite vaccines—a reality? *Vet Parasitol*. 2001;98(1-3):149–167.

[9] Chakravarti DN, Fiske MJ, Fletcher LD, Zagursky RJ. Mining genomes and mapping proteomes: identification and characterization of protein subunit vaccines. *Dev Biol (Basel)*. 2000;103:81–90.

[10] Toby GG, Golemis EA. Using the yeast interaction trap and other two-hybrid-based approaches to study protein-protein interactions. *Methods*. 2001; 24(3):201–217.

[11] Templin MF, Stoll D, Schrenk M, Traub PC, Vohringer CF, Joos TO. Protein microarray technology. *Trends Biotechnol*. 2002;20(4):160–166.

[12] Walker J, Flower D, Rigley K. Microarrays in hematology. *Curr Opin Hematol*. 2002;9(1):23–29.

[13] Grandi G. Antibacterial vaccine design using genomics and proteomics. *Trends Biotechnol*. 2001; 19(5):181–188.

[14] Rathod PK, Ganesan K, Hayward RE, Bozdech Z, DeRisi JL. DNA microarrays for malaria. *Trends Parasitol*. 2002;18(1):39–45.

[15] Knox DP, Redmond DL, Skuce PJ, Newlands GF. The contribution of molecular biology to the development of vaccines against nematode and trematode parasites of domestic ruminants. *Vet Parasitol*. 2001;101(3-4):311–335.

[16] Glynne RJ, Watson SR. The immune system and gene expression microarrays—new answers to old questions. *J Pathol*. 2001;195(1):20–30.

[17] Dhiman N, Bonilla R, O'Kane DJ, Poland GA. Gene expression microarrays: a 21st century tool for directed vaccine design. *Vaccine*. 2001;20(1-2):22–30.

[18] de Veer MJ, Holko M, Frevel M, et al. Functional classification of interferon-stimulated genes identified using microarrays. *J Leukoc Biol*. 2001; 69(6):912–920.

[19] Emslie KR, Molloy MP, Barardi CR, et al. Serotype classification and characterisation of the rotavirus SA11 VP6 protein using mass spectrometry and two-dimensional gel electrophoresis. *Funct Integr Genomics*. 2000;1(1):12–24.

[20] Thomas JJ, Bakhtiar R, Siuzdak G. Mass spectrometry in viral proteomics. *Acc Chem Res*. 2000; 33(3):179–187.

[21] Yao X, Freas A, Ramirez J, Demirev PA, Fenselau C. Proteolytic 18O labeling for comparative proteomics: model studies with two serotypes of adenovirus. *Anal Chem*. 2001;73(13):2836–2842.

[22] Toda T, Sugimoto M, Omori A, Matsuzaki T, Furuichi Y, Kimura N. Proteomic analysis of Epstein-Barr virus-transformed human B-lymphoblastoid cell lines before and after immortalization. *Electrophoresis*. 2000;21(9):1814–1822.

[23] Diaz JJ, Giraud S, Greco A. Alteration of ribosomal protein maps in herpes simplex virus type 1 infection. *J Chromatogr B Analyt Technol Biomed Life Sci*. 2002;771(1-2):237–249.

[24] Greco A, Bienvenut W, Sanchez JC, et al. Identification of ribosome-associated viral and cellular basic proteins during the course of infection with herpes simplex virus type 1. *Proteomics*. 2001;1(4):545–549.

[25] Rodriguez JM, Salas ML, Santaren JF. African swine fever virus-induced polypeptides in porcine alveolar macrophages and in Vero cells: two-dimensional gel analysis. *Proteomics*. 2001;1(11):1447–1456.

[26] Blattner FR, Plunkett G 3rd, Bloch CA, et al. The complete genome sequence of *Escherichia coli* K-12. *Science*. 1997;277(5331):1453–1474.

[27] Fleischmann RD, Adams MD, White O, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*. 1995; 269(5223):496–512.

[28] Fraser CM, Gocayne JD, White O, et al. The minimal gene complement of *Mycoplasma genitalium*. *Science*. 1995;270(5235):397–403.

[29] Bult CJ, White O, Olsen GJ, et al. Complete genome sequence of the methanogenic archaeon, *Methanococcus jannaschii*. *Science*. 1996;273(5278): 1058–1073.

[30] Tomb JF, White O, Kerlavage AR, et al. The complete genome sequence of the gastric pathogen *Helicobacter pylori*. *Nature*. 1997;388(6642):539–547.

[31] Paine K, Flower DR. Bacterial bioinformatics: pathogenesis and the genome. *J Mol Microbiol Biotechnol*. 2002;4(4):357–365.

[32] O'Connor CD, Farris M, Fowler R, Qi SY. The proteome of *Salmonella enterica* serovar Typhimurium: current progress on its determination and some applications. *Electrophoresis*. 1997;18(8): 1483–1490.

[33] Hecker M, Engelmann S. Proteomics, DNA arrays and the analysis of still unknown regulons and unknown proteins of *Bacillus subtilis* and pathogenic gram-positive bacteria. *Int J Med Microbiol*. 2000;290(2):123–134.

[34] Regula JT, Ueberle B, Boguth G, et al. Towards a two-dimensional proteome map of *Mycoplasma pneumoniae*. *Electrophoresis*. 2000;21(17):3765–3780.

[35] Vandahl BB, Birkelund S, Demol H. et al. Proteome analysis of the *Chlamydia pneumoniae* elementary body. *Electrophoresis*. 2001;22(6):1204–1223.

[36] Shaw AC, Gevaert K, Demol H, et al. Comparative proteome analysis of *Chlamydia trachomatis* serovar A, D, and L2. *Proteomics*. 2002;2(2):164–186.

[37] Piechaczek K, Dobrindt U, Schierhorn A, Fischer GS, Hecker M, Hacker J. Influence of pathogenicity islands and the minor leuX-encoded tRNA5Leu on the proteome pattern of the uropathogenic *Escherichia coli* strain 536. *Int J Med Microbiol*. 2000;290(1):75–84.

[38] Malhotra S, Silo-Suh LA, Mathee K, Ohman DE. Proteome analysis of the effect of mucoid conversion on global protein expression in *Pseudomonas aeruginosa* strain PAO1 shows induction of the disulfide bond isomerase, dsbA. *J Bacteriol*. 2000;182(24):6999–7006.

[39] Rosenkrands I, King A, Weldingh K, Moniatte M, Moertz E, Andersen P. Towards the proteome of *Mycobacterium tuberculosis*. *Electrophoresis*. 2000;21(17):3740–3756.

[40] Rosenkrands I, Weldingh K, Jacobsen S, et al. Mapping and identification of *Mycobacterium tuberculosis* proteins by two-dimensional gel electrophoresis, microsequencing and immunodetection. *Electrophoresis*. 2000;21(5):935–948.

[41] Jungblut PR, Muller EC, Mattow J, Kaufmann SH. Proteomics reveals open reading frames in *Mycobacterium tuberculosis* H37Rv not predicted by genomics. *Infect Immun*. 2001;69(9):5905–5907.

[42] Jungblut PR, Schaible UE, Mollenkopf HJ, et al. Comparative proteome analysis of *Mycobacterium tuberculosis* and *Mycobacterium bovis* BCG strains: towards functional genomics of microbial pathogens. *Mol Microbiol*. 1999;33(6):1103–1117.

[43] Mattow J, Jungblut PR, Schaible UE, et al. Identification of proteins from *Mycobacterium tuberculosis* missing in attenuated *Mycobacterium bovis* BCG strains. *Electrophoresis*. 2001;22(14):2936–2946.

[44] Wong DK, Lee BY, Horwitz MA, Gibson BW. Identification of fur, aconitase, and other proteins expressed by *Mycobacterium tuberculosis* under conditions of low and high concentrations of iron by combined two-dimensional gel electrophoresis and mass spectrometry. *Infect Immun*. 1999;67(1):327–336.

[45] Monahan IM, Betts J, Banerjee DK, Butcher PD. Differential expression of mycobacterial proteins following phagocytosis by macrophages. *Microbiology*. 2001;147(pt 2):459–471.

[46] Betts JC, Lukey PT, Robb LC, McAdam RA, Duncan K. Evaluation of a nutrient starvation model of *Mycobacterium tuberculosis* persistence by gene and protein expression profiling. *Mol Microbiol*. 2002;43(3):717–731.

[47] Betts JC, Dodson P, Quan S, et al. Comparison of the proteome of *Mycobacterium tuberculosis* strain H37Rv with clinical isolate CDC 1551. *Microbiology*. 2000;146(pt 12):3205–3216.

[48] Covert BA, Spencer JS, Orme IM, Belisle JT. The application of proteomics in defining the T cell antigens of *Mycobacterium tuberculosis*. *Proteomics*. 2001;1(4):574–586.

[49] Langenberg M-L, Tytgat GNJ, Schipper MEI, Rietra PJGM, Zanen HC. Campylobacter-like organism in the stomach of patients and healthy individuals. *Lancet*. 1984;ii(1348).

[50] Utt M, Nilsson I, Ljungh A, Wadstrom T. Identification of novel immunogenic proteins of *Helicobacter pylori* by proteome technology. *J Immunol Methods*. 2002;259(1-2):1–10.

[51] McAtee CP, Lim MY, Fung K, et al. Identification of potential diagnostic and vaccine candidates of *Helicobacter pylori* by two-dimensional gel electrophoresis, sequence analysis, and serum profiling. *Clin Diagn Lab Immunol*. 1998;5(4):537–542.

[52] Kimmel B, Bosserhoff A, Frank R, Gross R, Goebel W, Beier D. Identification of immunodominant antigens from *Helicobacter pylori* and evaluation of their reactivities with sera from patients with different gastroduodenal pathologies. *Infect Immun*. 2000;68(2):915–920.

[53] Jungblut PR, Bumann D, Haas G, et al. Comparative proteome analysis of *Helicobacter pylori*. *Mol Microbiol*. 2000;36(3):710–725.

[54] Nilsson I, Utt M, Nilsson HO, Ljungh A, Wadstrom T. Two-dimensional electrophoretic and immunoblot analysis of cell surface proteins of spiral-shaped and coccoid forms of *Helicobacter pylori*. *Electrophoresis*. 2000;21(13):2670–2677.

[55] Chakravarti DN, Fiske MJ, Fletcher LD, Zagursky RJ. Application of genomics and proteomics for identification of bacterial gene products as potential vaccine candidates. *Vaccine*. 2000;19(6):601–612.

[56] Haas G, Karaali G, Ebermayer K, et al. Immunoproteomics of *Helicobacter pylori* infection and relation to gastric disease. *Proteomics*. 2002;2(3):313–324.

[57] McAtee CP, Hoffman PS, Berg DE. Identification of differentially regulated proteins in metronidozole resistant *Helicobacter pylori* by proteome techniques. *Proteomics*. 2001;1(4):516–521.

[58] Yoder OC, Turgeon BG. Fungal genomics and pathogenicity. *Curr Opin Plant Biol*. 2001;4(4):315–321.
[59] Yoder OC, Turgeon BG. Molecular genetic evaluation of fungal molecules for roles in pathogenesis to plants. *J Genet*. 1996;75:425–440.
[60] Oliver R, Osbourn A. Molecular dissection of fungal phytopathogenicity. *Microbiology*. 1995;141 (pt 1):1–9.
[61] Hogan LH, Klein BS, Levitz SM. Virulence factors of medically important fungi. *Clin Microbiol Rev*. 1996;9(4):469–488.
[62] Lim D, Hains P, Walsh B, Bergquist P, Nevalainen H. Proteins associated with the cell envelope of *Trichoderma reesei*: a proteomic approach. *Proteomics*. 2001;1(7):899–909.
[63] Gutierrez JA. Genomics: from novel genes to new therapeutics in parasitology. *Int J Parasitol*. 2000;30(3):247–252.
[64] Colley DG. Parasitic diseases: opportunities and challenges in the 21st century. *Mem Inst Oswaldo Cruz*. 2000;95(suppl 1):79–87.
[65] Knox DP, Redmond DL, Skuce PJ, Newlands GF. The contribution of molecular biology to the development of vaccines against nematode and trematode parasites of domestic ruminants. *Vet Parasitol*. 2001;101(3-4):311–335.
[66] Jefferies JR, Campbell AM, van Rossum AJ, Barrett J, Brophy PM. Proteomic analysis of *Fasciola hepatica* excretory-secretory products. *Proteomics*. 2001;1(9):1128–1132.
[67] Almeida R, Norrish A, Levick M, et al. From genomes to vaccines: *Leishmania* as a model. *Philos Trans R Soc Lond B Biol Sci*. 2002;357(1417):5–11.
[68] Thiel M, Bruchhaus I. Comparative proteome analysis of *Leishmania donovani* at different stages of transformation from promastigotes to amastigotes. *Med Microbiol Immunol (Berl)*. 2001;190(1-2):33–36.
[69] Cohen AM, Rumpel K, Coombs GH, Wastling JM. Characterisation of global protein expression by two-dimensional electrophoresis and mass spectrometry: proteomics of *Toxoplasma gondii*. *Int J Parasitol*. 2002;32(1):39–51.
[70] Dlugonska H, Dytnerska K, Reichmann G, Stachelhaus S, Fischer HG. Towards the *Toxoplasma gondii* proteome: position of 13 parasite excretory antigens on a standardized map of two-dimensionally separated tachyzoite proteins. *Parasitol Res*. 2001;87(8):634–637.
[71] Fletcher JM, Nair SP, Ward JM, Henderson B, Wilson M. Analysis of the effect of changing environmental conditions on the expression patterns of exported surface-associated proteins of the oral pathogen *Actinobacillus actinomycetemcomitans*. *Microb Pathog*. 2001;30(6):359–368.
[72] Monahan IM, Betts J, Banerjee DK, Butcher PD. Differential expression of mycobacterial proteins following phagocytosis by macrophages. *Microbiology*. 2001;147(pt 2):459–471.
[73] Ragno S, Romano M, Howell S, Pappin DJ, Jenner PJ, Colston MJ. Changes in gene expression in macrophages infected with *Mycobacterium tuberculosis*: a combined transcriptomic and proteomic approach. *Immunology*. 2001;104(1):99–108.
[74] Kovarova H, Halada P, Man P, et al. Proteome study of *Francisella tularensis* live vaccine strain-containing phagosome in Bcg/Nramp1 congenic macrophages: resistant allele contributes to permissive environment and susceptibility to infection. *Proteomics*. 2002;2(1):85–93.
[75] Pizarro-Cerda J, Jonquieres R, Gouin E, Vandekerckhove J, Garin J, Cossart P. Distinct protein patterns associated with *Listeria monocytogenes* InlA- or InlB-phagosomes. *Cell Microbiol*. 2002;4(2):101–115.
[76] Truffa-Bachi P, Lefkovits I, Frey JR. Proteomic analysis of T cell activation in the presence of cyclosporin A: immunosuppressor and activator removal induces de novo protein synthesis [Erratum in Mol Immunol. 2000;37(5):261]. *Mol Immunol*. 2000;37(1-2):21–28.
[77] Nyman TA, Rosengren A, Syyrakki S, Pellinen TP, Rautajoki K, Lahesmaa R. A proteome database of human primary T helper cells. *Electrophoresis*. 2001;22(20):4375–4382.
[78] Fratelli M, Demol H, Puype M, et al. Identification by redox proteomics of glutathionylated proteins in oxidatively stressed human T lymphocytes. *Proc Natl Acad Sci USA*. 2002;99(6):3505–3510.
[79] Williams A. Applications of computer software for the interpretation and management of mass spectrometry data in pharmaceutical science. *Curr Top Med Chem*. 2002;2(1):99–107.
[80] Sidhu KS, Sangvanich P, Brancia FL, et al. Bioinformatic assessment of mass spectrometric chemical derivatisation techniques for proteome database searching. *Proteomics*. 2001;1(11):1368–1377.
[81] Vihinen M. Bioinformatics in proteomics. *Biomol Eng*. 2001;18(5):241–248.
[82] Chakravarti DN, Chakravarti B, Moutsatsos I. Informatic tools for proteome profiling. *Biotechniques*. 2002;(suppl):4–15.
[83] Nussbaum AK, Dick TP, Keilholz W, et al. Cleavage motifs of the yeast 20S proteasome beta subunits deduced from digests of enolase 1. *Proc Natl Acad Sci USA*. 1998;95(21):12504–12509.
[84] Holzhutter HG, Frommel C, Kloetzel PM. A theoretical approach towards the identification of cleavage-determining amino acid motifs of the 20S proteasome. *J Mol Biol*. 1999;286(4):1251–1265.
[85] Kuttler C, Nussbaum AK, Dick TP, Rammensee HG, Schild H, Hadeler KP. An algorithm for the prediction of proteasomal cleavages. *J Mol Biol*. 2000;298(3):417–429.

[86] Nussbaum AK, Kuttler C, Hadeler KP, Rammensee HG, Schild H. PAProC: a prediction algorithm for proteasomal cleavages available on the WWW. *Immunogenetics*. 2001;53(2):87–94.

[87] Kesmir C, Nussbaum AK, Schild H, Detours V, Brunak S. Prediction of proteasome cleavage motifs by neural networks. *Protein Eng*. 2002;15(4):287–296.

[88] Daniel S, Brusic V, Caillat-Zucman S, et al. Relationship between peptide selectivities of human transporters associated with antigen processing and HLA class I molecules. *J Immunol*. 1998;161(2):617–624.

[89] Brusic V, van Endert P, Zeleznikow J, Daniel S, Hammer J, Petrovsky N. A neural network model approach to the study of human TAP transporter. *In Silico Biol*. 1999;1(2):109–121.

[90] Chapman HA. Endosomal proteolysis and MHC class II function. *Curr Opin Immunol*. 1998;10(1):93–102.

[91] Sette A, Buus S, Appella E, et al. Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis. *Proc Natl Acad Sci USA*. 1989;86(9):3296–3300.

[92] D'Amaro J, Houbiers JG, Drijfhout JW, et al. A computer program for predicting possible cytotoxic T lymphocyte epitopes based on HLA class I peptide-binding motifs. *Hum Immunol*. 1995;43(1):13–18.

[93] Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S. SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics*. 1999;50(3-4):213–219.

[94] Parker KC, Shields M, DiBrino M, Brooks A, Coligan JE. Peptide binding to MHC class I molecules: implications for antigenic peptide prediction. *Immunol Res*. 1995;14(1):34–57.

[95] Parker KC, DiBrino M, Hull L, Coligan JE. The beta 2-microglobulin dissociation rate is an accurate measure of the stability of MHC class I heterotrimers and depends on which peptide is bound. *J Immunol*. 1992;149(6):1896–1904.

