Review

Computer-aided biotechnology: from immuno-informatics to reverse vaccinology

Sandro Vivona[1], Jennifer L. Gardy[2], Srinivasan Ramachandran[3], Fiona S.L. Brinkman[4], G.P.S. Raghava[5], Darren R. Flower[6] and Francesco Filippini[1]

[1] Molecular Biology and Bioinformatics Unit, Department of Biology, University of Padua, Padua, Italy
[2] Centre for Microbial Diseases and Immunity Research, University of British Columbia, Vancouver, British Columbia, V6T 1Z4, Canada
[3] G.N. Ramachandran Knowledge Centre for Genome Informatics, Institute of Genomics and Integrative Biology Mall Road, Delhi, India
[4] Department of Molecular Biology and Biochemistry, Room SSB 8166, Simon Fraser University, Burnaby, British Columbia, V5A 1S6, Canada
[5] Bioinformatics Centre, Institute of Microbial Technology, Sector 39-A, Chandigarh, India
[6] The Jenner Institute, University of Oxford, Compton, Berkshire, UK

Genome sequences from many organisms, including humans, have been completed, and high-throughput analyses have produced burgeoning volumes of 'omics' data. Bioinformatics is crucial for the management and analysis of such data and is increasingly used to accelerate progress in a wide variety of large-scale and object-specific functional analyses. Refined algorithms enable biotechnologists to follow 'computer-aided strategies' based on experiments driven by high-confidence predictions. In order to address compound problems, current efforts in immuno-informatics and reverse vaccinology are aimed at developing and tuning integrative approaches and user-friendly, automated bioinformatics environments. This will herald a move to 'computer-aided biotechnology': smart projects in which time-consuming and expensive large-scale experimental approaches are progressively replaced by prediction-driven investigations.

Introduction

From 'functional' bioinformatics to computer-aided biotechnology

In recent years, science has been characterized increasingly by computer-aided research: projects implementing early *in silico* analyses that shorten the length and shape the direction of subsequent 'wet-bench' tasks. The continuous flow of new genomic sequence and functional annotation data from every taxonomic lineage permits researchers to capture correlations based on large datasets, enabling the design of more-reliable analytical and predictive tools. For example, the alignment of multiple homologous sequences has helped in the identification of a large number of structural and functional signatures: protein domain profiles and pattern-based motifs indicative of catalytic sites, ligand-binding sites, sorting signals and more (for example, see [1–4]). This has generated several integrated databases containing combined sets of such signatures and of scanning tools capable of inferring possible functions or regulatory mechanisms from the presence of either canonical or degenerate signatures in a sequence of interest. Such 'sequence-to-function' approaches [5] have enhanced the production of further functional evidence by prediction-driven experiments. Early examples included using the identification of specific

Glossary

Antigen: a moiety exhibiting antigenicity; a substance recognized by an existing immune response and associated molecules such as T-cells or antibodies in a recall response. The antigen is typically, but not exclusively, a protein or carbohydrate component of a pathogen and is recognized by a component of the immune system.
B-cell epitopes: regions of the surface of a protein, or other biomacromolecule, recognized by soluble or membrane-bound antibody molecules.
Cellular immunity: immune response mediated by T-cells that constantly patrol the body seeking out foreign proteins originating from pathogens.
Conformational epitope: epitopic residues that are not arranged sequentially. When a protein folds, the residues forming a conformational epitope are juxtaposed, enabling the antibody to recognize its three-dimensional structure.
Continuous or linear epitope: a protein epitope consisting of 6 to 10 adjacent amino acid residues.
Epitopes: isolated peptides; localised regions within larger protein structures, carbohydrates, lipids or nucleotides – or a combination thereof.
Humoral immunity: immune responses mediated by B-cells via soluble or membrane bound antibodies.
Immunogenicity: property of a molecular (protein, lipid or carbohydrate, or a combination thereof), or supramolecular (virus, bacteria or protozoan parasite), moiety that enables it to induce a significant response from the immune system.
Immuno-informatics: a branch of bioinformatics focusing on immunology and vaccinology.
Linear data: data structures in which insertion and deletion is possible in linear fashion (e.g. arrays or linked lists).
Non-linear data: data structures in which linear insertion and deletion is not possible (e.g. trees, graphs or stacks).
Protective antigen: an antigen that triggers a protective immune response.
Quantitative matrix (QM): this is essentially a refined motif. The matrix (X,Y) represents the probability that an X residue will occur at a specific Y position. Summation of position specific values, from the matrix, for a given peptide yields the predicted binding score.
T-cell epitopes: short peptides bound by MHC molecules and subsequently recognized by T-cells.

E-mail addresses: francesco.filippini@unipd.it, fmemf@bio.unipd.it

0167-7799/$ – see front matter doi:10.1016/j.tibtech.2007.12.006 Available online 21 February 2008

signatures in a sequence to suggest functional assays able to identify otherwise long-sought functions [6] as well as to imply functions for disease gene products [7,8]. Later, more complex, genome-wide analyses have led to the identification of proteomic complements that underlie regulatory pathways or interaction network organization in model organisms [9,10]. This bioinformatics revolution has heralded the birth of 'computer-aided biotechnology'.

Immuno-informatics as an example of computer-aided biotechnology

The use of bioinformatics approaches to uncover functional information has had a great impact on molecular immunology and has enabled researchers in the field to address biological and biotechnological problems by turning their attention towards 'compound problems': problems that require the integration of diverse lines of both *in silico* and experimental evidence. With respect to sequence analysis, this scaling-up strategy requires the availability of reliably predicted features. To handle evidence diversity, immuno-informatics (see Glossary) uses strategies that span several areas of bioinformatics, including database creation and management [11,12], the definition and use of functional and structural signatures, and the development and use of predictive tools [12–14]. Moreover, a variety of algorithmic approaches has been used to develop immuno-informatic tools. The functional potential of these *in silico* approaches has found its paradigm in reverse vaccinology (RV, see below), an example of how rational integration of reliable strategies has turned into synergetic and applicative power. RV represented a revolution in immunology and a milestone in biotechnology; it illustrated how a compound biological problem such as vaccine design could be solved by identifying, addressing and integrating *in silico* several 'sub-problems' with a reasonable degree of success [15].

Below, we highlight several of the most important problems from an immunological viewpoint, and the *in silico* resources that can be used to address them. We then present examples with applicative potential.

Databases and predictive tools

Molecular basis of the immune response

Immunogenicity (see Glossary) is among the most important and interesting properties within vaccine design and discovery. In classifying measures of immunogenicity, it is useful to distinguish between humoral and cellular immunity (see Glossary). Antibodies mediate humoral immunity; in cellular immunity, a complex system of proteins converges on the interaction between major histocompatability complexes (MHCs) and T-cell receptors (TCRs) (see Figure 1). The TCR exhibits a wide range of selectivities and affinities; TCRs bind MHCs that are presented on the surfaces by other cells. These proteins bind small peptide fragments derived from both host and pathogen proteins. The recognition of such ternary complexes of MHC, peptide and TCR lies at the heart of the cellular immune response. An antigen (see Glossary) can be recognized either as a whole molecule by an antibody or as one or more peptide fragments by the T-cell.

