2

Immunoinformatics and Computational Vaccinology: A Brief Introduction

Paul D Taylor and Darren R Flower

The Jenner Institute, University of Oxford, Compton, Berkshire, RG20 7NN, UK
Darren.flower@jenner.ac.uk

Summary. Immunoinformatics has recently emerged as a buoyant and dynamic sub-discipline within the wider field of bioinformatics. Immunoinformatics is the application of bioinformatic methods to the unique problems of immunology and vaccinology. Immunoinformatics, as a principal component of incipient immunomic technologies, is beginning to foment important changes within immunology, as this key discipline tries to free itself from the empirical straight jacket that has characterised its development and attempts to grapple with the post-genomic revolution. Immunoinformatics is, importantly, also beginning to establish itself as a pivotal tool within vaccine discovery.

2.1 Introduction

Have you ever had a bout of the common cold? Do you suffer from Hay Fever or a nut allergy or Asthma? Have you ever had a more serious infectious disease? Are you a victim of a chronic autoimmune disease or even cancer? Now answer a seemingly distinct and unrelated question: have you ever used a computer? If you answered yes to either group of questions, have you ever thought of combining the two? The use of computers to fight infectious disease and other acute and chronic disease states may seem far fetched to many, but computers have long been used in the design of small molecule drugs, and now they are beginning to impact on the design and discovery of immunotherapeutics and prophylactic vaccines. We see this dramatic synergy made manifest through the discipline of immunoinformatics, a profound and exciting new computational science able to greatly accelerate the speed and effectiveness of vaccine and immunotherapeutic discovery.

The domain of infectious disease - allergy, in all its forms; autoimmune disease, such as rheumatoid arthritis; and even cancer - is the domain of Immunology. Immunology is, amongst other many other things, the study of how the body defends itself against infection, and from the standpoint of human disease, a proper appreciation of innate and adaptive immunity is crucial, as the immune system has evolved, at least in part, to combat infectious disease, the greatest source of preventable human mortality and morbidity. Immunology is thus a very broad branch of the biosciences which has led, directly or indirectly, to many pivotal advances in modern bio-medicine. Moreover, our knowledge of the molecular and cellular mechanisms which underlie immunity has also allowed for the development of new clinical and non-clinical technologies, which have an equally broad range of applications. While much of its focus remains strongly anthrocentric, or, at least, centred on the adaptive immune system of vertebrates, its societal importance can not be gainsaid, as it deals with the physiological function of the immune system in both health and disease; the malfunctioning of immunity in immunological disorders (autoimmune diseases, allograph rejection, hypersensitivities, and immune deficiency); and the in vitro, in vivo, and in situ, chemical and physiological properties of immunological components of the immune system.

However, immunology, and all its attendant disciplines, now find themselves at a turning point, whether or not practitioners realize it. After a hundred years of empirical research, immunology is increasingly poised to reinvent itself as a quantitative, genome-based science. Like most bioscience disciplines, immunology is increasingly facing the challenge of capitalizing on a potentially overwhelming cascade of new information delivered by high-throughput, post-genomic technologies. This data is both bamboozlingly complex and on a scale which has never been encountered before. It is also clear, at least to some, that such high throughput approaches will engender a paradigm-shift from traditional hypothesis-driven research to a new data-driven, information-focused approach, with new understanding emerging from the analysis of complex, intricate, multifaceted datasets.

In response to pressures such as these, there has been much recent interest in the development and deployment of informatics tools, which can analyze the data that arises from immunological research of all kinds. In turn, this has lead to the growth of two flavours of computational support for immunology. The first is straightforward bioinformatics support, technically indistinguishable from support given to other areas of biology, and includes genome annotation of both the human genome and diverse microbial species. For example, well in excess of 150 bacterial genomes have now been sequenced, and hundreds more are nearing completion [Paine & Flower 2002]. Another area of growth is immunotranscriptomics, or immunologically targeted Microarray analysis [Walker *et al.* 2002].

The other kind of support is more focussed: immunoinformatics. This is an exciting and dynamic specialism, which has emerged in recent years within the wider world of bioinformatics. It addresses the particular problems which arise within immunology, including the accurate prediction of immunogenicity, be that manifest as the identification of epitopes or the prediction of whole protein immunogens; this endeavour stands as the principal short- to medium-term goal of immunoinformatic investigation. The theoretical or mathematical modeling of the immune systems seeks, as a discipline, to address what some are wont to call *important scientific questions*: how might immunity work? What is the nature of host-pathogen interactions? Work of this sort is described in various Chapters elsewhere in this book: Chapter 4 by Lee and Perelson, which describes computational modeling of the function of B cell and T cell receptors; Chapter 13 by Denise Kirschner, which describes multi-scale modeling of the immune system in response to pathogens; and Chapter 17 by Melvin Cohn, which describes the self/non-self paradox. Immunoinformatics, on the other hand, is concerned with prosaic, nitty-gritty, nuts-and-bolts issues: is this particular amino acid sub-sequence of a protein an epitope? Is this protein within a viral genome more antigenic than another? Can we identify common virulence factors in the genomes of a distinct phylogenetic grouping of bacteria? It is with questions of this sort that immunoinformatics concerns itself; it is a discipline which rolls up its sleeves and gets on with the job.

Like bioinformatics, immunoinformatics is grounded in computer science. Increasingly, however, immunoinformatics integrates a whole array of cross-disciplinary techniques from physical biochemistry and biophysics; computational, medicinal, and analytical chemistry; structural biology and protein homology modeling, as well as many other branches of biological, physical, and computational science. Traditionally, it has emphasized problem solving and focused on data classification into discrete sets rather than predicting continuous, quantitative data, leading to the use of *black-box* neural networks for prediction and to databases such as SYFPEITHI [Rammensee *et al.* 1999]. Increasingly, however, approaches are turning towards more quantitative models, familiar from decades of QSAR analysis of drug molecules, which predict continuous binding measures. This approach is more overtly physico-chemical in nature, with a greater implicit emphasis on the explanation of underlying atomistic molecular mechanisms. These different points of view are highly complementary. Remaining conflicts between these differing perspectives are easily reconciled by methods from Drug Design. Such methods meet both objectives: seeking to explain and understand without sacrificing efficiency or loosing sight of the pragmatic and utilitarian purpose of the undertaking.

It is perhaps a cliché, or at least a truism, to say that the immune system is complex, complicated, and hierarchical, exhibiting considerable emergent behaviour at every level from subcellular to organismal. Yet, for all that, this aphorism retains an essential veracity. If it were not true, then the book you are reading, *in silico immunology,* would not need to be written. The complexity of the immune systems is confounding, and, though many might wish to deny it, our ignorance of it remains profound. Yet, at the heart of the immune system lie straightforward molecular recognition events: the coming together of two or more molecules to form stable complexes of measurable duration. In terms of atomistic interactions, these events are indistinguishable from the binding phenomena experienced by any macromolecule. The binding of an epitope to a major histocompatibility complex protein (MHC), or T Cell Receptor (TCR) to a peptide-MHC complex is, thus, in terms of underlying physico-chemical phenomena, identical in nature to any other molecular interaction in any other area of bioscience. It is only at higher levels - when tens, or thousands, or millions of different molecules come together – that immune systems exhibit, in time and space, complex and confusing emergent behavior. In seeking to understand immunology and address its problems, immunoinformatics can exploit the observation that the immune system is based on simple, understandable molecular events, and, within the biological context of the subject, it does seek, in the broadest sense, to model such phenomena. Much of the rest of this chapter will explore how.

