Are bacterial vaccine antigens T-cell epitope depleted?

Mark Halling-Brown¹, Clare E. Sansom¹, Matthew Davies², Richard W. Titball³ and David S. Moss¹

¹School of Crystallography, Birkbeck College, London WC1E 7HX, UK
²The Jenner Institute, University of Oxford, Compton, Newbury, Berkshire RG20 7NN, UK
³School of Biosciences, University of Exeter, Exeter EX4 4QD, UK

For many infectious diseases, protective immunity can be elicited by vaccination with pathogen-derived proteins. Peptides derived from these proteins are bound to major histocompatibility complex (MHC) molecules and presented to T-cell receptors to stimulate an immune response. We show here that, paradoxically, bacterial proteins known experimentally to elicit a protective immune response are relatively depleted in peptides predicted to bind to human MHC alleles. We propose three nonconflicting reasons for this: the lack of precision of current predictive software, the low incidence of hydrophobic residues in vaccine antigens or evolutionary pressure exerted on bacteria by the immune system. We suggest that there is little value in predicting candidate vaccines based on high MHC-binding epitope density.

Introduction

The recognition of a pathogen by the human adaptive immune system is accompanied by a series of responses that may lead to T-cell activation and pathogen elimination [1, 2]. The first stage of the T-cell response to microbes involves the processing of microbial proteins into peptides that are presented to the TCR (T-cell receptor) for recognition. Two separate pathways are involved in this process. In the MHC (major histocompatibility complex) class I pathway (Figure 1a), microbial proteins in the cytosol or the nuclear compartment of the infected cell are cleaved by cytosolic proteases such as proteasomes. The resulting peptides are transported into the ER (endoplasmic reticulum) by a TAP (transporter associated with antigen processing). In the ER, the peptides bind to MHC class I molecules, and the resulting protein complex can be presented at the cell surface for possible interaction with the TCR of CD8⁺ T cells. The resulting T-cell response can lead to destruction of the infected cell.

In the MHC class II pathway (Figure 1b), microbial proteins in vesicular compartments are cleaved into peptide fragments, some of which bind to MHC class II molecules and are transported to the cell surface. The MHC class II–peptide complexes are recognized by helper CD4⁺ T cells, which enable an adaptive immune response to the pathogen. It is well documented that exposure to microbial pathogens, including bacteria, can trigger both MHC class I and class II immune responses [3].

Reverse vaccinology, defined by Rappuoli [4] as ‘studying vaccine development in silico… by using genomic information’, is an increasingly popular strategy in vaccine design. Because of the central role that T-cell epitopes play in protective immunity, there is considerable interest in identifying such epitopes in microbial proteins for inclusion in subunit vaccines. MHC binding is a prerequisite for the development of a T-cell response, and many algorithms have been developed for predicting MHC-binding peptides from protein sequences [5–7]. Structure-based methods, which may be more useful for rarer alleles, are beginning to be developed, although recent work suggests that peptide binding might be driven entropically [8]. These algorithms are based on experimentally determined MHC–peptide affinity binding data that has been obtained using a variety of techniques. Generally, programs for predicting peptide binding to MHC class I molecules are more accurate than those for predicting peptide binding to MHC class II (Box 1; Table 1; Ref. [9]). Programs have also been developed for predicting protein cleavage by the proteasome (e.g. NetChop [10]) and TAP transport [11] in the class I pathway within the antigen presenting cell. Thus far, these programs have been relatively rarely used for vaccine design. Prediction of protein cleavage in the MHC class II pathway is more challenging because several proteases are involved, and their specificities are not well understood [12].

An alternative in silico strategy involves the identification of proteins with properties that have been associated with immunogenicity [13]. The most common criterion used in this approach has been to predict the subcellular location of proteins. This is based on the premise that microbial proteins on the cell surface or secreted are most likely to be vaccine candidates [4]. However, because of their exposure to the immune system this yields a vast pool of potential vaccine candidates to choose from, which is impracticable for in vivo trials.

No single predictive step can identify proteins that are able to induce protective immunity, which could thus provide the basis of potential vaccines. There is therefore great interest in combining the techniques developed thus far to provide an effective, rational approach to vaccine design. We have investigated the density of predicted MHC-binding peptides in known vaccine antigens; our results suggest...
that, paradoxically, proteins predicted to contain a large number of such peptides are less likely than random proteins to elicit an immune response. This has important implications for computer-aided vaccine design [14].

**Vaccine antigens are depleted in predicted T-cell epitopes**

We compared a dataset of 70 proteins experimentally known to confer immune protection against bacterial pathogens in at least one animal model of human disease [15] (the vaccine antigen dataset) against a control dataset of randomly selected sequences derived from the predicted proteomes of pathogenic bacteria. All of the sequences in these datasets were annotated with experimentally determined subcellular locations (Table 2). The majority of the vaccine antigens were either secreted from the bacterium (23 of 70; 33%) or located on the outer membrane or cell wall (39 of 70; 56%). This confirms the validity of the
common assumption that the most immunogenic proteins will be those that are most quickly and readily visible to the host immune system.