Identification of antigens relevant to the immune response

A starting step towards reaching the goal of obtaining an effective vaccine is the identification of an optimal composition of protective antigens (see Glossary) [16]. Both proteins and capsule polysaccharides have been used as protective antigens in various vaccine formulations but, in the latter case, the need to cultivate the pathogen and the emergence of molecular variations have made it difficult to effectively sustain vaccine availability [17]. Hence, major efforts worldwide are directed towards identifying and

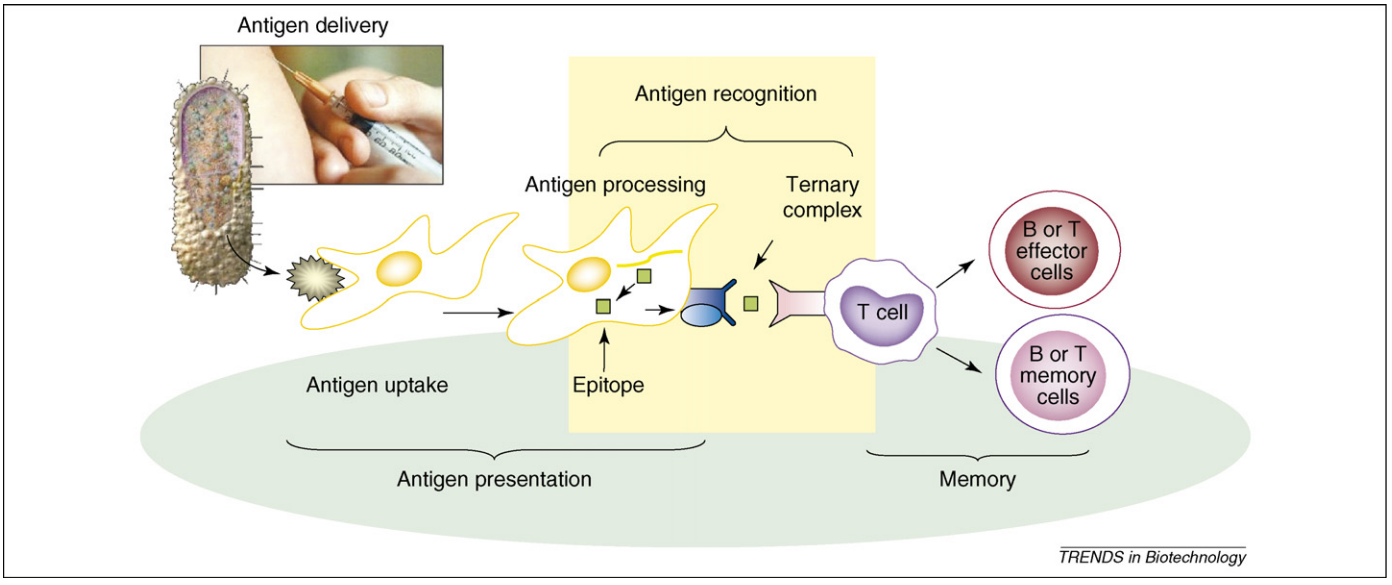

Figure 1. Molecular and cellular players in antigen presentation and recognition. Cellular immunity is mediated by T-cells, which constantly patrol the body, seeking out foreign proteins originating from pathogens. The uptake of an antigen is followed by intracellular processing, which results in the production of small peptide epitopes that are presented at the cell surface and are bound to major histocompatibility complex (MHC) molecules (blue). The T-cell receptor (TCR, red) exhibits a wide range of selectivities and affinities; TCRs bind to MHCs, which are presented on the surfaces by other cells. It is recognition of such ternary MHC–peptide–TCR complexes (i.e. antigen recognition, yellow background box) that lies at the heart of both the adaptive, and memory, cellular immune response.

testing protein antigens. Traditional efforts in this area screened expression libraries with patients' antibodies. However, the availability of complete genomic sequences of pathogens and a plethora of new bioinformatics tools have opened up new approaches to antigen identification. The rationale used focuses on identification of a pathogen's surface proteins. The interactions between a pathogen's surface-exposed or secreted proteins and host cells represent a pivotal step in pathogenesis and virulence. In pathogens, several of the key players are proteins involved in adhesion, invasion, secretion, signalling, annulling host responses, toxicity, motility and lipoproteins [16,18]. Some protein classes, such as fimbriae, are common to several species of pathogens, including Gram-negative and Gram-positive bacteria, whereas others are more specific. An effort is underway to elucidate the structures of surface molecules so as to unravel the molecular mechanisms of interactions [19–21]. Figure 2 shows a few examples of illustrated Protein Data Bank (PDB) structures.

Extensive research in the field defined several *in silico* markers that provide a rational basis for prediction. Such markers include regular expressions (sequence patterns) and matrix-based profiles specific to surface localized proteins. For example, the characterization of the trimeric autotransporters in Gram-negative bacteria facilitates their identification by a sequence profile [22], whereas a profile including the LPxTG motif identifies cell-wall anchoring domains in Gram-positives [23]. Finally, signatures specific to antigen subfamilies have been reported, such as two conserved motifs that characterize the ligand-binding regions of cell wall-anchored adhesins called MSCRAMMs (microbial surface components recognizing adhesive matrix molecules) [24].

More general approaches for the identification of cell-surface proteins exist, however. There are currently several computational methods available for the prediction of a protein's subcellular localization (SCL). Given that targeting of a protein to its ultimate destination is a complex process, reliable SCL prediction can be challenged by

