BIOINFORMATICS APPLICATIONS NOTE

Vol. 15 no. 4 1999
Pages 343–344

The domain-server: direct prediction of protein domain-homologies from BLAST search

János Murvai[1], Kristian Vlahovicek[1], Endre Barta[2], Subbiah Parthasarathy[1], Hedvig Hegyi[3], Friedhelm Pfeiffer[4] and Sándor Pongor[1]

[1]International Centre for Genetic Engineering and Biotechnology, 34012 Trieste, Italy, [2]Agricultural Biotechnology Center, 2100 Gödöllö, Hungary, [3]Department of Molecular Biochemistry and Biophysics, Yale Medical School, New Haven, CT 06520, USA and [4]Munich Information Center for Protein Sequences, GSF-Forschungszentrum für Umwelt und Gesundheit AG MIPS, Max-Planck-Institut für Biochemie, Martinsried, Germany

Received on November 12, 1998; revised on January 12, 1999; accepted on January 14, 1999

Abstract

Results: *A WWW server for protein domain homology prediction, based on BLAST search and a simple data-mining algorithm (Hegyi,H. and Pongor,S. (1993)* Comput. Appl. Biosci.*,* ***9****, 371–372), was constructed providing a tabulated list and a graphic plot of similarities.*
Availability: *http://www.icgeb.trieste.it/domain. Mirror site is available at http://sbase.abc.hu/domain. A standalone programme will be available on request.*
Contact: *pongor@icgeb.trieste.it*
Supplementary information: *A series of help files is available at the above addresses.*

Difficult' protein domain homologies — e.g. those that are not included in protein motif databases — can be best predicted by visual evaluation of database search results and scrutinizing database record annotations. This laborious procedure can be facilitated by a simple algorithm, FTHOM (Hegyi and Pongor, 1993), that systematically compares the alignments with the feature table of each database entry. The result is a ranked list of the most probable domain homologies. The problem of finding weak domain homologies can be best described as a sorting task. The strategy used by the original FTHOM algorithm is to re-sort the search output according to the name of the domains that the individual alignments hit in each database record (Hegyi and Pongor, 1993). In the present version we apply an additional sorting dimension, the sequence position within the query. The domain similarities are projected back to the query sequence, and so local similarities will produce peaks in a similarity versus sequence plot. The main improvements and modifications are the following:

(a) Use of BLAST 1.4 (Altschul *et al*., 1990) instead of FASTA (Lipman and Pearson, 1985) gave an increase in speed and sensitivity, the latter is due to the separate scoring of individual short alignments (contigs) and to the complexity-filtering (Wootton, 1994).

(b) Standardization of domain names. In sequence databases, protein domains are often described under similar but not identical names. Instead of these, we now employ standardized names developed for the SBASE protein domain library (Murvai *et al*., 1999). For PIR searches we have retained the domain names used in Protfam (Mewes *et al*., 1998).

(c) Preprocessing of the annotations. The feature table of the protein sequence database is now preprocessed into an indexed database and information is retrieved 'on the fly' as the program processes the search output. This makes it possible for us to include additional databases, such as the PIR International Sequence Database (Barker *et al*., 1998).

(d) Each domain similarity found is characterized by a number of parameters such as: (i) number of times the domain type was hit by the query (NSD), (ii) cumulative similarity score (SUM), (iii) average score (SUM/NSD), and (iv) maximal similarity score found (MSC). In addition, the number of times a given domain type occurs in the database (GN) is also given in the output. This makes it possible to find out how 'typical' a new domain similarity is. Namely, if the query is similar to the majority of the entries of a given domain type, the similarity is probably not accidental. For example, the query used in Figure 1A (C1S_HUMAN of Swissprot) is similar to 175 out of 285 trypsin-like domains in the PROTFAM database. On the other hand, the query is strongly homologous to only 5 entries out of 496 'COMPLEMENT FACTOR H REPEAT HOMOLOGY' (Sushi) domains in the database.

(e) Graphic plotting of selected domain similarity scores along the query sequence whereby significant domain similarities will show up as peaks. These plots — in a successful case — will reflect the domain architecture of the query (Figure 1B), in addition to the list of constituent domains (Figure 1A). We note that the evaluation requires biological knowledge. The server does not report the number of domains found within the query, e.g. the large peak in Figure 1B represents two complement factor H homology (Sushi) domains.