2.2 Immunogenicity: A Brief Primer

Immunogenicity is that property of a molecular, or supramolecular, moiety that allows it to induce a significant response from the immune system. Here a molecular moiety may be a protein, lipid, carbohydrate, or some combination thereof. A supramolecular moiety may be a virus, bacteria or protozoan parasite. An immunogen - a moiety exhibiting immunogenicity - is a substance which can elicit a specific immune response, while an antigen - a moiety exhibiting antigenicity - is a substance recognized, in a recall response, by the extant machinery of the adaptive immune response, such as T cells or antibodies. Thus, antigenicity is the capacity, exhibited by an antigen, for recognition by one or several parts of the antibody or TCR immune repertoire. Immunogenicity, on the other hand, is the ability of an immunogen to induce a specific immune response when it is exposed to initial surveillance by the immune system. These two properties are clearly coupled but properly understanding how they are inter-related is by no means facile.

Predicting actual antigenicity and/or immunogenicity of a complex protein remains problematic. It depends simultaneously upon the context in which it is presented and also the nature of the immune repertoire that recognizes it. Either or both of these components may be critical. For example, the immune response in many immunogens or antigens is focused to a handful of immunodominant structures, while much of the rest of the molecule may be unable to engender a response. Mutating an antigen may eliminate, reduce, or even enhance its inherent immunogenicity, or, of course, it might move it to other regions of the molecule. In seeking to assess immunogenicity, we must consider properties of the host and the pathogenic organism of origin, and not just the intrinsic properties of the antigen itself. The composition of the available immune repertoire will affect its response to a given epitope and alter its recognition of a particular target. When mounting a response *in vivo*, those elements of an immune repertoire capable of participating, in a given response, might have been deleted through their cross-reactivity with host antigens. Moreover, fundamental restrictions on the antibody repertoire, imposed by the limited number of V genes that encode the antigen-binding site of the antibody, may also limit responses. Overall, it is clear that antigenicity and immunogenicity have many interlinked causes. The induction of immune responses requires critical interaction between parts of the innate immune system, which respond rapidly and in a relatively nonspecific manner, and other, more specific components of the adaptive immune system, which can recognize individual epitopes.

Immunogenicity is currently the most important and interesting property for analysis and prediction by immunoinformatics. Immunogenicity can manifest itself through both arms of the adaptive immune response: humoral (mediated through the binding of whole protein antigens by antibodies) and cellular immunology (mediated by the recognition of proteolytically cleaved peptides by T cells). Humoral Immunogenicity, as mediated by soluble or membrane-bound cell surface antibodies, can be measured in several ways. Methods such as enzyme-linked immunosorbent assay (ELISA) or competitive inhibition assays yield values for the Antibody Titre, the concentration at which the ability of antibodies in the blood to bind an antigen has reached its half maximal value. One can also measure directly the affinity of antibody and antigen, using, for example, equilibrium dialysis. Likewise, measurements of cellular immunity through T cell responses have become legion. For class I presentation, arguably the most direct approach is to measure T cell killing. Cytotoxic T lymphocytes or CTL, can induce lysis in target cells. This can be measured using a radioisotope of chromium, which is taken into target cells and released during CTL lysis. For class II presentation, the proliferative response of CD4+ T cells, which acts indirectly by activating B cells or macrophages, can be measured using the incorporation of tritiated thymidine into T cell DNA during cell division. Alternatively, Enzyme-Linked ImmunoSpot, or ELISpot, assays measure production of cytokines or other molecules by class I and /or class II T-cells when exposed to antigen. More recently attention has migrated towards tetramers as tools for detecting T cell responses [Doherty *et al.* 2000]. MHC:peptide tetramers are formed from four biotinylated peptide-MHC complexes (pMHCs) bound to tetrameric avidin or streptavidin. These tetramers bind to TCRs with a proportionately higher affinity allowing antigen-T cell interactions to be assessed with greatly enhanced specificity.

Much of immunogenicity is determined by the presence of epitopes, the principal chemical moieties recognized by the immune system. Consequently, the accurate prediction of B cell and T cell epitopes is the pivotal challenge for immunoinformatics. Despite a growing appreciation of the role played by non-peptide epitopes, such as carbohydrates and lipids, nonetheless peptidic B cell and T cell epitopes (as mediated by the humoral or cellular immune systems respectively) remain the principal tools by which the intricacy of immune responses can be surveyed and manipulated, since it is the recognition of epitopes by T cells, B cells, and soluble antibodies that lies at the heart of the adaptive immune response. Such initial responses lead, in turn, to the activation of the cellular and humoral immune systems and, ultimately, to the effective destruction of pathogenic organisms.

While the prediction of B cell epitopes remains primitive and largely unsuccessful [Blythe & Flower 2005], a multitude of sophisticated, and successful, methods for the prediction of T cell epitopes have been developed [Flower *et al.* 2002]. These began with early motif methods [Doytchinova *et al.* 2004c], and have grown to exploit both qualitative and semi-quantitative approaches, typified by neural network classification methods, and a variety of more quantitative techniques [Doytchinova & Flower 2002c]. Most modern methods for T cell epitope prediction rely on predicting the affinity of peptides binding to MHCs. The T cell, a specialized type of immune cell mediating cellular immunity, constantly patrols the body seeking out foreign proteins derived from microbial pathogens. T cells express a particular receptor: the T cell receptor or TCR, which exhibits a wide range of selectivities and affinities. TCRs bind MHCs, which are presented on the surfaces of other cells. These proteins in turn bind small peptide fragments (epitopes) originating from both endogenous, or self, and exogenous, including pathogen-derived, protein sources. It is, as we have said, the recognition of such complexes that lies at the heart of both the adaptive, and memory, cellular immune response.

2.2.1 T cell and Antibody Repertoire

Until recently, the immune system was thought to discriminate rigidly between "self" and "non-self". This discrimination was believed to form the basis of protection of the host against pathogens. Such views are changing as studies indicate that determinants of self do not always induce absolute immune tolerance in the host. Under certain conditions peptides from self-antigens can be processed and displayed by MHC as targets for immune surveillance. This provides a rationale for the investigation of, say, self-epitopes as mediators of autoimmunity, or epitopes from cancer antigens as targets for immunotherapy or targeting epitopes from proteins which induce allergic reactions.

As we have said, MHCs bind peptides. These are themselves derived through the degradation, by protelotytic enzymes, of foreign and self proteins. Foreign epitopes originate from benign or pathogenic microbes, such as viruses and bacteria. Self epitopes originate from host proteins that find their way into the degradation pathway as part of the cell's intrinsic quality control procedures. The proteolytic pathway by which peptides become available to MHCs is very complex and many, many important details and molecular components remain to be elucidated. Yet, it is the complexity and degeneracy of the T cell presentation pathway that allows peptides with diverse post-translational modifications, such as phosphorylation or glycosylation, to form pMHCs, and thus, ultimately, to be recognized by TCRs. Moreover, MHCs are very catholic in terms of the molecules they bind and are not restricted to peptides. Chemically modified peptides and peptidomimetics are also bound by MHCs. It is also well known that many drug-like molecules bind to MHCs [Pichler 2002].

