In silico validation of a reverse vaccinology pipeline using a peptide binding dataset

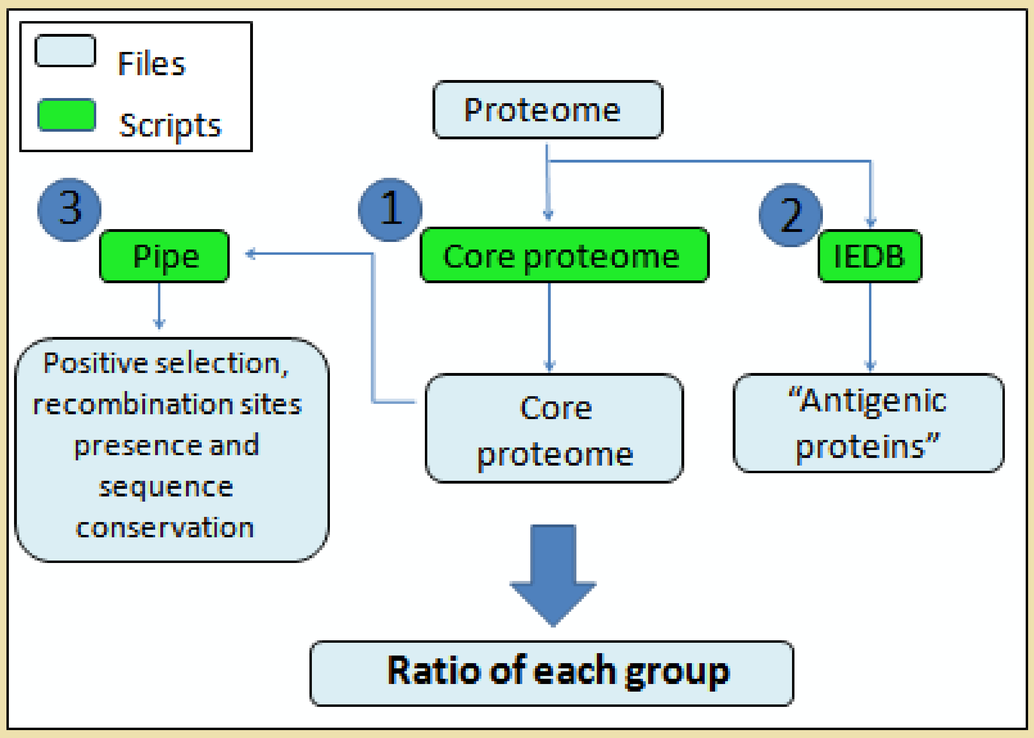
Hobeich Naya, Carlos
Final Project - Degree in Microbiology, Course 2013-2014
Department of Genetics and Microbiology, Fac Biosciences. Autonomous University of Barcelona
08193 Bellaterra (Cerdanyola del Vallès). Barcelona, SPAIN

Universitat Autònoma de Barcelona

Background

Reverse vaccinology involves several computational tools to predict, directly from the genome sequence, protein characteristics used to identify proteins which are worthy of laboratory investigation.
Some protein sequence characteristics have been used to look for potentially immunogenic proteins: Positive selection, presence of recombination sites or 100% conserved sequence.
In this work, an *in silico* protocol is established to assess the suitability of this strategy. An in house Perl script was created to calculate the proportion, expressed in percentage, of "antigenic proteins" from each group. Calling "antigenic protein" to those that match with a peptide binding dataset, extracted from the Immune Epitope Database[1], results will be compared with the proportion found in complete proteomes (referred henceforth as the benchmark dataset) to try to infer the relationship between the potential immunogenicity and the pertinence to a specific group

Workflow



1. **To construct the core proteome (set of proteins common to all strains) reciprocal best sequence alignment hit from all proteomes (by specie) was assumed to represent orthologous sequences. (In house method using BLAT[2]).**

2. **This script generates a list where the query that matched, totally or partially, to a positive annotated epitope sequence from the peptide binding dataset (from IEDB) is shown. (In house method also using BLAT)**

3. **Amino acid and/or DNA sequences from the core proteome undergo several analytical steps integrated in the pipeline: multiple sequence alignment, calculation of protein variability among selected proteomes, detection of recombination and selective pressure based on the DNA codon alignment. (In house method using ClustalW[3], PHYLIP[4], Pal2Nal[5] and HYPHY[6])**

Results

In Figure 1 and 2, for all strain of each species, the arithmetic mean of the proportion of the number of proteins that matched with an epitope sequence in IEDB can be seen. This result is expressed in percentage respect all proteins from each group.