[96] Parker KC, Bednarek MA, Coligan JE. Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide sidechains. *J Immunol*. 1994;152(1):163–175.

[97] Parker KC, Bednarek MA, Hull LK, et al. Sequence motifs important for peptide binding to the human MHC class I molecule, HLA-A2. *J Immunol*. 1992;149(11):3580–3587.

[98] Parker KC, Carreno BM, Sestak L, Utz U, Biddison WE, Coligan JE. Peptide binding to HLA-A2 and HLA-B27 isolated from *Escherichia coli*. Reconstitution of HLA-A2 and HLA-B27 heavy chain/beta 2-microglobulin complexes requires specific peptides. *J Biol Chem*. 1992;267(8):5451–5459.

[99] DiBrino M, Parker KC, Shiloach J, et al. Endogenous peptides bound to HLA-A3 possess a specific combination of anchor residues that permit identification of potential antigenic peptides. *Proc Natl Acad Sci USA*. 1993;90(4):1508–1512.

[100] DiBrino M, Parker KC, Margulies DH, et al. The HLA-B14 peptide binding site can accommodate peptides with different combinations of anchor residues. *J Biol Chem*. 1994;269(51):32426–32434.

[101] Parker KC, Biddison WE, Coligan JE. Pocket mutations of HLA-B27 show that anchor residues act cumulatively to stabilize peptide binding. *Biochemistry*. 1994;33(24):7736–7743.

[102] DiBrino M, Parker KC, Margulies DH, et al. Identification of the peptide binding motif for HLA-B44, one of the most common HLA-B alleles in the Caucasian population. *Biochemistry*. 1995;34(32):10130–10138.

[103] DiBrino M, Tsuchida T, Turner RV, Parker KC, Coligan JE, Biddison WE. HLA-A1 and HLA-A3 T cell epitopes derived from influenza virus proteins predicted from peptide binding motifs. *J Immunol*. 1993;151(11):5930–5935.

[104] Honma K, Parker KC, Becker KG, McFarland HF, Coligan JE, Biddison WE. Identification of an epitope derived from human proteolipid protein that can induce autoreactive $CD8^+$ cytotoxic T lymphocytes restricted by HLA-A3: evidence for cross-reactivity with an environmental microorganism. *J Neuroimmunol*. 1997;73(1-2):7–14.

[105] Bisset LR, Fierz W. Using a neural network to identify potential HLA-DR1 binding sites within proteins. *J Mol Recognit*. 1993;6(1):41–48.

[106] Brusic V, Schonbach C, Takiguchi M, Ciesielski V, Harrison LC. Application of genetic search in derivation of matrix models of peptide binding to MHC molecules. *Proc Int Conf Intell Syst Mol Biol*. 1997;5:75–83.

[107] Harrison LC, Honeyman MC, Trembleau S, et al. A peptide-binding motif for I-A(g7), the class II major histocompatibility complex (MHC) molecule of NOD and Biozzi AB/H mice. *J Exp Med*. 1997;185(6):1013–1021.

[108] Honeyman MC, Brusic V, Harrison LC. Strategies for identifying and predicting islet autoantigen T-cell epitopes in insulin-dependent diabetes mellitus. *Ann Med*. 1997;29(5):401–404.

[109] Honeyman MC, Brusic V, Stone NL, Harrison LC. Neural network-based prediction of candidate T-cell epitopes. *Nat Biotechnol*. 1998;16(10):966–969.

[110] Brusic V, Rudy G, Harrison LC. MHCPEP, a database of MHC-binding peptides: update 1997. *Nucleic Acids Res*. 1998;26(1):368–371.

[111] Rosenfeld R, Zheng Q, Vajda S, DeLisi C. Flexible docking of peptides to class I major-histocompatibility-complex receptors. *Genet Anal*. 1995;12(1):1–21.

[112] Sezerman U, Vajda S, DeLisi C. Free energy mapping of class I MHC molecules and structural

determination of bound peptides. *Protein Sci.* 1996;5(7):1272–1281.

[113] Vasmatzis G, Zhang C, Cornette JL, DeLisi C. Computational determination of side chain specificity for pockets in class I MHC molecules. *Mol Immunol.* 1996;33(16):1231–1239.

[114] Rognan D, Reddehase MJ, Koszinowski UH, Folkers G. Molecular modeling of an antigenic complex between a viral peptide and a class I major histocompatibility glycoprotein. *Proteins.* 1992; 13(1):70–85.

[115] Rognan D, Zimmermann N, Jung G, Folkers G. Molecular dynamics study of a complex between the human histocompatibility antigen HLA-A2 and the IMP58-66 nonapeptide from influenza virus matrix protein. *Eur J Biochem.* 1992;208(1):101–113.

[116] Rognan D, Scapozza L, Folkers G, Daser A. Molecular dynamics simulation of MHC-peptide complexes as a tool for predicting potential T cell epitopes. *Biochemistry.* 1994;33(38):11476–11485.

[117] Caflisch A, Niederer P, Anliker M. Monte Carlo docking of oligopeptides to proteins. *Proteins.* 1992;13(3):223–230.

[118] Lim JS, Kim S, Lee HG, Lee KY, Kwon TJ, Kim K. Selection of peptides that bind to the HLA-A2.1 molecule by molecular modelling. *Mol Immunol.* 1996;33(2):221–230.

[119] Androulakis IP, Nayak NN, Ierapetritou MG, Monos DS, Floudas CA. A predictive method for the evaluation of peptide binding in pocket 1 of HLA-DRB1 via global minimization of energy interactions. *Proteins.* 1997;29(1):87–102.

[120] Froloff N, Windemuth A, Honig B. On the calculation of binding free energies using continuum methods: application to MHC class I protein-peptide interactions. *Protein Sci.* 1997;6(6):1293–1301.

[121] Arora N, Bashford D. Solvation energy density occlusion approximation for evaluation of desolvation penalties in biomolecular interactions. *Proteins.* 2001;43(1):12–27.

[122] Doytchinova IA, Blythe MJ, Flower DR. Additive method for the prediction of protein-peptide binding affinity. Application to the MHC class I molecule HLA-A∗0201. *J Proteome Res.* 2002;1(3):263–272.

[123] Doytchinova IA, Flower DR. Toward the quantitative prediction of T-cell epitopes: CoMFA and CoMSIA studies of peptides with affinity to the class I MHC molecule HLA-A∗0201. *J Med Chem.* 2001;44(22):3572–3581.

[124] Doytchinova IA, Flower DR. Physicochemical explanation of peptide binding to HLA-A∗0201 major histocompatibility complex: a three-dimensional quantitative structure-activity relationship study. *Proteins.* 2002;48(3):505–518.

[125] Rognan D, Lauemoller SL, Holm A, Buus S, Tschinke V. Predicting binding affinities of protein ligands from three-dimensional models: application to peptide binding to class I major histocompatibility proteins. *J Med Chem.* 1999;42(22):4650–4658.

[126] Logean A, Sette A, Rognan D. Customized versus universal scoring functions: application to class I MHC-peptide binding free energy predictions. *Bioorg Med Chem Lett.* 2001;11(5):675–679.

[127] Altuvia Y, Schueler O, Margalit H. Ranking potential binding peptides to MHC molecules by a computational threading approach. *J Mol Biol.* 1995;249(2):244–250.

[128] Altuvia Y, Sette A, Sidney J, Southwood S, Margalit H. A structure-based algorithm to predict potential binding peptides to MHC molecules with hydrophobic binding pockets. *Hum Immunol.* 1997;58(1):1–11.

[129] Miyazawa S, Jernigan RL. Residue-residue potentials with a favorable contact pair term and an unfavorable high packing density term, for simulation and threading. *J Mol Biol.* 1996;256(3):623–644.

[130] Schueler-Furman O, Altuvia Y, Sette A, Margalit H. Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein Sci.* 2000;9(9):1838–1846.

[131] Betancourt MR, Thirumalai D. Pair potentials for protein folding: choice of reference states and sensitivity of predicted native states to variations in the interaction schemes. *Protein Sci.* 1999;8(2):361–369.

[132] Stryhn A, Pedersen LO, Romme T, Holm CB, Holm A, Buus S. Peptide binding specificity of major histocompatibility complex class I resolved into an array of apparently independent subspecificities: quantitation by peptide libraries and improved prediction of binding. *Eur J Immunol.* 1996;26(8):1911–1918.

[133] Stevens J, Wiesmuller KH, Walden P, Joly E. Peptide length preferences for rat and mouse MHC class I molecules using random peptide libraries. *Eur J Immunol.* 1998;28(4):1272–1279.

[134] Stevens J, Wiesmuller KH, Barker PJ, Walden P, Butcher GW, Joly E. Efficient generation of major histocompatibility complex class I-peptide complexes using synthetic peptide libraries. *J Biol Chem.* 1998;273(5):2874–2884.

[135] Zhao Y, Gran B, Pinilla C, et al. Combinatorial peptide libraries and biometric score matrices permit the quantitative analysis of specific and degenerate interactions between clonotypic TCR and MHC peptide ligands. *J Immunol.* 2001;167(4):2130–2141.

[136] Pinilla C, Rubio-Godoy V, Dutoit V, et al. Combinatorial peptide libraries as an alternative approach to the identification of ligands for tumor-reactive cytolytic T lymphocytes. *Cancer Res.* 2001; 61(13):5153–5160.

[137] Udaka K, Wiesmuller KH, Kienle S, et al. An automated prediction of MHC class I-binding peptides based on positional scanning with peptide libraries. *Immunogenetics*. 2000;51(10):816–828.

[138] Blythe MJ, Doytchinova IA, Flower DR. JenPep: a database of quantitative functional peptide data for immunology. *Bioinformatics*. 2002;18(3):434–439.

[139] Van Regenmortel MH, Daney de Marcillac G. An assessment of prediction methods for locating continuous epitopes in proteins. *Immunol Lett*. 1988;17(2):95–107.

[140] Ferreira-da-Cruz Mde F, Giovanni-de-Simone S, Banic DM, Canto-Cavalheiro M, Camus D, Daniel-Ribeiro CT. Can software be used to predict antigenic regions in *Plasmodium falciparum* peptides? *Parasite Immunol*. 1996;18(3):159–161.

[141] Thornton JM, Edwards MS, Taylor WR, Barlow DJ. Location of "continuous" antigenic determinants in the protruding regions of proteins. *EMBO J*. 1986;5(2):409–413.

[142] Barlow DJ, Edwards MS, Thornton JM. Continuous and discontinuous protein antigenic determinants. *Nature*. 1986;322(6081):747–748.

[143] Van Regenmortel MH, Pellequer JL. Predicting antigenic determinants in proteins: looking for unidimensional solutions to a three-dimensional problem? *Pept Res*. 1994;7(4):224–228.

[144] Alix AJ. Predictive estimation of protein linear epitopes by using the program PEOPLE. *Vaccine*. 1999;18(3-4):311–314.

[145] Pellequer JL, Westhof E. PREDITOP: a program for antigenicity prediction. *J Mol Graph*. 1993;11(3): 204–210.

[146] Saleh MT, Fillon M, Brennan PJ, Belisle JT. Identification of putative exported/secreted proteins in prokaryotic proteomes. *Gene*. 2001;269(1-2):195–204.

[147] Kumar A, Agarwal S, Heyman JA, et al. Subcellular localization of the yeast proteome. *Genes Dev*. 2002;16(6):707–719.

[148] Gromiha MM. A simple method for predicting transmembrane alpha helices with better accuracy. *Protein Eng*. 1999;12(7):557–561.

[149] Nishikawa K, Ooi T. Correlation of the amino acid composition of a protein to its structural and biological characters. *J Biochem (Tokyo)*. 1982; 91(5):1821–1824.

[150] Eisenhaber F, Frőmmel C, Argos P. Prediction of secondary structural content of proteins from their amino acid composition alone. II. The paradox with secondary structural class. *Proteins*. 1996; 25(2):169–179.

[151] Chiapello H, Ollivier E, Landes-Devauchelle C, Nitschke P, Risler JL. Codon usage as a tool to predict the cellular location of eukaryotic ribosomal proteins and aminoacyl-tRNA synthetases. *Nucleic Acids Res*. 1999;27(14):2848–2851.

[152] Chou KC, Elrod DW. Using discriminant function for prediction of subcellular location of prokaryotic proteins. *Biochem Biophys Res Commun*. 1998; 252(1):63–68.

[153] Cedano J, Aloy P, Pērez-Pons JA, Querol E. Relation between amino acid composition and cellular location of proteins. *J Mol Biol*. 1997;266(3):594–600.

[154] Nakashima H, Nishikawa K. Discrimination of intracellular and extracellular proteins using amino acid composition and residue-pair frequencies. *J Mol Biol*. 1994;238(1):54–61.

[155] Andrade MA, O'Donoghue SI, Rost B. Adaptation of protein surfaces to subcellular location. *J Mol Biol*. 1998;276(2):517–525.

[156] Reinhardt A, Hubbard T. Using neural networks for prediction of the subcellular location of proteins. *Nucleic Acids Res*. 1998;26(9):2230–2236.

[157] Nakai K. Predicting various targeting signals in amino acid sequences. *Bull Inst Chem Res Kyoto Univ*. 1991;69:269–291.

[158] Nakai K. Protein sorting signals and prediction of subcellular localization. *Adv Protein Chem*. 2000; 54:277–344.

[159] Briggs MS, Gierasch LM, Zlotnick A, Lear JD, DeGrado WF. In vivo function and membrane binding properties are correlated for *Escherichia coli* lamB signal peptides. *Science*. 1985; 228(4703):1096–1099.

[160] von Heijne G. Signal sequences. The limits of variation. *J Mol Biol*. 1985;184(1):99–105.

[161] Martoglio B, Dobberstein B. Signal sequences: more than just greasy peptides. *Trends Cell Biol*. 1998;8(10):410–415.

[162] Sjőstrőm M, Wold S, Wieslander A, Rilfors L. Signal peptide amino acid sequences in *Escherichia coli* contain information related to final protein localization. A multivariate data analysis. *EMBO J*. 1987;6(3):823–831.

[163] Edman M, Jarhede T, Sjőstrőm M, Wieslander A. Different sequence patterns in signal peptides from mycoplasmas, other gram-positive bacteria, and *Escherichia coli*: a multivariate data analysis. *Proteins*. 1999;35(2):195–205.

[164] Nielsen H, Engelbrecht J, Brunak S, von Heijne G. Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng*. 1997;10(1):1–6.

[165] Jagla B, Schuchhardt J. Adaptive encoding neural networks for the recognition of human signal peptide cleavage sites. *Bioinformatics*. 2000;16(3):245–250.

[166] Mouritsen S, Dalum I, Engel AM, et al. MHC class II-bound self-peptides can be effectively separated by isoelectric focusing and bind optimally to their MHC class II restriction elements around pH 5.0. *Immunology*. 1994;82(4):529–534.

[167] Dongre AR, Kovats S, deRoos P, et al. In vivo MHC class II presentation of cytosolic proteins

revealed by rapid automated tandem mass spectrometry and functional analyses. *Eur J Immunol.* 2001;31(5):1485–1494.

[168] Purcell AW, Gorman JJ. The use of post-source decay in matrix-assisted laser desorption/ionisation mass spectrometry to delineate T cell determinants. *J Immunol Methods*. 2001;249(1-2):17–31.

* Corresponding author.
E-mail: darren.flower@jenner.ac.uk
Fax: + 44 1635 577901; Tel: + 44 1635 577954

Journal of Biomedicine and Biotechnology • 2003:5 (2003) 291–298 • PII. S1110724303209256 • http://jbb.hindawi.com

RESEARCH ARTICLE

Use of Immunomatrix Methods to Improve Protein-Protein Interaction Detection

M. Walid Qoronfleh,* Ling Ren, Daryl Emery, Maria Perr, and Barbara Kaboord

Bioresearch Division, Perbio Science, 2202 North Bartlett Avenue, Milwaukee, WI 53202-1009, USA

Received 14 June 2002; accepted 18 December 2002

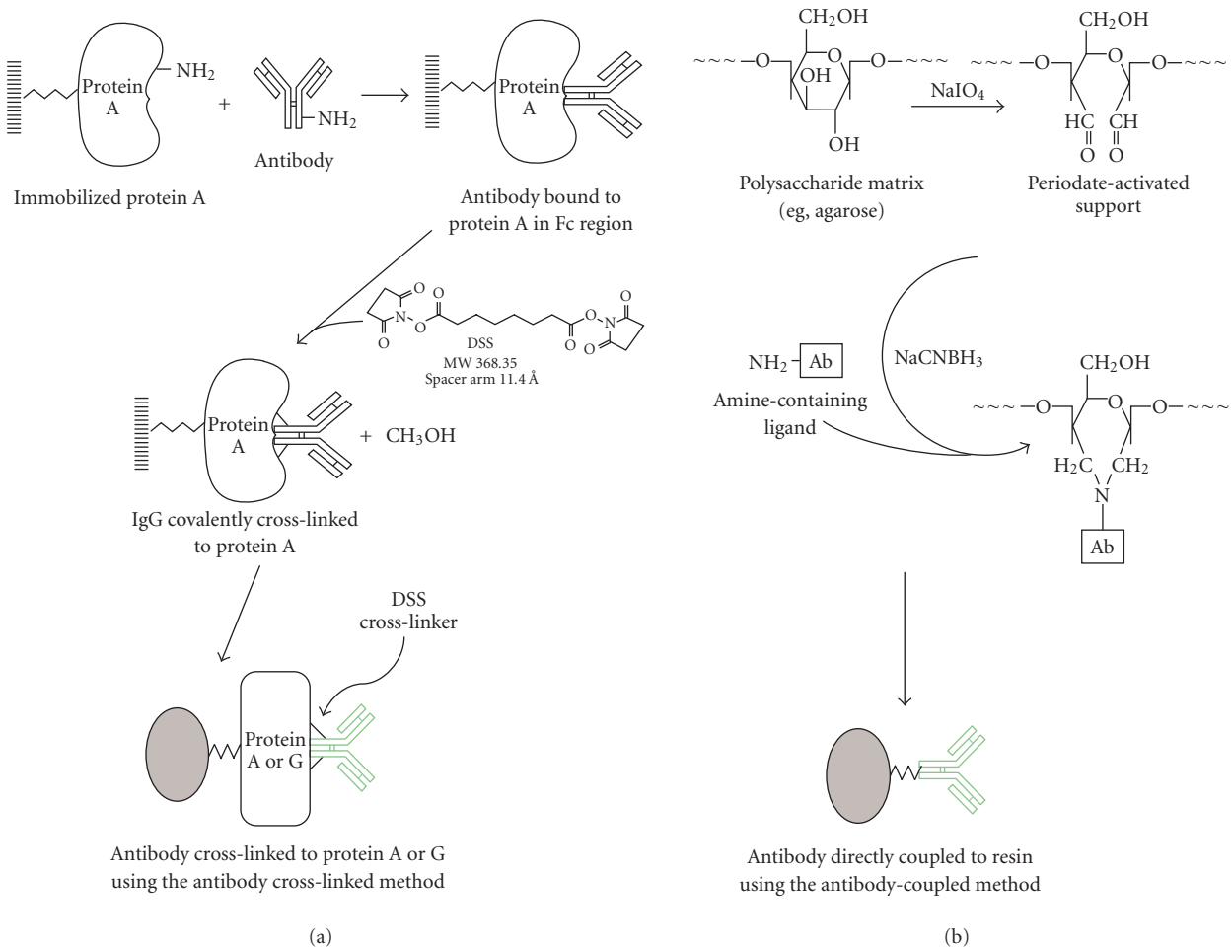
Immunoprecipitation (IP) and coimmunoprecipitation (co-IP) are key techniques for studying protein-protein interactions. These methods utilize immobilized protein A or protein G to isolate antibody-bound target antigens. The main disadvantage of traditional immunoprecipitation and coimmunoprecipitation is that the conditions used to elute the precipitated antigen also release the antibody, contaminating the antigen and destroying the antibody support. To overcome these problems, we describe two methods to generate a reusable antibody support by cross-linking the antibody to immobilized protein A or protein G, or by coupling it directly to the resin. Our studies have demonstrated that the immobilization efficiency for the antibody coupling method was similar for several species of antibody. Furthermore, we illustrate that using both methods of antibody immobilization yields IP and co-IP results similar to traditional protocols but eliminates the antibody heavy and light chains contamination.