A

```
                Feature name            NSD     Gn.  Sum.Score  Sum/NSD Overlap Max.
_______________________________________ ____   ____  _________  _______ ____________

TRYPSIN HOMOLOGY                         173    285     11367      65.7       100
C1R/C1S REPEAT HOMOLOGY                    9     38      1907     211.9      1070
TRYPSIN HOMOLOGY (FRAGMENT)               14     41      1121      80.1        96
COMPLEMENT FACTOR H REPEAT HOMOLOGY        5    496       697     139.4       475
EGF HOMOLOGY                               5    381       404      80.8       172
```

B.

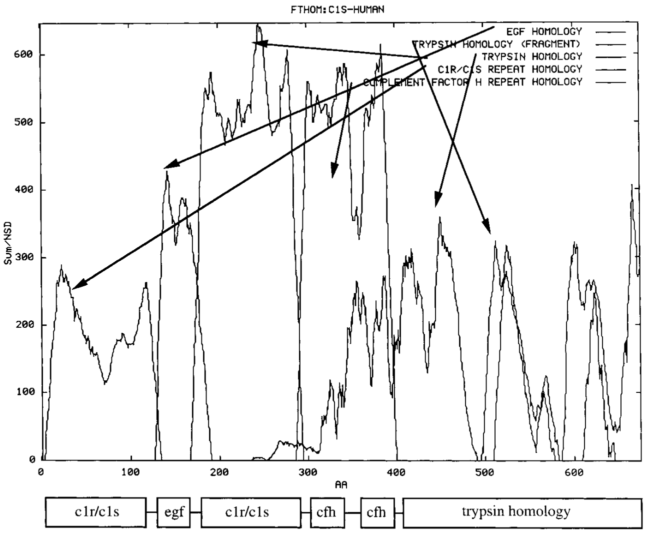

Fig. 1. *FTHOM* output obtained on the C1S protein (C1S_HUMAN, heavy and light chains) run against the PIR-International database. (**A**) Tabulated output of best domain homologies. (**B**) Graphic output of the same. The numerical values of the local homologies are multiplied by a common scaling factor (100 in this case) and smoothed with a window of 15 positions. (The picture in the output is in colours. The arrows are added only here, for better identification.) It is noted that the output reflects the correct domain structure of the protein, shown as a cartoon below the diagram (‘cofh’ indicates complement factor H homology, see text for other abbrevieations).

Acknowledgements

The ICGEB group is supported by EMBnet, the European Molecular Biology Network in the framework of EU grant ERBBIO4-CT96-0030. Work by MIPS was supported by grants from the Bundesministerium für Bildung, Forschung und Technologie (BMBF, FKZ 0311670, 01KW9703/7) and the European Commission (BIOCT-CT-96110). JM thanks Csanád Bachraty (Biological Research Center, 6723 Szeged, Hungary) for his help during the early phases of this project.

References

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Molec. Biol.*, **215,** 403–410.

Barker,W.C., Garavelli,J.S., Haft,D.H., Hunt,L.T., Marzec,C.R., Orcutt,B.C., Srinivasarao,G.Y., Yeh,L.S.L., Ledley,R.S., Mewes,H.W., Pfeiffer,F. and Tsugita,A. (1998) The PIR-international protein sequence database. *Nucleic Acids Res.*, **26,** 27–32.

Hegyi,H. and Pongor,S. (1993) Predicting potential domain homologies from FASTA search results. *Comput. Appl. Biosci.*, **9,** 371–372.

Lipman,D.J. and Pearson,W.R. (1985) Rapid and sensitive protein similarity searches. *Science*, **227,** 1435–1441.

Mewes,H.W., Hani,J., Pfeiffer,F. and Frishman,D. (1998) MIPS: a database for protein sequences and complete genomes. *Nucleic Acids Res.*, **26,** 33–37.

Murvai,J., Barta,E., Vlahovicek,K., Szepesvári,C., Acatrinei,C. and Pongor,S. (1999) The SBASE protein domain sequence library release 6.0. *Nucleic Acids Res.*, **27**, 257–259.

Wootton,J.C. (1994) Non-globular domains in protein sequences: automated segmentation using complexity measures. *Computat. Chem.*, **18,** 269–285.