Nucleic Acids Research, 1993, Vol. 21, No. 13 3111–3115

The SBASE protein domain library, release 2.0: a collection of annotated protein sequence segments

Sándor Pongor[1,2]*, Vesna Skerl[1,+], Miklós Cserzö[1,§], Zsolt Hátsági[2,3], György Simon[1] and Valeria Bevilacqua[1]

[1]International Centre for Genetic Engineering and Biotechnology, Area Science Park, 34012 Trieste, Italy, [2]ABC Institute for Biochemistry and Protein Research, 2100 Gödöallö, Hungary and [3]Department of Computer Sciences, The University of Chicago, Chicago, IL 60637, USA

ABSTRACT

SBASE 2.0 is the second release of SBASE, a collection of annotated protein domain sequences. SBASE entries represent various structural, functional, ligand-binding and topogenic segments of proteins [Pongor,S. et al. (1993) *Prot. Eng.*, in press]. This release contains 34,518 entries provided with standardized names and it is cross-referenced to the major protein and nucleic acid databanks as well as to the PROSITE catalog of protein sequence patterns [Bairoch,A. (1992) *Nucl. Acids Res.*, 20 suppl, 2013–2018]. SBASE can be used for establishing domain homologies using different database-search tools such as FASTA [Lipman and Pearson (1985) *Science*, 227, 1436–1441], FASTDB [Brutlag et al. (1990) *Comp. Appl. Biosci.*, 6, 237–245] or BLAST3 [Altschul and Lipman (1990) *Proc. Natl. Acad. Sci. USA*, 87, 5509–5513] which is especially useful in the case of loosely defined domain types for which efficient consensus patterns can not be established. SBASE 2.0 and a set of search and retrieval tools are freely available on request to the authors or by anonymous 'ftp' file transfer from <ftp.icgeb.trieste.it>.

INTRODUCTION

The most widely used approach to predict the biological function of a newly determined protein sequence is a simple database search. Detection of distant homologies such as may exist between multidomain proteins can pose problems, however, since biologically significant alignment patterns often constitute less than 25% of a domain sequence, the alignments may not always be significant in the mathematical sense (1,2). SBASE is a collection of protein domain sequences designed to facilitate the detection of such distant homologies.

The current release 2.0 of SBASE contains over 34 thousand annotated protein sequence segments consistently named by structure, function, biased composition, binding-specificity and/or similarity to other proteins. SBASE can be considered as a conversion of the protein sequence database into a format that facilitates detection of functional and structural similarities rather than sequence homologies. Searching this database with standard programs such as FASTA (3), FASTDB (4) or BLAST3 (5) yields information on the potential functions of the detected homology regions which may allow the detection of potential domain homologies and prediction of function. In addition, a set of search and retrieval tools is designed to facilitate the detection of domain homologies using SBASE 2.0.

DESCRIPTION OF THE DATA

SBASE 2.0 is a collection of over 34 thousand protein domain sequences with a total of almost 2 million amino acids, which is a considerable increase in comparison to release 1.0 (Table 1).

Definition of protein domains

Domains included in SBASE are protein sequence segments with known structure and/or function. The main classes of protein domains are defined as follows (Table 2): **Structural domains** are sequence segments with a known structure (like the protein modules (6) such as epidermal-growth-factor-like [EGF-like] and immunoglobulin-like [IG-like] domains), biased composition (e.g. serine/threonine-rich domains) as well as various sequence repeats. **Homology domains** are regions of homology to other proteins detected by the original authors. These homology regions are less well characterized then the 'established' structural domains but can be eventually used to define further domain types. **Ligand-binding domains** are sequence segments known to bind specific ligands (such as DNA, metals, sugars etc.). **Cellular location domains** are sequence segments known to be involved in targeting (signal peptides, nuclear-localization signals, chloroplast transit-peptides), as well as domains of transmembrane proteins (cytoplasmic, transmembrane and extracellular domains).

