Chapter 3
Identification of Candidate Vaccine Antigens In Silico

Darren R. Flower, Matthew N. Davies, and Irini A. Doytchinova

Abstract The identification of immunogenic whole-protein antigens is fundamental to the successful discovery of candidate subunit vaccines and their rapid, effective, and efficient transformation into clinically useful, commercially successful vaccine formulations. In the wider context of the experimental discovery of vaccine antigens, with particular reference to reverse vaccinology, this chapter adumbrates the principal computational approaches currently deployed in the hunt for novel antigens: genome-level prediction of antigens, antigen identification through the use of protein sequence alignment-based approaches, antigen detection through the use of subcellular location prediction, and the use of alignment-independent approaches to antigen discovery. Reference is also made to the recent emergence of various expert systems for protein antigen identification.

3.1 Introduction

The overwhelming case for vaccines and vaccination was long ago proven, yet vaccines remain stubbornly underused. Controversy continues to surround vaccines: it took over 10 years for a contentious connection between autism and the MMR vaccine 1998 to be finally and ambiguously discredited [1]. Yet, for all the prevalence of misinformation and muddled thinking, mass vaccination represents—by far and away—the most efficient, efficacious, and effective form of prophylactic medical intervention currently available to combat disease.

During most of the last century, in the developed world, over 600,000 people died on average annually from a combination of smallpox, diphtheria, polio, measles, and rubella; today this figure has fallen below 100. Smallpox in particular was always a dreaded killer. Indeed, even during the 1960s, at least 10 million cases

D.R. Flower (✉)
School of Life and Health Sciences, University of Aston, Aston Triangle, Birmingham, UK
e-mail: d.r.flower@aston.ac.uk

D.R. Flower and Y. Perrie (eds.), *Immunomic Discovery of Adjuvants and Candidate Subunit Vaccines*, Immunomics Reviews: 5, DOI 10.1007/978-1-4614-5070-2_3,

of smallpox were reported annually from across the globe, leading to about 2 million deaths a year. Yet, today, the disease has been completely eradicated. In the last 30 years, there have been no known cases. Poliomyelitis or polio is the other large-scale disease which has come closest to eradication. Its success too has been formidable: in 1991, the Pan American Health Organization effectively eradicated polio from the Western Hemisphere, since when the Global Polio Eradication Programme has significantly decreased the overall incidence of Poliomyelitis through the rest of the world. In 1988, there were approximately 350,000 cases spread through 125 countries; in the past years, global figures amounted to less than 2,000 annually.

Yet, in spite of such remarkable success, death from vaccine-preventable diseases remains unacceptably high [2]. There are over 70 common infectious diseases responsible for one in four deaths globally. Rotavirus and Pneumococcus are pathogens causing diarrhoea and pneumonia, the leading causes of infant deaths in underdeveloped countries. In the next decade, effective, widespread vaccination programs against such pathogenic microbes could save the lives of 7.6 million children under 5 years of age. Hepatitis B causes 600,000 deaths in adults and children aged over 5. Seasonal, non-pandemic influenza kills upwards of half a million globally each year. For those aged under 5 in particular, a series of diseases causes an extraordinary and largely preventable death toll. For example, tetanus accounts every year for 198,000 deaths, pertussis is responsible for over 290,000 deaths, Hib gives rise to in excess of 386,000 deaths, diphtheria accounts for 4,000 deaths, and yellow fever over 15,000 deaths. Arguably, the most regrettable, the most lamentable situation is that of measles. Measles accounts for the unneeded deaths of 540,000 under-fives and over 70,000 adults and older children.

Despite this, the situation is by no means bleak. By the close of 2008, approximately 42 million had been vaccinated against Hib and 192 million children against hepatitis B. During its first decade, vaccinations against polio, Hep B, Hib, measles, pertussis, and yellow fever funded by GAVI had prevented the unnecessary loss of over 5 million lives. There are approximately 50 vaccines licensed for use in humans, around half of these are widely prescribed. Yet, most of these vaccines target the prevention of common childhood infections, with the remainder addressing tropical diseases encountered by travellers to the tropics; only a relatively minor proportion combat endemic disease in under-developed countries. Balancing the persisting need against the proven success and anticipated potential, vaccines remain an area of remarkable opportunity for medical advance, leading directly to unprecedented levels of saved and improved lives.

From a commercial perspective, the vaccine arena has long been neglected, in part because of the quite astonishing success limned above; today, and in comparative terms at least, activity within vaccine discovery is feverish [3, 4]. During the last 15 years, tens of vaccines and vaccine candidates have moved successfully through clinical trials, and vaccines in late development number in the hundreds. In stark contrast to antibiotics, vaccine resistance is negligible and nugatory.

Despite the egregious and outrageous success enjoyed by vaccines, many major issues persist. The World Health Organisation long ago identified tuberculosis

(TB), HIV, and malaria as the three most significant life-threatening infectious diseases globally. No vaccine has been licensed for malaria or HIV, and there seems little realistic hope for such vaccines appearing in the immediate future. Bacille Calmette Guérin (BCG), the key anti-TB vaccine, is of limited efficacy [5]. Levels of morbidity and mortality generated by diseases already targeted by vaccines remain high. Influenza is the key example, with a global annual estimated death toll in the region of half a million.

In the twenty-first century, the world continues to be threatened by infectious and contagious diseases of many kinds: visceral leishmaniasis, Marburg's disease, West Nile, dengue, as well as SARS potentially pandemic H5N1 influenza, and over 190 human and emerging zoonotic infections, as well as the persisting threat from HIV, TB, and malaria mentioned above. All this is further compounded by the additional risk arising from antibiotic-resistant bacteria and bioterrorism, not to mention major quasi-incidental issues, such climate change, an accelerating growth in the world's population, increased travel, and the overcrowding seen within the burgeoning populations concentrated into major cities [6].

For reasons we shall touch on below, the discovery of vaccines is both more urgent and more difficult than it has ever been. In an era where conventional drug discovery has been seen to fail—or at least as seen by cupiditous investors, for whom the current model of pharmaceutical drug discovery is broken—vaccines are one of a number of biologically derived therapies upon which the future economic health of the pharmaceutical industry is thought to rest. The medical need, as stated above, is clear. Set against this is the unfortunate realisation that vaccines exist for most easily targeted diseases, those mediated by neutralising antibodies, and so outstanding vaccine-targets are those of more intractable diseases mediated primarily by cellular immunity. To address those properly requires what all discoveries required: hard work and investment; but they also need new ideas, new thinking, and new vaccine discovery technology. Amongst, these are computational techniques, the most promising of which are those targeting the discovery of novel vaccine antigens: the candidate subunit vaccines of tomorrow see Fig. 3.1.

3.2 Vaccines

Vaccines are agents—either molecular (epitope- or antigen-based vaccines) or supramolecular (attenuated or inactivated whole pathogen vaccines)—which are able to create protective immunity against specific pathogenic infectious microorganisms and any diseases to which they might give rise. Protective immunity can be characterised as an enhanced but highly specific response to consequent re-infection—or infection by an evolutionarily closely related micro-organisms—made by the adaptive immune system. Such increased or enhanced immunity is facilitated by the quantitative and qualitative augmentation of immune memory, which is able to militate against the pernicious effects of infectious disease. Vaccines synergise with the herd immunity they help engender, leading to reduced transmission rates as well as prophylaxis against infection.

Fig. 3.1 Whole antigen discovery. When looking at a reverse vaccinology process, the discovery of candidate subunit vaccines begins with a microbial genome, perhaps newly sequence, progresses through an extensive computational stage, ultimately to deliver a shortlist of antigens which can be validated through subsequent laboratory examination. The computational stage can be empirical in nature; this is typified by the statistical approach embodied in vaxijen [115]. Or this stage can be bioinformatic; this involves predicting subcellular location and expression levels and the like. Or, this stage can take the form of a complex mathematical model which uses immunoinformatic models combined with mathematical methods, such as metabolic control theory [153], to predict cell-surface epitope populations

The term "vaccine" derives from vacca (Latin for cow). The words vaccine and vaccination were coined specifically for anti-smallpox immunization by the discoverer of the technique, Edward Jenner (1749–1823). These terms were later extended by Louis Pasteur (1822–1895) to include a far more extensive orbit or remit, including the entire notion of immunisation against any disease [2, 3, 6].

Several fundamentally distinct varieties of vaccine exist. These include *inter alia* inactivated or attenuated whole pathogen-based vaccines; subunit vaccines are based on one or more protein antigens, vaccines based upon one or more individual epitopes, carbohydrate-based vaccines, and combinations thereof. Hitherto, the best-used and, thus, the most successful types of vaccine were built from attenuated—"weakened" or non-infective or otherwise inactivated—pathogenic whole organisms, be they bacterial or viral in nature. Well-known examples include the following: the BCG vaccine which acts prophylactically against tuberculosis and Albert Sabin's anti-poliomyelitis vaccine based on attenuated poliovirus. The vast majority of subunit vaccines are immunogenic protein molecules, and are typically discovered using a somewhat haphazard search process.

Concerns over the safety of whole-organism vaccines long ago prompted the development of other kinds of vaccine strategy, including those based upon antigens as the innate or immanent active biological constituent of either single or composite vaccines. The vaccine which targets Hepatitis B is a good exemplar of a so-called subunit vaccine as it is based on a protein antigen: the viral envelope hepatitis B surface antigen. Other types of as-yet-unproven vaccines include those based on epitopes and others based on antigen-presenting cells; many have entered clinical trials, but none have fulfilled their medical or commercial potential.

It is often difficult to capture the proper scientific meaning and use of recondite terms, often borrowed from common usage or archaic language. So, let us be more specific. An immunogen—a molecular moiety exhibiting the property of immunogenicity—is any material or substance capable of eliciting a specific immune response. An antigen, on the other hand, is a molecular moiety exhibiting the property of antigenicity. It is a substance or material recognised by a primed immune system. Such a persisting state of immune readiness may be mediated by humoral immunity (principally via the action of soluble antibodies) or by cellular immunity (as mediated by T-cells, antigen presenting cells (APCs), or other phagocytic cells), or a combination of both, in what is often referred to as a "recall" response.

Immunogenicity is vital: it is the signature characteristic or property that prompts a certain molecular moiety to evoke a significant immune response. Here, we shall strictly limit use of "immunogen" and "antigen" to a sole meaning. Here, an "antigen" or an "immunogen" will mean a protein that is capable of educing some kind of discernible response from the host immune system. Specifically, and for practical reasons, we will almost exclusively be referring to proteins derived from a pathogenic micro-organism.

At present, the prophylaxis engendered by all current effective vaccines—all except BCG—is primarily mediated by the humoral immune system, via soluble antibodies. However, the disease mechanisms of most serious diseases for which vaccines are not available are usually mediated by cellular immunity. Thus, for untreated disease, we seek to identify immunogenicity generated principally by cellular responses or by a combination of cellular and humoral responses, rather than by humoral immunity alone.

To some extent, subunit vaccines can be thought to represent something of a compromise between vaccines based on attenuated or otherwise inactivated whole-organisms and the many more recent and more innovative vaccine strategies typified by epitope or poly-epitope vaccines. Vaccines based around whole pathogens have long engendered safety concerns [7–9]. From the Lubeck disaster and the cutter incident [10–12] to the recent MMR debacle, issues over safety, real or imagined, have always dogged the development of vaccines [1, 9]. Indeed, during the eighteenth century the pre-vaccination practice of variolation against Smallpox prefigured much of the current debate over the perceived danger of vaccines [13].

