Journal of Computer-Aided Molecular Design, 8 (1994) 193–210
ESCOM

J-CAMD 234

HomologyPlot: Searching for homology to a family of proteins using a database of unique conserved patterns

J.M.R. Parker and R.S. Hodges*

Medical Research Council Group in Protein Structure and Function, Department of Biochemistry, University of Alberta, Edmonton, Alberta, Canada T6G 2H7

Received 25 May 1993
Accepted 3 September 1993

Key words: Homology; Family; Database

SUMMARY

A new database of conserved amino acid residues is derived from the multiple sequence alignment of over 84 families of protein sequences that have been reported in the literature. This database contains sequences of conserved hydrophobic core patterns which are probably important for structure and function, since they are conserved for most sequences in that family. This database differs from other single-motif or signature databases reported previously, since it contains multiple patterns for each family. The new database is used to align a new sequence with the conserved regions of a family. This is analogous to reports in the literature where multiple sequence alignments are used to improve a sequence alignment. A program called HomologyPlot (suitable for IBM or compatible computers) uses this database to find homology of a new sequence to a family of protein sequences. There are several advantages to using multiple patterns. First, the program correctly identifies a new sequence as a member of a known family. Second, the search of the entire database is rapid and requires less than one minute. This is similar to performing a multiple sequence alignment of a new sequence to all of the known protein family sequences. Third, the alignment of a new sequence to family members is reliable and can reproduce the alignment of conserved regions already described in the literature. The speed and efficiency of this method is enhanced, since there is no need to score for insertions or deletions as is done in the more commonly used sequence alignment methods. In this method only the patterns are aligned. HomologyPlot also provides general information on each family, as well as a listing of patterns in a family.

INTRODUCTION

Probably the most reliable method to predict and model the 3D structure of a new sequence is by sequence homology to an X-ray-defined protein structure. However, this is dependent on the accuracy of the alignment method, which in turn depends on the choice of parameters used in each method [1]. The use of multiple alignment algorithms, where several structurally similar protein sequences are used to identify conserved residues, promises to increase the probability of

*To whom correspondence should be addressed.

correct alignments [2]. These methods must still contend with the problems of assigning the appropriate similarity and gap scores. It is interesting that none of them allows specifically for the observed variability in the number of residues between conserved regions in proteins.

We propose an alternative method of identifying homologies and sequence alignments, based on a database of several conserved hydrophobic core patterns observed for each family of protein sequences. The program HomologyPlot, suitable for IBM or compatibles, uses this database to find homology to known families. The assumption is that hydrophobic residues in the interior or core of a protein direct the protein folding pattern and are conserved for all proteins in a family of similar protein structures. This is an extension of the role of hydrophobic residues in proteins, first proposed by Kauzman [3] in 1959. Further support for the importance of hydrophobic residues in protein folding comes from numerous reports that a correlation of sequence hydrophobicities is a measure of structural similarities [4–7]. Solvent-inaccessible residues have also been shown to identify a core of residues that are conserved for members of a protein family [8]. Not surprisingly, residue types most frequently observed in the core were the hydrophobic residues V, I, L, W, F, M and C. It has also been shown that hydrophobic microdomains of proteins are conserved in evolution and form a well-defined core structure [9]. Sequence alignments can be improved by not allowing insertions or deletions in hydrophobic cores of proteins [10,11]. Hydrophobic patterns have also been used to identify structural folds in proteins and are a distinctive feature for a particular class of proteins [12].

HomologyPlot provides a new perspective on the several comprehensive homology search algorithms already available [13–17]. The use of conserved sites in homology alignments has been described in detail for the IgG [6], globin [16,18] and serine protease [19] families. Reverse searching [20], or using fragment libraries [15] and structural motifs [21,22] to search for sequence homologies, has been described to identify homologous families of proteins. However, to the best of our knowledge, this is the first application of a database of these unique conserved patterns to identify family homologies as well as to provide sequence alignments. Using these conserved patterns in sequence alignments is similar to multiple sequence alignment procedures, since multiple sequences in a family were used to define the conserved core patterns. In addition, these patterns can be used to screen uncharacterized sequences in a protein sequence database and to scan for homologies to known protein sequences. An important feature of this program is the relatively small database size and the resulting search speed, which is typically one minute on a 286 IBM PC computer. The resulting alignment of patterns or sequences is automatic and quickly identifies important regions that can fix the sequence alignment to conserved regions of X-ray-defined structures. New patterns or motifs for coiled-coil and leucine zippers are presented.

METHODS

Amino acid substitution and similarities

Although there are many amino acid substitution and similarity scales, we have chosen to use a similarity grouping of amino acids, determined specifically in the conserved regions of protein sequences. We reasoned that the similarities of amino acids may be more restricted in evolutionary conserved regions of protein sequences. We have observed that a grouping of nonpolar, small or polar and charged residues was consistently found in conserved regions of protein families. Five groups (see Table 1, Homology Similarity Group) were selected as follows: group 1 contains the nonpolar residues V, I, L, F, Y, W, M and C and the residues A and T; group 2 contains the small/polar residues G, A, S, T, P, N and D, note that the residues A and T are also allowed in group 1; group 3 contains the negatively charged residues E and D and their neutral analogs Q and N; group 4 contains the residues H and Q; group 5 contains the positively charged residues R and K.

Family pattern database

The Dayhoff Atlas of Protein Sequence and Structure [23] identifies 181 superfamilies, comprising 314 families of protein sequences. We have compiled a database of protein family sequences where more than one sequence is known for that family. The families included in this version of HomologyPlot are: actin [23,24], alcohol dehydrogenase [23], antifreeze protein [25], aminoacyl tRNA synthetase [26], amyloid [27], aspartate transcarbamylase [23,28], aspartic proteinase [23,29–32], azurin [23,33,34], basic protease inhibitor [23], beta barrels [35,36], Bowman–Birk inhibitors [23], bradykinin [23], calcitonin [23], carbonic anhydrase B and C [23], carboxypeptidase [23], casein [23], catalase [23], channel proteins [37,38], citrate synthase [39], collagen [23], conotoxins [40], corticotropin [23], crystalin [23], crambin [23], cytochrome B [23], cytochrome C [23,41], dihydrofolate reductase [23], defensins [42], endothelins [43], enolase [44], epidermal growth factor [45,46], eucaryotic initiation factors [47], ferredoxin [23], ferritin [48], fibrinogen [23], flaggelin [23], flavodoxin [23], gastrin [23], globin [18,23,49–51], glyceraldehyde-3-phosphate dehydrogenase [23], G proteins [52], hemerythrin [23], high potential iron sulfur protein [23], IgG lambda C light [23], IgG

TABLE 1
COMPARISON OF HOMOLOGYPLOT AMINO ACID SIMILARITIES TO THE 250-PAM SCORE[a]

Residue	HomologyPlot similarity group					250-PAM similarity	
	1	2	3	4	5	HIGH	LOW
V	V					ILM	TA
I	I					VLMF	T
L	L					MIVF	
W	W					R	FY
F	F					YLI	MW
Y	Y					F	HWC
M	M					LIV	FRK
C	C						SY
A	A	A				STPG	VQEDN
T	T	T				SA	PGNDEKIV
G		G				SAD	NET
P		P				SA	TQHR
S		S				TPAGN	DERK
D		D	D			EQNGH	STAK
N		N	N			DHEQK	TAGR
E			E			DQNH	STAGK
Q			Q	Q		HEDNRK	PA
H				H		QNRDE	PYK
R					R	KHWQ	SPNM
K					K	RNQ	STDEHM

[a] Column 1 lists all 20 naturally occurring amino acids. Columns 2–6 list the residues in groups 1–5 for HomologyPlot. Columns 7 and 8 list all 250-PAM HIGH and LOW similarities, respectively. HIGH similarities have 250-PAM scores of greater than 1 and LOW similarities have 250-PAM scores of 0. Residue similarity in the 250-PAM columns is ranked from highest to lowest similarity.

kappa C light [23], IgG kappa V light [23], inhibin [53], insulin [54], interleukin [55], keratin [23], Kringle region [23], Kunitz-type trypsin inhibitor [56,57], lactate dehydrogenase [23], lectin [58], lysozyme [23], myosin [23], neurophysin [23], ovomucoid third domain [23,59], pancreatic polypeptide [54], pancreatic secretary trypsin inhibitor [23], penicillinase [23], phospholipase [23], plant toxin [23,60], plastocyanin [23], proinsulin-related protein [23,54], ras protein [23,61,62], retroviral protease [32,63,64], ribonuclease [23], rubredoxin [23], scorpion neurotoxin [23], serine proteases [19,23,65,66], snake toxin [23], streptokinase [23], subtilisin [23], super oxide dismutase thioredoxin [23,67], triose phosphate isomerase [23], troponin C [23], tryptophan synthetase [23], vasopressin [23].

