A Genome-Level Search for Bacterial Genes on Which Positive Selection May Operate: A Means for Identifying Possible Virulence Factors?

Junaid Gamieldien junaid@sanbi.ac.za

Winston Hide winhide@sanbi.ac.za

South African National Bioinformatics Institute, University of the Western Cape
Private Bag X17, Bellville, 7535, South Africa

1 Introduction

Positive selection is very important for the evolution of a new or improved protein function, as it increases the probability that an advantageous mutation becomes established in a population. Previous studies have shown that genes on which positive selection may operate can be identified by comparing synonymous substitution rates (K_S) and nonsynonymous substitution rates (K_A), where $K_A > K_S$.

It has been previously shown that positive selection may operate on the virulence genes of pathogens to enable them to maintain or improve their 'advantage' over their hosts. K_A/K_S analysis of genes coding for cell surface proteins of the malaria parasite, *Plasmodium* spp., have identified gene regions that have nonsynonymous substitution rates that are significantly higher than the synonymous rate [3]. Since host antibodies target these protein regions, it has been suggested that this phenomenon is evidence of positive selection-based diversification of the protein to maintain an advantage over the host immune system. A similar phenomenon has been demonstrated in highly pathogenic HIV1 versus mildly pathogenic HIV2 and in highly pathogenic versus mildly pathogenic SIV [4]. A broader study by Endo *et al.* [2] analyzed approximately 3600 groups of homologous sequences obtained from DNA sequence databases and found 17 candidate groups that showed evidence for positive selection. Of these, 9 groups were the surface antigens of parasites or viruses.

A number of bacterial virulence-associated genes have been demonstrated to be under positive selection. Recently, the cytotoxin-associated gene A (*cagA*) of *Helicobacter pylori*, found in nearly all isolates from patients with peptic ulcer disease, has been shown to be under positive selection [5]. It has also been suggested that two proteins from pathogenic *Neisseria* species and *Haemophilus* species, which play a role in iron acquisition from host transferrin, are under positive selection [1]. This indicates that positive selection may operate on pathogen virulence genes other than those coding for surface antigens.

We are developing a system for the identification of bacterial virulence genes at a raw, whole-genome level, using K_A/K_S ratios. Here, we present a brief introduction to the system and our preliminary results.

2 Methods

To test our hypothesis, we have performed two intraspecies K_A/K_S comparisons using genomic sequences from the different strains of *Mycobacterium tuberculosis* and *Neisseria meningitidis*, available from the Sanger Center and The Institute for Genomic Research.

The major steps in the method are:

a. Open reading frame prediction
b. Searching for identical genes in each intraspecies ORF set
c. DNA alignment (built from protein alignments)
d. Windowed calculation of K_S and K_A
e. Selection of genes that have regions where $K_A > K_S$
f. Sequence similarity searches of candidates for their functional annotation

3 Results

For *M. tuberculosis*, we have identified a number of genes under positive selection that may biologically be associated with virulence. A large percentage of these genes code for enzymes that produce modified fatty acids found uniquely in the cell walls of pathogenic mycobacteria. Others include a transcription regulator of virulence factors, transmembrane proteins, and a protein that has a potential role in adaptation of the pathogen to intracellular (macrophage) conditions. Analysis of the limited *N. meningitidis* genomic sequences produced similar results. These included genes coding for a hemoglobin receptor, which plays a role in acquisition of host iron; two host-cell adhesion proteins; and an enzyme involved in serum resistance of the pathogen. A number of genes were previously reported to code for virulence factors.

4 Summary

Since a number genes detected were previously reported to be virulence genes, and many other candidates may be associated with virulence based on their predicted function, there is suggestive evidence that the system has the potential to identify many virulence genes from the raw genomic data of a bacterial pathogen.

References

[1] Cornelissen, C.N. and Sparling, P.F., Iron piracy: acquisition of transferrin-bound iron by bacterial pathogens, *Mol. Microbiol.*, 14(5):843–850, 1994.

[2] Endo, T., Ikeo, K., and Gojobori, T., Large-scale search for genes on which positive selection may operate, *Mol. Biol. Evol.*, 13(5):685–690, 1996.

[3] Ohta, T., Cirumsporozoite protein genes of malaria parasites (*Plasmodium* spp.): evidence for positive selection on immunogenic regions, *Genetics*, 127(2):345–353, 1991.

[4] Shpaer, E.G., Mullins, J.I., Rates of amino acid change in the envelope protein correlate with pathogenicity of primate lentiviruses, *J. Mol. Evol.*, 37(1):57–65, 1993.

[5] Van der Ende, A., Pan, Z., Bart, A., Van der Hulst, R.W.M., Feller, L., Xiao, S., Tytgat, G.N.J., and Dankert, J., cagA-Positive *Helicobacter pylori* populations in China and The Netherlands Are Distinct, *Infect Immun*, 66(5): 1822–1826, 1998.