While the case for vaccines is unanswerable, we should not be complacent. Any live vaccine, however extensively attenuated, can revert to a pathogenic, disease-inducing form. This is currently an on-going issue for polio vaccination [14]. Other issues, particularly the chemical or biological contamination of vaccines during manufacture, remain enduring and persistent problems. Undesired immunogenicity, the type leading to severe and pathological immune responses, rather than enduring immune memory, is a concern for both whole-organism and subunit-based vaccines, as well as putative biologics [15]. Immunologists and vaccinologists have thus long sought alternatives to the use of whole organisms as vaccines. Subunit vaccines and conjugate vaccines are one such. Vaccines based

on epitopes, singly or in combination, are another. The diversity of innovations in vaccine design holds much potential for success, but, thus far at least, has proved spectacularly unsuccessful in a clinical context.

Logically, a vaccine that relies solely on, at most, a few well-chosen epitopes, should be effective, efficacious, and, above-all, safe. Epitopes, as peptides, may be cytotoxic and might possibly prompt some kind of inopportune immune response but cannot be infective or revert to infectivity. In many ways, epitopes are closer in size and share many properties with synthetic small molecules; possibly dealing with their pharmacokinetics as such may be better than thinking of them as biologic drugs. In practice, of course, epitope-based vaccines, like subunit vaccines, suffer from poor immunogenicity, necessitating the use of a complex combination of adjuvants and complicated delivery systems.

For diverse reasons, including immunogenicity, stimulating protective immune responses against intracellular pathogens remains problematic when using non-replicating vaccines. Why should this be? First, the immune response is very complex, involving both the innate and adaptive immunity, and significant interaction between them. In all probability, and particularly when viewed in the context of the whole population, many epitopes and danger signals are involved; likewise, the many different immune actors, be they acting at the cellular or molecular levels, interact with each other and are subject to complex mechanisms of genetic, epigenetic, and system-level control and regulation. It may be that only the large and complex organism-sized vaccines can induce the range of immune responses necessary across the population to induce protection, since they comprise a potential host of immunogenic molecular moieties, not just a single immunodominant epitope See Fig. 3.2.

In that which follows, we shall seek to explore the availability and accessibility of informatic techniques and informatic tools used to identify candidate subunit vaccines of microbial origin. Yet, we shall start by adding context with an examination of experimental approaches to antigen discovery: so-called reverse vaccinology. Reverse vaccinology already relies on informatics, but, in a sense at least, what we would like to do using informatics is to reproduce as much as is possible the steps inherent in successful reverse vaccinology in silico rather than in vitro.

3.3 Reverse Vaccinology and the Experimental Identification of Antigens

Reverse vaccinology, and the necessary computational support, is a much more prevalent means of identifying subunit vaccines [16]. See Fig. 3.1. Even today, many experimentalists retain a deep and atavistic distrust of all computation. Experimentalists seldom trust the reliability and dependability of computational methodology, choosing to trust instead in what they believe to be infallible, if actually rather elusive, empirical reliability of observations, experiments, and the whole paraphernalia of laboratory experimentation. Yet, things are in the process of

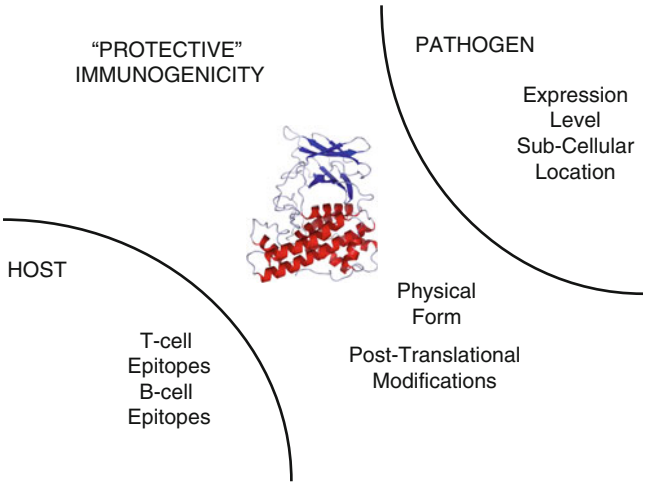

Fig. 3.2 Factors underlying immunogenicity. As elaborated in the text, the phenomenon of immunogenicity can be explored through the diversity of underlying factors contributing to the instigation of the immune response. The can be assigned to the host (epitope recognition), the pathogen (location and expression level), and also factors intrinsic to the protein antigen itself, such as the possession of post-translational danger signals

changing, and this change is likely to accelerate as we move forward into a future that looks more parsimonious and uncertain by the day.

Vaccines have come a long way from the days when they were prepared directly from the fluids of smallpox pustules or extracts of infected spinal cords. Yet vaccine discovery and development remains firmly empirical. Many modern vaccines still comprise entire inactivated pathogens. While vaccines targeting papillomavirus, tetanus, hepatitis B, and diphtheria are subunit vaccines, few are recombinant proteins devoid of contaminants. Some would argue that the only molecular vaccines are glycoconjugates: oligosaccharides conjugated to immunogenic carrier proteins.

Conventional empirical, experimental, laboratory-based microbiological ways to identify putative candidate antigens require cultivation of target pathogenic micro-organisms, followed by teasing out their component proteins, analysis in a series of in-vitro and in-vivo assays, animal models and with the ultimate objective of isolating one or two proteins displaying protective immunity.

Unfortunately, in reality, the process is more complex, and more confusing, and much more confounding as this brief synopsis might suggest. Cultivating pathogens outside the environment offered by their host organism can be difficult, even impossible. Not every protein is readily expressed in adequate quantities in vitro, and many proteins are only expressed in an intermittent basis during the time course of infection. Thus, a considerable number of potential, putative, and possible vaccine candidate antigens could be missed by conventional experimental approaches.

Reverse vaccinology [16–19] has the potential to analyse genomes for potential antigens, initially scanning "open reading frames" (ORFs), then selecting proteins because they are open to surveillance by the host immune system. This usually involves some complex combination of informatic-based prediction methodologies. Recombinant expression of the resulting set of identified molecules can overcome their reduced natural abundance, which has often prevented us recognising their true potential. By enlarging the repertoire of native antigens, this technology can help to foster the development of a new cohort of vaccines.

Reverse vaccinology was originally established and has been established by studying *Neisseria meningitidis*, which is responsible for meningococcal meningitis and sepsis. Vaccines are currently available for all serotypes, except that serogroup B. *N. meningitidis* ORFs were found initially [20, 21]; 570 proteins were then identified, 350 expressed in vitro and 85 found to be surface exposed. Seven proteins elicited immunity over many strains. The culmination of this work was a "universal" vaccine for serogroup B based on five antigens [22]. This proto-vaccine, when used with Alum as adjuvant, induced murine bactericidal antibodies versus 78 % of 85 meningococcal strains drawn from the world population of *N. meningitidis*. Strain coverage increases to over 90 % when used with CpG or MF59 as adjuvant.

Another key illustration is *Porphyromonas gingivalis*, an anaerobic gram-negative bacterium found in the chronic adult inflammatory gum disease periodontitis. Initially, 370 ORFS were identified [23]; of these, 120 protein sequences were open to immune surveillance and 40 were positive for several sera. Two antigens were found to be protective in mice.

Yet another fascinating instance is provided by *Streptococcus pneumoniae*, a prime cause of meningitis, pneumonia, and sepsis [24, 25]. In this study, 130 potential ORFs were initially identified, with 108 of these proteins being readily expressed. Finally, six proteins were seen to induce protection against the pathogen.

More recently, other and more advanced experimental techniques, such as microarrays, are beginning to come on-stream, opening up a gallimaufry of possible technologies to the new but maturing field of reverse vaccinology. The following gives but a taste of what is to come.

Using ribosome display to undertake in-vitro protein selection, Weichart et al. [26] identified within the methicillin-resistant COL strain of the virulent human pathogen *Staphylococcus aureus* 75 genes, the majority of which were secreted or surface-localized proteins; of these, 25 % had cell envelope function, 24 % were transporter proteins, and 9 % were virulence factors or toxins.

Using an ingenious combination of advanced proteomics techniques and in-vitro assays, Giefing et al. [27] identified 18 novel vaccine candidates which prevented infections in children and in the elderly caused by a variety of pneumococcus serotypes; four demonstrating major protection versus sepsis in animals. Two leads—StkP (a serine/threonine protein kinase) and PcsB (a structural protein with a role in cell wall separation of group B *Streptococcus*)—showed clear cross-protection as potential candidate vaccines against four separate pneumococcal serotypes.

Using a whole proteome microarray, and in order to identify protein antigens, Eyles et al. [28] probed serum from BALB/c mice previously immunized with a vaccine comprising: killed *Francisella tularensis* and two immunomodulatory adjuvants. Eleven out of the top twelve immunogenic antigens were known already as immunoreactive, although 31 further proteins were discovered using this experimental approach. In further work from this consortium, Titball and co-workers [29] constructed a protein microarray of 1,205 *Burkholderia pseudomallei* proteins, treated it with 88 patient samples, identifying 170 antigens. This smaller set was treated with a further 747 distinct sera from 10 groups of patients, identifying 49 putative candidate antigens.

This survey, brief though it is, helps to highlight the potential power of reverse vaccinology for vaccine discovery. However, since the number of antigens is high, given all the potential difficulties in characterising and expressing them, it is important to note that both computational and experimental techniques and methodologies will doubtlessly omit important and interesting proteins from further analysis, though not necessarily for the same or similar reasons. Thus, with the burgeoning discipline of reverse vaccinology, both computational and experimental techniques are in need of constant development and improvement.

3.4 Immunoinformatics

Compared to its role to drug discovery, genomics, and a host of other bioscience sub-disciplines, bioinformatics support for the preclinical discovery and development of vaccine is in its infancy; yet, as interest in vaccine discovery increases, the situation changes. There are two key types of bioinformatics support for vaccine design, discovery, and development. At the technical level, the first of these cannot be properly or meaningfully distinguished from general support for target discovery. It includes the annotation of pathogen genomes, more conventional host genome annotation, and the statistical analysis of immunological microarray experiments. The second form of support concentrates on immunoinformatics, that is, the informatics analysis of immunological problems, principally epitope prediction.

B-cell epitope prediction remains defiantly basic or is largely dependent on a sometimes unavailable knowledge of three-dimensional protein structure. Both structure- [30] and data-driven [31] prediction of antibody-mediated epitopes evince poor results. However, methods developed to predict T-cell epitopes now possess considerable algorithmic sophistication. Moreover, they continue to develop and evolve, as well as extend their scope and remit to address new and ever larger and more challenging epitope prediction problems. Presently, accurate and reliable T-cell epitope prediction is restricted to predicting the binding of peptides to the major histocompatibility complex (MHC). Class I peptide-MHC prediction can be reasonably accurate, or is for properly characterised, well-understood alleles [32]. Yet a number of key studies have demonstrated that class

II MHC binding prediction is almost universally inaccurate, and is thus erratic and unreliable [33–35]. A similar situation persists for structure-driven prediction of MHC epitopes [36, 37].

Irrespective of poor predictive performance, several other problems exist for epitope prediction. For T cell prediction in particular, a prime concern is with the availability or rather lack of availability of relevant data. It is now known that immunogenic T cell epitopes, thought previously to be peptides no more than 10 amino acids in length, can be 16 or more residues long. Longmer epitopes now greatly expand the number of possible peptides open to inspection by T cells [38–41]. The inadequate results generated by B cell epitope prediction algorithms may indicate that a fundamental reinterpretation of extant B cell epitope data is necessary before improved methods become feasible.

These factors, when taken together, are consistent with the notion that methods relying only on the possession of certain epitopes will not be fully effective when tasked with antigen or immunogen identification. This is supported by information indicating a lack of correspondence between selected antigens and experimentally verified protective proteins.