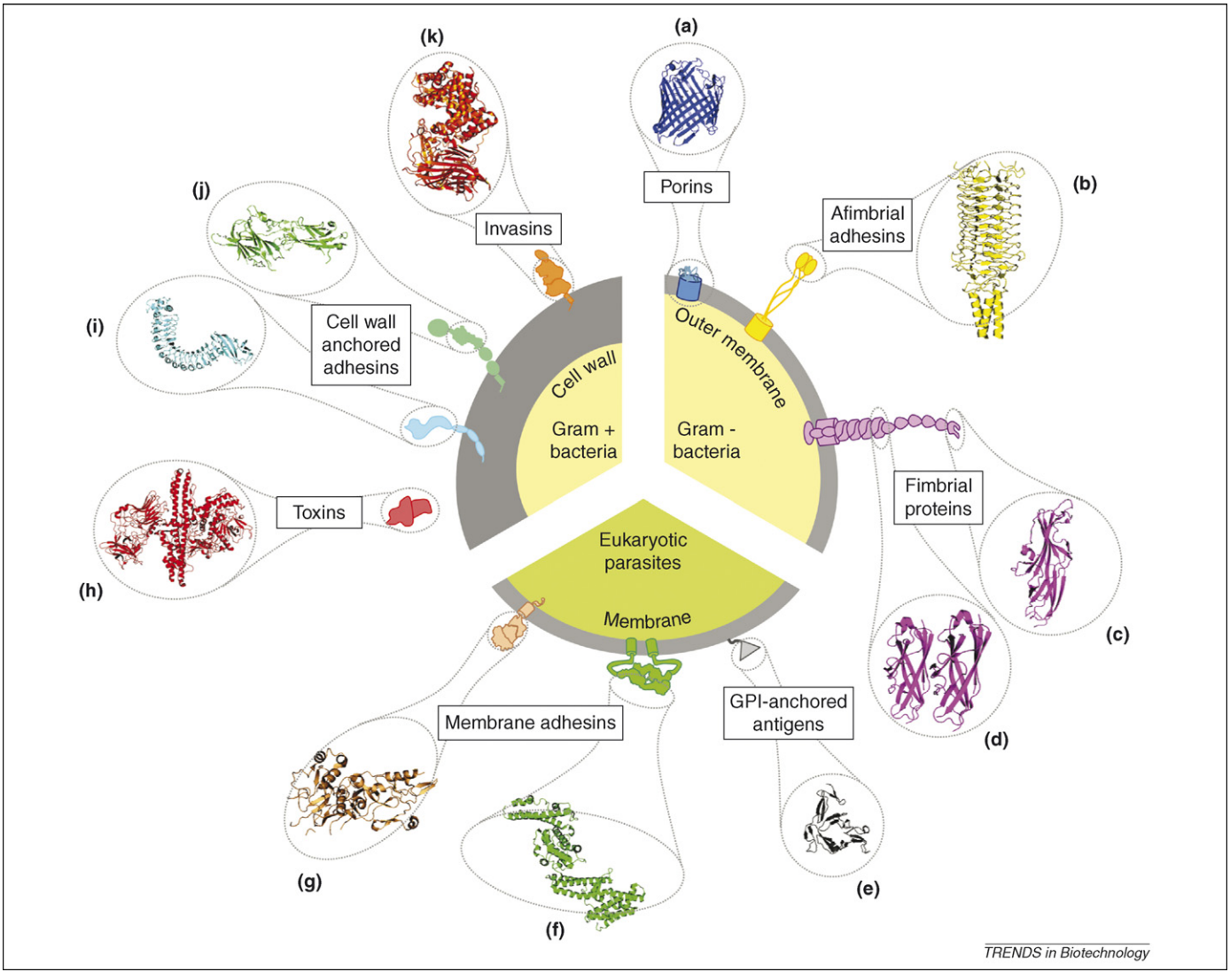

Figure 2. Structures of surface-exposed and secreted proteins relevant to the pathogen–host interface. Represented examples can be specific or common to different pathogen classes. The following indicates the antigen's name, pathogen's species and PDB codes, respectively, for each structure drawn: **(a)** OmpF, *Escherichia coli*, 1gfn; **(b)** YadA, *Yersinia pestis*, 1p9h; **(c)** PapG, *Escherichia coli*, 1j8s; **(d)** PapE-PapK, *Escherichia coli*, 1n12; **(e)** Psv25, *Plasmodium vivax*, 1z27; **(f)** EBA127, *Plasmodium falciparum*, 1zrl; **(g)** AMA-1, *Plasmodium vivax*, 1w8k; **(h)** BoNT-B, *Clostridium botulinum*, 1epw; **(i)** InlA, *Lysteria monocytogenes*, 1O6T; **(j)** SdrG, *Staphylococcus epidermidis*, 1r19; **(k)** SpnHI, *Streptococcus pneumoniae*, 1egu. Structural representations were generated with PyMOL (http://www.pymol.org).

integrative strategies (see below for a discussion of this approach to *in silico* analysis). Antigens can also be identified functionally. Adhesins, for example, mediate the first essential step of interaction between a pathogen and its cognate host receptor [25]. Therefore, developing strategies to abrogate this step have attracted considerable attention in recent years and several adhesins are currently being evaluated for inducing protective immunity. Important examples include HpaA from *Helicobacter pylori* [26], HMW1–HMW2-like adhesion proteins of non-typeable *Haemophilus influenzae* [27], the RgpA-Kgp proteinase-adhesin complex of *Porphyromonas gingivalis* [28,29], P97R1 adhesin from *Mycoplasma hyopneumoniae* [30], fibrinogen binding protein, fibronectin binding protein A, clumping factor A and collagen adhesin of *Staphylococcus aureus* [31], and FimH adhesin of *Escherichia coli* [32].

Why predict epitopes? The importance of data collection

Antigens can be processed into fragments called epitopes (see Glossary). The epitope is the immunological quantum, the smallest recognizable molecular entity in the immune system. Although the importance of non-peptide epitopes is now increasingly well understood, peptidic B-cell and T-cell epitopes (see Glossary), as mediated by the humoral or cellular immune systems, respectively, remain the principal tools by which the complexity of immune responses may be probed. MHC molecules are extremely polymorphic in their peptide-binding regions, and as a result, they exhibit widely varying binding specificity. Furthermore, different human leukocyte antigen (HLA) types are expressed at dramatically different frequencies by different ethnicities. A web-based tool to predict population coverage of T-cell epitope-based diagnostics and vaccines has been developed on the basis of MHC binding and/or T-cell restriction data. Accordingly, epitope-based vaccines or diagnostics can be designed to maximize population coverage, while minimizing complexity and variability of coverage obtained or projected in different ethnic groups [33].

It is the recognition of epitopes by T-cells, B-cells and soluble antibodies that lies at the heart of the immune response. Such recognition leads to activation of the cellular and humoral immune systems and, ultimately, to the effective destruction of pathogenic organisms. As the principal components of both subunit and poly-epitopic vaccines, the accurate prediction of B-cell and T-cell epitopes is thus a pivotal challenge for immuno-informatics. Successful application of T-cell epitope identification is now apparent in several areas: cancer therapy [34,35], epitope-vaccine strategy against infective agents, including HIV-1 [36], and from diagnostics [37] to allergic and autoimmune diseases [38,39]. Predicted T-cell epitopes have recently been validated [40].

By incorporating detailed knowledge about antigen processing into predictive methods, we can hope to significantly improve the reliability of epitope identification. Antigen presentation is a complex, multi-stage process. Specific models for individual steps in the processing and presentation pathway are available today. We can almost track and predict the whole path from antigen to presented epitope solely on the basis of its amino acid sequence. Proteasome restriction, binding to transporters associated with antigen presentation (TAP), MHC loading and TCR interaction can be modelled and incorporated into integrated prediction tools [41–44]. Such tools are limited by restricted experimental data [45], because the development of predictive methods depends heavily on large, accurate training sets [46]. This highlights the pivotal importance of reliable, detailed epitope databases. MHCPEP (MHC-binding peptides) [47] opened the way to the collection of sensible, usable data. AntiJen [48], MHCBN (MHC binders and non-binders) [49], BCIPEP (B-cell immunodominant peptides) [50], HaptenDB [51] and others have carried this work forward, several resources to be made available on the web. The collection of data concerning immune epitopes has been so crucial to become a research field itself; the need for a large but very reliable set of experimentally proved information puts the researchers in front of the problem of scrupulously screening a huge amount of literature. The Immune Epitope Database (IEDB) contains information on immune epitopes curated manually from the scientific literature, and it makes public a large set of quantitative peptide-binding affinity measurements relating to a set of different human and mammalian MHC class I alleles. A dedicated classifier was subsequently created to speed up the reference selection process without sacrificing sensitivity or specificity of the human expert classification [52–54].

Methods and approaches in epitope identification

Using experimental approaches to calculate the affinity between a given MHC molecule and an antigenic peptide is both difficult and time-consuming. Thus, various computational methods have been developed for the purpose.

Initially, 'direct' methods (i.e. aimed at predicting T-cell epitopes) were developed on the basis of the hypothesis that peptides recognized by T cells have amphipathic structures, with a hydrophobic side facing the MHC molecule and a hydrophilic side interacting with the TCR. The AMPHI algorithm and SOHHA (strip-of-helix hydrophobicity algorithm) are examples of direct methods [14,55]; recently, a direct method for predicting cytotoxic T lymphocyte (CTL) epitopes was developed, implementing machine-learning techniques that can discriminate T-cell epitopes and non-epitope MHC binders [55]. The above methods were superseded by the analysis of an MHC–peptide complex by X-ray crystallography, which resolved the extended conformation of a peptide bound to the MHC groove [56]. Presently, most available methods are considered to be indirect, because they predict MHC binders instead of T-cell epitopes. Current MHC binder prediction methods are based mainly on motifs, quantitative matrices (QMs, see Glossary), machine-learning techniques and *ab initio* prediction.

In motif-based approaches, an MHC-binding peptide of a particular allele is first examined for the presence of motifs comprising specific 'anchor residues' at specific 'anchor positions'. Antigen sequences are then scanned for these motifs to detect MHC binders [14]. However, both the absence of motifs in some MHC binders and the possible presence of secondary anchor residues at non-conserved positions can limit the usefulness of this