Motif database

A database of motifs was also compiled and included: N-linked glycosides, chymotrypsin reactive serine, trypsin reactive serine, thrombin reactive serine, phosphoglucomutase reactive serine, phosphorylase reactive serine, phosphatase active site, serine proteases, AMP-dependent kinase, ATP binding protein, nuclear location site, fibronectin recognition site, ATP binding site, EF Ca binding site, calmodulin-dependent kinase, phosphorylation site, GTP binding site, GTPase activity, IgG variable region, ADP beta-alpha-beta fold, GTP binding protein, zinc DNA binding protein, sugar transport, actin binding protein, RNA binding protein, tyrosine kinase, leucine zipper, coiled-coils, A motif for ATP binding, B motif for ATP binding, aminoacyl tRNA synthetase, DNA binding proteins, zinc finger, cell division cycle gene, ligand gated ion channel protein and phosphate binding loop. These sites or motifs have been described in the literature [21,22].

Compiling the database

Sequence homologies were obtained from multiple sequence alignments, described in the literature. Each family was analyzed individually, with more reliable information being obtained from families that contained more sequences. Patterns are defined as a series of consecutive highly conserved residues, observed in sequence alignments for a family of protein sequences. Patterns were recorded sequentially from the N-terminus to the C-terminus. Breaks were allowed where there were more than four consecutive residues without conserved homology. Two types of patterns were used, identification patterns (ID) and hydrophobic core (HCore) patterns. The ID patterns were highly conserved in all members of the family and are a minimum requirement to identify any family. The HCore patterns, which represent the full set of patterns, were also conserved but may not be represented in all members of that family. The ID patterns were always a subset of the HCore patterns and were used mainly to identify families. The HCore patterns were used primarily for sequence alignments. The following is a more detailed description of how patterns are located and assigned from sequence alignments.

First, conserved core regions are identified in multiple sequence alignments starting from the N-terminus and proceeding to the C-terminus. Core regions are defined as highly homologous regions, comprising 1–5 or 1–4 hydrophobic repeats that do not contain insertions or deletions compared to other sequences for that region in the same family of sequences. Nonpolar residues are represented by group 1, nonhomologous residues are presented by an asterisk. Highly conserved residues are also recorded as part of the pattern. Examples of these core regions are shown in Fig. 1 for pattern 1 (1***1**1), pattern 2 (1**1F), pattern 3 (1F**1), pattern 4 (1G**11), pattern 5 (1***1**1), pattern 6 (1**1***H), pattern 7 (F**1***11), pattern 8 (A1**11) and pattern 9 (1***Y). If a highly conserved region has several 1–5 or 1–4 repeats separated by four or more nonconserved residues, this region is represented by two or more patterns. An example of a single conserved region that has two patterns is given by the 18-residue region (see region containing patterns 2 and 3) in the sequence alignment of the proteins SWMb, HaHb, IGlbn and LegHb in Fig. 1. The highly conserved P residue (see pattern 2, proline) is separated by four or more residues from the next conserved residue 1 (K) (see pattern 3, the first conserved residue in this pattern is nonpolar (group 1); occasionally lysine (K) is observed at this position). For this reason, and because of the deletion in this region for WGlbn, this conserved region is assigned two patterns.

In some cases, exact residues or groups of residues adjacent to the core regions identified as described above are used in the pattern, if all sequences in the alignment have the same residue at that position and these residues are not separated by more than three residues from the N- or C-terminus of the core regions defined above. Examples of this are shown in Fig. 1 for group 3 and the residues serine (S) in pattern 1, proline (P) in pattern 2, histidine (H) in pattern 6 and

```
     1           Pattern1                Pattern2    Pattern3
     2        ..T..........S...................HL........KV.....
     3        .1S**3***1***1**1.............1**1F***P....1F**1..
     4          .hhhhhhhhhhhhhhtt   .hhhhhhhhhhhhhh hhhh.thhhh.t
SWMb          VLSEGEWQLVLHVWAKVEA--DVAGHGQDILIRLFKSHPETLEKFDRFKH
HaHb          VLSPADKTNVKAAWGKVGA--HAGEYGAEALERMFLSFPTTKTYFPHF-D
IGlbn         -LSADQISTVQASFDKVKG------DPVGILYAVFKADPSIMAKFTQFAG
WGlbn         GLSAAQRQVIAATWKDIAGNDNGAGVGKDCLIKHLSAHPQMA-AVFGFSG
LegHb         GFTEKQEALVNSSSQLFKQNPSN--YSVLFYTIILQKAPTAKAMFSFLKD

     1                       Pattern4     Pattern5            Pattern6
     2        ...............12.............5......,............
     3        ...............HG**11..1***1**1...........1**1***H
     4        ...hhhhhh.hhhhhhhhhhhhhhhhhhttt.     t  hhhhhhhhhh
SWMb          LKTEAEMKASEDLKKHGVTVLTALGAILKKK-----GHHEAELKPLAQSH
HaHb          LS-----HGSAQVKGHGKKVADALTNAVAHV-----DDMPNALSALSDLH
IGlbn         -KDLESIKGTAPFETHANRIVGFFSKIIGEL--P---NIEADVNTFVASH
WGlbn         ----AS---DPAVADLGAKVLAQIGVAVSHL--GDQGKMVAQMKAVGVRH
LegHb         SAGVV---DSPKLGAHAEKVFGMVRDSAVQLATGEWLDGKD-----GSIH

     1                    Pattern7                  Pattern8
     2        ............L.........................2...........
     3        ............F**1***11.................A1**11......
     4        hhtt       .hhhhhhhhhhhhhhhhh tttt.hhhhhhhhhhhhhhh
SWMb          ATKH--KIPIKYLEFISEAIIHVLHSRHPGDFGADAQGAMNKALELFRKD
HaHb          AHKL--RVDPVNFKLLSHCLLVTLAAHLPAEFTPAVHASLDKFLASVSTV
IGlbn         KPRG---VTHDQLNNFRAGFVSYMKAHT--DFA-GAEAAWGATLDTFFGM
WGlbn         KGYGNKHIKGQYFEPLGASLLSAMEHRIGGKMNAAAKDAWAAAYADISGA
LegHb         IQKG--VLDPH-FVVVKEALLKTIKEASGDKWSEELSAAWEVAYDGLAT-

     1        Pattern9
     2            1
     3        1***Y
     4        hhhhhhhht...
SWMb          IAAKYKELGYQG
HaHb          LTSKYR------
IGlbn         IFSKM-------
WGlbn         LISGLQS-----
LegHb         AIKAA-------
```

Fig. 1. Sequences of sperm whale myoglobin (SWMb), human alpha hemoglobin (HaHb), insect globin (IGlbn), worm globin (WGlbn) and bean leghemoglobin (LegHb). The four rows above the sequences contain the pattern numbers (row 1), the alternate residues in patterns (row 2), the database patterns (row 3) and the secondary structure (row 4; h for helix, s for sheet and t for turns). These sequence alignments have been described in the literature [18,23].

