Bioinformation
www.bioinformation.net
Volume 5
Issue 1
open access
Hypothesis

Discriminating antigen and non-antigen using proteome dissimilarity III: tumour and parasite antigens

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

[1]The Jenner Institute, University of Oxford, Compton, Newbury, Berkshire, United Kingdom; RG20 7NN Medical Genetics Section, University of Edinburgh, Edinburgh, United Kingdom EH4 2XU; [2]Aston University, Life and Health Sciences, Aston University, Aston Triangle, Birmingham, United Kingdom, B5 7ET. E-mail: D.R.Flower@aston.ac.uk; phone+44 (0)121 204 5182; * Corresponding author.

Received May 27, 2010; accepted June 09, 2010, published June 24, 2010

Abstract:
Computational genome analysis enables systematic identification of potential immunogenic proteins within a pathogen. Immunogenicity is a system property that arises through the interaction of host and pathogen as mediated through the medium of a immunogenic protein. The overt dissimilarity of pathogenic proteins when compared to the host proteome is conjectured by some to be the determining principal of immunogenicity. Previously, we explored this idea in the context of Bacterial, Viral, and Fungal antigen. In this paper, we broaden and extend our analysis to include complex antigens of eukaryotic origin, arising from tumours and from parasite pathogens. For both types of antigen, known antigenic and non-antigenic protein sequences were compared to human and mouse proteomes. In contrast to our previous results, both visual inspection and statistical evaluation indicate a much wider range of homologues and a significant level of discrimination; but, as before, we could not determine a viable threshold capable of properly separating non-antigen from antigen. In concert with our previous work, we conclude that global proteome dissimilarity is not a useful metric for immunogenicity for presently available antigens arising from Bacteria, viruses, fungi, parasites, and tumours. While we see some signal for certain antigen types, using dissimilarity is not a useful approach to identifying antigenic molecules within pathogen genomes.

Background:
Vaccines induce provoke protective immunity and come in both "living" and "non-living" varieties. Living vaccines are usually attenuated or weakened pathogenic microbes which retain aspects of natural infection including the ability to revert to a pernicious form. Non-living vaccines are either chemically or heat treated whole pathogens or pathogen components. Subunit vaccines comprise protein components isolated from pathogenic micro-organisms and have several advantages: they have longer shelf-lives, are more stable, and cannot regain pathogenic status. However, their identification can be arduous requiring the expense of much time and resource.

Tumour immunotherapy refers to using the host immune system to battle cancer. Strictly, a tumour is a solid lesion or neoplastic growth resulting from unregulated cell division, and may be benign, pre-malignant, or malignant; we use the term synonymously with cancer, and specifically here as a pseudonym for human tumour antigens extracted from the SEREX database. Tumour vaccines contain a specific protein derived from a tumour able to stimulate a protective immune response. Tumour vaccines are therapeutic rather than prophylactic in nature, and are typically injected subcutaneously or directly into cancerous tissue. They are a nascent form of personalised medicine, with different vaccines targeting different cancers, with the potential to identify antigens directly from the patient.

Diseases of parasite and protozoan origin cause significant mortality and morbidity: over 3.5 billion people currently suffer parasitic infection, primarily in tropical and subtropical countries, particularly pastoral regions of Asia, Latin America, and Africa; incidence in industrialized countries is relatively low. Parasites typically invade the body *via* mucosal surfaces. Buccal parasites, for example, after ingestion will either remain in the intestine or escape via the intestinal wall, invading other organs; while some will bore through the skin or enter *via* insect bites. The life-cycles of many parasites, particularly single-celled parasites, are complex with many stages involving eggs and larval forms yet usually reproduce within the host. This makes developing vaccines extremely problematic, and currently there are no licensed vaccines available targeting parasitic diseases.

