BIOINFORMATION
Discovery at the interface of physical and biological sciences

open access www.bioinformation.net Hypothesis

Volume 8(6)

Immunoinformatic evaluation of multiple epitope ensembles as vaccine candidates: E coli 536

Jade Rai, Ka In Lok, Chun Yin Mok, Harvinder Mann, Mohammed Noor, Pritesh Patel, & Darren R Flower*

Aston Pharmacy School, Life and Health Sciences, Aston University, Aston Triangle, Birmingham, B4 7ET, UK; Darren R Flower - E-mail: D.R.Flower@aston.ac.uk; Phone+44 (0)121 204 5182; * Corresponding author

Received March 27, 2012; Accepted March 28, 2012; Published March 31, 2012

Abstract:
Epitope prediction is becoming a key tool for vaccine discovery. Prospective analysis of bacterial and viral genomes can identify antigenic epitopes encoded within individual genes that may act as effective vaccines against specific pathogens. Since B-cell epitope prediction remains unreliable, we concentrate on T-cell epitopes, peptides which bind with high affinity to Major Histacompatibility Complexes (MHC). In this report, we evaluate the veracity of identified T-cell epitope ensembles, as generated by a cascade of predictive algorithms (SignalP, Vaxijen, MHCPred, IDEB, EpiJen), as a candidate vaccine against the model pathogen uropathogenic gram negative bacteria *Escherichia coli* (E-coli) strain 536 (O6:K15:H31). An immunoinformatic approach was used to identify 23 epitopes within the E-coli proteome. These epitopes constitute the most promiscuous antigenic sequences that bind across more than one HLA allele with high affinity (IC50 < 50nM). The reliability of software programmes used, polymorphic nature of genes encoding MHC and what this means for population coverage of this potential vaccine are discussed.

Background:
Human immunity comprises the innate and the adaptive immune response. Vaccination is principally concerned with humoral and cell mediated adaptive immune responses **[1].** Vaccination generates effective and appropriate adaptive immune responses, to be activated subsequent to immunisation. The overall goal is the generation of long lasting immunity against microbial pathogens via the production of diverse immune memory cells. The Major Histacompatibility Complex (MHC) is a large region encoded on human chromosome 6; genes of the MHC are also known as Human Leukocyte Antigens (HLA), and demonstrate the highest levels of sequence polymorphism within the human population **[2].** This genetic diversity is responsible for the adaptability of the human immune system to microorganisms and pathogens.

T lymphocytes are integral to the adaptive, cell-mediated immune response against foreign substances, falling into two broad groups: T-cells (CD4) and cytotoxic T-cells (CD8), depending on the co-receptors they express. CD4 cells associate with Class I MHC and CD8 cells bind to class II MHC. In order for an effective immune response to occur, various threshold need to be surpassed **[1].** Without effective processing of antigens by intracellular pathways and efficient presentation on the surface of Antigen Presenting Cells (APCs), then effective epitope-MHC-T-cell complexes cannot form. Effective T-Cell epitopes possess several key characteristics: the ability to be bound by MHCI and MHCII with high specificity and affinity; to form stable peptide-MHC complex that can bind with high affinity to T-cell receptors and subsequently activate T cells; antigenic peptides must be cleaved by the proteasome; for class I antigen presentation potential epitopes must be successfully transported by TAP; must be successfully secreted and presented on the cellular surface of Antigen Presenting Cells (APC); and to bind many MHCI and MHCII alleles with high affinity to provide wide population coverage. Epitopes which bind with high affinity to several MHC alleles tend to be stronger candidates for use within epitope-based vaccines. Currently, however, there is no real consensus as to what constitutes an optimal smallest set of 'promiscuous' epitope's for inclusion in a vaccine. Within the context of reverse vaccinology, computer software enables rapid processing of

genomic data for vaccine discovery **[1].** Gene sequences are identified as open reading frames, epitopes predicted within them, and filtered from several thousand epitopes to a few dozen candidate immunogens. Such immunogens then undergo experimental validation *in vitro*, *in vivo*, and in challenge models, before entering clinical trials. Current immunoinformatic software focuses on multi-step approaches for T-cell epitope identification, including: proteasome cleavage, TAP transport and MHC binding. These methods have the advantages of higher accuracy and a lower rate of false positive predictions, greatly reducing the number of peptides requiring testing.