We scored all nonamer peptides in each protein in each dataset for predicted class II MHC binding using two well-characterised algorithms: TEPITOPE [16,17] and MHCPred [18] (Box 1; Table 1). TEPITOPE has been used successfully to predict MHC II–restricted T-cell epitopes in Mycobacterium tuberculosis proteins [19] and tumour-associated antigens [20], and MHCPred has been used in the prediction of therapeutic vaccines for the treatment of cancer [21].

Using both TEPITOPE and MHCPred, our work showed that the proteins in the control dataset scored significantly more for predicted class II epitopes than those in the vaccine antigen dataset (R. Shaban, PhD thesis, University of London, London, UK, 2008; Table 2). For comparison, Table 2 also includes archaeabacteria that have probably never been encountered by the vertebrate immune system (e.g. bacteria from hydrothermal vents). These too gave class II epitope scores that were higher than that of the vaccine antigen dataset.

Therefore, the vaccine antigen dataset differs from the control dataset of randomly selected bacterial proteins in possessing a lower density of nonamers that bind to MHC class II alleles [15]. Furthermore, the vaccine antigen dataset is not atypical population of bacterial proteins per se; rather, their T-cell epitope content is typical of the secreted and outer membrane proteins, groups of proteins to which most of the this dataset belong to. Therefore, as groups, the secreted and outer membrane proteins are also predicted to contain a reduced number of MHC class II epitopes. Importantly, the overall pattern of predicted T-cell epitopes in the vaccine antigen dataset or in proteins of different subcellular locations did not depend significantly on the class II allele or alleles used in the prediction algorithm.

A similar analysis of the vaccine antigen and control datasets was carried out using well-characterised MHC class I–binding algorithms, including SYFPEITHI [22], NHLApred [23] and BIMAS [24] (Box 1; Table 1). Our analysis indicated that vaccine antigens were also significantly deficient in MHC class I epitopes compared with the control dataset (Table 2). This conclusion was also independent of the MHC class I allele used for epitope prediction.

Overall, these findings indicate that secreted and outer membrane proteins of bacterial pathogens, including
vaccine antigens, are relatively depleted in predicted MHC class I and/or MHC class II binding epitopes.

**Why are immunogenic vaccine targets poor MHC binders?**

**Role of software artefacts**

It is important to realize that we have shown antigenic proteins to be depleted in sequences that are predicted to bind to MHC molecules rather than in those that have been experimentally determined to do so. Empirically based prediction algorithms are variable in quality, and none are wholly precise. MHC class I prediction programs can predict binding to a range of common alleles—as opposed to predicting the T-cell response—with fair accuracy [14], but class II programs cannot yet be thought of as wholly precise. MHC class I prediction programs can bind to MHC molecules rather than in those that have been experimentally determined to do so. Empirically based prediction algorithms are variable in quality, and none are wholly precise. MHC class I prediction programs can predict binding to a range of common alleles—as opposed to predicting the T-cell response—with fair accuracy [14], but class II programs cannot yet be thought of as wholly reliable [25]. It is therefore possible that our results can be at least partly explained by software artefacts. These could arise from an overall lack of precision, the large number of rarer alleles [5] or the paucity of binding data for class II molecules. Furthermore, the relationship between peptide–MHC binding and T-cell response is poorly understood, so it is not possible to predict which known MHC-binding peptides will elicit such a response. However, the high statistical significance of the results, and the fact that consistent results were obtained using a range of prediction algorithms using different algorithms suggests that software artefacts are unlikely to be the most important cause of this phenomenon.

**Role of amino acid composition**

Part of the reason why these proteins are deficient in predicted T-cell epitopes might lie in the role that hydrophobic residues play in MHC class II–peptide binding. X-ray structural analysis and binding studies of MHC class II–peptide complexes have shown that these peptides will almost invariably possess one of seven hydrophobic amino acids (Phe, Ile, Leu, Met, Val, Trp or Tyr; FILMVWY) bound in pocket 1 of the MHC active site [17,26–28]. MHC class I molecules, by contrast, have ‘anchor’ residues at positions P2 and P9 that favour the binding of medium-sized hydrophobic residues (ILMV). We have discovered a strong positive correlation between the FILMVWY content of proteins and their CD4+ T-cell epitope scores, consistent over both datasets and all subcellular locations (see Table 2). We also found that vaccine antigen proteins and secreted and outer membrane proteins from the control dataset contain a significantly lower proportion of these residues than cytosolic or inner membrane proteins (Table 2). Amino acid composition is one of the determinants of subcellular location, partly because of the different environments in which the proteins are localized [29]. The composition of each subset in our control database did not differ significantly from that of similar subsets generated by another study [30] based on all gram-negative bacterial sequence entries in the SwissProt [31] database. Thus, the vaccine antigen dataset [15] differs from control proteins in possessing both fewer FILMVWY amino acids and fewer nonamers predicted to bind MHC class I and class II alleles, although, because most epitopes are amphipathic (both hydrophobic and hydrophilic), the reduction in hydrophobic residues cannot be the sole reason for the lack of observed epitopes within this dataset.