3.5 Genomic-Level Identification of Antigens

There are many means of identifying antigenic proteins. Most focus on the properties of protein sequence and structure, but arguably one of the most insightful is instead to examine properties, both local and global, of the underlying nucleic acid. One notable way is to look for evidence of the horizontal or lateral transfer of so-called pathogenicity islands or PAIs. Horizontal transfer, such as transformation, conjugation, or transduction, is distinct from the vertical transfer of genetic material from an ancestor within its lineage. It typically involves an organism incorporating genetic material from an evolutionarily distant organism without being its offspring.

PAIs are a specific type of genomic island; that is, part of a genome acquired through direct transfer between microbes. A genomic island can occur in distantly related species and may be mono- or multi-functional; there are many sub-classes classified by function. Other examples include antibiotic resistance islands, metal resistance, and secretion system islands. The gene products of PAIs are crucial to the propagation of disease pathogenesis, much as the PAIs themselves are key to the evolution of pathogenesis. Pathogen-associated type III and type IV secretion systems are, for example, often found together in the same PAI.

Detecting such large ($>$10 Kb) and discrete clusters of genes clusters, habitually possessing a characteristically atypical G/C content, at least when compared with the remainder of the genome, leads, in turn, to the individual identification within clusters of virulence-associated protein antigens. Prokaryotic PAIs are frequently associated with tRNA-encoding genes, many are flanked by repeat structures, and many contain fragments of mobile genetic elements such as plasmids and phages.

PAIs can be identified by combining analysis of nucleotide composition and phylogeny, amongst others. Composition-based approaches rely on the natural variation between genome sequences from different species. Regions of the genome with abnormal composition, as demonstrated by nucleotide or codon bias, may be potentially transferred horizontally. Such methods are prone to inaccuracies; these result from inherent genomic sequence variation, such as is seen in highly expressed genes, and the observation that over time the sequences of genomic islands alter to mirror the composition of host genomes.

Evolution-based approaches seek regions that may have been transferred horizontally by comparing related species. Put at its simplest: a putative genomic island present in one species, but absent from several related species, is consistent with horizontal transfer. Of course, the island may have been present in the last common ancestor shared by the species compared and subsequently been lost from the other species. A less likely explanation would be that the island arose by mutation and selection in this species and no other. To decide, a body of extra evidence would need to be explored, such as the size of the PAI, the mechanistic ease of deletion, the consistent presence of the island in more distantly related species, the relative pathogenicity of island-less species, and the divergence of the genome relative to that of other related species.

Many methods, which seek to quantify and leverage these somewhat vague notions, are now available [42–44]. Such analysis at the nucleic acid level shares many features in common with approaches used to identify CpG islands in eukaryotic genomes [45–48]. Recently, Langille et al. tested six sequence-composition genomic island prediction methods and found that IslandPath-DIMOB and SIGI-HMM had the greatest overall accuracy [49].

Island Path was designed to help identify prokaryotic PAIs, through the visualisation of common PAI characteristics such as mobile element-associated genes or atypical sequence composition [50]. SIGI-HMM is a very accurate sequence composition-based genomic island predictor, which combines a Hidden Markov Model (HMM) and codon usage measurement to identify genomic islands [51].

In another work, Yoon et al. coupled heuristic sequence searching methods, which aimed simultaneously to identify PAIs and individual virulence genes, with composition and codon-usage bias [52]. Exploiting a machine learning approach, Vernikos and Parkhill sampled the structural features of genomic islands using a hypothesis-free, bottom-up search, with the objective of explicitly quantifying the contribution made by each feature to the overall structure of different genomic islands [53]. Arvey et al. sought to identify large chromosomal regions with atypical features using a general divergence measureable to quantify the compositional difference between genomic segments [54]. IslandPick is a comparative genomic island predictor, rather than a composition-based approach, that can identify very probable genomic islands and very probable non-genomic islands within investigated genomes but does require that several phylogentically related genomes are available [49]. Observing PAIs as having a G + C composition closer to their host genome, Wang et al. used so-called genomic barcodes to identify PAIs.

These barcodes are based on the fact that the frequencies of 2-mers to 7-mers, and their reverse complement, are very stable across a whole genome when using a window size of over 1,000 bps and that this constituted a characteristic signature for genomes [55].

The ready detection of PAIs, as a tool in computational reverse vaccinology, has been greatly aided by the deployment of several web-based resources. A key example of a server that successfully integrates several accurate genomic island predictors is IslandViewer [56], which combines the methods: IslandPick [49], IslandPath [50], and SIGI-HMM [51] and is available at the URL: http://www.pathogenomics.sfu.ca/islandviewer/query.php. The GUI facilitates the visualisation of genomic islands and downloading of data at the gene and chromosome levels in a variety of formats.

Another important, web-accessible resource is PAIDB or the PAI database. This is a wide-ranging database of PAIs, containing 112 distinct PAIs and 889 GenBank accessions present in 497 strains of pathogenic bacteria [57]. PAIDB may be accessed via the URL: http://www.gem.re.kr/paidb.

Thus, alternative techniques and methodologies are required in order to select and to rank proteins likely to be protective antigens and thus candidate vaccines. Below, we shall explore three key approaches: subcellular location prediction, alignment-dependent sequence similarity searching, and alignment-independent empirical statistical approaches.

3.6 Identifying Antigens Using Sequence Similarity

In this section, we consider, perhaps, the clearest and cleanest way to identify potential new antigens in any microbial genome to alignment-dependent sequence similarity searching. There are two complimentary but distinct ways of identifying the immunogenicity of a protein from its sequence. One is to look for significant similarity to proteins of known immunogenicity. This idea seems so straightforward as to be almost facile. The other approach is somewhat less obvious conceptually but almost as straightforward logistically and involves seeking to identify antigens as proteins without discernible sequence similarity to any host protein. Let us turn to the first of these two alternatives.

Let us begin by stating or rather reiterating the obvious. If we know the sequence of an existing antigen or antigens, we can use sequence searching to find similar sequences in the target genome [58, 59]. Any candidate antigens selected by this process can then be selected for further verification and validation. The same old, familiar caveats apply here: are chosen thresholds appropriate? Are high-scoring matches an artefact or are they real and meaningful? The litany of such conditions is all too familiar to anyone well versed in sequence similarity searching. Clearly, when a sequence search is run, using BLAST or FASTA3, for example, an enormously long list of nearly identical proteins might ensue, or one that does not get any hits at all, or almost any intervening result might be obtained. As reflective

practitioners, we must judge which result can be classified as useful and which cannot, and in so doing, identify sets of suitable thresholds, above which we expect usefulness and below which we might anticipate little or no utility. Thresholds are contingent upon the sequence family studied, as well as being dependent solely on the problem investigated. Thus heuristically identified cut-offs are desirable, but much thinking and empirical investigation are required to select appropriate values.

Of course, the process adumbrated above presupposes that sufficient antigenic protein sequences are known. Compilation of this data is the role of the database. Recently, extensive literature mining, coupled with factory-scale experimentation, has created many functional immunology databases, although databases, such as SYFPEITHI [60, 61], focussing on cellular immunology—primarily MHC processing, presentation, and T cell recognition—have existed for 15–20 years. Arguably, the best extant database is the HIV molecular immunology database [62], although clearly the depth of the database is at the expense of generality and breadth. Other recent databases include MHCBN [63, 64] and EPIMHC [65], amongst many others. Two databases, warrant particular attention: AntiJen [66], formerly known as Jenpep [67, 68]; and IEDB [69].

Implemented as a relational postgreSQL database, AntiJen integrates a wide-ranging set of data items, much of which is not stored by other databases. In addition to the kind of cellular immunological information familiar from SYFPEITHI, such as MHC binding and T cell data, AntiJen additionally archives B cell epitopes and also includes a significant stockpile of quantitative data: kinetic, thermodynamic, as well as functional, including measurements of immunological peptide–protein and protein–protein interactions. The IEDB database is considerably more extensive than other equivalent database systems, benefiting from the input of 13 dedicated epitope sequencing projects. IEDB has come to eclipse other work in this area. Although both AntiJen and IEDB are full of epitope-focussed information of many flavours, they remain incomplete concerning immunogenic antigens. Fortuitously, specific antigen-orientated—rather than epitope-focussed—databases are starting to be available.

Arguably, the most obvious and most unambiguous example of an antigen is virulence factor (VF): proteins, such as toxins, able to induce disease directly by attacking a host. Analysis of known pathogens has allowed recurring VF systems of 40+ distinct proteins. Often, sets of VFs exist as discrete, distinct genome-encoded PAIs, as well as being more widely spread through the genome.

Clearly, antigens do not need to be VFs in order to be immunogenic and thus candidates for subunit vaccines. Instead, they need only be accessible to the immune system. They do not need to directly or indirectly mediate infection. Thus, other databases are needed which capture, collate, and archive the burgeoning plethora of antigen-orientated data. Recently, we have helped developed a very different database: AntigenDB [70]. It contains over 500 antigens collated from the primary scientific literature, as well as other sources. Another related database system has been christened VIOLIN (vaccine investigation and online information network) [71], which allows straightforward curation and the analysis and

comparison of research data across diverse pathogens in the context of human medicine, animal models, laboratory model systems, and natural hosts.

As we outline above, in addition to identifying sequence similarity to known antigens, another idea gaining ground is that the immunogenicity of an antigen is solely determined by the absence of similarity to host proteins. Some think this is the prime determinant of potential protein immunogenicity [72, 73]. Such ideas are supported by the belief that immune systems are actively educated to lack reactivity to self-proteins [74], a process—often termed "immune tolerance"—which is generated via epitope-specific mechanisms [75, 76].

What we really want is a meaningful measure of the "foreignness" of a protein correlating with its immunogenicity. Usually, "evolutionary distance" substitutes for "foreignness." Clearly, such an evolutionary distance must be specified in terms of biomacromolecular structures or sequences. But, is this practically useful for selecting candidate vaccines?

Another way to formulate this idea is to say that the probability that a protein is immunogenic is exclusively a product of its dissimilarity, at the whole-sequence or sequence-fragment level, to each and every protein contained within the host proteome. Most search software is well matched to this problem. In terms of fragment length, the typical length of an epitope might seem logical, since the epitope is the molecular moiety typically recognised during the initial phase of an immune response. Yet, even at the epitope level—say a peptide of 8–16 amino acid residues—even a single conservative mutation or mismatch in an otherwise identical match might prove significant. Single sequence alterations may totally abrogate or significantly enhance neutralising antibodies binding or recognition by the machinery of cellular immunology.

We have attempted to benchmark sequence similarity and correlate it with immunogenicity in order to explore the potential of this idea in a quantitative fashion. To that end, we examined the differences between sets of antigens and non-antigen using sequence similarity scores. We looked specifically at sets of 100 known non-antigenic and 100 antigenic protein sequences from six sources: bacteria, viruses, fungi, and parasites, as well as allergens and tumours [77–79], comparing pathogen sequence to those from humans and mice using BLAST [80].

Most non-antigenic and antigenic sequences were non-redundant; implying a lack of homologues between pathogens and host proteomes, although certain parasite antigens, such as catalases and heat shock proteins, had a much greater level of similarity. We were not able to determine a suitable and appropriate threshold based on the hypothesis of non-redundancy to the host's proteome, suggesting that this is not a viable solution to vaccine antigen identification.

However, rather than looking at nucleic acid sequences, or at protein sequences using an alignment-based approach, a new set of techniques, based upon alignment-free techniques, has been and is being developed; as this approach begins to show significant potential, we shall examine it next.

