ELSEVIER

Harnessing bioinformatics to discover new vaccines

Matthew N. Davies and Darren R. Flower

The Jenner Institute, University of Oxford, Compton, Berkshire, RG20 7NN, UK

Vaccine design is highly suited to the application of *in silico* techniques, for both the discovery and development of new and existing vaccines. Here, we discuss computational contributions to epitope mapping and reverse vaccinology, two techniques central to the new discipline of immunomics. Also discussed are methods to improve the efficiency of vaccination, such as codon optimization and adjuvant discovery in addition to the identification of allergenic proteins. We also review current software developed to facilitate vaccine design.

Immunomics is a comparatively new discipline that uses high throughput techniques to explore immune mechanisms. It combines several fields, including informatics, genomics, proteomics, immunology and clinical medicine. To date, a key focus of immunomics has been the development of algorithms for the design and discovery of new vaccines. The success of a vaccine can be measured by its strength, its specificity, the duration of the immune response and its capacity to create immunological memory. Most extant vaccines are mediated by protective responses based on neutralizing antibodies, yet the protective mechanisms of remaining disease targets are mainly T cell based. Vaccines can be live attenuated whole pathogens, subunits, or be epitope-based. It is possible to design attenuated pathogens by removing virulence factors or reducing their metabolic capacity to replicate at speed, both of which are amenable to computational design and discovery. Subunit and epitope-based vaccines are, as we reveal here, also amenable to discovery through informatics, as is the discovery of adjuvants, without which non-whole-organism vaccines are insufficiently immunogenic to be effective. Several *in silico* techniques have been developed to identify suitable vaccine candidates, principally proteins within pathogen genomes that have antigenic properties. Here, we outline currently available techniques and software for vaccine discovery, and examples of how such algorithms can be applied.

Corresponding author: Flower, D.R. (darren.flower@jenner.ac.uk)

T cell and B cell epitope prediction

With an ever-increasing number of pathogen genomes fully or partially determined, there is a pressing need for *in silico* methods that can identify potential vaccine candidates. For a vaccine to be effective, it must invoke a strong response from both T cells and B cells; therefore, epitope mapping is a central issue in their design [1,2]. *In silico* prediction methods have the capacity to accelerate epitope discovery greatly. B cell and T cell epitope mapping has led to so-called 'epitope fishing', the scanning of pathogenic genomes for potential epitopes using predictive algorithms [2]. Complex microbial pathogens, such as *Mycobacterium tuberculosis*, can interact within the immune system in a multitude of ways [3]. There are >4000 proteins in the *M. tuberculosis* genome; this means that experimental analysis of host–pathogen interactions would be prohibitive in terms of time, labour and expense. T cell and B cell epitope mapping can help to define such inter-relationships by examining them on a protein-by-protein basis.

T cell epitopes are antigenic peptide fragments derived from a pathogen that, when bound to a major histocompatibility complex (MHC) molecule, interact with T cell receptors after transport to the surface of an antigen-presenting cell. If sufficient quantities of the epitope are presented, the T cell can trigger an adaptive immune response specific for the pathogen. MHC class I (MHC-I) and class II (MHC-II) molecules form complexes with different types of peptide. The MHC-I molecule binds to a peptide of 8–15 amino acids in length within a single closed groove. The peptide is secured largely through interactions with anchoring residues at the N- and C-termini of the peptide, whereas the central region is

1359-6446/06/$ - see front matter doi:10.1016/j.drudis.2007.03.010

more flexible [4]. MHC-II peptides vary in length from 12–25 amino acids, and are bound by the protrusion of peptide side chains into cavities within the groove and through a series of hydrogen bonds formed between the main chain peptide atoms and the side chain atoms of the MHC molecule [5]. Unlike the MHC-I molecule, where the binding site is closed at each end, the peptide can extend out of both open ends of the binding groove. Experimentally determined affinity data have been used to develop a variety of MHC binding prediction algorithms, which can distinguish binders from non-binders based on the peptide sequence. These include motif-based systems [4], support vector machines (SVMs) [6–8], hidden Markov models (HMMs) [9], neural networks [10], quantitative structure–activity relationship (QSAR) analysis [11] and structure-based approaches [12,13]. MHC binding motifs are a straightforward and easily comprehended method of epitope detection but produce many false positive and many false negative results [4]. SVMs are machine learning algorithms based on statistical theory that seeks to separate data into two distinct classes (in this case binders and non-binders) [4–6]. HMMs are statistical models where the system being modelled is assumed to be a Markov process with unknown parameters [9]. In a HMM, the internal state is not visible directly but variables influenced by the state are. HMMs aim to determine the hidden parameters from observable ones. A HMM profile can be used to determine those sequences with 'binder-like' qualities. Bayesian neural networks can also be applied to the problem as they are better suited to recognizing complicated peptide patterns than some other algorithms [10]. Bayesian neural networks, in particular, have several advantages: they are robust, difficult to overtrain, minimize the risk of overfitting, are tolerant of noisy or missing data, automatically find the least complex model that explains the data and can automatically optimize their architecture. An alternate approach has tried to develop supertypes, peptides that bind to several MHC alleles with similar specificity. Identifying a supertype peptide can result in the development of a vaccine capable of covering a significant percentage of the population, irrespective of ethnicity. Reche and Reinherz used a position-specific scoring matrix of aligned MHC-I peptides to identify several potential supertypes [11].

