BIOINFORMATICS

Vol. 15 no. 5 1999
Pages 382–390

PIR-ALN: a database of protein sequence alignments

Geetha Y. Srinivasarao, Lai-Su L. Yeh, Christopher R. Marzec, Bruce C. Orcutt and Winona C. Barker

Protein Information Resource (PIR), National Biomedical Research Foundation, 3900 Reservoir Road NW, Washington, DC 20007, USA

Received on August 16, 1998; revised on January 19, 1999; accepted on February 26, 1999

Abstract

Motivation: *The Protein Information Resource (PIR) maintains a database of annotated and curated alignments in order to visually represent interrelationships among sequences in the PIR-International Protein Sequence Database, to spread and standardize protein names, features and keywords among members of a family or superfamily, and to aid us in classifying sequences, in identifying conserved regions, and in defining new homology domains.*

Results: *Release 22.0, (December 1998), of the PIR-ALN database contains a total of 3806 alignments, including 1303 superfamily, 2131 family and 372 homology domain alignments. This is an appropriate dataset to develop and extract patterns, test profiles, train neural networks or build Hidden Markov Models (HMMs). These alignments can be used to standardize and spread annotation to newer members by homology, as well as to understand the modular architecture of multidomain proteins. PIR-ALN includes 529 alignments that can be used to develop patterns not represented in PROSITE, Blocks, PRINTS and Pfam databases. The ATLAS information retrieval system can be used to browse and query the PIR-ALN alignments.*

Availability: *PIR-ALN is currently being distributed as a single ASCII text file along with the title, member, species, superfamily and keyword indexes. The quarterly and weekly updates can be accessed via the WWW at pir.georgetown.edu. The quarterly updates can also be obtained by anonymous FTP from the PIR FTP site at NBRF.Georgetown.edu, directory [ANONYMOUS.PIR.ALIGNMENT].*

Contact: *pirmail@nbrf.georgetown.edu;Geetha@nbrf.georgetown.edu*

Introduction

Multiple sequence alignments are widely acknowledged to be powerful tools in the analysis of sequence data. Crucial residues for activity and for maintaining protein secondary and tertiary structure are often conserved in sequence alignments. Alignments are also starting points for evolutionary studies (Häger *et al*., 1995; Bilaud *et al*., 1996; Kanai *et al*., 1997; Siezen and Leunissen, 1997), for defining new motifs and domains (Emery *et al*., 1996; Alkema *et al*., 1997; Ponting, 1997; Schultz *et al*., 1997, 1998), for structure prediction (Matsuo *et al*., 1996), for identifying and classifying new sequences (Cosman *et al*., 1990; Henikoff *et al*., 1996; Chuang *et al*., 1997; Henrique *et al*., 1997; Neuwald *et al*., 1997; Smith *et al*., 1997; Wissmann *et al*., 1997; Dosanjh *et al*., 1998), for elucidating structure–function relationships (Douglas *et al*., 1997; Meinnel *et al*., 1997), for homology model building (Adzhubei *et al*., 1998; Dodge *et al*., 1998), for deriving mutation data matrices (Srinivasarao *et al*., 1990; Jones *et al*., 1992; Benner *et al*., 1994) and for site-directed mutagenesis studies (Fiordalisi *et al*., 1994; Aberle *et al*., 1996).

The alignment database (PIR-ALN) was started at the Protein Information Resource in 1991 (Barker *et al*., 1992). The PIR-ALN is derived from the PIR protein sequence database (Barker *et al*., 1999) and has three types of alignments representing protein families, superfamilies and homology domains. The selection of data and inclusion of alignments into PIR-ALN are closely linked to the classification procedures of the PIR-International Protein Sequence Database (Srinivasarao *et al*., 1999).

The concept of ‘protein superfamily’ was introduced by Margaret Dayhoff in the 1970s and was used to partition the protein sequence database based on evolutionary considerations (Dayhoff *et al*., 1975; Dayhoff, 1976). This concept has recently been revised to take into account multidomain proteins (Barker *et al*., 1996). The scheme has been used to classify the sequences, based on global similarity, into hierarchical nested sets of superfamilies and families that are closed under transitivity, and to characterize homology domains based on local similarity. All members of a superfamily will have the same domain architecture, except for those in which part of the genetically encoded sequence is missing due to alternative splicing (Barker *et al*., 1996). Each superfamily alignment is composed of representative sequences from at least two different families (within the same superfamily).

A protein family is a group of sequences that can be aligned from end to end and are <55% different globally. This has been the threshold below which sequences can be unambiguously aligned by currently available multiple alignment methods (McClure *et al*., 1994). Such families are further clustered into superfamilies including more distantly related members.

A homology domain is a subsequence of a protein that is distinguished by a well-defined set of properties or characteristics and also occurs in at least two different superfamilies. Many protein sequences are composed of a number of distinct functional regions (domains) or of multiple copies of the same domain. Domains may be considered basic building blocks for multidomain proteins. Related domains may be shared in various combinations and arrangements among a large number of multidomain proteins (Baron *et al*., 1991; Bork, 1991; Patthy, 1991; Barker *et al*., 1995; Doolittle, 1995; Bork *et al*., 1996). Selected sequence segments corresponding to the same homology domain in different proteins are extracted and aligned in PIR-ALN. Currently, we have an alignment for every homology domain defined in the PIR-International Protein Sequence Database.

The alignments in PIR-ALN contain a selection of sequences both to keep the alignments at reasonable size and to ensure that there is no bias towards a group that has many sequences. The other sequences that could be included in the alignment are listed in the 'other members' field on the entry. No alignment algorithms that we have found can handle automatic addition of new distantly related members to superfamily and homology domain alignments satisfactorily. However, the automatically generated set of alignments including all members, called PROT-FAM (Mewes *et al*., 1997), is available from our collaborators at MIPS and these are cross-referenced in the PIR-ALN entry on the cross-reference record as MIPSALN. These are available for searching on the Web at the URL http://www.mips.biochem.mpg.de.

