Protein Engineering vol.5 no 6 pp.479–488, 1992

Construction of a dictionary of sequence motifs that characterize groups of related proteins

Atsushi Ogiwara, Ikuo Uchiyama, Yasuhiko Seto[1] and Minoru Kanehisa[2]

Institute for Chemical Research, Kyoto University, Uji, Kyoto 611 and [1]Protein Research Foundation, Peptide Institute, Ina, Minoh, Osaka 560, Japan

[2]To whom correspondence should be addressed

An automatic procedure is proposed to identify, from the protein sequence database, conserved amino acid patterns (or sequence motifs) that are exclusive to a group of functionally related proteins. This procedure is applied to the PIR database and a dictionary of sequence motifs that relate to specific superfamilies constructed. The motifs have a practical relevance in identifying the membership of specific superfamilies without the need to perform sequence database searches in 20% of newly determined sequences. The sequence motifs identified represent functionally important sites on protein molecules. When multiple blocks exist in a single motif they are often close together in the 3-D structure. Furthermore, occasionally these motif blocks were found to be split by introns when the correlation with exon structures was examined.

Key words: consensus patterns/genome analysis/homology search/multiple sequence alignment/prediction of protein function

Introduction

When the amino acid sequences of two proteins are similar, they probably belong to the same group of functionally related proteins. Thus, when a new protein sequence is determined, it is customary to perform a database search for similar sequences in the hope of obtaining a clue to its biological function. The search involves pairwise comparisons against individual sequences in the database. This is becoming more time-consuming with the rapid growth in database size. An alternative approach is to search a library of signature patterns, each of which uniquely identifies a group of related proteins. Whether all protein groups can be represented by such diagnostic patterns is arguable, but this approach is certainly more effective because the comparison is made against individual groups rather than individual sequences in the database.

It is common knowledge that functionally important sites are well conserved in the amino acid sequences of related proteins. Conserved regions are not necessarily contiguous in the primary structure, because a functional site in the 3-D structure can be composed of separate pieces of conserved segments. The conserved amino acid patterns, often called consensus patterns or sequence motifs (Taylor, 1988; Hodgman, 1989), are usually identified by the tedious method of multiple aligning and comparing a group of functionally related sequences. These published motifs are then manually collected, verified and organized in a motif library (Bairoch, 1989; Seto *et al.*, 1990). An additional constraint to the conserved regions is introduced in this study: the uniqueness of amino acid patterns when compared with all other sequences outside the group. This has enabled the design of an automatic procedure to define from the protein sequence database a collection of signature patterns that uniquely identify specific protein groups. This procedure is applied to the superfamily grouping of the PIR database and a library of sequence motifs is constructed that identifies specific superfamilies.

Materials and methods

Databases

The amino acid sequences were obtained from the PIR database release 26.0 (September 1990). The PIR database is divided into three sections (two sections before release 26.0): PIR1, annotated and classified entries; PIR2, preliminary entries; and PIR3, unverified entries. Only the PIR1 section is used when constructing a motif library. The releases of 19.0 (December 1988) to 29.0 (June 1991) were also used for comparison purposes. The 3-D coordinates of the protein structures were acquired from the Brookhaven Protein Data Bank (April 1991).

Functional groups of proteins

Suppose that a protein sequence database is divided into groups, each containing functionally related members, and that the diagnostic amino acid patterns that uniquely identify the membership to each functional group are required. The PIR superfamily classification is used to define a protein group, but there may also be other definitions. A superfamily is a group of proteins bearing significant sequence similarity and represents the probable evolutionary relationships of the proteins (Dayhoff, 1978; Dayhoff *et al.*, 1983). It is not always the case that a protein is uniquely assigned to one superfamily, because it can contain multiple domains with different functions. For simplicity, however, the PIR superfamily numbering scheme is used, which assumes that each protein in the database belongs to one, and only one, superfamily.

Dictionary of unique peptide words

A three-step procedure is employed to identify the sequence motifs. The first step involves an exhaustive search for unique peptide words (UPWs) which, in our definition, are short oligopeptide patterns that are well conserved and found exclusively in one protein group. A group is usually a single superfamily, but it can be extended to comprise a few superfamilies. In practice, as illustrated in Figure 1, we make a tally of all possible tetra-, penta- and hexapeptide patterns in the superfamilies of the PIR database. Let n_s and N_T be the numbers of sequences containing a given pattern in a given superfamily and in the entire database respectively. The pattern is unique to this superfamily when $n_s = N_T$. The pattern is conserved when $n_s = N_T \geq f \cdot m$, where m is the number of members belonging to the superfamily and f is the parameter defining the majority. We consider different cases ranging from $f = 1$ (100% conservation) to $f = 0.7$ (70% conservation). Although the distinction between 100 and 70% conservation is highly dependent on the superfamily size and variability of its members, the uniqueness is mostly determined by the size and variability of the entire database.