Computationally-driven genomic approaches to vaccine discovery remain challenging due to the high genetic diversity of pathogens. Gram-VE bacterium *Escherichia coli* are usually harmless, and are an important component of the gut microbiome; however, certain strains do cause disease and can be fatal. Specifically, uropathogenic Escherichia coli (UPEC) infection is a prevalent disease with potentially severe complications **[3].** Globally, UPEC accounts for over 4 out of 5 urinary tract infections. In the urinary tract, UPEC colonizes the bladder preferentially causing cystitis, and also infects the kidneys, causing pyelonephritis. Isolated initially from a patient with a urinary tract infection, the uropathogenic Escherichia coli strain 536 (O6:K15:H31), is now among the best-understood model pathogens; the availability of the O536 genome facilitates the computational identification of vaccine components. In this work, we describe the use of a novel immunoinformatic protocol to identify an optimal epitope ensemble as a putative candidate vaccine against uropathogenic *E coli*; we then explore the limitations of such strategies, highlighting the pressing need for experimental validation of this and other similar studies.

Methodology:

Rationale

Promiscuous high-affinity epitope candidates suitable for inclusion in a vaccine were sought using a novel epitope prediction pipeline implemented using web-accessible prediction servers, as described below. Each step in the pipeline generates data requiring filtering. The process of re-scaling has its own problems **[4].** A balance needs to be struck between conserving variability across epitope sequences and filtering out sub-optimal epitopes. Having a common measure describing the strength of epitope binding to MHC is important. It allows a direct comparison to be made between data obtained from different prediction servers.

Pipeline Structure

The structure of the pipeline is given in **(Figure 1)**. SignalP **[5]** predicts which proteins from a virtual proteome files will be secreted via the Type II secretion pathway. Proteins predicted to be secreted by the SignalP server were filtered to isolate the top 5 scoring antigenic proteins using VaxiJen **[6].** Class I and Class II peptide-MHC binding afinity for these proteins was then predicted using 3 reliable servers. MHCPred **[7]** (MHC Class 1 and 2) is an implementation of the additive method. CRB **[8]** predicts binding to 57 Human class I and 26 class II MHC alleles. IEDB **[9]** again predicts binding to a wide range of human MHC alleles. All three servers cover a wide range of human MHC alleles, all of which exist at high frequency in the global human population, and are based on a large collection of literature binding data. EpiJen predicts T-cell epitopes using a combination of proteasome cleavage, TAP transport, and MHC binding **[10].**

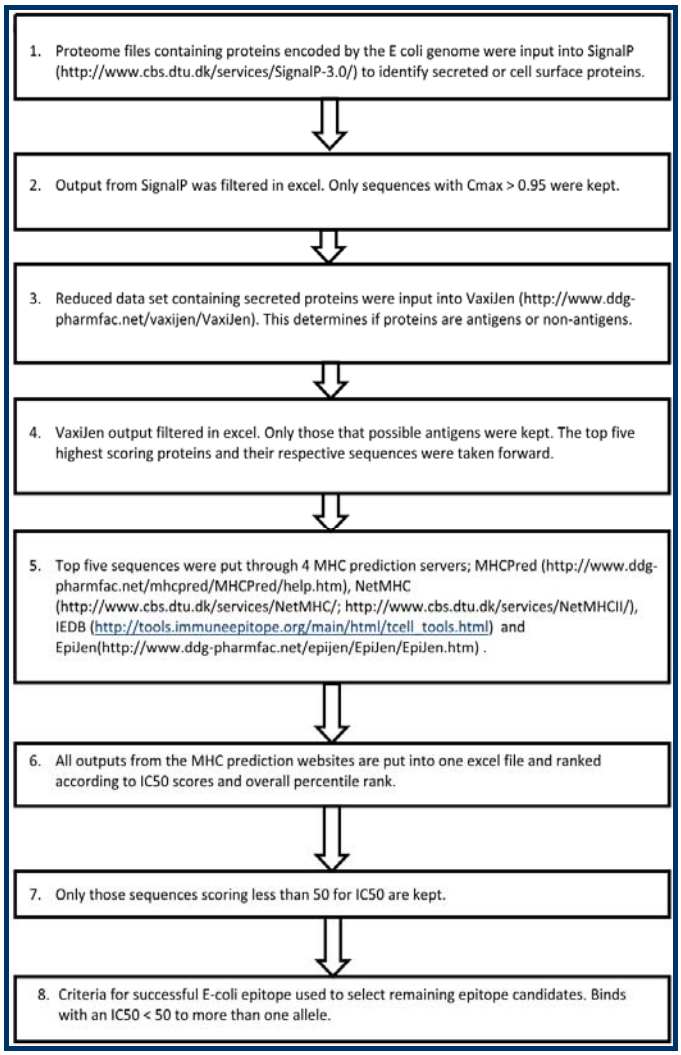

Figure 1: Epitope Prediction Pipeline, Flowchart summarising the epitope prediction pipeline described in methods.

