CABIOS APPLICATIONS NOTES

Vol.9. no.3. 1993
Pages 371–372

Predicting potential domain homologies from FASTA search results

Hedvig Hegyi and Sándor Pongor[1]*

Here we describe a simple and fast algorithm designed to extract recurrent domain-homologies from FASTA (Pearson and Lipman, 1988) search results obtained between a query and domains annotated in the SWISS-PROT database (Bairoch and Boeckmann (1992). The program evaluates FASTA outputs that contain the annotation part of the SWISS-PROT entries, for example such as can be produced with the GCG version of FASTA (Devereux *et al.*, 1984). Briefly, the length of overlap is calculated between the aligned region and all domains annotated in the feature table of a given entry. Each annotated domain is then given a fraction of the score values (init1, initn and opt) proportional to the length of the overlap. The program sums up the scores by feature names, then prepares a list of feature names ranked according to the assigned score values calculated from init1.

Figure 1 shows the analysis of human complement C1S heavy chain (C1SH) and light chain (C1SL) (Tosi *et al.*, 1987). In both outputs, the names of the expected domains appear in the top-ranking positions. The homology to serine proteases is apparent in the C1SL output (Figure 1B) even though this feature is not annotated in most enzymes of this class. The information is of an approximate nature, however. First, not all proteins and domain-types have annotated homologs in the database. For example, the feature-name REPEAT in Figure 1A designates in fact the 100-residue-long domain, recently described by Bork (1991). Second, the domain annotations in SWISS-PROT are not entirely uniform yet (e.g. 'C3B/C4B INTERACTION DOMAIN' and 'SCR' are alternative names). Third, the simple scoring method implies that the distribution of identities is uniform within the alignment, which is an obvious simplification. In spite of these caveats we believe that FTHOM can give a useful and quick qualitative summary of the domain-composition of a protein. The advantage of this approach lies in the fact that it uses the entire information given in an annotated database, and not only a consensus representation of the domains.

FTHOM is written in written in C for VAX environments, and in Pascal for IBM PCs. The source code can be obtained on request from the authors (h1546heg@ella.hu, pongor@icgeb.trieste.it) and will be available through anonymous ftp to ftp.icgeb.trieste.it

ABC Institute for Biochemistry and Protein Research, 2100 Gödöllö, Hungary and [1]International Centre for Genetic Engineering and Biotechnology, Area Science Park, 34012 Trieste, Italy
*To whom correspondence should be addressed

References

Bairoch, A. and Boeckmann, B. (1992) *Swiss-Prot Protein Sequence Database*, EMBL Data Library, European Molecular Biology Laboratory, Heidelberg, Germany, release 20.

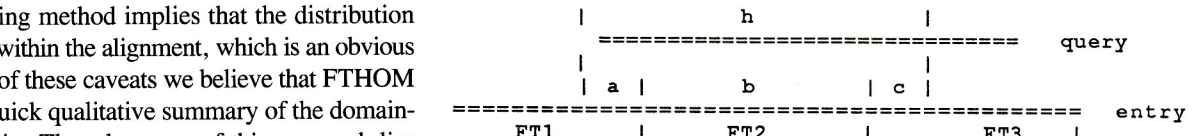
Bork, P. (1991) Complement components C1r/C1s, bone morphogenetic protein 1 and *Xenopus laevis* developmentally regulated protein UVS.2 share common repeatS. *FEBS Lett.*, **282**, 9–12.

Devereux, J., Haeberli, P. and Smithies, O. (1984) A comprehensive set of sequence analysis programs for the VAX. *Nucleic Acids Res.*, **12**, 387–395.

Pearson, W. R. and Lipman, D.J. (1988) Improved tools for biological sequence comparisons *Proc. Natl. Acad. Sci. USA*, **85**, 2444–2448.

Tosi, M., Duponchel, C., Meo T. and Julier, C (1987) Complete cDNA sequence of human complement C1s and close physical linkage of the homologous genes C1s and C1r. *Biochemistry*, **26**, 8516–8524.