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, leucine-rich alpha-2-glycoprotein repeat homology, homeobox homology, and kinase-related transforming protein superfamily.

Components of an entry

Each entry in the database has several fields. A sample entry as displayed by the ATLAS retrieval program is shown in Figure 1. The header indicates the beginning of a new entry and contains the unique alignment identifier. Each entry in the database can be cross-referenced by this unique identifier. The superfamily, family and homology domain alignments can be distinguished by >SA, >FA and >DA in the unique identifier, followed by a four-digit number. The *title* line identifies the superfamily, family, or homology domain, which is usually the superfamily name followed by the placement number. The *alternate names* line lists possible alternate names for the proteins or the homology domains. The *date* line shows the entry creation date, sequence revision date and text revision date, indicating when the entry was created and updated. The *members* line contains the identification codes of the sequences on the alignment, followed by member titles as they appear in the PIR protein sequence database. The *cross-references* line at this level has links to external databases like PDB and PROCLASS.

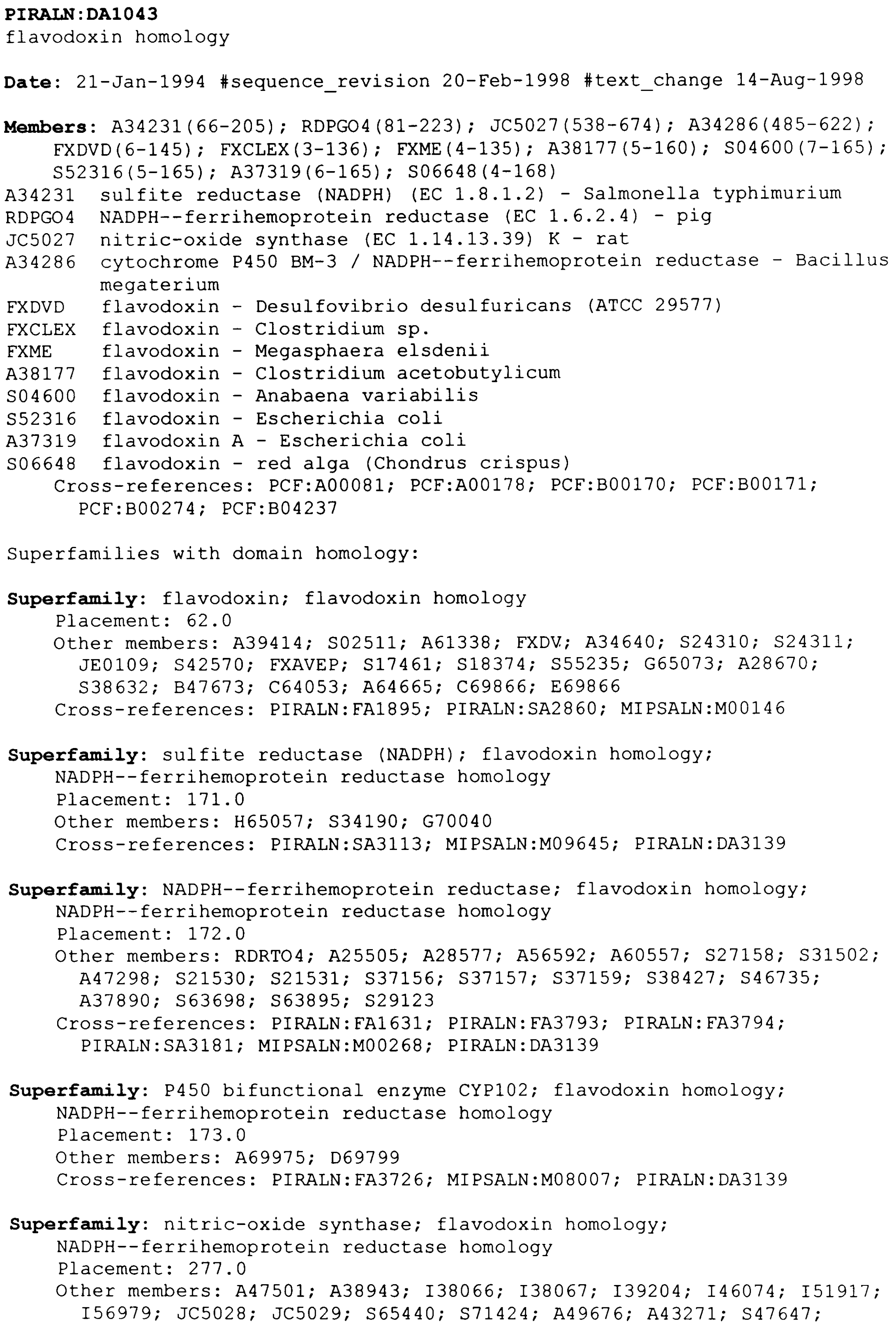
The classification section includes information on the superfamily name, followed by the placement number. (Placement numbers serve to order the superfamilies and sequences in the sequence database; they are adjusted with each quarterly release.) The *other members* line lists the PIR identification codes of sequences that are classified in the same group, but not included on the alignment. The *cross-references* line is used to link the alignment to other related alignments in PIR-ALN and MIPS-ALN databases. The homology domain alignment shown in Figure 1 is linked to several superfamily alignments. The *comment* line reports the numbers of superfamilies or families and their members. The *conserved regions* are computed from the alignment and regular expressions are derived. Keywords that are common to all members of the group are included in the *keywords* line. Other keywords found are included in the *other keywords* line. The number of sequences and number of positions on the alignment are also listed on the comment lines.

The alignment section contains the alignment of sequences in interleaved format. The alignment positions are marked on top of the alignment block. Completely conserved residues are indicated by an asterisk, partially conserved residues up to 51% by a period and those up to 82% by a colon at the bottom of the alignment block.

