Poster C-24

Comparative sequence analysis of microbial genomes to identify potential virulence genes in Mycobacterium tuberculosis.

Authors:
Gordon Jamieson (*IIDMM*)
Halimah Rabiu (*IIDMM*)
Venu Vuppu (*IIDMM*)
Nicola Mulder (*IIDMM*)

Short Abstract: Microbes are masters of adaptation to their environment, which results in ever changing targets for drug and vaccine research. This project aims to identify and characterise virulence genes from Mycobacterium tuberculosis through cataloguing of all microbial genes and determination of their phylogenetic distributions.

Long Abstract:

The availability of complete genome sequences for microbial pathogens allows for the characterisation of virulence genes in Mycobacterium tuberculosis (M. tuberculosis) through intensive comparative genomic and proteomic technologies. Major differences between pathogen and non-pathogen data sets from large-scale data mining have been compiled into an in-house relational database. An initial interpretation of the results will be presented.

M. tuberculosis, has been known for some 120 years, the disease continues to plague humanity and remains the leading cause of death due to infection worldwide. About 8 million new cases of active tuberculosis arise each year resulting in 3 million annual deaths. The project aim is to identify and characterise potential virulence genes from M. tuberculosis through cataloguing of all microbial genes and their phylogenetic profiles. Potential virulence genes in M. tuberculosis can then be selected based on their presence in pathogens and absence from non-pathogens.

Traditionally, orthologous genes across different species have been identified using reciprocal best hit methods, but this approach has its limitations. Here we present a combinatorial approach for identifying orthologous protein sets using more than one method. Approximately eighty complete microbial genome sequences from human pathogens and non-pathogens have been identified and their complete non-redundant protein sets retrieved. A local database of all protein sequences from the complete genomes was generated and each proteome set was run through sequence similarity searches (BLAST) against this database. Perl scripts were written to parse and load data from the BLAST output files into a MySQL database. InterPro data has also been retrieved for each of the selected genomes in this study to provide a second method for classification of microbial proteins into related sets. Protein sets are classified according to the sets of protein signature methods they match, and thus the domains they contain and families they belong to. The final orthologous sets are formed from the union of protein sets from the BLAST and InterPro results. We have the option here to impose stricter or more relaxed rules to define the final protein sets.

The preliminary results have begun to highlight major differences between the two groups, and promises to provide a means for selecting potential virulence genes. At least 200 M. tuberculosis genes have no significant hits to the local database, and at least 300 genes appear to be unique to closely related mycobacterial species. Interestingly, a large number of hypothetical proteins have been identified. It is these uncharacterised proteins that may offer insight in search of novel virulence genes. GO data has also been retrieved for each of the selected genomes in this study, which, together with the InterPro data, provides functional characterisation for all proteins, where available. Further characterisation of proteins will also be achieved from other information resources: literature data, protein sequence analysis and structures, and microarray data. The potential virulence genes of interest will undergo evolutionary studies to determine whether they evolved through horizontal transfer into the pathogens or gene loss from non-pathogens. The results of the study show how mining of genomics data can contribute to a deeper understanding of mycobacterial pathogenesis and outlines a computational approach that may be broadly applicable for studying pathogenesis in other organisms.