Protein Engineering vol.6 no.4 pp.391–395, 1993

The SBASE domain library: a collection of annotated protein segments

Sándor Pongor[1,2], Vesna Skerl[1,4], Miklós Cserzö[1,5], 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 Gödöllö, Hungary and [3]Department of Computer Sciences, The University of Chicago, Chicago IL, 60637, USA

[4]Permanent address: Institute for Nuclear Sciences—Vinca, PO Box 522, Belgrade, Yugoslavia

[5]Permanent address: Biological Research Center, Hungarian Academy of Sciences, 1518 Budapest 7, Hungary

SBASE is a database of annotated protein domain sequences representing various structural, functional, ligand binding and topogenic segments of proteins. The current release of SBASE contains 27 211 entries which are provided with standardized names in order to facilitate retrieval. SBASE 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) *Nucleic Acids Res.*, 20, Suppl., 2013–2118]. SBASE can be used to establish domain homologies through database search using programs 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 cannot be established. The use of SBASE is illustrated on the DNA binding protein Brain-4. The database 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>.

Key words: distant homologies/domains/protein sequence patterns

Introduction

Database search is probably the most widely used approach to predict the biological function of a newly determined protein sequence. If two large proteins share only a few common domains, detection of homology often becomes problematic. First, random identities present in the alignment may 'mask' the biologically important sequence patterns (Barker *et al.*, 1988; Baron *et al.*, 1991). Second, search programs such as FASTA (Lipman and Pearson, 1985), FASTDB (Brutlag *et al.*,1990) or BLAST3 (Altschul and Lipman, 1990) give only the best homology regions between the query and a database entry, so weaker homologies between the query and the same entry can remain undetected.

SBASE is a comprehensive collection of over 27 000 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 FASTA or FASTDB yields information on the potential functions of the detected homology regions which may allow or at least facilitate the detection of domain homologies and prediction of function.

Description of the database

Sources of domain sequence data

Sequence data in SBASE originate from three different sources: (i) from the SWISS-PROT protein sequence databank (Bairoch and Boeckmann, 1992); (ii) from the Protein Sequence Database of the Protein Identification Resource (PIR) (Barker *et al.*, 1992); and (iii) from the literature. These entries are either keyed in manually or are extracted and translated from the EMBL (Higgins *et al.*, 1992) or GenBank (Burks *et al.*, 1992) nucleotide sequence databases.

Definitions of domains

Domains included in SBASE are sequence segments with known structure and/or function. Several main types of domains are defined as follows. (i) Structural domains are sequence segments with a known structure [like the protein modules (Bork, 1992) such as epidermal growth factor-like (EGF-like) domains, immunoglobulin-like (IG-like) domains], biased composition (e.g. serine/threonine-rich domains) as well as various sequence repeats. (ii) 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. (iii) Ligand binding domains are sequence segments known to bind specific ligands (such as DNA, metals, sugars, etc.) (iv) 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). Redundancy of sequences in SBASE is kept to a minimum. In some cases, however, domains are defined in an overlapping fashion. For example, an extracellular domain (cellular location domain) may contain epidermal growth factor type repeats which are also represented as individual entries. In order to facilitate retrieval of sequences belonging to the same domain type SBASE domain names are standardized. Examples of these standardized names/keywords are listed in Table I. Table II lists the main domain types included in SBASE. Each of the main domain types is represented by domain group and the number of corresponding entries in SBASE are given in the table.

Domain boundaries are used as defined by the original authors or are determined by similarity to domains with defined boundaries. The statistical distribution of sequence lengths in SBASE is summarized in the histogram in Figure 1.