INTRODUCTION

Immunoprecipitation (IP) is a powerful immunochemical technique that has been used to study antigen characteristics such as antigen presence and quantity, relative molecular weight, rate of synthesis or degradation, posttranslational modifications, and interactions with proteins, nucleic acids, or ligands [1, 2, 3]. The IP procedure involves extracting antigens from cells in an appropriate lysis buffer, incubating the lysate with antibody to allow formation of immune complexes, and precipitating those complexes with immobilized protein A or protein G.

Coimmunoprecipitation (co-IP) is a key technique used to study protein-protein interactions [4]. Co-IP has been widely used to study receptor-ligand interactions [5], enzyme-substrate interactions [6], and interactions of subunits within a protein complex [7]. Co-IP of cell or tissue extract is also used to confirm yeast two-hybrid screening results [8, 9, 10]. Typically, an antibody specific for one protein is incubated with a cell lysate or a protein mixture to form an immune complex with the target protein (antigen). The target protein may be interacting with one or other more proteins to form a protein complex (co-complex). The entire co-complex is then precipitated using immobilized protein A or protein G.

Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE) followed by staining, autoradiography, or Western blot analysis is typically used to detect the interacting partners. If the antigen or its interaction partner(s) and the antibody heavy and light chains have similar relative molecular weights then, under reducing conditions, they will comigrate, making analysis of the IP results problematic. Several alternatives are currently used to circumvent this problem. One of these methods is to eliminate the reducing agent in the Laemmli buffer to cause the whole antibody molecule to migrate at the top of the gel, thus separating it from most proteins [11]. This technique, however, utilizes milder sample denaturing conditions which may not disrupt strong interactions within protein complexes and, therefore, may not be useful for co-IP experiments. A second alternative is to probe Western blots with biotinylated primary antibodies [12]. This method is generally less sensitive but must be exercised when the antibodies used for IP and immunoblotting have been generated in the same animal species. In this paper, we present two quick and easy IP and co-IP methods (seize technology) to eliminate antibody contamination in precipitated proteins: the antibody cross-linking method and the antibody coupling method (Seize Technology is a trademark of Pierce Biotechnology, Inc, Rockford, Ill—Scheme 1) that improve protein-protein interaction detection. The first approach uses a chemical cross-linker, disuccinimidyl suberate (DSS), to attach the Fc portion of an antibody to immobilized protein A or protein G. This novel procedure combines cross-linking and affinity chromatography to generate an oriented antibody-protein A or protein G support. The second method couples the antibody directly onto an activated support via lysine residues. This coupling procedure eliminates the need for protein A or protein G and offers universal coupling of all antibody species and subclasses; even chicken IgY and mouse IgG_{2a} can be coupled equally well. Moreover, the antibody supports generated by both methods are reusable.



Scheme 1. Diagram of the antibody bioconjugation chemistry. (a) Description of the antibody cross-linked method. (b) Explanation of the antibody-coupled method.

We have compared the traditional (non-cross-linked), the antibody cross-linked, and the antibody-coupled IP techniques. The traditional IP method gave higher recovery of target protein but contained strong contamination of antibody heavy and light chains while the improved methods showed no antibody contamination. In addition, several protein complexes were precipitated showing that the benefits of antibody-coupled resins extend to co-IP applications as well.

MATERIALS AND METHODS

Materials

Antibodies. Mouse monoclonal anti-T7-tag antibody was purchased from Novagen, Inc (Madison, Wis). The anti-MDM2 monoclonal antibody was bought from Oncogene Research Products (Boston, Mass). The mouse monoclonal antibody to 20S proteasome subunit $\alpha 6$ was purchased from Affinity Research Products Ltd (Exeter, UK). The polyclonal goat anti-GFP (green fluorescent protein) antibody was procured from Pierce Chemical Co (Pierce, Rockford, Ill). During the antibody immobilization processes, all centrifugation steps were performed at $80 \times g$ for 1 minute.

Plasmid DNAs. Plasmid DNA, pGEM-Hsp53, and pGEM-HsMDM2 were kindly provided by Dr Arthur Haas (Medical College of Wisconsin, Milwaukee, Wis), and pET-T7-tag-Max and pET-c-Myc were a gift from Dr Kent Wilcox (Medical College of Wisconsin).

SDS-PAGE. All precast SDS-PAGE gels utilized in our experiments were the Novex brand (Invitrogen, Carlsbad, Calif). Standard electrophoresis conditions recommended by the gel manufacturer were employed. Prestained protein molecular weight marker (BlueRanger) was obtained from Pierce.

Reagents. For reagents and supplies which are not described herein the vendor was Pierce.

Cross-linking antibodies to protein G agarose

Protein G agarose (Pierce) was dispensed into a spin column and washed 2–3 times with modified Dulbecco's phosphate-buffered saline (PBS, 8 mM Na_2PO_3, 2 mM

K_2PO_3, 140 mM NaCl, 10 mM KCl, pH 7.4). A 100–200-μg aliquot of monoclonal or polyclonal antibody was incubated with 400-μL protein G agarose (50% slurry) for one hour at room temperature (RT). After the unbound antibodies were washed away, antibody-bound protein G agarose was resuspended in 400-μL modified Dulbecco's PBS, and 0.1 mL of 13 mg/mL DSS cross-linker freshly prepared in dimethyl sulfoxide (DMSO) was added. The cross-linking reaction was performed at RT for one hour. The excess DSS was removed by washing the resin 4 times with 400-μL of Tris-buffered saline (TBS, 25 mM Tris, 150 mM NaCl, pH 7.2), 4 times with 0.1 M glycine (pH 2.8) to remove free antibody, and finally 3 times with TBS. The cross-linking efficiency was evaluated by A_{280}. The antibody-protein G agarose was stored as 50% slurry at 4°C.

Coupling of antibodies to agarose resin

Coupling of antibodies to agarose resin was performed using AminoLink Plus Coupling Gel (Pierce) in a spin column. Briefly, coupling gel was washed twice with PBS (100 mM Na_2PO_3, 150 mM NaCl, pH 7.2). Affinity-purified antibody diluted in PBS containing 50 mM sodium cyanoborohydride was added to the resin and the mixture was inverted at RT for 4 hours at 1-hour intervals. The flow-through was spun out and the resin was washed once with PBS to remove any uncoupled antibody. A 30-minute incubation with 1 M Tris-HCl, pH 7.4, and 50 mM sodium cyanoborohydride blocked the remaining sites on the resin. The resin was washed 6 times with 1 M NaCl and equilibrated in PBS containing 0.02% sodium azide for storage at 4°C. The flow-through and the first wash were evaluated by A_{280} to determine the coupling efficiency.

In vitro transcription/translation

The TNT T7/Sp6-Coupled Reticulocyte Lysate System (Promega Corp, Madison, WI) was used for the in vitro synthesis of ^{35}S-labeled proteins directly from DNA templates containing T7 or SP6 RNA polymerase promoters. The DNA template (typically 1 μg) was incubated with the transcription/translation mix in a total volume of 50 μL at 30°C for 90 minutes. The synthesized protein products were analyzed by SDS-PAGE and visualized by autoradiography.

Coimmunoprecipitation

Co-IP of 20S proteasome complex with mouse monoclonal antibody to 20S proteasome subunit α6

A 100-μL aliquot of mouse monoclonal anti-α6 (100–1000μg) was cross-linked onto protein G agarose (100-μL of settled resin) (see above procedure). Three flasks (75 cm^2) of 80% confluent HeLa cells ($\sim 5 \times 10^7$ cells) were lysed in 3 mL of M-PER mammalian lysis buffer (Pierce). The cell lysate was then diluted in an equal volume of modified Dulbecco's PBS and precleared with 100-μL protein G resin for 1 hour at 4°C with rotation. The immunoprecipitation was performed overnight at 4°C using 100-μL of the antibody-protein G resin (settled resin). The resin was washed three times with 400 μL of TBS, and the protein complexes were eluted three times with 100-μL 0.1 M glycine (pH 2.8). Elutions were pooled and concentrated using Ultrafree-0.5 centrifugal filtration devices (Millipore, Bedford, Mass). SDS-PAGE (12% gel) and silver staining [13] were carried out for protein detection. For comparison, traditional IP was performed using the same conditions without cross-linking the antibody to the protein G resin.

To demonstrate the antibody specificity, HeLa cell lysate (30 μL) was separated by 12% SDS-PAGE and transferred onto nitrocellulose membrane. The blot was probed with 1 μg/mL mouse monoclonal antibody to 20S proteasome subunit α6 and detected with SuperSignal West Pico chemiluminescent substrate (Pierce).

Co-IP of c-Myc and T7-tagged Max with mouse monoclonal antibody to T7-tag

Two hundred micrograms of mouse anti-T7 tag antibody were cross-linked to protein G agarose (200 μL settled resin). Plasmid DNA pET-T7-tag-Max and pET-c-Myc were used for the in vitro synthesis of ^{35}S-labeled proteins (see the procedures above). Equal amounts of ^{35}S-labeled T7-tagged Max and c-Myc (5 μL) were incubated together for 30 minutes at 30°C. This mixture was added to a spin column which contained antibody-protein G agarose (100-μL of settled resin) in 400-μL modified Dulbecco's PBS. The co-IP was carried out at 4°C for 2 hours with constant rotation. The resin was washed four times with 400-μL TBS and proteins were eluted three times with 0.1 M glycine (pH 2.8) and concentrated. The eluted protein complexes were resolved on 4–20% SDS-PAGE. The gel was washed in Milli-Q water for 5 minutes, soaked in Amplify (Amersham Biosciences, Piscataway, NJ) for 15–30 minutes, dried, and exposed to Kodak MS film with intensifying screens (Kodak, Rochester, NY) at −70°C overnight. Luciferase (^{35}S-labeled) was incubated with T7-tagged Max as a negative control.

Co-IP of human p53 and MDM2 with mouse monoclonal antibody to human MDM2

Plasmid DNA pGEM-Hsp53 and pGEM-HsMDM2 were used for the in vitro synthesis of ^{35}S-labeled proteins (see the procedure above). One hundred micrograms of anti-human MDM2 antibody were coupled to 200 μL agarose using the antibody coupling method (see the procedure above). Equal amounts of ^{35}S-labeled human p53 and MDM2 (4 μL) were incubated together for 30 minutes at 30°C. This mixture was added to the antibody-coupled agarose (60 μL settled resin) in 200 μL modified Dulbecco's PBS with protease inhibitors (Roche Molecular Biochemicals, Indianapolis, Ind) and rotated for 2 hours at 4°C. The resin was then washed four times with TBS and proteins were eluted three times with 0.1 M glycine (pH 2.8) and concentrated. The eluted protein complexes were resolved by 4–12% SDS-PAGE and

detected with autoradiography. Luciferase (^{35}S-labeled) was incubated with MDM2 as a negative control.

Immunoprecipitation

E coli BL21 cells (Novagen) containing 6×His-tagged GFP plasmid were induced and lysed with B-PER bacterial protein extraction reagent (Pierce). The 6×His-GFP protein was partially purified using nickel resin (Pierce). Anti-GFP antibody (130 μg) was coupled onto AminoLink Plus Coupling Gel, cross-linked to protein G agarose (see above procedures), or just mixed with protein G agarose (all used 100-μL of settled resin). The 6×His-GFP fusion protein (135 μg) purified with nickel resin was mixed separately with the three antibody resins in 100-μL modified Dulbecco's PBS in spin columns. The IP was carried out at 4°C for one hour with rotation. The resin-bound antigen was washed three times with 400-μL TBS, and the bound antigen was eluted three times with 100-μL of 0.1 M glycine (pH 2.8). One fifth of each elution fraction was resolved on 12% SDS-PAGE under reducing conditions and detected by Coomassie staining.

RESULTS AND DISCUSSION

Antibody conjugation chemistry

The two novel procedures for antibody immobilization combine linker and resin chromatography techniques to generate either a cross-linked or coupled antibody activated support. Chemical reactions in the antibody immobilization process are described in Scheme 1. For further details of the bioconjugate chemistry, see [14, 15]. The chemistry of conjugation is briefly discussed below. The scheme also depicts a representation of the antibody-bound resin for the two developed methods.

Cross-linked antibody method and co-IP applications

In the antibody cross-linked method, DSS is used to covalently link the Fc portion of the antibody to the protein G or protein A agarose, generating a reusable antibody support. DSS is a water-insoluble, noncleavable, homobifunctional N-hydroxysuccinimidyl (NHS) ester cross-linker [14, 15]. This cross-linker is widely used for conjugating radiolabeled ligands to cell surface receptors [16]. Accessible α-amine groups present on the N-termini of peptides and proteins react with NHS-esters. However, α-amines are seldom available on a protein, so the reaction with side chains of amino acids becomes important. While five amino acids contain amine groups in their side chains, only the ε-amine of lysine reacts significantly with NHS-ester. A phosphate-buffered system was chosen for our coupling buffer since any amine groups present in the buffer would quench the reaction. DSS was dissolved in DMSO at a concentration of 13 mg/mL and used at a final concentration of 2.6 mg/mL. Because it contains a primary amine, TBS was used as a blocking reagent and a washing buffer after the cross-linking reaction was completed. Glycine at pH 2.8 was used to remove any free antibody after the cross-linking step and to elute the antigen. The average cross-linking efficiency was over 80% (data is not shown).

Co-IP of proteins from cellular extracts is the most convincing evidence that two or more proteins physically interact with each other. The 26S proteasome is a key enzyme in the ubiquitin/ATP-dependent pathway of protein degradation [17, 18]. The catalytic core of this unusually large complex ($M_r \sim 700,000$) is formed by the 20S proteasome, a barrel-shaped structure comprised of four rings each containing seven subunits, $\alpha_7\beta_7\beta_7\alpha_7$ [19, 20]. The fourteen different subunits of mammalian 20S proteasome have molecular weights ranging from 18 to 33 kd [21]. The α6 subunit is located on the outer rings of the 20S proteasome [19]. The mouse monoclonal antibody to human 20S proteasome subunit α6 recognizes a 33- kd band on a Western blot of total HeLa cell lysate (Figure 1). The antisubunit α6 antibody was cross-linked to protein G agarose and used to coimmunoprecipitate the whole 20S proteasome complex from HeLa cell lysate. The eluted protein complex contains a series of proteins ranging from 18 to 33 kd, which is the typical pattern of 20S proteasome subunits (Figure 1a). The higher molecular weight proteins ranging from 45 to 100 kd (Figure 1a, lane 2) are the regulatory subunits of the 26S proteasome complex which were coimmunoprecipitated with the catalytic core complex [22]. As a comparison, a co-IP with the traditional method (non-cross-linked antibody) was performed using the same conditions. Although only one sixth of the total eluent was analyzed on SDS-PAGE, a strong contamination of antibody heavy and light chains was observed (Figure 1b). This result demonstrates that the cross-linked antibody can efficiently coimmunoprecipitate a large protein complex and eliminate the antibody contamination. Therefore, this method could be scaled up to affinity purify protein complexes for downstream assays and protein characterization studies.