The matrix of differences provides measures of the interrelationships among the sequences. The upper right portion of the matrix gives the number of differences between the sequences. The lower left portion represents the percent differences between sequences. For the homology domain alignments, the region of the sequence used in the alignment is shown under the sequence code. The matrix can be used to select one sequence from each family in a superfamily alignment for building profiles or for developing weighting schemes so that there is no bias towards any one family.

Alignment methods

CLUSTAL V and W (Higgins *et al*., 1992; Thompson *et al*., 1994), are easy-to-use multiple sequence alignment programs developed by Des Higgins. Multiple alignment programs perform well when sequences are <50% different



```
PIRALN:DA1043
flavodoxin homology

Date: 21-Jan-1994 #sequence_revision 20-Feb-1998 #text_change 14-Aug-1998

Members: A34231(66-205); RDPGO4(81-223); JC5027(538-674); A34286(485-622);
    FXDVD(6-145); FXCLEX(3-136); FXME(4-135); A38177(5-160); S04600(7-165);
    S52316(5-165); A37319(6-165); S06648(4-168)
A34231  sulfite reductase (NADPH) (EC 1.8.1.2) - Salmonella typhimurium
RDPGO4  NADPH--ferrihemoprotein reductase (EC 1.6.2.4) - pig
JC5027  nitric-oxide synthase (EC 1.14.13.39) K - rat
A34286  cytochrome P450 BM-3 / NADPH--ferrihemoprotein reductase - Bacillus
        megaterium
FXDVD   flavodoxin - Desulfovibrio desulfuricans (ATCC 29577)
FXCLEX  flavodoxin - Clostridium sp.
FXME    flavodoxin - Megasphaera elsdenii
A38177  flavodoxin - Clostridium acetobutylicum
S04600  flavodoxin - Anabaena variabilis
S52316  flavodoxin - Escherichia coli
A37319  flavodoxin A - Escherichia coli
S06648  flavodoxin - red alga (Chondrus crispus)
    Cross-references: PCF:A00081; PCF:A00178; PCF:B00170; PCF:B00171;
      PCF:B00274; PCF:B04237

Superfamilies with domain homology:

Superfamily: flavodoxin; flavodoxin homology
    Placement: 62.0
    Other members: A39414; S02511; A61338; FXDV; A34640; S24310; S24311;
      JE0109; S42570; FXAVEP; S17461; S18374; S55235; G65073; A28670;
      S38632; B47673; C64053; A64665; C69866; E69866
    Cross-references: PIRALN:FA1895; PIRALN:SA2860; MIPSALN:M00146

Superfamily: sulfite reductase (NADPH); flavodoxin homology;
    NADPH--ferrihemoprotein reductase homology
    Placement: 171.0
    Other members: H65057; S34190; G70040
    Cross-references: PIRALN:SA3113; MIPSALN:M09645; PIRALN:DA3139

Superfamily: NADPH--ferrihemoprotein reductase; flavodoxin homology;
    NADPH--ferrihemoprotein reductase homology
    Placement: 172.0
    Other members: RDRTO4; A25505; A28577; A56592; A60557; S27158; S31502;
      A47298; S21530; S21531; S37156; S37157; S37159; S38427; S46735;
      A37890; S63698; S63895; S29123
    Cross-references: PIRALN:FA1631; PIRALN:FA3793; PIRALN:FA3794;
      PIRALN:SA3181; MIPSALN:M00268; PIRALN:DA3139

Superfamily: P450 bifunctional enzyme CYP102; flavodoxin homology;
    NADPH--ferrihemoprotein reductase homology
    Placement: 173.0
    Other members: A69975; D69799
    Cross-references: PIRALN:FA3726; MIPSALN:M08007; PIRALN:DA3139

Superfamily: nitric-oxide synthase; flavodoxin homology;
    NADPH--ferrihemoprotein reductase homology
    Placement: 277.0
    Other members: A47501; A38943; I38066; I38067; I39204; I46074; I51917;
      I56979; JC5028; JC5029; S65440; S71424; A49676; A43271; S47647;
```

Fig. 1. A sample entry in the PIR-ALN database as displayed by the ATLAS retrieval program on the PIR Web site: pir.georgetown.edu

I56575; JN0457; I53165; I37361

Cross-references: PIRALN:FA2798; PIRALN:FA2799; PIRALN:SA2797; MIPSALN:M08077; PIRALN:DA3139

Superfamily: mioC protein; flavodoxin homology

Placement: 4269.0

Other members: QQEC16; C64085; S45108; B65061

Cross-references: PIRALN:SA3378; MIPSALN:M05215

Comment: This homology domain occurs in 6 classified superfamilies, 80 classified members.