It is also possible to take a structural approach to the problem. QSAR analysis techniques have been used to refine the peptide interactions with the MHC-I groove by incrementally improving and optimizing the individual residue-to-residue interactions within the binding groove. This has led to the design of so-called 'superbinders' that minimize the entropic disruption in the groove and, therefore, can stabilize even disfavoured residues at so-called 'anchor positions' [12]. Finally, molecular dynamics has been used to quantify the energetic interactions between the MHC molecule and peptide for both MHC-I and MHC-II by analyzing the three-dimensional structure of the MHC–peptide complex [13–15]. Other predictive techniques related to epitope mapping have attempted to use different stages of antigen processing and presentation as a basis of prediction, most specifically the binding of the TAP transporter in the MHC-I pathway [16,17].

The epitope of a B cell is defined by the discrete surface region of an antigenic protein bound by the variable domain of an antibody. The production of specific antibodies for an infection can boost host immunity in the case of both intracellular and extracellular pathogens. The binding region of the antibody is composed of three hypervariable loops that can vary in both length and sequence so that the antibodies generated by an individual cell present a unique interface [18]. Despite the extreme variability of the region, the antibody-binding site is more hydrophobic than most protein surfaces with a significant predilection for tyrosine residues. B cell epitopes can be divided into continuous (linear) and discontinuous (conformational), the conformational epitopes being regions of the antigen separated within the sequence but brought together in the folded protein to form a three-dimensional interface. To date, the amino acid distribution of the complementary antigen surface has been difficult to characterize, presenting no unique sequential or structural features on which to base a predictive system. It is partly for this reason that B cell epitope prediction has lagged far behind T cell prediction in terms of accuracy and comprehensibility, but also because much of the data on which predictions are based are still open to question owing to the poorly understood recognition properties of cross-reactive antibodies [18]. Despite these fundamental limitations, several B cell epitope prediction programs are available, including DiscoTope [19], 3DEX [20] and CEP [21].

Reverse vaccinology

An alternative technique to epitope mapping is reverse vaccinology. Conventional experimental approaches to vaccine discovery cultivate pathogens under laboratory conditions, dissecting them into their main component proteins [1]. Antigens offering protective immunity are then identified by testing individual components. The trouble with this approach is that many of the proteins that are expressed during infection are not necessarily expressed *in vitro*, meaning good candidate antigens can be overlooked. The antigens also need to be produced on a large scale, making the process extremely time consuming. Moreover, it might not be possible to cultivate a particular pathogen in the laboratory. By contrast, reverse vaccinology analyzes the entire genome of a pathogen to identify potentially antigenic extracellular proteins. The technique is typically more effective for prokaryotic than eukaryotic organisms as eukaryotic organisms tend to have larger, more-complicated genomes. Initially, an algorithm capable of identifying open reading frames (ORFs) scans the pathogenic genome. Programs that can do this include ORF-FINDER [22], GLIMMER [23] and GS-Finder [24]. After all ORFs have been identified, proteins with the characteristics of secreted or surface molecules must be identified. Unlike the relatively straightforward task of identifying ORFs, selecting proteins liable to immune system surveillance is challenging. Programs such as ProDom [25], Pfam [26] and PROSITE [27] can identify sequence motifs characteristic of certain protein families and can therefore help to predict if a protein belongs to an extracellular protein family. An alternative approach is to use protein subcellular location prediction, which attempts to characterize where in a cell a protein is located [28]. Extracellular proteins identified in this way can be tested as potential subunit vaccines.

Reverse vaccinology was pioneered by a group investigating *Neisseria meningitidis,* the pathogen responsible for sepsis and meningococcal meningitis. The pathogen has five major subgroups (A, B, C, Y and W135) that are defined by the chemical composition of their capsular polysaccharides. Vaccines based on

the capsular proteins have been developed for all of the serotypes with the exception of subgroup B. Subgroup B is not a suitable target for a similar vaccine because its capsular protein is nearly identical to a human protein. Any vaccine derived from this subgroup would probably have poor immunogenicity and could also generate autoantibodies. The *Neisseria meningitidis* genome was scanned for potential ORFs [29,30]. Out of the 570 proteins that were identified, 350 could be expressed successfully *in vitro*. It was also determined experimentally that 85 of the 350 were surface exposed. A further test of candidate suitability was undertaken, to test the suitability of the proteins for conferring protection against heterologous strains. This was done by evaluating the proteins for gene presence, phase variation and sequence conservation in a panel of 22 genetically diverse MenB strains that were representative of the global diversity of the natural *N. meningitidis* population. Seven identified proteins conferred immunity over a broad range of strains, demonstrating the viability of *in silico* analysis as an aid to finding candidates for the clinical development of a MenB vaccine.

