284–285 *Nucleic Acids Research, 1999, Vol. 27, No. 1*

Database of protein sequence alignments: PIR-ALN

Geetha Y. Srinivasarao*, Lai-Su L. Yeh, Christopher R. Marzec, Bruce C. Orcutt, Winona C. Barker and Friedhelm Pfeiffer[1]

Protein Information Resource, National Biomedical Research Foundation, 3900 Reservoir Road N.W., Washington, DC 20007, USA and [1]GSF-Forschungszentrum f. Umwelt und Gesundheit, Munich Information Center for Protein Sequences, am Max-Planck-Institut f. Biochemie, am Klopferspitz 18, D82152, Martinsried, Germany

Received September 1, 1998; Revised September 24, 1998; Accepted October 22, 1998

ABSTRACT

The Protein Information Resource (PIR) has been maintaining a database of curated protein sequence alignments since 1991. The collection includes superfamily, family and homology domain alignments. CLUSTAL V/W is used to generate multiple sequence alignments and ALNED, an interactive alignment editor, is used to check and correct them. The database has helped in classifying sequences, in defining new homology domains, and in spreading and standardizing protein names, features and keywords among members of a family or superfamily. The ATLAS information retrieval system can be used to browse and query the PIR-ALN alignments. The quarterly and weekly updates can be accessed via the WWW at http://www-nbrf.georgetown.edu/pir/

INTRODUCTION

PIR-ALN is a database of curated and annotated protein sequence alignments derived from the PIR-International Protein Sequence Database. The database includes alignments of protein sequences as superfamilies, families and homology domains. Sequences belong to the same homeomorphic superfamily if they are homologous from end-to-end (1). Each superfamily is further classified into families containing sequences that are at least 45% identical. Many protein sequences are composed of a number of distinct functional regions or 'domains', or multiple copies of the same domain. The sequence segments corresponding to the same homology domain in two or more superfamilies are extracted and aligned to form the homology domain alignments in PIR-ALN.

DATA SELECTION

The selection of data is tied very closely to the task of classifying sequences into families and superfamilies based on sequence comparision (2). Closely related sequences are first clustered into families and then the families are clustered into superfamilies. As a first step in clustering the incoming sequences, our collaborators at Munich Information Center for Protein Sciences (MIPS) dynamically maintain a database of FASTA scores from searching every sequence against the PIR-International Protein Sequence Database (3). High-scoring sequences are examined for percent identity and length of overlap. Software has been developed to determine if a new sequence belongs to an existing family or if a new family must be created. Using this approach, 95.2% of the sequences in PIR have been clustered at the family level. About 30% of the sequences fall into single-member families. For groups that have at least two members, multiple sequence alignments have been generated. The PROT-FAM database of family alignments is available for browsing and searching at MIPS http://www.mips.biochem.mpg.de (4).

Sequence families are then clustered into superfamilies and family, superfamily, and domain alignments are constructed at the PIR. An overview of the process is shown in Figure 1. The longest sequence from each PROT-FAM family is used to generate a new database called FAMBASE. Using this database improves the sensitivity of FASTA searching because distantly related sequences can be found among the top scores. FASTA results of searches against FAMBASE are analyzed to cluster families into superfamilies based on length of overlap and percent identity.

The DOMAINDB database contains the sequence segments represented in all the homology domain alignments in PIR-ALN. This database is searched to screen new sequences for already defined homology domains.

Once the sequences to be clustered are identified, the CLUSTALW (5) program is used to generate multiple sequence alignments. Since computer-generated alignments are not always biologically correct, ALNED, an interactive alignment editor has been developed in-house to view, edit and update the alignments.

CURRENT STATUS AND DATABASE ACCESS

The alignments in PIR-ALN contain a selection of sequences both to keep the alignments at a reasonable size and to ensure that there is no bias towards a group that has many sequences. Superfamily alignments are made when the superfamily has at least two members from different families. The PIR-ALN database has alignments of each type of homology domain defined as a feature in the PIR Protein Sequence Database.

For some superfamilies and homology domains with a large number of sequences that are highly divergent, several alignments containing representative sequences have been constructed. Some examples are immunoglobulin homology, SH3 homology

*To whom correspondence should be addressed. Tel: +1 202 687 2121; Fax: +1 202 687 1662; Email: geetha@nbrf.georgetown.edu

