Thieftest: a data-mining program for horizontal gene transfer detection

Ptitsyn A., Gamieldien, J. and Hide, W.
South African National Bioinformatics Institute,
ptitsyn@sanbi.ac.za

We have developed a data-mining program to search a whole bacterial genome for genes that may have been acquired by horizontal gene transfer. The program is based on the suggestion that results of a BLAST search can be interpreted as a measure of distance between a query sequence and a set of sequences. For example, if we have two contrasting sets of sequences, and a query sequence has a better BLAST search score for one set, then the query sequence may be phylogenetically closer to this set. Thus, all bacterial genomic sequences are supposed to score higher against a prokaryotic subset of Genbank than against a eukaryotic subset. Thus, if a bacterial sequence scores higher against a eukaryotic subset than against a bacterial one, this sequence may have been acquired by horizontal gene transfer. We have applied this program to search for possible cases of horizontal gene transfer between Mycobacterium tuberculosis and eukaryotic organisms (presumably human host). From over 3500 predicted TB genomic ORFs we have identified a few genes of possible eukaryotic origin. We have also used sensitive phylogenetic methods to verify the authenticity of a number of these candidates, adding further support to the validity of our approach. This program is implemented as a PERL script and requires BLAST II program to run.