alanine (A) in pattern 8. In addition, some sequences have the same residue at a particular position and only rarely an alternate, but many times a similar group residue in the same position in another sequence in that family. In this case, the alternate residue is recorded along with the most frequent residue in that pattern. Examples are shown in Fig. 1 for S,T (pattern 1) F,L (pattern 2), F,V (pattern 3) H,1 (pattern 4, histidine and nonpolar group), L,F (pattern 7), A,2 (pattern 8, alanine and small/polar group) and Y,1 (pattern 9, tyrosine and nonpolar group). In some isolated examples, unusual residues are found in otherwise nonpolar hydrophobic core conserved regions. These residues are included in the database and are identified by using this residue as an alternate in a pattern. One example is the unusual substitution of serine in pattern 1 for leghemoglobin (see Fig. 1). Note that leghemoglobin (LegHb) in this multiple sequence alignment has the sequence VNSSSQLF (pattern 1***S**1, homologous to the highly conserved pattern 1***1**1 in most of the sequences for this region in this family). Serine is permitted as an alternative residue in this pattern. Another example is shown by the substitution of lysine for a highly conserved nonpolar residue in pattern 3 (KF**1, homologous to the highly conserved pattern 1F**1) and in pattern 5 (1***1**5, homologous to the highly conserved pattern 1***1**1). Lysine or group 5 (KR) is permitted as an alternative residue in these positions.

There are two types of patterns called HCore and ID patterns. HCore regions are defined as all conserved patterns observed in the multiple sequence alignment. Not all protein sequences in this alignment will contain these patterns. For example, in leghemoglobin (Fig. 1, pattern 6), the hydrophobic residues (group 1) for the pattern 1**1***H are missing. ID patterns are defined as a subset of the HCore patterns which are found in all sequences in that family. For the globin family (see Fig. 1) patterns 2, 3 and 5 are found in all myoglobins, hemoglobins, leghemoglobins and globins. A new sequence is identified as a member of a known family if all of the ID patterns are located in the new sequence.

All patterns are recorded as a series of group numbers, exact residue names or asterisks which represent any amino acid residue. Database patterns for the globin family are shown as they appear in the database in Table 2. Brackets are used to represent alternate groups or exact residues.

Most family patterns in the database were composed of five to fifteen of these conserved patterns. Examples of the patterns are shown in Table 2 for myoglobin, cytochrome and serine protease families. Both of the ID and HCore patterns are identified in this table. Figures 1–3 show the same patterns and how they were derived from the multiple alignment of representative sequences from globins, cytochromes and serine proteases.

Search method to find homology to a family

Each new sequence was searched for conserved patterns in the following manner. The first pattern for the first family of the HomologyPlot database was searched for a match in the new sequence. If it was not found, the second database pattern in that family was searched for a match from the beginning of the new sequence. This was repeated until the first matching pattern was found. At that position in the new sequence, the search continued along the sequence until the next consecutive pattern of the first family in the database was found. It is important to note that only the highest number of consecutive patterns were scored, since it is assumed that the order as well as the largest matching number of patterns found will identify the correct family. This search can be run using either the ID or HCore patterns for a family.

The search was then repeated with the first pattern of the second family of the HomologyPlot

TABLE 2
FAMILY PATTERNS IN THE DATABASE FOR GLOBIN, CYTOCHROME AND SERINE PROTEASE PROTEIN SEQUENCES

	HCore	ID
Globin patterns		
1	1S(T)**3***1***1(S)**1	
2	1**1(H)F(L)***P	1**1(H)F(L)***P
3	1(K)F(V)**1	1(K)F(V)**1
4	H(1)G(2)**11	H(l)G(2)**11
5	1***1**1(5)	
6	1**1***H	1**1***H
7	F(L)**1***11	
8	A(2)1**11	
9	1***Y(1)	
Cytochrome patterns		
1	G(A)D***G(P)**11(S)	
2	C(A)**CH	C(A)**CH
3	1GP*1**11(K)	
4	1***1(3)1(K)	
5	1*W(I)*3**1	
6	1**P(G)**1(G)1	
7	2***M	2***M
8	11*11K(1)	
Serine protease patterns		
1	1*G(3)G	
2	P1*1*1	
3	C2**11	
4	11*22HC	11*22HC
5	1**G	
6	1***1*HP*1	
7	D1*11*1	
8	1(2)*1P	
9	1*GW(1)G(T)	
10	1**1*1*1	
11	1C*G	
12	GDSGG	GDSGG
13	G1121G(1)	
14	1**W(I)I(L)***1	

Group 1 residues are represented by 1, group 2 by 2, group 3 by 3 and group 5 by 5. Any residue is defined by an asterisk and alternate residues are given in brackets.

database from the beginning of the new sequence. This procedure was repeated until all of the families in the database had been searched. A successful search was found when all of the ID patterns of a family were found in the new sequence and the highest score or number of patterns was recorded.

Circular homology

In special cases, for example the lectin family, circular homology is found. This version of HomologyPlot allows one to cycle through all of the patterns to search for a higher homology score. In other words, the highest homology may not begin at the N-terminus of the family database pattern, but rather somewhere in the middle of the database patterns. By sequentially starting the search at the next pattern in the sequence of database patterns and continuing with the remaining patterns, all combinations of circular homology can be searched. For example, with a database family pattern that contained three patterns (1,2,3), circular homology will repeat the homology pattern search three times by looking for patterns 123, 231 and 312 in a new sequence.

RESULTS AND DISCUSSION

Amino acid substitutions

Most of the current methods to detect distant homologies in proteins rely on an amino acid pair score matrix. Higher scores are assigned to similar amino acid pairs than to dissimilar pairs in an alignment. The simplest is the unitary matrix (UM), where identical residues are assigned a value of one and nonidentical residues a value of zero. The genetic code matrix (GCM) is related to the maximum number of identities in the nucleotides coding for a pair of residues in an alignment. The most commonly used matrix is based on the mutations observed in families of closely related sequences and represents the probability of an exchange due to a mutation as a percent accepted mutation (PAM) [68]. It is instructive at this point to compare the 250-PAM and the HomologyPlot amino acid similarities as listed in Table 1. In HomologyPlot, group 1 (or nonpolar) residues are grouped as the residues VILNWYMCAT. This grouping is similar to the HIGH PAM grouping for the residues VILWFYMC. The exception is the residue R, which in

```
      1    Pattern1       Pattern2          Pattern3
      2    .A....P...S....A.......................K..........
      3    .GD***G**11....C**CH...........1GP*1**11..........
      4    ...hhhhhhhhhhh.tttt....tt...........tt.tt......tt.
Tuna       -GDVAKGKKTFVQK-CAQCHTVENGGKH--KVGPNLWGLFGRKTGQAEGY
Human      -GDVEKGKKIFIMK-CSQCHTVEKGGKH--KTGPNLHGLFGRKTGQAPGY
Rubrm      EGDAAAGEKVSK-K-CLACHTFDQGGAN--KVGPNLFGVFENTAAHKDNY
PA         ----EDPEVLFKNKGCVACHAI--EKK---MVGPAYKDVAAKFAGQA-GA
C6         -GDIANGEQVFTGN-CAACHS---VQQQKTLELSSLWKAK---SYLANFN

      1         Pattern4  Pattern5    Pattern6        Pattern7
      2    .......3K........I...........G..G.................
      3    ...12**11......1*W*3**1...1**P**11........2***M...
      4    .....hhhhh.........hhhhhhhhh.hhhh.........tt......
Tuna       S--YTDANKS---KGIVWNNDTLMEYLENPKKYI--------PGTKM-IF
Human      S--YTAANKN---KGIIWGEDTLMEYLENPKKYI--------PGTKM-IF
Rubrm      A--YSESYTEMKAKGLTWTEANLAAYVKNPKAFVLEKSGDPKAKSKM-TF
PA         EAELAQRIKNG--SQGVW------------------------GPIPM-PN
C6         GDE----------SAIVYQVTN--------------------GKNAMPAF

      1              Pattern8
      2                    1
      3    ...........11*11K
      4    .....hhhhhhhhhhhhhh
Tuna       AGIKKKGERQDLVAYLKSATS-
Human      VGIKKKEERADLIAYLKKATNE
Rubrm      K-LTKDDEIENVIAYLKTLK--
PA         A--VSDDEAQTLAKWVLSQK--
C6         GGRLEDDEIANVASYVLSKAG-
```