approach. QM-based methods perform better than motif-based ones, because they consider the contribution of each residue at each position in the peptides, instead of anchor positions and residues [57]; however, QM-based methods fail to handle non-linearity (see Glossary) in data. In fact, QMs are linear methods in which independent contribution (to affinity) of each position (residue) is assumed. Instead, real data are non-linear because they are not a simple summation of individual contributions of residues in binder. Machine-learning techniques have overcome this limit [58,59], because they can handle non-linearity in data and hence enable the prediction of MHC binders with a high accuracy [60]. A current major constraint of machine-learning techniques is that they require a large training set of MHC binder data, which is only available for a limited albeit increasing number of MHC alleles. A large dataset from the IEDB was used to establish a set of benchmark predictions with one neural network method and two matrix-based prediction methods [61]. Recently, discriminant analysis and multiple linear regression have been compared as algorithmic engines for the definition of QMs for binding affinity prediction. A matrix integrating the two methods proved powerfully predictive under cross-validation [62]. In addition, certain *ab initio* methods make predictions on the basis of the structural binding of the MHC molecule and peptides [63,64]. Such structure-based predictions are limited by a high computational cost and the need for three-dimensional imaging of the MHC-peptide complex. However, advances in computational power and the continuous flow of structural data are boosting structure-based prediction of both B-cell and T-cell epitopes. DiscoTope is a novel method that uses protein three-dimensional structural data for discontinuous epitope prediction. It is based on amino acid statistics, spatial information, and surface accessibility in a compiled dataset of discontinuous epitopes determined by X-ray crystallography of antibody–antigen protein complexes [65]. The PePSSI (peptide-MHC prediction of structure through solvated interfaces) algorithm has been developed for flexible structural prediction of peptide binding to MHC molecules. Several peptides that PePSSI predicts to be strong HLA-A2 binders are similarly predicted by another structure-based algorithm, PREDEP (prediction of MHC class I epitopes). Thus, structure-based prediction can identify potential peptide epitopes without known binding motifs, suggesting that side-chain orientation in binding peptides can be obtained using PePSSI [66].

Vaccine design on the basis of antigen and epitope identification

Epitope-based vaccine design A generally useful synthetic vaccine might contain one or more T-cell and B-cell epitopes, plus non-proteinaceous danger-signals. Furthermore, it can be an artificial poly-epitope vaccine or a natural antigen, possibly accompanied by administration of an adjuvant. Nonetheless, all effective vaccines engage directly with and are subsequently recognized by effectors of immune memory. Long gone are the days when vaccinology was a purely empirical endeavour. It is important to understand the hows and the whys of the immunological response to a pathogenic organism if we are going to successfully manipulate them during the design and development of new vaccines.

Vaccine design is also highly suited to the application of *in silico* techniques, for both the discovery and development of new and existing vaccines. The prediction of promiscuous MHC binders is crucial for the design of applicable subunit vaccines. The pioneering work from Parker *et al.* [67] and Rammensee *et al.* [68] opened the way to further investigations in this area. They developed a method to predict the relative binding strengths of all possible nonapeptides to MHC I molecules on the basis of experimental peptide binding data [67], and they provided a first listing of MHC ligands and peptide motifs [68]. Several methods were then developed to predict binders for a large number of MHC I [57,69] and MHC II [70,71] alleles. Further attempts have been made to develop comprehensive methods, and these also consider other components involved in antigen processing, such as the prediction of cleavage site [72,73] and TAP binders [42,44,74]. Several methods – these are mostly based on physiochemical properties – have been developed to predict continuous or linear and conformational epitopes (see Glossary). For instance, BEPITOPE is aimed at predicting the location of continuous B epitopes and patterns in proteins [75], whereas the CEP (conformational epitope prediction) server provides a web interface to a conformational epitope prediction algorithm that, apart from predicting conformational epitopes, also predicts antigenic determinants and sequential epitopes [76]. Existing methods are poor at predicting B-cell epitope [77,78], and thus immuno-informaticians are developing new tools, including machine-learning techniques [79]. Furthermore, a standardization of data formats through a broader collaboration between groups has also been suggested [80].

The discovery of discontinuous B-cell epitopes is a major challenge in vaccine design. New structure-based methods perform better at predicting residues of discontinuous epitopes than do methods based solely on sequence information. Predictions by such methods can guide experimental epitope-mapping in both the rational vaccine design and the development of diagnostic tools, and might lead to more efficient epitope identification [65].

Structure-based prediction of MHC–peptide association can also be of help in cancer vaccine design, because peptide vaccination for cancer immunotherapy requires identification of peptide epitopes derived from antigenic self-proteins associated with the tumour [66].

A combined immuno-informatics and structure-based modelling approach was followed to develop T-cell epitope-based vaccine design. MHCsim has been developed to generate specific MHC–peptide complex structures and to provide configuration files upon which to run molecular modelling simulations. It enables the automated construction of both MHC–peptide structure files and the corresponding configuration files required to execute a molecular dynamics simulation [81].

The role of secretory proteins of *Mycobacterium tuberculosis* in pathogenesis and stimulation of specific host responses is well documented. Therefore identification of T-cell epitopes from this set of proteins might serve to define candidate antigens with vaccine potential. A set of

Box 1. Machine-learning: how reliable data can affect tool development

The accuracy of a method depends on the technique and dataset used for training.

An artificial neural network (ANN) is an information processing architecture with three layers (input, hidden and output) of highly interconnected neurons; each connection is associated with a weight. Inspired by biological nervous system, ANNs have been widely used for pattern recognition and classification. Each neuron is designed to mimic its biological counterpart, accepting a weighted set of inputs and responding with an output. After training with known examples, the information for classification of new entities is stored in the form of weights associated with the connections between neurons. ANN allows development of a general classification method through which one can predict with high accuracy even in the absence of similarity between testing and training data. It can handle non-linear data and can perform very well in cases with a large number of training patterns. The major drawback is overfitting; in cases where learning was performed for too long or where training examples are rare, the learner may adjust to very specific random features of the training data that have no causal relation to the target function. In this process, the performance on the training examples still increases while the performance on unseen data becomes worse.

A hidden Markov model (HMM) is a statistical model initially developed for speech recognition and now widely used in bioinformatics. An HMM is a finite set of states, each of which is associated with a (generally multidimensional) probability distribution. Transitions among the states are governed by a set of transition probabilities. In a particular state, an outcome or observation can be generated according to the associated probability distribution. To an external observer, only the outcome is visible; the state itself is 'hidden'.

Support vector machines (SVMs) are universal approximators based on statistical learning and optimization theory. They support both regression and classification tasks and can handle multiple, continuous and categorical variables. To construct an optimal hyperplane, SVMs employ an iterative training algorithm that is used to minimize an error function. Hyperplanes are searched in the space of possible inputs; subsequently, these hyperplanes are used to separate positive and negative patterns. SVMs are based on statistical learning, which has the capability to handle linearity and non-linearity in a dataset. SVMs have minimum over-optimization even when training with a limited number of patterns.

While preparing training datasets with sequences we need to emphasize the following aspects: (i) quality, (ii) reliability, (iii) diversity, (iv) size, (v) redundancy and (vi) blind datasets.

(i) Quality: use well-annotated (preferably experimentally validated) data. It is best to avoid sequences with ambiguous annotation, conflicting experimental evidence, or those that were annotated through prediction themselves.
(ii) Reliability: it is preferable to collect sequences from reliable data sources (e.g. SWISS-PROT, RefSeq). Preference should be given to sequences annotated by more than one group or characterized by direct experimental evidence.
(iii) Diversity: it is important to obtain sequences from diverse sources (e.g. several species) to develop a widely applicable method.
(iv) Size: performance displays a positive correlation with the size of the dataset used for training. Thus attempts should be made to collect as many sequences as possible to develop accurate methods.
(v) Redundancy: redundant or highly similar sequences need to be removed from datasets to avoid biasing the algorithm towards groups of similar sequences.
(vi) Blind datasets: to avoid over-optimization, one should test a newly developed method on a blind or independent dataset not used in development of the method.

secretory proteins of *M. tuberculosis* H37Rv was analyzed computationally for the presence of HLA class I binding nonameric peptides. All possible overlapping nonameric peptide sequences from such secretory proteins were generated *in silico* and analyzed for their ability to bind to alleles belonging to the A, B and C loci of HLA class I. The structural basis for recognition of nonamers by different HLA molecules was studied by employing structural modelling of HLA class I–peptide complexes. This yielded a good correlation between structural analysis and binding prediction. Pathogen peptides that could behave as self- or partially self-peptides in the host were eliminated with the use of a comparative study with the human proteome, thus reducing the number of peptides for analysis [82].