3.7 Identifying Antigens through Subcellular Location Prediction

Proteins accessible to immune system surveillance are assumed to lie external to the microbial organism or be attached to its surface rather than being sequestered and sequestrated within the cell. For bacteria, this means being located on—or in—the outer membrane surface or being secreted. Thus, being able to accurately predict the physical location of a putative antigen can provide considerable insight into the likelihood that a particular protein will prove to be an immunogenic and possibly protective.

There are two basic kinds of prediction method for identifying subcellular location: manual rule construction and the application of data-driven machine learning methods. Data used to discriminate between compartments include sequence-derived features of the protein, such as hydrophobic regions; the amino acid composition of the whole protein; the presence of certain specific motifs; or a combination thereof. Accuracy differs significantly between different methods and different compartments, mostly resulting from the deficiency and inconsistency of data used to derive models. Gross overall sequence similarity is unable to predict protein sub-cellular location reliably or accurately. Even nearly identical protein sequences may be found in distinct locations, while there are many proteins which exist simultaneously at several distinct locations within the cell, often having equally distinct functions at these different sites [81].

Eukaryotes and prokaryotes have quite distinct subcellular compartments. The number of such compartments used in prediction studies varies. A common schema reduces prokaryotic to three compartments (cytoplasmic, periplasmic, and extracellular) and eukaryotic cells to four compartments (nuclear, cytoplasmic, mitochondrial, and extracellular). Other structural classifications evince in excess ten eukaryotic compartments. Ten compartments maybe a conservative estimate, such is the complex richness of sub-cellular structure. Any prediction method must account for permanent, transient, and multiple locations, and, in addition, multi-protein complexes and membrane-bound organelles as possible sites.

Numerous signal sequences exist. Several methods predict lipoproteins. The prediction of proteins translocated via the TAT-dependent pathway is important but has yet to be addressed properly. However, amongst binary, single-outcome approaches, SignalP is probably the most accurate and reliable method available. It uses neural networks to predict the presence and probable cleavage sites of type II or N-terminal Spase-I-cleaved secretion signal peptides [82–84]. This signal is common to both prokaryotic and eukaryotic organisms. SignalP has recently been enhanced with a HMM intended to discriminate cleaved from uncleaved signal anchors. A limitation of SignalP is its proclivity to over-predict: it cannot properly discriminate reliably between a number of very similar yet functionally different signal sequences, regularly predicting lipoproteins and integral membrane proteins as type II signals.

Many methods have been devised capable of dividing a genome or virtual-proteome between the various subcellular locations of a eukaryotic or prokaryotic cell. PSORT is a good example; it is a multicategory prediction procedure, comprising many different programmes [85–88]. PSORT I predicts 17 subcellular compartments, while PSORT II predicts ten different locations. iPSORT deals with several compartments: chloroplast, mitochondrial, and proteins secreted from the cell, while PSORT-B focuses solely on predicting bacterial sub-cellular locations.

Another effective programme is HensBC [89]. HensBC can assign gene products to one of four different types (nuclear, mitochondrial, cytoplasmic, or extracellular) with an accuracy of about eight out of ten for gram-negative bacteria. Another programme, SubLoc [90], predicts prokaryotic subcellular location divided between three compartments. Another programme is Gpos-PLoc [91], which integrates several basic classifiers. Other methods include Phobius [92], LipoP 1.0 [93], and TatP 1.0 [94]. A comparison of several such programmes, using 272 mycobacterial proteins as a gold standard [95], showed subcellular localisation prediction and possessed high predictive specificity.

We have developed a set of methods which predict bacterial subcellular location. Using a set of methods for lipoprotein, TAT secretion, and membrane protein prediction [96–102], three different Bayesian network architectures were implemented as software pipelines able to predict specific subcellular locations, and two serial implementations using a hierarchical decision structure, and a parallel implementation with a confidence-level-based decision engine [103]. The soluble-rooted serial pipeline performed better than the membrane-rooted predictor. The parallel pipeline outperformed the serial pipeline but was significantly less efficient. Genomic test sets proved more ambiguous: the serial implementation identified 22 more of the 74 proteins of known location yet more accurate predictions are made overall by the parallel implementation.

The implications of this work are clear. The complexity of subcellular structures must be integrated fully into sub-cellular location prediction. In extant studies, many important cellular organelles are not considered; different routes by which proteins can reach the same compartment are ignored; and proteins existing simultaneously at several locations are likewise discounted. Clearly, combining high specificity predictors for each compartment appropriately must be the way forward [103].

Many difficulties, problems, and quandaries persist; the most keenly felt is the lack of high-quality, verified, and validated datasets which unambiguously established the location of well-characterised proteins. This dearth is particularly serious for certain types of secreted protein, such as type III secretion. In a similar manner, considerably more work is required to accurately predict the locations for proteins of viral origin; while certain studies are encouraging [104, 105], the complexity of viral interaction with host organisms continues to confound attempts at analysis.

3.8 Identifying Antigens Using Alignment-Independent Methods

Predicting antigens in silico typically utilise bioinformatics tools. Such tools can identify signal peptides or membrane proteins or lipoproteins successfully, yet the majority of algorithms tend to depend on motifs characteristic of antigens or, more generally, sequence alignment as the principal arbiter of definitive and meaningful sequence relationships. This is potentially a problem of some magnitude, particularly given the wide range of evolutionary rates and mechanisms amongst microbial proteins. Certain protein families do not, however, show obvious or significant sequence similarity, despite having common biological properties, functions, and three-dimensional structures [106, 107].

Thus alignment-based approaches may not always produce useful and unequivocal results, since they assume a direct sequence relationship that can be identified by simple sequence search techniques. Immunogenicity, as a signature characteristic, may be encrypted within the structure and/or sequence instead. This may be encoded so cryptically or so subtlety as to completely confound or at least mislead conventional sequence alignment protocols. Discovery of utterly novel and previously unknown antigens will be totally stymied by the absence of similarity to known antigenic proteins.

Alignment-dependent methods tend to dominate bioinformatics and, by extension, immunoinformatics. Several authors have chosen to look at alternative strategies, implementing so-called alignment-independent or alignment-free techniques. The first authors to do so were Mayer et al., who reported that protective antigens had a different amino acid composition compared to control groups of non-antigens [108]. Such a result is unsurprising since it has long been known that the structure and sequence composition of proteins adapted to the different redox environments of different sub-cellular compartments [109].

Mayer's analysis was formulated primarily in terms of univariate comparisons of antigens versus controls for different properties. Subsequently, we explored bivariate comparison in terms of easily comprehensible scatter-plots. See Fig. 3.3 for representative examples. What their results ably demonstrate is the potential for the discrimination of antigens and non-antigens by the appropriate selection of orthogonal descriptors. The challenge, of course, is to identify a robust choice of descriptors which are capable of extrapolating as well interpolating when used predictively.

Progressing beyond this type of analysis, and synergising with our other work on alignment-independent representation [110–114], we have initiated the development of new methods to differentiate antigens—and thus potential vaccine candidates—and non-antigens, using more sophisticated alignment-free approach to sequence representation [115, 116]. Rather than focus on epitope versus non-epitope, our approach utilises data on protective antigens derived from diverse pathogens to create statistical models capable of predicting whole-protein antigenicity.

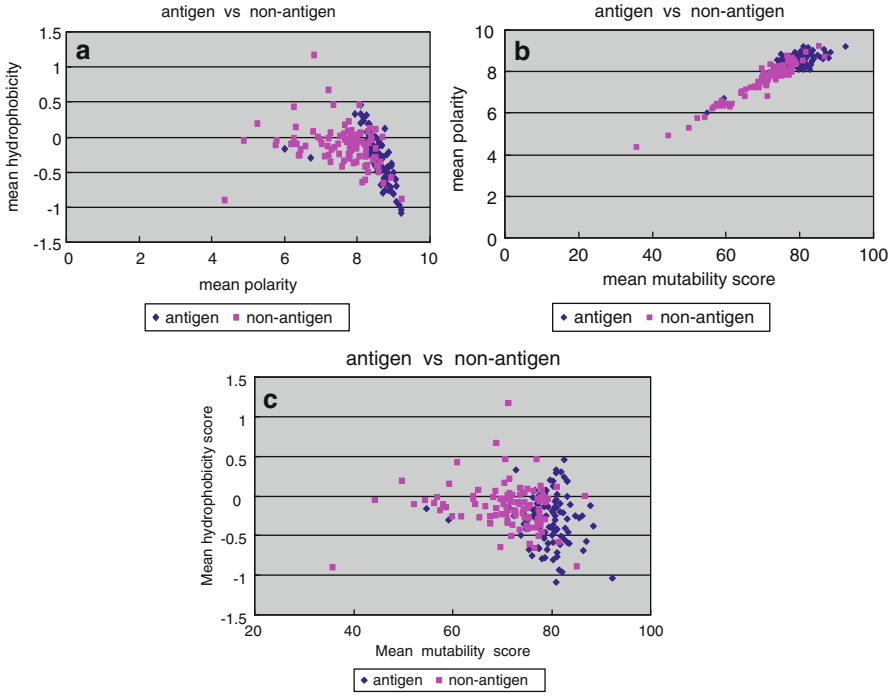

Fig. 3.3 Two scale plots of antigen and non antigen. (**a**) Proteins separated in terms of mean hydrophobicity versus mean polarity. (**b**) Proteins separated by mean polarity versus mean relative mutability. (**c**) Protein separated by mean hydrophobicity versus mean relative mutability score

Our alignment-independent method for antigen identification uses the auto cross covariance (ACC) transformation originally devised by Wold et al. [117, 118] to transform protein sequences into uniform vectors. The ACC transform has found much application in peptide prediction and protein classification [119–126]. In our method, amino acid residues are represented by the well-known and well-used z descriptors [127–129], which characterise the hydrophobicity, molecular size, and polarity of residues. Our method also accounts for the absence of complete independence between distinct sequence positions.

We initially applied our approach to groups of known viral, bacterial, and tumour antigens, developing models capable of identifying antigen. Extra models were subsequently added for fungal and parasite antigens. For bacterial, viral, and tumour antigens, models had prediction accuracies in the 70–89 % range [115, 116, 130]. For the parasite and fungal antigens, models had good predictive ability with 78–97 % accuracy. These models were incorporated into a server for protective antigen prediction called VaxiJen [115] (URL: http://www.darrenflower.info/VaxiJen). VaxiJen is an imperfect but encouraging start; future research will yield significantly more insight as well-characterised protective antigens increase significantly in number [70].

3.9 Antigen Selection and Immunogenicity

As we have said, a number of bioinformatics problems are unique to the discipline of immunology: the greatest of these is the accurate quantitative prediction of immunogenicity. This chapter has in its totality been suffused and pervaded by the idea of immunogenicity and the challenge of predicting this property in silico. Such an endeavour is confounding, yet exciting, and, as a key instrument in developing better, safer, more effective vaccines, is also of undisputed practical utility.

Successful immunogenicity prediction is at its simplest made manifest through the identification of B cell or T cell epitopes. Epitope recognition, when seen as a chemical event, may be understood in terms of the relationships between apparent biological function or activity and basic physicochemical properties. Delineating structure-activity or property-activity relationships of this kind is a key concern of immunoinformatics. At the other end of the spectrum, immunogenicity can be viewed is a cohesive, integrated, system property: a property of the entire and complete immune system and not a series of individual and isolated molecular recognition events. Thus, the task of predicting systems-level immunogenicity is in all likelihood manifold more demanding than predicting peptide-binding say.

The clinical manifestation of vaccine immunogenicity arises from the complex amalgam of many contributing extrinsic and intrinsic factors, which includes pathogen-side and host-side properties, as well as those just coming directly from proteins themselves. See Fig. 3.2. Protein-side properties include the aggregation state of candidate vaccines and the possession of PAMPs. Pathogen-side properties are clearly properties intrinsic to the pathogen, including expression levels of the antigen, the time-course of this expression, as well as its subcellular location. So-called host-side properties are innate recognition properties of host immunity, and most obviously include T cell epitopes or B cell epitopes.

