PFStats: A tool for protein analysis by decomposition of residue coevolution networks and amino acid reduced alphabets applications

Neli Fonseca, Lucas Bleicher, Marcelo Querino

Universidade Federal de Minas Gerais

Structural and functional insights about protein families can be obtained by amino acid conservation and correlation analysis. Futhermore, experimental research has suggested that protein folding can be achieved with fewer characters than the 20 naturally occurring amino acids. Our group has recently proposed a method to obtain functional sub-class determinants in protein families, called Decomposition of Residue Coevolution Networks (DRCN). DRCN is a sequence based method for analysis of protein families represented by multiple sequence alignments. We present a software for protein family analysis using DRCN, conservation analysis, alphabet reductions, and automatic annotation search. The algorithms were grouped in order to have a robust and intuitive application to the analysis of homologous proteins. The DRCN analysis consists of a unique required input file, a multiple sequence alignment (MSA), besides that a PDB file can be also used to visualize the results in the structure. The MSA quality is a crucial factor to achieve better results with the methodology, therefore, a filtering step is available to maximize its representativeness by removing fragments, poorly aligned sequences and redundancy. We have studied four protein family domains: lysozyme C/Alpha-lactoalbumin, phospholipases A2, nitrogen regulatory protein PII, and the DNA binding domain of the nuclear receptors IV; three MSAs aproaches extracted from Pfam and 19 amino acids reducted alphabets from literature. We have found insights about catalytic and binding sites in all of then. There's also information related to secondary structure, the hydrophobic putative channel, and dimerization sites. By looking for the anti-correlated edges, we could find a residue or a group of residues that separates two or more sub-classes. That's the case of the C122 in the phospholipase A2, this node form an anti-correlated hub that connects every community. Its presence occurs in 217 sequences, all from *Oikopleura dioica*, and all without the phospholipase catalyctic activity. The uses of reduced alphabet in DRCN analysis usually increase the number of residues in each community and in the most cases maintaining a consistent hypothesis for their biological role. But in these cases, nuclear receptors IV study, the uses of a reduced alphabet can hide clusters that share common positions with another community

Funded by FAPEMIG and CAPES