studia
biophysica Volume 129 (1989) Number 2-3, p. 231 - 240

A METHOD FOR PROPERTY PATTERN SEARCHES IN PROTEIN SEQUENCE DATA BASES, DEMONSTRATED BY DETECTION OF GTP-BINDING SITES

P. BORK, C. GRUNWALD
Central Institute for Molecular Biology, Academy of Sciences of the GDR, Division of Biomathematics, Robert-Roessle-Str. 10, Berlin 1115, GDR

Paper read at the Conference APPLICATION OF DATA BASES AND MICRO-COMPUTERS IN MOLECULAR BIOLOGY, Oct. 31 - Nov. 5, 1988, Jena, GDR

ABSTRACT

A database search for sequence sections that match a given pattern has been developed. In the pattern, not only the kind of the amino acid at a specific position may be specified but, by choice, also physicochemical and steric properties and positions of possible deletions. This allows to detect sequence sections that are not conserved in sequence but are so in structure and function. The property patterns can be automatically derived as consensus patterns from alignments of related sequence sections and reveal structurally and functionally important features of the examined sequences. Studies of GTP-binding sites resulted in well-defined consensus patterns.

INTRODUCTION

Until now, tertiary structures of only about 300 molecules (Brookhaven data bank, Cambridge, Release 44, BERNSTEIN /3/) have been determined. On the other hand, some thousand protein sequences are known (SWISSPROT Protein Sequence Data Base, Release 5: 5207 entries). Therefore, often only sequence data are available for investigations on a special protein. Since spatial structures of amino acid chains are determined by their sequences (CREIHGTON /4/), various methods have been developed to exploit sequence information for predictions of structure and function. Beside secondary structure prediction and alignment algorithms, the recognition and analysis of consensus patterns of related proteins/protein domains is such a method (for review see TAYLOR /8/). In the most simple case, a consensus pattern contains the information as to which amino acids are common in the examined protein sections (then to be called consensus sequence). It can be used to search for proteins with sequence patterns that match the consensus pattern. Ideally, these proteins have structures and functions similar to those of the sequences that served as the basis for the construction of the consensus motif.

The relationship existing between amino acid sequence and steric structure is not always evident: On the one hand, tertiary structures of short, identical sequence pieces in different proteins may be very dissimilar (KABSCH & SANDERS /6/, ARGOS /1/), reflecting interactions with other parts of the molecule and the surrounding solvent. On the other hand, very different sequences can form similar folds. This is one reason for the impossibility to determine well-defined consensus sequences, for example, of dinucleotide-binding sites (ARGOS & LEBERMAN /2/). Similar steric structures of dissimilar sequences may occur since different amino acids can play a similar role in the structure due to common steric and physicochemical properties (charges, hydrophobicity, extension of side chains). Therefore, consensus patterns should include information on these properties besides or instead of the specification of the kind of amino acid present at specific positions.

We have developed a method for deriving patterns of such properties (i.e. consensus patterns) from alignments of related sequences and for the subsequent database search for sequence sections that match these patterns. The characterisation of the residues of the patterns is based on 10 physicochemical and steric properties (see FIG. 1) given in ZVELEBIL et al. /9/. TAYLOR /7/ also used these properties to represent the results of alignments of related proteins ("search templates").

As an example, the results of searching for GTP-binding sites of proteins in the SWISSPROT-Database are presented here.

METHOD

To search for property patterns over the data base and to refine results, a FORTRAN-77 program set PAT consisting of the programs PCONSTR, PSEARCH and PEDIT was developed.

Input

To run the search program, a proper pattern had to be constructed. In FIG. 1 possibilities for characterising the individual amino acid residues (up to 35) of this pattern are demonstrated.

mismatches: 2

	aa	sp	co	hy	po	ne	po	ch	sm	ti	al	ar	pr
1	C		7	.	2	2	.	2	.	.	.	2	2
2	C		6	.	2	2	.	2	.	.	.	.	2
3	V		8	1	2	2	2	2	.	2	1	2	2
4	X		.	.	.	.	.	.	.	.	.	.	.
5	G	!	.	.	.	.	.	.	.	.	.	.	.
6	X		.	.	.	.	.	.	.	.	.	.	.
7	G	*	7	.	2	2	.	2	1	1	2	2	2
8	G	!	.	.	.	.	.	.	.	.	.	.	.
9	N		5	.	.	2	.	.	.	.	.	.	.
10	V		7	1	2	2	.	2	.	2	.	2	2
11	G		9	1	2	2	2	2	1	1	2	2	2

