Expert Opinion

Using databases and data mining in vaccinology

Matthew N Davies, Pingping Guan, Martin J Blythe, Jesper Salomon, Christopher P Toseland, Channa Hattotuwagama, Valerie Walshe, Irini A Doytchinova & Darren R Flower†
† *The Jenner Institute, University of Oxford, Compton, Berkshire, RG20 7NN, UK*

Throughout time functional immunology has accumulated vast amounts of quantitative and qualitative data relevant to the design and discovery of vaccines. Such data includes, but is not limited to, components of the host and pathogen genome (including antigens and virulence factors), T- and B-cell epitopes and other components of the antigen presentation pathway and allergens. In this review the authors discuss a range of databases that archive such data. Built on such information, increasingly sophisticated data mining techniques have developed that create predictive models of utilitarian value. With special reference to epitope data, the authors discuss the strengths and weaknesses of the available techniques and how they can aid computer-aided vaccine design deliver added value for vaccinology.

Keywords: antigen, database, data mining, epitope, immunoinformatics, vaccine

Expert Opin. Drug Discov. (2007) 2(1):19-35

1. Introduction

'*In science there is only physics; all the rest is stamp collecting*' is one rendering of a famous, if somewhat contemptuous, quotation from Lord Rutherford (1871 – 1937). Thus, a lucid distinction between data and understanding in science has long been clear. Increasingly, science is data rather than hypothesis driven. In the high-throughput era – be that astronomical, metrological or postgenomic – data generation is no longer the bottleneck; instead it is the ability to interpret data usefully that limits us. After a century of empirical research, immunology and vaccinology are poised to reinvent themselves as genome-based, high-throughput sciences. They too must face the challenge of capitalising on a potentially overwhelming inundation of new data, which is both dazzlingly complex and delivered on a hitherto inconscionable scale. It is only by fully embracing computation that such a goal will be achieved.

A vaccine is a molecular or supramolecular agent that elicits specific, protective immunity, an enhanced adaptive immune response to reinfection against pathogenic microbes and the diseases they cause, by the potentiation of immune memory and ultimately mitigating the effect of subsequent infection. Mass vaccination, which takes account of so-called herd immunity, is now widely accepted to be among the most efficacious prophylactic treatments for both infectious and many chronic diseases, including *inter alia* cancers and allergies. Although vaccines remain a small part of the global therapeutics market, their potential for growth, in an era characterised by rapid changes in the demographic need for new and improved vaccines, is unbounded. Historically, vaccination has been undertaken using attenuated whole-pathogen vaccines, such as Bacillus of Calmette and Guerin (BCG) for tuberculosis (TB) or Sabin's Polio vaccine. Safety concerns have seen the development of new strategies for vaccine development, separately focusing on antigen and epitope vaccines. The hepatitis B vaccine is an example of an antigen – or subunit – vaccine, and many epitope-based vaccines have entered clinical trials. A potentially useful vaccine might contain one or more immunogenic T-cell epitope, one or more

informa healthcare

immunogenic B-cell epitope, plus non-proteinacous 'danger signals', and may be an artificial polyepitope vaccine or a natural antigen, delivered as protein, via vectors, or as raw DNA, possibly together with an adjuvant (a molecule or preparation that exacerbates an immune response). Coupling immunogenicity prediction to the need for improved delivery mechanisms and adjuvants makes computational vaccinology both a challenging and an exciting area of therapeutic discovery.

Meaningful and quantitative molecular and functional data underlies attempts to predict those aspects of immunological systems that are capable of being predicted. Immunoinformatic computation can identify undiscovered links within data sets, yet currently it is not possible, for example, to identify unknown components of pathways or uncover protein function, other than by establishing sequence or structural similarity to extant protein exemplars. Thus, there are limits to what computational vaccinology can achieve as well as immense opportunities to exploit its potential. It is important to realise what can be done and what can not be done, what is useful and what is not. What immunoinformatics can offer is tools and methods that form part of a wider experimental and clinical endeavour. It offers a set of techniques replete with utilitarian value that can be leveraged by computational vaccinology to facilitate the design and discovery of vaccines. To be useful, data we wish to model and predict must be properly accumulated and archived. This is the role of the database. Once data has been stored, it must be analysed. This is the province of data mining. The authors explore these issues separately below.

2. Databases

Databases in immunology and, thus, vaccinology, have a long history. Databases such as Kabat [1] and IMGT [2] have concentrated on the compilation and high quality annotation of host side sequences and structures. In the 1960s, Elvin Kabat and Tai Te Wu began the Kabat Database, initiating the collection and alignment of human and mouse immunoglobulin light chain sequences. The Kabat Database is among the oldest biological databases, and was for some time the only database containing sequence alignment information. Concentrating initially on Human sequences, Marie-Paule Lefranc's IMGT database system has a rich – some say bewildering – complexity that sets it apart from other databases.

Recently, the Immuno Polymorphism Database (IPD) system [3], a set of specialist databases facilitating the study of polymorphic genes in the immune system, has recently emerged from the IMGT's long shadow. IPD focuses on a variety of data and importantly looks at non-human species, such as non-human primates, cattle and sheep and, thus, extends work in non-primates beyond laboratory model animals into commercially important farm livestock. IPD currently consists of four databases: i) IPD-KIR containing alleles of killer-cell immunoglobulin-like receptors; ii) IPD-MHC containing major histocompatibility complex (MHC) sequences from different species; iii) IPD-HPA containing human platelet alloantigens antigens; and iv) IPD-ESTDAB, which allows access to a database of melanoma cell lines. Another important database in the host area is VBASE2, which stores germ line sequences of human and mouse immunoglobulin variable (V) genes [4]. An omission to extant databases, such as IMGT or IPD, is the lack of a mouse MHC database with a sound and consolidated nomenclature. The mouse is, arguably, the most convenient model organism for vaccine development. However, efforts to compile such a database, or even concoct such a nomenclature, have made little progress since the late 1980s.

With the advent, and ceaseless progression, of genomics and the genomic sequencing of microbial genomes, the specialist pathogen database has emerged. These are now legion. Well in excess of 2500 genomes are now available: > 200 from bacteria, > 1200 from viruses, > 600 from plasmids, > 30 eukaryotes and > 500 from organelles. These values will be long superceded before this review is published in 2007. Pathogens represent a small, but exceptionally important, subset of these genomes. Apart from databases devoted to HIV and hepatitis C virus (HCV), oral pathogens are particularly well served [5,6]. Other representative examples are listed in **Table 1**.

Positioned between databases that concentrate on host or pathogens separately are resources that focus on host–pathogen interactions. This is an area ripe for exploration, and one that overlaps strongly with transcriptomic and proteomic databases. The best extant example is the area of so-called virulence factors (VF): elements that enable a pathogen to successfully colonise its host or cause disease. Analysis of pathogens – such as *Vibrio cholerae* or *Streptococcus pyogenes* – has identified coordinated 'systems' of toxins and virulence factors, which may consist of > 40 distinct proteins. Traditionally, VFs, which are usually, but exclusively, secreted or outer membrane proteins, have been classified as adherence/colonisation factors, invasions, exotoxins, transporters, iron-binding siderophores and miscellaneous cell surface factors. A broader definition groups VFs into three: i) 'true' VF genes; ii) VFs associated with the expression of 'true' VF genes; and iii) VF 'lifestyle' genes required for colonisation of the host [7].

The Virulence Factors Database (VFDB) contains 16 characterised bacterial genomes with an emphasis on functional and structural biology and can be searched using text, BLAST or functional queries [8]. The ClinMalDB-US database is being establishing following the discovery of multi-gene families encoding VFs within the subtelomeric regions of *P. falciparum* and *P. vivax* [9,10]. TVFac (Los Alamos National Laboratory Toxin & Virulence Factor database) contains genetic information on > 250 organisms and separate records for thousands of virulence genes and associated factors. The Fish Pathogen Database, set up by the Bacteriology & Fish Diseases Laboratory, has identified > 500 virulence genes using the fish as a model system. Pathogens studied include *Aeromonas hydrophila*, *Edwardsiella tarda* and many *Vibrio* species. *Candida albicans* Virulence Factor (CandiVF) is a

Table 1. Immunological databases

Host databases		
IMGT/HLA	http://www.ebi.ac.uk/imgt/hla/allele.html	Aligned and annotated HLA sequences following the WHO nomenclature
IMGT/TR	http://imgt.cines.fr/textes/IMGTrepertoire	Aligned and annotated T-cell receptor sequences
IPD database	http://www.ebi.ac.uk/ipd/index.html	Database providing a centralised repository for the data that define the human platelet antigens
Kabat	http://www.kabatdatabase.com/	Recently commercialised antibody database
VBASE	http://www.vbase2.org/	Integrated database of germ-line variable genes from immunoglobulin loci of human and mouse
ABG	http://www.ibt.unam.mx/vir/	Germline gene directories of mouse
V BASE	http://vbase.mrc-cpe.cam.ac.uk/	Database of human antibody genes
Pathogen databases		
APB	http://www.engr.psu.edu/ae/iec/abe/database.asp	Airbourne pathogen database maintained by aerobiological engineering department, Penn State
APDD	http://psychro.bioinformatics.unsw.edu.au/pathogen/index.php	Database designed to facilitate the detection of contaminant sequences in the human branch of dbEST that may be derived from Archaea
ARS	http://www.ars.usda.gov/research/projects/projects.htm?accn_no=406518	Developing database containing molecular sequence information from viral, bacterial and fungal plant pathogens
BROP	http://www.brop.org/	Bioinformatics resource for oral pathogens
EDWIP	http://cricket.inhs.uiuc.edu/edwipweb/edwipabout.htm	Ecological database of the world's insect pathogens
FPPD	http://fppd.cbio.psu.edu/	Fungal plant pathogen database catalogues the genotypes and phenotypes of fungal/oomycetes plant pathogens
LEGER	http://leger2.gbf.de/cgi-bin/expLeger.pl	Post-genome database for *Listeria* research
ORALGEN	http://www.oralgen.lanl.gov/	Database contains molecular information pertaining to oral pathogens, bacterial and viral
Pathema	http://www.tigr.org/pathema/index.shtml	In depth curatorial analysis of six target organisms from the list of NIAID category A – C pathogens
ShiBASE	http://www.mgc.ac.cn/ShiBASE/	Genomic data on *Shigella*, a group of Gram-negative, facultative intracellular pathogens
STDGen	http://www.stdgen.lanl.gov/	Compilation and analysis of molecular sequence information pertaining to sexually transmitted bacteria and viruses
VBI	http://phytophthora.vbi.vt.edu/	Microbial database at VBI hosts data from a range of plant pathogenic oomycetes, fungi and bacteria
VIDIL	http://insectweb.inhs.uiuc.edu/Pathogens/VIDIL/index.html	Viral diseases of insects in the literature database
Virulence factor databases		
VFDB	http://zdsys.chgb.org.cn/VFs/main.htm	Reference database for bacterial virulence factors
CandiVF TVfac	http://www.tvfac.lanl.gov/	Toxin and virulence factor database for > 250 organisms
PRINTS	http://www.jenner.ac.uk/BacBix3/PPprints.htm	All virulence factors situated in the PRINTS protein fingerprint
ClinMalDB-USP	http://malariadb.ime.usp.br/malaria/us/bioinformaticResearch.jsp	Developing database containing multi-gene families that encode virulent determinants
Fish pathogen database	http://dbsdb.nus.edu.sg/fpdb/about.html	> 500 virulence genes identified using fish as a model system
PHI-BASE	http://www.phi-base.org/	Integrated host–pathogen database

HCV: Hepatitis C virus; HLA: Human leukocyte antigen; MHC: Major histocompatibility complex; NIAID; National Institute of Allergy and Infectious Diseases; TAP: Transannular patch; TCR: T-cell receptor.

