When and where do protein folds come from? an evolutionary view

Song Yang[1], Phil Bourne[2]

Keywords: SCOP, fold evolution, disulfide bond, phylogenetic tree

Compared with the fast-growing numbers of protein sequences, the number of possible protein structures is quite limited. It has been proposed that there may be only 400-4000 protein folds existing in all organisms. Thus structure represents a more conserved and alternative measure by which evolution can be studied. Thus as more and more complete genomes of organisms in every kingdom are sequenced, it is possible to compare protein folds across various genomes in the tree of life to gain an evolutionary view.

We have used existing databases (SUPERFAMILY[1] and PEDANT[2]), which contain fold assignments for complete genomes (based on the protein fold classification defined by SCOP[3] and using the homology search methods PSI-BLAST or HMMs), to obtain protein fold counts in 17 Archaea, 123 Bacteria and 16 Eukaryota[4]. The accompanying graph shows a Venn diagram of fold distribution (defined as the second level in SCOP) in the three kingdoms. Among the total 753 folds, only one fold, d.199, is unique to Archaea. Since the only representative of d.199 in PDB is a transcription factor from bacteriophage T4, it is likely that Archaea obtained this fold from a virus. Therefore, Archaea have hardly any unique folds, if any. In contrast, Bacteria and Eukaryota appear to have invented a substantial number of new folds since the divergence of the three kingdoms. About 40% of the folds (24 of 61) unique to Eukaryota contain disulfide bonds, whereas the percentage of folds containing disulfide bonds found only in Bacteria is 31% (5 of 16), those folds found in both Eukaryota and Bacteria is 22% (36 of 165), and those common to all kingdoms is only 6% (30 of 491). This suggests that many protein domains stabilized by disulfide bonds emerged after the atmosphere became more oxygen-rich, under which conditions Eukaryota generated more disulfide bond-containing folds than the other groups.

On another front, we were able to build a phylogenetic tree based on the existence and abundance of folds in each genome. The tree is similar and comparable to phylogenetic trees built with gene sequences or other features. Detailed analyses of certain folds were also performed.

1 Department of chemistry and biochemistry, University of California, San Deigo
2 San Diego Supercomputer Center, Department of Pharmacology, UCSD, Burnham Institute

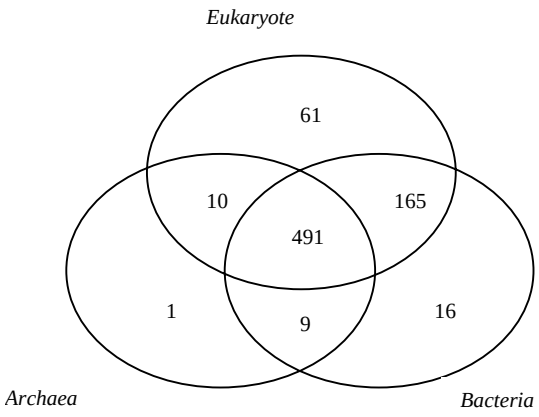

Figure 1: Protein fold distribution among three kingdoms

References

[1] Gough, J., Karplus, K., Hughey, R. and Chothia, C. 2003. Assignment of Homology to Genome Sequences using a Library of Hidden Markov Models that Represent all Proteins of Known Structure. *J. Mol. Biol*., 313: 903-919

[2] Frishman, D etc. 2003. The PEDANT genome database. *Nucleic Acids Research* 31: 207-211.

[3] Lo Conte L., Brenner S. E., Hubbard T.J.P., Chothia C., Murzin A. 2002. SCOP database in 2002: refinements accommodate structural genomics. *Nucl. Acid Res.* 30: 264-267.

[4] Gustavo Caetano-Anollés and Derek Caetano-Anollés. 2003. An Evolutionarily Structured Universe of Protein Architecture. *Genome Res.* 13: 1563-1571.