ABSTRACT

Motivation: Evolutionary relationships of proteins have long been derived from the alignment of protein sequences. But from the view of function, most restraints of evolutionary divergence operate at the level of tertiary structure. It has been demonstrated that quantitative measures of dissimilarity in families of structurally similar proteins can be applied to the construction of trees from a comparison of their three-dimensional structures. However, no convenient tool is publicly available to carry out such analyses.

Results: We developed STRUCLA (STRUcture CLAssification), a WWW tool for generation of trees based on evolutionary distances inferred from protein structures according to various methods. The server takes as an input a list of PDB files or the initial alignment of protein coordinates provided by the user (for instance exported from SWISS PDB VIEWER). The user specifies the distance cutoff and selects the distance measures. The server returns series of unrooted trees in the NEXUS format and corresponding distance matrices, as well as a consensus tree. The results can be used as an alternative and a complement to a fixed hierarchy of current protein structure databases. It can complement sequence-based phylogenetic analysis in the 'twilight zone of homology', where amino acid sequences are too diverged to provide reliable relationships.

Availability: The service is free for all users and available at http://asia.genesilico.pl/strucla/

Contact: asia@genesilico.pl; iamb@genesilico.pl

INTRODUCTION

Protein evolution is a topic of keen interest because of its importance for prediction of protein function. Many functional prediction methods rely on identification and characterization of similarity of sequence or structure between the protein of interest and those for which functional information is available (Mushegian et al., 1997; Irving et al., 2001). However, the identification of similarity is frequently not enough to assign a predicted function if an uncharacterized protein is related to a group of functionally different proteins. The functional prediction can be greatly improved by focusing on the phylogenetics of the family—i.e. inferring the history of divergence and convergence events by mapping the known functions onto the evolutionary tree (Eisen, 1998). Commonalities and variations observed in the family of proteins provide the basis for their classification and suggest the relative importance of amino acid residues and individual structural elements - information that can be used in modeling and structure-based prediction of protein function. This is of key interest in structural genomics, which aims to first determine the structure of proteins, and then investigate their function later (Irving et al., 2001). Evolutionary analysis of protein structures and sequences can be also used to infer the hypothetical ‘minimal ancestor’ of the family (Murzin, 1998) and the allowed pathways of elaboration of the common fold. This approach is helpful in modeling of folding pathways, which may be of particular value in understanding the principles that govern the folding of the polypeptide chain (Efimov, 1997).

SEQUENCE AND STRUCTURE-BASED TREEING

Phylogenetic methods have been criticized because of their dependence on multiple sequence alignments that are not always reliable and unbiased. Although some families of homologous proteins retain rather high sequence similarity, many diverge to such a degree that the statistical measure of the evolutionary relatedness based on sequence similarity becomes uncertain (Rost, 1997). The sequence variations within a family reflect the restraints of the tertiary structures: apart from the catalytic or binding residues, conserved amino acids are most often in the common core. It is now known that there are many distant relationships that can only be identified through
structure comparisons (Chothia and Lesk, 1986) and structural alignments have been long used as a ‘golden standard’ for improving sequence alignments (Gerstein and Levitt, 1998). However, there are many families comprising proteins that have diverged beyond any significant sequence similarity and still retain the architecture of their ancestral fold, like globins or PD-(D/E)XK nuclease (Bujnicki, 2000)—in such cases it is not feasible to infer their evolutionary history even if the multiple sequence alignment is obtained from superimposed structures.

It is known that the percent sequence identity and RMSD have shortcomings as measures of sequence and structural similarity (Levitt and Gerstein, 1998). In a seminal publication, Chothia and Lesk (1986) shown that there is a strong correlation between these two values, but it is not linear. Wood and Pearson (1999) demonstrated that the correlation between structural similarity and sequence similarity becomes linear when the ‘raw’ percent identity and RMSD are replaced by the statistical Z-scores. This correlation is however statistically less reliable when the percentage sequence identity is lower than 30%. Recently, it has been proposed that if the sequence similarity for a pair of proteins with known structure is measured as the mean distance between the sequences in the subsets of sequence space compatible with their structures, then even RMSD and identity become linearly related (Koehl and Levitt, 2002).