The assessment of epitope conservation is dramatically important. In vaccine design, a high level of conservation can provide broader protection across multiple strains of pathogens. In diagnostic analyses or in monitoring diseases, a specific marker is obtained by looking for an epitope with a low level of conservation. In both these opposite cases, a tool for epitope conservation evaluation is fundamental [83].

Antigen-based vaccine design Adhesins are key molecules for the design of vaccines against microbial infective agents [84]. The acellular pertussis vaccine, currently approved for use against whooping cough caused by *Bordetella pertussis*, contains the filamentous adhesin hemagglutinin as one of its components [85]. Identification of adhesins and adhesin-like proteins is experimentally arduous, making bioinformatics the ideal strategy for their identification. Homology-based approaches are powerful in quickly identifying homologous sequences, but they can fail when the sequences become diverse. To overcome this limitation, machine-learning techniques and compositional features of molecules have been used to develop several algorithms that allow more focussed predictions. These algorithms are non-homology methods based on sequence properties. A step in this direction is the development of SPAAN (software program for prediction of adhesins and adhesin-like proteins using neural network) [86], which is based on highly curated datasets and neural networks that were optimally trained for compositional attributes (taking into account the considerations described in Box 1). The final probability of a given protein likely to be an adhesin is the weighted average of individual probabilities emerging from the five networks, based on the accuracy of each network. Adhesins had not previously been discovered in *Mycobacterium tuberculosis*, but SPAAN was able to predict several adhesins, which were independently confirmed by functional evidence; one example is Rv1818c, which was shown to display adhesin-like characteristics [87,88] and is being examined for vaccine design now [89]. Thus, the search for candidate adhesin proteins can be a most promising approach for identifying novel vaccine candidates.

An alternative approach seeks to identify putative subunit vaccines by finding antigens with genomic sequences; VaxiJen uses an alignment-free approach to identify antigens directly [90,91]. Instead of concentrating on predicting antigenic regions, the method uses statistical models

based on discriminating positive and negative sets of bacterial, viral and tumour antigen datasets to predict whole-protein antigenicity. VaxiJen has shown impressive prediction accuracy of up to 89%. Whole-antigen prediction is best used in conjunction to other methods within RV, such as subcellular localization prediction.

Integrative approaches in immuno-informatics

Integrating predictive tools to challenge compound problems

Early *in silico* predictive tools focused on using basic computational techniques to identify simple sequence features. The precision of these tools was improved thanks to new computational techniques such as 'approximate nearest neighbor' and 'hidden Markov models' (HMMs) (Box 1); then, more complex methods were developed.

The first instances of these more complex approaches came in the form of consensus algorithms for the improved identification of secondary structure elements. The strategy of combining the results of multiple methods into a consensus prediction is still employed in current methods. In the best examples, the predictive performance of a consensus method can be up to 40% higher than the performance of individual methods [92].

Integrative predictive strategies can be used to address more challenging predictive tasks. Rather than using multiple tools to predict the same sequence feature and then combining those results into a high-confidence prediction for that feature, multiple tools can be used to predict multiple sequence features, and the results combined to generate a prediction of a more complex biological characteristic of interest, achieving higher precision than ever before. This is crucial to biotechnological projects in immuno-informatics, wherein *in silico* steps have to challenge compound problems at multiple levels (molecule, cell, tissue, organ, organism and environment) rather than predicting a specific molecular moiety. To ensure an ultimately reliable prediction, both sequence features and computational techniques must be carefully selected. Complex biological attributes such as localization or function can be influenced by a multitude of features, and it is worth researching and including as many as possible to bolster a method's predictive success. It is important to consider, in addition to features that are predictive of a specific characteristic, those features that can be used to rule out one or more potential classes. For example, the presence of a signal peptide is insufficient to reliably assign a subcellular localization, but it provides evidence that a protein is not cytoplasmic.

Individual sequence feature prediction methods can exhibit varying levels of performance, from quite precise (few false positives) to less stringent (increased recall at the expense of precision). In situations in which the precision of the methods to be integrated varies, the prediction generated by each method should be weighted accordingly, with more value placed on predictions generated by the more precise techniques.

Integrative approaches in immuno-informatics

So far, bioinformatics research has developed and used a wide range of integrative strategies and tools to infer functional information. Tools to infer protein functions are of outstanding interest to immuno-informaticians as well as most to biotechnologists, and – in the never-ending story of obtaining fine and reliable predictions – two recent examples of integrative approaches can be presented. Proteome Analyst [93] extracts keywords from a query protein's SwissProt record, or from the SwissProt records of its nearest homologues. These keywords are then passed to multiple classifiers, each of which has been trained to classify a protein according to a particular feature. The program outputs a list of features predicted for a query protein, including function, active residues, interaction partners, enzymatic activity, cofactors, localization, ontologies and several more. ProFunc [94] accepts three-dimensional proteins structures as input and analyzes them at both the sequence and the structural level for features such as motifs, gene neighbours, conserved residues, folds, and binding sites, all of which provide significant clues to the protein's unique function.

Given that protein function is strongly dependent on subcellular localization (SCL), SCL predictors can also play a pivotal role in basic and applied research projects. For instance, these tools can rapidly and confidently identify cell surface-exposed proteins, which represent the most accessible pool of potential vaccine targets (see next section).

Reliable prediction of protein SCL is best tackled by integrative approaches, because targeting of a protein to its ultimate destination is influenced by, or correlated with, multiple sequence features, such as signal peptides, membrane-spanning regions, functional motifs and differences in amino acid composition unique to specific cellular compartments. Computationally identifying one of these features alone is not sufficient to infer where a protein is targeted to within a cell. However, by identifying several such features in a protein – some of these might support a specific SCL, others might rule out other cellular compartments – enough information can be gathered to make a reliable assessment of a protein's likely localization. PSORTb represents an example of an integrative environment able to yield extremely high-precision predictions of the location to which a bacterial protein is ultimately targeted [95]. In fact, the precision of certain integrative localization prediction methods, including PSORTb, has now surpassed that of high-throughput proteomics techniques [96]. For a list of further SCL predictors, and PSORT versions other than PSORTb, see supplementary material or http://www.psort.org.

On the basis of the results of comparative evaluations of multiple predictors [95,96], it has been recommended that users wishing to infer the cell-surface proteome carry out their analysis using PSORTb [97] for bacterial proteins, or Proteome Analyst for eukaryotic proteins [98]. These methods, at the time of evaluation, displayed the lowest false-positive rates both overall and for the localization sites in question, and their robust approach to prediction and simple user-interfaces makes them ideal for both small and large-scale analysis.

Prediction of a bacterial surface proteome was crucial for the beginning of the RV story and represented an example of how an *in silico* step can replace large exper-

Box 2. Foundations of a reverse vaccinology *in silico* environment

The underlying principles governing RV projects can be described as follows: (i) surface exposure of antigens is a fundamental requirement for the host–pathogen interaction; (ii) antigen conservation among different pathogen strains and serotypes should be taken into account to formulate widely protective vaccines. Recently, this has emerged as a central issue; group B *Streptococcus*, a multi-serotype bacterial pathogen, can adopt gene variability as a way to escape the immune response [104]. This variability makes it unlikely that a single universal protective antigen will be identified, and thus suggests that a combination of different antigens in a vaccine formulation is required; a high level of conservation among these antigens is the only way to achieve broad protection; (iii) the absence of molecular mimicry between vaccine candidates and human proteins is a must to avoid poor immunogenicity or autoimmunity. The need to avoid such problems was actually the driving force behind the conception of the RV approach; (iv) when present, sequence homology to proteins associated with virulence or immunogenicity is a strong element for antigen selection, regardless of SCL prediction; another related principle is (v), the absence of antigens in non-pathogenic strains, such that antigens that are specific to pathogenic strains are selected as putatively correlating with virulence; (vi) the ease of high-throughput heterologous expression is also an important element to consider [15]. Following the expression difficulties in the *Neisseria meningitidis* experiment, a cut-off of two and four transmembrane helices was applied to the RV studies of *Chlamydia pneumoniae* and *Bascillus anthracis*, respectively (multiple transmembrane helices can be indicative of a surface-exposed SCL, but, conversely, they are known to impair heterologous expression in *E. coli*).