Conserved region: none

Alignment: #sequences 12 #positions 174

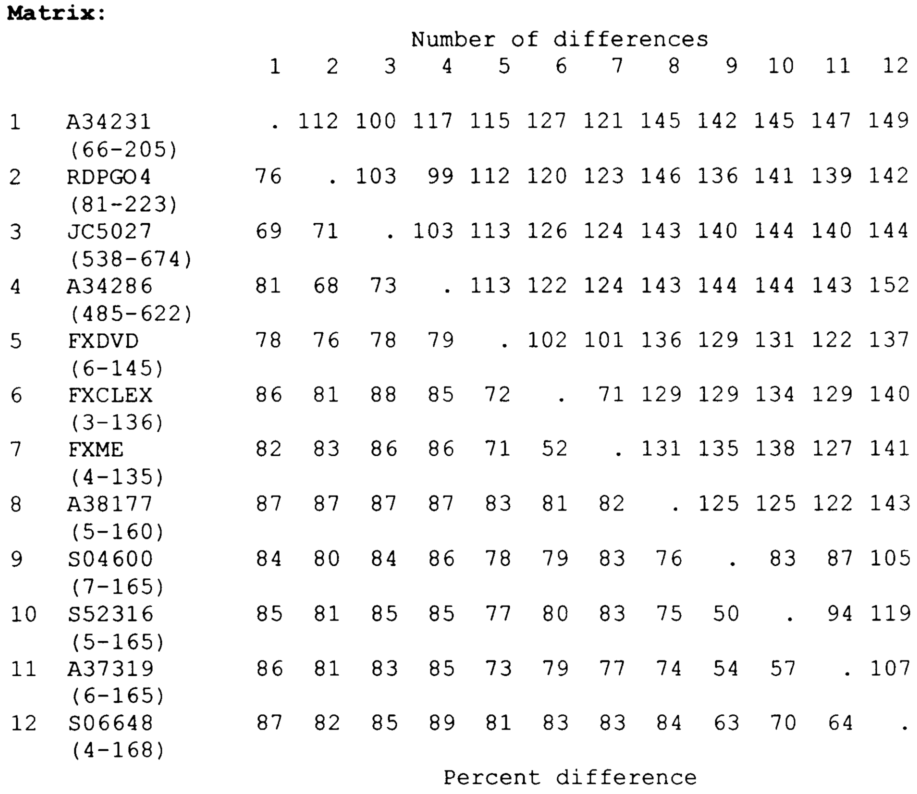
[wide alignment display]

```
                10        20        30        40        50        60
A34231  LISASQTGNARRVAEALRDDLLAANLNVTLVN-AGDYKFKQIAS-----EKLLVIVTSTQ
RDPGO4  VFYGSQTGTAEEFANRLSKDAHRYGMRGMAA-DPEEYDLSDLSSLPEIENALAVFCMATY
JC5027  VLFATETGKSEALARDLA-ALFSYAFNTKVV-CMEQYKANTL-----EEEQLLLVVTSTF
A34286  VLYGSNMGTAEGTARDLADIAMSKGFAPQVA-TL-DSHAGNLP----REGAVLIVTASYN
FXDVD   IVFGSSTGNTESIAQKLEELIAAGGHEVTLL-NAADASAENLA---DGYDAVLFGC-SAW
FXCLEX  IVYWSGTGNTEKMAELIAKGIIESGKDVNTI-NVSDVNIDEL----LNEDILILGC-SAM
FXME    IVYWSGTGNTEAMANEIEAAVKAAGADVESV-RFEDTNVDDV----ASKDVILLGC-PAM
A38177  ILYSSKTGKTERVAKLIEEGVKRSGNIEVKTMNLDAVDKKFL----QESEGIIFGT-PTY
S04600  LFYGTQTGKTESVAEIIRDEF---GNDVVTLHDVSQAEVTDL----NDYQYLIIGC-PTW
S52316  LFYGSSTCYTEMAAEKIRDII---GPELVTLHNLKD-DSPKLM---EQYDVLILG-IPTW
A37319  IFFGSDTGNTENIAKMIQKQL---GKDVADVHDIAKSSKEDL----EAYDILLLGI-PTW
S06648  IFFSTSTGNTTEVADFIGKTL---GAKADAPIDVDDVTDPQAL---KDYDLLFLGA-PTW
conser  . ... ::..:  *  .       :  .       .     .       . ....  ..

                70        80        90       100       110       120
A34231  GEGEPP--EEAVALHKFLFSKKAPKLENTAFAVFSLGDT-SYEFFCQSGKD-FDSKLAEL
RDPGO4  GEGDPT--DNAQDFYDWLQEAD-VDLTGVKYAVFGLGNK-TYEHFNA-MGKYVDKRLEQL
JC5027  GNGDCP--SNGQTLKKSLFMMKELGHT-FRYAVFGLGSS-MYPQFCAFAHD-IDQKLSHL
A34286  GHPP----DNAKQFVDWLDQASADEVKGVRYSVFGCGDK-NWATTYQKVPAFIDETLAAK
FXDVD   GMEDL---EMQDDFLSLFEEFNRIGLAGRKVAAFASGDQ-EYEHFCG-AVPAIEERAKEL
FXCLEX  GDEVL----EESEFEP-FIEEISTKISGKKVALFG-----SYGWGDGKWMRDFEERMNGY
FXME    GSEEL----EDSVVEP-FFTDLAPKLKGKKVGLFG-----SYGWGSGEWMDAWKQRTEDT
A38177  -YANI----SWEMKKW-IDESSEFNLEGKLGAAFSTAN--SIAGGSDIALLTILNHLMVK
S04600  NIGEL----QSDWEGL-YSELDDVDFNGKLVAYFGTGDQIGYADNFQDAIGILEEKISQR
S52316  DFGEI----QEDWEAV-WDQLDDLNLEGKIVALYGLGDQLGYGEWFLDALGMLHDKLSTK
A37319  YYGEA----QCDWDDF-FPTLEEIDFNGKLVALFGCGDQEDYAEYFCDALGTIRDIIEPR
S06648  NTGADTERSGTSWDEFLYDKLPEVDMKDLPVAIFGLGDAEGYPDNFCDAIEEIHDCFAKQ
conser  . .                      . .. .: :. ..   :      .   .   .

               130       140       150       160       170
A34231  GGERLLD---------------------RVDADVEYQAAAS---EWRARVVDVL
RDPGO4  GAQRIFD---------------------LGLGDDDGNLEE----DFITWREQFW
JC5027  GASQLAP---------------------TGEGDELSGQED----AFRSWAVQTF
A34286  GAENIAD---------------------RGEADASDDFEG----TYEEWREHMW
FXDVD   GATIIAE----------------------GLKMEGDASND--PEAVASFAEDVL
FXCLEX  GCVVVET----------------------PLIVQNEPDEA--EQDCIEFGKKIA
FXME    GATVIGT-----------------------AIVNEMPDNA--PE-CKELGEAAA
A38177  GM-LVYSG---GVAFGKPKT-HLGYVHINEIQENEDENARIFGERIANKVKQIF
S04600  GGKTVGYWSTDGYDFNDSKA--LRNGKFVGLALDEDNQSDLTDDRIKSWVAQLK
S52316  GVKCVGYWPTEGYEFTSPKPVIADGQLFVGLALDETNQYDLSDERIQSWCEQIL
A37319  GATIVGHWPTAGYHFEASKG-LADDDHFVGLAIDEDRQPELTAERVEKWVKQIS
S06648  GAKPVGFSNPDDYDYEESKS--VRDGKFLGLPLDMVNDQIPMEKRVAGWVEAVV
conser  *.  .                        ..   .             . ..
```

