486–488 *Nucleic Acids Research, 2003, Vol. 31, No. 1*
DOI: 10.1093/nar/gkg063

Integration of related sequences with protein three-dimensional structural families in an updated version of PALI database

V. S. Gowri, Shashi B. Pandit, P. S. Karthik, N. Srinivasan* and S. Balaji

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

Received August 22, 2002; Revised September 25, 2002; Accepted October 2, 2002

ABSTRACT

The database of Phylogeny and ALIgnment of homologous protein structures (PALI) contains three-dimensional (3-D) structure-dependent sequence alignments as well as structure-based phylogenetic trees of protein domains in various families. The latest updated version (Release 2.1) comprises of 844 families of homologous proteins involving 3863 protein domain structures with each of these families having at least two members. Each member in a family has been structurally aligned with every other member in the same family using two proteins at a time. In addition, an alignment of multiple structures has also been performed using all the members in a family. Every family with at least three members is associated with two dendrograms, one based on a structural dissimilarity metric and the other based on similarity of topologically equivalenced residues for every pairwise alignment. Apart from these multi-member families, there are 817 single member families in the updated version of PALI. A new feature in the current release of PALI is the integration, with 3-D structural families, of sequences of homologues from the sequence databases. Alignments between homologous proteins of known 3-D structure and those without an experimentally derived structure are also provided for every family in the enhanced version of PALI. The database with several web interfaced utilities can be accessed at: http://pauling.mbu.iisc.ernet.in/~pali.

INTRODUCTION

Analysis of a set of similarly folded proteins with distinct amino acid sequences, such as homologues, can help in identifying residues and regions of polypeptide chains that are likely to be important in the formation and stability of the fold (1–3). An organised database of three-dimensional (3-D) structure-based sequence alignments of homologous protein domain families should be helpful in protein design, engineering, studies on evolution and in generating rules for improved procedures for comparative modelling. Association of amino acid sequences, represented as ORFs in genomic data, with the appropriate structural families provides a quick picture about the gross three-dimensional shape of the proteins encoded in genomes.

About couple of years ago, we had developed a database referred to as PALI representing Phylogeny and ALIgnment of homologous protein structures (4). PALI contains structure-based sequence alignment of homologous proteins of known structures. We have also derived dendrograms relating the members of a family using dissimilarity measure based on 3-D structures apart from the traditional sequence similarity based measure. For the pairs of homologous proteins with sequence identity lower than about 30%, direct relationship between sequence and structural dissimilarity is not valid (3). Hence the structure-based measures are more suited than the sequence-based measures for generating dendrograms which models evolutionary relationships. This is consistent with the fact that base sequences of genes coding for homologous proteins shows the maximum variability and 3-D structures of homologues having high similarity with amino acid sequence showing intermediate extent of variability.

PALI is being updated regularly to reflect the ever increasing number of proteins of known structure (5). This paper reports an update of PALI made very recently. In addition to the update, in the present release of PALI, we have enhanced the information content in the database by integrating the protein domain structural families with the corresponding homologous sequences obtained from sequence databases. Thus, every family in PALI is now enriched with a substantial volume of sequence information from proteins without an experimentally derived structure.

UPDATED VERSION OF PALI

The list of protein domains in every homologous family has been derived by consulting the latest version of SCOP database (6). While both X-ray and NMR structures are included in PALI, care is taken not to include multiple entries of the structures of same protein (such as structures in different crystal forms or in different functional states). The coordinate

*To whom correspondence should be addressed. Tel: +91 803942837; Fax: +91 803600535/+91 803600683; Email: ns@mbu.iisc.ernet.in

Present address:
P.S. Karthik, Biomedical and Biotechnological Center, Faculty of Chemistry and Mineralogy, University of Leipzig, Johannisallee 29, 04103 Leipzig, Germany