Format

The SBASE domain library is composed of domain sequence entries. Each entry is composed of lines, with a format similar to that used by the EMBL and SWISS-PROT databases. A sample entry is shown in Figure 2. The name of the domain is contained in the DE (definition) lines and is also partly repeated in the ID

Table I. Examples of standardized domain names in SBASE compared to domain names in SWISS-PROT and PIR databases

SWISS-PROT	PIR	SBASE
Actin binding Actin binding region	Actin binding	Actin binding
EGF-type EGF-like EGF-repeat EGF-repeat	EGF-like	EGF-repeat
EGF-like, type X	EGF-like, type X	EGF-repeat, Type X
Type X EGF-like repeat	Type X homology with EGF	(X = A,B)
Type-2 repeats (EGF-like)		
Fibronectin type X repeats	Fibronectin-like, type X	FN*n*-repeat
Fibronecting type-X Fibornectin type X module Fibronectin-like, type X	(X = I, II, III)	(*n* = 1,2,3)
IG-like XX-type domain (XX = V, C2, J)	Ig XXX chain V region (XXX = kappa, lambda, heavy)	IG-like, type V
IG V-like domain IG V region homology	Ig V region homology	
B, IG-like IG-like IG-like domain IG-related	Ig V region homology	IG-like
Ala-rich	Alanine-rich	Ala-rich

(identification) line. The name of the parent protein (i.e. the protein originally containing the domain) is given in the DP line.

SBASE is distributed in the following two formats: the IG concatenated format (annotations + sequences) and the FASTA format (only sequences), along with a keyword-based Sun-UNIX retrieval tool and documentation. For release 1.0, the two formats occupy 17.2 and 2.2 Mb, respectively.

Cross-references

SBASE 1.0 is cross-referenced with the following databases: SWISS—PROT protein sequence databank (Bairoch and Boeckmann, 1992), PIR, the protein sequence database of the Protein Identification Resource (Barker *et al.*, 1992), EMBL (Higgins *et al.*, 1992), GenBank (Burks *et al.*, 1992), HIV, the human retrovirus and AIDS database (McKusick, 1990; Myers, 1990), OMIM, McKusick's database of the Mendelian Inheritance in Man, REBASE, the database of type 2 restriction enzymes (Roberts and Macelis, 1992) and PROSITE, the dictionary of Protein Sites and Patterns (Bairoch, 1992). The cross-references are given in the DR lines of the entry.

The number of SBASE entries corresponding to several characteristic domain groups and the number of those among them cross-referenced to the corresponding PROSITE patterns (if existent), are given in Table I. These data show that some of the annotated domain sequences in SBASE are, in fact, not detected by PROSITE.

Search/retrieval tools

SBASE was originally developed in the UNIX environment (SUN 4/390, Sun OS 4.1.1) containing the IntelliGenetics sequence analysis package (IntelliGenetics, 1991).

The FASTA and the IG-formatted versions of SBASE can be searched by standard search-tools FASTA (Barker *et al.*, 1988) and FASTDB (Baron *et al.*, 1991) respectively, as well as program SCAN (Simon *et al.*, 1992) which is based on a window-sliding algorithm. SCAN performs individual searches using overlapping parts of the query and presents the results as a list of best domain-homologies along the query sequence.

DRP is a menu-oriented keyword-based retrieval program which allows viewing and saving of SBASE entries using any word occurring in the annotations as a keyword. DRP is a C shell script developed under Sun OS 4.1.1. DRP uses the WAISINDEX and WAISSEARCH programs originally developed as a part of the WAIS Wide Area Information Software Server (WAIS, 1991).

Contents of SBASE release 1.0

SBASE 1.0 (April 1992) contains 27 211 domain sequence entries comprising 1 551 445 amino acids. The complete database (both annotations and sequences) requires 17.2 Mb, and SBASE in FASTA format (sequences only) 2.2 Mb of disk storage space. Program DRP is distributed together with SBASE. DRP uses additionally a set of index files requiring 12 Mb of disk space.

Distribution

SBASE is distributed by Anonymous ftp from <ftp.icgeb.trieste.it>. Individual entries are available through the gopher server (Alberti *et al.*, 1991–1992) of ICGEB. Copies of SCAN are also available from the authors <pongor@icgeb.trieste.it>.