In vitro binding and co-IP assays are very useful when studying the interactions of proteins that become complex only at a certain point in the cell cycle or of a subset of proteins belonging to a larger protein complex [22, 23]. Co-IP can also be used to confirm protein-protein interaction results from an in vivo yeast two-hybrid screen [8]. Max and c-Myc are a pair of interacting proteins that form heterodimers to regulate the transcription of genes which have been shown to contribute to carcinogenesis [24]. In this experiment, ^{35}S-labeled Max (T7-tagged) and c-Myc were translated in vitro in the presence of L-[^{35}S] methionine using a rabbit reticulocyte lysate. The two separately synthesized ^{35}S-labeled proteins were incubated at 30°C for 30 minutes. The anti-T7 tag antibody was cross-linked to protein G agarose and used to coimmunoprecipitate T7-tagged Max and c-Myc. Luciferase was used as a negative control for T7-tagged Max. Figure 2 shows that c-Myc can be coimmunoprecipitated with Max, whereas luciferase does not coimmunoprecipitate with Max. The entire co-IP experiment was performed in a single spin column, which limited contamination of radioactive

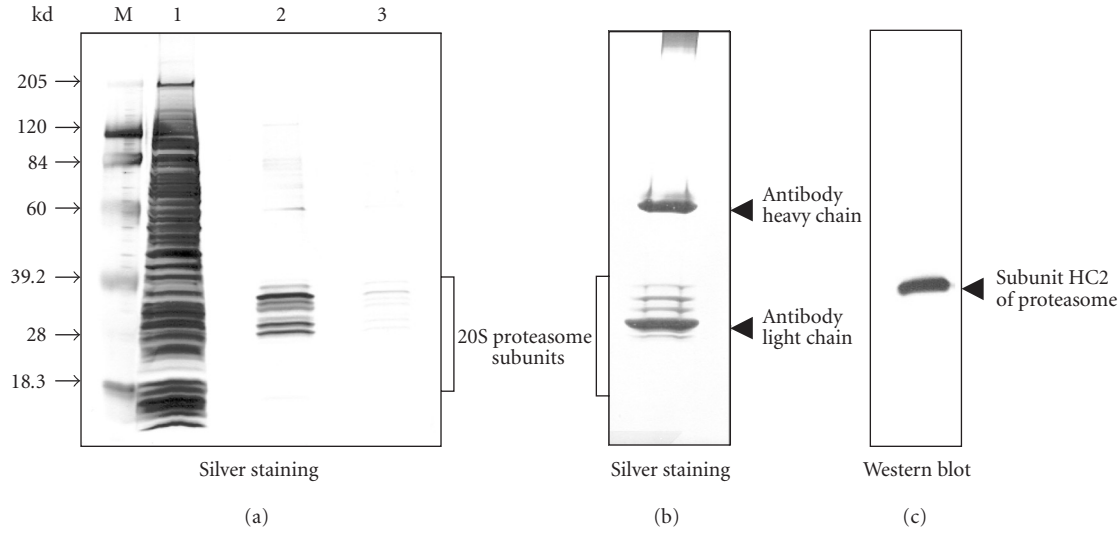


FIGURE 1. Comparison of co-IP using antibody cross-linked protein G agarose and non-cross-linked antibody with protein G agarose. Mouse monoclonal antiproteasome subunit α6 antibody (100-μL) and 100-μL protein G resin were used for both co-IPs. HeLa cell ($\sim 5 \times 10^7$ cells) lysate was precleared with protein G agarose and co-IP was performed at 4°C overnight. (a) Co-IP of 20S proteasome complex using antisubunit α6 antibody cross-linked protein G agarose. The eluted proteasome complex was concentrated and separated on 12% SDS-PAGE and silver stained. Lane M, BlueRanger prestained protein molecular weight marker mix; lane 1, crude HeLa cell lysate; lane 2, elutions 1 and 2; lane 3, elutions 3 and 4. (b) Co-IP of 20S proteasome complex using traditional antisubunit α6 antibody with protein G agarose. One sixth of total eluted proteasome-antibody complex was separated on 12% SDS-PAGE and silver stained. (c) Immunoblot detection of 20S proteasome subunit α6 in crude HeLa cell lysate using anti-α6 antibody.

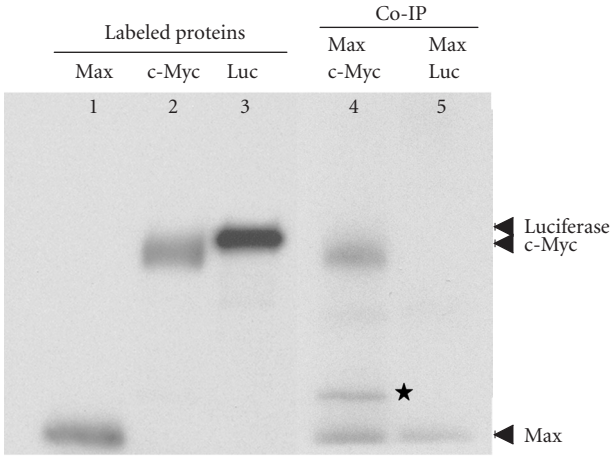


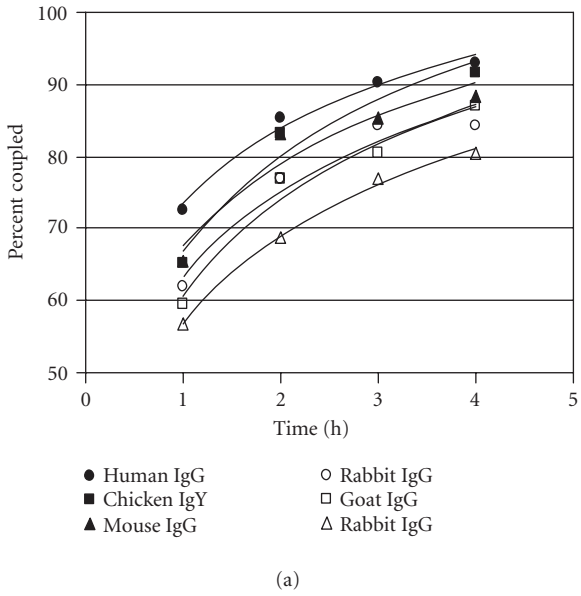
FIGURE 2. Coimmunoprecipitation of T7-tagged Max and c-Myc using anti-T7 tag antibody cross-linked protein G agarose. Max (T7-tagged), c-Myc, and luciferase were in vitro translated and ^{35}S-labeled (lanes 1, 2, and 3) using the TNT-Coupled Reticulocyte Lysate System. Before co-IP, Max and c-Myc, and Max and luciferase were mixed proportionally and incubated at 30°C for 1 hour. Co-IPs were carried out with anti-T7 tag antibody cross-linked protein G agarose at 4°C for 2 hours (lanes 4 and 5) and the eluted protein complexes were separated on 4–20% SDS-PAGE. The ^{35}S-labeled proteins were detected by autoradiography. Star (∗): degradation product of c-Myc.

material, and prevented the loss of resin between the washes, thereby enhancing the recovery [25]. As little as 25 μL of antibody-protein G agarose per sample was used up to five times without detectable loss of the activity. This advantage could be very useful for confirming in vivo yeast two-hybrid screening results because the protein G agarose cross-linked to an antibody against the "bait" protein can be used repeatedly to confirm the interaction between the "bait" protein and each "prey" protein. We also have used protein G agarose cross-linked with the c-Myc tag antibody to successfully coimmunoprecipitate SV40 large-T antigen with c-Myc-tagged p53 (data is not shown).

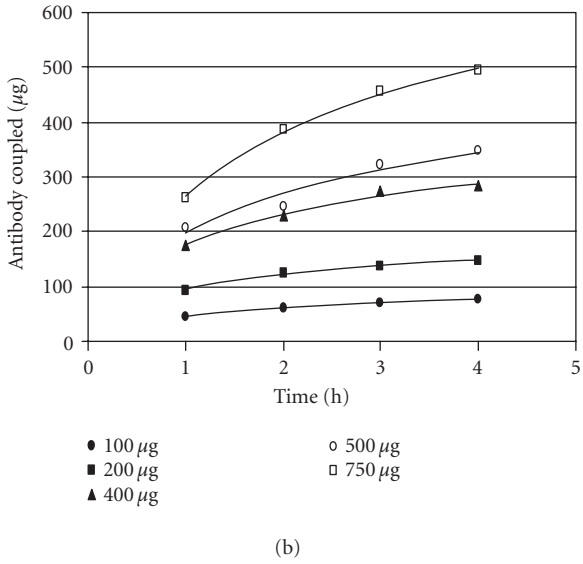
Although cross-linking antibody to protein G or protein A agarose is a good approach to immobilize the antibodies in the correct orientation, the cross-linking efficiency and specificity varied depending upon the concentrations of antibody, protein G agarose, and cross-linking reagent. Binding capacity after cross-linking also varied from antibody to antibody. The drop in binding capacity could be caused by a change in conformation when cross-linking occurs on the surface of the antibody molecule or when cross-linking occurs within the binding sites (see data below). Therefore, optimal DSS and antibody concentrations need to be determined empirically for each antibody. This led us to expand our investigation for a more universal method of antibody attachment with improved preservation of antibody binding activity.

Antibody-coupled method and co-IP applications

The antibody-coupled procedure utilizes reductive amination to directly link the antibody to the agarose



(a)



(b)

Figure 3. Coupling of mammalian and avian antibodies. (a) Antibody from various species (200 μg) was coupled to 200 μL of coupling gel (settled gel) at 1-hour intervals for 4 hours at RT. For the chicken antibody, 500 μL was used. (b) Normal chicken IgY antibody (100–750μg) was coupled to 200 μL of coupling gel (settled gel) at 1-hour intervals for 4 hours at RT.

bead. The coupling resin is provided in an activated state containing aldehyde groups formed by mild oxidation of adjacent diols using sodium meta-periodate [14]. Primary and secondary amine groups on the antibody react with the aldehydes to form Schiff bases that are then reduced by sodium cyanoborohydride to form secondary and tertiary amine linkages [14]. Since this procedure links the antibody to the resin in every direction, not all antibody molecules will present an active orientation. Typical coupling efficiencies for various species of antibodies are shown in Figure 3a. The coupling efficiencies were determined by spectrophotometric analysis of antibody solutions before and after coupling. On average, 88 percent of the antibody was coupled in 4 hours when using 200 μg of antibody and 200 μL of settled resin. Scalability was demonstrated when immobilization using 50 and 100-μL of settled coupling resin yielded comparable coupling efficiencies (data is not shown). All species of antibody exhibited the same relationship with respect to time and coupling. Therefore, this method is not limited to antibody species that only bind strongly to protein G or protein A. Figure 3b illustrates the relationship between protein concentration and the rate of the coupling reaction. As expected, the rate of coupling increases with increased protein concentration. These results show that this is a universal technique with no need to optimize for each antibody.

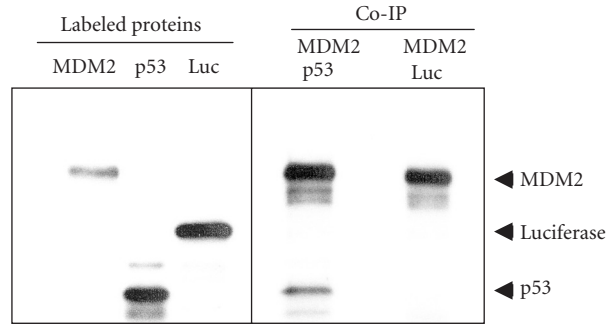


Figure 4. Coimmunoprecipitation of p53 and MDM2 using anti-MDM2 antibody-coupled agarose. MDM2, p53, and luciferase were in vitro translated and ^{35}S-labeled using TNT-Coupled Reticulocyte Lysate System. p53 and MDM2 were combined and incubated at 30°C for 30 minutes. Co-IP was performed at 4°C for 2 hours with 60 μL anti-MDM2 antibody-coupled agarose. Luciferase was used as a negative control protein to incubate with MDM2. Eluted proteins were resolved on 4–12% SDS-PAGE and visualized by autoradiography.

MDM2 oncoprotein plays a central role in the regulation of p53 tumor suppressor protein [26, 27, 28]. MDM2 binds to p53 and blocks its activity as a tumor suppressor and promotes its degradation in many tumor cells [6, 29]. In our experiment, human p53 and MDM2 genes were transcribed/translated and ^{35}S-labeled in a reticulocyte lysate. The co-IP result using coupled anti-MDM2 antibody shows that MDM2 interacts with p53 but not luciferase (Figure 4). We have used as little as 20 μL of the antibody-coupled resin and have reused the resin up to five times without obvious loss of activity.

Binding capacity of antibody-coupled resin or antibody cross-linked resin

The antibody cross-linked and antibody-coupled procedures eliminate the contamination problem by preventing the antibody from co-eluting with the antigen.

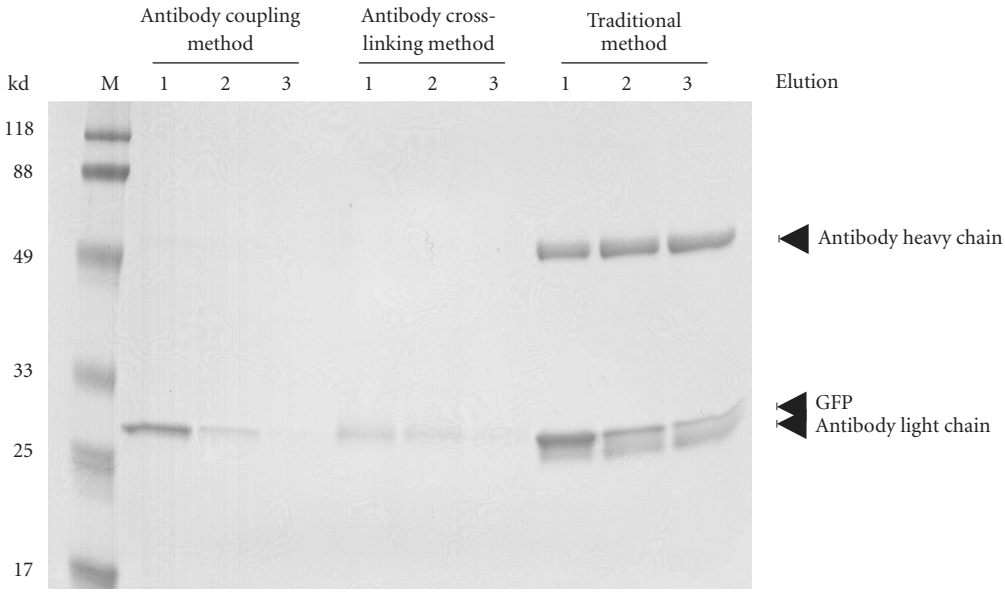


Figure 5. Comparison of IP using antibody cross-linked protein G agarose, antibody-coupled agarose, and non-cross-linked antibody with protein G agarose. In each case, 130 μg of affinity-purified goat anti-GFP antibody was used with 100-μL of settled protein G gel or coupling gel. IP was performed using 135 μg of partially purified GFP. Ten percent of the elution volume was electrophoresed on a 12% polyacrylamide reducing gel and stained with Coomassie. Lane M, BlueRanger molecular weight marker; lanes 1–3 elutions from different antibody immobilization methods.

Another benefit to both methods is that the antibody-resin is reusable, thereby conserving valuable antibody. We compared the traditional, antibody cross-linked and antibody-coupled procedures to evaluate the relative amount of antigen recovered using the same amount of goat anti-GFP antibody (Figure 5). Under these conditions, a significant increase in recovered antigen is seen when using the antibody-coupled procedure versus the antibody cross-linked method. Although the traditional method yielded a greater quantity of antigen, the presence of antibody light chains in the eluent distorted the recovered GFP band, since they have comparable molecular weights.

CONCLUSIONS

We have developed two methods to immunoprecipitate and coimmunoprecipitate proteins that eliminate antibody contamination. The first method properly orients the antibody for antigen recognition by binding its Fc portion to protein A or protein G resin. Subsequently, the antibody is cross-linked to the resin to prevent leaching into the eluent. The second method achieves the same goal yet is universal for all antibody classes and species because it couples the antibody directly to the resin matrix. Both methods give comparable results to traditional IP although overall antigen-binding efficiency is not always as high as with traditional IP. The precipitated proteins from our IP and co-IP methods can be easily detected and characterized in downstream applications such as mass spectrometry or enzymatic assays. Furthermore, the stabilized antibody-linked resin from either technique can be regenerated and reused multiple times thereby conserving precious antibody samples.

REFERENCES

[1] Williams NE. Immunoprecipitation procedures. *Methods Cell Biol*. 2000;62:449–453.
[2] Harlow Ed, Lane D. Using Antibodies: A Laboratory Manual. Cold Spring Harbor, NY: Cold Spring Harbor Laboratory Press; 1999.
[3] Otto JJ, Lee SW. Immunoprecipitation methods. *Methods Cell Biol*. 1993;37:119–127.
[4] Ransone LJ. Detection of protein-protein interactions by coimmunoprecipitation and dimerization. *Methods Enzymol*. 1995;254:491–497.
[5] Shi Y, Ullrich SJ, Zhang J, et al. A novel cytokine receptor-ligand pair. Identification, molecular characterization, and in vivo immunomodulatory activity. *J Biol Chem*. 2000;275(25):19167–19176.
[6] Honda R, Tanaka H, Yasuda H. Oncoprotein MDM2 is a ubiquitin ligase E3 for tumor suppressor p53. *FEBS Lett*. 1997;420(1):25–27.
[7] Yamauchi J, Kaziro Y, Itoh H. C-terminal mutation of G protein beta subunit affects differentially extracellular signal-regulated kinase and c-Jun N-terminal kinase pathways in human embryonal kidney 293 cells. *J Biol Chem*. 1997;272(12):7602–7607.

[8] Tanimura S, Ohtsuka S, Mitsui K, Shirouzu K, Yoshimura A, Ohtsubo M. MDM2 interacts with MDMX through their RING finger domains. *FEBS Lett.*1999;447(1):5–9.

[9] Wong C, Naumovski L. Method to screen for relevant yeast two-hybrid-derived clones by coimmunoprecipitation and colocalization of epitope-tagged fragments–application to Bcl-xL. *Anal Biochem*. 1997;252(1):33–39.

[10] Estojak J, Brent R, Golemis EA. Correlation of two-hybrid affinity data with in vitro measurements. *Mol Cell Biol*. 1995;15(10):5820–5829.

[11] Wiese C, Galande S. Elimination of reducing agent facilitates quantitative detection of p53 antigen. *Biotechniques*. 2001;30(5):960–963.

[12] Berryman M, Bretscher A. Immunoblot detection of antigens in immunoprecipitates. *Biotechniques*. 2001;31(4):744–746.

[13] Scopes PK, Smith JA. Analysis of protein. In: Ausubal FM, Brent R, Kingston RE, et al, Eds. *Current Protocols in Molecular Biology*. New York, NY: John Wiley & Sons;1994:10.6.3.

[14] Hermanson GT, Mallia AK, Smith PK. *Immobilized Affinity Ligand Techniques*. San Diego, Calif: Academic Press; 1992.

[15] Hermanson GT. *Bioconjugate Techniques*. San Diego, Calif: Academic Press; 1996.

[16] Cox GW, Mathieson BJ, Giardina SL, Varesio L. Characterization of IL-2 receptor expression and function on murine macrophages. *J Immunol*. 1990;145(6):1719–1726.

[17] Rechsteiner M, Hoffman L, Dubiel W, The multicatalytic and 26S proteases. *J Biol Chem*. 1993;268(9):6065–6068.

[18] Eytan E, Armon T, Heller H, Beck S, Hershko A. Ubiquitin C-terminal hydrolase activity associated with the 26S proteasome complex. *J Biol Chem*. 1993;268(7):4668–4674.

[19] Groll M, Ditzel L, Lowe J, et al. Structure of 20S proteasome from yeast at 2.4 A resolution. *Nature*. 1997;386(6624):463–471.

[20] Tanahashi N, Tsurumi C, Tamura T, Tanaka K. Molecular structure of 20S and 26S proteasomes. *Enzyme Protein*. 1993;47(4–6):241–251.

