Bioinformation
www.bioinformation.net

open access
Hypothesis

Discriminating antigen and non-antigen using proteome dissimilarity: bacterial antigens

Kamna Ramakrishnan[1], Darren R Flower [2,*]

The Jenner Institute, University of Oxford, Compton, Newbury, Berkshire, United Kingdom, RG20 7NN; [1]Medical Genetics Section, University of Edinburgh, Edinburgh, United Kingdom. EH4 2XU; [2]Life and Health Sciences, Aston University, Aston Triangle, Birmingham, United Kingdom. B5 7ET, Darren Flower- Email: D.R.Flower@aston.ac.uk. Ph no: +44 (0)121 204 5182 * Corresponding author

Received March 19, 2010; accepted April 14, 2010; published April 30, 2010

Abstract:

It has been postulated that immunogenicity results from the overall dissimilarity of pathogenic proteins versus the host proteome. We have sought to use this concept to discriminate between antigens and non-antigens of bacterial origin. Sets of 100 known antigenic and non-antigenic peptide sequences from bacteria were compared to human and mouse proteomes. Both antigenic and non-antigenic sequences lacked human or mouse homologues. Observed distributions were compared using the non-parametric Mann-Whitney test. The statistical null hypothesis was accepted, indicating that antigen and non-antigens did not differ significantly. Likewise, we were unable to determine a threshold able to separate meaningfully antigen from non-antigen. Thus, antigens cannot be predicted from bacterial genomes based solely on their dissimilarity to the human genome.

Background:

It is a verity now universally acknowledged that population-level vaccination is the most effective preventative measure discovered thus far for the control of infectious disease and for subsequently mitigating the effects of re-infection. A vaccine is a molecular or supramolecular agent able to engender specific protective immunity, an enhanced adaptive immune response to re-infection by pathogens. Immunity is characteristic of proper immune system functioning and manifests itself in the ability of the host to tolerate endogenous, somatic substances and to eliminate exogenous, foreign material. This discrimination defends us against diverse infectious diseases, since microbial products are readily seen by the immune system as alien products. Protection results in the potentiation of host survival and the destruction of pathogens **[1]**.

Vaccines themselves come in various guises. Some may be living but weakened - or attenuated - strains of micro-organisms that have been cultured under conditions which reduce their capacity to cause disease and thus give rise only to mild or undetectable infections; and include those acting against measles, rubella, yellow fever, mumps, and tuberculosis. Such vaccines reduce the intrinsic virulence exhibited by a virulent microorganism, typically by altering their growth conditions, yet leave their immunogenic properties largely unaffected. Vaccines may also be killed or inactivated organisms treated with heat or chemicals; examples include vaccines active against Influenza, cholera, bubonic plague, polio, hepatitis A, and Rabies. So-called subunit vaccines are a third form of vaccine. They comprise individual pathogenic proteins, and examples include vaccines against HBV, Human Papillomavirus, and *Haemophilus influenzae* B. Subunit vaccines, consisting of highly immunogenic carbohydrate, such as cell wall components; protein; or glyco-protein conjugates, stimulate measurable yet often quite weak immune responses, necessitating the use of adjuvants to raise initial levels of immunogenicity and also complex vaccination regimes to sustain enduring protection. Yet despite such drawbacks, subunit vaccines remain a popular objective and are thus a current focus for vaccine discovery.