Genomics is fashioning a new epoch of knowledge-led vaccine design and discovery. Known as reverse vaccinology, it combines advanced molecular biology technology with advanced *in silico* analysis of pathogen genomes, enabling the systematic identification of potential antigens within a pathogen. Key to this endeavour is the bioinformatics protocols used to detect antigens, such as those which predict sub-cellular location as the main determinant associated with antigens. As proposed by Kundac *et al.* **[1]**, another persuasive concept is the idea of dissimilarity of antigens versus non-antigens at the sequence level. In this paper, we extend our previous analysis **[2, 3]** beyond bacteria, viruses, and fungi, to explore parasite and tumour antigens.

Methodology:
Datasets of known antigens obtained previously from the literature were analysed **[2, 3, 4, 5]**. Non-antigens were randomly selected from Swiss-Prot so that they mirrored the species distribution within the antigen sets **[2, 3]**. Parasite and Tumour antigens used here are listed below in **Figure 1.** Additionally, genomes corresponding to Human, Mouse, and Parasite 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/], and tumour sequences from the SEREX datasets available from the Cancer Immunome Database [http://ludwig-sun5.unil.ch/CancerImmunomeDB/]. Tumour non-antigens were collected at random by selecting human proteins from Swiss-Prot. All the peptide sequences obtained were in FASTA format.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 5(1): 39-42 (2010)

Bioinformation Volume 5
www.bioinformation.net Issue 1
open access
Hypothesis

As before **[2, 3]**, antigen datasets, non-antigen datasets and parasite genomes were compared to the Human and Mouse Genome, and analysed with a local, standalone version of BLAST **[6]**, which afforded full management of E-value cut-offs. E-value thresholds were raised from 10 to 6000 to identify best matches even when these lacked statistical significance. We also analysed $(\log_{10}^{E\text{-value}})+1$ values obtained from BLAST. By using the statistical package Minitab, Release 14.1, we compared antigen and non-antigen sets, as random samples of two larger, independent populations, utilising a Mann–Whitney test.

Discussion:

Previously, we have examined the difference between sets of antigens and non-antigens for Bacterial, fungal, and species, and found no clear separation of the two, concluding that proteome dissimilarity does not provide a means of sifting out potential antigens from a newly sequenced pathogen genome. Here we have expanded our analysis, focusing on parasite and tumour antigens.

Compared to our previous sequence similarity analyses **[2, 3]**, parasite non-antigens evinced more noticeable dissimilarity to the human genome than did parasite antigens, suggesting a clearer separation than before. See **Figure 2.** However, there were seven antigenic proteins that demonstrated high similarity to the human genome compared to one equivalently similar non-antigen. We also evaluated the genomes of four different parasite species – *Cryptosporidium parvum, Distyostelium discoideum, Leishmania infantum* and *Trypanosoma brucei* – as a background reference, or a "control" as some would put it, for this comparison of antigens and non-antigens, comparing them to both human and mouse genomes. The distribution of matches between parasite and human genomes indicates that most proteins lie inside the range characterising the antigen proteins. While a signal is clearer here than before **[2, 3]**, the distribution is inverted compared to our expectation: a discernible proportion of the antigens were more similar, not less, than are the non-antigens. Thus for these proteins, antigenicity is encoded in a subtle and cryptic manner, not apparent simply from sequence comparison. It is possible, but likely, that using some form of similarity filter - rather than a dissimilarity filter - may provide a threshold able to indicate the potential antigenicity of parasite proteins.

We observed that the distributions of tumour antigen matches to the human and mouse genomes spread across a wide E-value range scale, much wider than seen previously **[2,3]**. **Figure 3** illustrates the analysis of tumour antigens, non-antigens, and reference genomes relative to the human genome. The distributions characterizing antigens and non-antigens were similar, yet antigens were by visual inspection proportionally more similar to the human genome relative to non-antigens. This observation is again inverted compared to our expectations, as was the case for parasites. This distribution may in part result from the presence of both antigens and non-antigens in the host. Thus identifying a threshold able to separate tumour antigens from non-antigens would again prove difficult.