Beyond these *in silico* criteria, DNA microarray analyses have been taken into consideration to tell us whether, when and to what extent antigens are expressed *in vitro* and *in vivo* during infection.

A step forward is to integrate, standardize, and automate analyses in order: (i) to minimize time costs in the process of vaccine candidate identification, without losing efficiency, but rather producing more information for finer discrimination and (ii) to provide the community with an accessible, dedicated RV environment. So far, NERVE (New Enhanced Reverse Vaccinology Environment) [105] represents the only attempt to bring together a similar body of principles into an automated pipeline resembling an RV process. Antigens are annotated according to multiple classification methods: (i) SCL prediction, (ii) adhesin probability (P_{ad}), (iii) number of transmembrane domains, (iv) homology to human proteins, (v) conservation in related strains, and (vi) putative function. A cut-off-based selection provides a final pool of vaccine candidates. This bundling of predictors illustrates the possibility of obtaining narrow vaccine candidate pools that are highly enriched in protective antigens. The flexibility of a modular RV environment allows for the inclusion of novel features likely to strengthen the capacity to represent biosequences as objects seen under the light of vaccine design, for example, the inclusion of more putative antigen features to be weighted in a score-based system. Quantitative methods can tell us not only which antigen is predicted to be vaccine candidate but also how suitable that seems to be. Then, the refinement of the classifying system is expected to further maximize the treatment of issues such as the risk of autoimmunity and ease of experimental steps, as well as new eventual questions that may prove relevant to antigen selection.

imental tasks; thus, it is clear that improved precision and reliability in the prediction of surface protein topology and class can further boost RV projects.

Reverse vaccinology

RV represents a successful and complete example of a computer-aided biotechnology. The availability of pathogen genomes and a pool of robust sequence analysis tools made it possible to implement the search for vaccine candidates on a genome scale as an *in silico* process [99]. Initially conceived as a way to escape major limitations in conventional experimental approaches, RV has become the standard approach to vaccine discovery in the current post-genomic era. Unlike the conventional identification of protective antigens amongst components of cultured pathogens, RV makes use of the whole spectrum of potential antigens. This allows vaccinologists to obtain pools of vaccine candidates that also include antigens that otherwise would be missed either because of poor or absent expression *in vitro* or because of the impossibility of culturing the pathogen. A reservoir of previously unrecognized vaccine candidates was instrumental in formulating a vaccine against *Neisseria meningitidis* serogroup B (MenB). Indeed, this pathogen subgroup was the only one whose surface-exposed polysaccharides were not included in the vaccines developed thus far. This is because of their significant similarity to human targets, which might potentially lead to weak immunogenicity or even autoimmunity. Sequencing the MenB genome and performing sequence alignment, motif scanning and SCL prediction analyses resulted in the identification of 570 putative surface-exposed proteins. Cloning and expression was unsuccessful for 220 of them, showing that it is important to take into account the relative ease of the high-throughput steps at the *in silico* stage. Serological analysis finally highlighted five vaccine candidates. This groundbreaking work kick-started the RV era in 2000 [15] and introduced a foundational approach for the *in silico* selection of vaccine candidates [100]. Since then, further RV projects from different groups around the world confirmed the validity of the 'reverse' approach and produced many vaccine candidates that showed promising preclinical results [101,102]. However, the manual curation of the *in silico* selection steps resulted in a certain variability in the methodologies applied (see supplemental Table S2). The need to standardize and automate RV resulted in the creation of a dedicated, new environment (see Box 2).

Conclusions

The continuous flow of 'omics' data and functional evidence from small-, medium- and large-scale 'basic science' and biotechnological projects is allowing bioinformaticians to move towards knowledge-based classification. Recently, sequence data from multiple strains of a single pathogen have enabled us to decipher virulence mechanisms. The comparison of multiple strains of a single species defines a 'pan-genome': a measure of the total gene repertoire that can pertain to a given microbial species. This will enable us to use a knowledge-based approach for the identification of vaccine candidates [103].

Improved computational techniques and combined integrative strategies are now able to provide researchers with high-confidence predictions for complex biological characteristics. This will herald a move to 'computer-aided biotechnology': prediction-driven smart projects in which time-consuming and expensive large-scale experimental approaches are progressively replaced by 'functional bioinformatics' investigations. To date, immuno-informatics

represents a powerful example of the functional bioinformatics approach. Reverse vaccinology demonstrates how immuno-informatics can and will underpin biotechnological research. The development of novel tools dedicated to the RV environment will help us address various sensible questions: to what extent, other than gene expression, are surface antigens actually accessible, considering antigen masking and protein degradation etc.? What does the antigen look like after eventual post-translation modification, especially in eukaryotic pathogens? How can we predict the best combination of vaccine candidates for a polyvalent vaccine formulation? Upon the analysis of the results obtained by the past RV works, the possibility of addressing these questions will greatly boost vaccine design in the future.

Acknowledgements

We wish to thank Christopher Woelk and Giovanni Mazzocco for useful discussions. S.V. and F.F. are supported by the Italian Ministry for University and Research (MIUR) and the University of Padua; J.L.G. and F.S.L.B. are supported by the Canadian Institutes for Health Research and the Michael Smith Foundation for Health Research; S.R. is a recipient of grant CMM0017, 'Drug target development using *in silico* biology', from the Council of Scientific and Industrial Research, India.

Supplementary data

Supplementary data associated with this article can be found, in the online version, at doi:10.1016/j.tibtech.2007.12.006.

References

1 McEntyre, J.R. and Gibson, T.J. (2004) Patterns and clusters within the PSM column in TiBS, 1992–2004. *Trends Biochem. Sci.* 29, 627–633

2 Boddy, M.N. *et al.* (1994) The p53-associated protein MDM2 contains a newly characterized zinc-binding domain called the RING finger. *Trends Biochem. Sci.* 19, 198–199

3 Aravind, L. and Koonin, E.V. (1998) The HD domain defines a new superfamily of metal-dependent phosphohydrolases. *Trends Biochem. Sci.* 23, 469–472

4 Filippini, F. *et al.* (2001) Longins: a new evolutionary conserved VAMP family sharing a novel SNARE domain. *Trends Biochem. Sci.* 26, 407–409

5 Oliver, S.G. (1996) From DNA sequence to biological function. *Nature* 379, 597–600

6 Filippini, F. *et al.* (1996) A plant oncogene as a phosphatase. *Nature* 379, 499–500

7 Emes, R.D. and Ponting, C.P. (2001) A new sequence motif linking lissencephaly, Treacher Collins and oral-facial-digital type 1 syndromes, microtubule dynamics and cell migration. *Hum. Mol. Genet.* 10, 2813–2820

8 Vacca, M. *et al.* (2001) MECP2 gene mutation analysis in the British and Italian Rett Syndrome patients: hot spot map of the most recurrent mutations and bioinformatic analysis of a new MECP2 conserved region. *Brain Dev.* 23, S246–S250

9 Carpi, A. *et al.* (2002) Comparative proteome bioinformatics: identification of a whole complement of putative protein tyrosine kinases in the model flowering plant *Arabidopsis thaliana*. *Proteomics* 2, 1494–1503

10 Li, D. *et al.* (2006) Protein interaction networks of *Saccharomyces cerevisiae*, *Caenorhabditis elegans* and *Drosophila melanogaster*: large-scale organization and robustness. *Proteomics* 6, 456–461

11 Flower, D.R. (2003) Databases and data mining for computational vaccinology. *Curr. Opin. Drug Discov. Devel.* 6, 396–400

12 Lefranc, M.P. (2004) IMGT-ONTOLOGY and IMGT databases, tools and Web resources for immunogenetics and immunoinformatics. *Mol. Immunol.* 40, 647–660