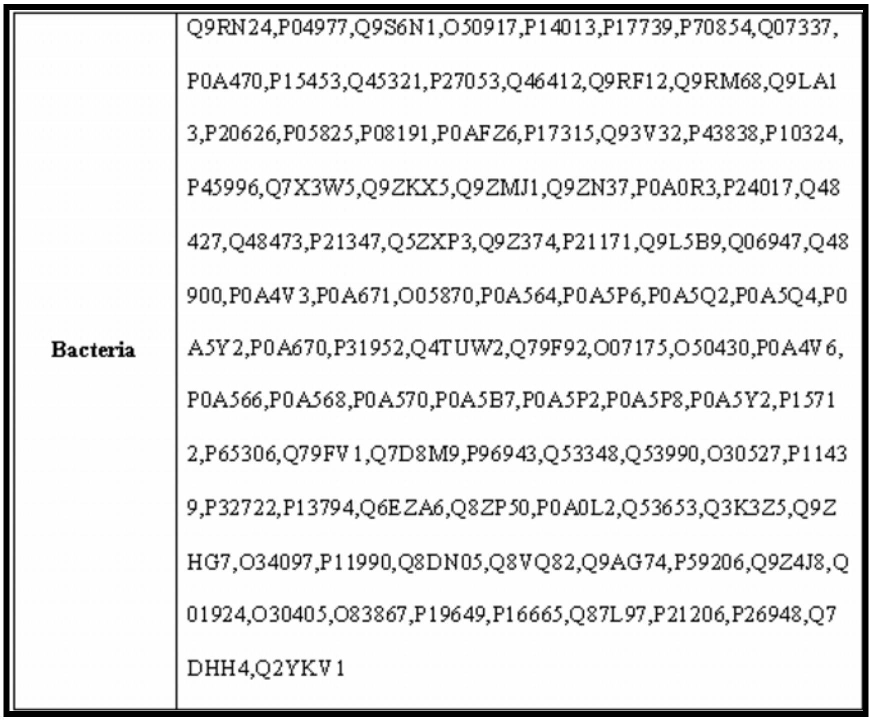
In the era of reverse vaccinology, we may pose ourselves the question: given the genome of a pathogen, how do we choose appropriate antigens as efficacious candidate subunit vaccines? The immunogenicity of an antigen arises no doubt from a complex interplay of factors. Traditionally, people have seen sub-cellular location as a prime determinant: immunogenic proteins need to be accessible to surveillance by the immune system; and as such they are likely to be secreted or located on the cell surface. Others have sought to use sequence similarity to identify antigens **[2]** or developed non-parametric methods to predict antigen status **[3, 4]**. Yet another alternative postulates that the immunogenicity of a protein is determined by its overall lack of similarity to the host proteome **[5]**. This idea is, in its simplicity and directness, quite compelling. Based upon this assumption, and through the use of the sequence similarity search tool BLAST, we attempt here to leverage sequence similarity to discriminate effectively antigen from non-antigen and establish a threshold that ostensibly separates the two classes. If successful, such an undertaking would constitute a significant advance in rational vaccine design. More specifically, we sought to analyse the degree of similarity between known bacterial antigens and the Human Genome and Mouse Genome, compared with a control group of non-antigens, and thus establish a systematic benchmark for such a strategy.

Methodology:

Sequence similarity comparison of antigen datasets, non-antigen datasets and various pathogenic genome sequences to the Human and Mouse Genome was performed using a local, standalone version of BLAST **[6]**, which allowed full control of E-value cut-off and of the substitution matrix used. E-value thresholds were raised from 10 to 6000 to identify best matches even when these lacked statistical significance. Seven substitution matrices were examined: BLOSUM45, BLSOUM62, BLOSUM80, BLOSUM90, PAM30, PAM70 and PAM250. Datasets of 100 known antigens obtained from the literature and 100 non-antigens from bacteria were analysed **[3].** Non-antigens were selected from SWISSPROT so that they mirrored the antigen sets, being selected stochastically from the same set of species **[3, 4]**. Bacterial antigens used are listed in **Figure 1.** Additionally, genomes corresponding to Human, Mouse, and Bacteria were downloaded from FTP sites at National Center for Biotechnology Information (NCBI) [http://www.ncbi.nlm. nih. gov/], European Bioinformatics Institute (EBI) [http://www.ebi.ac. uk/], and Ensembl [http://www.ensembl.org/]. Looking at ($\log_{10}{}^{\text{E-value}}$) +1 values from BLAST, the Mann–Whitney – or two-sample rank - test was used to compare the bacterial antigen and non-antigen sets, as random samples of two larger, independent populations, using the statistical package Minitab, Release 14.1.

Discussion:

A dataset comprising 100 known bacterial literature protein antigens, as well as a control set of 100 bacterial protein non-antigens [**3, 4**], were compared against the Human and Mouse Genomes, resulting in lists of the highest scoring similarities. For completeness, a similar analysis for five whole bacterial genomes (*Escherichia coli, Chlamydophila pneumoniae, Haemophilus influeanzae, Salmonella typhi* and *Mycobacterium tuberculosis*) was also undertaken. All sets analysed were similarly distant to both mouse and human genomes. Scores for antigen, non-antigen, and *E. coli* proteome versus the human proteome are plotted in **Figure 2.** It is clear that the three overall distributions strongly overlap, and that no cut-off point is apparent.