[21] Tanaka K, Tsurumi C. The 26S proteasome: subunits and functions. *Mol Biol Rep*. 1997;24(1-2):3–11.

[22] Chu-Ping M, Vu JH, Proske RJ, Slaughter CA, DeMartino GN. Identification, purification, and characterization of a high molecular weight, ATP-dependent activator (PA700) of the 20 S proteasome. *J Biol Chem*. 1994;269(5):3539–3547.

[23] DaFonseca CJ, Shu F, Zhang JJ. Identification of two residues in MCM5 critical for the assembly of MCM complexes and Stat1-mediated transcription activation in response to IFN-gamma. *Proc Natl Acad Sci USA*. 2001;98(6):3034–3039.

[24] Littlewood TD, Amati B, Land H, Evan GI. Max and c-Myc/Max DNA-binding activities in cell extracts. *Oncogene*. 1992;7(9):1783–1792.

[25] Brymora A, Cousin MA, Roufogalis BD, Robinson PJ. Enhanced protein recovery and reproducibility from pull-down assays and immunoprecipitations using spin columns. *Anal Biochem*. 2001;295(1):119–122.

[26] Juven-Gershon T, Oren M. Mdm2: the ups and downs. *Mol Med*. 1999;5(2):71–83.

[27] Freeman DA, Wu L, Levine AJ. Functions of the MDM2 oncoprotein. *Cell Mol Life Sci*. 1999;55(1):96–107.

[28] Haines DS. The mdm2 proto-oncogene. *Leuk Lymphoma*. 1997;26(3-4):227–238.

[29] Haines DS, Landers JE, Engle LJ, George DL. Physical and functional interaction between wild-type p53 and mdm2 proteins. *Mol Cell Biol*. 1994;14(2):1171–1178.

* Corresponding author.
E-mail: walid.qoronfleh@perbio.com
Fax: +1 414 227 3759; Tel: +1 414 227 3605

Journal of Biomedicine and Biotechnology • 2003:5 (2003) 299–307 • PII. S1110724303209268 • http://jbb.hindawi.com

RESEARCH ARTICLE

Optimization of Rolling-Circle Amplified Protein Microarrays for Multiplexed Protein Profiling

Weiping Shao,[1]* Zhimin Zhou,[1] Isabelle Laroche,[1] Hong Lu,[1] Qiuling Zong,[1] Dhavalkumar D. Patel,[2] Stephen Kingsmore,[1] and Steven P. Piccoli[1]

[1]*Molecular Staging, Inc, Suite 701, 300 George Street, New Haven, CT 06511, USA*
[2]*Thurston Arthritis Research Center, Department of Medicine, University of North Carolina at Chapel Hill, CB# 7280, 3330 Thurston Building, Chapel Hill, NC 27599-7280, USA*

Received 13 August 2002; accepted 10 December 2002

Protein microarray-based approaches are increasingly being used in research and clinical applications to either profile the expression of proteins or screen molecular interactions. The development of high-throughput, sensitive, convenient, and cost-effective formats for detecting proteins is a necessity for the effective advancement of understanding disease processes. In this paper, we describe the generation of highly multiplexed, antibody-based, specific, and sensitive protein microarrays coupled with rolling-circle signal amplification (RCA) technology. A total of 150 cytokines were simultaneously detected in an RCA sandwich immunoassay format. Greater than half of these proteins have detection sensitivities in the pg/mL range. The validation of antibody microarray with human serum indicated that RCA-based protein microarrays are a powerful tool for high-throughput analysis of protein expression and molecular diagnostics.

INTRODUCTION

Despite great advancements in genomics research, DNA microarrays are still of limited applicability as diagnostic tests, in part because proteins, not genes, are the ultimate effectors of a biological process. Proteins are often expressed at concentrations and varied structural forms having modifications that cannot be predicted from mRNA analysis. There is a great deal of interest in proteomics-based approaches that enable profiling the abundance of proteins, or investigation of protein-protein and protein-drug interactions. The concept of a protein microarray or biochip is attractive for rapid, high-throughput proteomics studies that utilize available sample volumes [1, 2, 3, 4]. Proteins that are present in abnormal concentrations (increased or decreased) in a disease state could be identified and validated as disease markers. Proteomics pattern analysis based on a multiplexed list of biomarkers may well make possible the diagnosis of certain diseases that do not have either effective screening options or biomarkers already detected by conventional immunoassays [5]. Identification of such disease markers will provide valuable information for detection, classification, and prognosis of diseases, as well as targets for treatment outcomes. High-throughput protein chips require arraying a wide range of probes that specifically recognize a single protein in complex mixtures such as serum, plasma, and other biological specimens. These probes are most frequently antibodies or antibody mimics [3, 4, 6, 7].

Given the immense promise of proteomics technologies, there are still limitations that need to be overcome, such as lack of sensitivity and enhanced information about the target. Because there is no PCR equivalent available to amplify proteins, the identification of low-abundance proteins may often be difficult or impossible, and important biomarker information could thus be lost. Rolling-circle amplification (RCA) is a unique signal amplification technology [8] that permits detection of multiple proteins with a broad dynamic range on protein microarrays [4, 9]. In this paper, we describe the development of RCA-based highly multiplexed and sensitive protein microarray immunoassays for detecting 150 proteins in an ELISA format. Application to human serum and plasma samples has established that the rolling-circle amplified protein microarray is a powerful tool for the profiling of expressed proteins in biological specimens.

MATERIAL AND METHODS

Preparation of microarrays. Glass slides were prepared and silanized according to procedures previously described [10]. Antibodies (R&D Systems, Minneapolis, Minn; Pharmingen, San Diego, Calif; BioSource International, Camarillo, Calif) were diluted to 0.5 mg/mL in PBS containing 0.05 mg/mL BSA, and 375 pL of each was spotted in quadruplicate onto the slides using a BioChip Arrayer (Packard BioSciences, Downers Grove, Ill). Each spot has a diameter of approximately 180 μm with a

center-to-center spacing of 270 μm. Each slide was divided by a teflon mask into 16 subarrays, each with a diameter of 0.65 cm. Within each subarray, 256 spots were printed at known locations, containing 40 different features (antibodies for specific antigens) in quadruplicate (160 spots). The remaining features comprised of controls for antibody immobilization, Cy5-labeled mouse IgG and internal calibrators (biotin mouse IgG).

RCA microarray immunoassays. Slides were blocked by adding 30 μL blocking buffer (0.5% nonfat dry milk, 0.2% BSA, and 0.05% Tween-20 in PBS) to each subarray and incubating for 1 hour at 37°C in a humidified chamber, and then washed by immersion in PBS containing 0.5% Brij 35 for 2 minutes. A 15 μL aliquot of sample (single analyte or mixture of multiple analytes; serum, plasma, or culture supernatant) was immediately added to each subarray, incubated for 30 minutes at 37°C, and then washed as described above. After analytes were captured, a 20 μL aliquot of the appropriate mixture of biotinylated detection antibodies at 0.1 μg/mL each was applied to individual arrays and incubated at 37°C for 30 minutes in a humidified chamber. Slides were washed twice for two minutes in PBS containing 0.5% Brij 35. A mouse monoclonal antibiotin antibody-primer conjugate was prepared by derivatization of a mouse monoclonal antibiotin IgG (Jackson ImmunoResearch Laboratories, Inc, West Grove, Pa) with a 5'-terminal amine-modified oligonucleotide primer, 5'-Thiol-AAA AAA AAA AAA AAA CAC AGC TGA GGA TAG GAC AT-3', as previously described [9]. The conjugate was annealed with 75 nM circle 1 (cyclic 5'-CTC AGC TGT GTA ACA ACA TGA AGA TTG TAG GTC AGA ACT CAC CTG TTA GAA ACT GTG AAG ATC GCT TAT TAT GTC CTA TC-3') in PBS containing 0.05% Tween-20 and 1 mM EDTA at 37°C for 30 minutes. Twenty microliters of the conjugate-circle mixture was applied to each array. Slides were incubated at 37°C for 30 minutes and then washed twice. RCA reaction mixture (20 μL) containing native T7 DNA polymerase (20 U/mL), 1 mM dNTPs, ssDNA-binding protein 22 μg/mL, 1x sequenase buffer, 8% DMSO, and 0.05 mM Cy5 decorator (5'-Cy5-TGT CCT ATC CTC AGC TGG-Cy5) was added to each subarray. Slides were incubated at 37°C for 45 minutes, then washed twice in PBS containing 0.5% Brij 35 and once in PBS, and spin-dried. Arrays were analyzed on an Axon *GenePix 4000B* scanner (Axon Instruments Inc, Foster City, Calif), and fluorescence quantitated by using GenePix Pro 3.0 quantitation software. Data was analyzed by using MSI Analyzer software (Molecular Staging, Inc).

RESULTS AND DISCUSSION

Fabrication of antibody microarray

Microarrays were printed on cyanosilane-coated glass slides. This surface was chosen based on experiments indicating that cyanosilane-activated slides have the advantages of (1) simpler preparation, (2) less-expensive reagents, and (3) more uniformity of bound antibody spot morphology over thiolsilane-coated and cross-linker (such as GMBS) activated glass slides (data not shown). The concentration of capture antibody for printing was optimized by comparing specific signal intensities from the same capture antibody printed on the chip at different concentrations chosen for microarray printing. Specific signals at 0.5 mg/mL reached the point of saturation (data not shown), indicating that antibodies saturated the area available for immobilization in the spot. The uniformity of antibody immobilization on the chip was assessed by measuring the coefficient of variation (CV) from the quadruplicate spots using Cy5-labeled anti-isotype antibodies (against the species used in the capture antibodies), previously reported as approximately 10% for each feature on the chip[4].

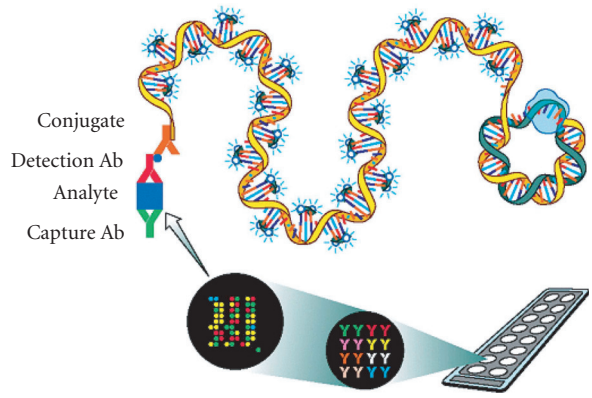


FIGURE 1. Schematic representation of microarray immunoassay with RCA signal amplification. It consists of (a) analyte capture by antibody immobilized on microarray, (b) detection by biotinylated secondary antibody, (c) binding of antibiotin antibody-oligonucleotide conjugate pre-annealed with circle 1, and (d) rolling-circle replication to generate long single-strand DNA which is hybridized with oligonucleotide decorator.

Improving the sensitivity of protein detection on antibody array

Rolling circle amplification enables amplification of signals tethered to proteins and nucleic acids [4, 9, 11]. Immunoassays incorporating RCA have shown a large increase in sensitivity when compared with the direct assay [4]. The technique consists of protein capture by antibody immobilized on microarrays, detection by biotin-labeled second antibody specific for the captured protein, binding of a universal antibiotin antibody oligonucleotide conjugate pre-annealed with circle, and RCA signal amplification where a long single-strand DNA is generated by rolling-circle replication in the presence of DNA polymerase and nucleotides, with hybridization of the RCA product to fluorescent-labeled complementary oligonucleotide probes occurring simultaneously (Figure 1). Besides the signal amplification by RCA, the sensitivity of the antibody array was further improved by two

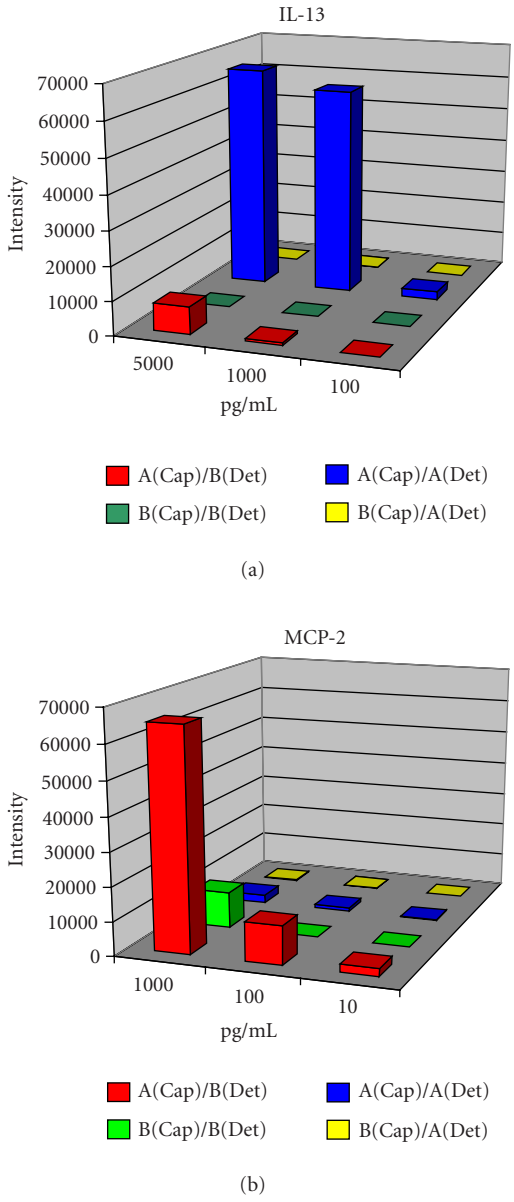


Figure 2. Optimization of antibodies from different sources (A and B) for choosing optimal pair for immunoassay on protein microarray with examples of (a) IL-13 and (b) MCP-2.

approaches: (1) choosing optimal capture and detection antibody pairs for the microarray immunoassay and (2) optimizing assay procedures.

Antibody microarray development commenced with selection of optimal antibody pairs. For individual features, if the matched antibody pair developed for conventional ELISA was available, it was preferentially selected for the chip development. However, in terms of sensitivity and specificity, antibody pairs do not always function equivalently on plastic ELISA plates and the glass-based microarrays. Antibody pairs from different manufacturers for the same antigen with similar sensitivity as ELISA assays perform quite differently in microarrays.

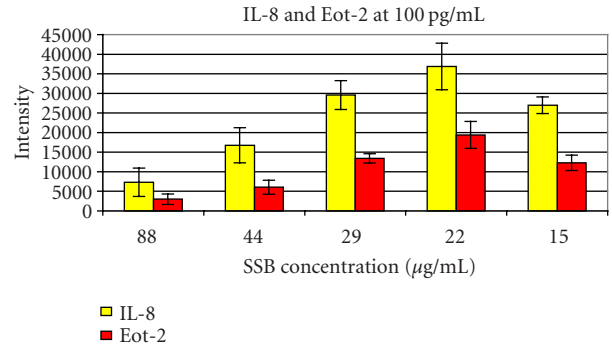


Figure 3. Titration of SSB in the RCA reaction mixtures. Cytokine IL-8 and Eot-2 at 100 pg/mL were detected.

Thus it has become necessary to select optimal pairs for antibody microarrays by comparing all available antibodies and pairs. As an example, Figure 2a shows the difference between two matched pairs of antibodies (for ELISA) from different sources with specificity for detecting the cytokine IL-13 using the antibody microarray. The pair from source A showed strong signal intensity with a sensitivity of less than 100 pg/mL. However, the pair from source B did not show signal even at concentrations as high as 5 ng/mL IL-13. Selecting the antibody pair from source A provided at least 50-fold increase of sensitivity when compared with the pair from source B. Alternatively, the ELISA antibody pair from the same set is not always the optimal for a microarray immunoassay. One example shown in Figure 2b is for the chemokine MCP-2. Neither of the ELISA-matched pairs showed sensitivity better than 100 pg/mL on the microarray slides. However, using a capture antibody from A and a detection antibody from B, the sensitivity for MCP-2 was improved to 10 pg/mL or better, a 10-fold increase in sensitivity relative to a single manufacturer's reagents. For features without matched pairs, monoclonal antibodies with different binding epitopes or polyclonal antibodies were extensively evaluated for sensitivity ranges permitting detection of biologically relevant changes in protein expression. Cross-reactivity and nonspecific signals were also extensively evaluated as described below.

The second approach of improving sensitivity was to optimize the assay procedures. This involved optimizing detection antibody concentrations, conjugate and circle concentrations, incubation temperatures and time, and RCA reaction solution composition. Analogous to sandwich ELISA assays, the conditions giving the best signal to noise ratio were found to be 0.1 μg/mL of detector with assay incubation time of 30 minutes at 37°C. Titration of conjugate and circle pre-annealed to conjugate revealed the optimal concentration to be 2 μg/mL conjugate with 75 nM circle. Other components, such as the concentration of single-strand DNA binding protein (SSB), were also optimized for maximum sensitivity in RCA reaction mixtures (Figure 3).

Table 1. Performance characteristics of the 150-feature protein microarray.

