Protein Seq Data Anal (1992) 5:39–42

Protein Sequences & Data Analysis

Improved detection of homology in distantly related proteins: similarity of adducin with actin-binding proteins

György Simon[1], Rudy Paladini[1], Sergio Tisminetzky[1], Miklòs Cserzö[1], Zsolt Hátsági[2,3], Alessandro Tossi[1], Sándor Pongor[1,3]

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

Received December 20, 1991 / Accepted March 27, 1992

Abstract. A novel and generally applicable method is described for the detection of homology in distantly related proteins using a new domain sequence database that contains over 20000 protein sequence segments of known function. The use of the method is illustrated on distantly related domains shared by complement components C1S and C1R, calcium-dependent serine proteinase and bone morphogenetic protein 1. New homologies are shown between human adducin and the actin-binding domains of alfa-actinin and dystrophin.

Introduction

If two distantly related sequences share only a few common domains, detection of homology is often problematic since random identities present in the alignment often "mask" the biologically important sequence patterns [1, 2]. Currently, detection of such homologies is possible only if the sequence pattern to be located is a priori known from multiple alignment. Even though collections of sequence patterns [3, 4] and search algorithms [5–7] are now available, their use is quite limited as currently there are no consensus patterns available for the majority of the known domains. Another limitation of pattern matching is the sensitivity to single mismatches that may lead to failure even if the rest of the pattern is entirely conserved.

Here we show that a good part of these problems can be efficiently solved through the use of a comprehensive library of known functional domains with only a moderate and quite acceptable increase in the required computation (CPU) time. This method does not require a priori knowledge of the pattern and can work also with domains for which consensus patterns cannot be developed.

Methods

All calculations were carried out on a SUN 4/390 computer under the SunOS 4.0.3 operating system (equivalent to UNIX BSD 4.2).

Correspondence to: G. Simon

Program *Scan*, written in C, compares all segments of a given length from a query sequence to a database. A window of length w slides along the sequence at increment steps i and a fast database search, according to the algorithm of Brutlag et al. [8], is carried out at every position of the window. The time requirement of this scanning procedure is proportional to O(W*L*N), where W is the length of the window, L is the length of the query and N is the length of the database. In the usual case, a separate database search can be performed at every residue of the query. In order to save CPU time, the value of i can be increased and the searches will thus be performed at every ith residue, i.e., the actual time requirement will decrease to O(W*L*N/i). For example, analysis of a 450 – residue sequence with a 50 – residue window in 200 steps using the *Sbase* domain library takes about 2 h on a SUN 4/390 workstation.

Sbase is an experimental library of annotated amino acid sequence segments. The domain sequences were collected from three sources: (a) protein sequence databases (Swiss-Prot, PIR), (b) journal articles and (c) as translations from nucleic acid databases (EMBL, GenBank). Domain boundaries were used as defined by the original authors or were determined by similarity to domains with defined boundaries. The present version of *Sbase* contains over 20000 entries (1287000 residues) and requires a storage capacity of 13 megabytes (Table 1). The entry structure follows that of the EMBL and Swiss-Prot databases [9], and contains information on the function of the segment, the name and the database entry of the original protein sequence from which the segment was derived, the source organism, as well as on the literature source (Fig. 1).

Results

Detection of domains in C1S heavy chain

The principle of the method is illustrated using the human complement 1S heavy chain sequence C1SH [10] as an example. C1SH contains one EGF-like domain, two C3B/C4B interaction repeats, and two 100-residue-long repeats. Figure 2 shows the scores obtained when C1SH was compared to a model database that included its four constituent domains. In this representation, local homologies appear as peaks. The efficiency of detection can be characterized by a signal-to-noise ratio which increases with the window size (Fig. 2a). However, we found that a window of 30–50 residues is satisfactory

Table 1. Examples of entries in the *Sbase* domain-library[a]