Bacteria	Q9RN24,P04977,Q9S6N1,O50917,P14013,P17739,P70854,Q07337,P0A470,P15453,Q45321,P27053,Q46412,Q9RF12,Q9RM68,Q9LA13,P20626,P05825,P08191,P0AFZ6,P17315,Q93V32,P43838,P10324,P45996,Q7X3W5,Q9ZKX5,Q9ZMJ1,Q9ZN37,P0A0R3,P24017,Q48427,Q48473,P21347,Q5ZXP3,Q9Z374,P21171,Q9L5B9,Q06947,Q48900,P0A4V3,P0A671,O05870,P0A564,P0A5P6,P0A5Q2,P0A5Q4,P0A5Y2,P0A670,P31952,Q4TUW2,Q79F92,O07175,O50430,P0A4V6,P0A566,P0A568,P0A570,P0A5B7,P0A5P2,P0A5P8,P0A5Y2,P15712,P65306,Q79FV1,Q7D8M9,P96943,Q53348,Q53990,O30527,P11439,P32722,P13794,Q6EZA6,Q8ZP50,P0A0L2,Q53653,Q3K3Z5,Q9ZHG7,O34097,P11990,Q8DN05,Q8VQ82,Q9AG74,P59206,Q9Z4J8,Q01924,O30405,O83867,P19649,P16665,Q87L97,P21206,P26948,Q7DHH4,Q2YKV1

Figure 1: Protein sequences compiled and annotated in-house as Bacterial antigens.

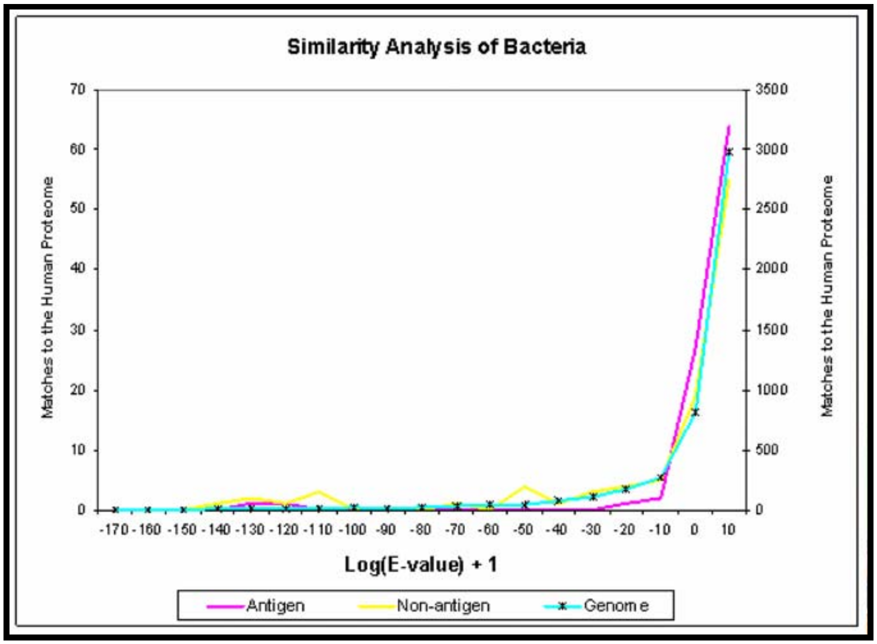


Figure 2: A sequence similarity comparison with the E-value as 6000 and BLOSUM 62 matrix, between the Antigen, Non-antigen and *E-coli* genome sequences. Two separate scales were used as the number of matches to the Human Genome varied from the antigen and non-antigen datasets to the genome. The blue line with the star marker symbolizes the genome is plotted on the scale to the right hand axis (Y axis).

We also analysed the effect of modifying BLAST parameters (data not shown), primarily by varying the substitution matrix. Such matrices purportedly measure the probability of evolutionary divergence of the protein to develop an evolutionary protein model. Technical practicalities limited us to 7 matrices: 4 BLOSUM series and 3 PAM series matrices. To a first approximation, the use of different matrices makes no operationally significant difference: although individual values varied, the shape and structure of the plotted score distributions remained largely unaltered. Moreover, the ability to discriminate antigens from non-antigens was essentially unaffected.