	Superfamily	1	...	94	95	96	97	...	
m	# Sequences	119	...	11	12	8	1	...	Total
	AAAA								
	:								
	QWYW	0	...	0	12	0	0	...	12
	:								
	WHFV	0	.	0	0	7	0	..	7
	:								
	VVVV								
				n_s					N_T

Fig. 1. Screening of unique peptide words. This figure shows the numbers of sequences containing given tetrapeptide patterns. The superfamily 95 has 12 member sequences and all contain the pattern QWYW, while all other sequences outside this superfamily do not possess this pattern. Thus, this pattern is unique to, and conserved in, the superfamily 95. The unique pattern WHFV is not 100% conserved in superfamily 96, but this pattern can be detected by setting a lower threshold value for the conservation

Consensus of unique peptide sentences

In the second step the order of unique peptide words in each sequence of a given group is examined and a consensus pattern constructed. As illustrated in Figure 2, each amino acid sequence is converted to an abstract structure, which may be called unique peptide sentences, consisting of the UPW pattern number and the number of residues separating the first residues of two successive UPWs. One amino acid mutation is allowed when searching for the occurrence of each UPW pattern. When the separation is smaller than the length of the preceding UPW there is actually an overlap between the two UPWs, as in patterns 3 and 4 in Figure 2. From a set of these sentences, some of which may lack specific UPWs and some of which may contain duplicates of the same UPW, a consensus sentence is constructed. This is a multiple alignment problem and an approximate procedure was devised by combining pairwise alignments. The optimal pairwise alignment can be obtained by the following dynamic programming algorithm which is similar to the RNA secondary structure prediction algorithm (Waterman and Smith, 1978; Kanehisa and Goad, 1982):

$$S_{i,j} = \max(S_{i,j} - \text{g.p.}(i,j,k,l) + w(i,j))\ (k < i,\ l < j,\ P_k = P_l)$$

where $S_{i,j}$ is the score up to the ith pattern P_i and jth pattern P_j, g.p. is the gap penalty and w is the weight for a match of two patterns. The resulting consensus pattern is represented by the order of UPWs with the upper- and lower-bound numbers of residues separating two successive UPWs (Figure 2).

Refinement of consensus

The consensus pattern obtained in the previous step is represented by the blocks of amino acid patterns, which we call motif blocks, separated by the upper- and lower-bound numbers of residues in the space region as follows:

< motif block1 > [min_spacer, max_spacer] < motif block2 >

As shown in Figure 3, this consensus is used again in the last step to compare each sequence in the group, to identify substitution patterns and to determine whether each block is conserved in all sequences. In practice, it is first decided whether a particular block exists or not, given the minimum fraction of matched residues, r, that constitute a block. Then, all substitution patterns are recorded. In the representation of our motif library, the plus sign designates that the block is conserved in all members of the group, while the minus sign indicates that some members lack the block. Substitution patterns are enclosed in braces.

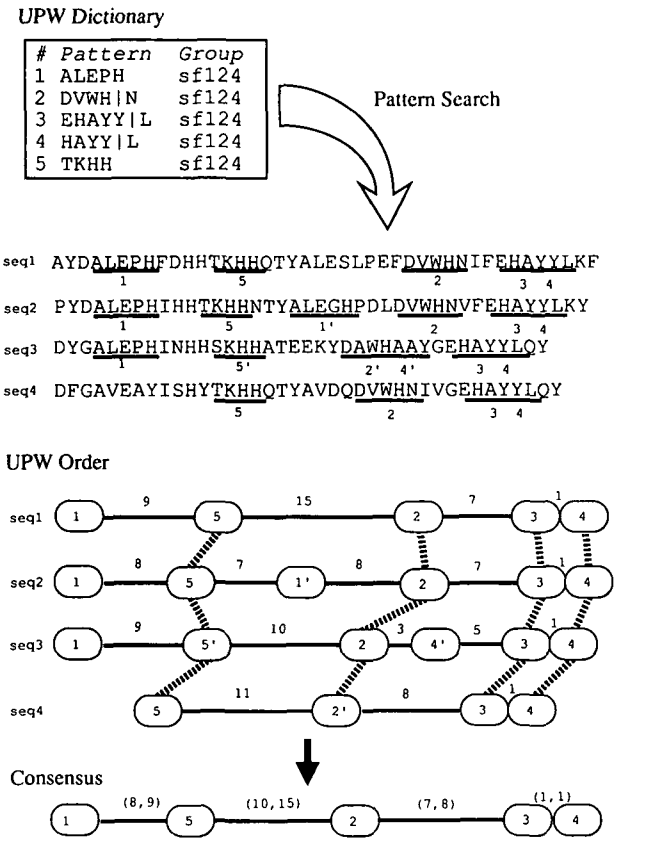

Fig. 2. An illustration of how the sequence motif is constructed from unique peptide words. First, the locations of unique peptide words in a given superfamily are examined for all member sequences. Then the consensus ordering of unique peptide words is obtained by a dynamic programming algorithm.

Results

Construction of a motif library

The PIR1 database release 26.0 contains 7235 sequences, totalling 2 221 416 residues, classified into 2350 superfamilies. The relatively large superfamilies that contained a set number of member sequences were considered. When the minimum value for the size of a superfamily in release 26.0 was defined as three or five members, there were 521 or 283 superfamilies respectively. As summarized in Table I, our procedure identified sequence motifs that characterized >50% of these superfamilies when the degree of conservation, f, was set at 80 or 70%. The motif library constructed with the minimum superfamily size of five members and $f = 80\%$ contained 145 sequence motifs (Table I). Out of the 145 motifs, 35 were characterized by single blocks while the rest contained multiple blocks, as shown in Table II. A complete listing of the 121 motifs containing < 10 blocks is shown in Appendix. Substitution patterns are obtained when $r = 80\%$.