FIG. 1 Input pattern - *example:* Property pattern of the NAD-binding site of some dehydrogenases

mismatches - number of allowed errors in matching specified properties
aa - amino acids (one letter code, "X" = variable)
sp - special characteristics ("!" = substitutions/mismatches forbidden, "*" = deletions allowed)
co - degree *of* conservation (see text)
hy, po ... - hydrophobic, polar, negative, positive, charged, small, tiny, aliphatic, aromatic, prolin ("1" = demanded, "2" = forbidden)

There are the following options for describing the properties of the residues: Of course (i) the type of amino acid may be specified (FIG. 1, second column); additionally one can specify (ii) positions where deletions of residues are allowed and positions where any substitution is forbidden as marked by an asterisk "*" and exclamation mark "!", respectively; (iii) the degree of conservation in case of a substitution *of* the specified amino acid residue according ZVELEBIL et al. /9/ by an integer I with I = 10 for demanding identity and I = 9 - n with n being the number of nonequivalent properties if a substitution of the residue is possible; (iv) any of the mentioned steric and physicochemical properties; they can be demanded or forbidden; and (v) the permitted number of mismatches of specified properties, exclusively residues which are labeled by an "!".

The construction of property patterns is possible a priori by inscription of properties into a given scheme or with the program PCONSTR on the basis of an alignment of known sequences.

Program PCONSTR

By means of the program PCONSTR it is possible to derive an optimal property pattern from an alignment of given sequence sections. "Optimal" means that the most rigorous pattern of properties is calculated; as many properties as possible will be specified. Loosenings are possible by editing the pattern or by allowing some mismatches.

Program PSEARCH

The PSEARCH-program carries out the search for proteins containing the property pattern over a protein sequence data bank (SWISSPROT, PIR, or own database in PIR-format).

Algorithm: All different variants of the motif, which result from admission of deletions by an "*", are calculated. Then all sequence entries of the data base are compared successively with the motif variants. If a part of a sequence is found that matches all specified properties with the exception of allowed mismatches, then the name of the protein, the detected section of the sequence, its position and the number and position of mismatches are listed.

The search is carried out either over the whole database or over a part of it according to a current list. A current list may be created during program run and contains the codes of the matched proteins. It can be used for further searches, for instance, with stronger patterns or if searches for proteins are carried out that contain more than one pattern (see our example).

Required CPU-times depend on the length of the sequence, the number of variants (hence the number of "*"), the number of "!" (increase speed because of aborting comparison if the labeled residues are not found at their positions) and the number of entries in the bank or current list. The search for a pattern of 21 amino acid residues (without allowing deletions) over the SWISSPROT-database (release 5, 5207 entries) required about seven minutes.

Program PEDIT

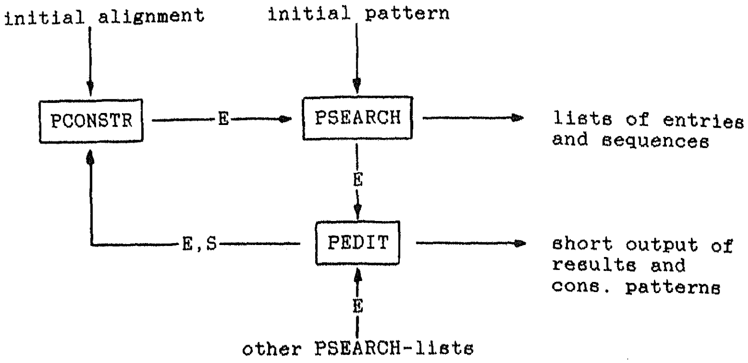
This program supplies alignments of the detected sequence sections, a summary of different searches and the derived consensus patterns.