Figure 1. *Continued*



```
Matrix:
                                 Number of differences
                       1   2   3   4   5   6   7   8   9  10  11  12

1    A34231            . 112 100 117 115 127 121 145 142 145 147 149
     (66-205)
2    RDPGO4           76   . 103  99 112 120 123 146 136 141 139 142
     (81-223)
3    JC5027           69  71   . 103 113 126 124 143 140 144 140 144
     (538-674)
4    A34286           81  68  73   . 113 122 124 143 144 144 143 152
     (485-622)
5    FXDVD            78  76  78  79   . 102 101 136 129 131 122 137
     (6-145)
6    FXCLEX           86  81  88  85  72   .  71 129 129 134 129 140
     (3-136)
7    FXME             82  83  86  86  71  52   . 131 135 138 127 141
     (4-135)
8    A38177           87  87  87  87  83  81  82   . 125 125 122 143
     (5-160)
9    S04600           84  80  84  86  78  79  83  76   .  83  87 105
     (7-165)
10   S52316           85  81  85  85  77  80  83  75  50   .  94 119
     (5-165)
11   A37319           86  81  83  85  73  79  77  74  54  57   . 107
     (6-165)
12   S06648           87  82  85  89  81  83  83  84  63  70  64   .
     (4-168)
                                   Percent difference
```

Figure 1. *Continued*

(Yeh *et al.*, 1993a,b; McClure *et al.*, 1994) and provide good starting points for alignments of more distantly related sequences.

All superfamily and homology domain alignments require editing to arrive at biologically realistic alignments. The following general criteria are used: (i) Within each sequence, gaps are minimally dispersed, especially at the ends. (ii) Among sequences, gaps are aligned whenever possible. (iii) The alignment should reflect conserved biological features such as active sites. In cases where the crystal structure is determined, the sequence from NRL_3D is used to help align the sequences based on structural features.

Although several alignment editors are available currently (Schuler *et al.*, 1991; Clark, 1992; Depiereux and Feytmans, 1992; Smith *et al.*, 1994; Attwood *et al.*, 1997; Thompson *et al.*, 1997), we have been developing our own editor that includes customized features for interacting with the PIR sequence databases. ALNED program in edit mode is used as an alignment editor to view, check, and correct alignments, and in update mode is used to create the distribution version of the PIR-ALN database, and to check for concurrency between PIR and PIR-ALN databases.

The ALNED Graphical User Interface is an interactive, menu-driven alignment editor written in FORTRAN for the VAX/VMS systems (Hunt, 1987). The editor can handle up to 200 sequences and sequence length of 15 000. The program reads in a file with pointers to the sequences in the protein sequence database and the corresponding gap specifications. An alignment can also be generated interactively. New complete sequences or segments either from any database available at PIR (e.g. NRL-3D), or from an external user-defined file, can be added to the alignment using the Needleman–Wunsch algorithm. All sequences in the other members line can be added automatically to the alignment.

Sequences can be grouped so that edits performed on one sequence, such as adding or deleting gaps, are propagated to all the members of the group simultaneously. This is helpful in working with superfamily alignments that have representatives from different families. One can also trim or extend the ends of the alignment, which is useful in deciding domain boundaries. One can lock the alignment vertically so regions that have already been aligned will not be affected by further editing. The program can search for sequence motifs, which can be very helpful when working with sequences that contain repeats or large unalignable regions in the middle. The program calculates the matrix of percent difference among the members and can rearrange the sequences according to the matrix.

The program also displays annotation information such as titles, classification, keywords, and features on the individual members of the alignment. This will guide the annotator in deciding which residues are crucial and what features can be standardized among the entries. The classification information and the matrix of percent difference will aid the annotators in checking the current classification and also in classifying new sequences.

Database updates

The primary challenge in keeping up the PIR-ALN database is to ensure concurrency of both the sequence and annotation information in both the PIR-International Protein Sequence Database and the PIR-ALN databases with every weekly update. Unlike some databases (such as Blocks) that can be generated with every update of the primary sequence database, PIR-ALN needs options for adding, deleting and modifying alignment entries. These database operations are handled by the ALNED program in update mode.

Each alignment entry is maintained as a file that has pointers to the PIR-International Protein Sequence Database along with the gap specifications and checksums for each sequence. The PIR-ALN alignment entry itself is generated with every update taking titles, species, alternate names, superfamily names, keywords and placement information from the member entries as well as the sequences from the PIR-International Protein Sequence Database. The species, superfamily names and keyword fields have controlled vocabularies (Barker, 1993) that can be browsed and used for searching the database from our Web site (pir.georgetown.edu/pirwww/search/lists.html).

In update mode, ALNED also checks the annotation of member entries for any reference to PDB and adds it to the cross-reference record of a PIR-ALN entry. The self cross-references to other PIR-ALN entries are determined using the member index and added while updating the database. ALNED also has built-in checking procedures that notify the user of logical inconsistencies, sequence revisions as well as changes in classification of the members. Checksums between PIR-International Protein Sequence Database and PIR-ALN are compared to determine sequence revisions. The error log will alert the annotator which alignments need to be modified. These checks are very important in cases when there is a sequence revision or if there are changes in the homology domain boundaries or when a member no longer belongs to the same family or superfamily. The ALNED program in update mode generates the distribution version of the PIR-ALN database.