The number of PROSITE cross-references illustrates the difference between the domain definitions used in the PROSITE catalog and SBASE, respectively. For example, SBASE 2.0 contains 286 epidermal growth factor like domains or EGF-

* To whom correspondence should be addressed at: International Centre for Genetic Engineering and Biotechnology, Area Science Park, 34012 Trieste, Italy

Permanent addresses: +Institute for Nuclear Sciences—Vinca, PO Box 522, Belgrade, Yugoslavia and §Biological Research Center, Hungarian Academy of Sciences, 1518 Budapest 7, Hungary

REPEATs (Table 2). Of these, 194 contain the consensus pattern 'EGF' given in PROSITE. They also contain cross references to other PROSITE entries, such as aspartate/asparagine hydroxilation sites, so the total number of PROSITE cross references in this group is 229. The difference between the information content of PROSITE and SBASE is thus twofold: a) SBASE contains a number of 'atypical' domain sequences that do not contain the corresponding PROSITE pattern; b) There are other loosely defined domain types, for which no consensus pattern is available yet (compare columns 2 and 4 in Table 2). In both cases, homology search against SBASE can give an indication of similarity, even though the significance of a similarity is not *a priori* tested, as in the case of PROSITE.

Table 1. Increase of data in SBASE release 2.0

RELEASE	DATE	RECORDS	AMINO ACIDS	SIZE [Mb]
1.0	2-APR-92	27 221	1 551 445	17,19
2.0	13-FEB-93	34 518 (+ 27%)	1 922 524 (+ 24%)	24,94 (+ 45%)

Source and origin of data

SBASE data originate from three different sources: *i)* from the SWISS-PROT protein sequence databank (7); *ii)* from the Protein Sequence Database of the Protein Identification Resource (PIR) (8); and *iii)* from the literature. The sequences are either translated from nucleotide sequence databases (9,10) or directly keyed in at the protein level. From a total of 34518 records in SBASE 2.0, 27212 (79%), 4892 (14%) and 2414 (7%) are of eukaryotic, prokaryotic and viral origin, respectively.

Redundancy of sequences in SBASE 2.0 is kept at a minimal level. In some cases, the domain definitions overlap, so the same sequence can be present in different entries. For example, short ligand binding domains are often parts of larger extracellular domains in transmembrane proteins, and their sequence will thus be represented by two separate records.

Standard names

In SBASE 2.0 the names of (a growing number of) domains types are standardized so as to facilitate the retrieval of all sequences of the same domain type (Table 3). The standardized name of the domain is given in the SN line of each entry. In order to

Table 2. Examples of domains in SBASE 2.0

Domain class Domain type	Number of records in SBASE 2.0	Records referenced to PROSITE	Pointers to PROSITE records of correct domain type
STRUCTURAL DOMAINS			
Repeats (all)	4675	996	–
EGF-repeats	286	229	194
IG-like repeats	370	12	7
Fibronectin repeats	188	24	16
Kringle domain	102	98	98
Sushi repeat	220	20	–
Heptad-repeat	24	1	–
Domains with biased composition			
Gly-rich	173	–	–
Pro-rich	115	–	–
Ser-rich	150	–	–
Cys-rich	106	–	–
OPA-repeat	7	–	–
Acidic	224	–	–
Basic	136	–	–
Hydrophilic	96	–	–
Hydrophobic	149	–	–
HOMOLOGY DOMAINS	2580	1016	–
LIGAND-BINDING DOMAINS	3916	1771	–
Calcium-binding	700	87	87
Zinc-fingers	1380	946	915
Other DNA-binding	881	341	349
RNA-binding	126	78	77
Lectin domains	73	14	0
Homeobox	184	163	163
Helix-turn-helix (H-T-H)	332	5	4
Helix-loop-helix (H-L-H)	51	34	34
CELL TOPOLOGY DOMAINS			
Extracellular	1844	451	–
Transmembrane	6800	210	–
Cytoplasmic	1978	205	–
Signal peptides	4079	152	–
Transit to organelles	557	4	–
Nuclear localization signals	58	0	0

(–: not determined or PROSITE KEY not specified)

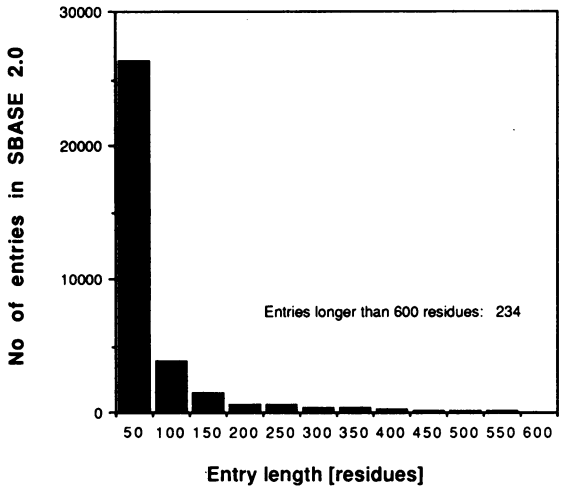

Figure 1. Length distribution of entries in SBASE 2.0 library. Domain lengths vary between 5 and 1000 amino acids.

facilitate the use of database search programs, this information is also inserted into the ID lines. (See Figure 2 for an example).