Table 1. Immunological databases (continued)

T-cell databases		
AntiJen	http://www.jenner.ac.uk/antijen/aj_tcell.htm	Quantitative binding data for MHC–ligand interactions, TCR–MHC complexes, TAP
EPIMHC	http://bio.dfci.harvard.edu/epimhc/	Curated Database of MHC Ligands
FIMM	http://research.i2r.a-star.edu.sg/fimm/	An integrated functional immunology database, focusing on MHC, protein antigens, antigenic peptides, and relevant disease information
HLA ligand database	http://hlaligand.ouhsc.edu/index_2.html	Legacy Repository of MHC binding data
HIV Immunology	http://www.hiv.lanl.gov/immunology	$CD8^+$ and $CD4^+$ T-cell HIV epitopes, proteome epitope maps
HCV Immunology	http://hcv.lanl.gov/content/immuno/immuno-main.html	$CD8^+$ and $CD4^+$ T-cell HCV epitopes, proteome epitope maps
IEDB	http://epitope2.immuneepitope.org/home.do	β-Version of biothreat pathogen T-cell epitope database β-release
JenPep	http://www.jenner.ac.uk/jenpep2/	Legacy Repository of MHC binding data
MHCBN	http://www.imtech.res.in/raghava/mhcbn	MHC-peptide binders and non-binders, TAP-peptide binders and non-binders, T-cell epitopes
MHCPEP	http://wehih.wehi.edu.au/mhcpep	MHC-presented epitopes
MPID-T	http://surya.bic.nus.edu.sg/mpidt/	Structual data for MHC–peptide–TCR interaction
SYFPEITHI	http://www.syfpeithi.de	MHC-presented epitopes, MHC-specific anchor and auxiliary motifs
B-cell databases		
AntiJen	http://www.jenner.ac.uk/antijen/aj_bcell.htm	Quantitative binding data for B-cell epitopes
BCIPEP	http://www.imtech.res.in/raghava/bcipep	B-cell epitope database
CED	http://web.kuicr.kyoto-u.ac.jp/~ced/	Conformational epitope database
EPITOME	http://www.rostlab.org/services/epitome/	Database of structurally inferred antigenic epitopes in proteins
IEDB	http://epitope2.immuneepitope.org/home.do	β-Version of biothreat pathogen B-cell epitope database β-release
HaptenDB	http://www.imtech.res.in/raghava/haptendb/	Database of haptens
HIV immunology	http://www.hiv.lanl.gov/immunology	B-cell HIV epitopes, pathogen proteome linear epitope maps, extensive literature citations regarding mAbs, curated epitope alignments
HCV immunology	http://hcv.lanl.gov/immuno/	B-cell HCV epitopes, pathogen proteome linear epitope maps, extensive literature citations regarding mAbs, curated epitope alignments
Allergen databases		
ALLALLERGY	http://www.allallergy.net/	Database for information on specific allergens
ALLERDB	http://sdmc.i2r.a-star.edu.sg/Templar/DB/Allergen/	Allergen database with integrated tools
Allergen database	http://allergen.csl.gov.uk/	Indexing facility for the retrieval of information on allergens and epitopes
Allergome	http://www.allergome.org/	List of allergen molecules and their biological functions
AllerMatch	http://www.allermatch.org/	Sequences database of allergenic proteins found in food
BIFS	http://www.iit.edu/~sgendel/	Database of food and foodbourne pathogens
CSL	http://www.csl.gov.uk/allergen/	Sequence database search and allergen analysis tools
FARRP	http://www.allergenonline.com/	Database of proteins of known and putative allergens (food, environmental and contact)

HCV: Hepatitis C virus; HLA: Human leukocyte antigen; MHC: Major histocompatibility complex; NIAID; National Institute of Allergy and Infectious Diseases; TAP: Transannular patch; TCR: T-cell receptor.

Table 1. Immunological databases (continued)

Allergen databases		
IMGT	http://imgt.cines.fr/	International immunogenetics information system
InformALL	http://foodallergens.ifr.ac.uk/	Database with information of allergenic foods
IUIS Allergen Nomenclature	http://www.allergen.org	Official list of recognised allergens
SDAP	http://fermi.utmb.edu/SDAP/	Structural database of allergenic proteins
SWISS-PROT AllergenIndex	http://www.expasy.org/cgi-bin/lists?allergen.txt	Nomenclature and index of allergen sequences in SWISS-PROT

HCV: Hepatitis C virus; HLA: Human leukocyte antigen; MHC: Major histocompatibility complex; NIAID; National Institute of Allergy and Infectious Diseases; TAP: Transannular patch; TCR: T-cell receptor.

small species-specific database that contains VFs that may be searched using BLAST or a human leukocyte antigen (HLA)-DR Hotspot Prediction server [11]. PHI-BASE is a noteworthy development, as it seeks to integrate a wide range of VFs from a variety of pathogens of plants and animals [12].

Recently, functional databases (those focusing on how the immune system operates), have begun to proliferate (see **Table 1**). A neat division can be made between databases that have as their primary focus T-cell epitope data and those that concentrate on B-cell epitopes. Historically, T-cell data came first, and has now reached a significant level of sophistication, whereas B-cell data has long lagged behind, although this imbalance is now being redressed.

3. T-cell databases

There have for a long time been databases, such as SYFPEITHI [13], which focus on properties of cellular immunology and look primarily at data relevant to MHC processing, presentation and T-cell recognition (see **Table 1**). In recent years there has been a flurry of new and improved databases that focus on cellular immunology. The authors discuss an illustrative cross-section of these below.

Arguably, the best database available for use is the HIV Molecular Immunology Database [14], although its obvious depth is at the expense of breadth and generality. It archives $CD4^+$ and $CD8^+$ T- and B-cell epitopes derived from the virus. Features of the database include viral protein epitope maps, sequence alignments, drug-resistant viral protein sequences and vaccine-trial data (Los Alamos National Laboratory [2001]). At present, the HIV $CD8^+$ T-cell epitope database contains 3150 entries describing 1600 distinct MHC class I–epitope combinations; the HCV database contains 510 entries describing 250 distinct MHC class I–epitope combinations [15]. It also includes detailed biological information regarding the response to the epitope, including its impact on long-term survival, common escape mutations, whether an epitope is recognised in early infection and curated alignments summarising the epitope's global variability. To coin a phrase: 'one day all databases will look this way'.

Other recent databases include MHCBN [16], which contains 18,790 MHC-binding peptides, 3227 MHC non-binding peptides, 1053 transannular patch (TAP) binders and non-binders and 6548 T-cell epitopes. EPIMHC [17] is a relational database of naturally occurring MHC-binding peptides and T-cell epitopes. At present, the database includes 4867 distinct peptide sequences from various sources, including 84 tumour antigens. Another database of interest is T-MPID [18]. Focusing primarily on T-cell receptor (TCR)/pMHC interactions, MPID-T is a manually curated MySQL database that includes experimentally determined structures of 187 pMHC complexes and 16 TCR/pMHC complexes. Two databases in particular, warrant special attention, albeit for different reasons. They are AntiJen [19] and IEDB [20].

AntiJen is a laudable – if flawed – attempt to integrate a wider range data than is archived by other databases. Implemented as a relational postgreSQL database, AntiJen, which is sourced from the primary literature and contains > 24,000 entries, includes quantitative kinetic, thermodynamic, functional and cellular data within the context of immunology and vaccinology. As well as T-cell and MHC-binding data, AntiJen holds > 3500 entries for linear and discontinuous B-cell epitopes, and includes measurements of peptide interactions with TAP transporter and peptide–MHC complex interactions with TCRs, as well as immunological protein–protein interactions.

Conversely, Immune Epitope Database and Analysis Resource (IEDB) is an NIH database that addresses issues of biodefence. It is on a larger scale than databases that have existed hitherto, and benefits from the input of 13 dedicated epitope sequencing products that exist, in part, to populate the database. A β-version of the new IEDB has recently come online that will focus on epitopes in potential bioterrorism agents or emerging infectious diseases. IEDB may yet eclipse all other efforts in functional immune databases.

4. B-cell databases

In recent years there has been a move to incorporate data on so-called B-cell epitopes into functional immunology databases (see **Table 1**). Antibodies, which mediate B-cell epitopes, are an

intrinsic component of the adaptive immune response for higher invertebrates. Their binding sites – or epitopes – are found within protein antigens and can be classified as 'linear' or 'discontinuous'. In addition to Korber's HIV and HCV databases, that also contain B-cell data, there are a variety of others which also archive B-cell epitopes and associated data. These include several examples discussed above, such as AntiJen and IEDB. Other examples focus exclusively on B-cell data; they include BciPep [21], CED [22] and Epitome [23].

The Bcipep contains 3031 linear B-cell epitopes sequences, sourced from primary literature via PUBMED links, within a web accessible PostgreSQL relational database. Epitopes are presented as annotated sequences with information on the antibody, immunogenicity, experimental method of determination, host protein and classification of antigen. Protein sequences are cross-referenced to the PDB and SWISS-PROT databases. The Discontinuous Epitope Database (CED) contains information on experimentally determined discontinuous epitopes from the peer-reviewed primary literature. The database contains ~ 200 manually curated entries, each corresponding to a single epitope annotated with information on immunogenicity, experimental methods, antibody and antigen, together with cross-references to other databases. Epitome is another database containing structurally inferred antigenic regions deduced from X-ray protein structures. It consists of all known antigen–antibody complex structures, with a description of residues involved in interactions, and their sequence/structural environments.