Following this success, reverse vaccinology has been applied to the genomes of several other pathogens. *Streptococcus agalactiae* is a Gram-positive streptococcus that can be found in the gut and female urogenital tract. It is a major cause of sepsis, pneumonia, meningitis and otitis media in young children. The vaccines that are currently available have a poor efficacy in infants and provide no crossprotection between different serotypes. Reverse vaccinology was undertaken for the group B *Streptococcus* (GBS) genome to explore the identification of a more effective and universal vaccine [31,32]. Mining of the genome identified 130 potential ORFs with significant homology to other bacterial surface proteins and virulence factors. 108 of 130 ORFs were successfully expressed and purified; six proteins were found to induce protective antibodies against pneumococcal challenge in a mouse sepsis model. All six of these candidates showed a high degree of crossreactivity against the majority of capsular antigens expressed *in vivo* that are believed to be immunogenic in humans.

Another example is *Porphyromonas gingivalis*, a Gram-negative anaerobic bacterium present in subgingival plaques present in chronic adult periodontitis, an inflammatory disease of the gums. Shotgun sequences of the genome identified ~370 ORFs [33] (although the actual number of ORFs within the genome is much higher). Seventy four of these had significant global homology to known surface proteins or an association with virulence. Forty six had significant similarity with other bacterial outer membrane proteins. Forty nine proteins were identified as surface proteins using PSORT [34] and 22 through motif analysis. This generated 120 unique proteins sequences. Of these, 107 were expressed in *Escherichia coli* and analyzed by western blotting using sera from human periodontitis patients and animal antisera. Forty candidates were shown to be positive for at least one of the sera. These were then purified and used to vaccinate mice, with only two of the antigens demonstrating significant protection. Both of these had a high degree of homology with outer membrane protein F (OprF) from *Pseudomonas aeruginosa,* a protein that is part of a vaccine undergoing human clinical trials. Finally, 141 ORFS were selected through *in silico* analysis from *Chlamydia pneumoniae*, an obligate intracellular bacterium that is associated with respiratory infections, cardiovascular disease and atherosclerotic disease, and is a significant cause of pneumonia in hospital and outpatient settings [35]. Fluorescence-activated cell sorting analysis combined with western blots and proteomic verification of elementary body total proteins on two-dimensional gel electrophoresis identified 53 putative surface-exposed proteins. If reverse vaccinology is applied appropriately in vaccine design, it can save enormous amounts of money, time and wasted labour.

Allergen discovery

Most allergic reactions depend on a series of intrinsic and extrinsic factors that control both the development and triggering of the condition [36]. Such reactions give rise to rhinitis, asthma and eczema and are usually induced by inhalation or ingestion of allergenic proteins. They are often caused by the type I hypersensitive reaction, induced by antigens that elicit specific immunoglobulin E (IgE) antibodies or from crossreactivity between common homologous allergens. Modified proteins used in food have led to increasing concern about the identification of allergenic proteins. The World Health Organization (WHO) and Food and Agriculture Organization (FAO) have defined an allergen as any protein that shares a contiguous sequence of at least six residues with a known allergen or has a homology with it of >35% over a region of 80 residues. This definition has been heavily criticized because it produces so many false positive results. IgE prediction has used similar approaches to those used for general epitope mapping. This includes motif-based approaches (the MEME–MAST program) [37], a k nearest neighbour classifier [38] and similarity searches against IgE epitope–epitope profiles [39]. However, none of the techniques yet developed have produced a reliable predictor of IgE epitopes, partly owing to the limited number of known epitopes. Many allergens seem to cluster in a limited number of protein families but few families seem to lack members with allergenic properties [40]. Attempts to induce immunities have been based on an allergen-specific immunotherapy known as structured treatment interruption (STI), whereby the patient is administered increasing amounts of allergen extract to augment their natural tolerance. The allergen extracts have been termed 'vaccines' by the WHO [41]. STI, although often effective, is time consuming and cannot be applied to all allergies [36]. Recombinant hypoallergenic allergen derivatives have been produced for several allergen sources [42]. Therefore, there is considerable interest in identifying target allergens that can be modified to dictate better the extent of the subsequent immune response. Although rational intervention is still problematic, the development of allergen databases should hopefully facilitate this.

Cancer vaccinology

It has been established that prophylactic vaccines can effectively block or suppress the growth of certain tumours. However, they are only effective against early microscopic tumours and not larger tumour masses [43]. It was shown that mammary carcinogenensis that was driven by the HER-2/neu oncogene in mice could be effectively blocked using the Triplex vaccine [43–45]. The vaccine was derived from cells that expressed the HER-2/neu antigen combined with two adjuvant signals, interleukin-12 and allogeneic MHC-I antigens. It is desirable to find the simplest and most effective means of vaccination that provides effective suppression of tumour growth. Although this can be done experi-

mentally, it is not the most efficient way of deriving a vaccination schedule. An algorithm was developed using a cellular automate approach that simulated the immune response to Triplex vaccine. The technique was optimized using all available experimental data [43–45]. The simulator then searched *in silico* to minimize the number of vaccines while not reducing the tumour prevention efficacy. This is the first instance where a simulator has been used to predict immune response in a vaccine. In the future, this technique will be applied to other vaccination strategies.