As we shall see below, the overall presentation process is long, complicated, and involves many subsidiary steps. There are several alternative processing pathways, but the principal ones seem linked to the two major types of MHC: Class I and Class II. Class I MHCs are expressed by almost all nucleated cells in the body. They are recognized by T cells whose surfaces are rich in CD8 co-receptor protein. Class II MHCs are only really expressed on so-called *professional antigen presenting cells* and are recognized by T cells whose surfaces are rich in CD4 co-receptors. MHCs are polymorphic. Generally, most humans have six classic MHCs: 3 Class I (HLA-A, B, and C) and 3 class II (HLA-DR, DP, and DQ), these proteins will have different sequences, or different HLA alleles, in different individuals. Different MHC alleles, both class I and Class II, have different peptide specificities. A simple way to look at this phenomenon is to say that MHCs bind peptides which exhibit certain particular sequence patterns and not others. Within the human population there are an enormous number of different, variant genes coding for MHC proteins, each of which exhibits discernibly different peptide-binding sequence selectivities. T cell receptors, in their turn, also exhibit different and typically weaker affinities for different peptide-MHC complexes. The combination of MHC and TCR selectivities thus determines the power of peptide recognition in the immune system and thus the recognition of foreign proteins and pathogens. This will be discussed more thoroughly in accompanying chapters. Whatever dyed-in-the-wool immunologists may say, such interactions form the quintessential nucleus of immune recognition, and thus the principal point of intervention by immunotherapeutics.

2.3 Epitopes and Epitology

The word *epitope* is widely used amongst biological scientists. Etymologically speaking, its roots are Greek, and, like most words, its meanings are diverse and in a state of constant flux. It is most often used to refer to any region of a biomacromolecule which is recognized, or bound, by another biomacromolecule. For an immunologist, the meaning is more restricted and refers to particular structures recognized by the immune system in particular ways. The region on a macromolecule, which undertakes the recognition of an epitope, is called a paratope. In terms of the physical chemistry of binding, then we need think only of equal partners in a binding reaction. B cell epitopes are regions of a protein recognized by Antibody molecules. T cell epitopes are short peptides which are bound by major histocompatibility complexes (MHC) and subsequently recognized by T cells.

A B cell epitope is a region of a protein, or other biomacromolecule, recognized by soluble or membrane-bound Antibodies. B-cell epitopes are classified as either linear or discontinuous epitopes. Linear epitopes comprise a single continuous stretch of amino acids within a protein sequence, while an epitope whose residues are distantly separated in the sequence and are brought into physical proximity by protein folding is called a discontinuous epitope. Although most epitopes are, in all likelihood, discontinuous, experimental epitope detection has focused on linear epitopes. Linear epitopes are believed to be able to elicit antibodies that can subsequently cross-react with its parent protein.

A T cell epitope is a short peptide bound, in turn, by MHC and TCR, to form a ternary complex. The formation of such a complex is the primary, but not sole, molecular recognition event in the activation of T cells. Many other co-receptors and accessory molecules, in addition to CD4 and CD8 molecules, are also involved in T cell recognition. The recognition process is not simple and remains poorly understood. However, it has emerged that the process involves the creation of the immunological synapse, a highly organised, spatio-temporal arrangement of receptors and accessory molecules of many types. The involvement of these accessory molecules, although essential, is not properly understood, at least from a quantitative perspective. Ultimately, the accurate modelling of all these complex processes will be required to gain full and complete insight into the process of epitope presentation.

We will explore how epitopes arise rather more fully below. The peptides presented by class I and class II MHCs differ, principally in terms of their length. Class I peptides are primarily derived from intracellular proteins, such as viruses. These proteins are targeted to the proteasome, which cuts them into short peptides. Subsidiary enzymes also cleave these peptides, producing a range of peptide lengths, of which the distribution used to be believed to fall neatly into the range: 8 to 11 amino acids. More recently, however, this has been shown that much longer peptides, currently up to 15 amino acids, can also be bound by MHCs and recognized by TCRs [Probst-Kepper *et al.* 2001]. For Class II, the receptor mediated intake of extracellular protein derived from a pathogen is targeted to an endosomal compartment, where such proteins are cleaved by cathepsins, a particular class of protease, to produce peptides which are typically somewhat longer than Class I. These, again, exhibit a considerable distribution of lengths, centred on a range of 15-20 amino acids. However, longer and shorter peptides can also be presented, via Class II MHCs, to immune surveillance. Peptide cleavage specificity exhibited by Cathepsins has also been investigated and some insight has been gained into cleavage motifs[Chapman 1998]. However, considerably more work is required before truly efficient predictive methods can be realized.

It is now generally accepted that only peptides that bind to MHC at an affinity above a certain threshold will act as T cell epitopes and that, to some extent at least, peptide affinity for the MHC correlates with T cell response. This particular issue is somewhat complicated and obscured by hearsay and dogma: as with many questions important to immunoinformatics, the key, systematic studies remain to be done. The behaviour of heteroclitic peptides, where synthetic enhancements to binding affinity are often reflected in enhanced T cell reactivity, seems compelling evidence of the relationship between affinity and immunogenicity. However, and whatever people may say, affinity of binding is an important component of recognition and thus of the overall process leading to the generation of an immune response. Not the only, or, necessarily, the most important part, but a key part nonetheless. Its importance is debated, particularly by people critical of the immunoinformatic endeavour. Nevertheless, its utility in this context is clear. Experimental immunologists and vaccinologists are constantly using nascent immunoinformatic approaches to select, filter, or prune lists of candidate peptides in order to identify functional epitopes.

Epitopes, whether B cell or T cell, are, as we have mentioned several times above, short continuous or discontinuous sequences or strings of amino acids. These may be of different length and exist in different contexts, but they remain sequences. As such they can be stored in functional immunological databases, much as whole sequences are stored in GenBank or Swiss-Prot. As we shall see in the next section, there are many of such resources, most available via the Internet.

2.4 Databases

For some time, the database has been the *lingua franca* or, more prosaically, the common language of bioinformatics. The creation, and the manipulation of databases, which contain biologically relevant information, is the most critical feature of contemporary bioinformatics. The same is true of immunoinformatics. This is manifest through its support for post-genomic bioscience and as a discipline in its own right. Functional data, as housed in databases, will rapidly become the principal currency in the dynamic information economy of 21^{st} Century immunology. Having said all that, there is nothing particularly new about immunological databases, at least in the sense that they do no more than apply standard data warehousing techniques in an immunological context. Nonetheless, the continuing development of an expanding variety of immunoinformatic database systems indicates that the application of bioinformatics to immunology is beginning to broaden and mature.

Example databases, such as IMGT [Robinson *et al.* 2003] or Kabat [Wu & Kabat 1970], have made the sequence analysis of important immunological macromolecules their focus for many years [Brusic *et al.* 2002]. Functional, or epitope-orientated, databases are somewhat newer, but their provenance is now well established. Generally speaking, such databases record data on T cell epitopes or peptide-MHC binding affinity. Arguably, the highest quality database currently available is the HIV Molecular Immunology database [Korber *et al.* 2001b], which focuses on the sequence and the sequence variations of a single virus, albeit one of unique medical important. However, the scope of the database is, in terms of the kinds of data it archives, less restricted than others, containing information on both cellular immunology (T cell epitopes and MHC binding motifs) and humoral immunology (linear and conformational B cell epitopes).

An early, and widely used, database is SYFPEITHI [Rammensee *et al.* 1999], another high quality development, which contains an up-to-date and useful compendium of T cell epitopes. SYFPEITHI also contains much data on MHC peptide ligands, peptides isolated from cell surface MHC proteins *ex vivo*, but purposely excludes binding data on synthetic peptide. MHCPEP [Brusic *et al.* 1998], a now defunct database, pooled both T cell epitope and MHC binding data in a flat file, introducing a widely used conceptual simplification, which combines together the bewildering variety of binding measures, reclassifying peptides as either *Binders* or *Non-Binders.* Binders are further sub-divided as High-binders, Medium binders, and Low binders. More recently, Brusic and coworkers have developed a much more complex and sophisticated database: FIMM [Schonbach *et al.* 2005]. This system integrates a variety of data on MHC-peptide interactions: in addition to T cell epitopes and MHC-peptide binding data, it also archives a wide variety of other data, including sequence data on MHCs themselves together with data on the disease associations of particular MHC alleles .