A *bona fide* candidate antigen should be available for immune surveillance and thus highly expressed, constitutively or transiently, as well as having several epitopes. A protein without immunogenicity would logically lack all or some of these characteristics. As a prediction problem, this is, to say the least, not uncomplicated; clearly consisting of a great variety of difficult-to-compute stages. In terms of mechanism, many of these stages are poorly understood. Yet, each can be addressed using standard computational and statistical tools. They can all be predicted, however, presupposing, of course, the presence of relevant data in sufficient quantity.

3.10 Expert Systems for Antigen Discovery

One of the strongest messages to emerge from this review is that immunogenicity is a strongly multi-factorial property: some protein antigens are immunogenic for one reason, or set of reasons, and other immunogenic proteins will be so for another possibly tangential reason or set of reasons. Each such causal manifold is itself

complex and potentially confusing. Thus, the prediction of immunogenicity is a problem in multi-factorial prediction, and the search for new antigens is a search through a multi-factorial landscape of contingent causes and discombobulating decoys.

Some of the evidence will be highly precise and quantitative. The kind provided by predictive immunoinformatics, for example. This typically yields exact values for, say, the binding affinity of a peptide to a protein component of the immune system, or an unequivocal yes or no answer to the question: is this peptide sequence an epitope? However, for each such exact prediction, we have some notional associated probability concerning how reliable we regard this result. Different methods evince a range of accuracy, which, in practice, equate to probabilities of reliability: we naturally have more confidence and assume a greater reliability for a highly accurate prediction versus one of average predictability, though it can still give wrong predictions and generally inaccurate predictors may work well for a specific subset of the data.

Other types of forms of evidence will have a distinctly more anecdotal flavour. Take, for example, the case of bacterial exotoxins. Together with endotoxins, such as LPS, and so-called superantigens, exotoxins form the principal varieties of toxin secreted by pathogenic bacteria. Exotoxins have evolved to be the most toxic substances known to science: in terms of the median lethal dose, botulinum toxin—the active ingredient of BOTOX and causative agent of botulism, amongst others—is about ten times as lethal as radioactive isotope polonium-210 and a million times more deadly than mainline poisons, such as arsenic or potassium cyanide. Virtually, all such potent bacterial exotoxins comprise two functionally distinct subunits, either separate proteins or distinct domains, usually denoted A and B. The A subunit is habitually an enzyme, such as a protease, which modifies specific protein targets, thus disrupting key cellular processes with host cells. The B subunit is a protein which binds to host cell surface lipids or proteins, enabling the toxin to be internalised efficiently. The high specificity of this dual action lends exotoxins much of their remarkable lethality.

Exotoxins are also extremely immunogenic, inducing the immune systems to produce high-affinity neutralising antibodies against them, and thus make excellent targets for vaccinology. A toxoid—a toxin which has been treated or inactivated, often by formaldehyde—is in essence a form of subunit vaccine and, as such, requires adjuvant to induce adequate immune responses. Vaccines targeting tetanus and diphtheria, which usually need boosting every decade, are based on toxoids, albeit typically combined with pertussis toxin acting as an adjuvant. Poisoning by exotoxins, on the other hand, requires treatment with antitoxin comprising pre-formed antibodies.

However, and say that we were offered a newly sequenced pathogen genome, is such a classification for AB toxins helpful when trying to identify a potential exotoxins? The answer is neither yes nor is it no, but lies somewhere between these extremes. Assuming we had extant knowledge or a reliable method predicting the presence of structural and functionally distinct domains, this very simple rule-of-thumb would become a useful tool for eliminating large numbers of possible

toxin molecules. It would not directly identify an antigen but would enormously reduce the workload inherent in their discovery.

As well as needing more and more reliable predictors, we also need a way of combining the information we gather from any set of reliable predictors to which we have access. Thus, when analysing a pathogen genome, what we seem to need, at least in order to identify immunogenic proteins, is both a set of reliable and robust tools and a cohesive expert system within which to embed them. Such systems, albeit still at a relatively crude and faltering level, do exist. Because there is an implicit hierarchy of one prediction being based on others, there is a need to balance and judge different pieces of probabilistic evidence. An effective expert system should be capable of such a feat.

To a first approximation, an expert system is a computer programme that undertakes tasks that might otherwise be prosecuted by a human expert ostensively by simulating the apparent judgement and behaviour of an individual or organization with expertise and experience within a particular discipline. An Expert System might make financial forecasts, or play chess; it might diagnose human illnesses or schedule the routes of delivery vehicles. To create an expert system, one first needs to analyse human experts and how they make decisions, before translating this into rules that a computer can follow. Such a system leverages both a knowledge base of accumulated expertise and a set of rules for applying such distilled knowledge to particular situations in order to solve problems. Sophisticated expert systems can be updated with new knowledge and rules and can also learn from the success of its prediction, again mirroring the behaviour of properly performing experts.

At the heart then of an Expert System is the need to combine evidence in order to reach decisions. Combining evidence, and reaching a decision based on that combined evidence, is no easier in the laboratory, be that virtual or actual, than it is in the court room. The problem of combining evidence is encountered across the disciplines, and various solutions have arisen in these different areas.

Within bioinformatic prediction, a particular variety of evidence combination, so-called meta-prediction, is a now a well-established strategy [131, 132]. This approach seeks to amalgamate the output of various predictors, typically internet servers, in an intelligent way so that the combined result is more accurate than any of those coming from a single predictor. Indeed, combining results from multiple prediction tools does often increase overall accuracy. A consensus strategy was first proposed by Mallios [133], who combined SYFPEITHI [60, 61, 134], ProPred [135, 136], and the iterative stepwise discriminant analysis meta-algorithm [137–139]. MULTIPRED [140] integrates HMMs and artificial neural networks (ANN). Six MHC class II predictors were combined by Dai and co-workers [141–143] basing its overall prediction on the probability distributions of the different scores. Trost et al. have used a heuristic method to address class I peptide-MHC binding [144]. Wang et al. [145] applied a consensus method to calculate the median rank of the top three predictive methods for each MHC class II protein initially evaluated so as to rank all possible 8-, 9-, and 10-mers from one protein. This rank was used to identify the top 1 % of peptides from each protein.

In probabilistic reasoning, or reasoning with uncertainty, there are many ways to represent espoused beliefs—or, in our domain, predictions—that effectively encode the uncertainty of propositions. These include fuzzy logic and the evidential method, among many others. For quantitative data, information fusion, in its various guises [146], is one robust route to effective combination. Another requires us to enter the world of Bayesian statistics, or, at least, a special thread within it.

Bayes theory, and the ever-expanding strand of statistics devolving from it, is concerned primarily with updating or revising belief in the light of new evidence, while so-called Dempster–Shafer theory [147] is concerned not with the conditional probabilities of Bayesian statistics but with the direct combination of evidence. It extends the Bayesian theory of subjective probability, by replacing Bayesian probabilities with belief functions that describe degrees of belief for one question in terms of probabilities for another and then combines these using Dempster's rule for merging degrees of belief when based on independent lines of evidence. Such belief functions may or may not have the mathematical properties of probabilities but are seemingly able to combine the rigor of probability theory with the flexibility of rule-based approaches.

Several Expert Systems of different flavours and hues have now become available within the vaccinology arena. Sundaresh et al. developed a specialist software package for the analysis of microarray experiments that could easily be classified as an Expert System and used it in the area of reverse vaccinology. This package, which was written in the open-source statistical package R, was used to help analyse a variety of complex microarray experiments on the bacteria *F. tularensis*, a category A bio-defense pathogen [148]. This programme implements a two-stage process for diagnostic analysis: selection of antigens based on significant immune responses coupled with differential expression analysis, followed by classification of measured antigen responses using a combination of k-Means clustering, support vector machines, and k-nearest neighbours.

We have already discussed VaxiJen [115, 116, 130], and the related server EpiJen [149], which combines various methods for identifying epitopes within extant proteins. These two servers can also be classified as vaccine-related Expert Systems. NERVE is another Expert System, which has been developed to help automate aspects of reverse vaccinology [150]. Using NERVE, the prioritisation of potential candidate antigens consists of several stages: prediction of subcellular localisation; is the antigen an adhesion?; identification of membrane-crossing domains; and comparison to pathogen and human proteomes. Candidates are filtered then ranked and putative antigens graded by provenance and its predicted immunogenicity.

The web-based Expert System, DyNAVacS [151], was developed to facilitate the efficient design of DNA vaccines and is available in the URL: http://miracle.igib.res.in/dynavac. It takes a structured approach for vaccine design, leveraging various key design parameters, including the choice of appropriate expression vectors, safeguarding efficient expression through codon optimization, ensuring high levels of translation by adding specific sequence signals, and engineering of CpG motifs as adjuvant mechanisms exacerbating immune responses. It also allows

restriction enzyme mapping, the design of primers, and lists vectors in use for known DNA vaccines.

VAXIGN is another Expert System developed to help facilitate vaccine design [152]. VAXIGN undertakes dynamic vaccine target prediction from sequence. Methodologically, it combines protein subcellular location prediction with prediction of transmembrane helices and adhesins, analysis of the conservation to human and/or mouse proteins with sequence exclusion from the genomes of non-pathogenic strains, and prediction of peptide binding to class I and class II MHC. As a test, VAXIGN has been used to predict vaccine candidates against uropathogenic *Escherichia coli*.

However, NERVE and its various and varied siblings are tasked with such a confounding and difficult undertaking that they are obliged to fall somewhat short of what is required. An obvious first step in tackling the greater problem is to address first subcellular location prediction. Then, we can look at antigen presentation, modelling for each component step, before building these into a fully functional model. We can also develop empirical approaches—such as VaxiJen [115, 116, 130]. We must also factor in antibody-mediated issues, properly address PAMPs, post translational danger signals, expression levels, the role of aggregation, and the capacity of molecular adjuvants to enhance the innate immunogenicity to usable levels. See Fig. 3.2.

3.11 Discussion and Conclusions

The value of vaccines is not yet unchallenged. However, most reasonable people would, in all probability, agree that they are a good thing, albeit with a few minor provisos. The idea underlying all vaccines is a strong and robust one: it is in the reification—that is, the realisation, manifestation, and instantiation—of this abstract concept that the trouble lies, if indeed trouble there is. Existing vaccines are by no means perfect; again, most sensible and well-informed people would no doubt acknowledge this also. One might argue that their intrinsic complexity, and the highly empirical nature of their discovery over decades, and the fraught nature of their manufacture, has much to answer in this regard.

Why should this be? In part, it is due to the extreme complexity of immune response to an administered vaccine, which is largely specific to each individual or at least is different in different sub-groups within the totality of the vaccinated population. The immune responses is comprised, at least for whole-pathogen vaccines, of the adaptive immune response to multiple B cell and T cell epitopes as well as the responses made by the innate immune responses to diverse molecular structures, principally PAMPs. When one considers also the degree to which such a repertoire of responses is augmented and modified by the action of additives, be they designed to increase the durability and stability of vaccines or be they adjuvants, which are intended to raise the level of immune reactions. Add in stochastic and coincidental phenomena, such as reversion to

pathogenicity, and we can see immediately that navigating our way through the vaccine minefield is no easy task. All such problems engendered by this intrinsic complexity are themselves compounded by our comparatively weak understanding of immunological mechanisms, since, if we understood the mechanism of responses well enough, we could and would have designed our vaccines to circumvent these issues.