The Scoring scheme used by FTHOM. FT1, FT2 and FT3 are features (domains) annotated in the query. h is the length of the homology region in the FASTA or BLAST alignment, a, b and c are the lengths of the overlaps with FT1, FT2 and FT3, respectively [expressed as number of amino acid residues]



$$Score(FT1) = \frac{a}{h} * Score(FASTA)$$

$$Score(FT2) = \frac{b}{h} * Score(FASTA)$$

$$Score(FT3) = \frac{c}{h} * Score(FASTA)$$

A C1S HEAVY CHAIN

(Peptide) FASTA of: C1s$Human from: 16 to: 437 April 21, 1992 14:22
Max. number of entries:100
Weighted scores

Feature name	FT freq	init1	initn	opt
REPEAT	13	405	727	981
REPEAT C3B/C4B INTERACTION REPEAT.	18	434	616	886
DOMAIN EXTRACELLULAR.	9	468	723	644
DOMAIN GLUTAMATE-RICH.	1	441	441	483
DOMAIN EGF-LIKE, TYPE B.	6	278	323	383
DOMAIN 9 REPEATS OF TYPE II EGF-LIKE REPEAT.	3	270	270	372
REPEAT TYPE II.	6	270	270	372
DOMAIN EGF-LIKE.	3	221	281	343
DOMAIN C3B/C4B INTERACTION DOMAIN (4 REPEATS).	2	188	188	294
DOMAIN C3B/C4B INTERACTION DOMAIN (20 REPEATS).	2	106	163	214
DOMAIN C5B-BINDING DOMAIN.	1	40	63	145
DOMAIN C3B/C4B INTERACTION DOMAIN (8 REPEATS).	1	78	78	140
DOMAIN ALPHA CHAIN.	1	117	117	138
DOMAIN EXTRACELLULAR, REPEAT-RICH REGION.	1	58	77	114
DOMAIN CONTAINS 16 EGF-LIKE REPEATS AND 3 CYS-RICH REPEATS.	1	98	123	108
DOMAIN LONG HOMOLOGOUS REPEAT B.	1	69	122	105
REPEAT EGF-LIKE 6.	3	69	120	100
DOMAIN DOMAIN III (7-EGF REPEATS).	2	84	127	93
SIMILAR WITH EGF PRECURSOR.	1	69	69	86

B C1S LIGHT CHAIN

(Peptide) FASTA of: C1s$Human from: 438 to: 688 April 21, 1992 15:18
Max. number of entries:100
Weighted scores

Feature name	FT freq	init1	initn	opt
DOMAIN SERINE PROTEASE.	28	2179	5043	6697
DOMAIN CATALYTIC.	1	77	166	301
SIMILAR WITH TRYPSIN.	1	71	202	269
DOMAIN BETA CHAIN.	1	69	199	256
DOMAIN SERINE-PROTEASE.	1	61	159	122
DOMAIN SEGMENT A.	1	36	96	112
DOMAIN SEGMENT B2.	1	23	62	72
DOMAIN SEGMENT B1.	1	20	53	62
DOMAIN SEGMENT C.	1	13	34	40
PROPEP PROLINE-RICH.	3	12	20	27
SITE MAY BE REMOVED BUT IS NOT NECESSARY FOR ACTIVATION.	1	5	14	19
CA_BIND	1	3	9	8
SITE CLEAVAGE SITE (BY FACTOR XA).	3	0	3	3
SITE CLEAVAGE (BY FACTOR XIA) (BY SIMILARITY).	2	0	2	2
SITE CLEAVAGE (BY FACTOR XIA).	2	0	2	2
BINDING SUBSTRATE.	2	2	4	2
SITE CLEAVAGE (BY FACTOR XA).	1	0	1	1
PROPEP ACTIVATION PEPTIDE.	1	0	1	1
SITE CLEAVAGE BY FACTOR XA, FACTOR XIIA,	1	0	1	1
SITE THROMBIN ACTIVATION SITE.	1	0	1	1

Fig. 1. FTHOM output obtained on the C1S heavy chain (**A**) and light chain (**B**) sequence from the analysis of the 100 top-ranking FASTA alignments.