Operation

The pipeline described was in part run in serial mode (SignalP and VaxiJen) and in parallel mode (MHCPred, IEDB, and CRB). Epijen was run as a control, and its outputs compared to those generated by MHCPred, IEDB, and CRB. For all data harvested from web-servers, extensive off-line integration and filtering was used subsequent to data gathering.

Exegesis

Only proteins which are secreted or expressed on the cell surface proteins are accessible to immune surveillance, as identified here by SignalP, and are thus of interest. Only a subset of these, antigenic proteins as identified by VaxiJen, are likely to possess immunodominant epitopes. Such epitopes are, as predicted by MHCPred, IEDB, CRB, and EpiJen, those binding promiscuously at high affinity to many MHC alleles. From the final ranked lists of epitopes generated by MHCPred, IEDB, CRB, and EpiJen, a set of epitopes showing high affinity across servers and across multiple alleles was selected using a greedy breadth-first search protocol.

Discussion:

We have described the identification of a multiple epitope ensemble as a candidate peptide vaccine. The virtual E coli proteome was assessed for potential epitopes using a pipeline composed of several servers: SignalP **[5]** (likelihood of protein secretion) followed by VaxiJen **[6]** (ranked antigenicity of secreted proteins) followed in parallel by MHCPred **[7],** IEDB **[8],** and NetMHC **[9]** (binding affinities of epitopes). EpiJen **[10]** was used as a control to assess the likelihood of epitope processing. The rationale for this pipeline is straightforward: only secreted proteins or those found on the cell surface, are accessible to immune surveillance, and of these, only a subset is likely to be antigenic, and are thus also likely to possess immunodominant epitopes **[1].** The main problem in filtering strong epitopes from weak epitopes is defining a suitable threshold. A balance must be maintained between conserving successful epitopes and keeping the number of epitopes manageable for subsequent *in vitro* testing stage. Thus, only epitope sequences that occurred in two or more server outputs were used. This utilised the strategy implicit in all meta-servers that consensus prediction is inherently more robust and reliable. Only very high affinity epitopes with an IC50 value less than or equal to 50 were used. Epitopes also had to bind to more than one class I and class II allele.

Using a breadth-first search strategy, the following 23 epitope sequences were identified as the smallest set of most promiscuous epitopes: ALAGVVPQY (active versus HLA0101, HLA0201); KLATLFLTA (HLA0101, HLA0201, HLA0203), MLSDAAPEV (HLA0201, HLA0202, HLA0203, HLA0206), SLISGSFLL (HLA0201, HLA0202, HLA0203, HLA0206), VMLNFKKTF (HLA0201, HLA1101), TLAAKDINV (HLA0202, HLA0203), LLSACIALA (HLA0202, HLA0203), ALSGDNNSV (HLA0202, HLA0203), NLNGGIQFV (HLA0202, HLA0203), TLSNSSSAV (HLA0202, HLA0203), FLSSSGGSA (HLA0202, HLA0203), ALITTLAIV (HLA0203, HLA0206), GIFAISALA (HLA0203, HLA1101, HLA6801), IVFSGSALA (HLA0203, HLA0206, HLA6801, HLA6802), FSLVTNEGI (HLA0203, HLADRB0701), FAISALAAT (HLA0206, HLA6801, HLADRB0101), AVLLLALLT (HLA0206, HLADRB0401), FNSLATMGV (HLA0206, HLADRB0701), KAVLLLALL (HLA6801, HLADRB0401), ISLAGSMKV (HLA6801, HLADRB0101, HLADRB0701), LSYVKSQRT (HLADRB0101, HLADRB0701), FNNSVSSMM (HLADRB0101, HLADRB0701), VQLVNSGTI (HLADRB0401, HLADRB0701).

The current work shows strongly similarity to that of Wieser *et al*. [11] They identified computationally immunodominant epitopes from six virulence-associated E coli antigens: Usp, Iha, FyuA, ChuA, IreA, and IroN; creating two wholly artificial genes were created, each encoding eight extended peptide epitopes, which when expressed recombinantly resulted in a vaccine active against pathogenic but not benign E. coli in the gut. The important difference here was that this study, led by experimentalists, was validated experimentally **[11].**

There are clear deficiencies in this and all other such studies. The poor predictivity of available software limits the ultimate reliability of all such studies. Servers facilitate vaccine development by assessing epitope binding affinity of different MHC alleles. Currently, the most reliable algorithms revolve around MHC I presentation with a heavy bias towards so-called HLA 'supertypes' **[12].** There are 9 major supertypes showing commonality of peptide anchor binding specificity. Predictions for MHC class II binding by T cell epitopes, is by contrast poor and unreliable. Overall, the predictivity of software mirrors the degree to which particular HLA alleles have been studied. A strict consensus surrounding what a promiscuous epitope is and what features might make such an epitope suitable for inclusion in a vaccine remain largely undefined. If incorrectly defined it could lead to strong potential epitopes being missed out. The consequences including a longer development process with increased cost and enhanced chances of a poorly efficacious vaccine. Extending on the above discussion, size and variability of the immune response to pathogens and vaccines alike probably precludes the identification of very small sets of epitopes, as is common in published studies of this kind, since they will not evoke responses of sufficient strength and duration in a large enough section of the population to induce either immune memory or the necessary herd immunity essential for all effective vaccines.