As each new release of the PIR1 database is produced, the motif library can be reconstructed by this automatic procedure. However, a long computation time is required because of the calculation of the many hexapeptide patterns in the initial screening of the UPWs. When the libraries shown in Table I were constructed without hexapeptide patterns, ~5% of the superfamilies could not be identified. This was a relatively small loss compared with the gain in computation time.

Superfamily assignment by sequence motifs

A procedure for superfamily assignment was established utilizing

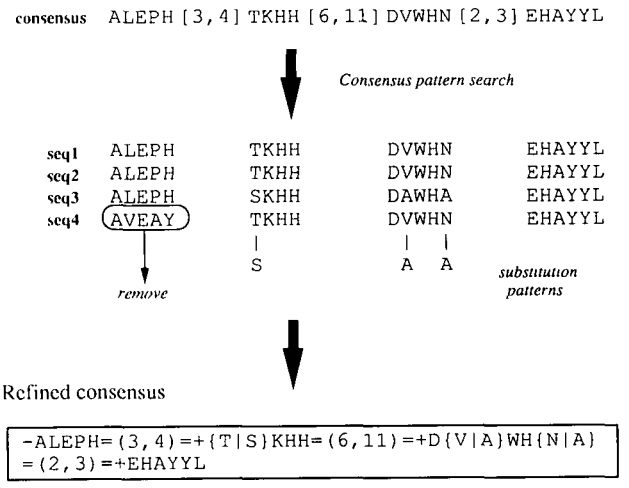

Fig. 3. Refinement of the sequence motif. The consensus pattern derived in the previous step is used to match against each member sequence of the superfamily. For those blocks that have more matches than the threshold value, substitution patterns are recorded.

our motif library, as follows: (i) begin the search using the first motif block. The criterion for the existence of a motif block is given by the parameter r, which specifies the minimum fraction of matched residues; (ii) if a motif block is found, check if the next motif block exists after the specified spacer length; and (iii) if a motif block is not found, skip this and continue searching for the next block. The search fails if no motif block is found. In the above procedure a sequence is considered assigned to a superfamily if any of the motif blocks match. No distinction is made between the conserved (+) and nonconserved (−) blocks.

Table III(a) shows the results of this procedure when applied to the PIR1 database release 26.0, which is the training data set used for constructing the motif library. When the block detection parameter $r = 100\%$, no entries were falsely assigned (false positives), but 140 entries could not be detected (false negatives) as belonging to one of the 145 superfamilies. At the level of $r = 80\%$ there were 70 false positives and 79 false negatives. When false positives were examined in more detail, all resulted from single motif blocks containing substitution patterns. Sequence motifs with multiple blocks or sequence motifs with single blocks without substitution patterns could be used safely for superfamily assignment.

Next, a test data set was prepared from release 29.0 of the PIR1 database by identifying new entries added after release 26.0. There were cases where several entries in multiple superfamilies were combined into a single superfamily or entries in a single superfamily were split into different superfamilies. In such cases, the multiple superfamilies are considered to be related and assignment to a related superfamily is the correct answer. The results using this new data set are summarized in Table III(b). Although the prediction ability (~68%) was not as great as had been expected, the search itself could be performed within a fraction of a second on a small workstation, which is two to three orders of magnitude faster than the FASTA homology search (Pearson and Lipman, 1988).

We modified the above procedure and stopped the search if any of the conserved (+) blocks were not found. The number of false positives could be decreased without affecting the number of false negatives in Table III(a) because this is how the conserved block was defined in the training set. However, this additional constraint has more effect on increasing the number of false negatives than decreasing the number of false positives in the test set of new sequence data in Table III(b) (data not shown).

If the motif library is to be used as an initial step in superfamily assignment, it is desirable to decrease the number of false negatives because false positives can easily be distinguished by sequence similarity in the subsequent step. There are still ~20% of false negatives in Table III(b), even with low values of r. It is possible to halve this by incorporating amino acid similarity scores, such as the PAM matrix (Dayhoff, 1978), when comparing motif blocks (data not shown).

Functional and structural aspects of sequence motifs

Because the sequence motifs identified represent well conserved regions within a group of related proteins, they are likely to correspond to functionally important sites. Table IV summarizes the percentages of biological sites, annotated in the PIR1 database, which correspond to motif blocks identified by our procedure. Table V is a listing of the single block sequence motifs that

Table I. Number of superfamilies (SFs) characterized by sequence motifs

Minimum SF size	No. of SFs	f (%)	No. of SFs characterized
5 members	283	80	145 (51%)
		70	162 (57%)
3 members	521	80	287 (55%)
		70	324 (62%)

Table II. Number of blocks constituting sequence motifs that characterize superfamilies

No of blocks	No. of SFs
1	35
2	27
3	12
4	10
5	12
6	10
7	7
8	5
9	3
≥10	24
Total	145

Minimum SF size: 5 members; $f = 80\%$.