Specie	Proteome	Core proteome	100% Conserved	Recombinant regions	Positive selection	Recombinant regions and positive selection
A. baumannii	8,83	10,52	15,58	12,80	12,14	19,00
B. pseudomallei	9,18	11,69	13,00	17,71	15,52	17,07
C. jejuni	11,20	14,47	0	17,95	18,18	22,22
C. trachomatis	17,99	14,38	9,09	36,00	22,16	54,40
C. botulinum	6,75	14,05	0	16,68	23,81	27,78
E. coli	17,88	21,81	17,86	28,78	29,18	35,83
F. tularensis	23,02	25,82	12,77	45,45	29,17	42,86
H. influenzae	17,17	19,26	4,35	21,66	24,20	24,69
H. pylori	13,09	14,63	0	15,90	16,16	16,43
L. monocytogenes	11,79	13,23	13,59	18,31	10,71	6,67
P. aeruginosa	9,64	10,95	16,76	15,80	14,10	16,38
S. enterica	16,39	18,93	19,78	27,01	24,76	27,22
S. aureus	9,76	11,63	13,00	22,59	15,00	23,08
S. pyogenes	11,08	14,54	9,52	22,76	17,84	19,19
T. pallidum	10,03	10,62	10,68	14,29	6,67	25,00
Y. pestis	17,07	20,27	19,01	31,81	20,84	31,47

Figure 1. Proportion of "antigenic proteins" by group.

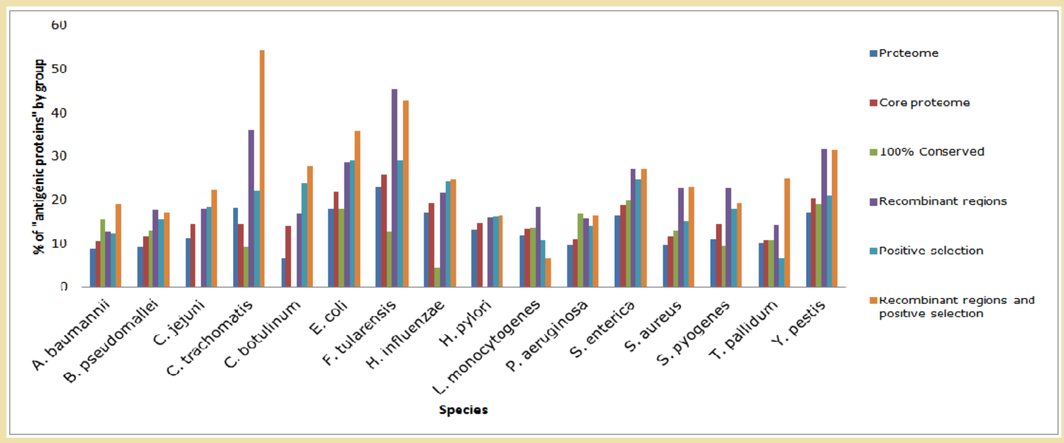


Figure 2. Bar chart from Figure 1 data.

The full data of each specie can be downloaded from http://bioinf.uab.es/hobeich/TFG/tablas_suplementarias.xls

Conclusions

If these results were representatives, we can conclude that, as found in most bibliography[7, 8, 9], positive selection, recombination sites or presence in core proteome could be a good characteristic when looking for candidates in the reverse vaccinology process.
These results could validate the use of a bioinformatic strategy based on selecting proteins under selection pressure for the election of vaccine candidates.
To ensure a better randomness in data a better species selection process and a bigger number of species are required.
The implementation of statistical analysis on data will allow to clearly distinguish those groups with higher proportions of potentially immunogenic proteins

Bibliography

1. Vita R, Zarebski L, Greenbaum JA, Emami H, Hoof I, Salimi N, et al. (2010). The immune epitope database 2.0. *Nucleic Acids Res*, 38(Database issue):D854–862.
2. Kent WJ. (2002). BLAT--the BLAST-like alignment tool. *Genome Res*., 12:656–664.
3. Larkin MA, et al. (2003). ClustalW and ClustalX. *Options*, 2:1–22.
4. Felsenstein, J. (1989). PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* 5: 164-166.
5. Suyama M, Torrents D, Bork P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*., 1;34(Web Server issue):W609–612.
6. Pond SLK, Frost SDW, Muse S V. (2005). HyPhy: hypothesis testing using phylogenies. *Bioinformatics*, 21:676–9.
7. Fitzpatrick DA, Creevey CJ, McInerney JO. (2005). Evidence of positive Darwinian selection in putative meningococcal vaccine antigens. *J Mol Evol*, 61:90–98.
8. Mes THM, van Putten JPM. (2007). Positively selected codons in immune-exposed loops of the vaccine candidate OMP-P1 of Haemophilus influenzae. *J Mol Evol*, 64:411–422.
9. Brehony C, Wilson DJ, Maiden MCJ. (2009). Variation of the factor H-binding protein of Neisseria meningitidis. *Microbiology*, 155(12):4155–4169.

Memory downloadable at http://bioinf.uab.es/hobeich/TFG

Acknowledgements

I would like to thank Oscar Conchillo-Solé and Dr. Daniel Yero for his advice
This project was carry out at the Institut de Biotecnologia i de Biomedicina at the Universitat Autònoma de Barcelona