Detection of domain homologies using SBASE

Current methods of detecting domain homologies are based on *a priori* known consensus structures, represented as consensus sequences or regular expressions (Abarbanel *et al.*, 1982; Patthy,

Table II. Examples of domain types in SBASE release 1.0 and cross-references with PROSITE 8

Domain type	Number of SBASE records	Referenced to PROSITE a	b
Structural domains			
EGF-like domains	224	171	138
Ig-like domains	315	12	7
Fibronectin type III repeats	162	2	Ns
Glycine rich domains	91	16	Ns
Proline rich domains	84	3	Ns
Other miscellaneous repeat domains	3602	Nd	Nd
Homology domains*	2039	840	Nd
Ligand binding domains	3274	1484	1169
DNA binding	1764	1123	1127
Zn-fingers	1113	869	859
Calcium binding domains	642	85	85
Lectin domains	24	7	7
RNA binding domains	104	70	69
Cellular topology domains			
Signal peptides	3441	138	Ns
Transit peptides (organellar)	445	3	Ns
Nuclear localization signals	66	0	Ns
Extracellular regions	1458	379	Ns
Transmembrane regions	5197	167	Ns
Cyloplasmic regions	1293	135	Ns
Origin of domains			
Eukaryota	22114	Nd	Ns
Prokaryota	3457	Nd	Ns
Viridae	1640	Nd	Ns
Total	27211	5180	Nd

a, number of SBASE records cross-referenced to any PROSITE entry; b, number of cross-references to the PROSITE key corresponding to the domain type. Ns, corresponding key not specified in PROSITE; Nd, not determined.
*Regions of homology to other proteins, excluding those listed as structural domains (e.g. EGF-like).

1987), frequency-matrices (Gribskov *et al.*, 1987; Staden, 1988) or homology blocks (Posfai *et al.*, 1989; Seto *et al.*, 1990; Smith *et al.*, 1990; Hennikoff and Hennikoff, 1991). Database searching (e.g. FASTA, FASTDB) in itself does not require prior knowledge of the consensus structure of a domain but suffers from many technical drawbacks. For example, the outputs are usually not easy to interpret in terms of domain homologies. This problem is partly solved by using a database in which the entries are labeled by structure and function, which makes it possible to get output lists that are easy to interpret. The second type of problem is caused by random identities that are outside the region of interest and that may outscore biologically important alignment patterns. This problem can be partly solved by using a domain database, since the entries only contain the 'important' (i.e. annotated) part. This solution is not always satisfactory however, since the domain library may contain a large number of closely homologous entries which will thus occupy the top positions in the list, obscuring the more distantly homologous entries. The problem can be helped however, if one conducts separate database searches against SBASE by sliding a window of given length (usually 10–30 residues) along the query (Simon *et al.*, 1992). This procedure results in an ordered list of local homologies along the sequence. In other words, the amino acid sequence is transformed into a sequence of possible domain homologies that is quite simple to evaluate. This can be especially useful in the case of nucleic acid binding proteins that are known to have domains of very loosely defined structure.

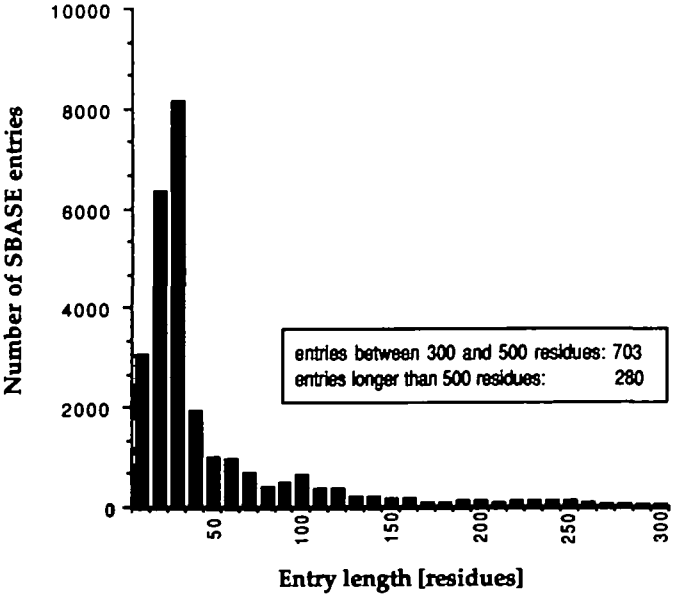

Fig. 1. Length distribution of entries in the SBASE domain library.