We also assessed the statistical significance of differences between distributions of scores using the Mann-Whitney test, a non-parametric statistical method, which tests the differences in medians rather than means, since the population being analysed was skewed and un-normalised. We used two different confidence intervals: 90% and 95%; p-values were the same for each. Although p-values for several of the PAM and BLOSUM matrices were marginally significant, overall - and in accordance with our visual interpretations - the test established the statistical null hypothesis that the apparent differences resulted solely from chance. The most promising results came with the use of the PAM30 matrix, but even here the result was again marginal and poorly predictive.

Overall, and given our original objectives, these results can be seen as disappointing, yet they leave no doubt that there is no obvious and significant difference between bacterial antigens and bacterial proteins or proteomes deemed to be non-antigens, and there is certainly no clear and unequivocal means to differentiate the two sets simply on the basis of scores. The non-antigen set seems to contain marginally more proteins that are more similar to the eukaryotic host proteomes than the antigen set, though this may be a statistical quirk due to the limited data set we use. It may thus be that when dissimilarity screening versus the human proteome is used in concert with other properties characteristic of antigens, such a technique predicated on dissimilarity will in part prove useful, at least in terms of rejecting a portion of the bacterial genome as low-likelihood candidates. However, as a tool for vaccine discovery, dissimilarity is in itself of limited utility, at least for Bacteria and as the sole arbiter of immunogenicity.

Thus, a key feature of future work would be to integrate this approach with other nascent techniques for antigen identification. Other work might include making use of more sophisticated approaches available for sequence similarity analysis, such as PatternHunter **[7].** Rather than use pairwise alignment, use could be made of multiple sequence alignment to derive distances, as implemented in say T-coffee [**8**]. Likewise, the more sophisticated scoring available in phylogenetic software, such as PHYLIP or PALM [**9**], could prove useful. In terms of analysing data, some virtue may be gained by using Operating Characteristic (ROC) plots and the 2 sample Kolmogorov Smirnov test to compare distributions.

Conclusion:

The hypothesis of Kundac *et al.* **[5]** is undeniably compelling. However, our examination provides evidence that contradicts this hypothesis, at least as we restate it: that non-redundancy is key to predicting candidate vaccine antigens. To abbreviate our story: we were unable to determine a suitable threshold able to differentiate antigens from non-antigens on the basis of similarity to the host's proteome. Both antigens and non-antigens were, to a first approximation, essentially equally dissimilar to host proteomes, and about as dissimilar as the background distribution provided by whole bacterial genomes. We conclude that straightforward dissimilarity searching is not a suitable approach for selecting candidate vaccine antigens from bacteria. We are currently pursuing more sophisticated and successful approaches, so far without success.

Protein immunogenicity arises from many factors. These include host-side properties - possession of B or T cell epitopes for example - and pathogen-side properties - protein expression levels and sub-cellular location - as well as its aggregation state and the possession of post-translational danger signals. A candidate vaccine should be highly expressed, available for immune surveillance, and possess epitopes that the host recognises. Predicting such diverse properties remains challenging, though several contributing factors can be reliably predicted. What we need is an integrative, systems biology approach to the problem. Our search for such an approach continues.

References :

[1] S Vivona *et al*. *Trends Biotechnol.*, **26**:190 (2008) [PMID : 18291542]
[2] H Ansari *et al*. *Nucleic Acids Res.*, **38**:D847 (2010) [PMID : 19820110]
[3] I Doytchinova & D Flower, *Vaccine*, **25**:856 (2007) [PMID : 17045707]
[4] I Doytchinova & D Flower, *BMC Bioinformatics*, **8**:4 (2007) [PMID : 17207271]
[5] D Kanduc *et al*. *Autoimmun. Rev*. **6**:290 (2007) [PMID : 17412300]
[6] S Altschul, *Nucleic Acids Research*, **25**:3389 (1997) [PMID : 9254694]
[7] M Li *et al*. *Genome Inform.* **14**:164 (2003) [PMID : 15706531]
[8] C Notredame *et al. J Mol Biol.* 302:205 (2000) [PMID : 10964570]
[9] S Chen *et al. PLoS One.* 4:e8116 (2009) [PMID : 19997614]

Edited by P. Reche
Citation: Ramakrishnan & Flower, Bioinformation 4(10): 445-447 (2010)