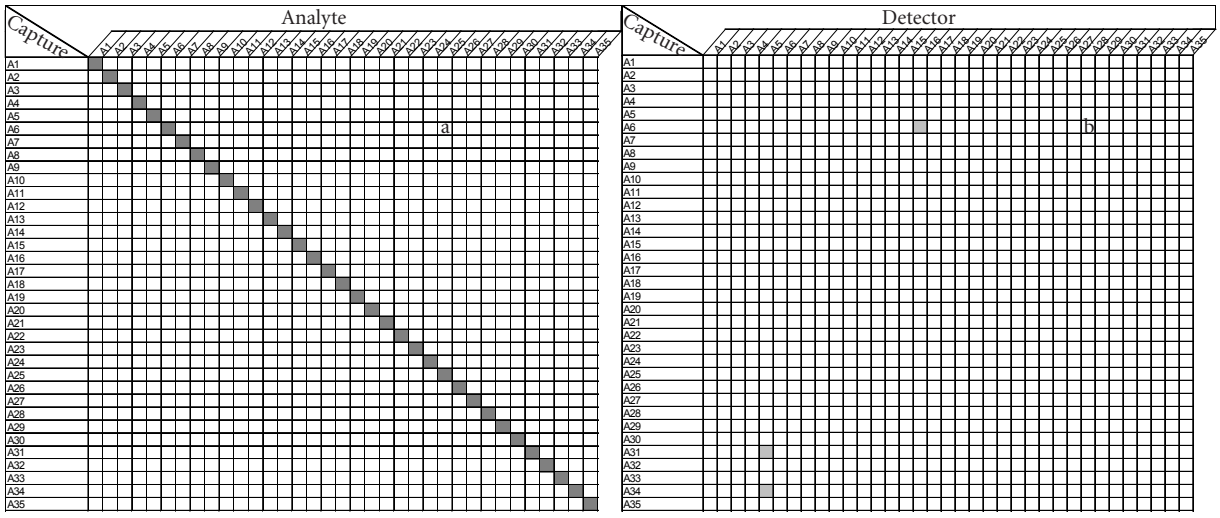
Analyte	Biological ED_{50} (pg/mL)	Mean serum levels (pg/mL)	Array sensitivity (pg/mL)	ELISA sensitivity (pg/mL)
ALCAM	na	52 000	100	24
ANG	na	360 000	10	6
AR	5000–15 000	45.4	100	na
BDNF	3000–10 000	27 793	10	20
BLC	5000–20 000	na	10	na
SVE-cadherin	na	2800	1000	313
CCL28	400 000–2 000 000	43.8	30	2.58
SCD23	na	1.3 IU/mL	10 000	0.15 U/mL
CD27	na	na	100	na
CD30	30 000–100 000	6.4 U/mL	10	na
CD40	na	102.1	300	15.6
CNTF	50 000–150 000	nd	300	8
CNTF Rα	200 000–400 000	na	100	na
CT-1	1000–4000	619.5	1000	na
CTACK	100 000–400 000	522	10	1.55
CTLA-4	2 000 000–4 000 000	nd	1000	313
DR6	na	na	10	na
EGF	100–400	336	10	0.7
ENA-78	3000–15 000	1449	100	15
Eot	10 000–20 000	68.6	10	5
Eot-2	10 000–50 000	249	10	1.83
Eot-3	250 000–1 000 000	9.7	100	2
Fas	10 000–40 000	9406	100	20
Fas ligand	200 000–500 000	nd	100	100
FGF acidic	100 000–300 000	na	10	na
FGF basic	100 000–250 000	nd	100	3
FGF-4	50–150	93.8	300	30
FGF-6	100–300	na	100	na
FGF-7	15 000–25 000	< 31.2	10	15
FGF-9	1000–2000	na	100	na
Flt-3 L	500–1000	93.9	10	7
Follistatin	100 000–400 000	2483	30	29
G-CSF	20–60	22	1000	20
GCP-2	na	156	100	1.6
GDNF	1000–3000	na	10	na
GM-CSF	20–80	1.72	10	3
Sgp130	3000–9000	306 000	100	80
GRO-α	1000–4000	93	10	10
GRO-β	3000–300 000	na	10	na
GRO-γ	10 000–100 000	na	10	na
HB-EGF	2000–5000	na	100	na
HCC-1	200 000–8 000 000	na	30	na
HCC4	150 000–750 000	na	100	na
HGF	20 000–40 000	721	100	40
HVEM	500 000–2 000 000	na	30	na
ICAM-1	na	211 000	10	350
ICAM-3	na	50 000	100	580
IFN-α	3.8×10^8 U/mg	< 10	100	15.6
IFN-γ	800–1500	< 15.6	10	8
IFN-ω	na	nd	100	9.375

Table 1. Continued.

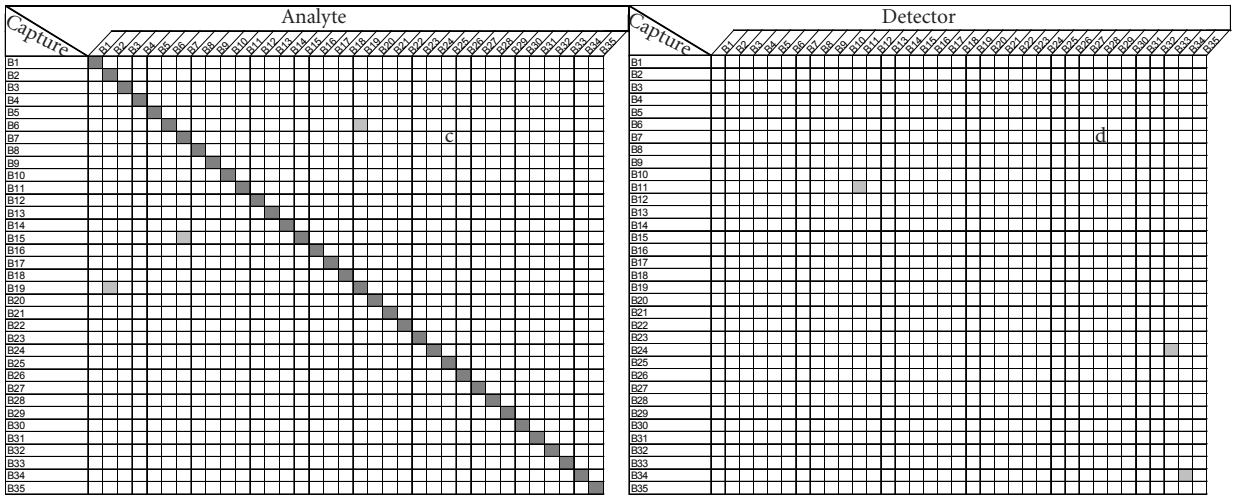
IGF-I	1000–3000	105 000	100	26
IGF-IR	na	na	100	na
IGF-II	5000–10 000	29 300	1000	na
IGFBP-1	1 000 000–4000 000	14 300	30	na
IGFBP-2	30 000–90 000	434 200	100	na
IGFBP-3	50 000–150 000	2 375 000	1000	50
IGFBP-4	30 000–90 000	na	100	na
I-309	3000–9000	nd	10	0.71
IL-1α	3.0–7.0	< 3.9	10	1
IL-1β	13–20	0.536	10	1
IL-1rα	20 000–60 000	418	100	22
IL-1 sRI	500 000–1000 000	na	100	na
IL-1 sRII	1 000 000–5000 000	11 000	10	10
IL-2	250–500	1.6	10	7
IL-2 Rβ	1 000 000–3000 000	na	300	na
IL-2 sRa	500 000–1 000 000	1346	10	10
IL-2 Rγ	3 000 000–6 000 000	na	30	na
IL-3	100–400	< 31.2	1000	7.4
IL-4	50–200	< 0.25	10	10
IL-5	100–200	< 7.8	10	3
IL-5 Rα	200 000–300 000	na	100	na
IL-6	200–800	1.62	10	0.7
IL-6 sR	5000–15 000	31 000	10	7
IL-7	200–500	2.82	10	0.1
IL-8	100–500	13.2	1	10
IL-9	500–1000	na	1000	na
IL-10	500–1000	2	10	3.9
IL-10 Rβ	na	na	10	na
IL-11	60–240	< 31.2	100	8
IL-12 (p40)	50–200	77	10	15
IL-12 (p70)	50–200	1.93	10	0.5
IL-13	3000–6000	< 62.5	10	32
IL-15	500–2000	2.14	100	2
IL-16	2 000 000–10000 000	171	100	6.2
IL-17	2000–6000	< 31.2	10	15
IL-18	na	126	100	12.5
IL-21	10 000–40 000	na	300	na
IP-10	20 000–60 000	89	10	1.7
I-TAC	1000–5000	177	10	13.9
Leptin/OB	400–2000	4760(M), 20676(F)	3000	7.8
LIF	500	na	100	8
SLIF-Rα	3 000 000–6000 000	4300	300	156
Lymphotactin	50 000–200 000	na	100	na
M-CSF	500–1500	670	10	9
M-CSF R	4000–12 000	na	300	na
MCP-1	5000–20 000	370	10	5
MCP-2	30 000–120 000	200	10	na
MCP-3	20 000–80 000	< 15.6	10	2
MCP-4	200 000–400 000	na	300	na
MDC	3000–9000	1089	10	62.5
MIF	50 000–100 000	1337	100	na

Table 1. Continued.

MIG	200 000–600 000	914	10	na
MIP-1α	2000–10 000	< 46.9	10	10
MIP-1β	10 000–30 000	80	10	11
MIP-1δ	2000–4000	na	100	na
MIP-3α	500–2000	26.3	30	0.47
MIP-3β	100 000–300 000	na	10	na
MMP-1	na	6490	300	52
MMP-2	na	1 169 000	300	na
MMP-7	na	2290	100	16
MMP-9	na	436 000	300	156
MMP-10	na	770	100	4.13
MPIF	200 000–500 000	536	100	3.3
MSP	10 000–30 000	4 nM	100	na
NAP-2	100 000–300 000	na	100	na
β-NGF	800–1500	265	10	na
NT-3	10 000–30 000	13.4	100	na
NT-4	5000–15 000	na	10	na
OSM	150–300	< 15.6	10	na
PARC	500 000–2000 000	na	100	na
PDGF-Rα	5 000 000–10 000 000	na	3000	na
PECAM-1	na	20 000	1000	na
PF4	5 000 000–15 000 000	14 700	300	na
PlGF	na	18	10	7
Prolactin	30–100	9300	1000	7.6 mIU/L
RANK	4000–10 000	na	300	na
RANTES	1000–5000	49 137	10	8
SCF	2 500–5000	984	10	9
SCF R	2 000 000–6000 000	na	100	na
SDF-1α	3000–9000	2000	10 000	18
SDF-1β	10 000–30 000	na	10 000	na
L-Selectin	na	954 000	10	300
ST2	na	500	30	na
TARC	3000–9000	331	100	5
TGF-α	100–400	22	100	na
TGF-β1	20–60	48 600	10	7
TGF-β3	10.0–30.0	2400	10	na
TIMP-1	8 nM	190 000	30	80
TIMP-2	na	106 000	10	na
TNF-α	20–50	1.25	3	4.4
TNF-β	20–50	< 156	10	16
TNF RI	45 000–90 000	1198	10	3
TNF-RII	4000–16 000	1725	100	1
TRAIL	4000–12 000	nd	5000	120
TRAIL R1	1000–3000	na	10	na
TRAIL R4	30 000–60 000	na	100	na
UPAR	50 000–150 000	2370	100	33
SVAP-1	na	125 900	100	300
VEGF	2000–6000	220	10	9
VEGF-R2	4000–8000	9768	1000	4.6



(A)



(B)

Figure 4. Cross-reactivity of capture antibodies and analytes ((a) and (c)), capture antibodies and detection antibodies ((b) and (d)) on the subarrays 3 (A) and 4 (B). Each subarray was printed with about 35 different capture antibodies (listed on the chart). Cross-reactivity was determined by using different matched secondary antibody and 50 ng/mL or 0 ng/mL of its cognate analyte, which has been described in the text. A shaded square off the diagonal indicates the presence of nonspecific signals (intensity ≥ 300). None of these nonspecific signals exceed the intensity of 600 fluorescence intensity units (images are scanned at PMT of 600 and power of 100% on Axon scanner).

The analytical sensitivity of the microarray immunoassay was determined by using serial dilutions of each single purified antigen. The sensitivity of detection, defined as the lowest concentration that delivered clear specific signal intensity above detection threshold (mean intensity of the negative control plus 2 SD), was calculated. Sixty-five (43%) of the 150 cytokine features had a sensitivity of ≤ 10 pg/mL, 52 (35%) had a sensitivity of ≤ 100 pg/mL, 27 (18%) had a sensitivity of ≤ 1000 pg/mL, and 6 (4%) had a sensitivity of ≥ 1,000 pg/mL (Table 1). Sixty-two of the cytokines represented on the chip are unique, with no corresponding commercially available ELISA kits. The sensitivity goal for these arrays was adequacy to detect biologically relevant (ie, 2-fold) increases in cytokine level above normal values in clinical fluids including serum and plasma.

Specificity of antibody microarray

Although antibodies are highly specific for their respective cognate proteins, it is possible that structurally related proteins might have similar epitopes. Cross-reactivity or nonspecific signals between immobilized

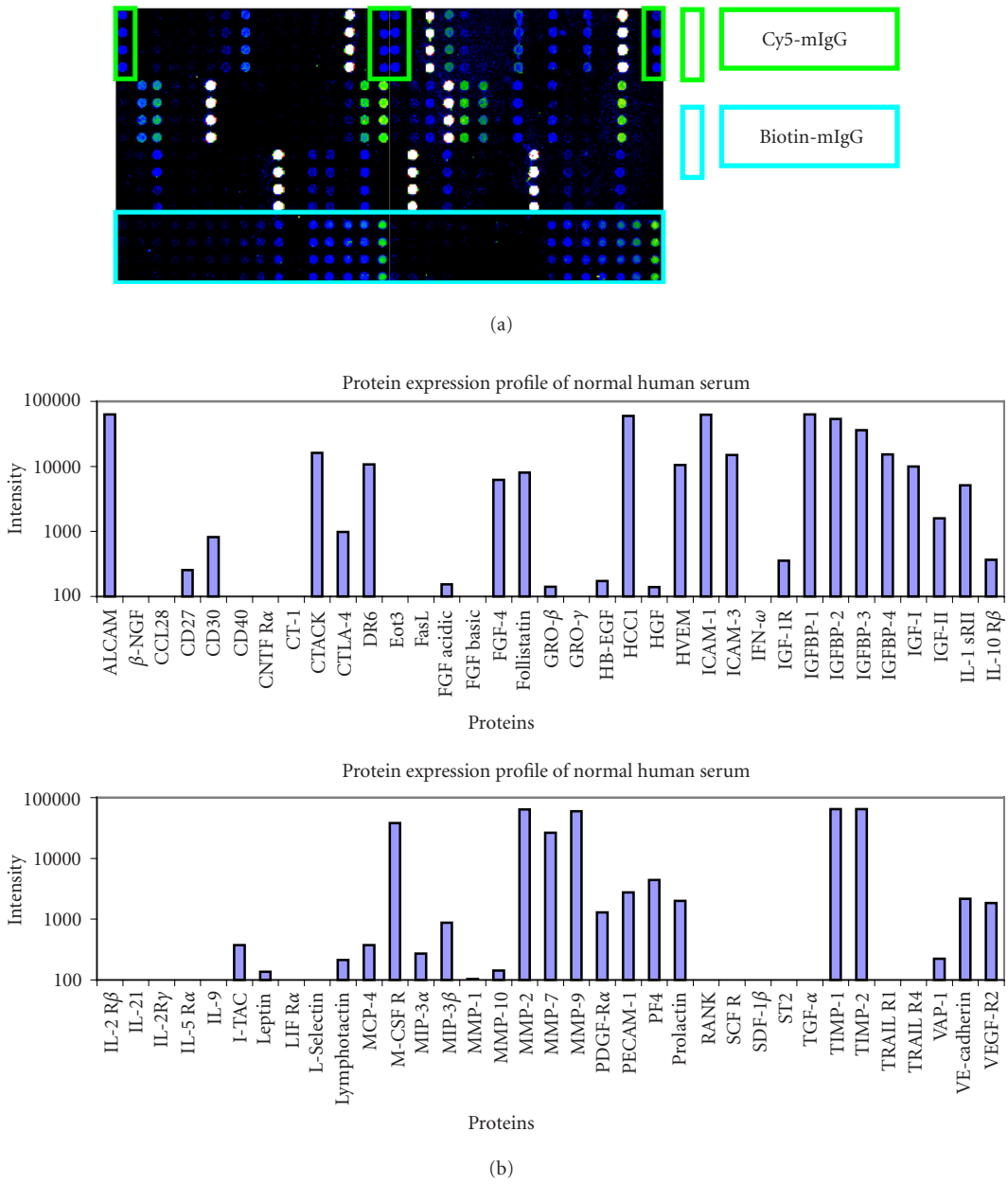


FIGURE 5. Protein expression profiling of normal human serum (Sigma) on 70-feature antibody microarray. Specific signals detected were consistent with the protein level in serum. (a) Images of detecting 70 proteins in antibody microarray obtained with Axon microarray scanner. (b) Fluorescence intensities were derived from microarray images with averaging four spots of the same feature.

capture antibodies and antigens or detection antibodies for 150 features were determined in three different ways. First, the cross-reactivity of individual antigens or detection antibodies was determined by analyzing a single antigen at 0 and 50 ng/mL with its corresponding paired detection antibody. Any signals on the 0 ng/mL array were categorized as cross-reactivity between the detection and capture antibodies. With 50 ng/mL analyte, signals should have only been observed at the specific locations. Other signals after background subtraction (the 0 ng/mL array) were assumed to be cross-reactivity from the analyte. Nonspecific signals were eliminated by segregating all features into multiple subarrays, thus the capture antibodies producing nonspecific signals were physically separated from those particular detection antibodies or antigens. This also allowed simultaneous detection of both components of a biological signaling system, such as a growth factor and its receptor.

Residual cross-reactivity and nonspecific signals were first minimized by optimization of washing and blocking conditions [4]. Next, a mixture of several analytes at concentrations of 10 ng/mL each was applied to the chip,

followed by detection with the complete detection antibody cocktail. Signals were only observed where both the analyte and its specific detection antibody were present. Then a mixture of all analytes in the subarray at 10 ng/mL each was added to the microarray, but detection was with an incomplete cocktail of antibodies. Signals were only detected on those features corresponding to the specific detection antibodies added. As shown in Figure 4, both cross-reactivity between capture antibodies and analytes and cross-reactivity between capture antibodies and detection antibodies in each subarray were minimized. Approximately 99% of nonspecific binding pairs had fluorescence intensity units of less than 300, used as the background signal minimum during the data quantitation. Interferences from biological specimen were screened using normal human serum and single detectors (biotinylated detection antibody) corresponding to each feature on the chip. Only specific signals of the proteins present in the specimen were detected by their associated detector antibody. Other nonspecific signals or background signals were negligible.

Validation of the antibody microarray

The antibody microarray was validated by running assays with normal human serum (Sigma, St. Louis, Mo) on our 70-feature protein chip. We simultaneously detected 70 proteins in normal human serum. Signals corresponding to specific proteins were identified (Figure 5), which are consistent with their presence with significant level in the normal human serum (Table 1). Proteins with low fluorescence signal intensities agree with their very low level in serum (if present), indicating that our antibody microarray provides a specific, robust, and high-throughput approach for global analysis of protein expression.

CONCLUSION

Immunoassays on microarrays hold appeal for studies requiring the ability to quantify many selected proteins simultaneously. Considerable progress has been made in this area recently in terms of increased assay sensitivity and complexity (ie, degree of multiplexing). Rolling-circle amplification facilitates the use of such arrays with its powerful degree of signal enhancement and simultaneous detection of a large number of defined analytes. Such robust systems will enable routine use of antibody arrays in both research and diagnostic modalities.

REFERENCES

[1] MacBeath G, Schreiber SL. Printing proteins as microarrays for high-throughput function determination. *Science*. 2000;289(5485):1760–1763.

