BIOINFORMATICS APPLICATIONS NOTE Vol. 17 no. 4 2001
Pages 375–376

PALI: a database of alignments and phylogeny of homologous protein structures

*S. Sujatha, S. Balaji and N. Srinivasan**

Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India

Received on August 11, 2000; revised on November 13, 2000; accepted on November 20, 2000

ABSTRACT

Summary: PALI is a database of structure-based sequence alignments and phylogenetic relationships derived on the basis of three-dimensional structures of homologous proteins. This database enables grouping of pairs of homologous protein structures on the basis of their sequence identity calculated from the structure-based alignment and PALI also enables association of a new sequence to a family and automatic generation of a dendrogram combining the query sequence and homologous protein structures.

Availability: PALI can be accessed on the web at: http://pauling.mbu.iisc.ernet.in/~pali

Contact: ns@mbu.iisc.ernet.in

INTRODUCTION

Variation in the amino acid sequences of evolutionarily related proteins is restrained by retention of the fold and function (e.g. Chothia and Lesk, 1986; Murzin *et al.*, 1995). This feature is exploited in the comparative modelling wherein a three-dimensional model of a protein is generated on the basis of related proteins of known structure (Srinivasan *et al.*, 1996; Sanchez and Sali, 1997). One could learn about variations in structural features as a function of sequence similarity within homologous proteins (e.g. Flores *et al.*, 1993) and such knowledge could be used in improved modelling. In order to aid performing such analyses we have set-up a database of **P**hylogeny and **ALI**gnment of homologous protein structures (PALI).

CONSTRUCTION OF PALI

The current release (1.2) of PALI consists of 604 families of homologous proteins involving over 2700 protein domains of known structure (Balaji *et al.*, 2001). The dataset has been largely derived by consulting the SCOP database (release 1.50) (Murzin *et al.*, 1995). The structure of PALI, and access details are summarised in Figure 1. Apart from the multiple structural alignment of members in a family, every member in a family is structurally aligned, using the program STAMP (version 4.2) (Russell and Barton, 1992), with every other member in the family (pairwise alignment). The dissimilarity in the homologous protein structures has been quantified by a structural distance metric defined by Johnson *et al.* (1990). Manual intervention was necessary to check the quality of the alignments. Some of the alignments, especially those involving distantly related proteins, were refined manually. The rigid-body alignments of homologous structures, performed using STAMP (Russell and Barton, 1992) and presented in PALI, are found to be highly similar to the alignments made by COMPARER (Sali and Blundell, 1990) which uses multistructural features and relationships.

Sequence similarity-based and structural similarity-based dendrograms have been generated for each family using PHYLIP package (Felsenstein, 1995). A comparison of these two kinds of dendrograms suggests that only for a small number of families the dendrograms differ (Balaji and Srinivasan, 2001).

A variety of features of PALI, especially the availability of pairwise alignments as well as sequence and structural similarity-based dendrograms, may be viewed as complementary to the other homologous protein structural databases such as HOMSTRAD (Mizuguchi *et al.*, 1998).

WEB ACCESS TO PALI

The following features are integrated into the web interface of PALI.

(1) The alignments and two types of dendrograms can be accessed family-wise. A search facility is also available to identify the family of interest.

(2) Pre-calculated structure-based and sequence-based dendrograms for all the families are available as *PostScript* files.

(3) A search tool is available to identify pairs of proteins characterised by a given range of sequence identity or structural similarity.

(4) PSI_BLAST (Altschul *et al.*, 1997) is integrated with PALI to help associating a new sequence to

*To whom correspondence should be addressed.