More recently, related, yet distinct, databases have begun to emerge, each addressing data on different aspects of molecular immunology. Kangueane and coworkers have developed a database that focuses solely on X-ray crystal structures of MHC-peptide complexes [Govindarajan *et al.* 2003], while Middleton et al. describe the Allele

Frequency Database which lists population frequencies of particular MHC alleles [Middleton *et al.* 2003]. All these databases are available via the Internet.

AntiJen [Toseland *et al.* 2005], formerly known as JenPep[Blythe *et al.* 2002, McSparron *et al.* 2003], is a database developed recently, which brings together a variety of kinetic, thermodynamic, functional, and cellular data within immunobiology. While it retains a focus on both T cell and B cell epitopes, AntiJen is the first functional database in immunology to contain continuous quantitative binding data on a variety of immunological molecular interactions, rather than the kind of subjective classifications described above. Data archived includes thermodynamic and kinetic measures of peptide binding to TAP and MHC, peptide-MHC complex binding to T cell receptors, and general immunological protein-protein interactions, such as the interaction of co-receptors, interactions with superantigens, etc. Although the nature of the data within AntiJen sets it apart from other immunology databases, there is, nonetheless, considerable overlap between other systems and AntiJen.

Moreover, AntiJen shares characteristics with several other newly-emergent non-immunological databases: thermodynamic binding databases, such as BindingDB [Chen *et al.* 2002], and a variety of other databases of different sorts, of which BIND [Bader & Hogue 2000] and Brenda [Schomburg *et al.* 2002] are prime examples. Such databases, which contain experimental measured binding affinities, are a relatively recent development. The focus of these databases is rigorously measured thermodynamic properties derived from experimental protocols such as Isothermal Titration Calorimetry (ITC), which can return not only free energies of binding, but also equivalent enthalpies, entropies, and heat capacities. As these protocols are well standardized, databases, such as BindingDB, can easily record precise experimental conditions.

In the domain of immunological experiments, AntiJen records IC_{50}, BL_{50}, t1/2 measurements, etc. For each such measurement, it also archives standard experimental details, such as pH, temperature, the concentration range over which the experiment was conducted, the sequence and concentration of the reference radiolabeled peptide competed against, together with their standard deviations. As it is rare to find a paper which records all such data in a reliable way, thus standardization remains a significant issue. It is also unclear how much more data remains to be collated.

Server Name	URL	
BIMAS	http://bimas.dcrt.nih.gov/molbio/hla_bind/	
EpiGenomix	http://epigenomix.com/pls/hiv/!www_user.front_end	
HLA-DR4 binding	http://www-dcs.nci.nih.gov/branches/surgery/sbprog.html	
LpPep	http://reiner.bu.edu/zhiping/lppep.html	
MHCPred	http://www.jenner.ac.uk/MHCPred	
MHC-THREAD	http://www.csd.abdn.ac.uk/~gjlk/MHC-Thread/	
NetMHC	http://www.cbs.dtu.dk/services/NetMHC/	
PREDEP	http://bioinfo.md.huji.ac.il/marg/Teppred/mhc-bind/	
PREDICT	http://sdmc.krdl.org.sg:8080/predict/	
ProPred	http://www.imtech.res.in/raghava/propred/	
RankPep	http://www.mifoundation.org/Tools/rankpep.html	
SVMHC	http://www.sbc.su.se/svmhc/	
SYFPEITHI	http://syfpeithi.bmi-heidelberg.com/Scripts/MHCServer.dll/EpiPredict.htm	
BIMAS	http://bimas.dcrt.nih.gov/molbio/hla_	bind MHC class I ligands
MAPPP	http://www.mpiib-berlin.mpg.de/MAPPP	MHC class I and proteasome cleavage
NetChop	http://www.cbs.dtu.dk	Proteasome cleavage
NetMHC	http://www.cbs.dtu.dk/services/NetMHC	HLA-A2 and H-2Kk
PAProC	http://www.paproc.de	Proteasome cleavage
ProPred	http://www.imtech.res.in/raghava/propred	HLA-DR
ProPred1	http://www.imtech.res.in/raghava/propred1	MHC class I ligands
SYFPEITHI	http://www.syfpeithi.de	MHC class I and II ligands
RANKPEP	http://www.mifoundation.org/Tools/rankpep.html	MHC class I & II ligands
SVMHC	http://www.sbc.su.se/svmhc	MHC class I ligands
Lib Score	http://www.ddbj.nig.ac.jp/analysesp-e.html	MHC class I ligands

Table 2.1. URLs of Prediction Servers

2.5 Immunoinformatic Datamining

A useful simplification of biological computation is to split methods between the areas of datamining and simulation. In truth, of course, there is a continuous spectrum of techniques stretching from one extreme to the other. Within immunology, a key example of data mining is the identification of peptide binding motifs, which seeks to characterize the peptide specificity of different MHC alleles in terms of dominant anchor positions with a strong preference for certain amino acids [Sette *et al.* 1989]. Such motifs are undoubtedly popular amongst immunologists, as they are simple to understand and just as simple to implement either visually or computationally. For example, human Class I allele HLA-A*0201, probably the best-studied allele, has anchor residues at peptide positions P2 and P9. At P2, acceptable amino acids would be L and M, and V and L at position P9. Secondary anchors, which are residues that are favourable, but not essential, for binding, may also be present. A seemingly uncountable number of papers have, over the past 15 years or so, successfully extended this to include the specificity patterns of many other alleles, both human and animal. However, despite this success, there are many fundamental problems with the motif approach.

The most significant of problem with motifs is that they are deterministic: a peptide either is, or is not, a binder. A brief reading of the literature shows that motif matches produce many false positives, and probably also produces an equal number of false negatives, although such negative results are seldom screened. Thus being motif-positive, as the jargon can put it, is neither necessary not sufficient for affinity for an MHC. Although it is clear that so-called primary anchors do often dominate binding, it is well known, that binding motifs, as descriptions of the process, are fundamentally flawed. Not hopeless, not useless, but partial and incomplete. In the sense that motifs are widely used and widely understood, they are indeed most useful, but as accurate predictors of binding they leave much to be desired. Such shortcomings have led many to seek other data mining solutions to the peptide-MHC affinity problem.

The development of data driven predictive methods in immunoinformatics is now two decades old. Early methods attempted to predict epitopes directly, and, in the absence of knowledge of the peptide preferences of MHC restriction, enjoyed limited success [Deavin *et al.* 1996, Flower 2003]. As described in chapter 8, several groups have used techniques from artificial intelligence research, such as artificial neural networks (ANNs) and hidden Markov models (HMMs), to tackle the problem of predicting peptide-MHC affinity [Brusic & Flower 2004]. ANNs and HMMs, are, for slightly different applications, particular favourites among bioinformaticians when looking for tools to build predictive models. However, the development of ANNs is often complicated by their preponderance for problems of interpretation, and also for overtraining and over-fitting. Of course, many other methods - indeed, in all probability, all methods - suffer similar or equivalent problems. Indeed, over-fitting is the curse of all data driven methods. Support vector machines are currently flavour-of-the-month. Whether this method, or indeed any other AIS-based approach, will ever escape the traps which have caught-out other techniques remains to be seen. A number of prediction servers are available over the web. See table 2.1.