Fig. 2. Sequences of cytochrome C from tuna (Tuna), human (Human), *R. rubrum* (Rubrm), *Pseudomonas Aero* (PA) and C6 (C6). The four rows above the sequences have the same meaning as in Fig. 1. These sequence alignments have been described in the literature [23,41].

```
     1      Pattern1    Pattern2         Pattern3   Pattern4
     2      ..3...............................................
     3      1*GG........P1*1*1...........C2**11....11*22HC....
     4      ....ss..tt..ttssss.........sssssssssttsssss.......
Trpsn       IVGGYTCGANTVPYQVSLNS-----GYHFCGGSLINSQWVVSAAHCYKS-
Elast       VVGGTEAQRNSWPSQISLQYRSGSSWAHTCGGTLIRQNWVMTAAHCVDRE
MCP         IIGGVESIPHSRPYMAHLDIVTEKGLRVICGGFLISRQFVLTAAHCKG--
SGT         VVGGTRAAQGEFPFMVRLS--------MGCGGALYAQDIVLTAAHCVSGS
Tonin       IVGGYKCEKNSOPWQVAVIN------EYLCGGVLIDPSWVITAAHCYSN-

     1           Pattern5             Pattern6
     2      ..................................................
     3      .......1**G...............1***1*HP*1..............
     4      ......ssss....tt.......sssssssss.tt..tttt.........
Trpsn       ----GIQVRLGQDNINVVEGNQQFISASKSIVHPSYNSNTL---------
Elast       ---LTFRVVVGEHNLNQNNGTEQYVGVQKIVVHPYWNTDDVA--------
MCP         ---REITVILGAHDVRKAESTQQKIKVEKQIIHESYNSVPN---------
SGT         GNNTSITATGGVVDLQS--GAAVKVRSTKVLQAPGYNGT-----------
Tonin       ----NYQVLLGRNNLFKDEPFAQRRLVRQSFRHPDYIPLIVTNDTEQPVH

     1          Pattern7           Pattern8          Pattern9
     2      .......................2....................1T....
     3      ....D1*11*1............1*1P..............1*GWG....
     4      ...t..sssss............................sssssss....
Trpsn       --NNDIMLIKLKSAASLNSRVASISLP-T--SCASAGTQCLISGWGNTKS
Elast       -AGYDIALLRLAQSVTLNSYVQLGVLPRA-GTILANNSPCYITGWGLTRT
MCP         --LHDIMLLKLEKKVELTPAVNVVPLP-SPSDFIHPGAMCWAAGWGKTGV
SGT         --GKDWALIKLAQPIN----QPTLKIA-T-TTAYNQ-GTFTVAGWGANRE
Tonin       DHSNDLMLLHLSEPADITGGVKVIDLP-T--KEPKVGSTCLASGWGSTNP

     1             Pattern10                    Pattern11
     2      ..................................................
     3      ........1**1*1*1.....................1C*G.........
     4      .........sssssss..hhhhhh..h.ttt...ttssss..........
Trpsn       SGTSYPDVLKCLKAPILSNSSCKS--AYPGQIT-SNMFCAGYL-QGGKDS
Elast       N-GQLAQTLQQAYLPTVDYAICSSSSYWGSTVK-NSMVCAGGD-G-VRSG
MCP         R-DPTSYTLREVELRIMDEKACVD--YR--YYEYKFQVCVGSP-TTLRAA
SGT         G-GSQQRYLLKANVPFVSDAACRS--AYGNELVANEEICAGYPDTGGVDT
Tonin       SEMVVSHDLQCVNIHLLSNEKCIE--TYKDNVT-DVMLCAGEM-EGGADT

     1       Pattern12           Pattern13
     2      ..........................1......................1
     3      ..GD*GGP.............G1121G...................1**W
     4      .tt.tt.sssst.....tsssssssss.......tt..sssss....hhh
Trpsn       CQGDSGGPVVCS-----GKLQGIVSWGSG--CAQKNKPGVYTKVCNYVSW
Elast       CQGDSGGPLHCLVN-GQYAVHGVTSFVSRLGCNVTRKPTVFTRVSAYIS̊W
MCP         FMGDSGGPLLCA-----GVAHGIVSYGHP-DAK---PPAIFTRVSTYVPT
SGT         CQGDSGGPMFRKDNADEWIQVGIVSWGYG--CARPGYPGVYTEVSTFASA
Tonin       CAGDSGGPLICD-----GVLQGITSGGATP-CAKPKTPAIYAKLIKFTSW

     1      Pattern14
     2      L........
     3      I***1....
     4      hhhhhh...
Trpsn       IKQTIASN-
Elast       INNVIASN-
MCP         INAVIN---
SGT         IASAARTL-
Tonin       IKKVMKENP
```

Fig. 3. Sequences of serine proteases from trypsin (Trpsn), elastase (Elast), mast cell protease (MCP), *Streptomyces griseus* (SGT) and tonin (Tonin). The four rows above the sequences have the same meaning as in Fig. 1. These sequence alignments have been described in the literature [19,23].

the 250-PAM score has a HIGH similarity to W. In HomologyPlot, group 2 (or small and polar) residues are grouped as the residues ATGPSDN. Again, these residues are similar to the HIGH PAM grouping for the residues ATGPSDN, except for the residues N and D where the PAM grouping also includes the residues HEQK. Group 3, or the negatively charged and their amide counterpart residues, are grouped as the residues DNEQ. Although these residues have the highest similar scores to each other in the PAM matrix, the HomologyPlot assignment is more restrictive since the HIGH PAM grouping also includes HRKG. Group 4 contains the residues HQ. Again, this assignment is very different than the PAM matrix. However, it is interesting that the residues H and Q have the highest similarities for each other in the PAM matrix. Group 5, or the positively charged residues, are grouped as RK. This is probably the greatest difference between the PAM and HomologyPlot assignments, since the HIGH 250-PAM group for RK also includes the residues HWQN.

The grouping of amino acids in the HomologyPlot assignment is similar to that described by Go and Miyazawa [69], where they presented a dimensionless scale, ordered according to polarity and volume. They suggested that amino acid replacements during protein evolution are very conservative and are grouped as polar (DNEQHRK), weakly polar (ATGPS) and nonpolar (VILWFYMC). In addition, they state that there are more unvaried amino acid residues in the protein interior or core than on the surface. The group 1 assignment in HomologyPlot is the same as the nonpolar assignment of Go and Miyazawa, with the addition of A and T. The group 2 assignment in HomologyPlot is the same as the weakly polar assignment of Go and Miyazawa, with the addition of D and N. It is not unreasonable to assign these residues as small and weakly polar also. We have separated the polar grouping of Go and Miyazawa into three groups, according to size and charge. The separate grouping of DNEQ (group 3) and RK (group 5) in our classification is not unusual and is similar to the assignments given by Eisenberg [70], Kyte and Doolittle [71] and by Janin [72]. In their assignments, the values of DEQN are similar and different than the values assigned to RK. The separate assignment of H and Q into group 4 residues was made because of the observed conservative substitutions for this group of residues seen in proteins, particularly in globin sequences. It should be noted that the 250-PAM scores for H and Q are more similar to each other than to other residues in the 250-PAM scoring.

Several examples of the groups used in HomologyPlot are listed below. Group 1 contains the nonpolar or hydrophobic residues VILFYWMC. In addition, A and T are also included in this group since these residues were frequently observed in conserved regions where nonpolar residues normally occurred in a family of sequences. Examples where A was a conserved substitution in a nonpolar (group 1) region are shown in Fig. 1 (patterns 3 and 9), Fig. 2 (patterns 3 and 8), and Fig. 3 (patterns 5 and 10). Similarly, examples where T was a conserved substitution in a conserved region are shown in Fig. 2 (pattern 1) and Fig. 3 (pattern 10). There were numerous other examples where substitution of A and T occurred in nonpolar regions.