Part of the answer to this cacophony of conflicting and confounding quandaries is the newly emergent discipline of vaccinomics. A proper understanding of the relationships between gene variants and vaccine-specific immune responses may help us to design the next generation of personalised vaccines. Vaccinomics addresses this issue directly. It seeks to identify genetic factors mediating or moderating vaccine-induced immune responses, which are known to be extremely variable within population. Much data indicate that host genetic polymorphisms are key determinants of innate and adaptive response to vaccination. HLA genes, non-HLA genes, and genes of the innate immunity all contribute, and do so in many ways, to the variation observed between individuals for immune responses to microbial vaccines. Vaccinomics offers many techniques that can help illuminate these diverse phenomena. Principal amongst these are population-based gene/SNP association studies between allele or SNP variation and specific responses, supplemented by the application of next-generation sequencing technology and microarray approaches.

Yet, and for all this nay-saying and gainsaying, vaccines and vaccination have demonstrated their worth time after time; yet, to justify the continuing faith we invest in them, new and better ways of making safer and more focussed vaccines must be found. Most current vaccines work via antibody-mediated mechanisms; and most target viruses and the diseases they cause. Unfortunately, the stock of such disease targets is dwindling. Low-hanging fruit has long since been cut down. Only fruit that is well out of reach remains. Vaccines based on APCs and peptides are new but unproven strategies; most modern vaccine development relies instead on effective searches for vaccine antigens.

One of the clearest points to emerge from such work is that there are many competing concepts, thoughts, and ideas that may confound or help efficient identification of immune reactive proteins. Certain such ideas we have outlined. Some are indisputably persuasive, even compelling, yet many strategies—and the technical approaches upon which they are based—have singly failed to deliver on their promise.

Long ago, and based on his lifetime's experience of all things immunological, Professor Peter CL Beverley sketched out a paradigm for protein-focussed vaccine development, which we have formalised further, and which schema is summarised in Fig. 3.4. Some of his factors overlap with the factors from Fig. 3.2. He identified many of the factors that potentially contribute to the immunogenicity of proteins, be they of pathogen origin or another source entirely, and also other features which might make proteins particularly suitable for becoming candidate vaccines. Of these, some are as-yet beyond prediction, such as the attractiveness for APCs or the inability to down-regulate immune responses. The status of proteins as evasins is

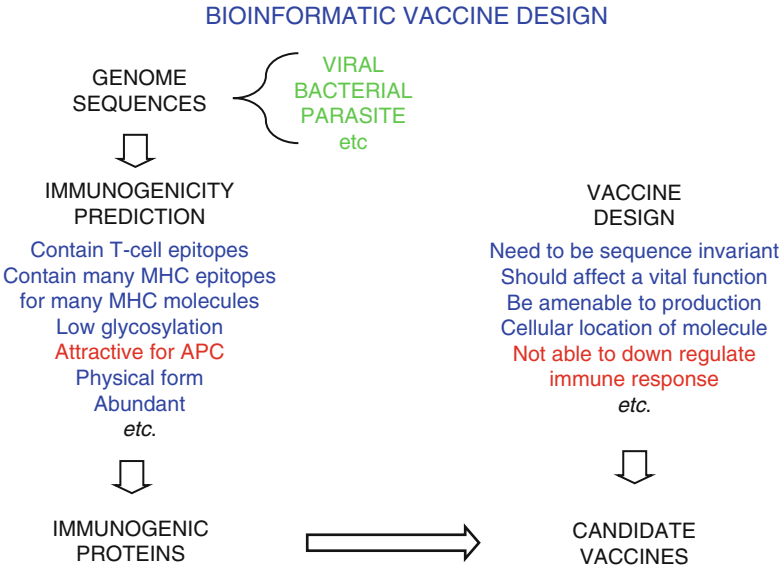

Fig. 3.4 The Beverley paradigm

currently only possibly addressable through sequence similarity-based approaches and likewise for the attractiveness for uptake by APCs is again, though possible there exist motifs, structural or sequence, which could be identified. Currently, the dearth of relevant data precludes prediction of such properties; and, while it is possible to predict some of these properties with some assurance of success, and others are predictable but only incidentally, overall, we are still some way from realising the dream embodied in Fig. 3.4.

Failure occurs for simple reasons: we deal with simplified abstractions and cannot hope to capture all that which is required for prediction by looking superficially at a single factor. Protein immunogenicity comes instead from the dynamic combination of innumerable contributing factors. This is by no means a facile or easily solved informatics conundrum. A vaccine candidate should have epitopes that the host recognises, be available for immune surveillance, and be highly expressed. Factors mediating protein immunogenicity are many; possession of B or T cell epitopes, post-translational danger signals, sub-cellular location, protein expression levels, and aggregation state amongst them. Predicting such diverse, complex, confounding properties is—and remains—a challenge.

Vaccine antigens, once discovered, should, ultimately, and with appropriate manipulation, together with an apt, apposite, and appropriate delivery system and the right choice of adjuvant, become first a candidate for clinical trials, before, hopefully, progressing to regulatory approval. We require an integrative, systems-biology approach to solve this problem. No single approach can be applied universally and with success; what we crave is the full integration of numerous equally

partial yet equally valid techniques and strategies which, in turn, draw upon a wealth of relevant, useful data. With an issue of such importance, even an incomplete solution should be sufficient.

References

1. Godlee F, Smith J, Marcovitch H (2011) Wakefield's article linking MMR vaccine and autism was fraudulent. BMJ 342:c7452
2. Flower DR, Davies MN, Ranganathan S: **Bioinformatics for Immunomics**, vol. 3, 1 edn: Springer; 2010.
3. Vivona S, Gardy JL, Ramachandran S, Brinkman FS, Raghava GP, Flower DR, Filippini F (2008) Computer-aided biotechnology: from immuno-informatics to reverse vaccinology. Trends Biotechnol 26(4):190–200
4. Davies MN, Flower DR (2007) Harnessing bioinformatics to discover new vaccines. Drug Discov Today 12(9–10):389–395
5. Lambert PH, Hawkridge T, Hanekom WA (2009) **New vaccines against tuberculosis**. Clin Chest Med 30(4):811–826, x
6. Flower D: **Bioinformatics for Vaccinology**, 1st edn: Wiley; 2008.
7. Plotkin SA (2001) Lessons learned concerning vaccine safety. Vaccine 20(suppl 1):S16–S19, discussion S11
8. Kwok R (2011) Vaccines: the real issues in vaccine safety. Nature 473(7348):436–438
9. Leask J (2011) Target the fence-sitters. Nature 473(7348):443–445
10. Day A (2009) **'An American tragedy'. The Cutter incident and its implications for the Salk polio vaccine in New Zealand 1955–1960**. Health History 11(2):42–61
11. Offit PA (2005) The Cutter incident, 50 years later. N Engl J Med 352(14):1411–1412
12. Nathanson N, Langmuir AD (1995) The Cutter incident. Poliomyelitis following formaldehyde-inactivated poliovirus vaccination in the United States during the Spring of 1955. II. Relationship of poliomyelitis to Cutter vaccine. 1963. Am J Epidemiol 142 (2):109–140, discussion 107–108
13. Flower DR (2008) Bioinformatics for vaccinology. Wiley, Chichester
14. Minor P (2009) Vaccine-derived poliovirus (VDPV): impact on poliomyelitis eradication. Vaccine 27(20):2649–2652
15. Flower DR (2009) Advances in predicting and manipulating the immunogenicity of biotherapeutics and vaccines. BioDrugs 23(4):231–240
16. Bambini S, Rappuoli R (2009) The use of genomics in microbial vaccine development. Drug Discov Today 14(5–6):252–260
17. Serruto D, Rappuoli R (2006) Post-genomic vaccine development. FEBS Lett 580 (12):2985–2992
18. Mora M, Donati C, Medini D, Covacci A, Rappuoli R (2006) Microbial genomes and vaccine design: refinements to the classical reverse vaccinology approach. Curr Opin Microbiol 9 (5):532–536
19. Serruto D, Adu-Bobie J, Capecchi B, Rappuoli R, Pizza M, Masignani V (2004) Biotechnology and vaccines: application of functional genomics to Neisseria meningitidis and other bacterial pathogens. J Biotechnol 113(1–3):15–32
20. Tettelin H, Saunders NJ, Heidelberg J, Jeffries AC, Nelson KE, Eisen JA, Ketchum KA, Hood DW, Peden JF, Dodson RJ et al (2000) Complete genome sequence of Neisseria meningitidis serogroup B strain MC58. Science 287(5459):1809–1815
21. Pizza M, Scarlato V, Masignani V, Giuliani MM, Arico B, Comanducci M, Jennings GT, Baldi L, Bartolini E, Capecchi B et al (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. Science 287(5459):1816–1820

22. Giuliani MM, Adu-Bobie J, Comanducci M, Arico B, Savino S, Santini L, Brunelli B, Bambini S, Biolchi A, Capecchi B et al (2006) A universal vaccine for serogroup B meningococcus. Proc Natl Acad Sci USA 103(29):10834–10839
23. Ross BC, Czajkowski L, Hocking D, Margetts M, Webb E, Rothel L, Patterson M, Agius C, Camuglia S, Reynolds E et al (2001) Identification of vaccine candidate antigens from a genomic analysis of Porphyromonas gingivalis. Vaccine 19(30):4135–4142
24. Wizemann TM, Heinrichs JH, Adamou JE, Erwin AL, Kunsch C, Choi GH, Barash SC, Rosen CA, Masure HR, Tuomanen E et al (2001) Use of a whole genome approach to identify vaccine molecules affording protection against Streptococcus pneumoniae infection. Infect Immun 69(3):1593–1598
25. Maione D, Margarit I, Rinaudo CD, Masignani V, Mora M, Scarselli M, Tettelin H, Brettoni C, Iacobini ET, Rosini R et al (2005) Identification of a universal Group B streptococcus vaccine by multiple genome screen. Science 309(5731):148–150
26. Weichhart T, Horky M, Sollner J, Gangl S, Henics T, Nagy E, Meinke A, von Gabain A, Fraser CM, Gill SR et al (2003) Functional selection of vaccine candidate peptides from Staphylococcus aureus whole-genome expression libraries in vitro. Infect Immun 71 (8):4633–4641
27. Giefing C, Meinke AL, Hanner M, Henics T, Bui MD, Gelbmann D, Lundberg U, Senn BM, Schunn M, Habel A et al (2008) Discovery of a novel class of highly conserved vaccine antigens using genomic scale antigenic fingerprinting of pneumococcus with human antibodies. J Exp Med 205(1):117–131
28. Eyles JE, Unal B, Hartley MG, Newstead SL, Flick-Smith H, Prior JL, Oyston PC, Randall A, Mu Y, Hirst S et al (2007) Immunodominant Francisella tularensis antigens identified using proteome microarray. Proteomics 7(13):2172–2183
29. Felgner PL, Kayala MA, Vigil A, Burk C, Nakajima-Sasaki R, Pablo J, Molina DM, Hirst S, Chew JS, Wang D et al (2009) A Burkholderia pseudomallei protein microarray reveals serodiagnostic and cross-reactive antigens. Proc Natl Acad Sci USA 106(32):13499–13504
30. Ponomarenko JV, Bourne PE (2007) Antibody-protein interactions: benchmark datasets and prediction tools evaluation. BMC Struct Biol 7:64
31. Blythe MJ, Flower DR (2005) Benchmarking B cell epitope prediction: underperformance of existing methods. Protein Sci 14(1):246–248
32. Lafuente EM, Reche PA (2009) Prediction of MHC-peptide binding: a systematic and comprehensive overview. Curr Pharm Des 15(28):3209–3220
33. Gowthaman U, Agrewala JN (2008) In silico tools for predicting peptides binding to HLA-class II molecules: more confusion than conclusion. J Proteome Res 7(1):154–163
34. El-Manzalawy Y, Dobbs D, Honavar V (2008) On evaluating MHC-II binding peptide prediction methods. PLoS One 3(9):e3268
35. Lin HH, Zhang GL, Tongchusak S, Reinherz EL, Brusic V (2008) **Evaluation of MHC-II peptide binding prediction servers: applications for vaccine research**. BMC Bioinformatics 9(suppl 12):S22
36. Knapp B, Omasits U, Frantal S, Schreiner W (2009) A critical cross-validation of high throughput structural binding prediction methods for pMHC. J Comput Aided Mol Des 23 (5):301–307
37. Zhang H, Wang P, Papangelopoulos N, Xu Y, Sette A, Bourne PE, Lund O, Ponomarenko J, Nielsen M, Peters B (2010) Limitations of Ab initio predictions of peptide binding to MHC class II molecules. PLoS One 5(2):e9272
38. Tynan FE, Burrows SR, Buckle AM, Clements CS, Borg NA, Miles JJ, Beddoe T, Whisstock JC, Wilce MC, Silins SL et al (2005) T cell receptor recognition of a 'super-bulged' major histocompatibility complex class I-bound peptide. Nat Immunol 6(11):1114–1122
39. Tynan FE, Borg NA, Miles JJ, Beddoe T, El-Hassen D, Silins SL, van Zuylen WJ, Purcell AW, Kjer-Nielsen L, McCluskey J et al (2005) High resolution structures of highly bulged viral epitopes bound to major histocompatibility complex class I. Implications for T-cell receptor engagement and T-cell immunodominance. J Biol Chem 280(25):23900–23909