In the prediction of MHC-binding, the main issues are the quality, quantity, ability to represent available data, the complexity of the selected predictive model relative to the natural complexity of the peptide-MHC interaction, and the training and testing of the predictive model. A good quality data set is critical to the creation of an accurate prediction system. Available data contains significant biases, as peptides are often pre-selected for experimental testing using binding motifs. Data is often intrinsically poor and requires data cleaning. Data quantity also has important implications for the selection of appropriate prediction methods. Guidelines have been given based on a recent comparative study of algorithm performance [Brusic & Flower 2004, Yu *et al.* 2002] and were suggested in the context of Artificial Intelligence techniques, which have well defined data requirements:

1. If there is no binding data at all, then speculative molecular modeling is the only option. Here, supertype analysis, as described later, can be useful.
2. When the number of available peptides is below 50, binding motifs are the most pragmatic solution.
3. With 50-100 peptides, quantitative matrices or SVMs can be used.
4. With data sets comprising over 100 peptides, HMMs or ANNs can be used.
5. With very large data sets, only really available for HLA-A*0201, ANNs can provide high specificity predictions, albeit at the price of slightly lowered sensitivity.

Our own QSAR methods have slightly different data requirements. For more information on these approaches see Chapter 8. The minimum set size is about 20 peptides, though models only begin to gain statistical significance at 40 peptides and above. When sets reach 200 or above, then it becomes possible to introduce reliable cross-terms: 1-2 and 1-3 side chain-side chain interactions in our case.

However, as we have explained above, it is not just quantity, but data diversity, that is an issue. As diversity in peptide sequence and binding affinity increases, so does the predictivity and generality of the models. Highly degenerate data or data with a very narrow affinity range often prove difficult. Predictive models should be tested before use, using internal cross-validation and the splitting of data into training and test sets.

2.6 Modelling T cell Mediated Antigen Presentation and Recognition

One of the most challenging problems in modern computational vaccinology is the effective modeling of the cellular presentation of antigenic epitopes. Professional antigen-presenting cells (APCs), such as dendritic cells or macrophages, endocytose and process protein antigens into peptides, which are subsequently presented on the cell surface associated with MHC molecules. This presentation can then result in the stimulation of cytotoxic or helper T cells. Conceptually, the phenomenon of

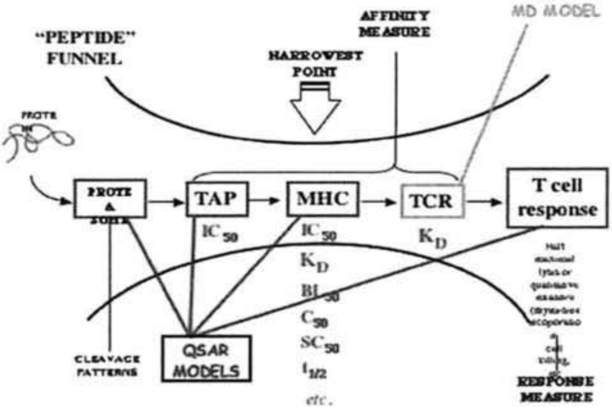

Fig. 2.1. The 'Simple' Class I Processing Pathway: A schematic showing a simplified view of class I antigen presentation. Peptides are generated initially from whole proteins via cleavage in the proteasome, followed by transport, into the endoplasmic reticulum (ER), by the transporter associated with processing (TAP). ERAAP trims peptides prior to binding by MHC molecules. MHCs are transported to the cell surface where they are recognized by T cells. The kind of measurable quantities, such as affinities or cleavage patterns, available for each step on the pathway are shown. This process approximates to a funnel with the principal bottleneck being binding by the MHC. The kind of model we have worked on (QSAR or MD model) is also indicated.

antigen presentation can be divided into three mechanistic stages. Firstly, antigen uptake: the recognition of antigen proteins by cell surface-receptors and the subsequent internalization of soluble, extracellular antigens. Secondly, antigen processing: intracellular enzymatic degradation and transport of endocytosed and cytoplasmic endogenous and exogenous proteins followed by peptide loading of MHC molecules. Thirdly, the exocytosis of MHC complexes, containing self and exogenous and endogenous antigenic peptides, and CD1, which presents potentially antigenic lipids. Put at its simplest, fragments of extracellular proteins are presented by class II MHCs and fragments of intracellular proteins are presented by class I MHCs; so-called cross-presentation refers to the presentation of extracellular antigens in the context of class I MHCs and vice-versa.

Much of what we adumbrate below will focus on class I antigen presentation; a somewhat simplified description of the many subsidiary steps involved in class I presentation is shown in Figure 2.1. Significant advances have been made recently in the modeling of class I presentation, particularly the prediction of proteasomal cleavage patterns and peptide binding to TAP. Together studies on proteasomal cleavage and TAP transport represent a good first attempt to produce useful predictive tools for the processing aspect of Class I restricted epitope presentation.

Cytosolic proteins, after labeling with ubiquitin, are transported to the proteasome, a multimeric protease responsible for most protein digestion within the cytosol, where they are cleaved into short peptides, typically 15 or fewer amino acids in length. Several methods have been development for predicting semi-stochastic proteasomal protein cleavage [Brusic & Flower 2004]. All perform statistical analysis of digested fragments from a small set of proteins, principally enolase-1, and augmented this sparse data-set with signals apparent in the termini of peptides eluted from cell surface MHCs. This developed the work of [Altuvia & Margalit 2000], who showed that the termini of peptides eluted from cell surface MHCs exhibit distinct sequence motifs at the C, but not the N, terminus, consistent with peptides undergoing N terminal trimming by other proteases subsequent to digestion by the proteasome. Several of these methods are available via the Internet. The predictive power shown by different prediction methods is only beginning to be evaluated objectively. [Saxova *et al.* 2003] evaluated three publicly available methods for proteasomal cleavage prediction, and found that the best method gave an accuracy value of 70% at the C-termini. Clearly, considerable progress is still required.

Peptides generated by the proteasome are subsequently bound by the transmembrane peptide transporter TAP, which translocates them from the cytoplasm to the endoplasmic reticulum (ER). In the ER peptides are bound by MHCs. A number of studies have been conducted into the peptide substrate specificities exhibited by the TAP transporter [Doytchinova *et al.* 2004a], leading to the development of several predictive models for the determination of peptides that bind to TAP. Most identify strong sequence patterns at the C-termini. This feature, also present in proteasome cleavage patterns, is consistent with a role for ERAAP in N-terminal *trimming* of peptides within the lumen of the endoplasmic reticulum.

So far, so good: a reasonably straightforward and uncomplicated linear pathway has been modeled with some success. However, there are, in reality, many other processing components and, indeed, whole presentation pathways, which greatly complicate the simple picture sketched out above. The growing complexity of antigen presentation is best exemplified by the class I processing pathway. See Figure 2.2. As well as the proteasome, peptides are cleaved by other cytosolic proteases, such as Tripeptidyl peptidase II (TPPII), currently the only well characterized protolytic enzyme known to be involved in presenting epitopes, although it is most probable that many others are involved. Peptides cleaved by the proteasome or TPPII are degraded by cytosolic proteases such as LAP and TOP. Peptides transported into the ER by TAP then bind to MHCs. This process is catalysed by a variety of chaperones, including Tapasin, calnexin, and ERp57. Peptides in the ER pool are trimmed by ERAAP and other proteases, such as L-RAP. Other anterograde and retrograde routes operate between the cytosol and the ER, by which means protein fragments can access different proteases, including puromycin resistant aminopeptidease.