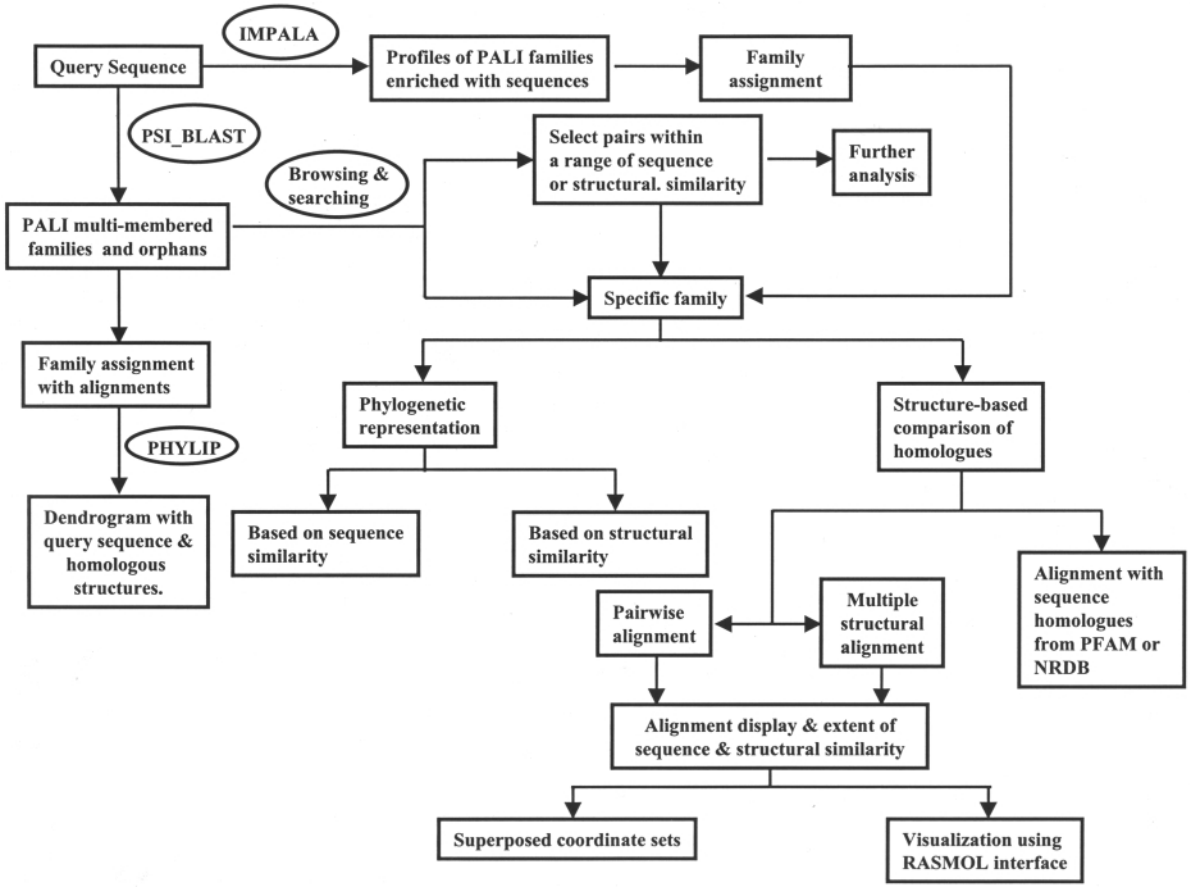

Figure 1. Organisation and access of PALI.

sets of domain regions of interest are gleaned from the protein structural entries in the protein data bank (7). Rigid body superposition of 3-D structures of all possible pairs of protein domains in a family as well as the multiple structural superposition has been performed using the program STAMP (8). The structural dissimilarity measures between any two protein domains in a family have been calculated using a modification of the metric proposed by Levitt and Gerstein (9) and Johnson *et al.* (10). The information about these metrices is provided in the PALI website and in our previous publications (4,5). Dendrograms have been generated, using the PHYLIP package of programs (11) on the basis of structural dissimilarity measures and sequence dissimilarity measures calculated for topologically equivalent residues defined by the structural deviation cut-off of 3 Å between C^{α} atoms.

PALI database is integrated with a web interface (12), which enables a user to probe the database conveniently. Apart from the alignments and dendrograms, it is possible to extract the superimposed set of atomic coordinates of homologous protein domains as well as visualization of the overlayed structures using an interface of RASMOL (13). Figure 1 shows the overall organization and capabilities of web access to PALI. Apart from various search tools, PSI_BLAST (14) and IMPALA (15) have been integrated with the web interface in order to enable a user to search the sequences of proteins of known structures in PALI and in the family profiles (see below), respectively. It is also possible to associate a query sequence with a PALI family and to automatically generate a dendrogram integrating the query with the members of the corresponding PALI family.

The present version of PALI (release 2.1) consist of 1661 families, which includes 817 orphans (single member families). Among 844 families with at least two members in each family, two membered families have highest representation (326). These 844 families consist of as many multiple structural alignments and 34 632 pairwise alignments calculated by considering only two protein domains at a time. All these structural superposition resulted in 2 805 450 residue–residue alignments. There are 1036 dendrograms in the present release of PALI coming from the families with three or more members.

ENHANCED INFORMATION CONTENT IN PALI USING HOMOLOGOUS SEQUENCES

In many of the sequence-based families, such as those in PFAM (16,17), with at least one member in the family with known 3-D structure, the majority of the homologous proteins in the family have no experimentally derived 3-D structure available. Thus, there is a need to increase the information content in every multiple structure-based sequence alignment

in PALI by integrating the PALI family with homologous sequences. This new feature has been incorporated in PALI starting from the present release of the database.

