Spatial representation of amino acid composition divergence in homologous protein families

Lucas Carrijo de Oliveira[1], Néli José da Fonseca Júnior[1], Lucas Bleicher[1],

1 Federal University of Minas Gerais

Abstract

Homologous protein families can be assessed by multiple sequence alignments (MSA), wherein each column represents an evolutionarily corresponding position among homologous proteins. Conserved positions, meaning invariable sites, indicate evolutionary constraints in amino acid substitutions, generaly due to structural and/or functional importance of such positions. Besides the fully conserved ones, there are some positions that are specificaly conserved in functional subclasses eventualy present in a family. In the same way, as one moves toward a phylogenetic tree, from root to leaves, some residues appears as being specificaly conserved in each clade, while others remain variable or unspecificaly conserved. By representing each residue (here calling residue a given amino acid in a specific position, like "His37") by the set of all sequences in a MSA having such a residue, and calculating the conditional probabilities of finding all other ones given the presence of that residue (e.g., probability of finding Asp71 in sequences having His37), it is possible to compare all possible sets of sequences on the basis of their amino acids composition. The present work introduces a distance based method to represent, in the N-dimensional space, the evolutionary divergence of amino acid composition in homologous protein families. For each residue, the method takes a specific sub-alignment (e.g, the subset of sequences in a MSA having that residue) and considers each column as a 20-dimensional vector, being each dimension the conditional probability of finding, at that position, each of the 20 amino acids. This way, each sub-alignment is represented as a set of 20-dimensional points in space. Two sub-alignments can than be compared by calculating the root mean square deviation (RMSD) between these two sets of points. An all against all distance matrix is generated and, by singular value decomposition (SVD), it is possible to define N-dimensional spatial coordinates from this distance matrix. By ploting these coordinates in a tridimensional Cartesian plane, one can visualize the pattern of amino acid composition divergence in homologous protein families, from more conserved to more specific residues, passing toward variable or unspecificaly conserved ones. Colouring points by frequency in MSA of their respective residues helps visualization of such an effect.

Funding: CAPES, CNPq