40. Burrows SR, Rossjohn J, McCluskey J (2006) Have we cut ourselves too short in mapping CTL epitopes? Trends Immunol 27(1):11–16
41. Ebert LM, Liu YC, Clements CS, Robson NC, Jackson HM, Markby JL, Dimopoulos N, Tan BS, Luescher IF, Davis ID et al (2009) A long, naturally presented immunodominant epitope from NY-ESO-1 tumor antigen: implications for cancer vaccine design. Cancer Res 69 (3):1046–1054
42. Guy L (2006) Identification and characterization of pathogenicity and other genomic islands using base composition analyses. Future Microbiol 1(3):309–316
43. Ou HY, Chen LL, Lonnen J, Chaudhuri RR, Thani AB, Smith R, Garton NJ, Hinton J, Pallen M, Barer MR et al (2006) A novel strategy for the identification of genomic islands by comparative analysis of the contents and contexts of tRNA sites in closely related bacteria. Nucleic Acids Res 34(1):e3
44. Ou HY, He X, Harrison EM, Kulasekara BR, Thani AB, Kadioglu A, Lory S, Hinton JC, Barer MR, Deng Z et al (2007) MobilomeFINDER: web-based tools for in silico and experimental discovery of bacterial genomic islands. Nucleic Acids Res 35:W97–W104, Web Server issue
45. Hackenberg M, Previti C, Luque-Escamilla PL, Carpena P, Martinez-Aroza J, Oliver JL (2006) CpGcluster: a distance-based algorithm for CpG-island detection. BMC Bioinformatics 7:446
46. Sujuan Y, Asaithambi A, Liu Y (2008) CpGIF: an algorithm for the identification of CpG islands. Bioinformation 2(8):335–338
47. Hutter B, Paulsen M, Helms V (2009) **Identifying CpG islands by different computational techniques**. OMICS 13(2):153–164
48. Su J, Zhang Y, Lv J, Liu H, Tang X, Wang F, Qi Y, Feng Y, Li X (2010) CpG_MI: a novel approach for identifying functional CpG islands in mammalian genomes. Nucleic Acids Res 38(1):e6
49. Langille MG, Hsiao WW, Brinkman FS (2008) Evaluation of genomic island predictors using a comparative genomics approach. BMC Bioinformatics 9:329
50. Hsiao W, Wan I, Jones SJ, Brinkman FS (2003) IslandPath: aiding detection of genomic islands in prokaryotes. Bioinformatics 19(3):418–420
51. Waack S, Keller O, Asper R, Brodag T, Damm C, Fricke WF, Surovcik K, Meinicke P, Merkl R (2006) Score-based prediction of genomic islands in prokaryotic genomes using hidden Markov models. BMC Bioinformatics 7:142
52. Yoon SH, Hur CG, Kang HY, Kim YH, Oh TK, Kim JF (2005) A computational approach for identifying pathogenicity islands in prokaryotic genomes. BMC Bioinformatics 6:184
53. Vernikos GS, Parkhill J (2008) Resolving the structural features of genomic islands: a machine learning approach. Genome Res 18(2):331–342
54. Arvey AJ, Azad RK, Raval A, Lawrence JG (2009) Detection of genomic islands via segmental genome heterogeneity. Nucleic Acids Res 37(16):5255–5266
55. Wang G, Zhou F, Olman V, Li F, Xu Y (2010) Prediction of pathogenicity islands in enterohemorrhagic Escherichia coli O157:H7 using genomic barcodes. FEBS Lett 584 (1):194–198
56. Langille MG, Brinkman FS (2009) IslandViewer: an integrated interface for computational identification and visualization of genomic islands. Bioinformatics 25(5):664–665
57. Yoon SH, Park YK, Lee S, Choi D, Oh TK, Hur CG, Kim JF (2007) Towards pathogenomics: a web-based resource for pathogenicity islands. Nucleic Acids Res 35:D395–D400, Database issue
58. Adamou JE, Heinrichs JH, Erwin AL, Walsh W, Gayle T, Dormitzer M, Dagan R, Brewah YA, Barren P, Lathigra R et al (2001) Identification and characterization of a novel family of pneumococcal proteins that are protective against sepsis. Infect Immun 69(2):949–958
59. Moxon ER, Hood DW, Saunders NJ, Schweda EK, Richards JC (2002) Functional genomics of pathogenic bacteria. Philos Trans R Soc Lond B Biol Sci 357(1417):109–116

60. Schuler MM, Nastke MD, Stevanovikc S (2007) SYFPEITHI: database for searching and T-cell epitope prediction. Methods Mol Biol 409:75–93
61. Rammensee H, Bachmann J, Emmerich NP, Bachor OA, Stevanovic S (1999) SYFPEITHI: database for MHC ligands and peptide motifs. Immunogenetics 50(3–4):213–219
62. Kuiken C, Korber B, Shafer RW (2003) HIV sequence databases. AIDS Rev 5(1):52–61
63. Lata S, Bhasin M, Raghava GP (2009) MHCBN 4.0: a database of MHC/TAP binding peptides and T-cell epitopes. BMC Res Notes 2:61
64. Bhasin M, Singh H, Raghava GP (2003) MHCBN: a comprehensive database of MHC binding and non-binding peptides. Bioinformatics 19(5):665–666
65. Reche PA, Zhang H, Glutting JP, Reinherz EL (2005) EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. Bioinformatics 21 (9):2140–2141
66. Toseland CP, Clayton DJ, McSparron H, Hemsley SL, Blythe MJ, Paine K, Doytchinova IA, Guan P, Hattotuwagama CK, Flower DR (2005) AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical, and cellular data. Immunome Res 1(1):4
67. McSparron H, Blythe MJ, Zygouri C, Doytchinova IA, Flower DR (2003) JenPep: a novel computational information resource for immunobiology and vaccinology. J Chem Inf Comput Sci 43(4):1276–1287
68. Blythe MJ, Doytchinova IA, Flower DR (2002) JenPep: a database of quantitative functional peptide data for immunology. Bioinformatics 18(3):434–439
69. Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, Damle R, Sette A, Peters B (2010) The immune epitope database 2.0. Nucleic Acids Res 38:D854–D862, Database issue
70. Ansari HR, Flower DR, Raghava GPS (2010) AntigenDB: an immunoinformatics database of pathogen antigens. Nucleic Acids Res 38:D847–D853
71. Xiang Z, Todd T, Ku KP, Kovacic BL, Larson CB, Chen F, Hodges AP, Tian Y, Olenzek EA, Zhao B et al (2008) VIOLIN: vaccine investigation and online information network. Nucleic Acids Res 36:D923–D928, Database issue
72. Kanduc D (2009) Epitopic peptides with low similarity to the host proteome: towards biological therapies without side effects. Expert Opin Biol Ther 9(1):45–53
73. Kanduc D (2005) Peptimmunology: immunogenic peptides and sequence redundancy. Curr Drug Discov Technol 2(4):239–244
74. Singh NJ, Schwartz RH (2006) Primer: mechanisms of immunologic tolerance. Nat Clin Pract Rheumatol 2(1):44–52
75. Miao CH (2007) Recent advances in immune modulation. Curr Gene Ther 7(5):391–402
76. Barron L, Knoechel B, Lohr J, Abbas AK (2008) Cutting edge: contributions of apoptosis and anergy to systemic T cell tolerance. J Immunol 180(5):2762–2766
77. Ramakrishnan K, Flower DR (2010) Discriminating antigen and non-antigen using proteome dissimilarity III: tumour and parasite antigens. Bioinformation 5(1):39–42
78. Ramakrishnan K, Flower DR (2010) Discriminating antigen and non-antigen using proteome dissimilarity II: viral and fungal antigens. Bioinformation 5(1):35–38
79. Ramakrishnan K, Flower DR (2010) Discriminating antigen and non-antigen using proteome dissimilarity: bacterial antigens. Bioinformation 4(10):445–447
80. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389–3402
81. Radisky DC, Stallings-Mann M, Hirai Y, Bissell MJ (2009) Single proteins might have dual but related functions in intracellular and extracellular microenvironments. Nat Rev Mol Cell Biol 10(3):228–234
82. Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protoc 2(4):953–971
83. Bendtsen JD, Nielsen H, von Heijne G, Brunak S (2004) Improved prediction of signal peptides: SignalP 3.0. J Mol Biol 340(4):783–795