Almost all the single and multi-membered PALI families are related to the corresponding family in the PFAM database using our approach employed previously (18). For this purpose, Position Specific Scoring Matrices (PSSMs are also referred to as profiles) have been derived for all the families in PFAM. Association between a PALI family and a PFAM family has been established primarily by querying every protein in the PALI database and in the database of PFAM family profiles using the program IMPALA (*E*-value cut-off: 0.00001) (15). Alignment between PALI and PFAM family members have been provided for most of the PALI families. However, in the difficult cases such as a PALI family is split into more than one PFAM family, a search has been made for the homologues of the PALI family in the non-redundant database (NRDB) of protein sequences using PSI_BLAST (*E*-value cut-off: 0.0001). The alignment thus obtained has been provided in the PALI database. A total of about 251 000 sequences have been associated with the structural families in PALI.

PSSMs and Hidden Markov Models (HMMs) have been generated for every PALI family enriched with a large number of homologous sequences. As the information content in these profiles is much larger than that of pure structural families it could be expected that these profiles are sensitive in detecting distant homologues. PALI web site integrated with IMPALA software enables a user to search a query sequence in the data set of PALI structures-sequences combined profile.

OUTLOOK AND DATA ACCESS

Apart from the major increase in the number of protein families and alignments in PALI, incorporation of the homologous sequence information provides a more complete picture about the extent of diversity and abundance in various families. This enhanced information content should enable robust deductions of relationship between sequence profiles and 3-D structural variability. Such deductions should help in the better understanding of evolution of various protein families as well as in the development of improved methods for protein fold recognition and comparative modelling.

PALI database can be accessed at http://pauling.mbu.iisc.ernet.in/~pali. PALI PSSMs and HMMs can be obtained freely by the academic community from the authors upon request.

ACKNOWLEDGEMENTS

We thank the Bioinformatics centre of our institute for providing convenient access to protein databank files. S.B.P. and S.B. are supported by Council of Scientific and Industrial Research, India. P.S.K. was supported by the Wellcome Trust, UK. This research is supported by the Department of Biotechnology, India and by the award of International Senior Fellowship in Biomedical Sciences to N.S. from the Wellcome Trust, UK.

REFERENCES

1. Chothia,C. and Lesk,A.M. (1986) The relation between the divergence of sequence and structure in proteins. *EMBO J.*, **5**, 823–826.
2. 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.
3. 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.*, **14**, 219–226.
4. Balaji,S., Sujatha,S., Kumar,S.S.C. and Srinivasan,N. (2001) PALI—a database of Phylogeny and ALIgnment of homologous protein structures. *Nucleic Acids Res.*, **29**, 61–65.
5. Balaji,S., Sujatha,S., Aruna,S., Mhatre,N.S. and Srinivasan,N. (2002) PALI (Release 1.3) *Nucleic Acids Res.*, http://www3.oup.co.uk/nar/database/summary/274.
6. 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.
7. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
8. 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.
9. Levitt,M. and Gerstein,M. (1998) A unified structural framework for sequence comparison and structure comparison. *Proc. Natl Acad. Sci. USA*, **95**, 5913–5920.
10. 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.
11. Felsenstein,J. (1995) PHYLIP (Phylogeny Inference Package) version 3.57c. Department of Genetics, University of Washington, Seattle, USA.
12. Sujatha,S., Balaji,S. and Srinivasan,N. (2001) PALI—a database of alignments and phylogeny of homologous protein structures. *Bioinformatics*, **17**, 375–376.
13. Sayle,R.A. and Milner-White,E.J. (1995) RasMol: biomolecular graphics for all. *Trends Biochem. Sci.*, **20**, 374–376.
14. 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.
15. Schaffer,A.A., Wolf,Y.I., Ponting,C.P., Koonin,E.V., Aravind,L. and Altschul,S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15**, 1000–1011.
16. Sonnhammer,E.L.L., Eddy,S.R., Birney,E., Bateman,A. and Durbin,R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res.*, **26**, 320–322.
17. Bateman,A., Birney,E., Durbin,R., Eddy,S.R., Howe,K.L. and Sonnhammer,E.L.L. (2000) Pfam protein families database. *Nucleic Acids Res.*, **28**, 263–266.
18. Pandit,S.B., Gosar,D., Abhiman,S., Sujatha,S., Dixit,S.S., Mhatre,N.S., Sowdhamini,R. and Srinivasan,N. (2002) SUPFAM—Database of potential protein superfamily relationships derived by comparing sequence-based and structure-based families: implications for structural genomics and function annotation in genomes. *Nucleic Acids Res.*, **30**, 289–293.