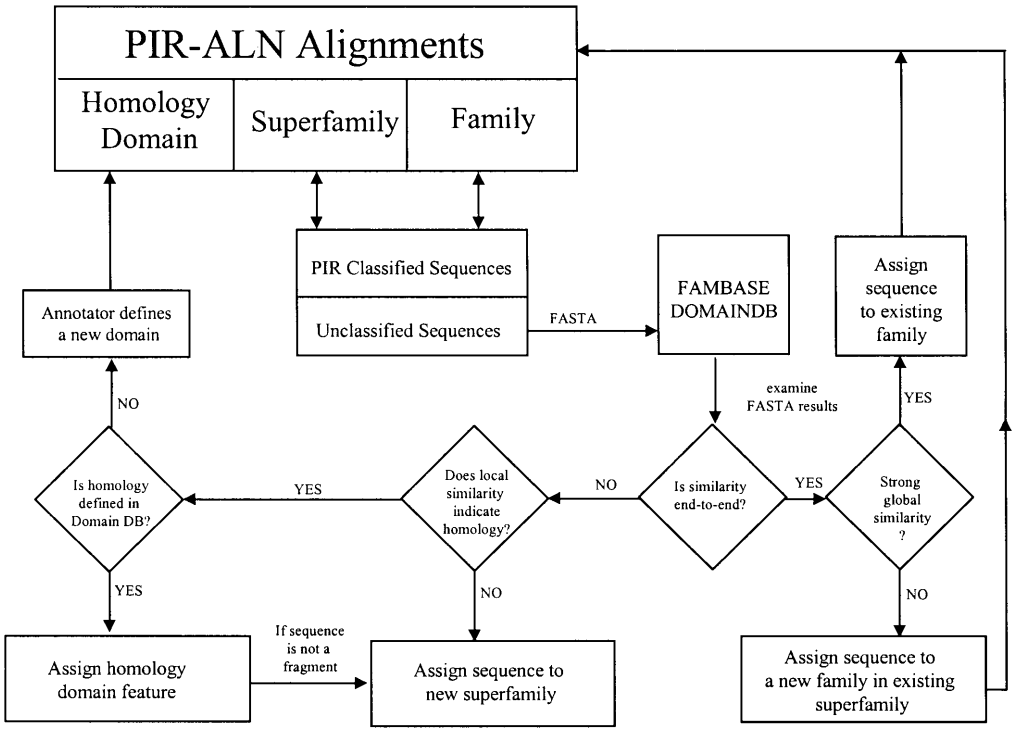

Figure 1. Flow of information to the PIR-ALN database.

Table 1. URLs for additional information on the PIR-ALN database

Information	URL
Database description	http://www-nbrf.georgetown.edu/pir/alndb.html
Statistics	http://www-nbrf.georgetown.edu/cgi-bin/nbrfbase
PIR search page	http://www-nbrf.georgetown.edu/pir/searchdb.html
Superfamily document	http://www-nbrf.georgetown.edu/pir/doc/sfdef.html
PIR-ALN search page	http://www-nbrf.georgetown.edu/nbrf/getaln.html
ATLAS CD distribution	http://www-nbrf.georgetown.edu/pir/atcd.html
ATLAS manual	http://www-nbrf.georgetown.edu/pir/doc/atlas.html

and kinase-related transforming protein superfamily. Currently, PIR-ALN has over 3500 alignments with >1000 superfamily and >350 homology domain alignments.

The URLs for current statistics, documentation of different fields, sample entry and database access are included in Table 1. The alignment database has been integrated with the ATLAS multidatabase information retrieval system, developed at PIR, which provides full access to the data.

The quarterly and weekly updates of the PIR-ALN alignment database can be accessed via the WWW. PIR-ALN is included on the 'Atlas of Protein and Genomic Sequences' CD-ROM available from the PIR-International centers in the US, Europe and Japan. It can also be obtained by anonymous FTP from the PIR FTP site at nbrf.georgetown.edu, directory [anonymous.pir.alignment].

The PIR-ALN database can be accessed on the PIR Web site in two ways. From the PIR entry request page, the PIR sequence entry will cross-reference a PIR-ALN entry if the sequence is a member of the retrieved alignment. Alternately, one can access the alignments directly through the PIR-ALN request page. The members and classification fields are hypertext linked to the PIR sequence database, so the user interacts with both the databases.

ACKNOWLEDGEMENTS

This work has been supported by NLM grant LM05798. The authors thank PIR staff for their contributions of alignments to the database and Katie Sidman and Desiree Goins for administrative support and help with Web page design. Work by MIPS was supported by grants from the Bundesministerium f. Bildung, Forschung und Technologie (BMBF, FKZ 0311670, 01KW9703/7) and the European Commission (BIOCT-CT-96110).

REFERENCES

1 Barker,W.C., Pfeiffer,F. and George,D.G. (1995) In Atassi,M.Z. and Apella,E. (eds), *Methods in Protein Structure Analysis*. Plenum Press, New York, pp. 473–481.
2 Barker,W.C., Garavelli,J.S., McGarvey,P.B., Marzec,C.M., Orcutt,B.C., Srinivasarao,G.Y., Yeh,L.-S.L., Ledley,R.S., Mewes,H.-W., Pfeiffer,F., Tsugita,A. and Wu,C. (1999) *Nucleic Acids Res.*, **27**, 39–43.
3 Mewes,H.W., Albermann,K., Heumann,K., Liebl,S. and Pfeiffer,F. (1997) *Nucleic Acids Res.*, **25**, 28–30.
4 Mewes,H.W., Hani,J., Pfeiffer,F. and Frishman,D. (1998) *Nucleic Acids Res.*, **26**, 33–37.
5 Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) *Nucleic Acids Res.*, **22**, 4673–4680.