Database access and distribution

ATLAS, a multidatabase retrieval program developed at PIR, can be used to query and retrieve alignments from the alignment database (Barker *et al.*, 1993). From the distribution versions of PIR-International Protein Sequence Database and PIR-ALN databases, a single multidatabase, multifield index is created. The ATLAS program uses this index to access several databases, including the PIR-International Protein Sequence Database and PIR-ALN simultaneously. Retrieval operations generate a current list (active subset of database entries) that may be modified by Boolean combinations of successive commands. The fields such as titles, members, superfamily names, species and keywords are indexed for quick and easy access to the data. The commands by which the PIR-ALN database can be accessed are given in Table 1. ATLAS is written in ANSI standard C and runs on several platforms. The various commands and capabilities of the ATLAS program are documented in http://pir.georgetown.edu/pirwww/product/atlascd.html.

Table 1. The list of commands in ATLAS that access the PIR-ALN database

Type:	Displays all the information contained in the entry specified by the user-specified unique alignment identifier.
Find:	Searches the title index and retrieves all alignments that include the user-specified terms in the title or the alternate names field.
Member:	Retrieves all the alignments in which a specified sequence (by sequence identifier code) has been used. Different segments of the same sequence can appear in different alignments (e.g., domains of a multidomain protein).
Keyword:	Searches the keyword index for all occurrences of a user-specified keyword or partial keyword.
Superfamily:	Searches the superfamily index for all occurrences of a user-specified superfamily name.
Copy:	Copies an alignment entry into an output file for display or printing. (Additional modifiers will be added to this command so that the user can retrieve the alignments in formats compatible with different alignment editors for further modification.)
Get/member:	Retrieves all members of the alignment from the PIR sequence database to a current list.
Species:	Retrieves all alignments that contain a sequence from user-specified species.

The PIR-ALN database can be accessed on the PIR Web site in two ways. From the PIR request page (http://pir.georgetown.edu/pirwww/search/searchdb.html), the PIR sequence entry will cross-reference the PIR-ALN entry if the sequence is a member of any alignment. Alternately, you can access the PIR-ALN from the alignment search page (http://pir.georgetown.edu/pirwww/search/textpiraln.html). The members and classification fields are hypertext linked to the PIR protein sequence database, so the user can move between the two databases.

Discussion

There are several second-generation databases derived from primary sequence databases that have recently been compared for their usability (Brown, 1998; Hofmann, 1998). PIR-ALN is derived from the PIR-International Protein Sequence Database and contains curated, gapped multiple sequence alignments representing a wide variety of proteins. The alignments in PIR-ALN contain all the information necessary to derive regular expressions (Bairoch, 1993), fingerprints (Attwood *et al*., 1998), weighted ungapped segments (Henikoff and Henikoff, 1996), HMMs (Sonnhammer *et al*., 1998), motifs (Bachinsky *et al*., 1997; Nevill-Manning *et al*., 1998), neural networks (Wu *et al*., 1998) and profiles (Gribskov and Veretnik, 1996). These can be used as diagnostic tools for searching the primary sequence databases for classification purposes and also to look for patterns in a user-defined sequence. Analyzing data from the ProClass database (Wu *et al*., 1998), we found that there are 529 unique alignments in PIR-ALN which can be used to develop patterns that are not represented in Blocks, PROSITE, Pfam and PRINTS databases. The alignments in PIR-ALN are linked to several of the above datasets via cross-references to ProClass database.

PIR-ALN is used internally by the PIR staff to define new domains and to check and refine domain boundaries. When working with highly divergent superfamilies, it is necessary to examine alignments to validate inclusion of sequences into the superfamily. In merging reports of the same sequence by different groups, alignments are very useful in identifying frameshift errors and in deciding which report is more reliable by comparing sequences from different species. PIR-ALN is also used as a tool for standardizing annotation information such as titles, features, classification and keywords across members of the alignment.

Plans for future enhancements include adding features, profiles and citations to the alignments, as well as providing the alignments in different formats compatible with other alignment editors and other bioinformatics tools. The feature information will be taken from annotations in the PIR-International Protein Sequence Database, which will be re-examined for consistency with the alignment. This will facilitate the development of an objective, self-consistent system for the assignment and verification of protein sequence features by homology as depicted by multiple sequence alignments.

PIR-ALN, a carefully reviewed, well-maintained alignment database with effective retrieval software, will serve as a useful resource for researchers in the biological, medical, and biotechnological sciences. It will also be useful for describing more precisely the evolutionary relationships of multidomain proteins and in comparative studies of whole genomes.

Acknowledgements

This work has been supported by NLM grant LM05798. The authors thank PIR staff for their contributions of alignments to the database and Kathryn Sidman for administrative support.

References

Aberle,H., Schwartz,H., Hoschuetzky,H. and Kemler,R. (1996) Single amino acid substitutions in proteins of the armadillo gene family abolish their binding to alpha-catenin. *J. Biol. Chem*., **271**, 1520–1526.

Adzhubei,I.A., Adzhubei,A.A. and Neidle,S. (1998) An integrated sequence-structure database incorporating matching mRNA sequence, amino acid sequence and protein three-dimensional structure data. *Nucleic Acids Res*., **26**, 327–331.

Alkema,M.J., Bronk,M., Verhoeven,E., Otte,A., van 't Veer,L.J., Berns,A. and van Lohuizen,M. (1997) Identification of Bmi1-interacting proteins as constituents of a multimeric mammalian polycomb complex. *Genes Dev*., **11**, 226–240.

Attwood,T., Payne,A.W.R., Michie,A.D. and Parry-Smith,D.J. (1997) A Colour INteractive Editor for Multiple Alignments—CINEMA. *EMBnet.news*, **3** (http://www2.ebi.ac.uk/embnet/news/).