Group 2 contains the small and polar residues ATGPSDN. Examples where these residues can substitute for each other are shown in Fig. 1 (pattern 4) and Fig. 2 (patterns 4 and 7). It should be noted that the residues A and T are included in both groups 1 and 2, since these residues can substitute as either nonpolar, small or polar residues.

Group 3 contains the negatively charged residues D and E and their neutral analogs N and Q. Examples of similarities in conserved regions in this group are seen in Fig. 1 (pattern 1) and Fig. 2 (pattern 5).

Group 4 contains the residues H and Q. Histidine was considered as a separate group because of its unique charge, because it is frequently found in active site or metal binding regions, and because histidine and glutamine were frequently found to substitute for each other. In particular, the myoglobin family of sequences has several H and Q conserved substitutions in over 25 different sequences [23].

Group 5 contains the positively charged residues R and K. It is interesting that in several cases of nonpolar conserved regions, this group, especially lysine, was found to substitute for nonpolar residues. This is probably a result of the nonpolar characteristics of the methylene side chains of lysine. Examples of these substitutions are shown in Fig. 1 (patterns 3 and 5) and Fig. 2 (patterns 3, 4 and 8).

Conserved pattern database

Patterns are defined as a consecutive list of conserved residues or groups of residues. Nonconserved residues define the boundary of the conserved patterns. Database patterns usually consisted of five to fifteen conserved residue patterns for each family of protein sequences. It is important to note that these patterns are usually found in secondary structure regions such as helix, sheet or turn regions. For example, Fig. 1 shows that nine out of nine database patterns for the globin family are in helical regions. Figure 2 shows that five out of eight database patterns for the cytochrome family are in helical regions and three patterns are in turn regions. Figure 3 shows that 10 out of 14 database patterns for the serine protease family are in sheet regions, one pattern is in a helical region and one pattern is in a turn region. Since most of the residues in these patterns occur in secondary structure regions and most of these residues are nonpolar and probably part of the hydrophobic core for a family of sequences, these patterns are called hydrophobic core patterns. However, there are some examples where polar and charged residues are also part of a pattern. In addition, binding sites or active regions usually show conserved homology and are included in these database patterns.

There are two types of patterns in the database called ID (identification patterns) and HCore (hydrophobic core patterns). HCore patterns are all of the conserved patterns found in a family. Subgroups of a family, in some cases, may not contain all of these HCore patterns. HCore patterns are used to align the new sequence to a known sequence in that family. ID patterns are used to identify the family of proteins. ID patterns are a subset of HCore patterns and are found in all members of a family. ID patterns allow for a more rapid search of all the families to identify a homology and are the minimum set of patterns that identify that family. This version of HomologyPlot contains a total of 280 ID patterns that are unique for the 84 families in the database. Most of the family ID database contained 2–4 patterns (70%) for each family, while 16% contained 5–7 patterns and 14% contained only one pattern. None of the patterns are repeated in the ID database. Most of the individual patterns contained 4–12 residues (82%), while 18% of the patterns had more than 12 residues.

Several database patterns are shown for myoglobin (Fig. 1 and Table 2), cytochrome (Fig. 2 and Table 2) and serine proteases (Fig. 3 and Table 2). Lectins and immunoglobulins are examples of other families where a large number of sequences are available.

In the globin family (Fig. 1 and Table 2) there are a total of nine patterns. Patterns 2, 3, 4 and 6 are the ID patterns. Histidines which are conserved in most globins are found in patterns 4 and 6. It is interesting that worm globins (see Fig. 1), as well as a few other globins, do not have the conserved histidine in pattern 4. This is taken into account by assigning a nonpolar residue as the alternate residue in that pattern. The conserved phenylalanine, which is found in most globin sequences, is present in pattern 3.

In the cytochrome family (Fig. 2 and Table 2) there are a total of eight patterns. Patterns 2 and 7 are the ID patterns. The conserved cysteine and histidine, in most cytochrome sequences, are shown in pattern 2. There are examples where the first C in pattern 2 is sometimes an A. The conserved methionine, in most cytochrome sequences, is shown in pattern 7. The variability of sequences in the cytochromes is due to the grouping of long, medium and short subfamily sequences in the superfamily for cytochromes. Although most of the cytochromes have both ID patterns, many of the cytochromes do not have all of the HCore patterns. For example, cytochrome PA (Fig. 2) does not have patterns 1, 5 and 6. Cytochrome C6 (Fig. 2) does not have patterns 3, 5 and 6. There are many examples in the database where only a subset of the HCore patterns for a family database are found in the alignments. However, this is to be expected, since these missing patterns are due to insertions or deletions of residues for some sub-family members.

In the serine protease family (Fig. 3 and Table 2) there are a total of 14 patterns. Patterns 4 and 12 are the ID patterns. The conserved histidine and cysteine, in most serine proteases, are shown in pattern 4. The active site region GDSGG is shown in pattern 12. As for the cytochrome example, not all of the sequences in the serine protease family contained all 14 patterns in the database. In particular, patterns 1 and 14 were not found in many sequences of the serine protease family.

Search examples

The reliability and specificity of the database patterns to predict family homologies and alignments are shown below for several examples using the HomologyPlot program. In addition, a search of the entire National Biomedical Research Foundation (NBRF) Protein Information Resource (PIR) [73] database (October 1991 issue, 31 848 sequences) with representative database patterns is shown. The PIR database was also searched by family names to record the total number of sequences in that family. For example, all globin-like sequences were identified by searching for names of the known members of that family, myoglobin, hemoglobin, leghemoglobin and globins. All cytochrome sequences were identified by searching for the names cytochrome C, C1, C2, C6, C550, C551, C552, C553, C554 and C555. It should be noted that not all patterns are required to identify a new sequence as a member of a family, as is shown below. However, a minimum number of patterns, which we have called the ID patterns, must be found in a new sequence to identify it as a member of a family.

APL globin and Lamprey globin

Recently, *Aplysia limacina* (APL) globin has been described as having a low sequence homology to Lamprey globin (LAM) [74]. Both APL and Lamprey globin have low homology (20%) to sperm whale myoglobin. When these sequences were analyzed with HomologyPlot, they were quickly identified as belonging to the globin family, both with 100% ID homology and 88 and 100% HCore homology, respectively. The next closest family homologies listed were 50% ID and 60% HCore homologies. Figure 4 shows the automatic alignment of family patterns generated by HomologyPlot. This agrees very well with the alignment assigned in the literature, except for pattern 7, where HomologyPlot assigns a different alignment (Fig. 4, row labeled HOM shows the HomologyPlot pattern alignment). Note that HomologyPlot has identified all ID patterns

```
     1       Pattern 1             Pattern2   Pattern3
     2       .1S**3***1***1**1...........1**1F***P....1F**1....
APL          SLSAAEADLAGKSWAPVFANKNANGLDFLVALFEKFPDSANFFADFKGKS
LAM          PLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLT
HOM          PLSAAEKTKIRSAWAPVYSTYETSGVDILVKFFTSTPAAQEFFPKFKGLT

     1                   Pattern4  Pattern5        Pattern6
     2       .............H***11..1***1**1.........1**1***H....
APL          VAD-IKASPKLRDVSSRIFTRLNEFVNNAANAGKMSAMLSQFAKEHV-GF
LAM          TADQLKKSADVRWHAERIINAVDDAVASMDDTEKMSMKLRNLSGKHAKSF
HOM          TADQLKKSADVRWHAERIINAVDDAVASMDDTEKMSMKLRNLSGKHAKSF

     1                   Pattern7        Pattern8    Pattern9
     2       .............F**1***11........21**11......1***Y.
APL          GVGSAQFENVRSMFPGFVASVAAPPAGADAAWTKLFGLIIDALKAAGK
LAM          QVDPQYFKVLAAVIADTV-------AAGDAGFEKLMSMICILLRSAY-
HOM          -------QVDPQYFKVLAAVIADTVAAGDAGFEKLMSMICILLRSAY-
```

Fig. 4. Alignment of globin sequences to *Aplysia limacina* (APL) globin and Lamprey globin (LAM). The row labeled HOM is the alignment predicted by HomologyPlot. The Lamprey sequence comprised residues 10–149. The two rows above the sequences list the pattern numbers (row 1) and the database patterns (row 2). These sequence alignments have been described in the literature [74].