Table III. Results of superfamily assignment

r	50%	80%	100%
(a) PIR1 release 26.0 (training set)			
True	1539	1480	1419
False positive	318	70	0
False negative	20	79	140
No opinion	5475	5635	5676
Total	7235	7235	7235
(b) new data in PIR1 release 29.0 (test set)			
True	175	172	161
False positive	66	37	28
False negative	40	43	54
No opinion	827	926	944
Total	1190	1190	1190

characterize 35 superfamilies, together with any known functional significance. Our procedure identified known consensus patterns, or closely related derivatives, such as the active site sequence GDSGG which is known to be exclusive to the serine protease superfamily (Dayhoff *et al.*, 1983).

The sequence motifs were obtained strictly from 1-D sequence information, the superfamily classification based on sequence similarity and the amino acid pattern searches. Among the 145 superfamilies with identified motifs, 21 superfamilies contained one or more member sequences with known 3-D structures; seven were characterized by single block motifs and 14 by multiple block motifs. Using the coordinate data from the Brookhaven Protein Data Bank (Bernstein *et al.*, 1977), it has been determined that multiple motif blocks come closer together in the 3-D structure. Typical examples are: L-lactate dehydrogenase (SF31; see Appendix for actual motifs), phosphoglycerate kinase (SF229), phospholipase A2 (SF281), neutral proteinase (SF385), carbonate dehydratase (SF472) and triose-phosphate isomerase (SF499). Figure 4 shows a stereo drawing of phospholipase A2 with two motif blocks at the active site.

Table IV. Correspondence of motif blocks with biological features

PIR feature	No. of entries	No. of sites	No. of hits	%Hits
Active site	79	185	58	31.35
Binding site	225	628	56	8.92
Modification site	92	187	44	23.53
Exon boundary	67	296	42	14 19

Correlation with intron positions

The correlation between conserved sequence patterns and exon structures has also been examined. A popular view suggests that introns existed in ancestral genes and have been removed under the exon shuffling mechanism (Holland and Blake, 1987) where an exon forms a structural or functional unit of a protein. Therefore, it was expected that the identified motif blocks may correspond to exon units. As shown in Table IV, however, quite a few introns were found to split functionally important motif blocks. Figure 5 shows typical examples where exon boundaries appear within the motif blocks. It is also noted that the intron positions around the motif block CGSCW of the papain (cysteine protease) superfamily (Ishidoh *et al.*, 1989) and around the motif block GDSGGP of the trypsin (serine protease) superfamily (Rogers, 1985) are not fixed within the respective member sequences. These observations appear to support the concept of intron insertions (Rogers, 1989), although all introns examined here may not fall into this category.

Discussion

Information about the functional properties of expressed protein products is often the main concern when DNA sequences are determined. The method presented in this paper is an attempt towards fully computerized interpretations of the sequence data.

Table V. Functional significance of single block sequence motifs

SF No.	PDB	Superfamily name	Pattern	Significance
14	2B5C	cytochrome b5	−HPGGEEVL	heme binding site
83		NADH dehydrogenase (ubiquinone) chain 2	+LS{L|M}GGLPP	
123	2SOD	superoxide dismutase (Cu−Zn)	−HFNP	metal binding site
155		chloramphenicol acetyltransferase	+HH{A|S}VCDG	chloramphenicol binding site
332	7RSA	pancreatic ribonuclease	−CKPV{N|B}TF{V|I}H	active site
333		α-amylase, subtilis type	+VDAVINH	
345		paramyxovirus hemagglutinin−neuraminidase	+NRKSCS	
346		influenza virus sialidase	−MNPNQK{I|L|T}	membrane attachment site
378	6CHA	trypsin	−{G|H}DSG{G|S}{P|V|S}	active site
381	9PAP	papain	+CG{S|G}CW	active site
576		potato proteinase inhibitor PTI	−CAGYKGCNY	
593		ras transforming protein	−DT{A|T}GQE	nucleotide binding site
614		corticoliberin−endorpholiberin	−DLTFHLLR	
630		somatostatin	−PRERKAGCKNF	active site
637		pituitary glycoprotein hormone β-chain	−CAGYC	hormone deficiency by mutation
654		gastrin	−GWMDF	amidation site
659		egg-laying hormone	−PRLRFY	active site
676	1MLT	melittin major	−LISWI	
711		interferon-α	−CAWE	active site
745		nonhistone chromosomal protein HMG-17	+RRSARLSA{K|R}P	
796	1CTF	ribosomal protein L12	+LGLKEAK	methylation site
839		collagen α-chain	−GPPG{P|A}P	
938	2MT2	metallothionein	−CSCCP	metal binding site
1078		photosystem II D2 protein	+VFLIYP	
1200		transposase repressor	−KLDRLGR	
1232		nif-specific regulatory protein	+ESELFG{H|V}	
1323		polyomavirus coat protein VP2	+V{M|L}LPL	
1351		parvovirus noncapsid protein	+{N|Q}FPFND	
1578		adenovirus fiber protein	−NPVYPY	
1583		adenovirus hexon-associated protein (IX)	+WAG{V|A}RQ	
1703		mammalian retrovirus gag polyprotein I	−GHPDQV{P|I}YI	
1722		AIDS sor protein	+SLQ{Y|F}LAL	
1724		AIDS nef protein	−KEKGGL	
1737		coronavirus E1 membrane glycoprotein	+SWW{S|A}FNPE	
1762		parainfluenza virus nonstructural protein	+V{I|M}MEEAW	

Fig. 4. Stereo drawing of phospholipase A2 where two motif blocks are shown in thick segments.