Domain sizes

Domain sequences collected in SBASE 2.0 range in length between 5 and 1000 amino acids with the mean length of 55.7 residues (Figure 1). The boundaries of the domains are either as previously defined in the original publications or determined by homology to domains with defined boundaries.

Cross-references

SBASE 2.0 has cross-references to several protein and nucleic acid databanks such as SWISS-PROT (7), PIR (8), EMBL (9), HIV (11), as well as OMIM (12), REBASE (13), PROSITE (14) and others (Table 4). In SBASE records the DR-lines contain the cross-reference data.

Record structure

SBASE 2.0 is a sequential database comprising 34518 domain entries separated by '^L' (form feed). SBASE records have a format similar to the ones used by SWISS-PROT and EMBL

Table 3. Examples of standard domain names in SBASE 2.0

SBASE standard name	Descriptive name (occurrence)
ANK-REPEAT	Ankyrin-like repeats (ankyrin, sex-determining protein FEM-1, Glp-1, Lin-12, Notch, Xotch, DNA-binding proteins)
ANNEXIN-REPEAT	Annexin-like domain (annexins, Ca-binding proteins which associate reversibly with membranes)
APPLE-REPEAT	Apple-domain (plasma serine proteases activated by factor XIIa, kallykrein, factor XI)
BNR-REPEAT	Bacterial neuramidinase homology domain, BNR-motif (neuraminidases, syalidases)
CUT-REPEAT	Cut protein-like domain (homeobox proteins)
EGF-REPEAT	Epidermal growth factor like domain (growth factors, membrane-bound proteins, extracellular matrix proteins, coagulation factors, complement system, cell adhesion proteins)
FN1-REPEAT	Fibronectin type I domain (fibronectins, coagulation factore XII, plasminogen activator)
FN2-REPEAT	Fibronectin type II domain (fibronectins, blood coagulation factor XII, cation-independent IGF-receptor, mannose receptor, 72 Kd type IV collagenase).
FN3-REPEAT	Fibronectin type III domain (fibronectins, transmembrane receptors)
HEPTAD-REPEAT	Heptad repeat regions, coiled-coil domains (filamentous proteins, actin-binding proteins, extracellular matrix)
HMG-BOX	HMG homology domain (high mobility group proteins, nonhiston chromosomal proteins, nucleolar transcription factors)
KRINGLE	Kringle domain (serine proteases, plasma proteins)
LDLR-REPEAT	Low density lipoprotein receptor (Cys-rich) repeat domain (low density lipoprotein receptor, complement C9, C8, perforin)
LEU-REPEAT	Leucine-rich repeat
LIM-REPEAT	LIM-domain (homeobox proteins, mammalian cys-rich protein CRIP, rhombotins, human cys-rich protein CRP)
LIN/NOTCH-REPEAT	lin/notch domain (lin-12, notch, glp-1)
GFR/TNFR-REPEAT	Cys-rich domains of neural growth factor receptor and tumor necrosis factor receptor (growth factor receptors, mammalian B-cell antigen CD40/Bp50/,.rat T-cell antigen OX40, mouse T-cell protein 4-1BB, vaccinia virus protein A53/SalF19R/)
OPA-REPEAT	Opa-domain (homeobox proteins, notch, zeste, mitosis initiation protein fs(1)ya)
SUSHI-REPEAT	C3b/C4b-binding domain, short consensus repeat Scr (apolipoprotein-H, complement system, blood coagulation system, transglutaminases, adhesion proteins)

(see a sample record on Figure 2). Each entry consists of: *i)* the annotation data sorted in several types of comment lines, *ii)* the name and *iii)* the the amino acid sequence. Every comment line follows these rules: it has a semicolon and a blank on positions 1–2, a two-character line-identifier on position 3–4, blanks on position 5–7 and data from position 8 to the end of line (position 70 or less). The type of data stored in a comment line is defined by the line-identifier (see Table 5). The line with the record name contains (only) the unique record identifier copied from the ID-line. The domain sequence is given in one-letter amino acid code, 70 characters per line, and it ends with a '1' sign. The complete database uses only upper case characters. Future releases of SBASE might appear in a slightly modified format.