Up to three output files of PSEARCH can be taken into account. With PEDIT results are compressed. A list is printed that contains sequence codes, an alignment of sequence sections together with their positions in the sequences, the number and position of mismatches and the calculated different property patterns (consensus patterns) that result from consideration of all sequence sections with no mismatches, all with none but one mismatch and so on up to the consideration of all findings. For an example, see FIG. 3. Additional lists of the aligned sequences permit editing (deletion or insertion of sequences) and calculation of a new property pattern with PCONSTR.

Use of the Programs

There are two main possibilities for using the program set PAT: (i) The search for sequences that contain a given property pattern and (ii) the determination of consensus patterns on the basis of an alignment of known functionally, structurally or evolutionary related sequence parts. Generally both methods will be used. For a scheme of the work with the programs see FIG. 2. After a first run of PSEARCH on the basis of aligned sequence sections or an a priori motif the results will enable one to determine a more appropriate pattern by consideration of newly found entries, eventually after experimental proof of their significance or comparison with other PSEARCH lists by means of

PEDIT. After editing, this pattern may be used as input for another search.



E ... editing of intermediate results possible
S ... experimental studies on function and structure of matched proteins, if necessary and possible

FIG. 2 Scheme of using programs

RESULTS AND DISCUSSION

For an example of the work with our programs we list here results of investigations on GTP-binding sites (detailed results of studies on nucleotide-binding sites will be published elsewhere). Three conserved sections are known to be characteristic for GTP-binding proteins: Section 1 is a nucleotide-ribose-binding site common to all nucleotide-binding proteins, section 2 is responsible for magnesium binding, section 3 for specific guanine-binding (JURNAK /5/). FIG 3. shows an output list of the program PEDIT (shortened). Listed are the results of searches for proteins that contain the three sequence sections. The words in the first column are the SWISSPROT-codes of the protein names. The next column contains the positions of the detected partial sequences in the whole sequence, followed by the sequence sections themselves and the number of mismatches. In the following column the positions of the second sequence sections are contained and so on. At the end of the list the resulting consensus patterns are listed.

CDGT$BACMA	577	IKAVIpKVAAGKTGVSVKtSS	2	74	KLYFgGDWQGiIDKIN	2	33	VdNKVNfS	2
	.		.	540	GNIVTIDGrGfGGTAG	2	.		.
DNAB$ECOLI	226	LIIVAARPSMGKTTFAMNLVE	0	142	IAEAgFDpQGRTSEDL	2	2	AGNKpfNK	2
	.		.	308	RNIYIDDSSGlTpTEV	2	6	pfNKQQAE	2
	.		.	.		.	172	rANKDEGP	1
	.		.	.		.	213	dLNKKTAG	1
EF1A$ARTSA	8	NIVVIGHVDSGKSTTTGHLIY	0	84	YYVTIIDAPGHRDFIK	0	150	GVNKMDST	0

(The rest of the list was shortened for reasons of typing space)