Some concerns remain over the interpretation of B-cell epitope data. In part, this comes from the experimental techniques used to identify B-cell epitopes, which measure antibody crossreactivity of synthetic peptide constructs to their parent antigen. If the experimental method used is not reliable, then its use will not facilitate the unequivocal identification of epitopes. Any prediction method based on such flawed data, however good its inherent accuracy, will be useless. Several papers [24-26] have thrown doubt over the reliability and interpretation of linear epitope identification methods using synthetic peptides. The authors' own analysis (Blythe and Flower, unpublished) corroborates such findings in that the logical explanation for these observations is a result of non-specific crossreactivity of antigen-specific antibodies, and binding of antibodies raised against denatured or degraded antigen *in vivo*. The entities identified by these epitope-mapping methods and presented in peer reviewed published information cannot be accepted unequivocally as 'true' epitopes, and must be considered primarily as artifacts of the experimental technique used to identify them. However, the term 'B-cell epitope' is so widely used within the published literature to refer to peptides recognised by antibodies that it seems likely that it will long remain ambiguous and misleading.

In addition, within the arena of humoral immunology, there exists other relevant databases. One of the most interesting is HaptenDB [27], which presently contains 2021 entries for 1087 haptens and 25 carrier proteins. Each entry details the nature of the hapten, together with its two- and three-dimensional structures, and carrier proteins. Data are also available for the coupling method, antihapten antibody production method, assay method and the specificities of antibodies. This opens up a new area of immunoinformatic enquiry, and many more databases focusing on haptens will surely follow.

5. Allergen databases

Pre-existing antibody or T cells, which bind to otherwise harmless environmental antigens, can give rise to allergic reactions, a condition that affects more than a third of the population. A common cause of allergic reactions such as asthma and hay fever is the binding of an antigen to IgE antibody on mast cells. Any protein that can stimulate such a reaction is known as an allergen. Desensitisation through vaccination against specific allergens is increasingly seen as a viable general treatment for allergy, one that overlaps considerably with so-called life style vaccines, which act against conditions ranging from chemical addiction to dental caries.

A number of allergen databases are now available. The International Union of Immunological Societies (IUIS) database lists clinically relevant allergens and isoallergens from > 150 species [28]. The Allergome database also lists allergen molecules and their biological functions [29], and additionally contains allergenic substances for which specific allergen proteins have yet to be identified. The Structural Database of Allergen Proteins (SDAP) contains both allergen sequences and structures and permits homology searches between known allergens and input user sequences [30].

Several databases have been developed for food and foodstuffs. The Food Allergy Research and Resource Programme database contains 1537 sequences of unique proteins for known, and putative, food, environmental and contact allergens and gliadins that may induce celiac disease. The InformAll database, formerly the PROTALL database, is maintained by a European consortium and archives information on plant food allergens involved in IgE-induced hypersensitivity reactions [31]. The Biotechnology Information of Food Safety (BIFS) [32] website contains a non-redundant list of allergen proteins designed to assess the potential allergenicity of proteins in specific foods, whereas the Central Science Laboratory (CSL) allergen database contains both food and inhalant and contaminant allergens. ALLALLERGY is a database that can be queried by food type and lists all chemical allergens as well as information on adverse reactions, crossreactivity and patient assessment.

Criticisms have been made of the currently available databases [33,34]. Comparisons between databases have suggested that there are inconsistencies in the number of allergen sequences for a particular foodstuff that arises partly from using different source databases, partly from different inclusion criteria, and partly from different treatment of duplicate sequences. Moreover, there is also no standard for the minimum patient number that must express sera where IgE

binds to an allergen for it to be counted as such. IUIS require at least five such patients, whereas other databases require only one. Difficulties in sequence annotation are further compounded when post-translational modifications are the source of allergenicity rather than the protein itself. Ultimately, it may be impossible to integrate fully all available databases; yet a non-redundant set of allergens is clearly needed, together with a full annotation of their structural, functional and clinical features.

6. Data mining

No commercial vaccine has ever been designed solely, or even significantly, using computational techniques. Likewise, there are no commercially available epitope-based or peptide vaccines. Most vaccines are still largely based around attenuated whole pathogen or, at best, are subunit vaccines. Almost all vaccines, with the exception of BCG, are mediated by antibodies rather than cellular immunity. Of course, immunoinformatics is best at predicting T-cell epitopes, whereas prediction of B-cell epitopes is, as shall be seen, more problematic than most suspect. Such is life. Nonetheless, T-cell biology is the focus of immeasurable interest within immunology and vaccinology. Hope, as they say, springs eternal. In the following sections the authors will adumbrate developments, difficulties and dilemmas in the prediction of epitopes and other important features of the immune response.

7. T-cell epitope prediction

MHC bind short peptides derived from host and pathogen proteins and present them on the cell surface for inspection by T cells. Peptides that are recognised by such cells are termed T-cell epitopes and peptide binding seems the most selective step in the recognition process. There are many sequence-based methods for the prediction of T-cell epitopes, most relying on the prediction of peptide–MHC binding [35]. Successfully modelling the peptide specificity exhibited by MHCs allows preselection of candidate peptides, which, in turn, can help identify immunogenic epitopes.

There are two kinds of MHC molecule. Class I MHC alleles have a binding groove, which is closed at both ends, making it possible to predict exactly which residues are positioned in the binding groove. Numerous methods have been applied to the problem of predicting MHC binding. MHC class I binding prediction is regarded as being very successful, with reported prediction accuracies of < 95% [36]. Attempts at predicting binding to class II MHC, which has a peptide groove open at both ends, show significantly lower accuracies, although many efforts using both traditional and novel approaches have been applied.

Beginning with the identification of so-called binding motifs in the 1980s that characterise the peptide specificity of different class I MHC alleles in terms of dominant anchor positions with a strong preference for a restricted group of amino acids: the best-studied human MHC protein, HLA-A*0201, has anchor residues at peptide positions P2 (accepting leucine and methionine) and P9 (accepting valine and leucine). Motifs are widely popular, and widely exploited, being simple to use and simple to understand. For example, the HIV and HCV databases [14,15] offer a simple tool (MotifScan) for identifying HLA motifs in query proteins, highlighting them on a protein alignment. MotifScan is based on motif libraries from SYFPEITHI and motifs extracted from the literature. An extension of MotifScan is the HIV Epitope Location Finder (ELF), where motifs are mapped onto alignments together with extensive database listings of class I HIV and HCV epitopes.

There are fundamental technical problems with motifs. Because peptides are viewed simplistically as either binders or non-binders, motifs generate many false positives and false negatives. It is obvious now that the whole peptide contributes to affinity, not just a few anchor residues, and so to T-cell-mediated immunogenicity. Effective models of binding must use rather more intricate, complex representations of the biophysical phenomena of binding. A succession of ever more sophisticated methods have been applied to the problem [35,37]: various empirical methods, notably Parker's BIMAS [38]; Artificial Neural Networks; Hidden Markov Models; and robust multivariate statistics, such as Partial Least Squares, to name but a few. These have lead to a swathe of web-server implementations available via the internet (see **Table 2**). Recent developments have pursued three distinct paths: one a data-driven approach based on sequence, one that combines together different aspects of the antigen presentation pathway, and the other a potentially more general structure-based approach.

Support Vector Machines (SVMs), originally introduced by Vapnik [39], are an artificial intelligence technique that is taking the world of bioinformatics by storm. Their inherent accuracy is seemingly compelling. A single SVM is a binary classifier that learns a decision boundary between two classes (for example, epitope versus non-epitope) using an appropriate amino acid representation. To find a boundary between classes, an SVM maximises the margin between them, choosing a linear separation in feature space. A *kernel function* $K(x_i, x)$ projects the data from the input space to the feature space. The formulation of SVMs is described lucidly in many books and publications (e.g., [40]). As a result of their success, a whole tranche of SVM-based methods for class I epitope prediction have been developed [36,41-46]. Most SVM methods undertake discriminant analysis, but an encouraging performance with quantitative prediction using support vector regression has been seen [47].

In contrast to sequence-based SVM methods, the most significant other trend in epitope prediction has been the use of the structures of MHCs and MHC–peptide complexes. These techniques use two main approaches: docking, with or without scoring, and molecular dynamics (MD) simulation. Leveraging earlier work [48,49], several methods seek to apply static docking methods (scoring functions derived from

Table 2. Prediction servers.

Predictive servers			
ABCpred	http://www.imtech.res.in/raghava/abcpred/	Artificial neural net linear B-cell epitope predictor	
AllerPredict	http://sdmc.i2r.a-star.edu.sg/Templar/DB/Allergen/	Allergen protein predictor	
BcePred	http://www.imtech.res.in/raghava/bcepred/bcepred_submission.html	Prediction of continuous B-cell epitopes in continous data using physiochemical properties	
BIMAS	http://thr.cit.nih.gov/molbio/hla_bind	Predicts HLA–peptide half time of disassociation	
ELF	http://www.hiv.lanl.gov/content/hiv-db/ELF/epitope_analyzer.html	MotifScan summaries, integrating known epitopes and HIV/HCV proteome epitope maps	07/04/06
EpiPredict	http://www.epipredict.de/Prediction/prediction.html	The binding contribution of every amino acid side chain in a class II-ligand is described by allele-specific two-dimensional data bases	
EpiVax	http://www.epivax.com	Prediction of class I and II conserved and promiscuous epitopes	
CTLPred	http://www.imtech.res.in/raghava/ctlpred	Prediction of CTL epitopes based on artificial neural networks and support vector machines	
IEDB Binding, MHC class I	http://www.immuneepitope.org/analyze/html/mhc_binding.html	Prediction of class I–peptide binding using three different methods	
IEDB Binding, MHC class II	http://www.immuneepitope.org/tools/matrix/iedb_input	Prediction of class II–peptide binding	
MHC-Thread	http://www.csd.abdn.ac.uk/~gjlk/MHC-Thread/	Program predicts peptides that are likely to bind to class II MHC molecules based on threading technique	
MHCPred	http://www.jenner.ac.uk/MHCPred	Predicted MHC–peptide or TAP–peptide IC_{50} binding values	28/04/05
MHC2Pred	http://www.imtech.res.in/raghava/mhc2pred	Prediction of promiscuous MHC class II binders	
MMBPred	http://www.imtech.res.in/raghava/mmbpred	Prediction of promiscuous MHC class I binders and prediction of mutations that will allow high-affinity binding	
MotifScan	http://www.hiv.lanl.gov/content/immunology/motif_scan	Summary and location of anchor motifs	
NetCTL	http://www.cbs.dtu.dk/services/NetCTL	Predicted proteasome or immunoproteasome cleavage	05/01/06
NetMHC	http://www.cbs.dtu.dk/services/NetMHC	Predicts MHC binding propensity of peptides	09/08/05
PAProC II	http://www.paproc.de/	Prediction of 20S human proteasomes	
PREDEPP	http://margalit.huji.ac.il/	Predicted MHC–peptide binding based on structure	
ProPred	http://www.imtech.res.in/raghava/propred	Predicted MHC class II–peptide binding (there are other related tools at the imtech web site)	
ProPred-I	http://www.imtech.res.in/raghava/propred1	Predicted MHC class I–peptide binding, optional proteasome/immunoproteasome cleavage filter	
RankPep	http://bio.dfci.harvard.edu/Tools/rankpep.html	Predicts peptide binders to MHCI and MHCII molecules using Position Specific Scoring Matrices (PSSMs)	
SVMHC	http://www-bs.informatik.uni-tuebingen.de/SVMHC	MHC class I predictions based on support vector machines and known MHC-binding peptides	
SYFPEITHI	http://www.syfpeithi.de/Scripts/MHCServer.dll/EpitopePrediction.htm	Predicted epitopes, binding motifs and epitope alignments for MHC proteins	