[2] Zhu H, Bilgin M, Bangham R, et al. Global analysis of protein activities using proteome chips. *Science*. 2001;293(5537):2101–2105.

[3] Haab BB, Dunham MJ, Brown PO. Protein microarrays for highly parallel detection and quantitation of specific proteins and antibodies in complex solutions. *Genome Biol*. 2001;2(2):1–13.

[4] Schweitzer B, Roberts S, Grimwade B, et al. Multiplexed protein profiling on microarrays by rolling-circle amplification. *Nat Biotechnol*. 2002;20(4):359–365.

[5] Petricoin EF 3rd, Ardekani AM, Hitt BA, et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*. 2002;359(9306):572–577.

[6] Huang RP, Huang R, Fan Y, Lin Y. Simultaneous detection of multiple cytokines from conditioned media and patient's sera by an antibody-based protein array system. *Anal Biochem*. 2001;294(1):55–62.

[7] Wiese R, Belosludtsev Y, Powdrill T, Thompson P, Hogan M. Simultaneous multianalyte ELISA performed on a microarray platform. *Clin Chem*. 2001;47(8):1451–1457.

[8] Lizardi PM, Huang X, Zhu Z, Bray-Ward P, Thomas DC, Ward DC. Mutation detection and single-molecule counting using isothermal rolling circle amplification. *Nat Genet*. 1998;19(3):225–232.

[9] Schweitzer B, Wiltshire S, Lambert J, et al. Immunoassays with rolling circle DNA amplification: a versatile platform for ultrasensitive antigen detection. *Proc Natl Acad Sci USA*. 2000;97(18):10113–10119.

[10] Falipou S, Chovelon JM, Martelet C, Margonari J, Cathignol D. New use of cyanosilane coupling agent for direct binding of antibodies to silica supports. Physicochemical characterization of molecularly bioengineered layers. *Bioconjug Chem*. 1999;10(3):346–353.

[11] Nallur G, Luo C, Fang L, et al. Signal amplification by rolling circle amplification on DNA microarrays. *Nucleic Acids Res*. 2001;29(23):e118.

* Corresponding author.
E-mail: weipings@molecularstaging.com
Fax: + 1 203 776 5278; Tel: + 1 203 772 5054

Journal of Biomedicine and Biotechnology • 2003:5 (2003) 308–314 • PII. S1110724303210032 • http://jbb.hindawi.com

RESEARCH ARTICLE

Diagnosis of Ovarian Cancer Using Decision Tree Classification of Mass Spectral Data

Antonia Vlahou,[1,2]* John O. Schorge,[3] Betsy W. Gregory,[1,2] and Robert L. Coleman[3]

[1]*Department of Microbiology and Molecular Cell Biology, Eastern Virginia Medical School, Norfolk, VA 23501, USA*
[2]*Virginia Prostate Center, Eastern Virginia Medical School and Sentara Cancer Center, Norfolk, VA 23501, USA*
[3]*Division of Gynecologic Oncology, Department of Obstetrics and Gynecology, University of Texas Southwestern, Dallas, TX 75390, USA*

Received 24 October 2002; revised 16 February 2003; accepted 19 February 2003

Recent reports from our laboratory and others support the SELDI ProteinChip technology as a potential clinical diagnostic tool when combined with *n*-dimensional analyses algorithms. The objective of this study was to determine if the commercially available classification algorithm biomarker patterns software (BPS), which is based on a classification and regression tree (CART), would be effective in discriminating ovarian cancer from benign diseases and healthy controls. Serum protein mass spectrum profiles from 139 patients with either ovarian cancer, benign pelvic diseases, or healthy women were analyzed using the BPS software. A decision tree, using five protein peaks, resulted in an accuracy of 81.5% in the cross-validation analysis and 80% in a blinded set of samples in differentiating the ovarian cancer from the control groups. The potential, advantages, and drawbacks of the BPS system as a bioinformatic tool for the analysis of the SELDI high-dimensional proteomic data are discussed.

INTRODUCTION

Ovarian cancer has the highest fatality-to-case ratio of all gynecologic malignancies [1, 2]. This is attributed to the lack of early warning signs and efficacious early detection techniques [1, 3]. Another problem hindering the successful management of the disease is the paucity in prognosticators that could assist the selection of treatment modality. One of the most promising routes towards improvement in the detection and surveillance of ovarian cancer is the identification of serum markers. Utilization of the CA125 as an ovarian cancer serum marker has improved cancer detection rates during the last few years [1, 2, 3]. Nevertheless, CA125 does not diagnose early-stage cancers with high accuracy and is prone to false positives. Therefore, the need to identify additional serum markers for ovarian cancer is paramount to the successful management of this disease.

A major obstacle in finding a diagnostic biomarker is the tremendous molecular heterogeneity that exists for nearly all human cancer, suggesting that simultaneous screening of a patient specimen for multiple biomarkers will be required to improve the early detection/diagnosis of cancer. DNA chip technologies address this problem at the genomic level, and provide accessibility to gene expression profiles. However, since proteins are, for the most part, the mediators of a cell's function, the study of the changes in proteins that result from a pathological lesion, such as cancer, would appear to be a rich source of potential cancer biomarkers.

Most of the previous studies in search of diagnostic biomarkers have employed two-dimensional electrophoresis (2DE) which can resolve hundreds to thousands of proteins present in complex protein mixtures, such as cell lysates and body fluids. Although some successes have been reported in detecting potential ovarian cancer-associated biomarkers [4, 5, 6, 7], this classical proteomic technique is very time consuming, not highly reproducible, and not easily adaptable to a clinical assay format.

A recently developed mass spectrometry proteomic approach, the SELDI (surface-enhanced laser desorption/ionization) ProteinChip System (Ciphergen Biosystems, Inc, Fremont, Calif), appears to hold promise for biomarker discovery and as a potential clinical assay format [8, 9]. (The SELDI system and its applications are described in the report by Reddy and Dalmasso [10]; and a recent review by Wright [11]). Using this system, distinct protein patterns of normal, premalignant, and malignant cells were found for ovarian, esophageal, prostate, breast, and hepatic cancers [12, 13, 14]. Potential biomarkers for breast and bladder cancers were also detected in nipple aspirate fluid and urine, see respectively [15, 16], by the SELDI system.

Recent reports also support that analysis of the SELDI data by "artificial intelligence" algorithms can lead to the identification of protein "fingerprints" specific for prostate, ovarian, and breast cancers, significantly increasing the accuracy in differentiating cancer from the non-cancer groups [17, 18, 19, 20]. These studies employed

Table 1. Demographics of the cancer and control groups included in the study.

	n	Mean age	Age range	Cancer stage	n
Cancer	44	55.9	20–85	Stage I	10
Normal	34	43.7	28–59	Stage II	4
Benign	61	46.8	20–83	Stage III	21
				Stage IV	9

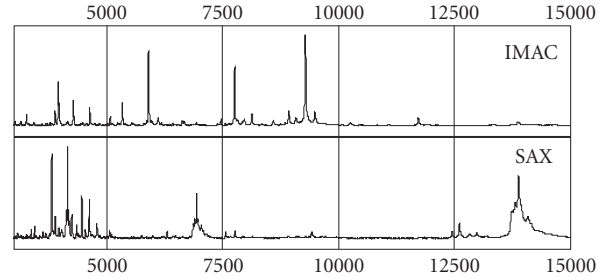


Figure 1. Protein spectra of one serum sample processed on the IMAC metal binding chip array and on the positively charged SAX chip array. Note that several different proteins are captured by the two different chip chemistries.

different algorithms to analyze the SELDI data, including a genetic algorithm [19], a decision tree [17, 18], and a support vector machine algorithm [20]. Each method appeared to be effective in developing accurate classification systems.

The high dimensionality of the data generated by SELDI requires a mathematical algorithm to analyze the data without overfitting. Since the SELDI protein profiling approach is new, it is difficult to determine up-front which algorithm to select for the data analysis and development of a "diagnostic" classifier. It is also fair to assume that different bioinformatic tools may be required for different cancer or disease systems. The objective of this study was to evaluate the commercially available classification algorithm (biomarker pattern software [BPS]) developed by Ciphergen Biosystems Inc for analysis of the SELDI serum protein profiling data from patients with ovarian cancer, benign pelvic diseases, and normal women. The potential, advantages, and drawbacks of this approach as well as suggestions for improvement are discussed.

METHODS

Serum samples

Serum samples were obtained from patients with epithelial ovarian cancer prior to treatment administration ($n = 44$), benign pelvic diseases ($n = 61$), and from women with no evidence of pelvic disease ($n = 34$) enrolled through the Division of Gynecologic Oncology, University of Texas, Southwestern Medical Center. Informed consent was obtained from all patient and control groups. The demographics of the patients and the stage distribution of the ovarian cancers are presented in Table 1. Benign conditions included benign pelvic masses (endometriosis, cystadenomas, hydrosalpinx, lipoma, Brenner tumor, fibroids, endometrial polyp). The sera were aliquoted and stored at −80°C.

SELDI processing of serum samples

Serum samples were applied on the strong anion exchange (SAX) and immobilized-copper (IMAC) chip surfaces. In brief, 21 μL of serum were mixed with 30 μL 8M urea in 1% CHAPS-PBS pH 7.4 buffer for 30 minutes at 4°C, followed by the addition of 100 μL of 1M urea in 0.125% CHAPS-PBS buffer and 600 μL of binding buffer compatible with the type of surface in use (PBS for IMAC and 20 mM Hepes containing 0.1% Triton for SAX). Fifty μL of the diluted samples were then applied onto the chips using a bioprocessor. Following a 30-minute incubation, nonspecifically bound molecules were removed by 3 brief washes in binding buffer followed by 3 washes with HPLC-gradient H_2O. Sinapinic acid (2X 1 μL of 50% SPA in 50% ACN-0.1%TFA) was applied to the chip array surface and mass spectrometry was performed using a PBS2 SELDI mass spectrometer (Ciphergen Biosystems Inc). Protein data were collected by averaging a total of 192 laser shots. Mass calibration was performed using the all-in-one peptide standard (Ciphergen Biosystems Inc) which contains vasopressin (1084.2 daltons), somatostatin (1637.9 daltons), bovine insulin β-chain (3495.9 daltons), human insulin recombinant (5807.6 daltons), and hirudin (7033.6 daltons). All samples were processed in duplicate.

Processing of SELDI data

Protein peaks were labeled and their intensities were normalized for total ion current (mass range 2–200 kd) to account for variation in ionization efficiencies, using the SELDI software (version 3.1). Peak clustering was performed using the Biomarker Wizard software (Ciphergen Biosystems) and the following specific settings: spectral data from IMAC surface; signal/noise (first pass): 4, minimum peak threshold: 10%, mass error: 0.3%, and signal/noise (second pass): 2 for the 2–20 kd mass range and signal/noise (first pass): 5, minimum peak threshold: 10%, mass error: 0.3%, and signal/noise (second pass): 2.5 for the 20–100 kd mass range. Spectral data from the SAX surface were analyzed with the same set of settings with the difference that the minimum peak threshold was set to 5%. With these labeling parameters, a total of 122 protein clusters (45 from the IMAC and 77 from the SAX surface) were generated. Peak mass and intensity were exported to an excel file, and the peak intensities from each duplicate spectra were averaged. Pattern recognition and sample classification were performed using the BPS. The decision tree described in the result section was generated using the Gini method nonlinear combinations. A 10-fold cross-validation analysis was performed as an

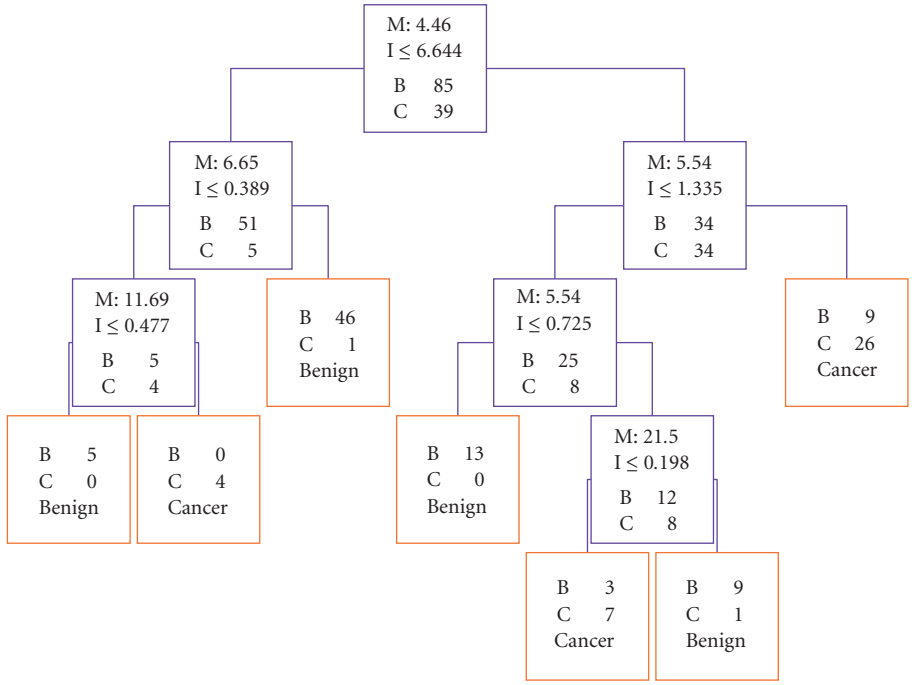


FIGURE 2. Decision tree classification of the ovarian cancer (C) and noncancer (normal and benign or B) groups. The blue boxes show the decision nodes with the peak mass (M in kd), the peak intensity (I) cutoff levels, and the number of samples. The 5.54, 6.65, and 11.7 kd masses were detected on the IMAC chip, and the 4.4 and 21.5 kd on the SAX chip. These five masses form the splitting rules. Cases that follow the rule are placed in the left daughter node. The red boxes are the terminal nodes with the classification being either cancer or benign (normal + benign).

initial evaluation of the test error of the algorithm. Briefly, this process involves splitting up the dataset into 10 random segments and using 9 of them for training and the 10th as a test set for the algorithm. Multiple trees were initially generated from the 122 classifiers by varying the splitting factor by increments of 0.1. These trees were evaluated by cross-validation analysis. The peaks that formed the main splitters of the tree with the highest prediction rates were then selected, the tree was rebuilt based on these peaks alone and evaluated by the test set. The values of P were calculated based on t-test (Biomarker Wizard software). The value $P < .05$ was considered to be statistically significant.

RESULTS

One hundred thirty-nine serum samples were assayed by SELDI mass spectrometry. Both SAX and IMAC surfaces could effectively resolve low-mass (< 20 kd) protein peaks, although the SAX surface appeared superior in resolving larger (> 20 kd) protein peaks. Figure 1 shows representative protein spectra from one serum sample processed on SAX and IMAC chips.

Of a total of 139 serum samples, 124 (85 controls and 39 cancers) were randomly selected to form the learning set and 15 (10 controls and 5 cancers) to form the blinded test set for the algorithm. Five peaks were selected by the BPS algorithm to discriminate cancer from the noncancer groups. Figure 2 is the decision tree that was generated from the learning set to classify the two groups. Three peaks (5.54, 6.65, and 11.7 kd) detected on the IMAC chip and 2 (4.4 and 21.5 kd) detected on the SAX surface form the main splitters. Their mass spectra and gray-scale/gel views are shown in Figures 3, 4, 5, 6, and 7. These peaks have significantly different intensity levels between the cancer and benign or normal controls with the exception of the 6.65 and 21.5 kd peaks, which did not differ significantly between cancers and benigns (Table 2). A 10-fold cross-validation analysis was performed as an initial evaluation of the accuracy of the algorithm in predicting ovarian cancer. A specificity of 80% and sensitivity of 84.6% were obtained (Table 3). In the test set, sensitivity and specificity of 80% were obtained (Table 3). The misclassified samples in the test set included one benign (uterine fibroid), one normal, and a stage III C cancer.

DISCUSSION

The high degree of genetic heterogeneity associated with human cancers makes it likely that panels of multiple biomarkers will be needed to improve early detection/diagnosis. This entails the development of high-throughput proteomic and genetic approaches as well as of reliable bioinformatic tools for data analysis.

Table 2. Statistical comparison of the intensity levels of the peaks used in the decision tree between the cancer and control groups. C-N: cancer versus normal; C-B: cancer versus benign; and C-B/N: cancer versus normal and benign.

MW (kd)	*P* (C-N)	*P* (C-B)	*P* (C-N/B)
4.47	< 0.001	< 0.001	< 0.001
5.54	< 0.001	< 0.001	< 0.001
6.65	< 0.001	0.13	< 0.001
11.69	< 0.001	0.017	< 0.001
21.5	< 0.001	0.43	< 0.001

Table 3. Performance of the decision tree in predicting ovarian cancer. Numbers in parentheses denote the number of correctly classified sample out of total number of samples in the group.

	Sensitivity%	Specificity%
Learning set	94.9 (37/39)	85.9 (73/85)
Cross-validation	84.6 (33/39)	80.0 (68/85)
Test set	80.0 (4/5)	80.0 (8/10)

The SELDI proteinChip system offers the advantage of rapid and simultaneous detection of multiple proteins from complex biologic mixtures. We employed this system in combination with the BPS classification algorithm for protein profiling of ovarian cancer in serum. Using this approach, a classifier that was 80% accurate in discriminating patients with ovarian cancer from patients with benign disease and healthy controls from a blinded test set was generated. Evaluation of the classifier by cross-validation and the analysis of the independent test set offers statistical confidence of the potential of this approach as an ovarian cancer detection tool. However, the sample size included in this study decreases the validity of generalized conclusions. Complete evaluation of this classifier will require testing its prediction rates for larger "blinded" and independent serum sets.

The BPS software was found to be relatively simple to use. However, BPS, like other mathematical algorithms, is prone to data overfitting, and also is not reliable when a large number of variables relative to samples sizes are included in the analysis. A preselection process of the most significant variables using statistical analysis (eg, ROC curve, ANOVA) may help in alleviating this problem.

