MULTIGENOME MAGPIE

Terry Gaasterland and Alexander Sczyrba
Laboratory of Computational Genomics
The Rockefeller University, New York, New York, USA
{gaasterl,asczyrba}@genomes.rockefeller.edu

The MAGPIE genome analysis system provides a flexible environment to define and combine the results of different searches for one or many DNA or protein query sequences. We have augmented the MAGPIE system [GS96] to collect and analyze matches between query sequences and complements of proteins from multiple genomes. The resulting system, Multi-genome MAGPIE, enables us to pose questions about phylogenetic distributions of orthologs and paralogs for each ORF in a genome; to gather statistics about sets of ORFs with common patterns of genomic distribution. We using Multi-genome MAGPIE for the following purposes: enhanced functional annotation of genomes; clustering and annotation of full length cDNA sequences; identification of proteins shorter than 100 amino acids; identification of inconsistencies in previously annotated genomes. For each type of task, we have defined a canonical MAGPIE configuration.

The underlying data structure that we use to compute and query multi-genome views of sets of ORFs is the *genomic signature* [GR98a, GR98b, GR98c]. A genomic signature for an ORF is a vector of integers, each of which represents the degree of protein sequence similarity between the query ORF and the closest protein in a particular genome. The integer is computed by using any combination of thresholds on similarity score elements (*e.g,* P-value, percent identity, percent of query ORF involved in the similarity). Consider an *E.coli* ORF with very strong similarity to proteins in 6 bacterial genomes, strong similarity to proteins in 4 archaeal genomes, and no similarity to proteins from 2 eukaryotic genomes. Its genomic signature would be 111111222200, where "0" means "no match, and where the signature positions represent the bacterial, archaeal, and eukaryotic genomes in order. To study phylogenetic domain patterns, we can compute *phylogenetic domain signatures* from genomic signatures by adding each domain to a phylogenetic signature if the genome matches another genome in that domain. For example, the signature above becomes B (for Bacteria) at level-1 and BA (Bacteria+Archaea) at level-2. To refine the analysis, we augment genomic signatures with numbers of matches in each genome at each level to build *genome signature matrices.* For three levels and 12 genomes, each cell of a 3x12 genomic signature matrix contains the number of similar ORFs in a particular genome at a particular level. Our system is general and uses any source of pairwise similarities as input.

The specific tasks enabled by the augmented system are the following:

- Presentation of a multi-genome view of a coding sequence translated from a genome.
- Presentation of a multi-genome view of a region of DNA (contig, cosmid, chromosome, genome).
- Computation and querying of *genomic signatures* for a set of ORFs.
- Prediction of co-expressed genes based on clusters of genomic signatures.

We have populated Multi-genome MAGPIE views of available public genomes with a cross-genome database of pairwise comparisons built with PSI-BLAST. See http://genomes.rockefeller.edu/magpie/mtub for an example. The cross-genome database of pairwise matches was built in collaboration with Andrej Šali and Roberto Sánchez, who are using the data to drive the construction of 3D models of proteins across all fully sequenced genomes (MODBASE) [SS98]. Future work includes merging Multi-genome MAGPIE and MODBASE.

References

[GR98a] Gaasterland, T. & Ragan, M. (1998a). Constructing multi-genome views of whole microbial genomes. *Journal of Microbial and Comparative Genomics,* **3**, 305–312.

[GR98b] Gaasterland, T. & Ragan, M. (1998b). A prokaryotic view of the yeast genome. *Journal of Microbial and Comparative Genomics,* **3**, 219–235.

[GR98c] Gaasterland, T. & Ragan, M. (1998c). Phyletic and functional patterns of ORF distribution among prokaryotes. *Journal of Microbial and Comparative Genomics,* **3**, 199–217.

[GS96] Gaasterland, T. & Sensen, C. W. (1996). Fully automated genome analysis that reflects user needs and preferences — a detailed introduction to the MAGPIE system architecture. *Biochimie,* **78**, 302–310.

[SS98] Sánchez, R. & Šali, A. (1998). Large-scale protein structure modeling of the *Saccharomyces cerevisiae* genome. *Proc. Natl. Acad. Sci. USA,* **95**, 13597–13602.