13 Brusic, V. and Petrovsky, N. (2003) Immunoinformatics–the new kid in town. *Novartis Found. Symp.* 254, 3–13 discussion 13–22, 98–101, 250–252

14 Korber, B. *et al.* (2006) Immunoinformatics comes of age. *PLoS Comput. Biol.* 2, e71

15 Pizza, M. *et al.* (2000) Identification of vaccine candidates against serogroup B meningococcus by whole-genome sequencing. *Science* 287, 1816–1820

16 Svennerholm, A.M. and Lundgren, A. (2007) Progress in vaccine development against *Helicobacter pylori*. *FEMS Immunol. Med. Microbiol.* 50, 146–156

17 Serruto, D. and Rappuoli, R. (2006) Post-genomic vaccine development. *FEBS Lett.* 580, 2985–2992

18 Rodrìguez-Ortega, M.J. *et al.* (2006) Characterization and identification of vaccine candidate proteins through analysis of the group A Streptococcus surface proteome. *Nat. Biotechnol.* 24, 191–197

19 Niemann, H.H. *et al.* (2004) Adhesins and invasins of pathogenic bacteria: a structural view. *Microbes. Infect.* 6, 101–112

20 Remaut, H. and Waksman, G. (2004) Structural biology of bacterial pathogenesis. *Curr. Opin. Struct. Biol.* 14, 161–170

21 Wilson, J.W. *et al.* (2002) Mechanisms of bacterial pathogenicity. *Postgrad. Med. J.* 78, 216–224

22 Cotter, S.E. *et al.* (2005) Trimeric autotransporters: a distinct subfamily of autotransporter proteins. *Trends Microbiol.* 13, 199–205

23 Telford, J.L. *et al.* (2006) Pili in Gram-positive pathogens. *Nat. Rev. Microbiol.* 4, 509–519

24 Ponnuraj, K. *et al.* (2003) A 'dock, lock, and latch' structural model for a staphylococcal adhesin binding to fibrinogen. *Cell* 115, 217–228

25 Sharon, N. (2006) Carbohydrates as future anti-adhesion drugs for infectious diseases. *Biochim. Biophys. Acta* 1760, 527–537

26 Nystrom, J. and Svennerholm, A.M. (2007) Oral immunization with HpaA affords therapeutic protective immunity against *H. pylori* that is reflected by specific mucosal immune responses. *Vaccine* 25, 2591–2598

27 Winter, L.E. and Barenkamp, S.J. (2006) Antibodies specific for the high-molecular-weight adhesion proteins of nontypeable *Haemophilus influenzae* are opsonophagocytic for both homologous and heterologous strains. *Clin. Vaccine Immunol.* 13, 1333–1342

28 Frazer, L.T. *et al.* (2006) Vaccination with recombinant adhesins from the RgpA-Kgp proteinase-adhesin complex protects against *Porphyromonas gingivalis* infection. *Vaccine* 24, 6542–6554

29 Yasaki-Inagaki, Y. *et al.* (2006) Production of protective antibodies against *Porphyromonas gingivalis* strains by immunization with recombinant gingipain domains. *FEMS Immunol. Med. Microbiol.* 47, 287–295

30 Chen, A.Y. *et al.* (2006) Evaluation of the immunogenicity of the P97R1 adhesin of *Mycoplasma hyopneumoniae* as a mucosal vaccine in mice. *J. Med. Microbiol.* 55, 923–929

31 Castagliuolo, I. *et al.* (2006) Mucosal genetic immunization against four adhesins protects against *Staphylococcus aureus*-induced mastitis in mice. *Vaccine* 24, 4393–4402

32 Poggio, T.V. *et al.* (2006) Intranasal immunization with a recombinant truncated FimH adhesin adjuvanted with CpG oligodeoxynucleotides protects mice against uropathogenic *Escherichia coli* challenge. *Can. J. Microbiol.* 52, 1093–1102

33 Bui, H.H. *et al.* (2006) Predicting population coverage of T-cell epitope-based diagnostics and vaccines. *BMC Bioinformatics* 7, 153

34 Lazoura, E. *et al.* (2006) Enhanced major histocompatibility complex class I binding and immune responses through anchor modification of the non-canonical tumour-associated mucin 1–8 peptide. *Immunology* 119, 306–316

35 Pietersz, G.A. *et al.* (2006) Design of peptide-based vaccines for cancer. *Curr. Med. Chem.* 13, 1591–1607

36 Liu, Z. *et al.* (2003) Epitope-vaccine strategy against HIV-1: today and tomorrow. *Immunobiology* 208, 423–428

37 Braga-Neto, U.M. and Marques, Jr E.T., (2006) From functional genomics to functional immunomics: new challenges, old problems, big rewards. *PLoS Comput. Biol.* 2, e81

38 Ali, F.R. and Larche, M. (2005) Peptide-based immunotherapy: a novel strategy for allergic disease. *Expert Rev. Vaccines* 4, 881–889

39 Atagunduz, P. *et al.* (2005) HLA-B27-Restricted CD8+ T Cell response to cartilage-derived self peptides in ankylosing spondylitis. *Arthritis Rheum.* 52, 892–901