EF2$MESAU	21	NMSVIAHVDHGKSTLTDSLVC	0	98	FLINLIDSPGHVDFSS	0	156	MMNKMDRA	0
EFG$ECOLI	12	NIGISAHIDAGKTTTTERILF	0	82	HRINIIDTPGHVDFTI	0	129	QANKYkVP	1
	.		.	494	IRQKVTDVEGKhAKQS	1	140	FVNKMDRM	0
	.		.	624	PEENTGDVIGDLSRRR	0	270	FKNKGvQA	1
EFTU$ECOLI	14	NVGTIGHVDHGKTTLTAAITT	0	75	RHYAHVDCPGHADYVK	0	134	FLNKCDMV	0
ETXC$STAAU	97	SKdnVGKVTGGKTCMYGGITK	2	77	YKDEVVDVYGsNYYVN	1	136	YeNKRNTI	1
EXO6$BPT4	31	KTLITGRNGGGKSTMLEAITF	0	506	VFDGsFDAEGiKGVAN	2	69	STNKKELL	0
GBAI$HUMAN	35	KLLLLGAGESGKSTIVKQMKI	0	117	QGVLpDDLSGvIRRLW	2	255	CnNKWfTD	2
	.		.	195	LHFKMFDVGGQRSERK	0	268	FLNKKD[illegible]	0
GBAS$HUMAN	42	RLLLLGAGESGKSTIVKQMRI	0	217	VNFHMFDVGGQRDERR	0	21	EANKKie[illegible]	2
	.		.	.		.	290	FLNKQDLL	0
GN41$BPT4	192	LNVLMAGVNVGKSlgLCSLAA	2	305	PTIIIVDYLGiCKSCR	1	165	YMNKArKV	1
IF2$BACST	246	VVTIMGHVDHGKTTLLDAIRH	0	291	KKITFLDTPGHEAFTT	0	349	AINKMDKP	0
NIFH$AZOVI	4	QCAIyGKGGIGKSTTTQNLVA	1	119	LdFVFYDVLGDVVCGG	1	199	LANKLgTQ	1
NRDA$ECOLI	658	STCTTpMSCCGKCrVTMVICN	2	446	PLNDVNDENGEIALCT	0	336	QINKLmYT	1
POLG$FMDV1	1212	VVCLrGKSGQGKSfLANVLAQ	2	1261	QTVVVMDDLGQNpdGK	2	831	AYNKApfT	2
POLN$SINDV	721	TIGVIGTPGSGKSAIIKStVT	1	907	VKQLqIDYPGHEVMTA	1	109	ITNKNlhE	2
PPCK$CHICK	232	GSGYGGNSLLGKKCFALRIAS	0	312	IAWMKFDELGNLRAIN	0	51	EeNKKiLD	2
	.		.	.		.	386	WKNKDWTP	0
RAS$DICDI	5	KLVIVGGGGVGKSALTIQLIQ	0	51	CLLDILDTAGQEEYSA	0	114	VGNKADLD	0
RAS1$DROME	5	KLVVVGPGGVGKSALTIQLIQ	0	51	CLLDILDTAGQEEYSA	0	114	AGNKCDLA	0
RASN$HUMAN	5	KLVVVGAGGVGKSALTIQLIQ	0	51	CLLDILDTAGQEEYSA	0	114	VGNKCDLP	0
RHO$APLCA	7	KLVIVGDGACGKTCLLIVFSK	0	53	VELALWDTAGQEDYDR	0	115	VGNKKDLR	0
SYTI$ECOLI	572	SSLMIsTAMKGKApYCQVLTH	2	30	ARWTdDDLYGiIRAAK	2	69	SVNKIlKD	1
TALA$BKPOV	423	YWLfkGPIDSGKTTLAAGLLD	2	9	eSMELMDLLGlERAAW	2	227	GVNKEYLL	0
TDC1$BOVIN	31	KLLLLGAGESGKSTIVKQMKI	0	140	SEYQLNDSAGyYLSDL	1	263	FLNKKDVF	0
	.		.	190	LNFRMFDVGGQRSERK	0			.
UVRD$ECOLI	24	NLLVLAGAGSGKTrVLVHrIA	2	336	RIkTWQDNGGaLAECA	2	62	FTNKAaAE	1
V58$BSMV	264	TGIISGvPGSGKSTIVRTLLK	1	407	dTVIITDYDGETDETe	2	186	LINKfQQF	1
VDNP$BPT7	307	VIMVTeGSGMGKSTFVRQqAL	2	107	QVaDYrDQNGNIVSQK	2	534	EYNKETGW	0

```
VE1$PAPV1     435 CLLIfGPPNTGKSmFCTSLLK 2   113 RQLFsQDDSGlELSLL 2     2 AdNKGTeN 2
VE59$LAMBD    190 IHAfIGRNGCGKTTILNGMIG 1    37 FYLTVFDEhGEKCdIG 2    17 EKNKAfLR 1
VNCA$ADEA2    329 TIwLfGPATTGKTNIAEAIAH 2   262 QIkAaLDNAGKIMSLT 2   143 GGNKVvdE 2
YBL4$EPBAR     67 AYVITGTAGAGKSTsVSCLhH 2   196 TNVIVVDEAGtLSVhI 2   260 QVNKIreC 2
YP2$YEAST      10 KLLLIGNSGVGKSCLLLRFSD 0    57 VKLQIWDTAGQERFRT 0   119 VGNKCDLK 0

                              No mismatch allowed

motif             NTGVAANNSMGKTTMVMQMTT         TYNTTVDSGGRLSYTT         MMNKQNMT
remark                      !!                        !  !                 !!
cons. degree      5667694545..677745855         555556.56.645555         56..5655
hydrophobic       ...1.1........11..1..         ....1...........         .1......
positive          .22222..2....222..2..         ..2..2.22.......         2....2..
negative          222222.2.2..2222.222.         2...22..2......2         .2....2.
polar             .....2......1.....2..         ..........1.....         .....1..
charged           .22222.......222..2..         .....2..2.......         ........
smal              .....1.......1....2..         ................         ........
tiny              ...2.1........2...2..         ..2.2.....2.....         .....2..
aliphatic         .....22.....22.......         ..........2.2...         ........
aromatic          222222.22...22.22..2.         .......22..22.22         .2..222.
prolin            222222..22..222222222         .22222.2..222222         22..222.

                              1 mismatch allowed

motif             NMGVTATNNMGKTTMVQQMTT         TQQTTTDTTGQTTYTN         QMNKQTQT
remark                      !!                        !  !                 !!
cons. degree      5566695555..677755755         555556.55.555555         55..5555
hydrophobic       ...1.1........11..1..         ................         ........
positive          ..22.2..2....222..2..         ..2..2.2........         ........
  ...
```