Structural domains	
EGF-like domains	124
Ig-like domains	149
Fibronectin type III repeats	79
Other repeat units	2994
Glycine-rich domains	65
Ser/Thr-rich domains	121
Ligand-binding domains	
Zn-fingers	126
Calcium-binding domains	501
Other-ligand binding domains	2151
Cellular topology domains	
Signal peptides	3121
Transit peptides (chloroplast and mitochondrial)	390
Extracellular regions	1180
Transmembrane regions	4036
Cytoplasmic regions	965
Prokaryotic sequences	2677
Eukaryotic sequences	18353

[a] The present version of the *Sbase* domain library contains over 22426 sequences (1287119 residues) and occupies 13 Mb of storage space

for detecting homologies to most known domain types. Shorter windows (10–15 residues) show short motifs more efficiently but may miss some long and very sparse motifs, like fibronectin type III domains. The value of the increment step *i* is the resolution of the plot. Its value was usually taken as 1 or 2, i.e., a database search was performed at every (or every second) residue.

When C1S was compared with the *Sbase* domain library, using a 50-residue window, similartiy with the repeat regions in calcium-dependent serine proteinase CASP [11], complement component C1R [12], and bone morphogenetic protein BMP1 [13] became apparent (Fig. 2b). (Very recently, the same homology was shown by P. Bork, using a more complex method based on multiple alignment and amino acid property patterns [14].

In practice we used two search strategies: (1) to find local homologies to known domains the query can be compared to the domain library; and (2) to find new common motifs with regions that are not included in the domain-library, the query can be compared to a database of functionally related proteins. In both cases, the time requirement of the analysis can range from 10 min to a few hours, depending on the run parameters of the database search [8]. We illustrate these strategies with the example of the rat adducin sequence.

Adducin contains regions similar to actin-binding domains of dystrophin and alfa-actinin

Adducin is a membrane skeletal protein of unknown function that interacts with spectrin, calmodulin and actin [15, 16]. The sequence of rat adducin has recently been elucidated [17]. The database search revealed no closely related proteins, only spurious similarities to coiled-coil domains of various filamentous proteins were found. As similarity of coiled-coil regions is in any case rather unspecific because of the approximate heptade repeats present in these sequences, we conducted a search for further homologies with known domains. We compared 50-residue segments of the rat adducin sequence (551 residues in total) with the complete *Sbase* domain library. In addition to the already known similarity with coiled-coil domains (not shown), this analysis revealed a similarity with a well-conserved N-terminal actin-binding region of dystrophins [18] for the region between residues 380 and 460 of rat adducin (Fig. 3). In a search for further local homologies we selected the

```
; ID   7LES$DROME-381-501
; AC   M00135
; DT   21-Jun-1991 (MANUAL)
; DE   REPEAT FIBRONECTIN TYPE III
; DE   SEVENLESS PROTEIN (GENE NAME: SEV).
; OS   FRUIT FLY (DROSOPHILA MELANOGASTER).
; OC   EUKARYOTA; METAZOA; ARTHROPODA; INSECTA; DIPTERA.
; RA   NORTON, P.A., HYNES, R.O., REES, D.J.G.
; RL   CELL 61:15-16(1990).
; DR   SWISS-PROT; 7LES$DROME AA 381-501
7LES$DROME-381-501
rtqpqlerapradgqstpltirwamhfpehylasrpfniqyqfvdhhgeeldleqedqdasgetgssawf
nladydcdeyymceilealipytqyrfrfelpfgenrdevlyspatpayqtl


; ID   ZNFP$LYCVA-32-53
; AC   SB22364
; DT   27-Jun-91
; DE   ZINC-FINGER
; DP   ZINC FINGER PROTEIN (GENE NAME: Z).
; OS   LYMPHOCYTIC CHORIOMENINGITIS VIRUS (STRAIN ARMSTRONG).
; OC   VIRIDAE; SS-RNA ENVELOPED VIRUSES; NEGATIVE-STRAND; ARENAVIRIDAE.
; DR   SWISS-PROT; ZNFP$LYCVA; AA 32-53
; DR   EMBL; M27693; LCVZFP.
; DR   PIR; A32592; A32592.
; RA   SALVATO M.S., SHIMOMAYE E.M.;
; RL   VIROLOGY 173:1-10(1989).
ZNFP$LYCVA-32-53
ckscwqkfdslvrchdhylcrhl
```