A collection of sequence motifs with associated biological meanings in evolutionary, functional and structural aspects may be considered a dictionary for such purposes. At the same time, the motif search approach is expected to solve the speed and sensitivity problems in the current homology search approaches. Because motifs represent more organized information, concentrated and extracted from primary databases, the search against a motif library is much faster than the search against a sequence database. It is also possible to incorporate various types of motif in the library, not only those to identify membership of a superfamily, but also other sequence patterns which are too weak to be detected by standard database search methods.

Until now, sequence motifs have been found by manually examining a set of related sequences, although there have been a few attempts to automate the procedure (Staden, 1989; Smith and Smith, 1990; Smith *et al.*, 1990). The essence of our automatic method is the concept of uniqueness. For a protein with 100 residues there are 20^{100} possible amino acid sequences. In nature, however, the repertoire of real amino acid sequences appears to be quite limited in comparison to this theoretical number. The protein sequences sequenced to date amount to 10 million residues, three times larger than 20^5 or 3.2 million pentapeptide patterns. In reality, ~40% of the possible pentapeptide patterns are not used in the known sequences. Thus, actual proteins seem to have evolved from a limited set of amino acid sequences, conserving functionally important residues. This has been the working hypothesis in this study. As expected, motif blocks, constructed from unique peptide words, were found to be well correlated with functionally important sites of protein molecules. In addition, separate blocks tend to be close together in space to form an active site.

For the motif library to be more useful, it is necessary to increase the number of identified superfamilies, i.e. to reduce the number of no opinions (70–80%) in Table III. One approach is to use lower levels of conservation, *f*, as shown in Table I.

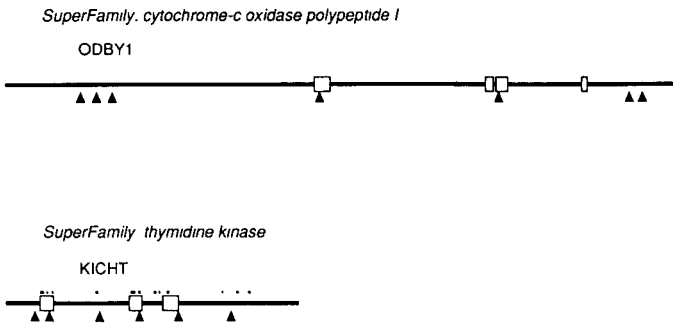

Fig. 5. Motif blocks (□) and intron positions (▲) in selected enzyme sequences. These examples demonstrate that conserved regions with functional importance are sometimes split by introns.

Another is to relax the condition of uniqueness which was strictly required in this analysis. A few exceptions can be allowed in other superfamilies and/or patterns could be identified that are unique to multiple superfamilies. In our preliminary analyses of the latter case, the pattern YGDTDS was found in two superfamilies (DNA-directed DNA polymerases of adenovirus and herpes virus) which share very little sequence homology. The possibility of combining multiple superfamilies based on short sequence motifs is thus inferred. The pattern HPDKGG was found exclusively in the three superfamilies: large T antigen, middle T antigen and small t antigen of polyoma and related viruses. However, this pattern was actually located in the exon shared by the three antigens.

Acknowledgements

This work was supported by a grant-in-aid for scientific research on the priority area 'Genome Informatics' from the Ministry of Education, Science and Culture, Japan

References

Bairoch,A. (1989) *Prosite: a dictionary of protein sites and patterns.* EMBL, release 4.

Bernstein,F.C., Koetzle,T.F., Williams,G.J.B., Meyer,E.F.,Jr, Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. and Tasumi,M (1977) *J. Mol. Biol.*, **112**, 535–542.

Dayhoff,M.O. (1978) *Atlas of Protein Sequence and Structure*, vol. 5, suppl. 3. National Biomedical Research Foundation, Washington, DC.

Dayhoff,M.O., Barker,W.C. and Hunt,L.T. (1983) *Methods Enzymol.*, **91**, 524–545

Hodgman,T.C. (1989) *CABIOS*, **5**, 1–13.

Holland,S.K. and Blake,C.C.F. (1987) *BioSystems*, **20**, 181–206.

Ishidoh,K., Kominami,E., Katunuma,N. and Suzuki,K. (1989) *FEBS Lett.*, **253**, 103–107.

Kanehisa,M.I. and Goad,W.B. (1982) *Nucleic Acids Res.*, **10**, 265–278.

Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl Acad. Sci. USA*, **85**, 2444–2448.

Rogers,J. (1985) *Nature*, **315**, 458–459.

Rogers,J.H. (1989) *Trends in Genet.*, **5**, 213–216.

Seto,Y., Ikeuchi,Y. and Kanehisa,M. (1990) *Proteins*, **8**, 341–351.

Smith,H.O., Annau,T.M. and Chandrasegaran,S.C. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 826–830.

Smith,R.F. and Smith,T.F. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 118–122.

Staden,R. (1989) *CABIOS*, **5**, 293–298.

Taylor,W.R. (1988) *Protein Engng*, **2**, 77–86.