Codon optimization

DNA vaccines are plasmids capable of expressing antigenic peptides within the host [46,47]. These are seen as attractive alternatives to conventional vaccines as they generate both a cellular and a humoral immune response; this combined response has proven particularly effective in combating intracellular pathogens. There are ways to optimize the efficiency of a DNA vaccine beyond the choice of antigen. The immunogenicity of DNA vaccines has been successfully enhanced by techniques such as codon optimization [48], CpG motif engineering [49,50] and the introduction of promoter sequences [51,52]. One of the most effective of these techniques has been codon optimization, which can enhance the efficiency of protein expression. Translationally-optimal codons are those that are recognized by abundant transfer RNAs [53] and, within a phylogenetic group, the frequency of particular codons in a gene is highly correlated with the level of gene expression. Immunogenicity depends on the effective translation and transcription of the antigenic protein, and it is possible to enhance this by selecting optimal codons for the translation of the vaccine. A similar technique, CpG optimization, can be used to optimize the codons with respect to CG dinucleotides. Pattern recognition receptors that form part of the innate immune system can often distinguish prokaryotic DNA from eukaryotic DNA by detecting unmethylated CpG dinucleotides in particular base contexts, termed 'CpG motifs'. The presence of these motifs in the sequence can be highly advantageous so long as it does not interfere with the process of codon optimization [54].

Adjuvant discovery

Another technique for optimizing the efficacy of vaccines is to develop an efficient adjuvant: a substance that enhances the immune response when delivered with the vaccine [55,56]. It is possible that some adjuvants function as immune potentiators, triggering an early innate immune response that enhances the effectiveness of the vaccine by increasing its uptake. Adjuvants can also enhance vaccination by improving the depot effect, that is the colocalization of the antigen and immune potentiators, by delaying the spread of the antigen from the site of infection so that absorption occurs over a prolonged period [57]. Aluminium hydroxide (also known as alum) is the only adjuvant currently licensed in humans. Aluminium-based adjuvants prolong antigen persistence due to the depot effect, which stimulates the production of IgG1 and IgE antibodies [58], and triggers the secretion of interleukin-4. There are also several small-molecule, drug-like adjuvants, including imiquimod, resiquimod, and other imidazoquinolines. Likewise, many proteins have been identified as potential adjuvant molecules. These are amenable to detection by sequence-based approaches and, potentially, by alignment-free techniques. This 'adjuvant hunting' can be likened to reverse vaccinology but with adjuvants instead of antigens as its quarry.

Small-molecule adjuvant discovery is amenable to techniques used routinely by the pharmaceutical industry. Three-dimensional virtual screening is a fast and effective way of identifying molecules by docking a succession of ligands into a defined binding site [59]. A large database of small molecules can be screened quickly and efficiently in this way. Using 'targeted' libraries containing a specific subset of molecules is often even more effective. It is possible to use 'privileged fragments' to construct combinatorial libraries, those which are expected to have an increased probability of success. A pharmacophore is a three-dimensional map of properties common to active conformations of a set of ligands exhibiting a particular activity; it can be used to discover new molecules with similar properties. Small molecules that have been investigated for adjuvant properties in this way include monophosphoryl lipid A, muramyl dipeptide, QS21, poly(DL-lactide-co-glycolide) and Montanide ISA-51 [60]. More recently, molecules that selectively interfere with chemokine-mediated T cell migration have shown the potential to function as adjuvants by downregulating the expression of costimulatory molecules, limiting T cell activation. Small-molecule chemokine receptor antagonists have been identified and shown to be effective at blocking chemokine function *in vivo* [61,62] although no compound has reached a Phase II clinical trial to date.

Available software

Several programs are available that can help to design and optimize vaccines (a full list of prediction servers is given in Table 1). In this section, some of the most effective algorithms for each form of vaccine design are discussed. For T cell epitope prediction, many programs are available. A sensible approach for a new user would be to use MHCBench [63], an interface developed specifically for evaluating the various MHC binding peptide prediction algorithms. MHCBench enables users to compare the performance of various programs with both threshold-dependent and -independent parameters. The server can also be extended to include new methods for different MHC alleles. Recently, the Immune Epitope Database (IEDB; http://immuneepitope.org/home.do) has also begun a service by which MHC-I prediction programs can be benchmarked [64]. B cell prediction is more problematic owing to the difficulties in correctly defining both linear and discontinuous epitopes from the rest of the protein. VaxiJen uses an alignment-free approach to predict antigens directly [65]. Instead of concentrating on epitope and non-epitope regions, the method used bacterial, viral and tumour protein datasets to derive statistical models for predicting whole protein antigenicity. The models showed prediction accuracy up to 89%, indicating a far higher degree of accuracy than has, for example, been obtained previously for B cell epitope prediction.