An example of analysis is illustrated in Figure 3 using brain specific protein 4 (Brain-4, BRN-4_RAT in SWISS-PROT) as the query. Brain-4 is a DNA binding protein of 336 amino acids that contains two well defined domains, the homeobox and the

```
; ID   ANX1$MOUSE-122-182 ANNEXIN-REPEAT.
; AC   SB01712
; DT   2-APR-1992
; DE   ANNEXIN-REPEAT.
; DP   ANNEXIN I (LIPOCORTIN I) (CALPACTIN II) (CHROMOBINDIN 9) (P35)
; DP   (PHOSPHOLIPASE A2 INHIBITORY PROTEIN).
; OS   MUS MUSCULUS (MOUSE).
; OC   EUKARYOTA; METAZOA; CHORDATA; VERTEBRATA; TETRAPODA; MAMMALIA;
; OC   EUTHERIA; RODENTIA.
; DR   SWISS-PROT; ANX1$MOUSE; P10107; AA 122-182
; DR   EMBL; X07486; MMLCIR.
; DR   EMBL; M24554; MMLCI.
; DR   PIR; S02181; S02181.
; DR   PIR; A32299; A32299.
; DR   PROSITE; PS00223; ANNEXIN.
; RA   SAKATA T., IWAGAMI S., TSURUTA Y., SUZUKI R., HOJO K., SATO K.,
; RA   TERAOKA H.;
; RL   NUCLEIC ACIDS RES. 16:11818-11818(1988).
ANX1$MOUSE-122-182
LRGAMKGLGTDEDTLIEILTTRSNEQIREINRVYREELKRDLAKDITSDTSGDFRKALLALA1
```

Fig. 2. Sample entry from the SBASE domain library. ID: entry identifier; DE: domain name; DP: name of parent protein; OS: the source organism; OC: taxonomic information; RA and RL: literature source; and DR: cross reference to other sequence databases

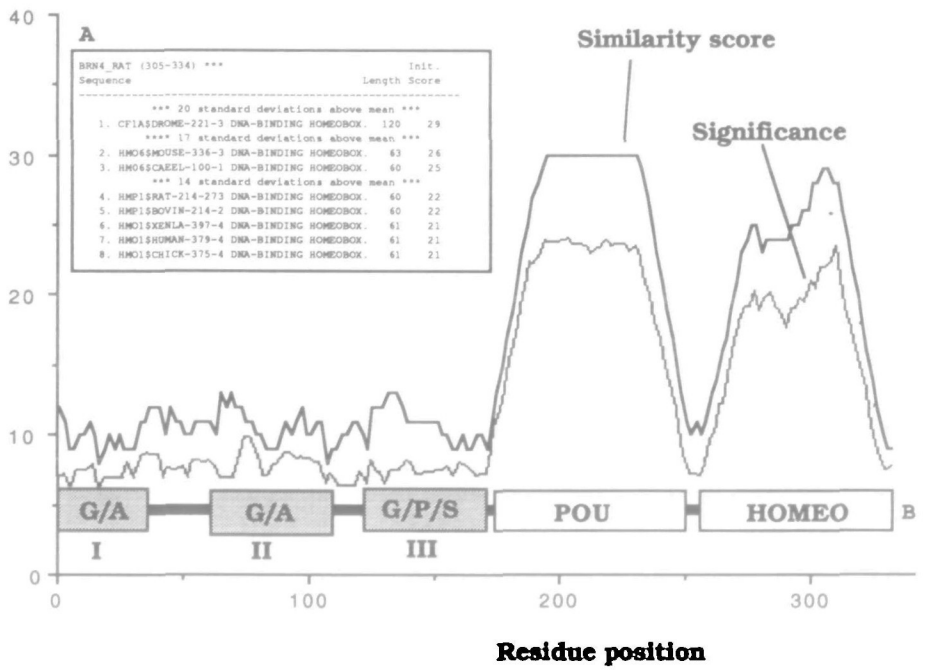

Fig. 3. Graphic representation of the homology search of the Brain-4 protein sequence, obtained by a SCAN (IntelliGenetics, 1991) search of the SBASE domain library, using a window of length 30 and increment of 2 residues (database search performed at every second residue). The result for each 30-residue long query fragment (window) is given as a value of the position of its beginning. The y-axis corresponds to the value of the initial score and the statistical significance value (Brutlag *et al.*, 1990). Inset A: partial list of best homologies for the homeobox domain. Inset B: graphic representation of the detected homologies.