CTL: Cytotoxic T cell; HCV: Hepatitis C virus; HLA: Human leukocyte antigen; MHC: Major histocompatibility complex; TAP: Transannular patch.

Table 3. The percentage of epitope predictions by different algorithms.

Algorithms	Class I HLA	Class II HLA	Mouse class I MHC	Mouse class II MHC
Additive sAA	93%	91%	85%	45%
Additive iAA	88%	–	70%	–
BIMAS	90%	–	30%	–
ComPred	73%	–	45%	–
netMHC	83%	–	–	–
ProPred	–	71%	–	–
RANKPEP	73%	46%	40%	69%
SVMHC	75%	–	–	–
SYFPEITHI	83%	50%	65%	–

–: Model not available for testing.
HLA: Human leukocyte antigen; MHC: Major histocompatibility complex.

computational chemistry or threading methods derived from structural bioinformatics) to identify MHC binders [50-55]. In general, the docking of small molecules performs well when distinguishing ligands from non-ligands [56], but perform poorly when predicting binding affinity; trends that are preserved when dealing with MHC prediction.

To address this, several workers have used molecular dynamics as a route to affinity prediction [57-60]. However, the continuing issues with the availability of computing resources that are able to sustain the requisite size and duration of simulation necessary for obtaining free energies of binding for medium sized systems, such peptide–MHC complexes has hampered development of easily deployable techniques. To address this, Wan and co-workers have used high-performance computing deployed via the nascent GRID to power more realistic simulations of a series of systems of ascending scale [61-63]; their work offers hope of ultimate success, but such success remains a distant prospect. Another interesting piece of work [64] combines MD with multivariate statistics to produce a hybrid approach to prediction. This effort highlights the lack of dominant interactions between peptide and MHC in MD simulations.

Recently, an attempt has been made to incorporate several components of the class I antigen presentation pathway, such as proteasome cleavage [65] and TAP binding [44,66], into composite approaches to T-cell epitope prediction [67-70]. Likewise, a CoMSIA method [71] has been explored for distinguishing true epitopes from non-epitopes that bind MHCs with high affinity [72]. These methods, which show encouraging improvements, compared with MHC-only approaches, use subsidiary stages, such as TAP binding, as additional filters to reduce the number of possible epitopes.

For class II MHC molecules, the open binding groove allows much longer peptides (< 25 residues) to bind to class II alleles than that which can bind to class I MHCs, which is thought to be limited to at most 15 amino acids. However, the grooves of MHC class II alleles will only accommodate 9 – 11 residues of the target peptide. Thus, class II peptides have the potential to bind to the MHC groove in one of several registers (potential alignments between the groove and antigenic peptide). Moreover, several studies have indicated that residues outside the binding groove (flanking residues) can also influence binding [73,74]. This greatly complicates attempts to produce predictive models of class II binding. Recently, pattern recognition has been applied to the class II problem, such as Artificial Neural Networks [75,76] and SVMs [45,77]. Typically, a binding core is first estimated or declared, and subsequently the binding affinity is predicted for an unknown peptide from this estimate. This two-step procedure restricts the task to a fixed-length formulation, thus, avoiding problems of handling variable length peptides.

Approaches for solving the dynamic variable-length nature of the class II prediction problem have shown promise. Methods include an iterative 'meta-search' algorithm [78], an iterative Partial Least Squares method [79], Hidden Markov Models [80,81], an Ant Colony search [82] and a Gibbs sampling algorithm [83]. Some of these novel approaches have significantly out performed conventional approaches.

Recently, a limited number of comparative validations have been performed that attempt to benchmark the performance of different MHC binding and T-cell epitope predictors [84,85]. Brusic has produced some useful rules of thumb for selecting appropriate techniques given the volume of data available: they suggest molecular modelling when data on binding is absent, motifs when it is scanty, and AI techniques when it is more plentiful. In addition to their accuracy, SVMs also need rather less data for training than some equivalent methods, such as ANN. Sette *et al.* used a data set of 48,828 quantitative affinity measurements (as accumulated by his group over the last 15 years) for peptides binding to 48 different mouse, human, macaque and chimpanzee MHC class I alleles, to evaluate a neural network method and two matrix-based prediction methods developed in-house and these were then compared with predictions made via the web. Although differences in respective data sets

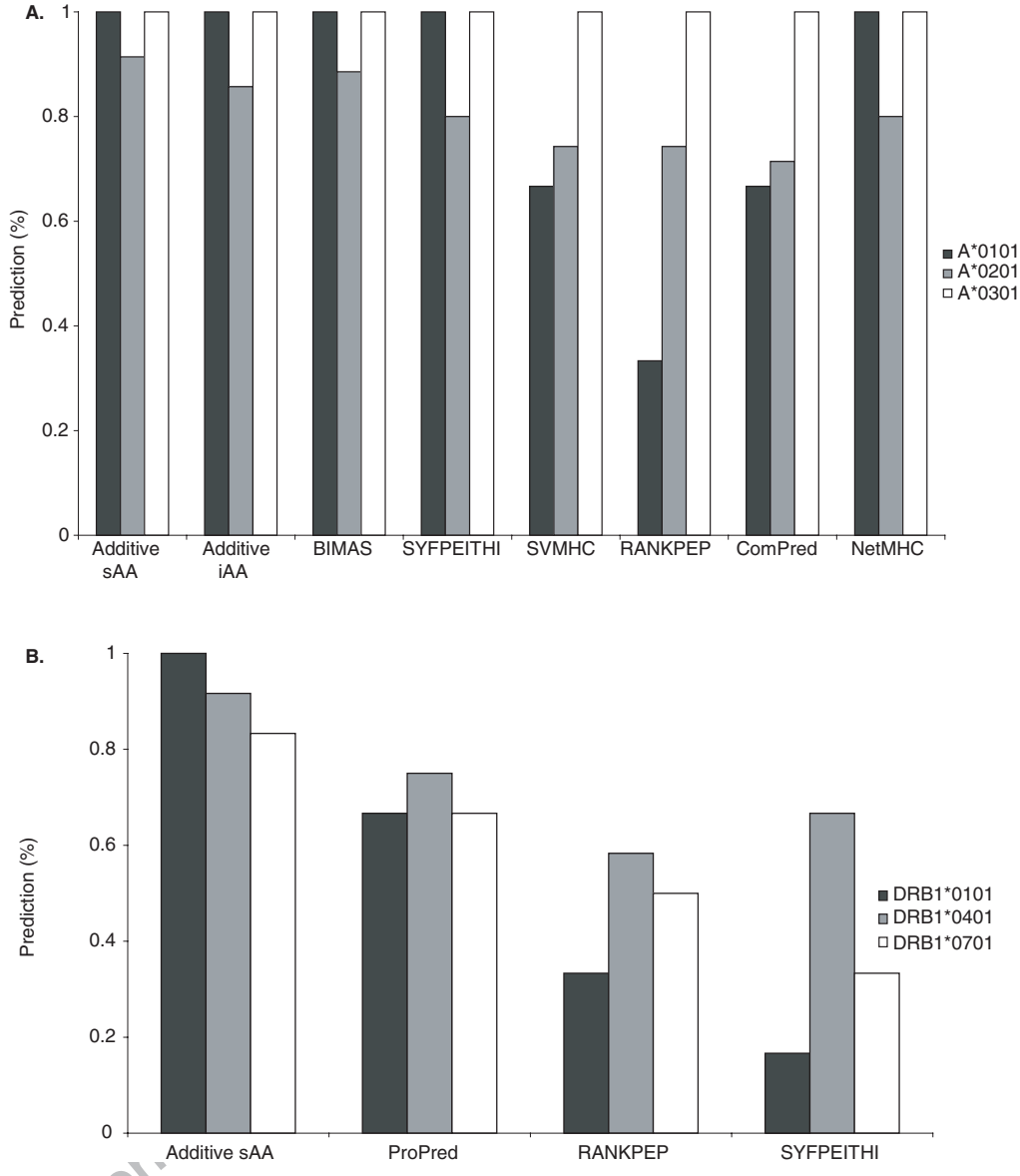

Figure 1. Data mining in immunology: comparison of epitope prediction methods. A. HLA class I epitope predictions. **B.** HLA class II epitope predictions. **C.** Mouse class I epitope predictions. **D.** Mouse class II epitope predictions.