The New Enhanced Reverse Vaccinology Environment (NERVE) program has been developed to automate and refine the process of reverse vaccinology further, in particular, the process of identifying surface proteins [66]. In NERVE, the processing of potential ORFs is a six-step process. It begins with predicting subcellular localization, followed by calculating the probability of the protein having adhesion-like properties, identifying transmembrane

TABLE 1

Servers for prediction problems in immunology and vaccinology

T cell epitope prediction	Web address	Server description
SYFPEITHI	http://www.syfpeithi.de	Predictive server featuring MHC-presented epitopes, MHC-specific anchor and auxiliary motifs
BIMAS	http://bimas.dcrt.nih.gov/molbio/hla_bind/	A HLA peptide binding predictor
EpiDirect	http://epipredict.de/index.html	Prediction method for MHC-II restricted T cell epitopes and ligands
HIV	http://hiv.lanl.gov/content/hiv-db/ALABAMA/epitope_analyzer.html	An HIV epitope location finder
Imtech	http://imtech.res.in/raghava/mhc/	Matrix optimization technique for the prediction of MHC binding
Lib Score	http://hypernig.nig.ac.jp/cgi-bin/Lib-score/request.rb?lang=E	Peptide position-specific library prediction server
MHC Bench	http://www.imtech.res.in/raghava/mhcbench/	Evaluation of MHC binding peptide predictive algorithms
MHCPred 20	http://www.jenner.ac.uk/MHCPred/	Quantitative T cell epitope prediction server with both human and mouse models
MHC-THREAD	http://www.csd.abdn.ac.uk/~gjlk/MHC-Thread/	Predictor for peptides that are likely to bind to MHC-II molecules
NetMHC	http://www.cbs.dtu.dk/services/NetMHC/	Produces a neural network prediction of binding affinities for HLA-A2 and H-2Kk
PREDEP	http://margalit.huji.ac.il/	MHC-I epitope prediction
ProPred	http://www.imtech.res.in/raghava/propred/	Prediction of MHC-II binding regulation in an antigen sequence using quantitative matrices
RANKPEP	http://www.mifoundation.org/Tools/rankpep.html	Prediction of binding peptides to MHC-I and MHC-II molecules
SMM	http://zlab.bu.edu/SMM/	Prediction of high affinity HLA-A2 binding peptides
SVMHC	http://www.sbc.su.se/~pierre/svmhc/	A prediction tool for MHC-I binding peptides
B cell epitope prediction		
ePitope	http://www.epitope-informatics.com/	Epitope prediction server for the identification and targeting of protein B cell epitopes
Bcepred	http://www.imtech.res.in/raghava/bcepred/	B cell epitope prediction methods based on physico-chemical properties on a nonredundant dataset
ABCpred	http://www.imtech.res.in/raghava/abcpred/	B cell epitope prediction using artificial neural networks
Antigenic	http://liv.bmc.uu.se/cgi-bin/emboss/antigenic	Predicts antigenic sites in proteins
Discotope	http://www.cbs.dtu.dk/services/DiscoTope/	Server predicting discontinuous B cell epitopes from three-dimensional protein structures
CEP	http://202.41.70.74:8080/cgi-bin/cep.pl	Server for predicting conformational epitopes
VaxiJen	http://www.jenner.ac.uk/VaxiJen/	Prediction of protective antigens and subunit vaccines
Reverse vaccinology		
LipPred	http://www.jenner.ac.uk/LipPred/	Server for the identification of lipoproteins
ORF-FINDER	http://www.ncbi.nlm.nih.gov/projects/gorf/	Program locates all open reading frames of a selectable minimum size in a sequence
GLIMMER	http://www.cbcb.umd.edu/software/glimmer/	Program uses interpolated Markov models to identify the coding regions
GLIMMERHMM	http://www.cbcb.umd.edu/software/GlimmerHMM/	Similar program to GLIMMER but specialized towards eukaryotic genomes
GS-Finder	http://tubic.tju.edu.cn/GS-Finder/download.htm	Algorithm to identify translation start sites of the ORFs in bacterial genomes
Zcurve	http://tubic.tju.edu.cn/Zcurve_B/	Server for recognizing protein coding genes in bacterial and archaeal genomes
GeneMark	http://exon.gatech.edu/GeneMark/	Family of gene prediction servers for prokaryotes, eukaryotes and viruses
NERVE	http://www.bio.unipd.it/molbinfo/	*In silico* identification of the best vaccine candidates from whole proteomes of bacterial pathogens
Allergen discovery		
AlgPred	http://www.imtechresin/raghava/algpred/	Prediction of allergens based on similarity of known epitope with any region of protein
AllerPredict	http://sdmci2ra-staredusg/Templar/DB/Allergen/	ALLERDB is an allergen database integrated with analytical tools
AllerMatch	http://www.allermatch.org/	Comparison of amino acid sequence of a protein of interest with sequences of allergenic proteins
WebAllergen	http://weballergen.bii.a-star.edu.sg/	Sequence compared against a set of prebuilt allergenic motifs

TABLE 1 (*Continued*)