At the other end of the process, extracellular proteins undergo antigen capture mediated by receptor-mediated endocytosis, entering the class I pathway through mechanisms of cross-presentation. The exact nature and number of such receptors remains obscure. Accurate modeling of this process is complicated by the observation

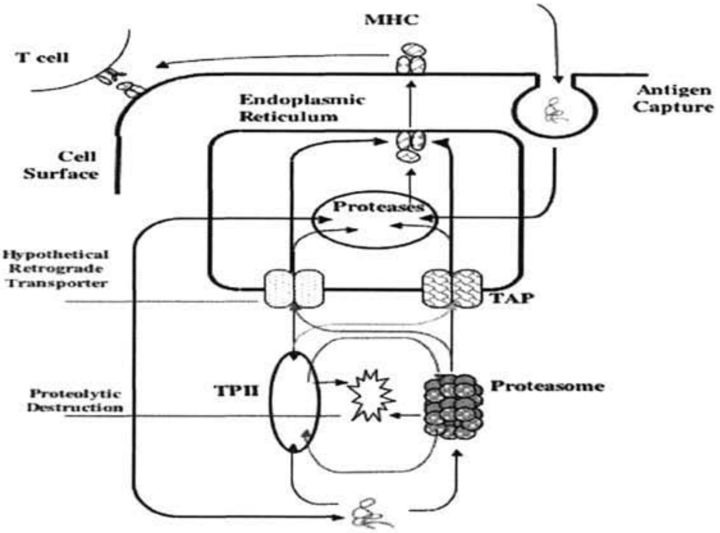

Fig. 2.2. Complex Antigen Presentation Pathway: A more realistic schematic of the class I antigen presentation pathway compared to Figure 2.1. This incorporates several proteases as well as different mechanisms of cross-presentation. Peptides are created or bound at indicated stage. The route taken by peptides is shown by arrows. Two points are of note: first, the synergistic interaction of TPPII and proteasome in the proteolytic creation of peptides and second the retrograde access to the cytosol.

that all cellular compartments or pools, whether conceptual or membrane bound, are leaky.

So far, not so good; at least from the view-point of someone trying to model the process. The accurate prediction of antigen processing and presentation depends on a proper understanding of the molecular mechanisms underlying the overall pathway. In order to develop a general model of cell surface epitope presentation, each of these steps would require its own predictive model for both the thermodynamics of peptide specificity and substrate-dependent peptide kinetics. The process requires decomposition into a set of peptide cleavage and peptide binding steps, each of which would then be open to modeling. This would not, in itself, account for the complexities of antigen presentation. Rather, we will need to supplement individual bioinformatic models with well understood mathematical models, such as those prototyped on reaction kinetics within multifurcating, multi-enzyme pathways: so called "metabolic control theory" [Fell 2005], which can account for substrate flux within multi-step, multi-component metabolic pathways, and allow for the ready incorporation of quantitative aspects of individual bioinformatic models. An effective model of this type, however hard to realize in practice, would, in all probability, better help us to understand why certain peptides come to dominate presentation: the apparently intractable problem of epitope immunodominance.

The presentation of peptides by the MHC is often viewed as the most discriminatory step of the presentation process. However, peptide recognition by the T cell is also vitally important. If we define recognition as the interaction of TCR and pMHC, then many complex subsequent steps are involved in the actual activation of T cells. Recognition is not an isolated event and the context in which an antigen is encountered by a T cell will determine if TCR engagement leads to full activation or tolerance. In the presence of costimulation, antigen presentation by an activated APC will lead to full activation. However, antigen presentation by a resting APC will lead to tolerance. Moreover, molecular recognition of toll-like receptor ligands by other receptors also has a key role in activating APCs and promoting the activation of T cells.

2.7 Immunoinformatics and Systems Biology

At the level of the antigen, and more specifically the protein antigen, immunogenicity is contingent upon properties of the molecule itself, as well as properties of both the host and the pathogen, be that microbe or cancer cell. It is, therefore, a collective property of the entire system of interacting cells and organisms. The response of the host is mediated by the recognition of T cell and B cell epitopes, as well as the recognition of more mechanistically-generic danger signals. The level of expression of the antigen and its subcellular location within the pathogenic organism are also potentially key arbiters of immunogenicity. The argument runs thus: a poorly expressed, under-represented protein in an inaccessible compartment of a microbial cell is unlikely to be an important antigen, however potent its individual epitopes may be. How such antigens interact with components of the presentation pathway is also important: both viral and bacterial proteins are known to interfere with processes of antigen presentation: some down regulate MHC production, for example, others interfere with peptide transport.

Immunonomics is a newly coined term which subsumes both the theoretical and experimental study of immunology and related disciplines in a post-genomic context. We have already described how the complex process of antigen presentation and subsequent T Cell recognition is beginning to be modeled. Such attempts, while noteworthy, are still floundering due to lack of relevant data. There is still an obvious need for experiments which directly support the development of useful and accurate *in silico* models. Immunoinformaticians need quality data to work from; existing data is seldom satisfactory. Informaticians can no longer exist solely on the crumbs dropped from the experimentalist's table. Instead, there is a clear and palpable requirement for experiments specifically addressing the kind of predictions that immunoinformaticans need to make. Antigen Presentation is being addressed by experiment as well as through the development of theoretical methods. Such experiments are, typically, still operating at the phenomenological level: describing the phenomenon but not dissecting it.

Another aspect of immunogenicity prediction, focuses on system properties of individual gene products from pathogenic micro-organisms. These seek to predict post-translational modifications, subcellular localization, and expression levels. Together these appear to be important factors in the identification of potential antigens. Cell surface proteins, or ones secreted into the extracellular milieu, are more directly open to surveillance by the immune system. Poorly expressed genes are unlikely to be potent antigens because, again, they will not be seen by the immune system. The presence of post-translational modifications can often act as danger signals or are potent immunomodulators either as components of T cell epitopes or through their binding to other receptors. The identification of pathogen proteins which are highly expressed and/or found or outside the pathogen cell and/or contain post-translational modifications is a highly complementary approach to the detection of epitopes, which can be used to select potential antigens with or without knowledge of T cell or B cell responses. An alternative to this approach is to attempt the direct identification of antigens without reference to any mechanistic detail. Here one might endeavor to discriminate between sets of known antigens and sets of known non-antigens or random sets of proteins. While conceptually simple and straightforward, this approach is untested: at present, neither large sets of antigens nor appropriate descriptors are forthcoming.

The effective prediction of protein expression levels in a pathogenic microbe is a potentially important indicator of putative immunogenicity. However, there are inherent difficulties in both the process of prediction itself and in even knowing what an appropriate expression level is. Clearly, under certain conditions, such as starvation, patterns of expression will change dramatically, being up-regulated or down-regulated significantly. Generally, we can assume that the successful surveillance of a microbial protein by the immune system will be linked, in part at least, to its presence in sufficient quantities. There are many ways to predict expression levels but the best studied is codon usage [Karlin & Mrazek 2000]. Different organisms display different codon biases: the preference for one codon rather than another when coding for amino acids. Moreover, there is also a correlation between the choice of which codon is used and the level and rate at which a protein is expressed. The ability to predict different expression levels under different conditions is difficult and requires at least a partial understanding of the whole hierarchy of immune regulation: transcription factors and their binding sites, operons, promotors, mulit-component regulatory networks, etc. To address this pivotal challenge will require the combined ingenuity and imagination of experimentalists, theorists, immunoinformaticans, computer scientists, and mathematicians.