The reliability of computational T-cell epitope identification is tested using recently published epitopes and whole protein sequences. Such a test better reflects how epitope prediction algorithms are applied in real-life situations by experimental immunologists working at the lab bench. The use of artificial test sets, as compiled from databases, is often flattering. Four epitope data sets were extracted from the literature: 40 human class I, 24 human class II epitopes. 20 mouse class I and 52 mouse class II epitopes were also included. Where possible, only recently determined epitopes were used. A total of 154 epitopes were sourced from 87 protein sequences. 43 human class I and 33 class II human epitopes were obtained, mainly restricted by A1, A3, A2, DRB*0101, *0401 and *0701 alleles. 26 mouse class I and 52 class II epitopes were obtained, mostly restricted by H-2Kb, H-2Db, I-Ab, I-Ad, I-Ak, I-Ed and I-Ek alleles. Ten epitope prediction algorithms were compared: BIMAS, SYFPEITHI, RANKPEP, PREDEP, ProPred, ComPred, netMHC, SVMHC, SMM and MHCPred (see **Table 2**). Not all methods have the same set of allele models. Only SYFPEITHI, RANKPEP and MHCPred predicted human and mouse class I and II MHC alleles; other algorithms are restricted to class I. MHCPred supports two different models: single amino acid model (additive sAA), which only considers individual amino acid contributions, and amino acid plus interactions model (additive iAA), which accounts for interactions between neighbouring residues. Both models were evaluated for human class I epitopes. Receiver operating characteristic (ROC) curves were used to analyse accuracy. An epitope was predicted if a predicted affinity or score was above a threshold, which was chosen to be the top 20 peptides output. ROC curves give a graphical representation of the true positive rate (sensitivity) versus the false positive rate (specificity) at different threshold levels. ROC curves, which plot (x, y) values of 1-Specificity versus Sensitivity, take the shape of a plateau curve. The area under the curve (Aroc) indicates the data quality. The closer the Aroc value to 1, the better the prediction. Aroc values between 0.8 and 0.9 indicate good prediction, values higher than 0.9 indicate excellent prediction, while values at or below 0.6 indicate complete lack of predictivity.All algorithms had good predictivity for human class I epitopes. Motif and matrix methods were particularly effective. The additive sAA model and BIMAS were the best algorithms and had the highest predictivity for individual alleles. Class II epitope predictions had lower redictivities than those of class I. This may be due to the variable length of class II epitopes and problems of data quality. The additive sAA model and ProPred exhibited high predictivity, followed by RANKPEP and SYFPEITHI. The additive sAA model and SYFPEITHI were best for mouse class I predictions. SYFPEITHI can predict all H-2Db epitopes, and the additive sAA model had the highest predictivity of 81% for H-2Kb predictions. For mouse class II prediction, RANKPEP had a very high predictivity of 85% compared with MHCPred, which only predicted 57%.

HLA: Human leukocyye antigen.

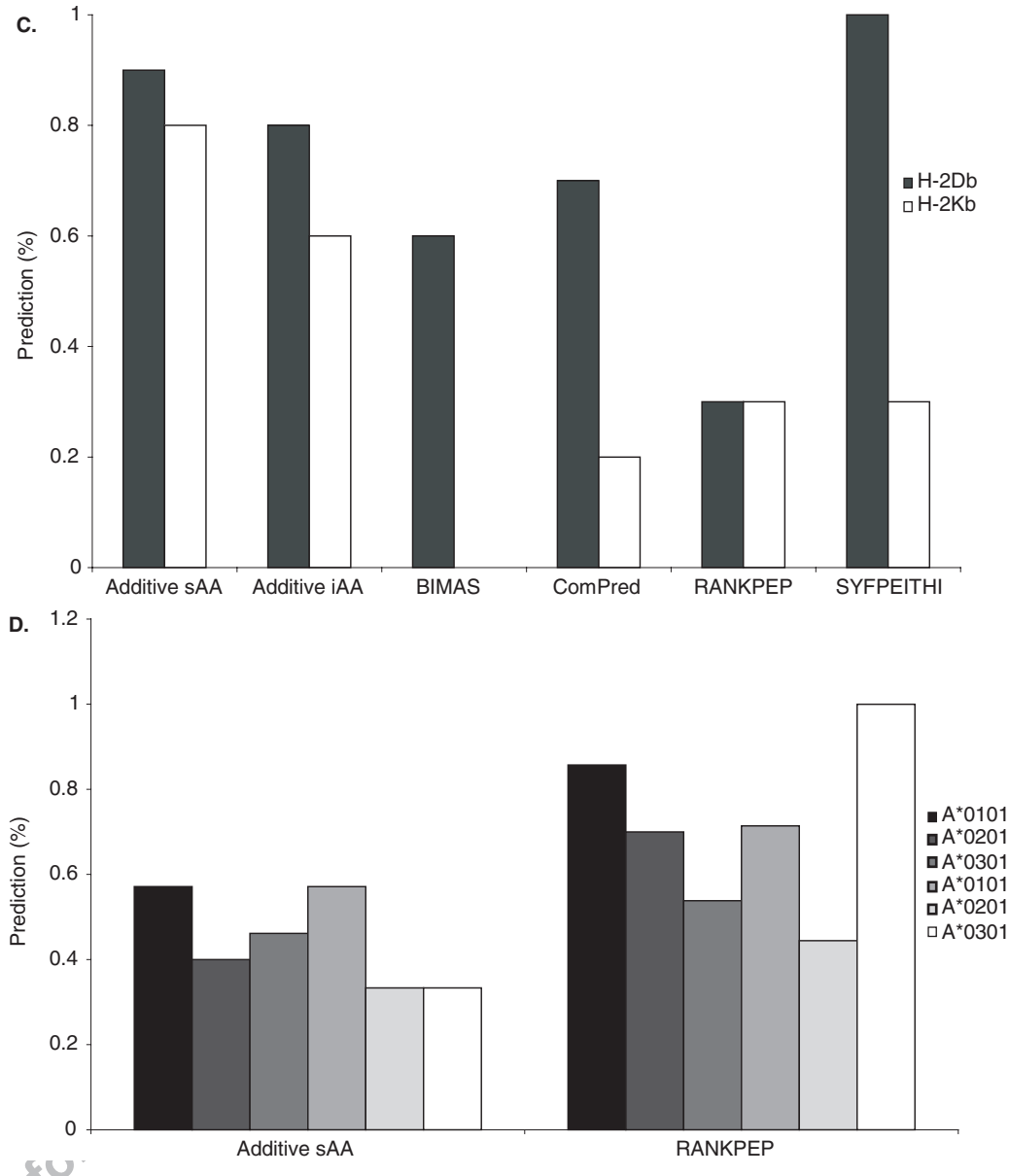

Figure 1. Data mining in immunology: comparison of epitope prediction methods (continued). A. HLA class I epitope predictions. **B.** HLA class II epitope predictions. **C.** Mouse class I epitope predictions. **D.** Mouse class II epitope predictions.

The reliability of computational T cell epitope identification is tested using recently published epitopes and whole protein sequences. Such a test better reflects how epitope prediction algorithms are applied in real-life situations by experimental immunologists working at the lab bench. The use of artificial test sets, as compiled from databases, is often flattering. Four epitope data sets were extracted from the literature: 40 human class I, 24 human class II epitopes. 20 mouse class I and 52 mouse class II epitopes were also included. Where possible, only recently determined epitopes were used. A total of 154 epitopes were sourced from 87 protein sequences. 43 human class I and 33 class II human epitopes were obtained, mainly restricted by A1, A3, A2, DRB*0101, *0401 and *0701 alleles. 26 mouse class I and 52 class II epitopes were obtained, mostly restricted by H-2Kb, H-2Db, I-Ab, I-Ad, I-Ak, I-Ed and I-Ek alleles.Ten epitope prediction algorithms were compared: BIMAS, SYFPEITHI, RANKPEP, PREDEP, ProPred, ComPred, netMHC, SVMHC, SMM and MHCPred (see Table 2). Not all methods have the same set of allele models. Only SYFPEITHI, RANKPEP and MHCPred predicted human and mouse class I and II MHC alleles; other algorithms are restricted to class I. MHCPred supports two different models: single amino acid model (additive sAA), which only considers individual amino acid contributions, and amino acid plus interactions model (additive iAA), which accounts for interactions between neighbouring residues. Both models were evaluated for human class I epitopes. Receiver operating characteristic (ROC) curves were used to analyse accuracy. An epitope was predicted if a predicted affinity or score was above a threshold, which was chosen to be the top 20 peptides output. ROC curves give a graphical representation of the true positive rate (sensitivity) versus the false positive rate (specificity) at different threshold levels. ROC curves, which plot (x, y) values of 1-Specificity versus Sensitivity, take the shape of a plateau curve. The area under the curve (Aroc) indicates the data quality. The closer the Aroc value to 1, the better the prediction. Aroc values between 0.8 and 0.9 indicate good prediction, values higher than 0.9 indicate excellent prediction, while values at or below 0.6 indicate complete lack of predictivity.All algorithms had good predictivity for human class I epitopes. Motif and matrix methods were particularly effective. The additive sAA model and BIMAS were the best algorithms and had the highest predictivity for individual alleles. Class II epitope predictions had lower redictivities than those of class I. This may be due to the variable length of class II epitopes and problems of data quality. The additive sAA model and ProPred exhibited high predictivity, followed by RANKPEP and SYFPEITHI. The additive sAA model and SYFPEITHI were best for mouse class I predictions. SYFPEITHI can predict all H-2Db epitopes, and the additive sAA model had the highest predictivity of 81% for H-2Kb predictions. For mouse class II prediction, RANKPEP had a very high predictivity of 85% compared with MHCPred, which only predicted 57%.

HLA: Human leukocyye antigen.

show a transparent comparison, they concluded that tools based on combinatorial peptide libraries perform very well. This may be due to the lack of tendentious self-reinforcement compared with techniques built from data sets originally selected by motif.

Although published comparisons tend to favour methods promulgated by the group undertaking the evaluation, the authors have examined the predictivity of 10 algorithms by testing if they were able to identify T-cell epitopes within whole protein sequences (see **Table 3** and **Figure 1**). The authors' approach mimics real-life situations and assesses algorithms as they might be used by experimentalists. The average predictivity of human class I epitopes was higher than class II predictions for all the algorithms, also human epitope predictions are more accurate than those in mice. Due to the small number of algorithms offering mouse class II predictions, only the additive model and RANKPEP were compared, in which RANKPEP had the higher predictivity of 68%, whereas MHCPred scored 45%.

