PROTEIN EVOLUTION
The Origin and Evolution of the Protein Universe

DR. L. ARAVIND
Investigator, Computational Biology Branch
National Center for Biotechnology Information, NLM, NIH

Thursday, June 26, 2003,
Lecture: 2:00 pm to 3:00 pm Room NE-301, Med/Dent Bldg. (Max ~50 people)

Georgetown University Medical Center
Washington D.C. 20057

Please notify PIR soon via e-mail at jlc57@georgetown.edu, subject: "Protein Evolution" and indicate the session(s) you are interested in attending.

Ever since the pioneering work of Zuckerkandl and Pauling, protein sequences and structures have been used extensively to infer organismal phylogeny and to predict biochemical functions. However, the origin and early evolution of proteins, and the diversification of protein families during the divergence of the major lineages of life, are poorly understood. The recent explosion of complete genome sequences and protein structures provides the necessary material to construct and test hypothesis regarding these issues in protein evolution. As these models cannot be easily reconstructed using conventional phylogenetic methods alone, diverse methodologies were applied to glean the requisite information. These include: 1) Detection of distant relationships between proteins through a combination of sequence and structural analysis, and classification of the protein world into monophyletic assemblages of domains. 2) Identification of sequence or structural features that are shared derived characters for particular clades within these monophyletic assemblages of domains. 3) Comparisons of the phyletic distributions of domains to derive the domain complements in the last common ancestors of the organismal lineages under consideration. 4) Detection of anomalies in tree topologies and phyletic distributions of particular proteins to decipher evolutionary forces involved, such as, lateral transfer and gene loss.

As a result, the most conserved sets of proteins and constituent domains, traceable to the Last Universal Common Ancestor (LUCA) of all life forms were identified. Through a comparison of the homologous domains in this set, the pre-LUCA stages of protein evolution were reconstructed. One of the conclusions that became apparent was that even before the extant translation apparatus was in place, complex protein domains, resembling extant forms, were already being synthesized. These investigations also lead to the hypothesis that early enzymatic domains with specific activities emerged from generalized RNA (ribozyme) interacting modules. The ribozymes were eventually displaced through the acquisition of strategically placed residues that allowed the protein to acquire catalytic activity.

The subsequent evolution of proteins was explored by identifying certain general principles that explain the origin of several distinct classes of domains. These include, the emergence of lineage specific alpha-helical domains through the duplication and collapse of simple helical segments, fusion of pre-existing domains into single folding units, and stabilization of incipient protein folds through metal chelation and disulfide bond...
formation. Additionally, various tendencies in the adaptive radiation of protein domains, such as lineage specific expansions, colonization of new functional niches, and emergence of novel domain architectures following massive lateral transfer events were identified. Use of these tendencies to explain the origin of eukaryotes and particular biological systems during the subsequent diversification of the eukaryotes is illustrated. Application of these studies in identifying and predicting functions of previously uncharacterized components of the eukaryotic chromatin, RNA metabolism- and signal transduction- systems is demonstrated.