

Phylogenomics of metazoans based on SVD analysis

Arun Seetharam and Gary W. Stuart
 Department of Life Sciences, Indiana State University.

Introduction:

Many molecular phylogenies are based on sequences sampled from one gene or a few conserved genes. In recent times large numbers of genome sequencing projects have generated a wealth of sequence data providing us the opportunity to consider building phylogenies using whole genomes. Traditional methods rely on alignment for deriving phylogenetic relatedness, but this becomes impractical when we consider comparing whole genomes. We recently developed a phylogenetic method that provides accurate comparisons for a high fraction of sequences within whole genomes without the prior explicit identification of orthologous or homologous sequences. This method was employed here in an attempt to identify the closest relative to the vertebrates. A recent study [2] proposed tunicates as the closest relative to the vertebrates, followed by cephalochordates. To do this they compared 146 conserved genes, represented in 14 different deuterosomes and 24 slowly evolving outgroup species. We compared the whole genome protein sequences of these deuterosomes to check whether this relationship is supported within genomic scale data using a different phylogenetic method.

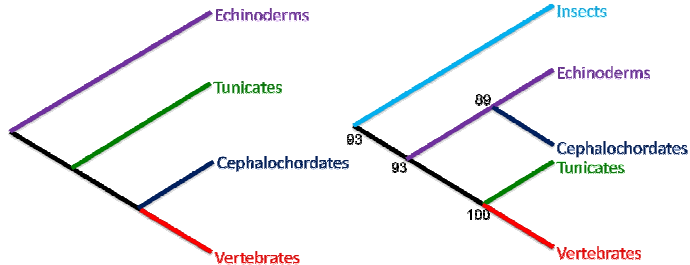


Fig 1: Traditional deuterosome evolution model based on complexity of organisms

Fig 2: Delsuc² et al., suggested tunicates to be closest to vertebrates

Results:

The results obtained based on our preliminary work suggested cephalochordates to be the closest to vertebrates and tunicates represent the earliest chordate lineage.

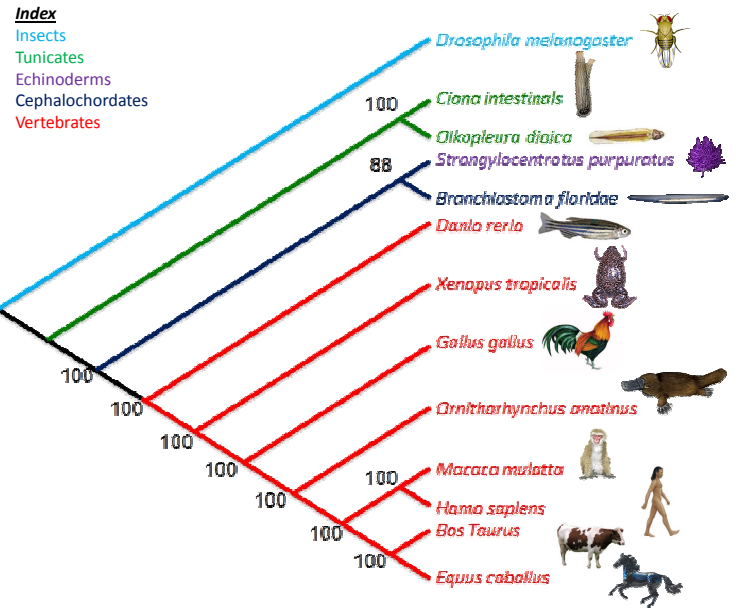


Fig 3: Preliminary results from 13 deuterosomes, using whole genome protein sequences. Through SVD analysis

Method

Data Collection: Whole genome sequences were downloaded from NCBI, ENSEMBL and JGI. All important species that had their full genome sequenced and are important to understand vertebrate phylogeny were included for the analysis. Only predicted protein sequences were used.

Matrix generation: The program AACODE was used to convert the sequences into a tetrapeptide frequency matrix, by setting word size to 4.. Each protein was represented as a column in the matrix within which each of the 160,000 possible tetrapeptides was enumerated. Each column could also be represented as a vector in multi-dimensional space.

Singular Value Decomposition: The frequency matrix was then subjected to SVD analysis using an SVD [1] program that decomposed the matrix into 3 different component matrices designated as U, Σ and V. The leading matrix U defines conserved motifs as correlated peptides, the Σ diagonal matrix defines the singular values, and the last matrix V provides the protein vector definitions.

Pairwise Distance measurement and generating tree: All protein vector definitions of a species are summed together to get the species vector definitions, and the relatedness between them is measured as the angle made between the vectors. This angle is used to compute pairwise distance using the COSDIST program. Phylip-Neighbor was then used to make trees from the pairwise distance values

Conclusions:

Our results were not in accordance with what was expected after including newly sequenced species of Appendicularian (*Oikopleura dioica*). Instead it strongly agreed with the traditional tree suggesting cephalochordates to be the closest to vertebrates. However, Echinoderms and cephalochordates shared a single ancestor as suggested by Delsuc et al. [2]. The branching within vertebrates supports the idea that the tree generated is of high quality since it agreed with well established relationships with strong support values.

References:

- Berry, M.W., Large scale singular value computations. *Int. J. Supercomput. Appl.* 6,13-49, 1992/
- Delsuc F., Brinkmann H., Chourrout D. and Philippe H., Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Science.* 439, 965-968, 2006.
- Halanych, K.M., The phylogenetic position of the pterobranch hemichordates based on 18S rDNA sequence data. *Mol. Phylogenet. Evol.* 4, 72-76, 1995.
- Stuart G.W., Moffett K. and Leader J.J., A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Mol Biol Evol.* 19, 554-562, 2002