T cell epitope prediction	Web address	Server description
DASARP	http://www.slv.se/templatesSLV/SLV_Page____9343.asp	Detection based on automated selection of allergen-representative peptide
Codon optimization		
DyNAVacS	http://miracle.igib.res.in/dynavac/	Program identifies a suitable expression vector and performs codon optimization
DNA 20	http://www.dnatwopointo.com/commerce/misc/opt.jsp	Custom gene synthesis including codon optimization for increased protein expression
Upgene	http://www.vectorcore.pitt.edu/upgene/upgene.html	A web-based DNA codon optimization algorithm
GeneMaker	http://www.blueheronbio.com/genemaker/codon.html	Algorithm matches codon usage of sequence with that of the host

domains, and comparing with the human proteome and with that of the selected pathogen, after which the protein is assigned a putative function. The vaccine candidates are then filtered and ranked based on these calculations. While it is generally accepted that determining ORFs is a relatively straightforward process, the algorithm used to define extracellular proteins from other proteins needs to be carefully selected. One of the most effective programs that can be used for this purpose is HensBC, a recursive algorithm for predicting the subcellular location of proteins [28]. The program constructs a hierarchical ensemble of classifiers by applying a series of if–then rules. HensBC can assign proteins to one of four different types (cytoplasmic, mitochondrial, nuclear or extracellular) with ~80% accuracy. Classification of Gram-negative bacterial proteins showed 83.2% accuracy at determining between five subcellular locations: cytoplasmic, intermembrane, periplasmic, outer membrane and extracellular. The algorithm is nonspecialized and can be applied to any genome. Any protein identified as being extracellular could be a potential vaccine candidate.

Allergen prediction software has not yet been effective. However, a web server named AlgPred has been developed that enables the prediction of allergens from the amino acid sequence of the protein [67]. It enables users to choose any of the following approaches: (i) the scanning of IgE epitopes, (ii) a motif-based approach, (iii) an SVM-based method using the amino acid composition of the protein, (iv) a hybrid approach, and (v) a BLAST search on ARPs. Owing to the limitations of the individual programs, it is recommended that a consensus approach be used for allergen prediction. Allermatch [68] is a server that enables the user to predict potential allergenicity based on the FAO and WHO allergen criteria. The most comprehensive approach to vaccine optimization is taken by DyNAVacS, an integrative bioinformatics tool that optimizes codons for heterologous expression of genes in bacteria, yeasts and plants [54]. The program is also capable of mapping restriction enzyme sites, primer design and designing therapeutic genes. The program calculates the optimal code for each amino acid encoded by a stretch of DNA by using codon usage tables, which contain codon frequencies for a variety of different genomes.

Concluding remarks

Vaccine design and development is an inherently laborious process but the programs and techniques outlined here have the potential to simplify the process greatly. The techniques also have the potential to identify candidate proteins that would be overlooked by conventional experimentation. In particular, reverse vaccinology has proved effective in the discovery of antigenic subunit vaccines that would otherwise remain undiscovered. Projects such as ImmunoGrid (http://www.immunogrid.org) and ViroLab (http://www.virolab.org:8080/virolab) are attempting to push the boundaries of what can be achieved through simulation of the immune system. ImmunoGrid seeks to simulate immune processes by combining experimental and computational studies, whereas ViroLab is trying to develop a virtual laboratory for infectious diseases by examining the genetic causes underlying human illness. Both harness GRID technology. GRID is a computing model that distributes processing power across large parallel infrastructures. GRID will enable larger and more-ambitious simulations than have been possible hitherto.

References

1 De Groot, A.S. (2006) Immunomics: discovering new targets for vaccines and therapeutics. *Drug Discov. Today* 11, 203–209

2 De Groot, A.S. and Berzofsky, J.A. (2004) From genome to vaccine – new immunoinformatics tools for vaccine design. *Methods* 34, 425–428

3 McMurry, J. *et al.* (2005) Analyzing *Mycobacterium tuberculosis* proteomes for candidate vaccine epitopes. *Tuberculosis (Edinb.)* 85, 95–105

4 Rammensee, H. *et al.* (1999) SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* 50, 213–219

5 Jardetzky, T.S. *et al.* (1996) Crystallographic analysis of endogenous peptides associated with HLADR1 suggests a common, polyproline II-like conformation for bound peptides. *Proc. Natl. Acad. Sci. U. S. A.* 93, 734–738

6 Donnes, P. and Elofsson, A. (2002) Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics* 3, 25. doi: 10.1186/1471-2105-3-25 (http://www.biomedcentral.com)

7 Liu, W. *et al.* (2006) Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models. *BMC Bioinformatics* 7, 182. doi: 10.1186/1471-2105-7-182 (http://www.biomedcentral.com)

8 Wan, J. *et al.* (2006) SVRMHC prediction server for MHC-binding peptides. *BMC Bioinformatics* 7, 463. doi: 10.1186/1471-2105-7-463 (http://www.biomedcentral.com)

9 Noguchi, H. *et al.* (2002) Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules. *J. Biosci. Bioeng.* 94, 264–270

10 Burden, F.R. and Winkler, D.A. (2005) Predictive Bayesian neural network models of MHC class II peptide binding. *J. Mol. Graph. Model.* 23, 481–489

11 Reche, P.A. and Reinherz, E.L. (2005) PEPVAC: a web server for multi-epitope vaccine development based on the prediction of supertypic MHC ligands. *Nucleic Acids Res.* 33, W138–W142