In a recent comparison, four combined epitope prediction methods, as described above, were tested using 99 recently identified epitopes [67]: EpiJen recognised 61 out of 99, SMM [68] 57 from 99, NetCTL [69] identified 49 and WAPP [70] found 33. The positive predictive value (true positives/[true positives + false positives]) was low for all methods: 21% for NetCTL, 17% for EpiJen and WAPP, and 16% for SMM. Quite stringent criteria, certainly more so than usual, were used. Methods of this kind represent the present state-of-the-art in epitope prediction. Differences in published acceptance criteria continue to cloud the unequivocal assessment of comparative performance. Of course, arguably the most convincing validation is the prospective, rather than retrospective, use of predictive methods to guide experiments, as most experimentalists are unimpressed by statistical evaluations using extant data. Sette's group has recently reported a significant example of such work [86]. They evaluate the accuracy of predicted class I T-cell epitopes derived from vaccinia virus in the H-2(b) mouse, and find that conventional methods predict the majority of murine responses. As well as predicting epitopes from known proteins, predictive methods can also be used to design novel, non-natural sequences that might act as super-agonists, antagonists, or blockers of MHC-mediated T-cell responses. Doytchinova *et al.* used a model of A2 binding to design super binders with affinities up to 2.5 orders of magnitude greater than the most affine natural peptide sequences [87]. They were also able to systematically alter the amino acid identity of anchor positions showing that peptides with ≥ 10 residues other than canonical anchors can be bound at, or above the affinity threshold concomitant with putative immunogenicity.

8. B-cell epitope prediction

Whereas T-cell epitope prediction progresses encouragingly (albeit with some as yet unresolved difficulties for class II), the prediction of B-cell epitopes abounds with real dilemmas. B-cell epitopes are classified by the relative location of epitope residues within the primary sequence of the antigen. It must also be said that their verity and exegesis depends on their means of experimental determination. The successful prediction of these features within proteins would be valuable to those areas of immunology that seek to explore the nature of the B-cell response. Using data sets representing the most stringent examples of peer-reviewed publications describing linear epitope-mapped protein sequences, Blythe and Flower [88] have explored the validity of epitope prediction methods based on sequence profiles using amino acid scales. This is a very simple methodology that is used routinely to visualise protein features such as hydrophobicity and secondary structure. Using 484 amino acid scales and 50 epitope-mapped protein sequences, as defined by polyclonal antibodies, analysis of both single sequence and combined profiles indicated categorically the underlining approach was inadequate.

Notwithstanding issues with the interpretation of existing data, methods continue to be published. Recently, epitope prediction methods based on decision trees, nearest neighbour learning [89], and freed-forward artificial neural networks [90], have been proposed. All three algorithms were trained using a greater amount of data sourced from a private database. Athough their methods were shown to be able to correctly classify a high proportion of it under cross-validation, demonstrating that this approach may have merit, the prediction performance was determined using a small set of 10 HIV-1 proteins, thereby potentially limiting the reliability of their performances. ABCpred [91] is based on a RNN form of ANN. It was trained using a set of 700 non-redundant linear epitopes of length 16. Non-epitopes were selected at random from Swiss-Prot. In a blind data set of 187 epitopes, the algorithm had sensitivity and specificity values of 71.66 and 61.50%. BepiPred [92] combines Parker's hydrophilicity scale with a Hidden Markov Model (HMM) trained on known linear epitopes and was assessed using epitope-mapped HIV proteins. Using ROC curves, BepiPred was shown to perform marginally better than the HMM and single propensity scales alone, although an A_{roc} value of 0.600 does not indicate true predicitivity. The assessment of the algorithm against the epitopes of a single organism does not, in itself, demonstrate the general competence of this algorithm. BepiPred is clearly not a quantum leap in performance over single-scale profiles. DiscoTope [93], which predicts discontinuous rather than linear epitopes, is based on the analysis of 76 antibody–antigen complex crystal structures. It uses a residue propensity scale combined with structural protrusion calculations to predict residues within discontinuous epitopes. Using a threshold that selects a 95% specificity, DiscoTope detects 15% of the epitope residues. The Algorithm achieves improved performance compared with existing discontinuous epitope prediction.

One must distinguish linear epitopes from those defined by antibody–antigen complex X-ray crystal structures. Crystallographic structures are probably the most reliable representation of epitopes as all interactions between antibody and antigen are described fully [26,94]. Structurally, antibody and antigen are unequal partners; atypical features

of the antibody–antigen interaction include the amino acid composition and hydropathy of the antibody's paratope. Although the complex is typical in terms of buried surface area, density of hydrogen bonds, and complementarity, the amino acid composition of the paratope is distinct from that of the epitope regions and other, non-obligate protein interfaces. The contribution to the paratope area of aromatic residues, particularly tyrosine, is significantly greater than other residues. On average, tyrosine forms a third of the paratope area. The hydrophobic content of the paratope is also significantly higher than that of other transient protein interfaces, and of epitopes, which are the mostly hydrophilic. Yet, by comparing the amino acid composition of epitopes to other similar sized surface regions evidence, their lack of distinguishing features, necessary for unequivocal prediction, is clear. This supports the Multi-determinant Regulatory Model proposed by Benjamin, in which any part of the accessible surface of a globular protein antigen can be recognised by antibodies, and that the entire exposed surface represents a 'continuum' of overlapping potential epitopes [95]. Epitope prediction based on data from single-specificity monoclonal antibodies may not facilitate the prediction of surface regions that represent either immunodominant or neutralising antigenic sites.

9. Conclusion

When reviewing computational support for vaccinology, a field in transition is seen. Straightforward sequence analysis and genome annotation has been used for some time; this is now being supplemented by the emergence of a variety of different databases and potent prediction techniques, most notably for T-cell epitopes. Of course, the computational identification of either T- or B-cell epitopes, does not close the book on immunological prediction. Obviously, there are many other things besides. For example, there has been a lot interest in identifying *in silico* so-called supertypes [96-98], antigens [99] and allergens [100]. There is also much interest in the differences between the constitutive and the immuno-proteasome [101,102].

An area the authors have not really touched on is that of tumour vaccines. In many ways, this offers the best chance of immunoinformatics having the kind of direct input that it so justly warrants. For a start, T cells have a key role in protecting against cancer. Recently, many strategies to identify tumour antigens have emerged, leading to the characterisation of various types of cancer antigens, over expressed antigens, differentiation antigens, germline antigens, mutated antigens and antigens of viral origin. The discovery that many class I MHC-binding epitopes derived from tumour antigens are recognised by $CD8^+$ T cells has greatly fomented the development of new vaccines. Epitopes of 9 – 12 amino acids are easily synthesised and can be used directly for immunisation. Likewise, class II cancer-derived epitopes have been shown to be recognised by $CD4^+$ T cells. Moreover, T-cell responses have been linked with tumour regression, opening up the possibility of personalised, or, at least, boutique, medicine. In this context, the idea of identifying epitopes for single alleles becomes of great importance.

Many prediction methods are either problematic (i.e., B-cell epitopes) or are only starting to be developed (i.e., antigens). Yet, T-cell epitope prediction works. When there is enough data to build a good model and the prediction method is of sufficient sophistication, it is more efficient to use prediction than to perform exhaustive experiments. The fact that such strategies are not more widely adopted says as much for the 'luddite' mindset prevalent among many experimental immunologists as it does for the poor exposure of prediction methods. Positive publicity may come through the high profile of IEDB, and its satellite projects; it is to be hoped that its kudos will not totally eclipse other methods of equal scientific stature and veracity. Good data is essential. The fact that the need for good data remains unfulfilled can be seen as very much a condemnation of the lack of coordination amongst researchers competing with each other for funding. Nevertheless, the development of computational approaches to MHC supertypes is a welcome attempt to overcome problems with data quantity. Likewise, structure-based methods not founded solely on extant binding data offer hope of addressing issues of both data quantity and quality.

10. Expert opinion

When one works in a Cinderella field, such as immunoinformatics or computational vaccinology, then it is sometimes difficult for outsiders to properly assess the relative merits of *in silico* vaccine design compared to more mainstream experimental studies, however transitory and uninformative they may ultimately prove. The potential – albeit largely unrealised – is huge, but only if people are willing to take up the technology and use it appropriately. People's expectations of computational work are often largely unrealistic and highly tendentious. Some expect perfection, and are soon disappointed, rapidly becoming vehement critics. Others are highly critical from the start and are nearly impossible to reconcile with informatic methods. However, neither appraisal is correct. Informatic methods do not replace, or even seek to replace, experimental work, only to help rationalise experiments, saving time and effort. They are slaves to the data used to generate them. They require a degree of intellectual effort equivalent in scale yet different in kind to that of so-called experimental science. The two disciplines – experimental and informatics – are, thus, complementary albeit distinct. Ultimately, as system biology takes hold and insinuates itself into every corner of biology, informatics will find its place, although that may take some time.

Acknowledgements

The authors thank V Brusic, P Borrow and S Ellis for discussions and collaboration. The authors also thank DJ Clayton, SL Hemsley, C Zygouri, A Worth, and H McSparron for their help and assistance.

Bibliography

Papers of special note have been highlighted as either of interest (•) or of considerable interest (••) to readers.