As well as visual inspection of the distributions of antigens and non-antigens for tumour and parasite, we also undertook a statistical comparison using the Mann–Whitney Test. At a 95% confidence level, the test gave a value of 0.000 for Parasite and 0.001 for Tumour. All previous assessments accepted the null hypothesis. Since these values are less than 0.05, this is indicative of a statistically significant discrepancy between the two antigen-versus-non-antigen distributions from both tumours and parasites. While the apparent significance for parasite and tumour was marked, it was somewhat at odds with the visual inspection of the histograms of similarity values. Although there may be a statistically significant difference in both paired distributions, there is again no useable cut-off capable of distinguishing antigen from non-antigen.

Parasite	O15808,Q6JH13,Q6RXY4,Q27396,Q3LUQ3,Q3LUQ2,Q3LUQ1,Q3LUP5,Q9BHA2,O15709, Q9GN06,Q8IHK7,Q94661,Q4QHG9,Q25306,Q9TVI8,Q4QEG6,Q9BJC7,Q4QHB5,Q9GRP6, Q4QD68,Q5ZQK2,Q8MLY1,O77301,Q9NJS2,O96368,Q9U8F4,Q8MUC1,Q8ISG0,Q05870, Q964D2,Q9NBD2,Q27354,P09792,P39097,O76993,O76992,O96346,Q24797,Q86G73,Q9BH 49,Q25029,Q2I7F0,Q6PZ53,Q966V8,Q8I867,Q18L65,O00816,Q9XZH7,O15681,Q967S9,P90 708,Q4DPU1,Q01137,P16026,Q00277,Q86LH7,Q8I1K3,Q7YZP0,O96986,Q6SJP6,Q9GRG4, Q86M44,O18714,P27730,P62884,Q9UB12,Q9TVP5,P43150,P36400,Q7KFG4,P21978,P2949 8,Q2I7E9,Q9XZG2,P13828,Q26883,Q7JPX5,Q26519,Q9U9R9,P08677,Q4VDQ1,O96555,Q9 NL18,P15744,Q8MPM6,Q08377,Q8I0N2,Q27298,Q95NF7,P13404,Q07828,P13403,Q27002, O00933,Q5Y808,P90661,Q25540,Q26971,O00930,Q17115,Q5XXE0,Q00708,Q26675,Q9N62 0,P26624,P27591,Q76LS6,Q25021,Q8I476,Q9GPL0,P06915,P48501,P06198,P19331,P27730
Tumour	10,13,14,19,24,79,81,82,97,103,104,108,109,112,113,116,126,129,131,133,137,161,163,180,1 89,190,195,196,216,234,244,251,349,355,360,374,375,417,432,434,435,453,457,463,551,565, 677,698,802,807,939,955,990,1014,1016,1020,1026,1047,1058,1069,1078,1080,1084,1093,10 99,1101,1110,1129,1137,1158,1163,1166,1179,1185,1188,1190,1192,1209,1210,1214,1220,12 33,1246,1252,1253,1260,1262,1279,1289,1320,1325,1328,1424,1438,1546,2018,2024,2032,,2 033,2034

Figure 1: Protein sequences compiled and annotated as antigens of Parasite and Tumour origin. Parasite sequences are denoted by their Swiss-Prot/TrEMBL codes. Tumour sequences are denoted by their SELEX codes.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 5(1): 39-42 (2010)

Bioinformation **Volume 5**
www.bioinformation.net Issue 1 *open access* **Hypothesis**

