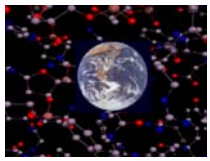
Benchmarking ortholog identification methods using function similarity

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

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



Abstract

We present a strategy to test the quality of ortholog identification methods. This strategy is based on the assumption that orthologs have a highly similar function. The conservation of function between proteins in different species can be measured by using the functional annotation provided with the protein sequences but also with functional data like co-expression data, protein interaction data and chromosomal position data. These data were used within the orthology benchmarking in a direct way, looking at only one protein in each species, and in a pairwise way, looking at two proteins in each species having some kind of relationship (e.g. co-expressed, interacting, neighboring). The benchmarking methods were used to test a number of ortholog identification methods, both methods that identify one-to-one orthologous relationships and methods that identify many-to-many relationships.

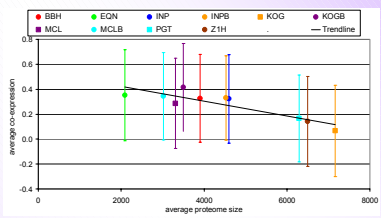
'Protein World' data set

For a fair comparison of all of the covered methods, the same data set was used at all time. This 'Protein World' data set was created by comparing all of the currently known and predicted proteins (SpTrEMBL) 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. The software package used to do the calculations is called 'Biofacet', formerly known as 'LASSAP' ('LArge Scale Sequence compArison Package'). The data set is available through the dutch bioinformatics portal site http://www.bioasp.nl. As for both human and mouse good expression data and other functional data were available, we used the orthologous relationships between these two species for our study.

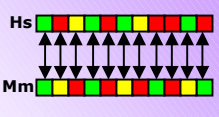
Compared ortholog identification methods

Best Bidirectional Hit (BBH), Equal SwissProt Names (EQN), InParanoid (INP), KOG (KOG), OrthoMCL (MCL), Phylogenetic Tree (PGT) and Z>100 (Z1H)

Direct conservation of function

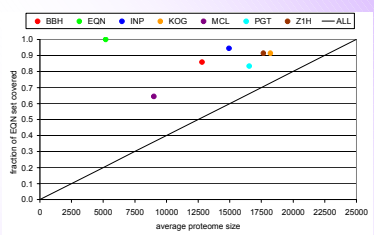


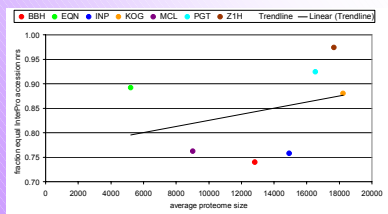
Co-expression



Equal SwissProt name

ABCD_HUMAN ~ ABCD_MOUSE



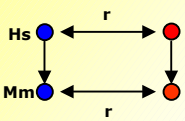


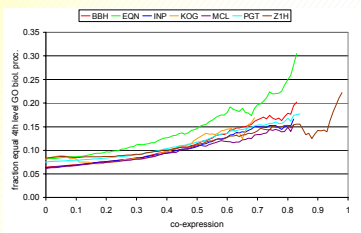
Equal InterPro acc. nr.

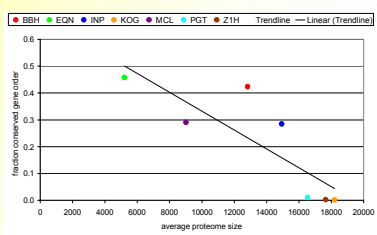
5H6_HUMAN (IPR000276) ~ 5H7_HUMAN (IPR000276)

Pairwise conservation of function

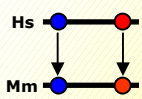
Conservation of co-expression







Conservation of gene order



Conservation of interaction