Connecting the proteomes of model organisms

Classification of the protein universe

Tim Hulsen[1], Peter Groenen[2] and Martijn Huynen[1]

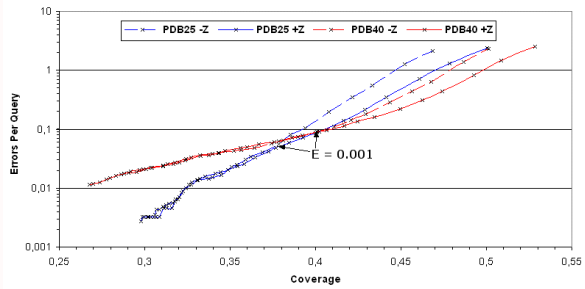
[1] NCMLS, p/a CMBI, Toernooiveld 1, 6525 ED Nijmegen
[2] NV Organon, Molenstraat 110, 5342 CC Oss

Abstract

In a large-scale computation, all of the currently known and predicted proteins have been compared through the Smith-Waterman algorithm with calculation of Z-values. This advanced algorithm is more precise than the more common BLAST and FASTA algorithms, which were used to build other protein relation databases. This database consisting of more than 100 proteomes will be used to create an index of intra- and interspecies relationships, by using clustering algorithms and methods to determine paralogs and orthologs. The outcome of this study will be combined with other information such as gene expression data.

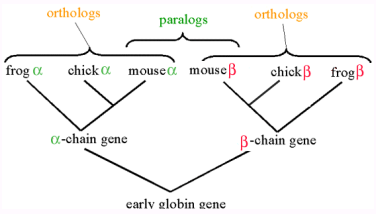
Smith-Waterman with Z-value calculation

Most DNA or protein comparison databases are built using algorithms like BLAST and FASTA, because of their speed. The Smith-Waterman algorithm used in this study is a slower but more advanced algorithm. The Z-value is an attempt to estimate the statistical significance of a Smith-Waterman dynamic alignment score (SW-score) through the use of a Monte-Carlo process. In the latter approach highly similar sequences are shuffled randomly 200 times and similarities determined and its significance measured to the already obtained sub-optimal alignments. Z-values offer a very precise definition of protein similarities and partly reduce the bias induced by the composition and length of the sequences (e.g. database size).
The added value of the Z-value was determined by comparing the hits from a normal Smith-Waterman with the hits from a Smith-Waterman with Z-value calculation. Using two ASTRAL SCOP (release 1.61) PDB sets, sets consisting of PDB protein sequences with 25% and 40% maximum structural identity, the numbers of proteins that are paralogs and occurring in the same SCOP class are calculated, while varying the E-value threshold. For an E-value above 10^{-3}, the Z-value calculation indeed gives less false positives and a better coverage compared to a Smith-Waterman without Z-value calculation.



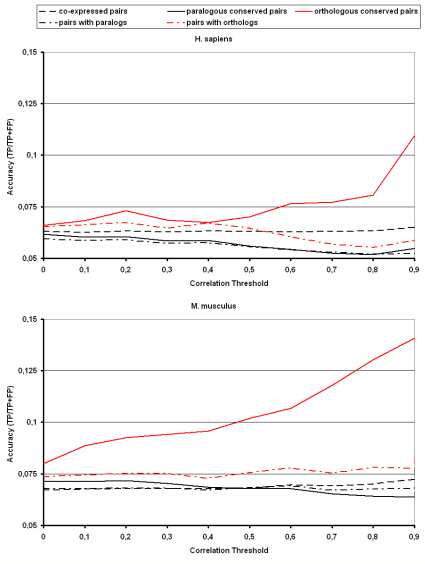
Orthology determination

Orthologs are genes in different species that evolved from a common ancestral gene by speciation. Normally, orthologs retain the same function in the course of evolution. Identification of orthologs is critical for reliable prediction of gene function in newly sequenced genomes.



A simple and fast method to determine orthologs is the 'best bidirectional hit': take the best hit for a particular species from a Smith-Waterman (or BLAST) comparison and see if this protein has the first protein as a best hit too. If this is the case, the two proteins are very probable orthologs.
For better orthology determination, several more advanced algorithms exist, like determination with use of phylogenetic trees. An important part of this project will be to find out what is the best method for orthology prediction.

Orthology combined with expression data



The orthologous (and paralogous) relationships between proteins determined by the outcome of our computation will be combined with information derived from expression data. One example is the comparison of these homologous protein pairs with co-expression of the corresponding genes.
For all pairs of genes in human and mouse, the correlation between their expression profiles was calculated. Then, while varying the minimal correlation coefficient, the accuracy was obtained by measuring the number of pairs with a correlation coefficient higher than the threshold that are on the same KEGG Pathway map (release 25). This was done for 'unrelated' co-expressed pairs, pairs that have orthologs in the other species, pairs that have paralogs, paralogous conserved pairs and orthologous conserved pairs. Only the latter group shows a linear relation between the co-expression and the accuracy. This could mean that orthology determination in combination with co-expression could be used for function prediction.