40 Lundegaard, C. *et al.* (2006) The validity of predicted T-cell epitopes. *Trends Biotechnol.* 24, 537–538
41 Stevanovic, S. (2005) Antigen processing is predictable: from genes to T cell epitopes. *Transpl. Immunol.* 14, 171–174
42 Tenzer, S. *et al.* (2005) Modeling the MHC class I pathway by combining predictions of proteasomal cleavage, TAP transport and MHC class I binding. *Cell. Mol. Life Sci.* 62, 1025–1037
43 Doytchinova, I. and Flower, D. (2003) The HLA-A2 supermotif: A QSAR definition. *Org. Biomol. Chem.* 1, 2648–2654
44 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
45 Flower, D.R. (2003) Towards *in silico* prediction of immunogenic epitopes. *Trends Immunol.* 24, 667–674
46 Brusic, V. *et al.* (2004) Computational methods for prediction of T-cell epitopes - a framework for modelling, testing, and applications. *Methods* 34, 436–443
47 Brusic, V. *et al.* (1998) MHCPEP, a database of MHC-binding peptides: update 1997. *Nucleic Acids Res.* 26, 368–371
48 Blythe, M.J. *et al.* (2002) JenPep: a database of quantitative functional peptide data for immunology. *Bioinformatics* 18, 434–439
49 Bhasin, M. *et al.* (2003) MHCBN:A comprehensive database of MHC binding and non-binding peptides. *Bioinformatics* 19, 665–666
50 Saha, S. *et al.* (2005) Bcipep: a database of B-cell epitopes. *BMC Genomics* 6, 79
51 Singh, M.K. *et al.* (2006) HaptenDB: a comprehensive database of haptens, carrier proteins and anti-hapten antibodies. *Bioinformatics* 22, 253–255
52 Vita, R. *et al.* (2006) Curation of complex, context-dependent immunological data. *BMC Bioinformatics* 7, 341
53 Peters, B. and Sette, A. (2007) Integrating epitope data into the emerging web of biomedical knowledge resources. *Nat. Rev. Immunol.* 7, 485–490
54 Wang, P. *et al.* (2007) Automating document classification for the Immune Epitope Database. *BMC Bioinformatics* 8, 269
55 Bhasin, M. and Raghava, G.P.S. (2004) Prediction of CTL epitopes using QM, SVM and ANN techniques. *Vaccine* 22, 3195–3204
56 Tong, J.C. *et al.* (2004) Modeling the structure of bound peptide ligands to major histocompatibility complex. *Protein Sci.* 13, 2523–2532
57 Singh, H. and Raghava, G.P.S. (2003) ProPred1: Prediction of promiscuous MHC class-I binding sites. *Bioinformatics* 19, 1009–1014
58 Brusic, V. *et al.* (2002) Prediction of promiscuous peptides that bind HLA class I molecules. *Immunol. Cell Biol.* 80, 280–285
59 Dönnes, P. and Elofsson, A. (2002) Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics* 3, 25
60 Bhasin, M. and Raghava, G.P. (2004) SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. *Bioinformatics* 20, 421–423
61 Peters, B. *et al.* (2006) A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Comput. Biol.* 2, e65
62 Doytchinova, I.A. and Flower, D.R. (2007) Predicting class I major histocompatibility complex (MHC) binders using multivariate statistics: comparison of discriminant analysis and multiple linear regression. *J. Chem. Inf. Model.* 47, 234–238
63 Schueler-Furman, O. *et al.* (2000) Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein Sci.* 9, 1838–1846
64 Doytchinova, I.A. and Flower, D.R. (2001) Toward the quantitative prediction of T-cell epitopes: coMFA and coMSIA studies of peptides with affinity for the class I MHC molecule HLA-A*0201. *J. Med. Chem.* 44, 3572–3581
65 Haste Andersen, P. *et al.* (2006) Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci.* 15, 2558–2567
66 Schiewe, A.J. and Haworth, I.S. (2007) Structure-based prediction of MHC-peptide association: Algorithm comparison and application to cancer vaccine design. *J. Mol. Graph. Model.* 26, 667–675
67 Parker, K.C. *et al.* (1994) Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol.* 152, 163–175
68 Rammensee, H.G. *et al.* (1995) MHC ligands and peptide motifs: first listing. *Immunogenetics* 41, 178–228
69 Bhasin, M. and Raghava, G.P.S. (2007) A hybrid approach for predicting promiscuous MHC class I restricted T cell epitopes. *J. Biosci.* 32, 31–42
70 Sturniolo, T. *et al.* (1999) Generation of tissue-specific and promiscuous HLA ligand database using DNA microarrays and virtual HLA class II matrices. *Nat. Biotechnol.* 17, 555–561
71 Singh, H. and Raghava, G.P.S. (2001) ProPred: prediction of HLA-DR binding sites. *Bioinformatics* 17, 1236–1237
72 Kesmir, C. *et al.* (2002) Prediction of proteasome cleavage motifs by neural networks. *Protein Eng.* 15, 287–296
73 Bhasin, M. and Raghava, G.P.S. (2005) Pcleavage: A SVM based method for prediction of constitutive and immuno proteasome cleavage sites in antigenic sequences. *Nucleic Acids Res.* 33, W202–W207
74 Bhasin, M. and Raghava, G.P.S. (2004) Analysis and prediction of affinity of TAP binding peptides using cascade SVM. *Protein Sci.* 13, 596–607
75 Odorico, M. and Pellequer, J.L. (2003) BEPITOPE: predicting the location of continuous epitope and patterns in proteins. *J. Mol. Recognit.* 16, 20–22
76 Kulkarni-Kale, U. *et al.* (2005) CEP: a conformational epitope prediction server. *Nucleic Acids Res.* 33, W168–W171
77 Blythe, M.J. and Flower, D.R. (2005) Benchmarking B cell epitope prediction: underperformance of existing methods. *Protein Sci.* 14, 246–248
78 Saha, S. and Raghava, G.P.S. (2004) BcePred: prediction of continuous B-cell epitopes in antigenic sequences using physico-chemical properties. In *Artificial Immune Systems* (Nicosia, G. *et al.*, eds), pp. 197–204, Springer Publisher
79 Saha, S. and Raghava, G.P.S. (2006) Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins* 65, 40–48
80 Greenbaum, J.A. *et al.* (2007) Towards a consensus on datasets and evaluation metrics for developing B cell epitope prediction tools. *J. Mol. Recognit.* 20, 75–82
81 Todman, S.J. *et al.* (2008) Toward the atomistic simulation of T cell epitopes Automated construction of MHC: Peptide structures for free energy calculations. *J. Mol. Graph. Model.* 26, 957–961
82 Vani, J. *et al.* (2006) A combined immuno-informatics and structure-based modeling approach for prediction of T cell epitopes of secretory proteins of *Mycobacterium tuberculosis*. *Microbes Infect.* 8, 738–746
83 Bui, H.H. *et al.* (2007) Development of an epitope conservancy analysis tool to facilitate the design of epitope-based diagnostics and vaccines. *BMC Bioinformatics* 8, 361
84 Wizemann, T.M. *et al.* (1999) Adhesins as targets for vaccine development. *Emerg. Infect. Dis.* 5, 395–403
85 Colombi, D. *et al.* (2006) Haemagglutination induced by *Bordetella pertussis* filamentous haemagglutinin adhesin (FHA) is inhibited by antibodies produced against $FHA_{430\text{-}873}$ fragment expressed in *Lactobacillus casei*. *Curr. Microbiol.* 53, 462–466
86 Sachdeva, G. *et al.* (2005) SPAAN: a software program for prediction of adhesins and adhesin-like proteins using neural networks. *Bioinformatics* 21, 483–491
87 Brennan, M.J. *et al.* (2001) Evidence that mycobacterial PE_PGRS proteins are cell surface constituents that influence interactions with other cells. *Infect. Immun.* 69, 7326–7333
88 Delogu, G. *et al.* (2004) Rv1818c-encoded PE_PGRS protein of *Mycobacterium tuberculosis* is surface exposed and influences bacterial cell structure. *Mol. Microbiol.* 52, 725–733
89 Chaitra, M.G. *et al.* (2007) Evaluation of T-cell response to peptides with MHC class I-binding motifs derived from PE_PGRS 33 protein of *Mycobacterium tuberculosis*. *J. Med. Microbiol.* 56, 466–474
90 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
91 Doytchinova, I.A. and Flower, D.R. (2007) Identifying candidate subunit vaccines using an alignment-independent method based on principal amino acid properties. *Vaccine* 25, 856–866
92 Arai, M. *et al.* (2004) ConPred II: a consensus prediction method for obtaining transmembrane topology models with high reliability. *Nucleic Acids Res.* 32, W390–W393

93 Szafron, D. *et al.* (2004) Proteome Analyst: custome predictions with explanations in a web-based tools for high-throughput proteome annotations. *Nucleic Acids Res.* 32, W365–W371
94 Laskowski, R.A. *et al.* (2005) ProFunc: a server for prediction protein function from 3D structure. *Nucleic Acids Res.* 33, W89–W93
95 Gardy, J.L. and Brinkman, F.S.L. (2006) Methods for predicting bacterial protein subcellular localization. *Nat. Rev. Microbiol.* 4, 741–751
96 Rey, S. *et al.* (2005) Assessing the precision of high-throughput computational and laboratory approaches for the genome-wide identification of protein subcellular localization in bacteria. *BMC Genomics* 6, 162
97 Gardy, J.L. *et al.* (2005) PSORTb v.2.0: expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. *Bioinformatics* 21, 617–623
98 Lu, Z. *et al.* (2004) Predicting subcellular localization of proteins using machine-learned classifiers. *Bioinformatics* 20, 547–556
99 Rappuoli, R. (2000) Reverse vaccinology. *Curr. Opin. Microbiol.* 3, 445–450
100 Serruto, D. *et al.* (2004) From genome to vaccine. In *Genomics, Proteomics and Vaccines* (Grandi, G., ed.), pp. 185–201, John Wiley & Sons Ltd
101 De Groot, A.S. and Rappuoli, R. (2004) Genome-derived vaccines. *Expert Rev. Vaccines* 3, 59–76
102 Davies, M.N. and Flower, D.R. (2007) Harnessing bioinformatics to discover new vaccines. *Drug Discov. Today* 12, 389–395
103 Muzzi, A. *et al.* (2007) The pan-genome: towards a knowledge-based discovery of novel targets for vaccines and antibacterials. *Drug Discov. Today* 12, 429–439
104 Maione, D. *et al.* (2005) Identification of a universal Group B streptococcus vaccine by multiple genome screen. *Science* 309, 148–150
105 Vivona, S. *et al.* (2006) NERVE: new enhanced reverse vaccinolgy environment. *BMC Biotechnol.* 6, 35