Attwood,T.K., Beck,M.E., Flower,D.R., Scordis,P. and Selley,J.N. (1998) The PRINTS protein fingerprint database in its fifth year. *Nucleic Acids Res*., **26**, 304–308.

Bachinsky,A.G., Yarigin,A.A., Guseva,E.H., Kulichkov,V.A. and Nizolenko,L.Ph. (1997) A bank of protein family patterns for rapid identification of possible functions of amino acid sequences. *Comput. Applic. Biosci*., **13**, 115–122.

Bairoch,A. (1993) The PROSITE dictionary of sites and patterns in proteins, its current status. *Nucleic Acids Res*., **21**, 3097–3103.

Barker,W.C., George,D.G., Srinivasarao,G.Y. and Yeh,L.-S. (1992) Database of protein sequence alignments. *FASEB J*., **6**, A348.

Barker,W.C., George,D.G., Mewes,H.-W., Pfeiffer,F. and Tsugita,A. (1993) The PIR-International databases. *Nucleic Acids Res*., **21**, 3089–3092.

Barker,W.C., Pfeiffer,F. and George,D.G. (1995) Superfamily and domain: organization of data for molecular evolution studies. In Atassi,M.Z. and Appella,E. (eds), *Methods in Protein Structure Analysis*. Plenum Press, New York, pp. 473–481.

Barker,W.C., Pfeiffer,F. and George,D.G. (1996) Superfamily classification in PIR-International Protein Sequence Database. *Methods Enzymol*., **266**, 59–71.

Barker,W.C. *et al*. (1999) The PIR-International Protein Sequence Database. *Nucleic Acids Res*., **27**, 39–43.

Baron,M., Norman,D.G. and Campbell,I.D. (1991) Protein modules. *Trends Biochem. Sci*., **16**, 13–17.

Benner,S.A., Cohen,M.A. and Gonnet,G.H. (1994) Amino acid substitution during functionally constrained divergent evolution of protein sequences. *Protein Eng*., **7**, 1323–1332.

Bilaud,T., Koering,C.E., Binet-Brasselet,E., Ancelin,K., Pollice,A., Gasser,S.M. and Gilson,E. (1996) The telobox, a Myb-related telomeric DNA binding motif found in proteins from yeast, plants and human. *Nucleic Acids Res*., **24**, 1294–1303.

Bork,P. (1991) Shuffled domains in extracellular proteins. *FEBS Lett*., **286**, 47–54.

Bork,P., Downing,A.K., Kieffer,B. and Campbell,I.D. (1996) Structure and distribution of modules in extracellular proteins. *Q. Rev. Biophys*., **29**, 119–167.

Brown,S.M. (1998) Analyzing protein families and domains on the Web. *BioTechniques*, **25**, 596–598.

Chuang,M.-H., Pan,F.-M. and Chiou,S.-H. (1997) Sequence characterization of gamma-crystallins from lip shark (*Chiloscyllium colax*): Existence of two cDNAs encoding gamma-crystallins of mammalian and teleostean classes. *J. Protein Chem*., **16**, 299–307.

Clark,S.P. (1992) MALIGNED: a multiple sequence alignment editor. *Comput. Applic. Biosci*., **8**, 535–538.

Cosman,D., Lyman,S.D., Idzerda,R.L., Beckmann,M.P., Park.L.S., Goodwin,R.G. and March,C.J. (1990) A new cytokine receptor superfamily. *Trends Biochem. Sci*., **15**, 265–270.

Dayhoff,M.O. (1976) The origin and evolution of protein superfamilies. *Fed. Proc*., **35**, 2132–2138.

Dayhoff,M.O., McLaughlin,P.J., Barker,W.C. and Hunt,L.T. (1975) Evolution of sequences within protein superfamilies. *Naturwissenschaften*, **62**, 154–161.

Depiereux,E. and Feytmans,E. (1992) MATCH-BOX: a fundamentally new algorithm for the simultaneous alignment of several protein sequences. *Comput. Applic. Biosci*., **8**, 501–509.

Dodge,C., Schneider,R. and Sander,C. (1998) The HSSP database of protein structure—sequence alignments and family profiles. *Nucleic Acids Res*., **26**, 313–315.

Doolittle,R.F. (1995) The multiplicity of domains in proteins. *Annu. Rev. Biochem*., **64**, 287–314.

Dosanjh,M.K., Collins,D.W., Fan,W., Lennon,G.G., Albala,J.S., Shen,Z. and Schild,D. (1998) Isolation and characterization of RAD51C, a new human member of the RAD51 family of related genes. *Nucleic Acids Res*., **26**, 1179–1184.

Douglas,D.A., Shi,Y.E. and Sang,Q.A. (1997) Computational sequence analysis of the tissue inhibitor of metalloproteinase family. *J. Protein Chem*., **16**, 237–255.

Emery,P., Durand,B., Mach,B. and Reith,W. (1996) RFX proteins, a novel family of DNA binding proteins conserved in the eukaryotic kingdom. *Nucleic Acids Res*., **24**, 803–807.

Fiordalisi,J.J., Al-Rabiee,R., Chiappinelli,V.A. and Grant,G.A. (1994) Site-directed mutagenesis of kappa-bungarotoxin: implications for neuronal receptor specificity. *Biochemistry*, **33**, 3872–3877.

Gribskov,M. and Veretnik,S. (1996) Identification of sequence patterns with profile analysis. *Methods Enzymol*., **266**, 198–212.

Häger,K.-P., Braun,H., Czihal,A., Müller,B. and Bäumlein,H. (1995) Evolution of seed storage protein genes: Legumin genes of *Ginkgo biloba. J. Mol. Evol*., **41**, 457–466.

Henikoff,J.G. and Henikoff,S. (1996) Blocks database and its applications. *Methods Enzymol*., **266**, 88–105.