84. Choo KH, Tan TW, Ranganathan S (2009) **A comprehensive assessment of N-terminal signal peptides prediction methods**. BMC Bioinformatics 10(suppl 15):S2
85. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K (2007) WoLF PSORT: protein localization predictor. Nucleic Acids Res 35:W585–W587, Web Server issue
86. Chen Y, Yu P, Luo J, Jiang Y (2003) Secreted protein prediction system combining CJ-SPHMM, TMHMM, and PSORT. Mamm Genome 14(12):859–865
87. Gardy JL, Spencer C, Wang K, Ester M, Tusnady GE, Simon I, Hua S, deFays K, Lambert C, Nakai K et al (2003) **PSORT-B: improving protein subcellular localization prediction for Gram-negative bacteria**. Nucleic Acids Res 31(13):3613–3617
88. Nakai K, Horton P (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. Trends Biochem Sci 24(1):34–36
89. Bulashevska A, Eils R (2006) Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. BMC Bioinformatics 7:298
90. Chen H, Huang N, Sun Z (2006) SubLoc: a server/client suite for protein subcellular location based on SOAP. Bioinformatics 22(3):376–377
91. Shen HB, Chou KC (2007) Gpos-PLoc: an ensemble classifier for predicting subcellular localization of Gram-positive bacterial proteins. Protein Eng Des Sel 20(1):39–46
92. Kall L, Krogh A, Sonnhammer EL (2007) Advantages of combined transmembrane topology and signal peptide prediction–the Phobius web server. Nucleic Acids Res 35:W429–W432, Web Server issue
93. Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H, Krogh A (2003) Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci 12(8):1652–1662
94. Bendtsen JD, Nielsen H, Widdick D, Palmer T, Brunak S (2005) Prediction of twin-arginine signal peptides. BMC Bioinformatics 6:167
95. Restrepo-Montoya D, Vizcaino C, Nino LF, Ocampo M, Patarroyo ME, Patarroyo MA (2009) Validating subcellular localization prediction tools with mycobacterial proteins. BMC Bioinformatics 10:134
96. Taylor PD, Attwood TK, Flower DR (2006) Toward bacterial protein sub-cellular location prediction: single-class discrimminant models for all gram- and gram+ compartments. Bioinformation 1(8):276–280
97. Taylor PD, Attwood TK, Flower DR (2006) Multi-class subcellular location prediction for bacterial proteins. Bioinformation 1(7):260–264
98. Taylor PD, Toseland CP, Attwood TK, Flower DR (2006) Alpha helical trans-membrane proteins: enhanced prediction using a Bayesian approach. Bioinformation 1(6):234–236
99. Taylor PD, Toseland CP, Attwood TK, Flower DR (2006) Beta barrel trans-membrane proteins: enhanced prediction using a Bayesian approach. Bioinformation 1(6):231–233
100. Taylor PD, Toseland CP, Attwood TK, Flower DR (2006) A predictor of membrane class: discriminating alpha-helical and beta-barrel membrane proteins from non-membranous proteins. Bioinformation 1(6):208–213
101. Taylor PD, Toseland CP, Attwood TK, Flower DR (2006) TATPred: a Bayesian method for the identification of twin arginine translocation pathway signal sequences. Bioinformation 1 (5):184–187
102. Taylor PD, Toseland CP, Attwood TK, Flower DR (2006) LIPPRED: a web server for accurate prediction of lipoprotein signal sequences and cleavage sites. Bioinformation 1 (5):176–179
103. Taylor PD, Attwood TK, Flower DR (2006) Combining algorithms to predict bacterial protein sub-cellular location: parallel versus concurrent implementations. Bioinformation 1 (8):285–289
104. Scott MS, Oomen R, Thomas DY, Hallett MT (2006) Predicting the subcellular localization of viral proteins within a mammalian host cell. Virol J 3:24
105. Shen HB, Chou KC (2007) Virus-PLoc: a fusion classifier for predicting the subcellular localization of viral proteins within host and virus-infected cells. Biopolymers 85(3):233–240

106. Flower DR, North AC, Attwood TK (1993) Structure and sequence relationships in the lipocalins and related proteins. Protein Sci 2(5):753–761
107. Flower DR (1993) Structural Relationship of Streptavidin to the Calycin Protein Superfamily. FEBS Lett 333(1–2):99–102
108. Mayers C, Duffield M, Rowe S, Miller J, Lingard B, Hayward S, Titball RW (2003) Analysis of known bacterial protein vaccine antigens reveals biased physical properties and amino acid composition. Comp Funct Genomics 4(5):468–478
109. Andrade MA, O'Donoghue SI, Rost B (1998) Adaptation of protein surfaces to subcellular location. J Mol Biol 276(2):517–525
110. Secker A, Davies MN, Freitas AA, Clark EB, Timmis J, Flower DR (2010) Hierarchical classification of G-protein-coupled receptors with data-driven selection of attributes and classifiers. Int J Data Min Bioinform 4(2):191–210
111. Davies MN, Secker A, Halling-Brown M, Moss DS, Freitas AA, Timmis J, Clark E, Flower DR (2008) GPCRTree: online hierarchical classification of GPCR function. BMC Res Notes 1:67
112. Davies MN, Secker A, Freitas AA, Clark E, Timmis J, Flower DR (2008) Optimizing amino acid groupings for GPCR classification. Bioinformatics 24(18):1980–1986
113. Davies MN, Secker A, Freitas AA, Mendao M, Timmis J, Flower DR (2007) On the hierarchical classification of G protein-coupled receptors. Bioinformatics 23(23):3113–3118
114. Davies MN, Gloriam DE, Secker A, Freitas AA, Mendao M, Timmis J, Flower DR (2007) Proteomic applications of automated GPCR classification. Proteomics 7(16):2800–2814
115. Doytchinova IA, Flower DR (2007) **VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines**. BMC Bioinformatics 8:4
116. Doytchinova IA, Flower DR (2007) Identifying candidate subunit vaccines using an alignment-independent method based on principal amino acid properties. Vaccine 25 (5):856–866
117. Wold S, Jonsson J, Sjostrom M, Sandberg M, Rannar S (1993) DNA and peptide sequences and chemical processes multivariately modeled by principal component analysis and partial least-squares projections to latent structures. Anal Chim Acta 277(2):239–253
118. Wold S, Eriksson L, Hellberg S, Jonsson J, Sjostrom M, Skagerberg B, Wikstrom C (1987) Principal property-values for 6 nonnatural amino-acids and their application to a structure activity relationship for oxytocin peptide analogs. Can J Chem 65(8):1814–1820
119. Dimitrov I, Garnev P, Flower DR, Doytchinova I (2010) Peptide binding to the HLA-DRB1 supertype: a proteochemometrics analysis. Eur J Med Chem 45(1):236–243
120. Kontijevskis A, Petrovska R, Yahorava S, Komorowski J, Wikberg JE (2009) Proteochemometrics mapping of the interaction space for retroviral proteases and their substrates. Bioorg Med Chem 17(14):5229–5237
121. Prusis P, Lapins M, Yahorava S, Petrovska R, Niyomrattanakit P, Katzenmeier G, Wikberg JE (2008) Proteochemometrics analysis of substrate interactions with dengue virus NS3 proteases. Bioorg Med Chem 16(20):9369–9377
122. Strombergsson H, Kryshtafovych A, Prusis P, Fidelis K, Wikberg JE, Komorowski J, Hvidsten TR (2006) Generalized modeling of enzyme-ligand interactions using proteochemometrics and local protein substructures. Proteins 65(3):568–579
123. Strombergsson H, Prusis P, Midelfart H, Lapinsh M, Wikberg JE, Komorowski J (2006) Rough set-based proteochemometrics modeling of G-protein-coupled receptor-ligand interactions. Proteins 63(1):24–34
124. Lapinsh M, Prusis P, Uhlen S, Wikberg JE (2005) Improved approach for proteochemometrics modeling: application to organic compound–amine G protein-coupled receptor interactions. Bioinformatics 21(23):4289–4296
125. Wikberg JE, Mutulis F, Mutule I, Veiksina S, Lapinsh M, Petrovska R, Prusis P (2003) Melanocortin receptors: ligands and proteochemometrics modeling. Ann N Y Acad Sci 994:21–26

126. Lapinsh M, Prusis P, Lundstedt T, Wikberg JE (2002) Proteochemometrics modeling of the interaction of amine G-protein coupled receptors with a diverse set of ligands. Mol Pharmacol 61(6):1465–1475
127. Hellberg S, Sjostrom M, Skagerberg B, Wold S (1987) Peptide quantitative structure-activity-relationships, a multivariate approach. J Med Chem 30(7):1126–1135
128. Jonsson J, Eriksson L, Hellberg S, Sjostrom M, Wold S (1989) Multivariate parametrization of 55 coded and non-coded amino-acids. Quant Struct Act Rel 8(3):204–209
129. Sandberg M, Eriksson L, Jonsson J, Sjostrom M, Wold S (1998) New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids. J Med Chem 41(14):2481–2491
130. Doytchinova IA, Flower DR (2008) Bioinformatic approach for identifying parasite and fungal candidate subunit vaccines. Open Vaccine J 1(1):4
131. Friedberg I, Harder T, Godzik A (2006) JAFA: a protein function annotation meta-server. Nucleic Acids Res 34:W379–W381, Web Server issue
132. Pawlowski M, Gajda MJ, Matlak R, Bujnicki JM (2008) MetaMQAP: a meta-server for the quality assessment of protein models. BMC Bioinformatics 9:403
133. Mallios RR (2003) A consensus strategy for combining HLA-DR binding algorithms. Hum Immunol 64(9):852–856
134. Dong HL, Sui YF (2005) Prediction of HLA-A2-restricted CTL epitope specific to HCC by SYFPEITHI combined with polynomial method. World J Gastroenterol 11(2):208–211
135. Mustafa AS, Shaban FA (2006) ProPred analysis and experimental evaluation of promiscuous T-cell epitopes of three major secreted antigens of Mycobacterium tuberculosis. Tuberculosis (Edinb) 86(2):115–124
136. Singh H, Raghava GP (2001) ProPred: prediction of HLA-DR binding sites. Bioinformatics 17(12):1236–1237
137. Mallios RR (2001) Predicting class II MHC/peptide multi-level binding with an iterative stepwise discriminant analysis meta-algorithm. Bioinformatics 17(10):942–948
138. Mallios RR (1999) Class II MHC quantitative binding motifs derived from a large molecular database with a versatile iterative stepwise discriminant analysis meta-algorithm. Bioinformatics 15(6):432–439
139. Mallios RR (1998) Iterative stepwise discriminant analysis: a meta-algorithm for detecting quantitative sequence motifs. J Comput Biol 5(4):703–711
140. Zhang GL, Khan AM, Srinivasan KN, August JT, Brusic V (2005) Neural models for predicting viral vaccine targets. J Bioinform Comput Biol 3(5):1207–1225
141. Huang L, Karpenko O, Murugan N, Dai Y (2007) Building a meta-predictor for MHC class II-binding peptides. Methods Mol Biol 409:355–364
142. Karpenko O, Huang L, Dai Y (2008) A probabilistic meta-predictor for the MHC class II binding peptides. Immunogenetics 60(1):25–36
143. Huang L, Karpenko O, Murugan N, Dai Y (2006) A meta-predictor for MHC class II binding peptides based on Naive Bayesian approach. Conf Proc IEEE Eng Med Biol Soc 1:5322–5325
144. Trost B, Bickis M, Kusalik A (2007) Strength in numbers: achieving greater accuracy in MHC-I binding prediction by combining the results from multiple prediction tools. Immunome Res 3:5
145. Wang P, Sidney J, Dow C, Mothe B, Sette A, Peters B (2008) A systematic assessment of MHC class II peptide binding predictions and evaluation of a consensus approach. PLoS Comput Biol 4(4):e1000048
146. Salim N, Holliday J, Willett P (2003) Combination of fingerprint-based similarity coefficients using data fusion. J Chem Inf Comput Sci 43(2):435–442
147. Basir O, Karray F, Zhu H (2005) Connectionist-based Dempster-Shafer evidential reasoning for data fusion. IEEE Trans Neural Netw 16(6):1513–1530

148. Sundaresh S, Randall A, Unal B, Petersen JM, Belisle JT, Hartley MG, Duffield M, Titball RW, Davies DH, Felgner PL et al (2007) From protein microarrays to diagnostic antigen discovery: a study of the pathogen Francisella tularensis. Bioinformatics 23(13):i508–i518
149. Doytchinova IA, Guan P, Flower DR (2006) EpiJen: a server for multistep T cell epitope prediction. BMC Bioinformatics 7:131
150. Vivona S, Bernante F, Filippini F (2006) NERVE: new enhanced reverse vaccinology environment. BMC Biotechnol 6:35
151. Harish N, Gupta R, Agarwal P, Scaria V, Pillai B (2006) DyNAVacS: an integrative tool for optimized DNA vaccine design. Nucleic Acids Res 34:W264–W266, Web Server issue
152. He Y, Xiang Z, Mobley HL (2010) Vaxign: the first web-based vaccine design program for reverse vaccinology and applications for vaccine development. J Biomed Biotechnol 2010:297505
153. Fell DA (2005) Enzymes, metabolites and fluxes. J Exp Bot 56(410):267–272