Fig. 1. Sample entries from the *Sbase* domain library. *ID* Entry identifyer; *DE* domain name; *DP* name of parent protein; *OS* the source organism; *OC* taxonomic information; *RA* and *RL* literature source; *DR* cross reference to other sequence databases

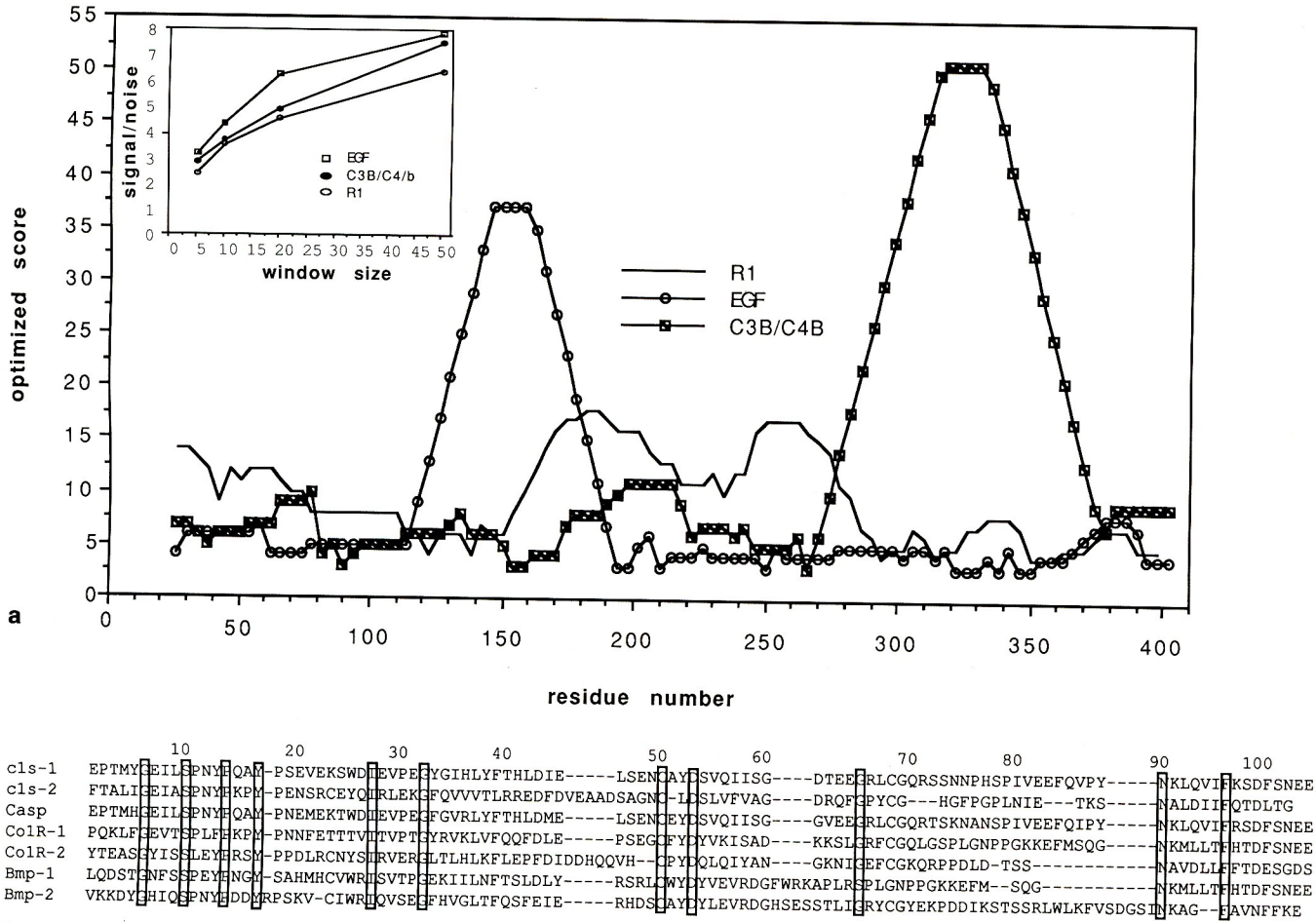

Fig. 2a, b. Segment analysis of the complement component C1S heavy chain sequence. Optimal scores of comparison plotted as a function of sequential position, showing the EGF-like domain, C3B/C4B interaction repeat, and R1 long repeat of C1S. **a** Signal to noise ratio (peak height divided by average baseline value) as a function of window size w. **b** Multiple alignment of the long repeats of C1S with those found in C1R, calcium sensitive proteinase and bone morphogenetic protein

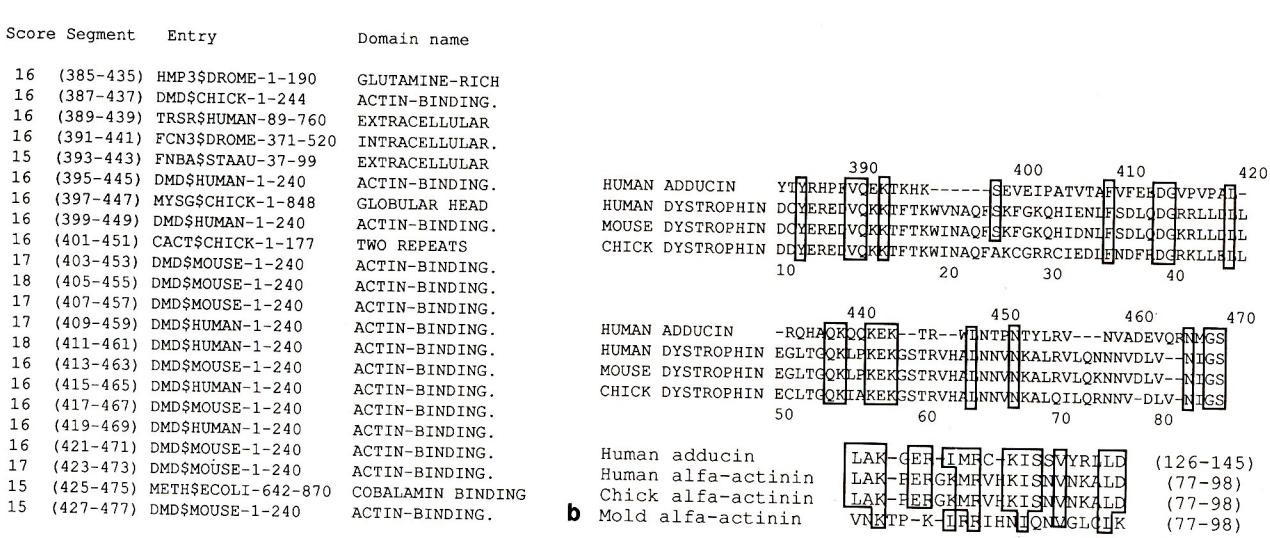

Fig. 3a, b. Comparison of the rat adducin sequence to the *Sbase* domain library. **a** List of best scoring domains (detail). Segment analysis of human adducin (sample output). **b** Alignment of human adducin region 380–480 with actin-binding domains of dystrophin and of adducin segment 125–145 with alpha-actinins

known actin-binding proteins as a database (70411 residues in total) and scanned the adducin sequence with a 15-residue window. This analysis confirmed the similarity to dystrophins, but also showed an additional region of homology between residues 120 and 150 to the actin-binding domains of alfa-actinins. Multiple alignment of these sequences revealed a motif conserved among several species which maps to the actin-binding domain of alfa-actinins (Fig. 3b). As actin cross-linking proteins are known to contain both coiled-coil repeats and actin-binding domains [18], it appears that rat adducin may have a domain structure similar to this group of proteins. We note that this homology could not be detected using the Prosite [3] or KeyBank [4] collection of sequence patterns (data not shown).