POU domain. The homeobox is a well-conserved protein domain of 60 amino acids (Gehring and Hiromi, 1986; Gehring, 1987; Schofield, 1987; Scott *et al.*, 1989) which was first identified in a number of *Drosophila* homeotic and segmentation proteins. It has since been found in many other organisms including vertebrates and yeast (mating type proteins). Some of the proteins containing the homeobox are known to be transcription factors, e.g. liver-specific transcription factor LF-B1 (HNF1-α and β). Homeobox domains can be further classified into subfamilies according to their homologies to the *Drosophila* genes *engrailed, antennapedia* and *paired*. The 'POU' domain is 70–75 residues long and is found upstream of a homeobox domain in some eukaryotic transcription factors; it is thought to confer site-specific DNA binding and to mediate protein–protein interaction on DNA (Herr *et al.*, 1988; Levine and Hoey, 1988; Robertson, 1988; Rosenfeld, 1991; Schoeler, 1991; Treacy *et al.*, 1991). Apart from the mammalian brain-specific proteins Brain-1 (Brn-1), Brain-2 (Brn-2) and Brain-3, the POU domain can be found, among others, in Oct-1 (or OTF-1, NF-A1), a transcription factor for small nuclear RNA and histone H2B genes. The results summarized in Figure 3 show high levels of homology at regions corresponding to the homeobox and POU domains. The list of homeobox homologies (Figure 3, inset A) shows that this domain is similar to a closely related group of homeoboxes present in POU-containing transcription factors. This group seems to be more closely related to the paired type subfamily than to an *engrailed* or *antennapedia* subfamily of homeoboxes (data not shown). The N-terminal region shows spurious similarities that correspond to various domains rich in glycine, alanine and proline, as indicated with shaded boxes in inset B. The best

homologies found in these regions are as follows:

Region I (1–32):	Modulating domain of the androgen receptor from mouse, rat and man;
Region II (65–112):	Steroid binding domain of estrogen receptors; RNA binding domain of fibrillarin from man, rat and frog, a component of the small nuclear ribonucleoprotein particle;
Region III (125–176):	Modulating domain of retinoic acid receptors.

Even though these homologies are not mathematically significant, it is worthwhile to point out that they all refer to a well-defined biochemical context, nuclear proteins involved in nucleic acid binding. This information can be used to design biological experiments and to attempt building consensus patterns for new homology groups. Finally we mention that SBASE can be useful as a catalog of annotated segments in order to develop consensus representations, using programs such as PIMA (Smith and Smith, 1990), PROTOMAT (Hennikoff and Hennikoff, 1991) or MOTIF (Smith *et al.*, 1990).

Future work

Further work on SBASE will be carried out in two main directions. First, standardization of domain names will continue so as to allow easy retrieval of more domain types. Second, automated procedures are being developed in order to identify new homologs of SBASE entries in protein and nucleic databases which will be added to the database as domain candidates.

Acknowledgements

The authors thank Professors Arturo Falaschi (ICGEB, Trieste) and Amos Bairoch (University of Geneva) for help and advice. 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

Abarbanel,R.M., Wieneke,P.R., Mansfield,E., Jaffe,D.A. and Brutlag,D.L. (1982) *Nucleic Acids Res.*, **12**, 263–280.

Alberti,R , Anklesaria,F., Lindner,P., McCahill,M. and Torrey,D. (1991–1992) *The Internet Gopher Protocol: a distributed document search and retrieval protocol*. University of Minnesota Microcomputer and Workstation Networks Center.

Altschul,S.F. and Lipman,D.J. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 5509–5513.

Bairoch,A. (1992) *Nucleic Acids Res.*, **20**, suppl., 2013–2018.

