Whole Genome Phylogenies via Singular Value Decomposition (SVD)

Karen Moffett, Mathew Kay, Steve Baker, and Gary Stuart (LSSTUAR@SCIFAC.INDSTATE.EDU)
Department of Life Sciences, Indiana State University, Terre Haute, IN 47802

An SVD-based analysis [1] utilizing all possible overlapping tripeptides was used to produce accurate phylogenetic trees from whole genome data. Both gene trees and species trees were derived from the same analysis, providing straightforward biomolecular species definitions based on total genome content. In one application, the complete set of mitochondrial proteins from 95 metazoan organisms was used to generate an 8,000 word by 1230 protein data matrix representing all proteins in terms of their tripeptide frequencies. Following SVD and dimension reduction, pairwise cosine values were calculated for each pair of proteins from the derived "protein vector" matrix and used to generate a distance matrix to be used in the NEIGHBOR program of PHYLIP. A Neighbor-Joining (NJ) gene tree was then produced. In this tree, the vast majority of proteins were well recognized and placed within large monophyletic or nearly monophyletic groups. To generate the species tree, the 12 or 13 protein vectors from each organism were added to make supervectors that represent species in 100 dimensional space. Again, following the production of a cosine-based distance matrix, an NJ tree was generated. In this tree, the major metazoan phyla were correctly grouped, and the phyla themselves were reasonably placed. Within the chordate branch, the birds, reptiles, fish, non-primate mammals, and primate mammals are all well grouped and reasonably placed. Within primates, humans are grouped with chimps, then gorilla, then gibbon, then orangutan, and then baboon. This order corresponds to an accepted order [2], except that gibbon and orangutan should perhaps be reversed. Also notable is that hippos branch with whales: a relatively new hypothesis supported by recent molecular data [3]. In addition, chondricthyes branch within the osteichthyes, instead of as a sister group. This supports another recent hypothesis based on molecular data [4]. In a second application, 5 whole bacterial genomes were analyzed in a similar way (over 7,800 proteins). The derived species tree for these 5 organisms was found to be correct. The two archaebacteria (*M. jannashii* and *P. horikoshii*) were seen to be more related to each other than either was to the three eubacteria (*H. pylori, H. infuenzae*, and *B. burgdorferi*). Also, within the eubacteria, H. pylori and *H. influenzae* were seen to be more related to each other than either was to *B. burgdorferi*, as expected. Further refinements of the method are planned and should improve the accuracy with which the relative functional meaning of both genes and species are estimated.

1. Berry, et al. 1999, *SIAM Review* 41:2, 335-362.
2. Horai et al. 1995, *Proc. Natl. Acad. Sci. USA* 92: 532-536.
3. Xu et al. 1996, *Mol. Biol. Evol*. 13: 1167-1173.
4. Rasmussen and Arnason 1999, *J. Mol. Evol.* 48: 118-123.