(Table 1, globin patterns 2, 3, 4 and 6). Also note that the HCore pattern 9 that is found in all myoglobin, hemoglobin and Lamprey sequences, is missing in the APL sequence. In addition to identifying the correct family homology, patterns 1–9 are also used to predict the structural alignment of APL and Lamprey globin to a known X-ray-derived structure of the same family. In this case sequences in patterns 1–8 for APL and patterns 1–9 for Lamprey globin (see Fig. 4) are superimposed on the sequences of patterns 1–9 for sperm whale myoglobin (see Fig. 1). These alignments agree with the structural alignments described in the literature [18,74].

There are a total of 385 hemoglobin and 60 myoglobin sequences in the PIR database. A search of this database with all nine HomologyPlot globin database patterns (Table 2) provided 442 sequences, 371 of which (96%) were hemoglobins, 60 (100%) were myoglobins and eight were other globin sequences. Only three other sequences were found, kinase-related transforming proteins, BPLF1 protein from Epstein–Barr virus and RNA replicase polyprotein. The reasons for the homology of these proteins to the globin family patterns must await comparisons with the X-ray-determined structures of these proteins. When only patterns 1, 2, 3, 4 and 6 (Table 2) were used to search the PIR database, 457 sequences were found, 382 of which (99%) were hemoglobin and 60 (100%) were myoglobin sequences. This shows that not all globin sequences have all patterns in the HomologyPlot database, since more sequences were found using only five patterns. This search also shows that even five out of the nine patterns for the globin database still provide a specific search for the globin family.

Cytochromes C551 and C2

Cytochromes C551 (*R. vannielii*) and C2 (*R. gelatinose*) are members of the cytochrome family with 50 and 10% sequence homology to tuna cytochrome C, respectively [23,41]. When these sequences are analyzed with HomologyPlot, they are identified as belonging to the cytochrome family, both with 100% ID homology. HCore homology is recorded as 62 and 100%, respectively. The closest homologies for other families were 50% ID and 50% HCore homologies. Figure 5 shows the automatic alignment of family patterns generated by HomologyPlot. Note that, even though patterns 1, 4 and 5 are missing for the C551 sequence, this sequence is still assigned to the cytochrome family. Note also that pattern 6 (Table 2, cytochrome pattern 6) matches both the C2 and C551 sequence. In most cytochrome sequences the pattern is 1**P**11. However, cytochrome C551 has two of the glycine substitutions in this pattern. In addition to identifying the correct family homology, sequences in patterns 1–8 (Fig. 5) for cytochromes C551 (*R. vannielii*) and C2 (*R. gelatinose*) are used to predict the structural alignment of these sequences to sequences in patterns 1–8 for an X-ray-derived structure of this family (tuna cytochrome C, Fig. 2). These alignments agree with the structural alignments described in the literature [41].

Thrombin and EGF binding protein

Thrombin (THROM) and EGF binding protein (EBP) are members of the serine protease family with 16 and 33% homology to bovine trypsin, respectively. These proteins are identified by HomologyPlot as serine proteases, both with 100% ID homology. HCore homology is recorded as 100 and 92%, respectively. The next closest family homologies were identified as having 50% ID and 60% HCore homologies. Figure 6 shows the automatic sequence alignment for family patterns of these proteins. Note that the family pattern 13 (Table 1, serine protease pattern 13) is missing in the EBP sequence and is due to a substitution of N for W. In addition to identifying the correct family homology, sequences in patterns 1–14 (Fig. 6) for thrombin and EBP are used to predict the structural alignment of these sequences to sequences in patterns 1–14 for an X-ray-derived structure of this family (trypsin, Fig. 3).

The motif GDSGGP is recognized as being a fingerprint for serine protease sequences [15]. A search of the PIR database for this motif provided 210 sequences containing the motif. However, only 26 of the 35 reported serine protease sequences [19] were found. A search for the more general sequence GDSGG motif, used in the HomologyPlot database, gave 215 sequences from the PIR database. In this case, 29 of the 35 reported serine protease sequences [19] were found. We have also searched the PIR database for the second HomologyPlot ID database pattern (see Table 2, pattern 4, 11*22HC). In this case, 212 sequences were found with 33 of the 35 known serine protease sequences identified. It is interesting that, when both of the ID patterns (Table 2) for the serine protease patterns are searched in the PIR database, 176 sequences were found with 29 of the 35 reported serine protease sequences identified. Analysis of the 35 known serine protease sequences with all of the HCore patterns identified only nine sequences. However, 31 of the family sequences contained 12 or more of the 14 HomologyPlot database patterns identified for the serine

```
      1     Pattern1    Pattern2        Pattern3
      2    .GD***G**11....C**CH.........1GP*1**1**5.....
CytcC2     AGDPVKGEQVFKQ--CKICHQVGPTAKNGVGPEQNDVFGQKAGAR
Cytc551    ----ATPAELATKAGCAVCHQPTA--K-GLGPSYQEIAKKYK---

      1     Pattern4  Pattern5    Pattern6 Pattern7
      2    ....1***N1..G**W*3**1...1**P**11.2***M.......
CytcC2     PGFNYSDANLASGLTWDEATLDKYLENPKAVV-PGTKM----VFV
Cytc551    ------------GQAGAPALMAERVRKGSVGIFGKLPMTPTPPAR

      1               Pattern8
      2    ..........11*111
CytcC2     GLKNPQDRADVIAYLKQLSGK
Cytc551    -IS-DADLKLVIDWILKTP--
```

Fig. 5. Alignment of cytochrome sequences for cytochrome C551 from *R. vannielii* (Cytc551) and cytochrome C2 from *R. gelatinosa* (CytcC2). The two rows above the sequences list the pattern numbers (row 1) and the database patterns (row 2). These sequence alignments have been described in the literature [23,41].

```
       1  Pattern1  Pattern2       Pattern3   Pattern4
       2  1*GG........P1*1*1.........C2**11....11*22HC......
Throm     IVEGQDAEVGLSPWQVMLFRKSPQELLCGASLISDRWVLTAAHCLLYPPW
EBP       VVGGFNCEKNSQPWQVAVYYQ--KEHICGGVLLDRNWVLTAAHCYVDQ--

       1           Pattern5              Pattern6
       2  ..........1**G................1***1*HP*1..........
Throm     NKNFTVDDLLVRIGKHSRTRYERKVEKISMLDKIYIHPRYNWKEN-----
EBP       --------YEVWLGKNKLFQ-EEPSAQHRLVSKSFPHPGFNMSLLMLQTT

       1         Pattern7           Pattern8        Pattern9
       2  ........D1*11*1............1*1P................1*G
Throm     -----LDRDIALLKLKRPIELSDYIHPVCLPDKQTAAKLLHAGFKGRVTG
EBP       PPGADFSNDLMLLRLSKPADITDVVKPIALP-T--KE--PKPGSTCLASG

       1                    Pattern10             Pattern11
       2  WG.................1**1*1*1..................1C*G.
Throm     WGNRRETWTTSVAEVQPSVLQVVNLPLVERPVCKASTRIRITNDMFCAGY
EBP       WGSITPTR-----WQKSDDLQCVFITLLPNENCAKVYLQKVTDVMLCAGE

       1            Pattern12           Pattern13
       2  ............GDSGG...............G1121G............
Throm     KPGEGKRGDACEGDSGGPFVMKSPYNNRWYQMGIVSWGEG-CDRNGKYGF
EBP       M---GGGKDTCAGDSGGPLICD------GILQGTTSNGPEPCGKPGVPAI

       1        Pattern14
       2  ......1**WI***1.....
Throm     YTHVFRLKKWIQKVIDRLGS
EBP       YTNLIKFNSWIKDTMMKNA-
```

Fig. 6. Alignment of serine protease sequences from thrombin (Throm) and EGF binding protein (EBF). The two rows above the sequences list the pattern numbers (row 1) and the database patterns (row 2). These sequence alignments have been described in the literature [19].

protease family. Most of the variations in the pattern homology were found in pattern 6 (see Fig. 6), where only 50% of the known members of this family contained that pattern.

