Data file Format Descriptions

Pairwise association matrices

Pariwise association matrices provide strength of functional association between all available gene pairs in an organism. Individual matrices are given for each type of functional association (i.e. phylogenetic profiles, co-expression, clustering on the chromosome, etc.). For each type of association two matrices are given – raw scores, and association-rank rescaled scores. Matrices are given for download as ZIP-compressed text files.

- First line of each file specifies a tab-separated list genes (1 to n)
- Each subsequent line gives a tab-separated list of association strengths between k^{th} gene (where k is the line number) and all other genes in the organism. Diagonal matrix elements (i.e. association of k^{th} gene with itself are meaningless).

Orthology datasets

Custom BLAST-based orthology datasets are given for *E. coli* and *S. cerevisiae* genomes. Each line of the orthology file (ZIP-compressed text file) reports on a single orthology mapping: from a protein *x* in the target genome (*E. coli* and *S. cerevisiae*) to a protein *y* from a set of query genomes. The information is given in the following tab-separated columns:

- NCBI GI number of the protein *x*
- NCBI Taxonomy Id of the query genome (*Y*), followed by ':', followed by accession number of the orthologous protein *y* in that query genome
- Forward BLAST hit number – the rank of protein *y* in the result of a BLAST query looking for homologs of protein x in genome *Y*
- Forward BLAST hit E-value
- Forward BLAST hit score
- Reverse BLAST hit number – the rank of the protein *x* in the result of a BLAST query looking for homologs of protein *y* in the target genome.
- Reverse BLAST hit E-value
- Reverse BLAST hit score

Note that if gene *x* does not appear in the result of reverse query (*y* against target genome), all reverse BLAST values (hit number, E-value, score) are assigned 0.

Metabolic neighborhoods

Metabolic neighborhoods for enzymes of *E. coli* and *S. cerevisiae* metabolism are given in the following format:

Each line describes a neighborhood of a single metabolic enzyme. Neighborhood layers are separated by ';' symbol. The 0^{th} layer is the enzyme itself. Each gene in the subsequent layers is given in the following format:

$g(1,n_1, n_2,.., n_k)$

where "g" is gene name, and n_i specifies total number of gene pairs connected by i^{th} metabolite in the shortest path connecting neighborhood gene g with the missing enzyme.