

## Abstract

In a large-scale computation, carried out by GENE-IT (HQ in Paris, France), 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** is being used to create an index of intra- and interspecies relationships, by using clustering algorithms and methods to determine **paralogs** and **orthologs**. It is available through the dutch bioinformatics portal site <http://www.bioasp.nl>.

The outcome of this study is being combined with other information such as gene expression and chromosomal position data. We are showing here that the paralogous conservation of chromosomal distance and coexpression between genes in human, together with the orthologous conservation of chromosomal distance, can be used to improve function prediction.



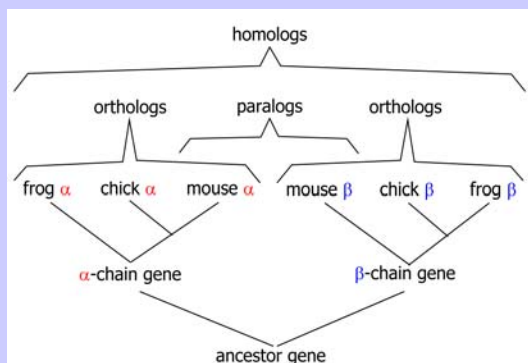
## Orthology determination within eukaryotes

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 COG 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 ( $Z > 20$ ,  $RH > 1/2 * Q$ ).

## Orthology

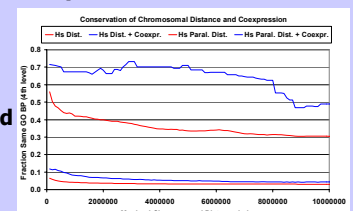
**Orthologs** are genes in different species that evolved from a common ancestral gene by **speciation**. Normally, orthologs retain the same function in the course of evolution. Identification of orthologs is critical for reliable prediction of gene function in newly sequenced genomes.



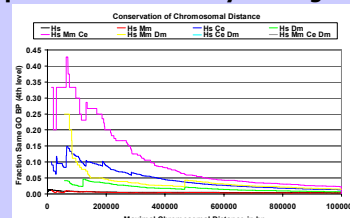
A simple and fast method to determine orthologs is the **'best bidirectional hit'**: take the best hit for a particular species from a Smith-Waterman (or BLAST) comparison and see if this protein has the first protein as a best hit too. If this is the case, the two proteins are very probable orthologs. However, several **more advanced** algorithms exist, like determination with use of **multiple alignments** and **phylogenetic trees**. Although methods like this are more time-consuming and intensive, they should provide a far better insight into gene evolution and orthology than the best bidirectional hit approach.

## Conservation of chromosomal distance and coexpression

The dataset mentioned above was used to test the **conservation of chromosomal distance** between gene pairs, and their **coexpression**. When only looking at human gene pairs, it is already clear that the shorter the distance between two genes and the larger their coexpression, the bigger the chance that these genes are involved in the same **biological process** (GO 4th level). But this number can be further improved using **paralogous conservation**, which means that there has to be a human paralogous gene pair which has the same properties for distance and coexpression.



Finally, **orthologous conservation** (human-mouse, human-worm and human-fly) shows effects that are quite alike. When only looking at chromosomal distance,



the use of multiple orthology datasets improves our accuracy from  $\sim 1\%$  up to **43%** for a combination of mouse and worm. The (mouse-)worm-fly combinations had too low numbers and were

left out of the graph. One of our future plans is to combine this with coexpression and things like conserved **relative orientation** ( $\rightarrow, \rightarrow, \rightarrow \leftarrow, \leftarrow$ ).