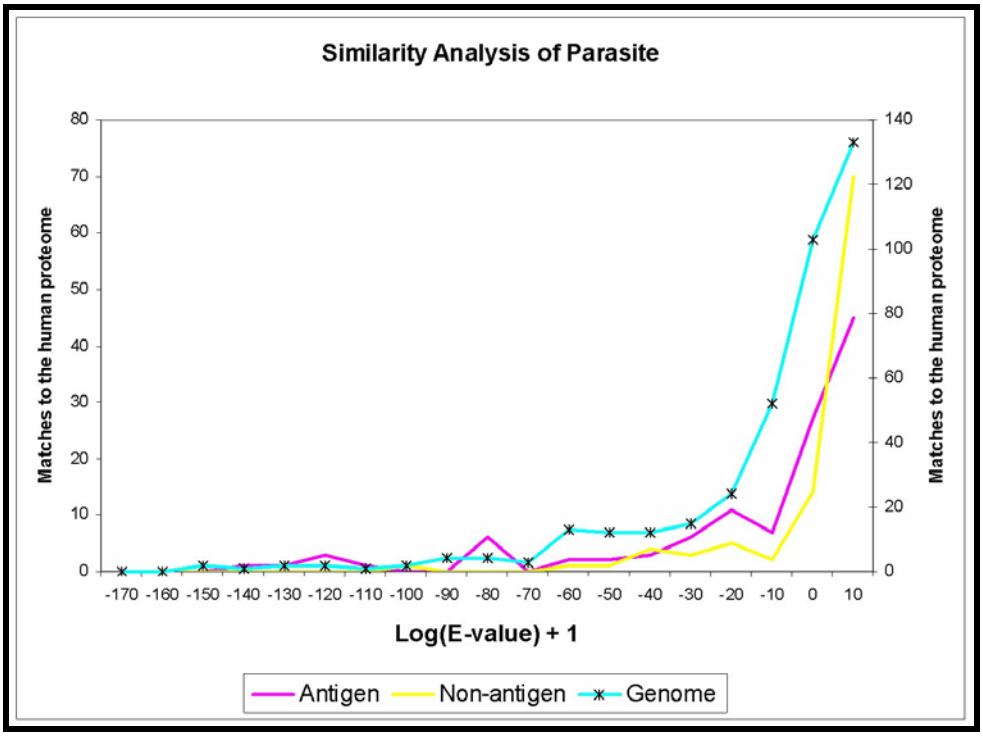

Figure 2: A sequence similarity comparison with the E-value as 6000 and BLOSUM 62 matrix, between the Antigen, Non-antigen and *Cryptosporidium parvum* genome sequences. Two separate scales were used as the number of matches to the Human Genome varied from the antigen and non-antigen datasets to the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