Bairoch,A. and Boeckmann,B. (1992) *Nucleic Acids Res.*, **20**, suppl., 2019–2022.

Barker,W.C., Hunt,L.T. and George,D.G. (1988) *Protein Seq. Data Anal.*, **1**, 363–373.

Barker,W.C., George,D.G., Mewes,H.-W. and Tsugita,A. (1992) *Nucleic Acids Res.*, **20**, suppl., 2023–2026.

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

Bork,P. (1992) *Curr. Opin. Struct. Biol.*, **2**, 413–421.

Brutlag,D.L., Dautricourt,J.-P., Maulik,S. and Relph,J. (1990) *Comp. Appl. Biosci.*, **6**, 237–245.

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) *Nucleic Acids Res.*, **20**, suppl., 2065–2069.

Gehring,W.J. (1987) *Science*, **236**, 1245–1252.

Gehring,W.J. and Hiromi,Y. (1986) *Annu. Rev. Genet.*, **20**, 147–173.

Gribskov,M., McLachlan,A.D. and Eisenberg,D. (1987) *Proc. Natl Acad. Sci. USA*, **84**, 4355–4358.

Hennikoff,S. and Henikoff,J.G. (1991) *Nucleic Acids Res.*, **19**, 6565–6572.

Herr,W., Sturm,R.A., Clerc,R.G., Corcoran,L.M., Baltimore,D., Sharp,P.A., Ingraham,H.A., Rosenfeld,M.G., Finney,M., Ruvkun,G. and Horvitz,H.R. (1988) *Genes Dev.*, **2**, 1513–1516.

Higgins,D.G., Fuchs,R., Stoehr,P.J. and Cameron,G.N. (1992) *Nucleic Acids Res.*, **20**, suppl., 2071–2074.

IntelliGenetics, IG—Molecular Biology Software System, Release 5.4 for UNIX, February 1991.

Keen,G.M., Kelly,M., Kristofferson,D. and Lawrence,J. (1992) *Nucleic Acids Res.*, **20**, suppl., 2065–2069.

Levine,M. and Hoey,T. (1988) *Cell*, **55**, 537–540.

Lipman,D.J. and Pearson,W.R. (1985) *Science*, **227**, 1436–1441.

McKusick,V.M. (1990) *Mendelian Inheritance in Man*. John H. Hopkins University Press, Baltimore, MD.

Myers,F. (1990) *Human Retrovirus and Aids Database*. Los Alamos National Laboratory, USA.

Patthy,L. (1987) *J. Mol. Biol.* ,**198**, 567–577.

Posfai,J., Bhagwat,A.S., Posfai,G. and Roberts,R.J. (1989) *Nucleic Acids Res.*, **17**, 2421–2435.

Robertson,M. (1988) *Nature*, **336**, 522–524.

Roberts,R.J. and Macelis,D. (1992) *Nucleic Acids Res.*, **20**, suppl., 2167–2180.

Rosenfeld,M.G. (1991) *Genes Dev.*, **5**, 897–907.

Schoeler,H.R. (1991) *Trends Genet.*, **7**, 323–329.

Schofield,P.N. (1987) *Trends Neurosci.*, **10**, 3–6.

Scott,M.P., Tamkun,J.W. and Hartzell,G.W. III. (1989) *Biochim. Biophys. Acta*, **989**, 25–48.

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

Simon,G., Paladini,R., Tisminetzky,S., Cserzo,M., Hatsagi,Z., Tossi,A. and Pongor,S. (1992) *Protein Sec. Data Anal.*, **5**, in press.

Smith,R.F. and Smith,T.F. (1990) *Proc. Natl Acad. Sci. USA*, **87**, 188–222.

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

Staden,R. (1988) *Comput. Appl. Biosci.*, **4**, 53–60.

Treacy,M.N., He,X. and Rosenfeld,M.G. (1991) *Nature*, **350**, 577–584.

WAIS (The Wide Area Information Server) Project (1991) Thinking Machines Corporation, Cambridge, MA.

Received on September 16, 1992; revised on January 18, 1993; accepted on January 30, 1993