**Role of evolutionary pressures**

Over the millennia, pathogens have evolved using a range of mechanisms to avoid the adaptive immune responses of their vertebrate hosts. We describe here other mechanisms through which bacteria might have evolved to evade the adaptive immune system: by mutating their peptide sequences so that they either do not bind to MHC molecules or alternatively bind to them and produce a T-cell response protective to the bacterium (e.g. Th2 or regulat-
ory T cell). The validity of the former hypothesis might be tested in several ways. For example, microorganisms that could never have encountered the mammalian adaptive immune system would not be expected to show depletion of T-cell epitopes (e.g. free-living bacteria). Indeed, Table 2 shows that archaeabacteria are rich in T-cell epitopes. Similarly, bacteria that have only recently emerged as mammalian pathogens (e.g. zoonoses) might show a less marked depletion of T-cell epitopes than those with a long-established history of causing disease. In a wide-ranging review, Finlay and McFadden [32] describe many ways in which bacteria interact with or avoid the immune system to survive within the host. Among these are mechanisms for blocking antigen presentation by MHC molecules, antigenic variation and molecular mimicry such that the host is unable to mount a timely response to the pathogen [33,34]. The depletion of epitopes in surface located or secreted proteins might be a consequence of the evolutionary pressure for bacteria to eliminate sequences that elicit adaptive immune responses. Host and parasite species co-evolve so that each undergoes genetic changes in response to the other; this is one variant of the so-called ‘Red Queen’s Race’ hypothesis [35]. It might be assumed that the pathogens would have an advantage in this race because of a much shorter life cycle than the host. However, the key to the vertebrate immune system is that it is adaptable to a variety of pathogens through somatic selection from a vast and degenerate repertoire of T cells and B cells. Co-evolution on the part of the mammalian host is thought to explain much of the observed hyper-polymorphism of the MHC [36]. In pathogens, this phenomenon has been best studied with rapidly evolving viruses. For instance, studies have shown that, in the presence of the MHC class I allele HLA-A3, HIV-1 will generate epitope mutations that impair the immune response generated by that allele, and a similar phenomenon can occur during chronic infection with HCV (hepatitis C virus) [36,37]. Therefore, bacterial proteins that are preferentially exposed to the mammalian immune system might have evolved sequences with fewer protease cleavage sites or MHC-binding motifs. Clearly the deletion of potential epitopes from the pathogen proteome must be balanced by the need for it to maintain the structural and functional integrity of the protein. However, evolutionary pressures on bacteria to evade immune surveillance might lead to modulation of the immune response rather than elimination of the pathogen. For example, we speculate that peptides from a bacterium that is resident in macrophages may have sequences that stimulate T-cell production of interleukin-10, which tends to deactivate macrophages [38].

Implications for future research and vaccine design

T-cell epitope prediction has been suggested as a valuable step in the prediction of vaccine antigens [11,39,40]. However, at least at present, prediction of MHC-binding peptides is very commonly used as a surrogate for the much more complex prediction of T-cell response. The amino acids that interact with the TCR are those that are not involved in the MHC binding because they face away from the MHC-binding groove rather than toward it. They are also predominantly hydrophilic in nature. Unfortunately, the variable nature of the complementary determining (CDR) loops on the TCR that interact with the peptide amino acids make it very difficult to predict which peptides will bind strongly to the TCR.

Our findings indicate that an epitope mapping strategy that identifies those proteins within a proteome that contain the maximum number of MHC-binding segments might not add significant value to other steps involved in in silico vaccine design, including the crucial but simple one of predicting protein subcellular location [41]. Selecting proteins for inclusion in vaccines that possess multiple epitopes might be important to ensure that protective immune responses develop across a genetically diverse human or animal population. However, we now suggest that bacterial proteins that generate protective immunity might do so, not through a high density of MHC-binding epitopes, but through a limited number of epitopes with optimal properties associated with an efficient T-cell response. These might include, for example, the strength of peptide binding to the MHC molecule, the ability of a peptide to activate a T cell or the cell surface half-life of the MHC–peptide complex. Clearly, additional work is needed to investigate the nature of the MHC-binding epitopes that are both deleted and retained by surface exposed and secreted proteins before MHC binding prediction can completely fulfill its obvious potential as an important step in modeling the T-cell response and in reverse vaccinology. We may also need to be predicting strong MHC–peptide binding. The required strength might depend on the concentration of the relevant pathogen-derived protein in the cell. Without a dominant binder, having many relevant binding peptides might not result in a successful T-cell stimulation because of competition between binders and with self-peptides and not yield a critical concentration of any single MHC–peptide complex on the cell surface. When some of these issues are more clearly understood, in silico prediction methods could become a more effective and commonplace component of vaccine design.

Conflict of Interest Statement
The authors declare no competing financial interests.

Acknowledgements
M.H.B. was funded through the European Commission ImmunoGrid project (http://www.immunogrid.org), STREP project i/d FP6-2004-IST-028669.

References
6 Rajapakse, M. et al. (2007) Predicting peptides binding to MHC class II molecules using multi-objective evolutionary algorithms. BMC Bioinformatics 22, 459