Lectins and circular homology

HomologyPlot also allows for a circular homology search option. Using the concanavalin A sequence without the circular homology option, the lectin family homology is rated second at 50% family homology, which represents a distant or low homology. Circular homology is easily identified in any search by using one or two cycles of circular homology and by observing changes in the position (rating) of families listed. A move to a higher rating probably means some circular homology is present. A circular homology search for concanavalin A gives 94% homology to the lectin family.

Coiled-coil motif and leucine-zipper motif

We were interested in the motif for the coiled-coil tropomyosin structures. The pattern for this motif in the HomologyPlot database is 1**1***1**1***1**1***1**1***1**1 (32 residues), where 1 represents any of the residues VILFYA. This is based on a heptad repeat for coiled-coil sequences proposed earlier [75]. There are a total of 52 tropomyosin sequences in the PIR database. A search of the PIR database with the HomologyPlot coiled-coil motif provided 536 sequences, 52 of which were tropomyosin sequences. Examples of protein sequences other than tropomyosin were cytochrome oxidase and cytochrome b, NADH dehydrogenase, nicotinic acetylcholine receptor, H^+-transporting ATP synthase, DNA and RNA directed polymerase, myosin, Glyceraldehyde-3-phosphate dehydrogenase, keratin, genome polprotein, env protein, dopamine receptor, antifreeze protein, spike glycoprotein, hypothetical protein-trypanosoma, elastin, and transform proteins. Many of these are coiled-coil, like sequences reported previously [76]. It is interesting that a preliminary search of the PIR database with a coiled-coil motif of 67 residues gave only the following proteins: hypothetical protein from trypanosoma and elastin sequences. A similar attempt to use a 60-residue coiled-coil motif gave only 26 out of 52 known tropomyosin sequences. An analysis of these 26 tropomyosin sequences revealed that the longer heptad repeat of 67 residues was interrupted by the amino acid threonine at the C-terminus and by serine at the N-terminus of the proposed pattern for tropomyosin.

We have distinguished between coiled-coil and leucine-zipper-type motifs since only three leucine-zipper-like transforming protein sequences were found with the coiled-coil motif. There are several families of leucine-zipper proteins, one of which is the bZIP family [77] which is characterized by the basic region at the N-terminus of the leucine zipper. The HomologyPlot motif for leucine-zipper sequences is R(N)*******R*R(K)*******L(F)******1. There are a total of 60 sequences for the transforming proteins myc(38), fos(7), jun(10) and fra(5). A search of the database using the HomologyPlot bZIP motif provided 463 sequences, 29 of which were for myc, 7 for fos, 10 for jun and 5 for fra proteins. Examples of other homologous sequences were DNA directed RNA polymerases, ribulose biphosphate carboxylase, ribosomal proteins, nicotinic acetylcholine receptor, genome polyprotein, colicin, insulin-like growth factor and myosin.

L******L******L******L******L is a second leucine-zipper motif in the HomologyPlot database. A search of the PIR database with this HomologyPlot motif gave 283 sequences, comprising myc(2), fos(7), jun(8) and fra(3) transforming protein sequences in addition to tropomyosin, keratin, myosin and elastin sequences. It is interesting to note that, although this heptad repeat for leucine zippers does recognize fos, jun and fra proteins as well as coiled-coil proteins, only 2 of the 38 known myc sequences are found. Analysis of several myc sequences revealed that leucine in this motif was frequently substituted by alanine or valine. This kind of analysis, to find patterns or motifs, can be used to understand the sequence differences in common motifs, which can identify subclasses within a certain motif. A knowledge of the sequence patterns, for example in coiled-coil or leucine-zipper motif, will provide more information about the various functional differences of subclasses of homologous protein sequences.

CONCLUSIONS

The use of HomologyPlot to identify sequence homology to a known family of proteins has several important advantages. The first point is that, since many of these patterns were derived from multiple alignment of sequences containing X-ray-defined structures, a quick optimal alignment of conserved regions is achieved for a new sequence to that of the X-ray-defined structure for that family. Second, by using a database of multiple patterns to identify a homology alignment, the problems of insertions, deletions and gap scores used in other sequence alignment algorithms are avoided. Third, conserved regions and residues for a family are identified and define where insertions, deletions or mutations are not allowed in modeling procedures. Fourth, circular homology (for example lectins) can be easily and quickly identified. Fifth, these patterns can be used to rapidly search the entire protein sequence database for proteins that have not been characterized.

In addition, the database provides information and references to the protein families and related protein sequences. It is the ID pattern that uniquely identifies a family and should always be run first in a homology search. Even if the family is not identified conclusively, the patterns found can be used to identify subdomains or sequences found in other protein families. For most searches, 10–20 examples of pattern matches are found. This number can be scanned rapidly for significant homologies, compared to the several hundred sequences found by conventional homology algorithms, especially for sequences with low or distant similarity to known structures. HomologyPlot also allows searches for motifs. The HomologyPlot database is readily accessible to the user, and can be updated, modified and refined as more sequences and homology families are identified.

ACKNOWLEDGEMENTS

The program HomologyPlot is being marketed by Synthetic Peptides, Inc., University of Alberta, Medical Science Building, room 355, Edmonton, Alberta, Canada T6G 2H7. The program SEQSEE was used to search the PIR database to confirm the HomologyPlot patterns used in this paper and was made available by Dr. B. Sykes, D. Wishart and R. Boyko at the University of Alberta. These programs are an integral part of the work done at the Protein Engineering Networks of Centers of Excellence, located at the University of Alberta.

REFERENCES

1 Argos, P., Vingron, M. and Vogt, G., Protein Eng., 4 (1991) 375.
2 Schuler, G.D., Atshchul, S.F. and Lipman, D.J., Protein Struct. Funct. Genet., 9 (1991) 180.
3 Kauzman, W., Adv. Protein Chem., 14 (1959) 1.
4 Henrissat, B., Saloheimo, M., Lavaitte, S. and Knowles, J.K.C., Protein Struct. Funct. Genet., 8 (1990) 251.
5 Altschuh, D., Protein Eng., 2 (1988) 193.
6 Taylor, W.R., J. Mol. Biol., 188 (1986) 233.
7 Sweet, R.M. and Eisenberg, D., J. Mol. Biol., 171 (1983) 479.
8 Hubbord, T.J.P. and Blundell, T.L., Protein Eng., 1 (1987) 159.
9 Zielenkiewicz, P. and Rabczenko, A., Protein Eng., 2 (1988) 115.
10 Kanaoka, M., Kishimoto, F., Ueki, Y. and Umeyama, H., Protein Eng., 2 (1989) 347.
11 Umezawa, Y. and Umeyama, H., Chem. Pharm. Bull., 36 (1988) 4652.
12 Bowie, J.U., Clarke, N.D., Pabo, C.O. and Saver, R.T., Protein Struct. Funct. Genet., 7 (1990) 257.
13 Taylor, E.C., Horton, M.R. and Krause, P.R., Comput. Biomed. Res., 24 (1991) 72.
14 Sander, C. and Schneider, R., Protein Struct. Funct. Genet., 9 (1991) 56.
15 Sato, Y., Ikeuchi, Y. and Kanehisa, M., Protein Struct. Funct. Genet., 8 (1990) 341.
16 Barton, G.J. and Steinberg, M.J.E., J. Mol. Biol., 212 (1990) 389.
17 Argos, P., J. Mol. Biol., 193 (1987) 385.
18 Bashford, D., Chothia, C. and Lesk, A.M., J. Mol. Biol., 196 (1987) 199.
19 Greer, J., Protein Struct. Funct. Genet., 7 (1990) 317.
20 Pearson, W.R., Methods Enzymol., 183 (1990) 63.
21 Bairoch, A., Nucleic Acids Res., 19 (1991) 2241.
22 Aitken, A., Identification of Protein Consensus Sequences. Active Site Motifs, Phosphorylation and other Post-translational Modifications, Ellis Horwood, New York, NY, 1990, p. 152.
23 Dayhoff, M.D., Atlas of Protein Sequence and Structure, Vol. 5, Supplement 3, National Biomedical Research Foundation, Washington, DC, 1978.
24 Tellam, R.L., Morton, D.J. and Clarke, F.M., Trends Biochem. Sci., 14 (1989) 130.
25 Davies, P.L. and Hew, C.L., FASEB J., 4 (1990) 2460.
26 Webster, T.A., Lathrop, R.H. and Smith, T.F., Biochemistry, 26 (1987) 6950.
27 Nishi, M., Sanke, T., Nagamatsu, S., Bell, G.I. and Steiner, D.F., J. Biol. Chem., 265 (1990) 4173.
28 Scully, J.L. and Evans, D.R., Protein Struct. Funct. Genet., 9 (1991) 191.