Likewise, failure properly to account for sequence similarity to other bacteria and to human proteome imposes severe constraints on the ultimate utility of any such study. We need to correct for any clear sequence similarity to other components of the gut microbiome, since vaccines also active against other, useful commensal bacteria would be of little utility. The problem here is that microbiome is large, complex, and variable on a local and global basis. Comparison with metagenomic analyses of the microbiome might prove the answer in this regard. Nor would it be good to induce autoimmunity against components of the human proteome that bare sequence similarity to an epitope ensemble. Unfortunately, proven and robust protocols able to identify such similarity are currently lacking.

Conclusion:

For all or nearly all in computational studies addressing real-world problems, there is a pressing need for experimental validation. An increasing number of papers describing *in silico* analyses of genomes and proteomes, producing epitope ensembles as putative candidate vaccines, are now being published **[13-19].** Most are sound, and, like the present study, even rigorous; yet their value cannot easily be quantified. As a consequence, their significance is questionable. Other studies **[11, 20]** by contrast combine immunoinformatic-driven vaccine design with experimental validation in various animal models, giving credence to their computational results. It has been said that theoreticians cannot "exist solely on morsels swept contemptuously from the experimentalists' table" **[1],** but equally prediction without validation will exert little influence and convince no one. Continuing to publish experimentally unverified papers is almost counterproductive in this context. Moreover, current research methodology is largely embodied in web-servers; operating such virtualised systems is facile, and the concomitant analysis of results straightforward. As the current paper amply demonstrates, the technical strictures inherent in the *in silico* design of epitope ensembles as candidate vaccines are quite within the ambit of undergraduate researchers. In sum, we can say that we have at once generated a sound prediction of an epitope ensemble vaccine and at the same time sought to highlight the severe limitations of all such studies in the absence of proper experimental validation.

References:

[1] Flower DR, *Bioinformatics for vaccinology*. 2008 p82-94, p168-173.

[2] Lafuente EM & Reche PA, *Curr Pharm Des.* 2009 **15**: 3209 [PMID: 19860671]

[3] Sivick KE & Mobley HL, *Infect Immun*. 2010 **78**: 568 [PMID: 19917708]

[4] MacNamara A *et al. PLoS Comput Biol.* 2009 **5**: e1000327 [PMID: 19300484]

[5] Bendtsen JD *et al. J Mol Biol*. 2004 **340**: 783 [PMID: 15223320]

[6] Doytchinova IA & Flower DR, *BMC Bioinformatics.* 2007 **8**: 4 [PMID: 17207271]

[7] Guan P *et al. Nucleic Acids Res.* 2003 **31**: 3621 [PMID: 12824380]

[8] Lund O *et al. Immunological Bioinformatics*. Massachusetts: The MIT Press. 2005 p111

[9] Greenbaum J *et al. Nucleic Acids Res.* 2008 **36**: W513 [PMID: 18515843]

[10] Doytchinova IA *et al. BMC Bioinformatics* 2006 **7**: 131 [PMID: 16533401]

[11] Wieser A *et al. Infect Immun*. 2010 **78**: 3432 [PMID: 20498257]

[12] Sidney J *et al. BMC Immunol.* 2008 **9**: 1 [PMID: 18211710]

[13] Akhoon BA *et al.* Microb *Pathog*. 2011 **51**: 77 [PMID: 21349321]

[14] Chandra S *et al. Bioinformation*. 2010 **5**: 155. [PMID: 21364778]

[15] Gupta A *et al. Bioinformation* 2011 **5**: 386 [PMID: 21383906]

[16] Sinha S *et al. Bioinformation*. 2011 **5**: 320 [PMID: 21383918]

[17] Jahangiri A *et al.* Vaccine. 2011 **29**: 6948 [PMID: 21791233]

[18] Barh D *et al. Bioinformation.* 2010 **5**: 77 [PMID: 21346868]

[19] John L *et al. Appl Biochem Biotechnol.* 2012 [Epub ahead of print] [PMID: 22434357]

[20] Seyed N *et al. PLoS Negl Trop Dis*. 2011 **5**: e1295 [PMID: 21909442]

Edited by P Kangueane

Citation: Rai *et al.* Bioinformation 8(6): 272-275 (2012)