Table 4. Cross-references to other databases in SBASE

DATABASE	No of pointers in SBASE 2.0
EMBL	51 555
PIR	43 855
SWISS-PROT	34 518
PROSITE	6 707
PDB	438
MIM	5 149
TFD	1 572
FLYBASE	1 354
ECOGENE	1 216
EC-2D-GEL	360
HIV	58
REBASE	14

```
; ID   ANP_HEMAM-39-163 C-TYPE LECTIN (LONG FORM) ANTIFREEZE PROTEIN PR
; AC   SB02151
; DT   13-FEB-1993
; SN   C-TYPE LECTIN (LONG FORM)
; DE   C-TYPE LECTIN (LONG FORM) ANTIFREEZE PROTEIN PRECURSOR (AFP).
; DP   ANTIFREEZE PROTEIN PRECURSOR (AFP).
; OS   HEMITRIPTERUS AMERICANUS (SEA RAVEN).
; OC   EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; PISCES; GNATHOSTOMATA;
; OC   OSTEICHTHYES; ACTINOPTERYGII; SCORPAENIFORMES.
; DR   SWISS-PROT; ANP_HEMAM; P05140; AA 39-163
; DR   EMBL; J02593; HAAFP.
; DR   EMBL; J05100; HAAFPG.
; DR   PIR; A24602; A24602.
; DR   PIR; A34313; A34313.
; DR   PROSITE; PS00013; PROKAR_LIPOPROTEIN.
; RA   NG N.F.L., TRINH K.-Y., HEW C.-L.;
; RL   J. BIOL. CHEM. 261:15690-15695(1986).
ANP_HEMAM-39-163
PNCPAGWQPLGDRCIYYETTAMTWALAETNCMKLGGHLASIHSQEEHSFIQTLNAGVVWIGGSACLQAGA
WTWSDGTPMNFRSWCSTKPDDVLAACCMQMTAAADQCWDDLPCPASHKSVCAMTF1
```

Figure 2. A sample entry from the SBASE 2.0 protein domain library.

3A Segment analysis of human adducin

Score	Segment	Entry	Domain name
16	(385-435)	HMP3$DROME-1-190	GLUTAMINE-RICH
16	(387-437)	DMD$CHICK-1-244	ACTIN-BINDING.
16	(389-439)	TRSR$HUMAN-89-760	EXTRACELLULAR
16	(391-441)	FCN3$DROME-371-520	INTRACELLULAR.
15	(393-443)	FNBA$STAAU-37-99	EXTRACELLULAR
16	(395-445)	DMD$HUMAN-1-240	ACTIN-BINDING.
16	(397-447)	MYSG$CHICK-1-848	GLOBULAR HEAD
16	(399-449)	DMD$HUMAN-1-240	ACTIN-BINDING.
16	(401-451)	CACT$CHICK-1-177	TWO REPEATS
17	(403-453)	DMD$MOUSE-1-240	ACTIN-BINDING.
18	(405-455)	DMD$MOUSE-1-240	ACTIN-BINDING.
17	(407-457)	DMD$MOUSE-1-240	ACTIN-BINDING.
17	(409-459)	DMD$HUMAN-1-240	ACTIN-BINDING.
18	(411-461)	DMD$HUMAN-1-240	ACTIN-BINDING.
16	(413-463)	DMD$MOUSE-1-240	ACTIN-BINDING.
16	(415-465)	DMD$HUMAN-1-240	ACTIN-BINDING.
16	(417-467)	DMD$MOUSE-1-240	ACTIN-BINDING.
16	(419-469)	DMD$HUMAN-1-240	ACTIN-BINDING.
16	(421-471)	DMD$MOUSE-1-240	ACTIN-BINDING.
17	(423-473)	DMD$MOUSE-1-240	ACTIN-BINDING.
15	(425-475)	METH$ECOLI-642-870	COBALAMIN BINDING
15	(427-477)	DMD$MOUSE-1-240	ACTIN-BINDING.
15	(429-479)	KAR3$YEAST-1-109	GLOBULAR.