29 Abad-Zapatero, C., Rydel, T.J. and Erickson, J., Protein Struct. Funct. Genet., 8 (1990) 62.
30 Gilliland, G.L., Winborns, E.L., Nachman, J. and Wlodawer, A., Protein Struct. Funct. Genet., 8 (1990) 82.
31 James, M.N.G. and Sielecki, A., Biol. Macro Assem., 3 (1987) 413.
32 Tang, J., James, M.N.G., Hsu, I.N., Jenkins, J.A. and Blundell, T.L., Nature, 271 (1978) 618.
33 Ryden, L. and Lundgren, J.O., Nature, 261 (1976) 344.
34 Murata, M., Richardson, J.S. and Sussman, D.J.L., Proc. Natl. Acad. Sci. USA, 82 (1985) 3073.
35 Luthy, R., McLachlan, A.D. and Eisenberg, D., Protein Struct. Funct. Genet., 10 (1991) 229.
36 Cowan, S.W., Newcomer, M.E. and Jones, T.A., Protein Struct. Funct. Genet., 8 (1990) 44.
37 Montal, M., FASEB J., 4 (1990) 2623.
38 Wistow, G.J., Pisano, M.M. and Chepelinsky, A.B., Trends Biochem. Sci., 16 (1991) 170.
39 Henneke, C.M., Danson, M.J., Hough, D.W. and Osguthorpe, D.J., Protein Eng., 2 (1989) 597.
40 Olivera, B.M., Rivier, J., Scott, J.K., Hillyard, D.R. and Cruz, L.J., J. Biol. Chem., 266 (1991) 22067.
41 Dickerson, R.E., Sci. Am., 242 (1980) 137.
42 Ganz, T., Selsted, M.E. and Lehrer, R.I., Eur. J. Haematol., 1 (1990) 1.
43 Sokolovsky, M., Trends Biochem. Sci., 16 (1991) 261.
44 Chin, C.C.Q., J. Protein Chem., 9 (1990) 427.
45 Baron, M., Norman, D.G. and Campbell, I.D., Trends Biochem. Sci., 16 (1991) 13.
46 Shoyab, M., Plowman, G.D., McDonald, V.L., Bradley, J.G. and Todaro, G.J., Science, 243 (1989) 1074.
47 Linder, P., Lasko, P.F., Ashbumer, M., Leary, P., Nielson, P.J., Nishi, K. and Schnier, J., Nature, 337 (1989) 121.
48 Ragland, M., Briant, J.F., Gagnon, J., Laulhere, J.P., Massenet, O. and Theil, E.C., J. Biol. Chem., 265 (1990) 18339.
49 Pastors, A. and Lesk, A.M., Protein Struct. Funct. Genet., 8 (1990) 133.
50 Hockenhull-Johnson, J.D., Stern, M.S., Martin, P., Dass, C., Desiderio, D.M., Wittenberg, J.B., Vinogradov, S.N. and Walz, D.A., J. Protein Chem., 10 (1991) 609.
51 Suzuki, T. and Furukohi, T., J. Protein Chem., 9 (1990) 69.
52 Clark, B.F.C., Jensen, M., Kjeldgaard, M. and Thirup, S., In Hook, J.B. and Poste, G. (Eds.) Protein Design and Development of New Therapeutics and Vaccines, New Horizons in Therapeutics, Smith Kline and French Laboratories Research Symposia Series, 1990, pp. 179–208.
53 Mason, A.J., Hayflick, J.S., Ling, N., Esch, F., Ueno, N., Ying, S.Y., Guillemin, R., Niall, H. and Seeburg, P.H., Nature, 318 (1985) 659.
54 Blundell, T.L. and Humbel, R.E., Nature, 287 (1980) 781.
55 Priestle, J.P., Schar, H.P. and Grutter, M.G., Proc. Natl. Acad. Sci. USA, 86 (1989) 9667.
56 Caldwell, J.B., Strike, P.M. and Kortt, A.A., J. Protein Chem., 9 (1990) 493.
57 Onesti, S., Brick, P. and Blow, D.M., J. Mol. Biol., 217 (1991) 153.
58 Sharon, N. and Lis, H., FASEB J., 4 (1990) 3198.
59 Laskowski Jr., M., Apostol, I., Ardelt, W., Cook, J., Gilleto, A., Kelly, C.A., Lu, W., Park, S.J., Qasim, M.A., Whadey, H.E., Wieczorek, A. and Wynn, R., J. Protein Chem., 9 (1990) 715.
60 Katzin, B.J., Collins, E.J. and Robertus, J.D., Protein Struct. Funct. Genet., 10 (1991) 251.
61 Barbacid, M., Annu. Rev. Biochem., 56 (1987) 779.
62 Santos, E. and Nebreda, R., FASEB J., 3 (1989) 2151.
63 Toh, H., Ono, M., Saigo, K. and Miyata, T., Nature, 315 (1985) 691.
64 Pechik, I.V., Gustchina, A.E., Antireever, N.S. and Fedorov, A.A., FEBS Lett., 247 (1989) 118.
65 Umezawa, Y. and Umeyama, H., Chem. Pharm. Bull., 36 (1988) 4652.
66 James, M.N.G., Delbare, L.T.J. and Brayer, G.D., Can. J. Biochem., 56 (1978) 396.
67 Eklund, H., Gleason, F.K. and Holmgren, A., Protein Struct. Funct. Genet., 11 (1991) 13.
68 George, D.G., Barker, W.C. and Hunt, L.T., Methods Enzymol., 183 (1990) 333.
69 Go, M. and Miyazawa, S., Int. J. Pept. Protein Res., 15 (1980) 211.
70 Eisenberg, D., Weiss, R.M., Terwilliger, T.C. and Wilcox, W., J. Chem. Soc., Faraday Symp., 17 (1982) 105.
71 Kyte, J. and Doolittle, R.F., J. Mol. Biol., 157 (1982) 105.
72 Janin, J., Nature, 277 (1979) 491.
73 Barker, W.C., George, D.G. and Hunt, L.T., Methods Enzymol., 183 (1990) 31.
74 Pastore, A., Lesk, A.M., Bolognesi, M. and Onesti, S., Protein Struct. Funct. Genet., 4 (1988) 240.
75 Hodges, R.S., Sodek, J., Smilie, L.B. and Jurasek, L., Cold Spring Harbor Symp. Quant. Biol., 37 (1972) 299.
76 Lupas, A., Van Dyke, M. and Stock, J., Science, 252 (1991) 1162.
77 Kerppola, T.K. and Curran, T., Curr. Opin. Struct. Biol., 1 (1991) 71.