1. JOHNSON G, WU TT: Kabat database and its applications: 30 years after the first variability plot. *Nucleic Acids Res.* (2000) **28**:214-218.
2. GIUDICELLI V, DUROUX P, GINESTOUX C *et al.*: IMGT/LIGM-DB, the IMGT comprehensive database of immunoglobulin and T cell receptor nucleotide sequences. *Nucleic Acids Res.* (2006) **34**:D781-D784.
•• **Definitive immunological host sequence/structure database.**
3. ROBINSON J, WALLER MJ, STOEHR P, MARSH SG: IPD – the Immuno Polymorphism Database. *Nucleic Acids Res.* (2005) **33**:D523-D526.
4. RETTER I, ALTHAUS HH, MUNCH R, MULLER W: VBASE2, an integrative V gene database. *Nucleic Acids Res.* (2005) **33**(Database issue):D671-D674.
5. NAKANO Y, SHIBATA Y, KAWADA M *et al.*: A searchable database for proteomes of oral microorganisms. *Oral Microbiol. Immunol.* (2005) **20**(6):344-348.
6. CHEN T, ABBEY K, DENG WJ, CHENG MC: The bioinformatics resource for oral pathogens. *Nucleic Acids Res.* (2005) **33**(Web Server issue):W734-W740.
7. WASSENAAR TM, GAASTRA W: Bacterial virulence: can we draw the line? *FEMS Microbiol. Lett.* (2001) **9995**:1-7.
8. CHEN LH, YANG J, YU J *et al.*: VFDB: a reference database for bacterial virulence factors. *Nucleic Acids Res.* (2001) **33**(Database issue):D325-D328.
9. GARDNER MJ, HALL N, FUNG E *et al.*: Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* (2002) **419**(6906):498-511.
10. DEL PORTILLO HA, FERNANDEZ-BECERRA C, BOWMAN S *et al.*: A superfamily of variant genes encoded in the subtelomeric region of *Plasmodium vivax*. *Nature* (2001) **410**(6830):839-842.
11. TONGCHUSAK S, CHAIYAROJ SC, VEERAMANI A, KOH JL, BRUSIC V: CandiVF – *Candida albicans* Virulence Factor Database. *Int. J. Pept. Res. Ther.* (2005) **11**:271-277.
12. WINNENBURG R, BALDWIN TK, URBAN M, RAWLINGS C, KOHLER J, HAMMOND-KOSACK KE: PHI-base: a new database for pathogen–host interactions. *Nucleic Acids Res.* (2006) **34**(Database issue):D459-D464.
13. RAMMENSEE H, BACHMANN J, EMMERICH NP, BACHOR OA, STEVANOVIC S: SYFPEITHI: database for MHC ligands and peptide motifs. *Immunogenetics* (1999) **50**:213-219.
14. YUSIM K, RICHARDSON R, TAO N *et al.*: Los alamos hepatitis C immunology database. *Appl. Bioinformatics* (2005) **4**(4):217-225.
15. KUIKEN C, KORBER B, SHAFER RW: HIV sequence databases. *AIDS Rev.* (2003) **5**(1):52-61.
16. BHASIN M, SINGH H, RAGHAVA GP: MHCBN: a comprehensive database of MHC binding and non-binding peptides. *Bioinformatics* (2003) **19**(5):665-666.
17. RECHE PA, ZHANG H, GLUTTING JP, REINHERZ EL: EPIMHC: a curated database of MHC-binding peptides for customized computational vaccinology. *Bioinformatics* (2005) **21**(9):2140-2141.
18. TONG JC, KONG L, TAN TW, RANGANATHAN S: MPID-T: database for sequence–structure–function information on T cell receptor–peptide–MHC interactions. *Appl. Bioinformatics* (2006) **5**:111-114.
• **Key resource for structural informatics.**
19. TOSELAND CP, CLAYTON DJ, MCSPARRON H *et al.*: AntiJen: a quantitative immunology database integrating functional, thermodynamic, kinetic, biophysical and cellular data. *Immunome Res.* (2005) **1**(1):4.
• **Database that attempt to integrate immunological data.**
20. SATHIAMURTHY M, PETERS B, BUI HH *et al.*: An ontology for immune epitopes: application to the design of a broad scope database of immune reactivities. *Immunome Res.* (2005) **1**(1):2.
•• **Advent of a new era in functional immune databases.**
21. SAHA S, BHASIN M, RAGHAVA GP: Bcipep: a database of B-cell epitopes. *BMC Genomics.* (2005) **6**(1):79.
22. HUANG J, HONDA W: CED: a conformational epitope database. *BMC Immunol.* (2006) **7**:7.
23. SCHLESSINGER A, OFRAN Y, YACHDAV G, ROST B: Epitome: database of structure-inferred antigenic epitopes. *Nucleic Acids Res.* (2006) **34**(Database issue):D777-D780.
24. LAVER WG, AIR GM, WEBSTER RG, SMITH-GILL SJ: Epitopes on protein antigens: misconceptions and realities. *Cell* (1990) **61**(4):553-556.
25. SCHWAB C, TWARDEK A, LO TP, BRAYER GD, BOSSHARD HR: Mapping antibody binding sites on cytochrome c with synthetic peptides: are results representative of the antigenic structure of proteins? *Protein Sci.* (1993) **2**(2):175-182.
26. VAN REGENMORTEL MH: Mapping epitope structure and activity: from one-dimensional prediction to four-dimensional description of antigenic specificity. *Methods* (1996) **9**(3):465-472.
27. SINGH MK, SRIVASTAVA S, RAGHAVA GP, VARSHNEY GC: HaptenDB: a comprehensive database of haptens, carrier proteins and anti-hapten antibodies. *Bioinformatics* (2006) **22**(2):253-255.
• **New database that greatly extends the scope of immunoinformatic inquiry.**
28. KING TP, HOFFMAN D, LOWENSTEIN H, MARSH DG, PLATTS-MILLS TA, THOMAS W: Allergen nomenclature. *Allergy* (1995) **50**(9):765-774.
29. MARI A, RICCIOLI D: The allergome web site – a database of allergenic molecules. aim, structure and data of a web-based resource. 60th Annual Meeting American Academy of Allergy, Asthma &

Immunology. San Francisco. *J. Allergy Clin. Immunol.* (2004) **113**(2 Pt 2):S301.

30. IVANCIUC O, SCHEIN CH, BRAUN W: SDAP: database and computational tools for allergenic proteins. *Nucleic Acids Res.* (2003) **31**(1):359-362.
31. MILLS EN, VALOVIRTA E, MADSEN C *et al.*: Information provision for allergic consumers – where are we going with food allergen labelling? *Allergy* (2004) **59**(12):1262-1268.
32. GENDEL SM: Sequence databases for assessing the potential allergenicity of proteins used in transgenic foods. *Adv. Food Nutr. Res.* (1998) **42**:63-92.
33. BRUSIC V, MILLOT M, PETROVSKY N, GENDEL SM, GIGONZAC O, STELMAN SJ: Allergen databases. *Allergy* (2003) **58**:1093-1100.
34. GENDEL SM, JENKINS JA: Allergen sequence databases. *Mol. Nutr. Food Res.* (2006) **50**:633-637.
35. FLOWER DR, MCSPARRON H, BLYTHE MJ *et al.*: Computational vaccinology: quantitative approaches. *Novartis Found Symp.* (2003) **254**:102-120.
36. DONNES P, ELOFSSON A: Prediction of MHC class I binding peptides, using SVMHC. *BMC Bioinformatics* (2002), **3**:25.
37. FLOWER DR: Towards *in silico* prediction of immunogenic epitopes. *Trends Immunol.* (2003) **24**(12):667-674.
38. PARKER KC, BEDNAREK MA, COLIGAN JE: Scheme for ranking potential HLA-A2 binding peptides based on independent binding of individual peptide side-chains. *J. Immunol.* (1994) **152**(1):163-175.
39. VAPNIK VN: The nature of statistical learning theory. New York. *Springer* (1995).
40. SCHÖLKOPF B, SMOLA AJ: Learning with kernels: support vector machines, regularization, optimization and beyond. Cambridge, Mass. *MIT Press* (2002).
41. ZHAO Y, PINILLA C, VALMORI D, MARTIN R, SIMON R: Application of support vector machines for T cell epitopes prediction. *Bioinformatics* (2003) **19**(15):1978-1984.
42. RIEDESEL H, KOLBECK B, SCHMETZER O, KNAPP EW: Peptide binding at class I major histocompatibility complex scored with linear functions and support vector machines. *Genome Inform.* (2004) **15**(1):198-212.
43. BHASIN M, RAGHAVA GP: Prediction of CTL epitopes using QM, SVM and ANN techniques. *Vaccine* (2004) **22**(23-24):3195-3204.
44. BHASIN M, RAGHAVA GP: Analysis and prediction of affinity of TAP binding peptides using cascade SVM. *Protein Sci.* (2004) **13**(3):596-607.
45. BHASIN M, RAGHAVA GP: SVM based method for predicting HLA-DRB1*0401 binding peptides in an antigen sequence. *Bioinformatics* (2004) **20**(3):421-423.
46. CUI J, HAN LY, LIN HH *et al.*: Prediction of MHC-binding peptides of flexible lengths from sequence-derived structural and physicochemical properties. *Mol. Immunol.* (2007) **44**(5):866-877.
47. LIU W, MENG X, XU Q, FLOWER DR, LI T: Quantitative prediction of mouse class I–MHC peptide binding affinity using support vector machine regression (SVR) models. *BMC Bioinformatics* (2006) **7**:182.
48. ALTUVIA Y, SCHUELER O, MARGALIT H: Ranking potential binding peptides to MHC molecules by a computational threading approach. *J. Mol Biol.* (1995) **249**:244-250.
49. SCHUELER-FURMAN O, ALTUVIA Y, SETTE A, MARGALIT H: Structure-based prediction of binding peptides to MHC class I molecules: application to a broad range of MHC alleles. *Protein Sci.* (2000) **9**:1838-1846.
50. JOJIC N, REYES-GOMEZ M, HECKERMAN D, KADIE C, SCHUELER-FURMAN O: Learning MHC I–peptide binding. *Bioinformatics* (2006) **22**(14):e227-e235.
51. BUI HH, SCHIEWE AJ, VON GRAFENSTEIN H, HAWORTH IS: Structural prediction of peptides binding to MHC class I molecules. *Proteins* (2006) **63**:43-52.
52. BORDNER AJ, ABAGYAN R: *Ab initio* prediction of peptide–MHC binding geometry for diverse class I MHC allotypes. *Proteins* (2006) **63**:512-526.
53. TONG JC, ZHANG GL, TAN TW, AUGUST JT, BRUSIC V, RANGANATHAN S: Prediction of HLA-DQ3.2β ligands: evidence of multiple registers in class II binding peptides. *Bioinformatics* (2006) **22**(10):1232-1238.
54. TONG JC, BRAMSON J, KANDUC D, CHOW S, SINHA AA, RANGANATHAN S: Modeling the bound conformation of *Pemphigus Vulgaris*-associated peptides to MHC Class II DR and DQ alleles. *Immunome Res.* (2006) **2**:1.
55. TONG JC, TAN TW, RANGANATHAN S: Modeling the structure of bound peptide ligands to major histocompatibility complex. *Protein Sci.* (2004) **13**(9):2523-2532.
56. HATTOTUWAGAMA CK, DAVIES MN, FLOWER DR: Receptor–ligand binding sites and virtual screening. *Curr. Med. Chem.* (2006) **13**(11):1283-1304.
57. FAGERBERG T, CEROTTINI JC, MICHIELIN O: Structural prediction of peptides bound to MHC class I. *J. Mol. Biol.* (2006) **356**:521-546.
58. ZACHARIAS M, SPRINGER S: Conformational flexibility of the MHC class I α1–α2 domain in peptide bound and free states: a molecular dynamics simulation study. *Biophys. J.* (2004) **87**(4):2203-2214.
59. PETRONE PM, GARCIA AE: MHC-peptide binding is assisted by bound water molecules. *J. Mol. Biol.* (2004) **338**(2):419-435.
60. DAVIES MN, SANSOM CE, BEAZLEY C, MOSS DS: A novel predictive technique for the MHC class II peptide-binding interaction. *Mol. Med.* (2003) **9**(9-12):220-225.
61. WAN S, COVENEY P, FLOWER DR: Large-scale molecular dynamics simulations of HLA-A*0201 complexed with a tumor-specific antigenic peptide: can the α3 and β2m domains be neglected? *J. Comput. Chem.* (2004) **25**(15):1803-1813.
62. WAN S, COVENEY PV, FLOWER DR: Molecular basis of peptide recognition by the TCR: affinity differences calculated using large scale computing. *J. Immunol.* (2005) **175**(3):1715-1723.

