Genomic scale analysis of horizontal gene transfer from bacteria to unicellular eukaryotes

David R. Westhead*, John W. Whitaker and Glenn A. McConkey

Faculty of Biological Sciences, University of Leeds

1 Introduction

We have recently developed metaTIGER (http://www.bioinformatics.leeds.ac.uk/metatiger/), a WWW resource devoted to the study of metabolic networks and their evolution. Using our metaSHARK software for metabolic annotation of genomes , we have annotated metabolic enzyme functions in over 500 species, including more than 100 eukaryotes. The WWW site has easy to use facilities for viewing and comparing the metabolic networks in different organisms via highlighted pathway images (using information from KEGG) and tables. metaSHARK provides genome annotations using raw nucleic acid sequences as input, predicting gene models using matching of profile hidden Markov models (HMMs), and is therefore able to annotate eukaryotes for which gene predictions are not available. This allows access to species for which only preliminary genome or EST sequence data exists, or for which gene structures are hard to predict owing to lack of suitable training information. We expect these features to be increasingly useful as new sequencing methods allow the sequencing of many more genomes for which detailed manual annotation will be impossible.

The main novelty of the metaTIGER site is the inclusion of significant amounts of evolutionary information. For each metabolic function (defined by Enzyme Commission (E.C.) number) the corresponding enzyme sequences were used to create a maximum likelihood phylogenetic tree, resulting in a comprehensive database of 2,257 trees. These trees were created from just the most conserved parts of the enzyme, and were limited to include only sequences which are very confident hits to the profile/HMM (E value $< 10^{-30}$). In addition, for each genome a (sub-)sequence was only associated with an E.C. number if it was the best match to the corresponding profile/HMM within the genome and was not a more confident match to any other profile/HMM. Thus the sequence sets for our trees eliminate the inclusion of paralogs as far as possible, and are designed to give the highest possible quality phylogenetic trees. The site contains facilities for viewing the trees using the state-of-the-art tree viewer iTOL . In addition there are tree query facilties which allow the user to search for trees with particular clade structures, for instance including potential horizontal gene transfers, or to find sequences suitable for concatenation to infer consensus organism phylogenies.

We have recently used the facilities described above to make the most comprehensive survey yet available of the horizontal transfer of metabolic genes between bacteria and unicellular eukaryotes. The 30 eukaryotic species with complete genome sequences considered derive from 10 eukaryotic genera, and reveal significant levels of confidently asserted transfers. The species set contains several significant groups of parasites of human and economic importance, and some of the transfers we find are potentially related to pathogenicity and the adaptation to a parasitic lifestyle.