Another important aspect of the prediction of immunogenicity is the accurate identification of Post-Translational Modifications (PTMs). These can take many forms and many potentially contribute to the molecular basis of immunogenicity. PTMs can include glycosylation and lipidation. Glycosylated proteins can be targets for binding by cell surface receptors based on sugar binding leptin domains. Glycosylated epitopes can also be bound by TCRs and antibodies. Lipids can act as epitopes directly through their presentation by CD1. PTMs can also be transitory, such as phosphorylation, or more permanent, such as modified amino acids. Many of these can be part of functional epitopes recognized by the immune system. Glycosylation of a protein, for example, is dependent on the presence of sequence patterns or motifs (Ser/Thr-X-Asn for N-linked glycosylation and Ser/Thr for O-linked) but this is not enough to correctly predict them. If these motifs are present at solvent inaccessible regions of a protein rather on the surface then they will not be glycosylated. Moreover, the other residues which surround these patterns will also affect the specificity of the glyscosylating enzymes: Pro as the X in the Ser/Thr-X-Asn motif for N-linked glycosylation will essentially prevent glycosylation. Glycosylation, in particular, is also very dependent on context, and it is thus a system property of an organism, and can vary considerably in terms of the nature and extent of the different sugars that can become attached to proteins, at least in eukaryotic systems.

Arguably, the most useful, and thus the best studied, of what we might broadly term system approaches to the identification of immunogens, has been the prediction of subcellular location. There are two basic types of prediction method: the manual construction of rules based on knowledge of what determines subcellular location and the application of data-driven machine learning methods, which automatically identify factors determining subcellular location by discriminating between proteins from different known location. Accuracy differs markedly between different methods and different compartments, due to a paucity of data or the inherent complexity that determines protein location. Such methods are often classified according to the input data required and how the prediction rules are constructed. Input data which is used to discriminate between compartments include: the amino acid composition of the whole protein; sequence derived features of the protein, such as hydrophobic regions; the presence of certain specific motifs; or a combination thereof. Phylogenetic profiles can also be used to predict protein location, as the location of close protein homologues can be assumed to be similar.

Signal complexity is a more complicated problem. A very complex signal will require considerable data so that one might be confident in the model. A simple signal, on the other hand, may prove difficult because many, otherwise unrelated, proteins may posses a sorting signal which appears similar, but only by chance. For example, the SWISS-PROT sequence database contains about twice as many non-perioxisomal proteins with a PTS1 sorting signal than real perioxisome-located proteins.

Another challenge is the difference in locations evinced by different organisms. PSORT, a knowledge-based, multi-category program for the prediction of subcellular location, and often regarded as the gold standard for such predictions, is composed of several different programs. Of special interest in this context is PSORT-B, which generates predictions for subcellular location in bacteria. It reports precision values of 96.5% and recall values of 74.8%. PSORT-B is a multi-category method which makes use of six algorithms: SCL-BLAST, which uses protein homology to identify location; PROSITE, which detects motifs; HMMTOP, which predicts membrane proteins; outer membrane β-barrel proteins are identified using specific sequence patterns; SubLocC, is an SVM that uses protein amino acid composition to assign a cytoplasmic or non-cytoplasmic location; and a Hidden Markov Model trained to identify signal peptide cleavage sites. The results of these 6 methods are combined using a Bayesian Network.

Another well known method of interest here is SignalP, which is based on neural networks and predicts N-terminal Spase-I-cleaved secretion signal sequences and their cleavage site. The signal predicted is the type-II signal peptide common to both eukaryotic and prokaryotic organisms, for which there is wealth of data, in terms of both quality and quantity. A recent enhancement of SignalP is a Hidden Markov Model version which is also able to discriminate uncleaved signal anchors from cleaved signal peptides. One of the limitations of SignalP is over-prediction, as it is unable to discriminate between several very similar signal sequences, regularly predicting membrane proteins and lipoproteins as type-II signals. Many other kinds of signal sequence exist. A number of methods have been developed to predict lipoproteins, for example. The prediction of proteins that are translocated via the TAT-dependent pathway is also important but has not been addressed in any depth.

2.8 Immunoinformatics and Vaccinology

Vaccines are molecular entities which can, in effect, mimic infectious organisms so that such microbes can later be recognised and destroyed by the human body or other host, without harm to itself, during subsequent infection. Based on sound, experimental data, immunoinformaticians are using statistical and artificial intelligence methods to identify computationally antigenic proteins and epitopes from pathogenic micro-organisms – bacteria, virus, parasites, or fungi – which the immune system can then recognize, tagging these invading microbes for eventual destruction. However, in order to realise the burgeoning power of these advances still requires much effort.

Vaccines can provide both therapeutic and prophylatic treatments of autoimmune diseases, allergy, and cancer, as well as infectious disease. In light of the many perceived threats to human health, views about infectious disease, in particular, are altering rapidly, leading to a radical reappraisal of the role of vaccines in the fight against pathogenic micro-organisms. Immunovaccinology is the name given to a rational form of vaccinology based very firmly upon our increasing understanding of the fundamental mechanisms which underpin immunology. It must also exploit the potential power of post-genomic technologies. Humanity has sought to address infection through the systematic use of biological and chemical entities: small molecule drugs and supramolecular vaccines. It is now generally accepted that mass vaccination, taking account, as it does, of the principal of herd immunity, is amongst the most effective prophylactic approaches to the treatment, or rather, pretreatment, of infectious disease.

The discovery of vaccination is generally attributed to Edward Jenner, who noted that milkmaids, who had contracted cowpox, a virus related to smallpox, seemed immune to the disease. On 14^{th} May 1796, he used the fluid from a cowpox pustule to build protective immunity against smallpox in James Phipps, an 8 year old boy. Jenner then infected him with smallpox. The boy did not become ill. Later, *Vaccination* - the word Jenner had invented for his treatment (from the Latin *vacca*, a cow) – was adopted for immunization against any disease. In 1980, the World Health Organisation was able to announce the total eradication of smallpox through worldwide vaccination.

However, vaccination has, until relatively recently, been a highly empirical science, relying on poorly understood, non-mechanistic approaches to the development of new vaccines. As a consequence of this, relatively few effective vaccines were developed, and deployed, during most of the two centuries that have elapsed since Jenner's work in the closing years of the Eighteenth century. In the post war years, when antibiotics were king, the threat posed by serious infectious disease, at least in First World countries, seemed to all but vanish, as vaccines and antimicrobial drugs combined to almost eliminate it. The present era is characterized by worries over a variety of burgeoning threats to human well-being: bio-terrorism, climate change, antibiotic resistance, etc. These changes have led, amongst other things, to the re-emergence of diseases such as TB, and exotic emergent diseases, e.g. SARS or avian flu.

A vaccine is a molecular, or super-molecular, agent which elicits specific, protective immunity against pathogenic microbes and the diseases they cause. Protective immunity is an enhanced adaptive immune response to re-infection, as potentiated by immune memory, which, ultimately, mitigates the effect of subsequent infection. Historically, vaccines have been attenuated whole pathogen vaccines such as Sabin's Polio vaccine or BCG for TB. Recently, safety concerns have led to the development of other strategies, focusing separately on subunit/antigen and epitope vaccines (see Figure 2.3). Hepatitis B vaccine is an example of an antigen - or subunit - vaccine, and many epitope-based vaccines have now entered clinical trials. Nevertheless, despite much effort, both publicly and commercially funded, efficacious vaccines are not yet available for many major pathogens such as Shigella, H. pylori, or Meningococcus B.

