Identification of conserved sites in protein folds

ID - 211

Bordo Domenico[1], Demartini Paolo[1], Polacco Moshe[1]

[1]Bioinformatics and Structural Proteomics, National Cancer Research Institute

Motivation

Homologous protein maintain their three-dimensional topology even if sequence similarity tends to be obliterated by the increasing evolutionary distance. However, homologous proteins may display regions in which substantial amino acid conservation is observed. There are two main reasons for this conservation. Residues may be important for structural reasons, or may be involved in the function carried out by the protein. In the HSSP database the structural information obtained by the delucidation of the protein three-dimensional structure has been combined with that obtained by standard sequence homology search performed by programs such as BLAST.

Therefore, each HSSP entry represent a molecular family of protein sharing the same three-dimensional topology, but that may contain several subfamilies of paralog proteins having distinct functionalities. Aim of this work is to develop a tool that integrates structural and sequence information of a given molecular family having known three-dimensional structure. In particular, the average amino acid variation due natural evolution (evolutionary clock hyphosesis) is considered and substracted by the amino acid variability observed in a specific protein family, in order to identify sites having reduced evolutionary rate as possible candidates for sites bearing either structural or functional relevance.

Methods

The proposed approach is implemented by a suite of PERL scripts and of C++ programs that obtain information form the HSSP database and use this to query the SwissProt database. The evolutionary distance separating a pair of orthologous protein sequences is obtained by querying an appropriate databases in which an evaluation of the evolutionary time between the two genomes has been implemented. The approach can be applied to any entry of the HSSP database. An user-friendly graphic interface is presently under development.

Results

Phylogenetic analysis are based on a stochastic model of sequence diversification, under which protein evolution is considered a set of identical and independent Markov chains, each one describing the evolution of a single position in the sequence. The availability of a multiple sequence aligment contained in HSSP, and the estimation of the evolutionary distance from the main sequence (of which the 3D structrue is known), allows to identify sites which display, coherently in a subset of the HSSP family, a degree of conservation significantly higher than that expected under the Markov chain hypothesis. The method is intended to help the identification of functionally relevant sites in protein, and to suggest paralogous relathionships within the same molecular family, possibly due to a functional diversification occurred as single step during the evolution of a molecular subfamily.

Email: domenico.bordo@istge.it