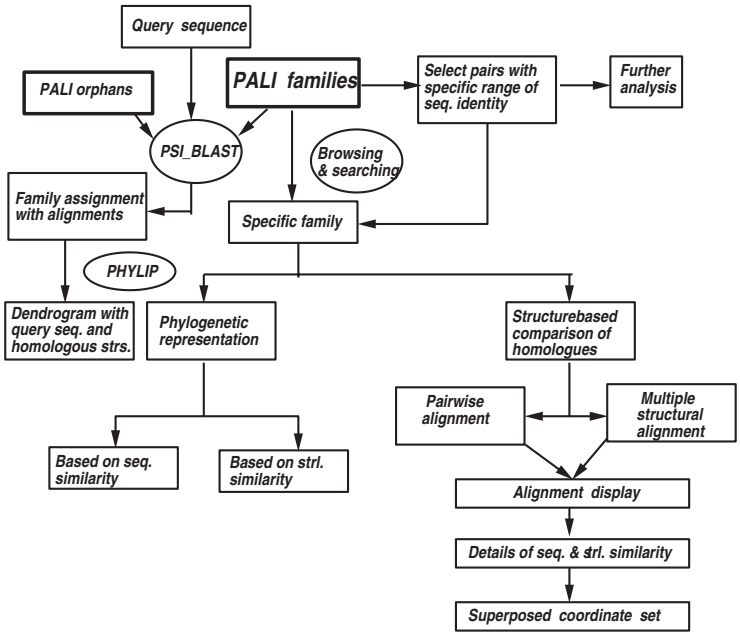

Fig. 1. An overview of the organisation and access to PALI.

one of the existing families. More than 600 single member families ('orphans') are also included in PSI_BLAST search.

(5) A dendrogram generating tool is also available that can automatically incorporate a query sequence on to the phylogenetic relationship of an existing homologous protein family. A combination of PSI_BLAST and PHYLIP is used for this purpose.

ACKNOWLEDGEMENTS

We thank Mr Sai Chetan Kumar for his help in updating the PALI database. S.Sujatha and S.Balaji thank the Wellcome Trust, UK and CSIR, India for financial support. This work is supported by a Senior Fellowship grant to N.Srinivasan from the Wellcome Trust, UK.

REFERENCES

Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhanng,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Balaji,S. and Srinivasan,N. (2001) Use of a database of structural alignments and phylogenetic trees in investigating the relationship between sequence and structural variability among homologous proteins. *Protein Eng.*, in press.

Balaji,S., Sujatha,S., Sai Chetan Kumar,S. and Srinivasan,N. (2001) PALI—a database of phylogeny and alignment of homologous protein structures. *Nucleic Acids Res.*, **29**, 61–65.

Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.

Felsenstein,J. (1995) PHYLIP (Phylogeny Inference Package). Version 3.57c. Department of Genetics, University of Washington, Seattle, USA.

Flores,T.P., Orengo,C.A., Moss,D.S and Thornton,J.M. (1993) Comparison of conformational characteristics in structurally similar protein pairs. *Protein Sci.*, **2**, 1811–1826.

Johnson,M.S,, Sutcliffe,M.J. and Blundell,T.L. (1990) Molecular anatomy: phyletic relationships derived from three-dimensional structures of proteins. *J. Mol. Evol.*, **1**, 43–59.

Mizuguchi,K., Deane,C.M., Blundell,T.L. and Overington,J.P. (1998) HOMSTRAD: a database of protein structure alignments for homologous families. *Protein Sci.*, **7**, 2469–2471.

Murzin,A.G., Brenner,S.E., Hubbard,T. and Chothia,C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. *J. Mol. Biol.*, **247**, 536–540.

Russell,R.B. and Barton,G.J. (1992) Multiple protein sequence alignment from tertiary structure comparison: assignment of global and residue confidence levels. *Proteins*, **2**, 309–323.

Sali,A. and Blundell,T.L. (1990) Definition of general topological equivalence in protein structures. A procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J. Mol. Biol.*, **212**, 403–428.

Sanchez,R. and Sali,A. (1997) Advances in comparative protein-structure modelling. *Curr. Opin. Struct. Biol.*, **7**, 206–214.

Srinivasan,N., Guruprasad,K. and Blundell,T.L. (1996) In Sternberg,M.J.E. (ed.), *Protein Structure Prediction—A Practical Approach.* Oxford University Press, Oxford, pp. 111–140.