Discussion

Database searches and pattern searches are two extreme methods that are available to predict the biological function of a newly determined protein sequence. Database searches by the current algorithms (cf [8]) are rapid and require no prior knowledge on the expected alignment; however, they are not very useful for detecting distantly related domains. Pattern matching methods, on the other hand, are strongly limited by the number of known sequence patterns, which is small as compared to the number of known domains. Here we describe a method that requires slightly more CPU time than simple database search or pattern matching, but that makes it possible to detect local homologies which may not be detected by those methods. As the domain library contains several examples for most known domain types, this type of analysis is less sensitive to single mismatches than pattern matching. The results appear as a list of local homologies to known domains that can yield information on possible structural similarities, binding functions or cellular location of individual sequence regions. In other terms, the amino acid sequence is transformed into a sequence of possible domain homologies that is quite simple to evaluate. The procedure of pattern verification is not automated, however. The experimenter still has to decide if a local homology is "meaningful" or not, which may introduce a certain measure of subjectivity.

Copies of *Scan* and *Sbase* can be obtained on request from S. Pongor at ICGEB, Trieste (pongor@genes.icgeb.trieste.it).

Acknowledgements. The authors thank Professors Arturo Falaschi and Francisco Baralle (ICGEB, Trieste) for useful discussions.

References

1. Barker WC, Hunt LT, George DG (1988) Protein Seq Data Anal 1:363–373
2. Baron M, Norman DG, Campbell ID (1991) Trends Biochem Sci 16:13–17
3. Bairoch A (1989) PROSITE: a dictionary of protein sites and patterns. Release 4.0, EMBL Biocomputing Technical Document 4. EMBL, Heidelberg, FRG
4. Moore LJ, Moore JT (1990) "KEYBANK: intelligenetics databank of tested nucleic acid and protein sequence patterns". Release 6.0. Intelligenetics, Mountain View
5. Abarbanel RM, Wieneke PR, Mansfield E, Jaffe DA, Brutlag DL (1982) Nucleic Acids Res 12:263–280
6. Patthy L (1987) J Mol Biol 198:567–577
7. Gribskov M, McLachlan AD, Eisenberg D (1987) Proc Natl Acad Sci USA 84:4355–4358
8. Brutlag DL, Dautricourt J-P, Maulik S, Relph J (1990) Comp Appl Biosci 6:237–245
9. Bairoch A (1990) Swiss-Prot Protein Sequence Database. EMBL Data Library. Release 14, EMBL, Heidelberg, FRG
10. Tosi M, Duponchel C, Meo T, Julier C (1987) Biochemistry 26:8516–8524
11. Kinoshita H, Sakiyama H, Tokinaga K, Imajoh-Ohmi S, Hamada Y, Isono K, Skiyama S (1989) FEBS Lett 250:411–415
12. Leytus SP, Kurachi K, Sakariasen KS, Davie EW (1986) Biochemistry 25:4855–4863
13. Wozney JM, Rosen V, Celeste AJ, Mitsock LM, Whithers MJ, Kriz RW, Hewick RM, Wang EA (1988) Science 242:1528–1534
14. Bork P (1991) FEBS Lett 282:9–12
15. Gardner K, Bennett V (1986) J Biol Chem 261:1339–1348
16. Bennet V, Gardner K, Steiner JP (1988) J Biol Chem 283:5860–5869
17. Tripodi G, Piscone A, Borsani G, Tisminetzky S, Salardi S, Sidoli A, James P, Pongor S, Bianchi G, Baralle F (1991) Biochem Biophys Res Commun 177:939–947
18. Matsudaira P (1991) Trends Biochem Sci 16:87–92