Petricoin et al [19] recently reported the successful application of a genetic algorithm for the analysis of SELDI proteomic data from ovarian cancer patients. In this study, five discriminatory peptides were detected, molecululr mass range 500–2500 daltons, and the accuracy in predicting ovarian cancer in a blinded set of samples was 97.4%. We focused on the analysis of potential biomarkers in higher mass ranges (> 2000 daltons). Furthermore, in contrast to the case where BPS algorithm is processed, that is, labeled peak information is analyzed, the genetic algorithm employed by Petricoin et al analyzes time-of-flight

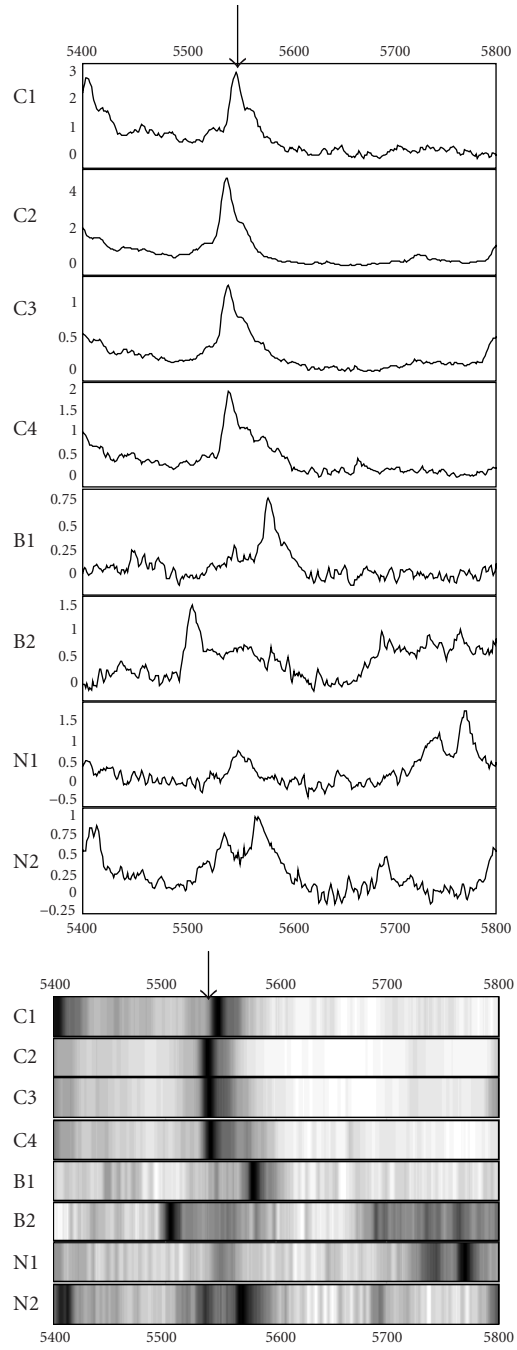


Figure 3. Spectra (top) and grey-scale or gel views (bottom) of the peaks (arrows) forming the splitting rules. The protein peak was detected on IMAC chip. The peak appears to be upregulated in the cancer (C1–C4) compared to the benign (B1-B2) and normal (N1-N2) groups.

"raw" SELDI data. In this case, prerequisite for the further identification of the potential discriminatory markers is the coupling of the genetic algorithm with a peak identification system where the raw data are translated into protein peak information. BPS employs the peak identification system of the SELDI software facilitating

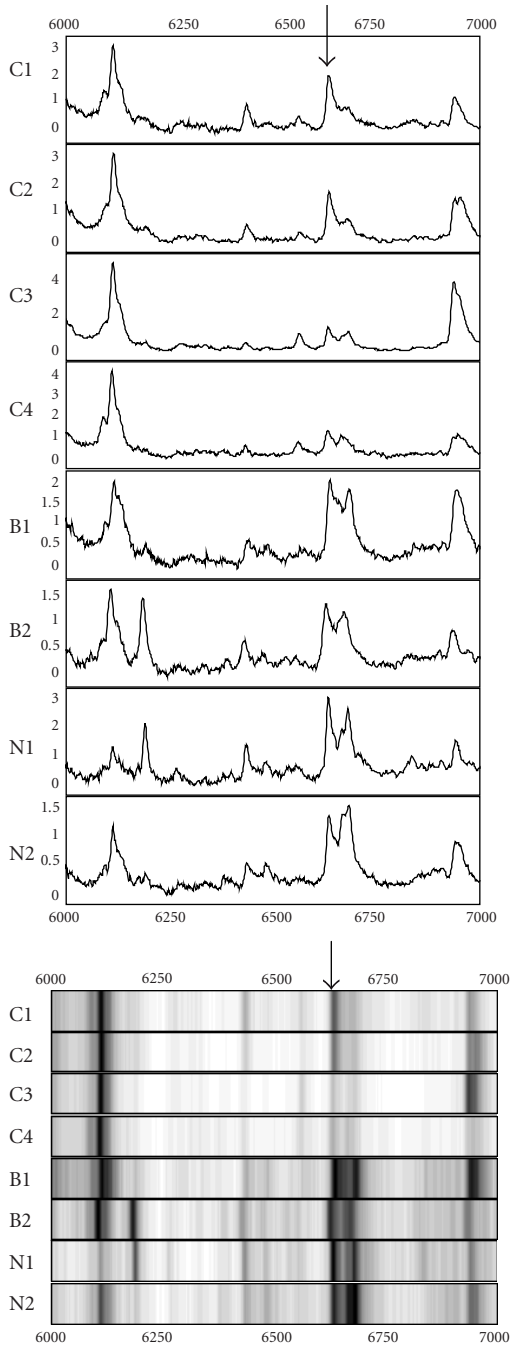


Figure 4. Spectra (top) and grey-scale or gel views (bottom) of the peaks (arrows) forming the splitting rules. The protein peak was detected on IMAC chip. The peak appears to be downregulated in the cancers.

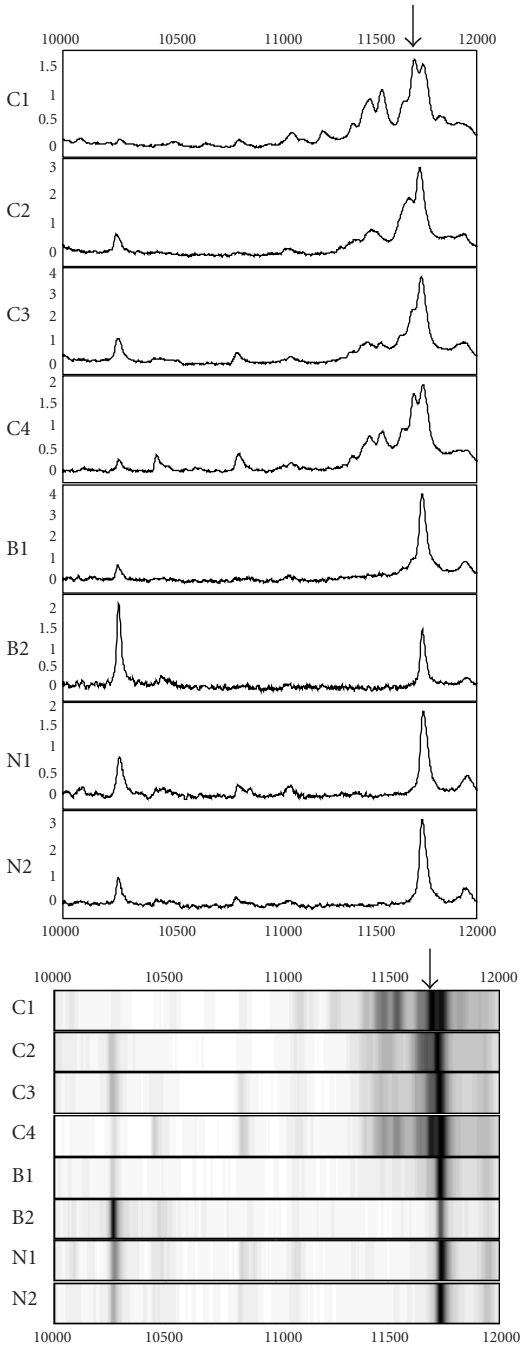


Figure 5. Spectra (top) and grey-scale or gel views (bottom) of the peaks (arrows) forming the splitting rules. The protein peak was detected on IMAC chip. The peak appears to be upregulated in cancer (C1–C4) compared to the benign (B1-B2) and normal (N1-N2) groups.

biomarker detection. It should be noted, however, that careful and precise selection of the peak labeling settings and normalization of peak intensities are considered critical for biomarker identification and for the efficient and reliable performance of any learning algorithm used in conjunction with the SELDI system.

Besides providing a preliminary evaluation of the suitability of BPS for the comparison of SELDI data, our study also demonstrates the potential of combining spectral data from different types of surfaces as a means to increase protein resolution. Although, compared to SELDI,

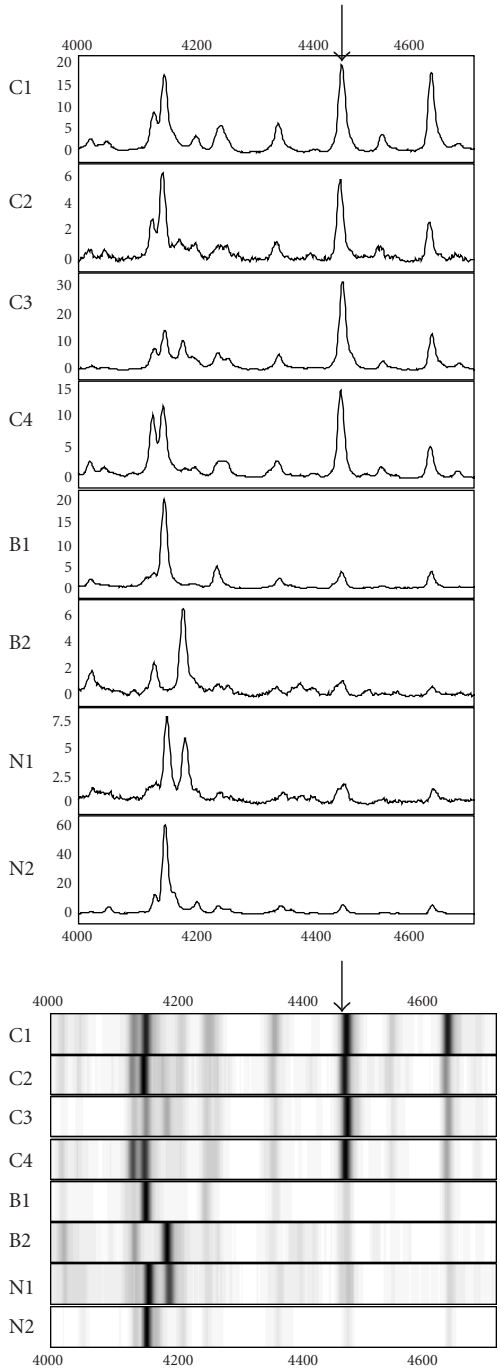


Figure 6. Spectra (top) and grey-scale or gel views (bottom) of the peaks (arrows) forming the splitting rules. The protein peak was detected on the SAX surface. The peak appears to be up-regulated in the cancer (C1–C4) compared to the begin (B1-B2) and normal (N1-N2) groups.

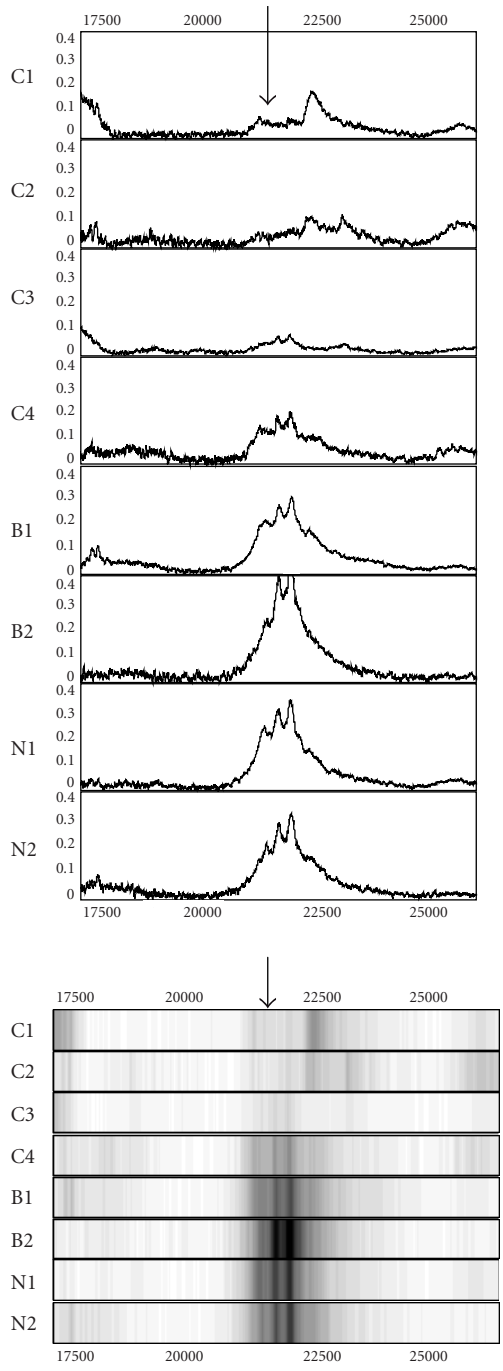


Figure 7. Spectra (top) and grey-scale or gel views (bottom) of the peaks (arrows) forming the splitting rules. The protein peak was detected on the SAX surface. The peak appears to be down-regulated in the cancers.

the resolving power of 2D gel electrophoresis remains unchallenged, we have found that this combinatorial approach can significantly enhance biomarker discovery and increase test accuracy for ovarian and breast cancers from 70–75% up to 90% [21].

In conclusion, the BPS software appears to be potentially suitable for analysis of the high-dimensional SELDI spectral data. Avenues for improvement of the algorithm performance include optimization of the peak labeling process as well as preselection of the most significant peaks by statistical approaches. More extended studies

will be required to validate the potential and reliability of BPS as a bioinformatic tool for proteomic studies. It should also be emphasized that comparative analysis of different types of algorithms will be of paramount importance for the better evaluation of their performance and the selection of the bioinformatic features needed for effective biomarker discovery and discrimination of cancer.

ACKNOWLEDGMENTS

This study was supported by grants from the Gustavus and Louise Pfeiffer Research Foundation, the Early Detection Research Network, NCI (CA85067), and the Virginia Prostate Center.

REFERENCES

[1] Hensley ML, Castiel M, Robson ME. Screening for ovarian cancer: what we know, what we need to know. *Oncology (Huntingt)*. 2000;14(11):1601–1616.

[2] Holschneider CH, Berek JS. Ovarian cancer: epidemiology, biology, and prognostic factors. *Semin Surg Oncol*. 2000;19(1):3–10.

[3] Menon U, Jacobs IJ. Recent developments in ovarian cancer screening. *Curr Opin Obstet Gynecol*. 2000;12(1):39–42.

[4] Jones MB, Krutzsch H, Shu H, et al. Proteomic analysis and identification of new biomarkers and therapeutic targets for invasive ovarian cancer. *Proteomics*. 2002;2(1):76–84.

[5] Bergman AC, Benjamin T, Alaiya A, et al. Identification of gel-separated tumor marker proteins by mass spectrometry. *Electrophoresis*. 2000;21(3):679–686.

[6] Alaiya AA, Franzen B, Fujioka K, et al. Phenotypic analysis of ovarian carcinoma: polypeptide expression in benign, borderline and malignant tumors. *Int J Cancer*. 1997;73(5):678–683.

[7] Thompson S, Turner GA. Elevated levels of abnormally-fucosylated haptoglobins in cancer sera. *Br J Cancer*. 1987;56(5):605–610.

[8] Hutchens TW, Yip TT. New desorption strategies for the mass spectrometric analysis of macromolecules. *Rapid Commun Mass Spectrom*. 1993;7:576–580.

[9] Merchant M, Weinberger SR. Recent advancements in surface-enhanced laser desorption/ionization-time of flight-mass spectrometry. *Electrophoresis*. 2000;21(6):1164–1177.

[10] Reddy G, Dalmasso EA. SELDI proteinchip® array technology: protein-based predictive medicine and drug discovery applications. *J Biomed Biotechnol*. 2003;2003(4):237–241.

[11] Wright GL Jr. SELDI proteinchip MS: a platform for biomarker discovery and cancer diagnosis. *Expert Rev Mol Diagn*. 2002;2(6):549–563.

[12] Wright GL Jr, Cazares LH, Leung SM, et al. Proteinchip(R) surface enhanced laser desorption/ionization (SELDI) mass spectrometry: a novel protein biochip technology for detection of prostate cancer biomarkers in complex protein mixtures. *Prostate Cancer Prostatic Dis*. 1999;2(5-6):264–276.

[13] Paweletz CP, Gillespie JW, Ornstein DK, et al. Rapid protein display profiling of cancer progression directly from human tissue using a protein biochip. *Drug Dev Res*. 2000;49:34–42.

[14] Cazares LH, Adam BL, Ward MD, et al. Normal, benign, preneoplastic, and malignant prostate cells have distinct protein expression profiles resolved by surface enhanced laser desorption/ionization mass spectrometry. *Clin Cancer Res*. 2002;8(8):2541–2552.

[15] Paweletz CP, Trock B, Pennanen M, et al. Proteomic patterns of nipple aspirate fluids obtained by SELDI-TOF: potential for new biomarkers to aid in the diagnosis of breast cancer. *Dis Markers*. 2001;17(4):301–307.

[16] Vlahou A, Schellhammer PF, Mendrinos S, et al. Development of a novel proteomic approach for the detection of transitional cell carcinoma of the bladder in urine. *Am J Pathol*. 2001;158(4):1491–1502.

[17] Adam BL, Qu Y, Davis JW, et al. Serum protein fingerprinting coupled with a pattern-matching algorithm distinguishes prostate cancer from benign prostate hyperplasia and healthy men. *Cancer Res*. 2002;62(13):3609–3614.

[18] Qu Y, Adam BL, Yasui Y, et al. Boosted decision tree analysis of surface-enhanced laser desorption/ionization mass spectral serum profiles discriminates prostate cancer from noncancer patients. *Clin Chem*. 2002;48(10):1835–1843.

[19] Petricoin EF, Ardekani AM, Hitt BA, et al. Use of proteomic patterns in serum to identify ovarian cancer. *Lancet*. 2002;359(9306):572–577.

[20] Li J, Zhang Z, Rosenzweig J, Wang YY, Chan DW. Proteomics and bioinformatics approaches for identification of serum biomarkers to detect breast cancer. *Clin Chem*. 2002;48(8):1296–1304.

[21] Vlahou A, Laronga C, Wilson L, et al. A novel approach toward development of a rapid blood test for breast cancer. *Clin Breast Cancer*. 2003;4(3):203-209.

* Corresponding author.
Current address: Foundation for Biomedical Research, Academy of Athens, Athens, Greece
E-mail: vlahoua@bioacademy.gr
Fax: + 30 210 6597545; Tel: + 30 210 6597519