12 Doytchinova, I.A. *et al.* (2005) Towards the chemometric dissection of peptide-HLA-A*0201 binding affinity: comparison of local and global QSAR models. *J. Comput. Aided Mol. Des.* 19, 203–212
13 Davies, M.N. *et al.* (2006) Statistical deconvolution of enthalpic energetic contributions to MHC-peptide binding affinity. *BMC Struct. Biol.* 6, 5. doi: 10.1186/1472-6807-6-5 (http://www.biomedcentral.com)
14 Davies, M.N. *et al.* (2003) A novel predictive technique for the MHC class II peptide-binding interaction. *Mol. Med.* 9, 220–225
15 Wan, S. *et al.* (2004) Large-scale molecular dynamics simulations of HLA-A*0201 complexed with a tumor-specific antigenic peptide: can the α_3 and β_2m domains be neglected? *J. Comput. Chem.* 25, 1803–1813
16 Doytchinova, I.A. and Flower, D.R. (2006) Class I T-cell epitope prediction: Improvements using a combination of proteasome cleavage, TAP affinity, and MHC binding. *Mol. Immunol.* 43, 2037–2044
17 Zhang, G.L. *et al.* (2006) PRED^{TAP}: a system for prediction of peptide binding to the human transporter associated with antigen processing. *Immunome Res.* 2, 3. doi: 10.1186/1745-7580-2-3 (http://www.immunome-research.com)
18 Blythe, M.J. and Flower, D.R. (2004) Benchmarking B-cell epitope prediction: underperformance of existing methods. *Protein Sci.* 14, 246–248
19 Haste Andersen, P. *et al.* (2006) Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci.* 15, 2558–2567
20 Schreiber, A. *et al.* (2005) 3D-Epitope-Explorer (3DEX): localization of conformational epitopes within three-dimensional structures of proteins. *J. Comput. Chem.* 26, 879–887
21 Kulkarni-Kale, U. *et al.* (2005) CEP: a conformational epitope prediction server. *Nucleic Acids Res.* 33, W168–W171
22 Rombel, I.T. *et al.* (2002) ORF-FINDER: a vector for high-throughput gene identification. *Gene* 282, 33–41
23 Delcher, A.L. *et al.* (1999) Improved microbial gene identification with GLIMMER. *Nucleic Acids Res.* 27, 4636–4641
24 Ou, H.Y. *et al.* (2004) GS-Finder: a program to find bacterial gene start sites with a self-training method. *Int. J. Biochem. Cell Biol.* 36, 535–544
25 Servant, F. *et al.* (2002) ProDom: Automated clustering of homologous domains. *Brief. Bioinform.* 3, 246–251
26 Bateman, A. *et al.* (2000) The Pfam protein families database. *Nucleic Acids Res.* 28, 263–266
27 Falquet, L. *et al.* (2002) The PROSITE database. *Nucleic Acids Res.* 30, 235–238
28 Bulashevska, A. and Eils, R. (2006) Predicting protein subcellular locations using hierarchical ensemble of Bayesian classifiers based on Markov chains. *BMC Bioinformatics* 7, 298–310
29 Tettelin, H. *et al.* (2000) Complete genome sequence of *Neisseria meningitidis* serogroup B strain MC58. *Science* 287, 1809–1815
30 Pizza, M. *et al.* (2000) Whole genome sequencing to identify vaccine candidates against serogroup B meningococcus. *Science* 287, 1816–1820
31 Wizemann, T.M. *et al.* (2001) Use of a whole genome approach to identify vaccine molecules affording protection against *Streptococcus pneumoniae* infection. *Infect. Immun.* 69, 1593–1598
32 Maione, D. *et al.* (2005) Identification of a universal Group B streptococcus vaccine by multiple genome screen. *Science* 309, 148–150
33 Ross, B.C. *et al.* (2001) Identification of vaccine candidate antigens from a genomic analysis of *Porphyromonas gingivalis*. *Vaccine* 19, 4135–4142
34 Rey, S. *et al.* (2005) PSORTdb: a protein subcellular localization database for bacteria. *Nucleic Acids Res.* 33, D164–D168
35 Montigiani, S. *et al.* (2002) Genomic approach for analysis of surface proteins in *Chlamydia pneumoniae*. *Infect. Immun.* 70, 368–379
36 Crameri, R. and Rhyner, C. (2006) Novel vaccines and adjuvants for allergen-specific immunotherapy. *Curr. Opin. Immunol.* 18, 761–768
37 Stadler, M.B. and Stadler, B.M. (2003) Allergenicity prediction by protein sequence. *FASEB J.* 17, 1141–1143
38 Soeria-Atmadja, D. *et al.* (2004) Statistical evaluation of local alignment features predicting allergenicity using supervised classification algorithms. *Int. Arch. Allergy Immunol.* 133, 101–112
39 Ivanciuc, O. *et al.* (2003) SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res.* 31, 359–362
40 Soeria-Atmadja, D. *et al.* (2006) Computational detection of allergenic proteins attains a new level of accuracy with *in silico* variable-length peptide extraction and machine learning. *Nucleic Acids Res.* 34, 3779–3793