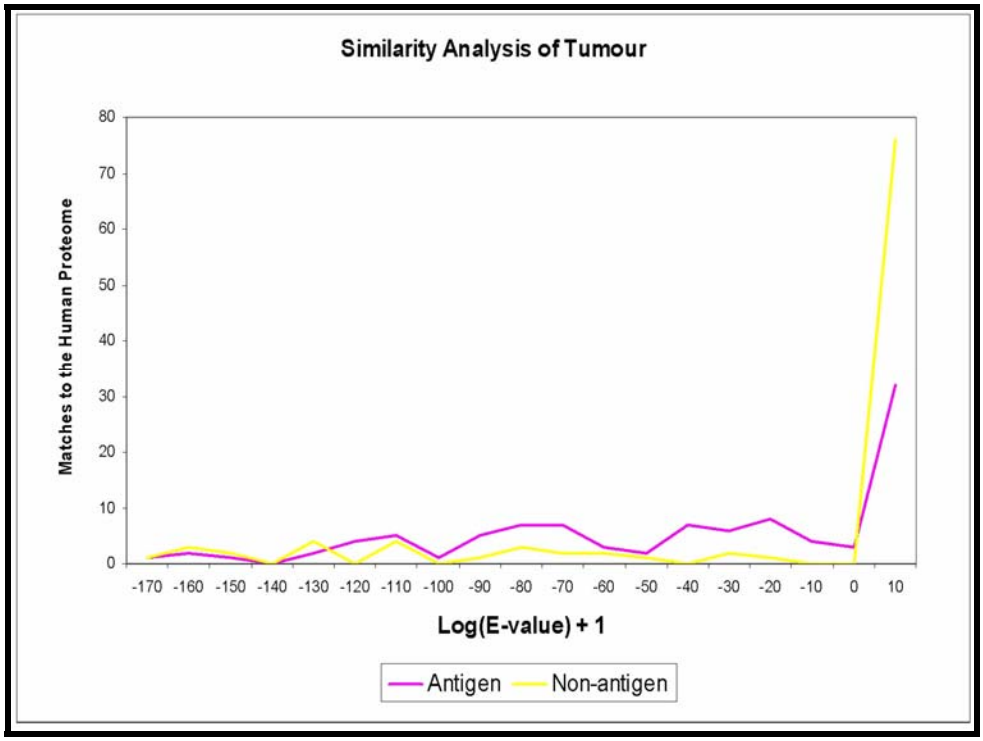

Figure 3: A sequence similarity comparison between Tumour antigens and the whole human proteome. E-value set at 6000 and BLOSUM 62 matrix, between tumour Antigen and Non-antigen. All self-matching identities were excluded from the results. A background control at the genome level was not conducted.

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 5(1): 39-42 (2010)

Bioinformation **Volume 5**
www.bioinformation.net Issue 1 **Hypothesis**
open access

Conclusion:
Antigens are the basis of subunit vaccines, and their identification by computationally-driven reverse vaccinology is vital, particularly in the era of emergent zoonotic infections and recrudescent disease, since many recently discovered pathogens cannot be cultivated, thus precluding the facile experimental identification of their immunogenic antigens. To this endeavour, immunoinformatics searches continually for robust and celeritous approaches to antigen prediction. One such is based on global dissimilarity searching, as suggested by Kanduc *et al.* **[1]**.

When compared to our previous analyses **[2, 3]**, both tumour and parasite distributions had a clearer and more discernible signal than before. However, both were inverted relative to our expectations, with antigens being more similar to the human genome than non-antigens. We felt this was counter-intuitive, but there may be evolutionary arguments consistent with this observation. Being so close to mammalian hosts, parasites may need to have evolved more complexity at the functional, and thus sequence and structural, levels, in order to allow their own primitive immune systems, to recognise and protect themselves from self-inflicted damage. Of course, such may be wholly specious arguments, and the true source of the apparent differences manifest as signal, may come from a statistical quirk or an observed sampling bias.

The present work is not definitive, and there is much further work that could be done, though radically different results seem unlikely. We envisage repeating our study as more antigens become available **[7]**; looking perhaps at functionally or immunologically congeneric subsets within the overall data; using more sensitive and sophisticated similarity assessment operating at both the global and local levels; and combining this approach with other methodology in a more extensive, rigorous, and comprehensive analysis.

References:
[1] D Kanduc *et al.* Autoimmun Rev6:290 (2007) [PMID : 17412300]
[2] K Ramachrisnan & D Flower, *Bioinformation* **4** (10): 445 (2010)
[3] K Ramachrisnan & D Flower, *Bioinformation* **4** (10): 445 (2010)
[4] I Doytchinova & D Flower *Vaccine* **25**:856 (2007) [PMID : 17045707]
[5] I. Doytchinova & D Flower *The Open Vaccine journal*, 1:22 (2008)
[6] S Altschul, *Nucleic Acids Research* **25**:3389 (1997) [PMID : 9254694]
[7] H Ansari *et al*, Nucleic Acids Res **38**:D847 (2010) [PMID : 19820110]

Edited by P. Kangueane
Citation: Ramakrishnan *et al.* Bioinformation 5(1):39-42 (2010)

ISSN 0973-2063 (online) 0973-8894 (print)
Bioinformation 5(1): 39-42 (2010)