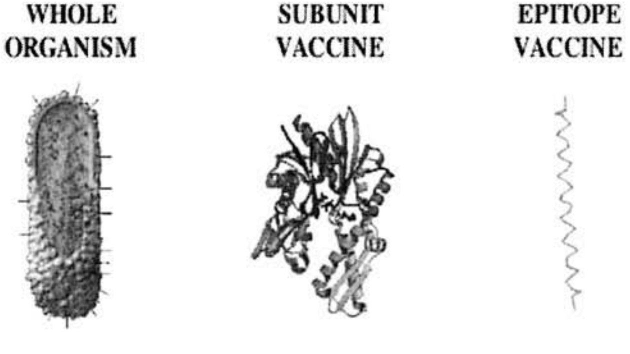

Fig. 2.3. Types of Vaccine: The three main kinds of vaccine component: whole organism attenuated pathogen; subunit whole protein vaccine; polyepitope vaccine. These three are the principal kind of core components of modern vaccines. Epitopes can, potentially, also be carbohydrate or lipid based or a mixture. Modern vaccines are delivered in a variety of ways, such as DNA or as part of a viral vector. Vaccines are also often delivered with adjuvants, molecules which can exacerbate an initial immune response.

Ultimately, the utilitarian value of epitope and immunogenicity prediction will need to be demonstrated through their usefulness in experimental vaccine discovery programmes. All of the methods, we have adduced, focus primarily on the discovery of T cell epitopes, which can prove useful, amongst other things, as diagnostic markers of microbial infection and as the potential basis of epitope vaccines. Many workers have, in recent years, used computational methods as part of their strategy for the identification of both Class I and Class II restricted T cell epitopes. However, it is certainly encouraging that many experimental immunologists are now beginning to see the need for informatics techniques. Computer-based data and knowledge management is essential if this data deluge is not to overwhelm the post-genomic vaccinologist.

There is a clear need to produce more accurate prediction algorithms, which cover more Class I and Class II alleles in more species. Yet, for these improved methodologies to be ultimately effective, i.e. that they are taken up and used routinely by experimental immunologists, these methods must also be tested rigorously for a sufficiently large number of peptides that their accuracy can be shown to work to statistical significance. To do this requires more than new algorithms and software, it requires the confidence of experimentalists to exploit the methodology and to commit laboratory experimentation. Yet most of these tools remain daunting for laboratory-based immunologists. The use of these methods should be routine. It is not only a matter of training and education. These methods must, ultimately, be made more accessible and robust.

2.9 Discussion

From a societal standpoint, immunology is rightly viewed as an important - even a paramount - science. Immunologists are sometimes regarded as a discipline apart. Immunology has a high standing in the wider scientific community: its journals have high impact factors, and it is a large and, generally speaking, a well funded discipline. Immunology is intimately connected with disease: infectious, most obviously, but also autoimmune disease, inherited and multi-factorial genetic disease, cancer, and allergy. Yet, for all its prestige, immunology finds itself at a pivotal point in its history. After more than a century of empirical research, it is on the brink of reinventing itself as a post-genomic science. How will it cope? One obvious way is through embracing computational science.

Immunoinformatics is an amalgam of many different disciplines. Operationally, it has grown from bioinformatics and much of immunoinformatics is ostensibly the application of standard bioinformatic techniques, such as MicroArray analysis or comparative genomics, to the context of immunology. There are, however, several areas which are unique to immunology. Amongst these, the accurate prediction of immunogenicity, be that manifest as the identification of epitopes or the prediction of whole protein antigenicity. It can be fairly described as both the high frontier of immunoinformatic investigation and a grand scientific challenge: it is difficult, yet exciting, and, as a central tool in the drive to develop improved vaccines and diagnostics, is also of true practical value. It requires not only an understanding of immunology but also the integration of many other disciplines, both experimental (physical biochemistry, cell biology, *etc.*) and theoretical (computer science, *etc.*).

We have discussed several distinct areas of immunoinformatic research, yet, there are many others, such as predicting B cell epitopes and adjuvant discovery among them. Immunoinformatics is changing quickly, with many groups trying to improve databases and algorithms. However, despite the steady increase in studies reporting the real-world use of prediction algorithms, there is still an on-going need for truly convincing validations of the underlying approach. Why should this be? As we have seen, predicting T-cell epitopes remains a daunting challenge. We still need to understand the underlying cell biology and model accurately the complexities of the class I and class II antigen presentation pathway. We also still need to understand and accurately model the underlying physical chemistry, in terms of both thermodynamics and kinetics, of peptide binding to MHCs and of TCRs binding to pMHCs.

We have come to a turning point, where a number of technologies have obtained the necessary level of maturity: post-genomic strategies on the one hand and predictive computational methods on the other. Progress will occur in two ways. One will involve closer connections between immunoinformaticians and experimentalists seeking to discover new vaccines. In such a situation, work would progress through a cyclical process of using and refining models and experiments, at each stage moving closer towards a common goal of effective, cost-efficient vaccine development. The other way is the devolved model, where methods are made accessible and used remotely via the web and the GRID.

However, when deprived of direct collaboration, there is still a clear and obvious need for experimental work to be conducted in support of the development of accurate *in silico* methods. Recent work from our laboratory shows the way. Peptides, as reported in literature binding experiments and epitope identification exercises, have heavily biased sequence compositions, resulting from a process of pre-selection which leads to spiraling self-reinforcement. Since only part of a given selection will bind, this rapidly converges to a very limited, and thus incomplete, model of binding dominated by the selection criteria used. These problems would be resolved by a properly designed training set. We have addressed this experimentally, beginning by correlating 90 literature peptide IC_{50}s with cell surface BL_{50} measurements [Doytchinova *et al.* 2004c]. Using models derived from these values, we predicted super-binders with pico-molar affinities much greater than reported values. Using analogues of super binders with modified anchor positions, we then evaluated the relative dominance of anchor positions in a fully systematic manner. Our ability to combine *in vitro* and *in silico* analysis allows us to improve both the scope and power of our predictions in a way that would be impossible using only data from the literature. To ensure we produce useful, quality *in silico* models, rather than worthless and unusable methods, we need to value the predictions generated by immunoinformatics for themselves and conduct experiments appropriately.

The innate and inherent complexity of the immune system is confounding at all levels. Nevertheless, the work of many skilled immunoinformaticains has attempted and nonetheless clearly succeeded in producing useful, if doubtless imperfect, models with true utilitarian value. Progress is, and will continue, being made. We should feel confident that the great synergy arising within this discipline will be of true benefit to Immunology, leading to clear improvements in vaccine candidates, diagnostics, and laboratory reagents. Methods able to predict immunogenicity accurately will become landmark tools for the immunologist and vaccinologist working in the world of tomorrow.

Acknowledgements

We should like to thank the following for the help. Firstly, members of the bioinformatics group at the Jenner Institute, University of Oxford, both past and present. Of these, we should like to single out Dr Irini Doytchinova for special mention, as her contribution to the group has been without equal. The others include, in no particular order: Valerie Walshe, Kelly Paine, Christopher P Toseland, Shelley L Hemslay, Christianna Zygouri, Dr Pingping Guan, Dr Helen McSparron, Dr Matt Davies, and Dr Channa Hattotuwagama, Dr Shunzhou Wan, Martin Blythe, and Debra Clayton (nee Taylor). We should also like to thank the following for their help and stimulating discussions: Prof Peter Beverley, Prof Vladimir Brusic, Dr Persephone Borrow, Dr Shirley Ellis, Dr Kevin Rigley, Dr Simon Wong, Dr Helen Bodmer, Dr Sam Hou, Dr Elma Tchillian, and Dr Josef Walker. Finally, we should also like to thank all our other colleagues and co-workers at the Edward Jenner Institute for Vaccine Research and the Institute for Animal Health, Compton for their close, nurturing, and supportive collaboration.