41 Linhart, B. and Valenta, R. (2005) Molecular design of allergy vaccines. *Curr. Opin. Immunol.* 17, 646–655
42 Bousquet, J. *et al.* (1998) Allergen immunotherapy: therapeutic vaccines for allergic diseases. A WHO position paper. *J. Allergy Clin. Immunol.* 102, 558–562
43 Motta, S. *et al.* (2005) Modelling vaccination schedules for a cancer immunoprevention vaccine. *Immunome Res.* 1, 5. doi: 10.1186/1745-7580-1-5 (http://www.immunome-research.com)
44 Pappalardo, F. *et al.* (2005) Modeling and simulation of cancer immunoprevention vaccine. *Bioinformatics* 21, 2891–2897
45 Lollini, P.L. *et al.* (2006) Discovery of cancer vaccination protocols with a genetic algorithm driving an agent based simulator. *BMC Bioinformatics* 7, 352. doi: 10.1186/1471-2105-7-352 (http://www.biomedcentral.com)
46 Babiuk, L.A. *et al.* (2000) Nucleic acid vaccines: research tool or commercial reality. *Vet. Immunol. Immunopathol.* 76, 1–23
47 Babiuk, L.A. *et al.* (2003) Induction of immune responses by DNA vaccines in large animals. *Vaccine* 21, 649–658
48 Uchijima, M. *et al.* (1998) Optimization of codon usage of plasmid DNA vaccine is required for the effective MHC class I-restricted T-cell responses against an intracellular bacterium. *J. Immunol.* 161, 5594–5599
49 Klinman, D.M. *et al.* (1997) Contribution of CpG motifs to the immunogenicity of DNA vaccines. *J. Immunol.* 158, 3635–3639
50 Booth, J.S. *et al.* (2007) Innate immune responses induced by classes of CpG oligodeoxynucleotides in ovine lymph node and blood mononuclear cells. *Vet. Immunol. Immunopathol.* 115, 24–34
51 Lee, A.H. *et al.* (1997) Comparison of various expression plasmids for the induction of immune response by DNA immunization. *Mol. Cells* 7, 495–501
52 Xu, Z.L. *et al.* (2001) Optimization of transcriptional regulatory elements for constructing plasmid vectors. *Gene* 272, 149–156
53 Henry, I. and Sharp, P.M. (2007) Predicting gene expression level from codon usage bias. *Mol. Biol. Evol.* 24, 10–12
54 Harish, N. *et al.* (2006) DyNAVacS: an integrative tool for optimized DNA vaccine design. *Nucleic Acids Res.* 34, W264–W266
55 Singh, M. and O'Hagan, D.T. (2002) Recent advances in vaccine adjuvants. *Pharm. Res.* 19, 715–728
56 Stills, H.F., Jr (2005) Adjuvants and antibody production: dispelling the myths associated with Freund's complete and other adjuvants. *ILAR J.* 46, 280–293
57 Gupta, R.K. (1998) Aluminum compounds as vaccine adjuvants. *Adv. Drug Deliv. Rev.* 32, 155–172
58 Singh, M. and Srivastava, I. (2003) Advances in vaccine adjuvants for infectious diseases. *Curr. HIV Res.* 1, 309–320
59 Schellhammer, I. and Rarey, M. (2004) FlexX-Scan: fast, structure-based virtual screening. *Proteins* 57, 504–517
60 Charoenvit, Y. *et al.* (2004) A small peptide (CEL-1000) derived from the β-chain of the human major histocompatibility complex class II molecule induces complete protection against malaria in an antigen-independent manner. *Antimicrob. Agents Chemother.* 48, 2455–2463
61 Godessart, N. (2005) Chemokine receptors: attractive targets for drug discovery. *Ann. N. Y. Acad. Sci.* 1051, 647–657
62 Johnson, Z. *et al.* (2004) Chemokine inhibition–why, when, where, which and how? *Biochem. Soc. Trans.* 32, 366–377
63 Singh, H. and Raghava, G.P. (2001) ProPred: prediction of HLA-DR binding sites. *Bioinformatics* 17, 1236–1237
64 Peters, B. *et al.* (2006) A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Comput. Biol.* 2, e65. doi: 10.1371/journal.pcbi.0020065 (http://compbiol.plosjournals.org)
65 Doytchinova, I.A. and Flower, D.R. (2007) VaxiJen: a server for prediction of protective antigens, tumour antigens and subunit vaccines. *BMC Bioinformatics* 8, 4. doi: 10.1186/1471-2105-8-4 (http://www.biomedcentral.com)
66 Vivona, S. *et al.* (2006) NERVE: new enhanced reverse vaccinology environment. *BMC Biotechnol.* 6, 35. doi: 10.1186/1472-6750-6-35 (http://www.biomedcentral.com)
67 Saha, S. and Raghava, G.P. (2006) AlgPred: prediction of allergenic proteins and mapping of IgE epitopes. *Nucleic Acids Res.* 34, W202–W209
68 Fiers, M.W. *et al.* (2004) Allermatch, a webtool for the prediction of potential allergenicity according to current FAO/WHO Codex alimentarius guidelines. *BMC Bioinformatics* 5, 133. doi: 10.1186/1471-2105-5-133 (http://www.biomedcentral.com)