FIG. 3: Results of the search for GTP-binding proteins (output of program PEDIT). Given the codes of SWISSPROT-entries, an alignment of the matched sequence sections, their positions in the sequences and the numbers and positions of mismatches, marked with lower case letters. (In the shortened list those proteins are not shown that stem from other species only or that are closely related to the listed ones, for instance, various ras-proteins.)

Altogether, our studies resulted in a detailed determination of the three motifs characterising GTP-binding properties. Already pattern 1 discriminates the closely related GTP- from ATP-binding sites. This has (without additional restrictions to the detected sequence sections) not been possible on the basis of consensus sequences: ARGOS & LEBERMANN /2/.

For further studies, an expansion of the property set could be valuable (for instance, by a property of "being glycine" which is, like prolin, a somewhat "exotic" amino acid).

Three general features of using the results of pattern searches can be derived:

- Consensus patterns (i. e. the property patterns) can be calculated and used for further searches in an extended database.

- Due to the lack of non-common properties, the derived consensus patterns make clear important structural and functional features of the studied sequence sections.

- Since steric structure and function of a protein depend on the properties of its amino acids, the pattern searching permits detection of structural motifs, detection of relationships between distantly related proteins and prediction of their structures and functions (for instance, of hypothetical proteins). TAYLOR /7/ proposed the creation of a consensus sequence database that may be used to determine structure and function of newly sequenced proteins.

ACKNOWLEDGEMENTS

The authors wish to thank their colleagues from the Division of Biomathematics for helpful discussions. The work was supported by Prof. Pfeil and Dr. Schoepp.

REFERENCES

/1/ ARGOS, P.
J. Mol. Biol. 197 (1987) 331 - 348

/2/ ARGOS, P.; LEBERMAN, R.
Eur. J. Biochem. 152 (1985) 651 - 656

/3/ BERNSTEIN, F.C.; KOETZLE, T.F.; WILLIAMS, G.J.B.; MEYER, E.F.; BRICE, M.D.; RODGERS, J.R.; KENNARD, O.; SHIMANOUCHI, T.; TASUMI, M.
J. Mol. Biol. 112 (1977) 535 - 542

/4/ CREIHGTON, T.E.
Freeman, New York 1983

/5/ JURNAK, F.
Science 230 (1985) 32 - 36

/6/ KABSCH, W.; SANDER, C.
Proc. Natl. Acad. Sci. USA 81 (1984) 1075 - 1078

/7/ TAYLOR, W.R.
J. Mol. Biol. 188 (1986) 233 - 258

/8/ TAYLOR, W.R.
Prot. Engin. 2 (1988) 77 - 86

/9/ ZVELEBIL, M.J.; BARTON, G.J.; TAYLOR, W.R.; STERNBERG, M.J.E.
J. Mol. Biol. 195 (1987) 957

Dr. P. BORK, Division of Biomathematics, Central Institute of Molecular Biology, Academy of Sciences of the GDR, Robert-Rössle-Str. 10, DDR-1115 Berlin, GDR