3B Alignment of rat adducin region 380-
actin-binding domain of dystrophin

```
                          390             400       410
RAT ADDUCIN       YTYRHPFVQEKTKHK------SEVEIPATVTAFVFEEDGVPV
HUMAN DYSTROPHIN  DCYEREDVQKKTFTKWVNAQFSKFGKQHIENLFSDLQDGRRL
MOUSE DYSTROPHIN  DCYEREDVQKKTFTKWINAQFSKFGKQHIDNLFSDLQDGKRL
CHICK DYSTROPHIN  DDYEREDVQKKTFTKWINAQFAKCGRRCIEDLFNDFRDGRKL
                    *    ** **  *                 *    **

                               440           450          460
HUMAN ADDUCIN     -RQHAQKQQKEK--TR--WLNTPNTYLRV---NVADEVQRNM
HUMAN DYSTROPHIN  EGLTGQKLPKEKGSTRVHALNNVNKALRVLQNNNVDLV--NI
MOUSE DYSTROPHIN  EGLTGQKLPKEKGSTRVHALNNVNKALRVLQKNNVDLV--NI
CHICK DYSTROPHIN  ECLTGQKIAKEKGSTRVHALNNVNKALQILQRNNVDLV -NI
                          **  ***  **   **  *  *     *  * *
```

Figure 3. Comparison of the rat adducin sequence to the SBASE 1.0 domain library. **A.** List of best scoring domains (detail). **B.** Alignment of adducin region 380–470 with actin binding domains of dystrophin.

Table 5. Types of comment lines in SBASE 2.0 records

LINE IDENTIFIER	CONTENT OF THE COMMENT LINE
ID	Unique record IDentifier. If the SWISS-PROT name is available, it is followed by the starting and the ending positions of the domain e.g. A20_HUMAN-286-317). In release 2.0 we strarted to store, in the rest of the ID-line, a short domain description for the sake of easier interpretation of database-search data.
AC	ACcession number–serial record number in format 'SBddddd' ('d' = digit). Attention: the AC-numbers in release 2.0 do not correspond to the ones in the previous release.
DT	DaTe.
SN	Standard Name.
DP	Definition of the Parent protein.
DE	DEfinition of the domain (same as SN + DP in short).
OS	Source Organism name.
OC	Taxonomy line.
DR	Database Reference (cross-reference).
RA	Authors of the literature Reference.
RL	Literature Reference.

RETRIEVAL AND SEARCH TOOLS

Programs

SBASE was originally developed in a UNIX environment (SUN 4/390, Sun OS 4.1.1) containing the Intelligenetic's sequence analysis package (15). SBASE 2.0 is accompanied by utility programs DRP and CHOPPER.

DRP (16) is a menu-oriented keyword-based retrieval program which allows viewing and saving of SBASE entries using as a keyword any word that occurres in the annotations. DRP is a C shell script developed under Sun OS 4.1.1 and uses the WAISINDEX and WAISSEARCH programs originally developed as a part of the WAIS Wide Area Information Software Server (17).

CHOPPER (18) is a homology search program designed to detect the best homologies of a given length between the query and a sequence database (SBASE). CHOPPER is based on a window-sliding algorithm and performs individual searches using overlapping parts of the query. The results are presented a) as a list of best domain-homologies and b) an ordered list of local homologies along the sequence. In other words, the procedure transforms the amino acid sequence into a sequence of possible domain homologies that is quite simple to evaluate. This can be especially useful in the case of binding-domains of very loosely defined structure. **Figure 3** shows the partial output obtained with CHOPPER on the sequence of an actin-binding protein, adducin. Homology to actin-binding domains was predicted using CHOPPER in region 380–470 of this protein (18). This region is found homologous to the actin binding domains of alpha-actinin and desmosin (Figure 3B), even though it does not contain the known consensus pattern of the actin binding domains. The similarities detected in this manner may not be mathematically significant in all cases, however they may orientate the design of biological experiments and also may help to develop new consensus patterns using programs such as PIMA (19), PROTOMAT (20), MOTIF (21) or PATCO (22).