Waterman,M.S. and Smith,T.F. (1978) *Math. Biosci.*, **42**, 257–266.

Received on February 29, 1992; revised on May 4, 1992; accepted on July 6, 1992

Appendix

Dictionary of sequence motifs characterizing superfamilies

```
< Superfamily number > Superfamily name
       Motif block - Spacer - Motif block - ...
    Notations used:
       + Conserved block;  - Nonconserved block;
       [ ]  Minimum and maximum length of a spacer;
       { }  Substitution patterns or a deletion by minus sign;
       ( )  Possible insertion pattern

< 14 > cytochrome b5
       -HPGGEEVL
< 31 > L-lactate dehydrogenase
       -PVD{I|V}L=[47,47]=-G{E|Q}HGD
< 50 > glyceraldehyde-3-phosphate dehydrogenase
       +GFGR{I|-}GR=[129,134]=+SNASCTTN{C|S}LAP=[14,14]=+{L|M}MTTVH=[30,31]=
       +TGAA{K|R}A{V|T}=[92,95]=+{S|A}WYDNE
< 60 > acyl-CoA oxidase
       -TVGDIG=[21,21]=-RFFM=[153,159]=+ACGGHG
< 68 > glutamate dehydrogenase (NAD(P)+)
       +AEG{A|S}N=[24,31]=+N{A|C}GGV
< 83 > NADH dehydrogenase (ubiquinone) chain 2
       +LS{L|M}GGLPP
< 96 > cytochrome-c oxidase polypeptide I
       -{Q|E}HLFWFFGHPEVYI=[126,127]=-VV{A|G}HFHYVLS
< 97 > cytochrome-c oxidase polypeptide II
       +G{H|F|Y}QWYW=[83,91]=-YGQCSE{I|L}
< 98 > cytochrome-c oxidase polypeptide III
       -SPWPL=[111,113]=+PLLNT=[105,106]=-YWHFVDV
< 123 > superoxide dismutase (Cu-Zn)
       -HFNP
< 130 > herpesvirus ribonucleotide reductase large chain
       +EPWH=[120,127]=-SNLCTEI=[296,311]=+GLKT{G|I}MYY
```

```
< 155 > chloramphenicol acetyltransferase
      +HH{A|S}VCDG
< 203 > thymidine kinase
      +GPMF{S|A}GK{S|T}{T|S}EL=[56,62]=+{V|I}IGIDE{G|A}QFF
< 229 > phosphoglycerate kinase
      -RVDFNVP=[32,33]=-SHLGRP=[94,99]=-DAFGTAHR=[218,219]=-SHVSTGGGASLELLEGK
< 247 > herpesvirus DNA-directed DNA polymerase
      -{F|L}DIEC=[284,412]=-DFASLYPS=[92,96]=-YGFTGV=[58,83]=+{I|V}YGDTD
< 269 > influenza virus RNA-directed RNA polymerase 3
      -HLEVC=[520,529]=-MKWGME=[1,1]=-RRCLLQS=[65,69]=-QLEGFSAESR
< 281 > phospholipase A2
      -YGC{Y|N}C{G|S}=[11,11]=-DRCC
< 332 > pancreatic ribonuclease
      -CKPV{N|B}TF{V|I}H
< 333 > alpha-amylase, subtilis type
      +VDAVINH
< 345 > paramyxovirus hemagglutinin-neuraminidase
      +NRKSCS
< 346 > influenza virus sialidase
      -MNPNQK{I|L|T}
< 378 > trypsin
      -{G|H}DSG{G|S}{P|V|S}
< 381 > papain
      +CG{S|G}CW
< 382 > calpain
      -DLKTDGF=[4,4]=+CR{S|N}MV=[76,76]=+DFDNF=[2,2]=+CLV{R|K}L=[24,24]=+WL{Q|L}LTM
< 385 > neutral proteinase
      +{A|P}AVDAHY=[46,46]=+YGDGDG=[96,100]=+GGVH{T|I}NSGI=[1,1]=+NK{A|Q}AY
< 444 > ribulose-bisphosphate carboxylase small chain
      -MQVWP=[54,55]=-YYDGRYWTMWKLPMFG=[31,31]=-R{Q|E}VQC
< 472 > carbonate dehydratase
      -QFHFHWGS=[4,4]=-GSEHTV
< 499 > triose-phosphate isomerase
      +GHSERR=[63,65]=+AYEP{V|L}WAIGTG{K|L}=[32,33]=+YGGS{V|A}=[18,18]=+VGGASL{K|E}
< 503 > glucose-6-phosphate isomerase
      +IGIGGS=[108,119]=+{V|I}GGRYS
< 576 > potato proteinase inhibitor PTI
      -CAGYKGCNY
< 593 > ras transforming protein
      -DT{A|T}GQE
