Conservation of chromosomal distance, relative orientation and co-expression of eukaryotic gene pairs

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

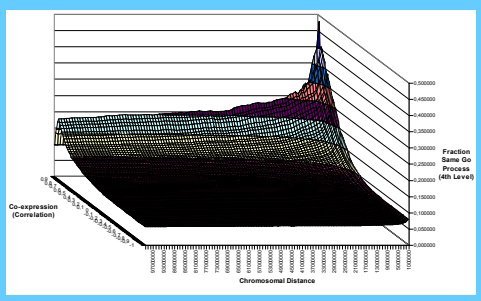
In a large-scale computation, all of the currently known and predicted proteins have been compared through the Smith-Waterman algorithm with calculation of Z-values. This advanced algorithm is more precise than the more common BLAST and FASTA algorithms, which were used to build other protein relation databases. This database called 'Protein World' is used here to define orthologous and paralogous relationships within the eukaryotic proteomes. We show that the paralogous and orthologous conservation of chromosomal distance, relative orientation and co-expression of gene pairs in eukaryotes can be used to improve function prediction.

Orthology determination within eukaryotes

Orthologies were determined by grouping all proteins over the 9 eukaryotic species covered in Protein World which have a Z-value above 20 compared to one of the human proteins, and having a region of homology larger than 50% of the query length. The resulting 24,263 groups were used to create ClustalW multiple alignments, and Neighbor-joining phylogenetic trees. An orthology determining algorithm then used the trees to define the orthologous relationships per species pair. To test the surplus value of our set compared to the NCBI KOG database, we compared the GO molecular function annotation of both protein sets. It turned out that 67% of the pairs in our set share a 4th level molecular function, whereas for the KOG pairs this is only 51%. Moreover, our pairwise relationships contain more proteins than are in the current KOG set for these 9 eukaryotes. Paralogous relationships within each species were determined by using the same parameters as mentioned above.

Using chromosomal distance and co-expression of gene pairs to predict their functional relation

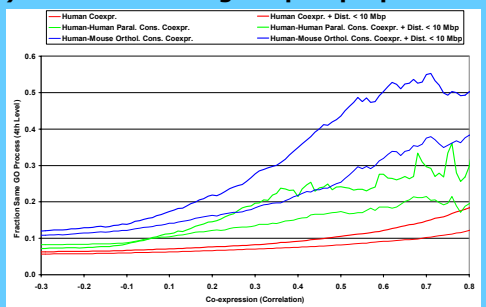
When two genes are co-expressed, they are likely to be involved in the same process. This means that if the function of a certain gene is unknown, the functional information of a co-expressed gene can be useful when annotating the novel gene. In prokaryotes, two genes that are in one operon often are involved in the same process too. In eukaryotes, operons are absent. However, the distance on the chromosome between two genes still gives some information about these genes being in the same biological pathway.



The figure shows how co-expression (x axis) and chromosomal distance (y axis) of gene pairs in human are related to the involvement of these two genes in the same biological process (z axis). The latter property is measured by using the GO biological process database. If the studied genes share a 4th level GO biological process, they are marked as being related.

Using the conservation of gene pair properties to improve functional predictions

The orthologous and paralogous relations that were determined for all eukaryotes (see above) were used to study both the paralogous (human-human) conservation and orthologous (human-mouse) conservation of gene pair properties.



If a co-expressed human gene pair has a paralogous gene pair that is co-expressed too, the chance that they are involved in the same biological process will increase (thin green line). It will be higher even more when both pairs are within a certain distance on the chromosome (here 10 million bp, thick green line). The results for orthologous human-mouse conservation are even better (blue lines). The red lines show the results for gene pairs without looking at conservation. The conservation of relative orientation (→←,←→,→→) seems to be important only when the chromosomal distance is very small, which is something that will be the subject of further research.