Distribution

SBASE 2.0 (April 1993) is distributed by Anonymous 'ftp' file transfer from <ftp.icgeb.trieste.it>. The complete database (with both annotations and sequences) requires 24.94 Mb while the reduced version in FASTA-format (with only sequences) 5.09 Mb of disk storage space. Individual entries are available through the gopher server (23) of ICGEB. Program DRP is available together with SBASE but it uses a set of index-files requiring additional disk space. Copies of CHOPPER are also available from the authors <pongor@icgeb.trieste.it>.

Future work

Further work on SBASE will be carried out in two main directions. First, the standardization of domain names will continue in the next releases so as to allow simple retrievals for a larger number of domain types. Second, automated procedures are being developed in order to identify in other databases new SBASE entries' homologs that could be added to SBASE as domain candidates.

ACKNOWLEDGEMENTS

The authors thank Profs. Arturo Falaschi and Francisco Baralle (ICGEB, Trieste). SBASE was established in 1990 and is maintained collaboratively by the International Center for Genetic Engineering and Biotechnology, Trieste, Italy and the ABC Institute for Biochemistry and Protein Research, Gödöllö, Hungary.

REFERENCES

1. Barker,W.C., Hunt,L.T. and George,D.G. (1988) *Protein Seq. Data Anal.*, **1**, 363–373.
2. Baron,M., Norman,D.G. and Campbell,I.D. (1991) *Trends in Biochemistry*, **16**, 13–17.
3. Lipman,D.J. and Pearson,W.R. (1985) *Science*, **227**, 1436–1441.
4. Brutlag,D.L., Dautricourt,J.-P., Maulik,S. and Relph,J. (1990) *Comp. Appl. Biosci.*, **6**, 237–245.
5. Altschul,S.F. and Lipman,D.J. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 5509–5513.
6. Bork,P. (1992) *Curr. Opin. Struct. Biol.*, **2**, 413–421.
7. Bairoch,A. and Boeckmann,B. (1992) *Nucl. Acids Res.*, **20 suppl**, 2019–2022.
8. Barker,W.C., George,D.G., Mewes,H.-W. and Tsugita,A. (1992) *Nucl. Acids Res.*, **20 suppl**, 2023–2026.
9. Higgins,D.G., Fuchs,R., Stoehr,P.J. and Cameron,G.N. (1992) *Nucl. Acids. Res.*, **20 suppl**, 2071–2074.
10. Burks,C., Cinkosky,M.J., Fischer,W.M., Gilna,P., Hayden,J.E.-D., Keen,G.M., Kelly,M., Kristofferson,D. and Lawrence,J. (1992) *Nucl. Acids Res.*, **20 suppl**, 2065–2069.
11. Myers,F. (1990) Human retrovirus and AidsDatabase, Los Alamos National Laboratory, USA.
12. McKusick,V.M. (1990) Mendelian Inheritance in Man. John H Hopkins University Press, Baltimore, MD.
13. Roberts,R.J. and Macelis,D. (1992) *Nucl. Acids Res.*, **20 suppl**, 2167–2180.
14. Bairoch, A. (1992) *Nucl. Acids Res.*, **20 suppl**, 2013–2018.
15. IntelliGenetics, IG–Molecular Biology Software System, Release 5.4 for UNIX, February 1991.
16. Pongor,S., Skerl,V., Cserzö,M., Hátsági,Z., Simon,G. and Bevilacqua,V. (1993) *Prot. Eng.*, in press.
17. The Wide Area Information Server Project, (1991) Thinking Machines Corporation, Cambridge.
18. Simon,G., Paladini,R., Tisminetzky,S., Cserzo,M., Hatsagi,Z., Tossi,A. and Pongor,S. (1992) *Protein. Sec. Data Anal.*, **5**, 39–42.
19. Smith,R.F. and Smith,T.F. (1990) *Proc. Natl. Acad. Sci. USA*, **87**, 118–122.
20. Henikoff,S. and Henikoff,J.G. (1991) *Nucl. Acids Res.*, **19**, 6565–6572.
21. Smith,H.O., Annau,T.M. and Chandrasegaran,S. (1990) *Proc. Natl. Acad. Sci.*, **87**, 826–830.
22. Polner,G., Skerl,V. and Pongor,S. (1993) *Prot. Seq. Data Anal.*, in press.
23. Alberti,R., Anklesaria,F., Lindner,P., McCahill,M. and Torrey,D.: 'The Internet Gopher protocol: a distributed document search and retrieval protocol.', University of Minnesota Microcomputer and Workstation Networks Center, 1991–1992.