Phylogenetic Pattern Recognition of Fungal Protein Families

Sangsoo Kim and Gir-Won Lee

Department of Bioinformatics, Soongsil University, Seoul

Great sequence diversity among fungal species makes studying fungal comparative genomics interesting. It is witnessed by the growing number of fungal genome projects. Currently tens of fungal genome sequences and the protein sequence sets thereof are available for public use. Systematic phylogenetic analyses of these protein sequences may allow the identification of the sequence features specific to subphyla of the fungal phylum. This may lead to a deeper understanding of the evolution of not only fungi but also other eukaryotes.

We have set out to cluster the fungal protein sequences from completed genomes, identifying orthologous and paralogous groups. Predicted proteome sets for 45 fungal completed genomes have been downloaded from NCBI, Broad Institute, and TIGR. The sequence sets comprised of Saccharomycotina, Schizosaccharomyces, Pezizomycotina, and Basidomycota, as well as Encephalitozoon and Rhizopus, but excluded any subspecies. A total of 429,878 protein sequences were initially screened for matches with protein domain databases such as CDD, Pfam, SMART, and COG. Among them, 267,988 were partitioned into 14,023 non-exclusive domain groups. This calculation was performed by employing a PC cluster of around 100 nodes at Korea Bioinformation Center (*http://www.kobic.re.kr*), Korea Research Institute of Bioscience and Biotechnology, Taejon, KOREA. This process inevitably puts different families of proteins that share the same domain into one group. In order to split such different families into different sets, we employed BLASTCLUST (*http://www.ncbi.nlm.nih.gov/Web/Newsltr/Spring04/blastlab.html*), which clusters a group of sequences based on pair-wise BLAST similarity scores and sequence coverage. Compared to stricter methods of detecting orthologs such as SYNERGY [1] or HomoloGene [2], this procedure has pros and cons. Although it does not guarantee for orthologous clustering, the paralogous groups can be clustered together. In other words, the paralogous relationships between orthologous sets are preserved in a cluster. We plan to set up a process that dissects the cluster based on species phylogenetic tree, as implemented in SYNERGY.

As a test case, we looked into inorganic pyrophosphatase family. In *S. cerevisiae*, both cytoplasmic (IPP1) and mitochondrial (PPA2) forms are known. An unrooted neighbor-joining (NJ) tree showed a cluster of 40 proteins (cytoplasmic) and another cluster of 16 proteins (mitochondrial). The proteins from species like Encephalitozoon, Neurospora, Aspergillus and others formed long branches. Predictions of subcellular localization of these proteins were ambiguous. It is also known that Encephalitozoon is a parasite with a mitochondrion. In HomoloGene database that clusters protein sequences from not only fungi but also animal and plants together, split this protein family into three clusters. Based on the dissimilarity between each pair of these sequences, they were mapped to a multidimensional space through a process known as multidimensional scaling. Its principal coordinates are chosen in such a way that the first component (PC1) captures the largest fraction of the total variance. The cytoplamic and mitochondrial proteins were located at the opposite sides of each other in PC1, while the proteins forming long branches in the NJ tree were found in the middle. We scanned the multiple sequence alignment and identified 41 residue positions whose diversities among these proteins were congruent with the clustering pattern in PC1. When those residues were mapped to a crystal structure reported for *S. cerevisiae*, they were clustered at four different regions in the periphery of the three-dimensional structure. It appears that the mutations that led to the divergence between cytoplasmic and mitochondrial forms of inorganic pyrophosphatases avoided the active site and were not randomly distributed along the sequence but clustered in several three-dimensional regions.

Here we have shown that combination of phylogenetic analysis and pattern recognition technique such as multidimensional scaling could offer interesting insight into the understanding protein sequence evolution.

References

[1] I. Wapinski, A. Pfeffer, N. Friedman, and A. Regev *Bioinformatics* **23**, i549 (2007).

[2] K.D. Pruitt, D.R. Maglott *Nucleic Acids Research* **29**, 137 (2001).