Johnson et al. (1990) derived a measure that gives a quantitative indication of three-dimensional structural divergence over evolutionary distances. They used the tertiary and primary structures of the immunoglobulin fold domains, globins, cytochrome c-related proteins, serine proteases, eye-lens crystallins and Rossmann fold proteins to demonstrate that there exist a nearly linear relationship between the ‘traditional’ distance measure determined from amino acid sequences and their distance measures based on structure comparison. They concluded that the trees derived from proteins structures may be less reliable than those from sequences of closely related homologs, but they can provide relationships between distantly related proteins where sequence comparisons are unreliable or generate statistically insignificant results. This strategy has been successfully applied to infer evolutionary history of proteins with statistically insignificant sequence similarity, yet similar structures suggesting common origin, namely the immunoglobulin fold superfamily (Halaby et al., 1999) and the PD-(D/E)XK superfamily of nuclease (Bujnicki, 2000). Other methods have been also developed to compare and classify protein structures and alternative measures of structure-based measures of protein divergence have been proposed (Schulz, 1977; Gerstein and Altman, 1995; Grishin, 1997; Yang and Honig, 2000; Carugo and Pongor, 2001, 2002). While some of them rely on traditional structure comparison and involve calculation of normalized or ‘improved’ RMSD, others compare intrinsic features and do not require structural superposition. The review of the methodology is beyond the scope of this article, for an excellent critical review on various protein structure similarity measures see (May, 1999).

STRUCLA SERVER
STRUCLA (STRUcture CLAssification) is a WWW-based e-mail server implementing several methods for evolutionary inference based on protein structures. The list of available methods for protein structure comparison is continuously expanded - for the up-to-date list of available algorithms and methods, please consult the STRUCLA website http://asia.genesilico.pl/strucla/.

The program reads in a multiple structure file in the PDB format and writes an output file (sent to the e-mail address provided by the user) containing all pairs of superimposed structures in a PDB format, the set of distance matrices in the PHYLIP format, and the unrooted tree files in the NEXUS format. The input file can include multiple structures or, alternatively, PDB accession numbers of proteins of interest. Such file can be generated using any molecular graphics program, which can open and save more than one set of coordinates at a time. For that purpose, we recommend using the SwissPDBViewer program (http://www.expasy.ch/spdbv/) (Guex and Peitsch, 1997) that offers an option of semi-automatic superposition and editing of protein coordinates, which can be used to limit the input data to the truly superimposable domains.

For these methods, which require comparison of the atomic coordinates, the optimal superposition is sought separately for all pairs in the dataset, i.e. no unique reference structure is used. This is because for highly diverged protein families the definition of a ‘mean structure’ makes little sense due to the great diversity in peripheral elements and problems with definition of the common core. Such problems in structure superposition have been widely studied (Godzik, 1996; Feng and Sippl, 1996) and cannot be fully solved by the existing automatic methods.

In the prototype of STRUCLA, we implemented the LGScore program (http://www.sbc.su.se/~are/lgscore/; Cristobal et al., 2001).

The superposition is used to compute a distance matrix according to the selected method. The distance matrix is included in the output file and can be used as an input file for other programs. STRUCLA implements the Neighbor-Joining method of Saitou and Nei (1987) to infer the unrooted tree, which is exported in the NEXUS format. If more than one method is selected, STRUCLA also infers the consensus tree, using the CONSENSE program from the PHYLIP package (Felsenstein, 1989). The majority-rule consensus tree consists of monophyletic groups that occur as often as possible among the primary
trees obtained with different methods. The consensus tree has at each branch a number indicating how many times the group which consists of the species in that branch occurred among the primary trees. This value may be used as an indicator of the ‘robustness’ of the topology of the consensus tree. An example input file and graphical presentation of the output, as well as guidelines for the users are provided on the STRUCLA website.

Structure-based evolutionary analysis at the level of superfamilies may help discriminate between the cases of convergence and divergence of protein structures, which is often not possible from studying pairwise superpositions and alignments (Schulz, 1977). For instance, the structure-based evolutionary analysis of S-adenosylmethionine-dependent methyltransferases and typical Rossmann-fold proteins allowed to determine the monophyletic origin of the methyltransferase family and verify several conflicting hypotheses inferred from earlier studies (Bujnicki, 1999). STRUCLA may be used for rationalizing classification of protein structures in databases, comparison and selection of templates for homology modeling, and benchmarking various measures of protein structure similarity and their relationship to similarity calculated from sequence comparisons.

ACKNOWLEDGEMENTS

This work was supported by KBN (grant 3P04A01124). J.M.B. is an EMBO/HLDI Young Investigator and a fellow of Foundation for Polish Science.

REFERENCES


