Exploration of Novel Methods to Visualize Genome Evolution

NASA's Applied Information Systems Research Program (NNG04GP90G)

Goal of our research is to develop and test approaches that can be used to visualize and dissect the mosaic nature of genomes. This will allow tracing histories of individual genes, detection of co-evolving traits, and correlation of the molecular record with the fossil and geological records. In the past we have developed tools that map the support of individual gene families for the different possible phylogenetic trees. However, because of the huge number of possible trees topologies, this approach cannot readily be extended to the depiction of many genomes. To facilitate the processing of phylogenetic information, we will organize the data from analyses of multiple genomes into matrices. For each gene family we will utilize either the significantly supported bipartitions, or the support for the different quartets that can be formed by selecting four of the analyzed species. All of these data matrices will be of very high dimension ([number of orthologous genes] x [number of possible tree topologies / quartets / bipartitions]).

We will explore different tools to analyze these matrices. First, we will consider the local linear embedding algorithm (LLE), which maps each point in the high dimensional space into a point in two-dimensional space via an encoding based on a covariance matrix of the distances to a selected set of nearest neighbors. This algorithm performs well in providing low dimensional projections of high dimensional data retaining the essential structure of the original data. Another algorithm that will be considered is the self-organizing map (SOM) algorithm. This neural network-based algorithm attempts to detect the essential structure of the input data based on the similarity between the points in the high dimensional space. We will investigate the usefulness of these and other approaches (principal component analysis (PCA), multidimensional scaling (MDS), support vector based kernel PCA, and ISOMAP) to produce informative two-dimensional maps depicting phylogenetic relationships among genomes, and we will develop tools to make gene types and tree topologies readily recognizable in each map. We aim for analyses and diagrams that depict gene families with similar histories as neighboring, whereas genes with different histories are classified into separate clusters. We will explore the utility of the different approaches using well-studied test cases: the eukaryotic genome where currently three large groups of genes are recognized based on their different evolutionary history (those form the host genome and those from the endosymbionts that evolved into mitochondria and plastid), and selected bacterial genomes where genes involved in the same function of metabolic pathway often have the same evolutionary history.