•• **GRID-enabled supercomputing brings immunological molecular dynamics simulations within reach.**

63. WAN S, COVENEY PV, FLOWER DR: Peptide recognition by the T cell receptor: comparison of binding free energies from thermodynamic integration, Poisson-Boltzmann and linear interaction energy approximations. *Philos. Transact. A Math Phys. Eng. Sci.* (2005) **363**(1833):2037-2053.
64. DAVIES MN, HATTOTUWAGAMA CK, MOSS DS, DREW MG, FLOWER DR: Statistical deconvolution of enthalpic energetic contributions to MHC–peptide binding affinity. *BMC Struct. Biol.* (2006) **6**:5.

65. SAXOVA P, BUUS S, BRUNAK S, KESMIR C: Predicting proteasomal cleavage sites: a comparison of available methods. *Int. Immunol.* (2003) **15**(7):781-787.

66. DOYTCHINOVA I, HEMSLEY S, FLOWER DR: Transporter associated with antigen processing preselection of peptides binding to the MHC: a bioinformatic evaluation. *J. Immunol.* (2004) **173**(11):6813-6819.

67. DOYTCHINOVA IA, GUAN P, FLOWER DR: EpiJen: a server for multistep T cell epitope prediction. *BMC Bioinformatics* (2006) **7**:131.

68. PETERS B, SETTE A: Generating quantitative models describing the sequence specificity of biological process with the stabilized matrix method. *BMC Bioinformatics* (2005) **6**:132.

69. LARSEN MV, LUNDEGAARD C, LAMBERTH K *et al.*: An integrative approach to CTL epitope prediction: a combined algorithm integrating MHC class I binding, TAP transport efficiency and proteasomal cleavage predictions. *Eur. J. Immunol.* (2005) **35**:2295-2303.

70. DÖNNES P, KOHLBACHER O: Integrated modeling of the major events in the MHC class I antigen processing pathway. *Protein Sci.* (2005) **14**:2132-2140.

71. HATTOTUWAGAMA CK, TOSELAND CP, GUAN P *et al.*: Toward prediction of class II mouse major histocompatibility complex peptide binding affinity: *in silico* bioinformatic evaluation using partial least squares, a robust multivariate statistical technique. *J. Chem. Inf. Model* (2006) **46**(3):1491-1502.

72. DOYTCHINOVA IA, FLOWER DR. Modeling the peptide–T cell receptor interaction by the comparative molecular similarity indices analysis-soft independent modeling of class analogy technique. *J. Med Chem.* (2006) **49**(7):2193-2199.

73. CARSON RT, VIGNALI KM, WOODLAND DL, VIGNALI DA: T cell receptor recognition of MHC class II-bound peptide flanking residues enhances immunogenicity and results in altered TCR V region usage. *Immunity* (1997) 7(3):387-399.

74. GODKIN AJ, SMITH KJ, WILLIS A *et al.*: Naturally processed HLA class II peptides reveal highly conserved immunogenic flanking region sequence preferences that reflect antigen processing rather than peptide–MHC interactions. *J. Immunol.* (2001) **166**(11):6720-6727.

75. NOGUCHI H, HANAI T, HONDA H, HARRISON LC, KOBAYASHI T: Fuzzy neural network-based prediction of the motif for MHC class II binding peptides. *J. Biosci. Bioeng.* (2001) **92**(3):227-231.

76. BURDEN FR, WINKLER DA: Predictive Bayesian neural network models of MHC class II peptide binding. *J. Mol. Graph Model* (2005) **23**(6):481-489.

77. YANG ZR, JOHNSON FC: Prediction of T cell epitopes using biosupport vector machines. *J. Chem. Inf. Model* (2005) **45**(5):1424-1428.

78. MALLIOS RR: Predicting class II MHC/peptide multi-level binding with an iterative stepwise discriminant analysis meta-algorithm. *Bioinformatics* (2001) **17**(10):942-948.

79. DOYTCHINOVA IA, FLOWER DR: Towards the *in silico* identification of class II restricted T cell epitopes: a partial least squares iterative self-consistent algorithm for affinity prediction. *Bioinformatics* (2003) **19**(17):2263-2270.

80. MURUGAN N, DAI Y: Prediction of MHC class II binding peptides based on an iterative learning model. *Immunome Res.* (2005) **1**:6.

81. NOGUCHI H, KATO R, HANAI T *et al.*: Hidden Markov model-based prediction of antigenic peptides that interact with MHC class II molecules. *J. Biosci. Bioeng.* (2002) **94**(3):264-270.

82. KARPENKO O, SHI J, DAI Y: Prediction of MHC class II binders using the ant colony search strategy. *Artif. Intell. Med.* (2005) **35**(1-2):147-156.

83. NIELSEN M, LUNDEGAARD C, WORNING P *et al.*: Improved prediction of MHC class I and class II epitopes using a novel Gibbs sampling approach. *Bioinformatics* (2004) **20**(9):1388-1397.

84. YU K, PETROVSKY N, SCHONBACH C, KOH JY, BRUSIC V: Methods for prediction of peptide binding to MHC molecules: a comparative study. *Mol. Med.* (2002) **8**(3):137-148.

85. PETERS B, BUI HH, FRANKILD S *et al.*: A community resource benchmarking predictions of peptide binding to MHC-I molecules. *PLoS Comput. Biol.* (2006) **2**(6):e65.

•• **Key benchmarking study comparing predictions with ~ 50K binding affinities.**

86. MOUTAFTSI M, PETERS B, PASQUETTO V *et al.*: A consensus epitope prediction approach identifies the breadth of murine T(CD8$^+$)-cell responses to vaccinia virus. *Nat. Biotechnol.* (2006) **24**(7):817-819.

• **Binding predictions prove their worth.**

87. DOYTCHINOVA IA, WALSHE VA, JONES NA, GLOSTER SE, BORROW P, FLOWER DR: Coupling *in silico* and *in vitro* analysis of peptide–MHC binding: a bioinformatic approach enabling prediction of superbinding peptides and anchorless epitopes. *J. Immunol.* (2004) **172**(12):7495-7502.

• **Immunoinformatics can be used to design as well as predict; definitive proof that motifs have been superceded.**

88. BLYTHE MJ, FLOWER DR: Benchmarking B-cell epitope prediction: Underperformance of existing methods. *Protein Sci,* (2005) **14**:246-248.

• **Traditional B-cell epitope prediction fails, and fails badly.**

89. SOLLNER J, MAYER B: Machine learning approaches for prediction of linear B-cell epitopes on proteins. *J. Mol. Recognit.* (2006) **19**:200-208.

90. SOLLNER J: Selection and combination of machine learning classifiers for prediction of linear B-cell epitopes on proteins. *J. Mol. Recognit.* (2006) **19**(3):209-214.

91. SAHA S, RAGHAVA GP: Prediction of continuous B-cell epitopes in an antigen using recurrent neural network. *Proteins* (2006) **65**:40-48.

92. LARSEN JE, LUND O, NIELSEN M: Improved method for predicting linear B-cell epitopes. *Immunome Res.* (2006) **2**:2.

93. ANDERSEN PH, NIELSEN M, LUND O: Prediction of residues in discontinuous B-cell epitopes using protein 3D structures. *Protein Sci.* (2006) **15**(11):2558-2567.

94. VAN REGENMORTEL MH: Immunoinformatics may lead to a reappraisal of the nature of B-cell epitopes and of the feasibility of synthetic peptide vaccines. *J. Mol. Recognit.* (2006) **19**:183-187.

95. BENJAMIN DC, BERZOFSKY JA, EAST IJ *et al.*: The antigenic structure of proteins: a reappraisal. *Annu. Rev. Immunol.* (1984) **2**:67-101.

96. DOYTCHINOVA IA, FLOWER DR: *In silico* identification of supertypes for class II MHCs. *J. Immunol.* (2005) **174**(11):7085-7095.

97. DOYTCHINOVA IA, GUAN P, FLOWER DR: Identifiying human MHC supertypes using bioinformatic methods. *J. Immunol.* (2004) **172**(7):4314-4323.

98. LUND O, NIELSEN M, KESMIR C *et al.*: Definition of supertypes for HLA molecules using clustering of specificity matrices. *Immunogenetics* (2004) **55**(12):797-810.

99. DOYTCHINOVA IA, FLOWER DR: Predicting candidate vaccine antigens using alignment-free method based on principal amino acid properties. *Vaccine* (2007) **25**:856-866.

- **Prediction starts to escape from the epitope ghetto.**

100. AALBERSE RC, STADLER BM: *In silico* predictability of allergenicity: from amino acid sequence via 3D structure to allergenicity. *Mol. Nutr. Food Res.* (2006) **50**(7):625-627.

101. KUTTLER C, NUSSBAUM AK, DICK TP, RAMMENSEE H-G, SCHILD H, HADELER KP: An algorithm for the prediction of proteasomal cleavages. *J. Mol. Biol.* (2000) **298**(3):417-429.

102. NUSSBAUM AK, KUTTLER C, HADELE KP, RAMMENSEE H-G, SCHILD H: PAProC: a prediction algorithm for proteasomal cleavages available on the WWW. *Immunogenetics* (2001) **53**(1):87-94.

Affiliation

Matthew N Davies[1], Pingping Guan[2], Martin J Blythe[1], Jesper Salomon[1], Christopher P Toseland[3], Channa Hattotuwagama[4], Valerie Walshe[1], Irini A Doytchinova[5] & Darren R Flower[†1]
[†]Author for correspondence
[1]The Jenner Institute, University of Oxford, Compton, Berkshire, RG20 7NN, UK
Tel: +44 01635 577954;
Fax: +44 01635 577901/577908;
E-mail: Darren.Flower@jenner.ac.uk
[2]John Innes Centre, Norwich, NR4 7UH, UK
[3]National Institute for Medical Research, Mill Hill, London, NW7 1AA, UK
[4]GlaxoSmithKline, New Frontiers Science Park (North), Fourth Avenue, Harlow, Essex, CM19 5AW, UK
[5]Faculty of Pharmacy, Medical University of Sofia, Dunav st. 2, 1000 Sofia, Bulgaria