Predicting genome wide gene orthology and paralogy relationships

Ensembl provides highly integrated annotation on complete sequenced genomes, with a particular emphasis on vertebrates. Having an almost exhaustive gene set for each species allows us to run genome wide comparisons between the genes of each species.

We describe the method used to compare the different genes from all species at once, and how we infer paralogy and orthology relationship between each pair of genes between or within species. The knowledge of these putative homologues, especially orthologues, will greatly increase our understanding of the different genomes.

The pipeline that is used in the Ensembl framework to predict gene orthology/paralogy relationships is presented, as well as different measures of its accuracy. Compared to most of the current methods based around reciprocal best hits for pairs of species, the method presented relies on the phylogenetic tree reconstruction of gene families. This gene trees offer a global view of the evolutionary history of a gene in context of the related genes both from the same and the other species. They also capture ancient duplication events, extending the range of the estimated paralogies and hence improving gene inclusion.

We also present data analysis at genome scale such as distribution of duplication events along the species tree, comparisons of GO term distributions, and the identification of heterotachy.