< 596 > bcl transforming protein
      +AGRTGYDNREIVMKYIHYKLSQRGYEWDAGD=[32,32]=+ARTSPL=[6,9]=+AGPALSPVPP=[8,8]=+AGD
      DFSRRYR=[1,1]=+DFAEMS{S|R}QLHLTPFTARGRFATVVEELFRDGVNWGRIVAFFEFGGVMCVESVNREMSP
      LVDNIALWMTEYLNRHLHTWIQDNGGW
< 614 > corticoliberin-endorpholiberin
      -DLTFHLLR
< 630 > somatostatin
      -PRERKAGCKNF
< 631 > vasopressin-neurophysin
      +CLPCGPGGKGRCFGP{S|N}ICC=[1,1]=+{D|E}ELGCF=[1,1]=+GTAEALRCQEE{N|I}YLPSPCQSGQK
      {P|A}CGS{G|E}{G|A}{R|A}{C|A}{A|L}
< 632 > corticotropin-lipotropin
      -M{E|G}HFRWG=[59,81]=-HFRW{G|S}
< 635 > proenkephalin
```

```
      +{Q|Y}KRYGG=[13,14]=+QKRYGGF{L|M}
< 636 > thyrotropin-gonadotropin alpha chain
      -GCCFSRAYPTP=[4,4]=-KTMLV=[8,8]=-TCCVAK=[20,20]=-CSTCY
< 637 > pituitary glycoprotein hormone beta chain
      -CAGYC
< 654 > gastrin
      -GWMDF
< 659 > egg-laying hormone
      -PRLRFY
< 676 > melittin major
      -LISWI
< 711 > interferon alpha
      -CAWE
< 739 > ubiquitin
      +MQIFVKTLTGKTITLEVE=[6,6]=-NVK{A|S}KIQDKEGIPPDQQRLIFAGKQLEDGRTL=[1,1]=
      -DYNIQKESTLHLVLRLR
< 742 > histone H2B
      -KQ{V|T}HPDTG=[8,8]=+{M|L}NSFVND=[1,1]=+F{E|G}RIA=[20,20]=-QTAVRL=[1,1]=
      -LPGELAKHAV{S|T}EGT{K|R}{A|S}VTKY{T|S}
< 743 > histone H3
      -A{R|H}TKQTA{R|C}KST{G|C}{G|R}KAPRKQL{A|V}{T|S}KAA=[6,6]=-{T|S}GGVKK{P|S}HR=
      [1,1]=-{R|K}P{G|D}TVAL{R|H|L}EI{R|H}{R|K}{Y|F}{Q|H}{K|-}STELLI{R|C|H}K{L|A}
      PFQRLV{R|Q}EIAQDFKT{D|E}LRFQ{S|R}{S|A}A=[8,8]=-EA{Y|N}{L|R}V{G|R|A|S}LFEDTN
      LCAIH{A|G}K{R|C}V{T|S}{I|V}{M|Q|I}PKD{I|M}Q=[2,2]=-RRIRGERA
< 744 > histone H4
      +KGGKG{L|M}GK{G|V}GAKRH(s){R|S}K=[1,2]=+{L|N}{R|K}{D|A}{N|S}I{Q|E}GITKPAIRRLA
      RRGGVKRIS=[19,19]=-RD{A|S}VT{Y|-}=[11,11]=-{M|L}DVVY{A|S}LKRQGRT{L|I}YGFGG
< 745 > nonhistone chromosomal protein HMG-17
      +RRSARLSA{K|R}P
< 775 > ribosomal protein S12
      +{T|N}PKKPNSALRK{V|I}=[1,1]=+RVRL{T|S}=[5,5]=-TAYIPGIGHNLQEHSVVL=[1,1]=
      +RGGRVKDLPGV{R|K}Y=[18,18]=+RSKYGVK{K|R}
< 792 > ribosomal protein L2
      +GGGHK{R|Q}=[19,19]=-TIEYDPNR=[137,137]=+MNP{V|I}DHP{H|G}GG{G|H}EG=[1,1]=
      -APIGRK
< 796 > ribosomal protein L12
      +LGLKEAK
< 807 > ribosomal protein L14
      +{V|C}ADNSGAR=[97,99]=-SLAPEV
< 810 > ribosomal protein L16
      -FPDKP=[8,8]=+RMG{S|K}GKG=[2,2]=-E{Y|K}WVAVVK
< 838 > feather keratin
      -CGPTPLANSCNEPC{V|L}(f)RQC={11,11}=+VVVT{L|F}PGPILSSFPQ
< 839 > collagen alpha chain
      -GPPG{P|A}P
< 844 > alpha crystallin
      -R{L|I}FDQ=[73,107]=-SREFHR
< 854 > tubulin
      -GGGTGSG=[32,32]=-V{V|I}EPYN=[119,119]=+DPR{H|N}G
< 855 > tropomyosin
      -QAEADKK=[116,116]=+QLKEAKHI
< 872 > kappa casein
      -RSPAQ=[34,34]=-AIPPKK
< 905 > amyloid protein
```

```
      +M{R|K}EAN=[4,4]=+DKYFHARGNYDAA
< 935 > ferritin
      +ASY{T|V}YLS=[3,3]=+{Y|F}FDRDD{V|I}AL=[42,42]=-DEWG=[16,16]=-NQALLDLH=[7,7]=
      +DPHLCDF{L|I}E
< 938 > metallothionein
      -CSCCP
< 941 > proline-rich protein
      -QGPP(s){P|Q}PG{K|N}PQGPPPQGG=[11,13]=-QGPPP{P|Q}(g)G(n){K|R|G}PQGPP{P|Q}PG
      {K|N}PQGPP{P|A}QG{G|D}
< 949 > heat shock protein 70