Henikoff,S., Endow,S.A. and Greene,E.A. (1996) Connecting protein family resources using the proWeb network. *Trends Biochem. Sci*., **21**, 444–445.

Henrique,D., Tyler,D., Kintner,C., Heath,J.K., Lewis,J.H., Ish-Horowicz,D. and Storey,K.G. (1997) cash4, a novel achaete-scute homolog induced by Hensen’s node during generation of the posterior nervous system. *Genes Dev*., **11**, 603–615.

Higgins,D.G., Bleasby,A.J. and Fuchs,R. (1992) CLUSTAL V: improved software for multiple sequence alignment. *Comput. Applic. Biosci*., **8**, 189–191.

Hofmann,K. (1998) Protein classification & functional assignment. *Trends Guide To Bioinformatics*, **(Trends Supplement)**, 18–21.

Hunt,L.T. (1987) *Protein Identification Resource Newsletter, No. 3*., National Biomedical Research Foundation, Washington, DC.

Jones,D.T., Taylor,W.R. and Thornton,J.M. (1992) The rapid generation of mutation data matrices from protein sequences. *Comput. Applic. Biosci*., **8**, 275–282.

Kanai,S., Kikuno,R., Toh,H., Ryo,H. and Todo,T. (1997) Molecular evolution of the photolyase–blue-light photoreceptor family. *J. Mol. Evol*., **45**, 535–548.

Matsuo,Y., Yamada,A., Tsukamoto,K., Tamura,H.-O., Ikezawa,H., Nakamura,H. and Nishikawa,K. (1996) A distant evolutionary relationship between bacterial sphingomyelinase and mammalian DNase I. *Protein Sci*., **5**, 2459–2467.

McClure,M.A., Vasi,T.K. and Fitch,W.M. (1994) Comparative analysis of multiple protein-sequence alignment methods. *Mol. Biol. Evol*., **11**, 571–592.

Meinnel,T., Lazennec,C., Villoing,S. and Blanquet,S. (1997) Structure-function relationships within the peptide deformylase family. Evidence for a conserved architecture of the active site involving three conserved motifs and a metal ion. *J. Mol. Biol*., **267**, 749–761.

Mewes,H.-W., Albermann,K., Heumann,K., Liebl,S. and Pfeiffer,F. (1997) MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Res*., **25**, 28–30.

Neuwald,A.F., Liu,J.S., Lipman,D.J. and Lawrence,C.E. (1997) Extracting protein alignment models from the sequence database. *Nucleic Acids Res*., **25**, 1665–1677.

Nevill-Manning,C.G., Wu,T.D. and Brutlag,D.L. (1998) Highly specific protein sequence motifs for genome analysis. *Proc. Natl Acad. Sci. USA*, **95**, 5865–5871.

Patthy,L. (1991) Modular exchange principles in proteins. *Curr. Opin. Struct. Biol*., **1**, 351–361.

Ponting,C.P. (1997) Tudor domains in proteins that interact with RNA. *Trends Biochem. Sci*., **22**, 51–52.

Schuler,G.D., Altschul,S.F. and Lipman,D.J. (1991) A workbench for multiple alignment construction and analysis. *Proteins*, **9**, 180–190.

Schultz,J., Ponting,C.P., Hofmann,K. and Bork,P. (1997) SAM as a protein interaction domain involved in developmental regulation. *Protein Sci*., **6**, 249–253.

Schultz,J., Milpetz,F., Bork,P. and Ponting,C.P. (1998) SMART, a simple modular architecture research tool: Identification of signaling domains. *Proc. Natl Acad. Sci. USA*, **95**, 5857–5864.

Siezen,R.J. and Leunissen,J.A.M. (1997) Subtilases: the superfamily of subtilisin-like serine proteases. *Protein Sci*., **6**, 501–523.

Smith,C.M., Shindyalov,I.N., Veretnik,S., Gribskov,M., Taylor,S.S., Ten Eyck,L.F. and Bourne,P.E. (1997) The protein kinase resource. *Trends Biochem. Sci*., **22**, 444–446.

Smith,S.W., Overbeek,R., Woese,C.R., Gilbert,W. and Gillevet,P.M. (1994) The genetic data environment an expandable GUI for multiple sequence analysis. *Comput. Applic. Biosci*., **10**, 671–675.

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.

Srinivasarao,G.Y., George,D.G. and Barker,W.C. (1990) The 1989 mutation data matrix. *Biophys. J*., **57**, 429a.

Srinivasarao,G.Y., Yeh,L.-S., Marzec,C.R., Orcutt,B.C., Barker,W.C. and Pfeiffer,F. (1999) Database of protein sequence alignments: PIR-ALN. *Nucleic Acids Res*., **27**, 284–285.

Thompson,J.D., Higgins,D.G. and Gibson,T.J. (1994) ClustalW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*., **22**, 4673–4680.

Thompson,J.D., Gibson,T.J., Plewniak,F., Jeanmougin,F. and Higgins,D.G. (1997) The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*., **25**, 4876–4882.

Wissmann,A., Ingles,J., McGhee,J.D. and Mains,P.E. (1997) *Caenorhabditis elegans* LET-502 is related to Rho-binding kinases and human myotonic dystrophy kinase and interacts genetically with a homolog of the regulatory subunit of smooth muscle myosin phosphatase to affect cell shape. *Genes Dev*., **11**, 409–422.

Wu,C.H., Shivakumar,S., Shivakumar,C.V. and Chen,S.-C. (1998) GeneFIND Web server for protein family identification and information retrieval. *Bioinformatics*, **14**, 223–224.

Yeh,L.-S., Srinivasarao,G.Y. and Barker,W.C. (1993a) Multiple sequence alignment methods. *Biophys. J*., **64**, A220.

Yeh,L.-S., Srinivasarao,G.Y. and Barker,W.C. (1993b) A study of multiple sequence alignment methods. Fifth International Symposium and Workshops, June 15–20, Baltimore, MD, USA.