      -IDLGTTYSCVGV=[49,50]=+K{R|P}LIGR=[11,11]=-KHWPF
< 953 > homeotic protein
      +H{F|T}N{R|H}YL{T|M}R{R|P}R=[8,8]=+{C|N}LTERQI{K|E}
< 972 > zein
      -Q{Q|H}QFLPFNQL=[6,6]=+{A|T|S}YLQ{Q|A}Q=[14,24]=-LPFNQL=[31,32]=-{Q|P}QLLP
      {F|Y}=[13,13]=-{L|Q}QQLLPF
< 993 > circumsporozoite protein
      +{N|K}KL{K|N}QP=[174,259]=-{P|Q}CSVTCG=[39,40]=-VSNSLG
< 1060 > methyl-accepting chemotaxis protein
      +DLSSRTE{Q|E}QA=[5,5]=+TAASMEQLTATV{K|G}QNA{D|E}NA{R|H}{Q|H}AS=[35,35]=
      +{D|E}I{I|T}{S|A}VI{D|N}{G|S}IAFQTNILALNAAVEAAR{A|-}GEQGRGFAVVA{G|S}EVR{N|T}
      LA{S|Q}RSAQAA{K|R}EI=[36,36]=+{D|H}IM{G|Q}EIA{S|A}AS{D|E}EQ{S|Q}RGI=[12,12]=
      +VTQQNA{S|A}LV
< 1078 > photosystem II D2 protein
      +VFLIYP
< 1122 > fimbrial protein
      +VLSVQGASAPV{K|E}KKSFFSKFTRLNMLRL{A|V}RAVIP{A|V}AVLMM{F|L}FP{Q|E}LAMAA=
      [1,3]=+G{Q|-}DLMA{S|K}GN{T|D}TVKATFGKDSS{V|I}VKWVVLAEVLVGAVMYMMTKNVKFL
      {A|V}GFAIISVFIAV{G|V}M{A|S}VVGL
< 1200 > transposase repressor
      -KLDRLGR
< 1232 > nif-specific regulatory protein
      +ESELFG{H|V}
< 1317 > large T antigen
      -ICQQA=[43,48]=-AGVAW=[78,78]=-FEDVKG=[28,28]=+{K|P}VNLE
< 1323 > polyomavirus coat protein VP2
      +W{M|L}LPL
< 1324 > papillomavirus L1 protein
      +RLVW{A|G|C}=[7,7]=+RGQPLG
< 1328 > papillomavirus E1 protein
      -MVQWA{Y|F}D=[139,139]=+S{H|Q}FWL
< 1351 > parvovirus noncapsid protein
      +{N|Q}FPFND
< 1362 > hepatitis B virus gene X protein
      +MAARL{C|Y}C=[120,124]=+FVLGGCRHK
< 1477 > herpesvirus glycoprotein B
      -CYSRP=[12,16]=+{E|Y}GQLG=[85,87]=-QRRNQ=[62,62]=+NPFG{A|G}=[27,27]=-PMKALYP
< 1578 > adenovirus fiber protein
      -NPVYPY
< 1583 > adenovirus hexon-associated protein (IX)
      +WAG{V|A}RQ
< 1688 > insect iridescent virus repetitive protein
      +IGSSST=[12,12]=+LQISG{T|R}
< 1692 > yellow fever virus genome polyprotein
```

```
      -DRGWGN{G|H}CGLFGKG=[304,312]=-{T|H}AWDF
< 1697 > togavirus structural polyprotein
      +VGRE{K|L}=[206,209]=+PHG{W|L}PH=[42,42]=+C{L|I}TPY
< 1703 > mammalian retrovirus gag polyprotein I
      -GHPDQV{P|I}YI
< 1714 > type E retrovirus env polyprotein
      -KPCVKL=[420,466]=-GI{V|L}QQQ=[95,99]=-Q{Q|E}EKN
< 1722 > AIDS sor protein
      +SLQ{Y|F}LAL
< 1724 > AIDS nef protein
      -KEKGGL
< 1731 > AIDS orf-X protein
      -NSGEETIGEAF=[10,10]=-NREAVNHLPRELIFQVWQRSW=[1,1]=-YWHD=[9,9]=-
      KYRYLC=[32,32]=-PPPPGL
< 1737 > coronavirus E1 membrane glycoprotein
      +SWW{S|A}FNPE
< 1746 > paramyxovirus nucleocapsid protein
      +AGLASF=[62,62]=+{L|I}WSYAMGV
< 1748 > parainfluenza virus polymerase-associated nucleocapsid phosphoprotein
      +LGVIQS=[121,121]=+{T|V}RFDP
< 1749 > simian paramyxovirus P protein
      -TVKIMDPG=[20,20]=-VSGPGD=[13,13]=-DELARP=[61,61]=-IIRSAI
< 1762 > parainfluenza virus nonstructural protein
      +V{I|M}MEEAW
< 1776 > influenza virus nonstructural protein NS2
      -RMSKMQL=[5,5]=-DLNGMITQ=[36,36]=-QKFEEIRWLI=[9,9]=-TENSFEQITF=[12,12]=
      -EIRTFSFQLI
< 1863 > tobacco mosaic virus coat protein
      -FDTRNRII=[18,18]=-RVDDATVAIR
```