BIOINFORMATICS APPLICATIONS NOTE Vol. 16 no. 12 2000
Pages 1155–1156

A simple probabilistic scoring method for protein domain identification

János Murvai, Kristian Vlahoviček and Sándor Pongor

Protein Structure and Function Group, International Centre for Genetic Engineering and Biotechnology, Area Science Park, 34012 Trieste, Italy

Received on July 19, 2000; revised and accepted on August 16, 2000

Abstract

Summary: *A simple heuristic scoring method is described for assigning sequences to known domain types based on BLAST search outputs. The scoring is based on the score distribution of the known domain groups determined from a database versus database comparison and is directly applicable to BLAST output processing.*

Availability: *The scoring system will be incorporated into the server www.icgeb.trieste.it/sbase/*

Contact: *murvai@icgeb.trieste.it; pongor@icgeb.trieste.it*

Prediction of domains in new protein sequences is an important task. Current methods use pattern databases (for a review see, Atwood, 2000) which are based on multiple alignments represented in various forms. In a search for fast prediction methods, we previously found that the number of significant similarities found by BLAST between a query and the known members of a domain-group, (NSD) and the average similarity score of the hits (AVS) are efficient indicators of domain homologies (Hegyi and Pongor, 1993; Murvai *et al.*, 1999). Here we describe a simple probabilistic scoring system that can be used for the automated prediction of domain homologies.

The ease by which a sequence can be assigned to a given domain group will depend on: (i) how similar the constituent sequences are to each other (the '*tightness*' of the group); and (ii) how different they are from other sequences ('*separation*' from other groups). We carried out a database versus database comparison using the 1515 domain groups (79 478 sequences) of the SBASE 7.0 domain sequence library. For each domain sequence we counted the significant BLAST similarities it had with the domains of a give type (NSD), and computed an average similarity score (AVS) from the corresponding values, as described (Hegyi and Pongor, 1993; Murvai *et al.*, 1999). Figure 1a shows the AVS versus NSD plot of the ABC transporter domains (Higgins *et al.*, 1988). We estimate the probability of an NSD value being a positive hit, from the cumulative frequency distribution of the known NSD values (curve $P_{\mathrm{p}}^{\mathrm{NSD}}(\cdot)$, Figure 1d). The $P_{\mathrm{p}}^{\mathrm{NSD}}(\cdot)$ curve of a domain group can be simply calculated from the results of the database versus database comparison as the fraction of NSD values below a given value ($0 < P_{\mathrm{p}}^{\mathrm{NSD}}(\mathrm{NSD}) < 1$). In an analogous way, we can estimate the probability of an NSD value indicating that a domain is not a false positive $P_{\mathrm{nfp}}^{\mathrm{NSD}}(\mathrm{NSD})$, from the NSD values of those sequences that do not contain the given domain but produce a significant BLAST hit with some of its members (curve $P_{\mathrm{nfp}}^{\mathrm{NSD}}(\cdot)$, Figure 1d). Similar *p*-curves ($P_{\mathrm{p}}^{\mathrm{AVS}}(\cdot)$, $P_{\mathrm{nfp}}^{\mathrm{AVS}}(\cdot)$) can be computed from the AVS values of the database versus database comparison. The empirical cumulative score C is then computed as follows:

$$C = TP_{\mathrm{NSD}} + P_{\mathrm{p}}^{\mathrm{NSD}}(\mathrm{NSD}), + P_{\mathrm{nfp}}^{\mathrm{NSD}}(\mathrm{NSD}), \\ + TP_{\mathrm{AVS}} + P_{\mathrm{p}}^{\mathrm{AVS}}(AVS) + P_{\mathrm{nfp}}^{\mathrm{AVS}}(AVS), \quad (1)$$

where TP_{NSD} and TP_{AVS} are premium values given if NSD and AVS are above the group threshold T_{NSD} and T_{AVS}; respectively. The values of both $TP_{\mathrm{T,NSD}}$ and $TP_{\mathrm{T,AVS}}$ were arbitrarily taken as 2, so the value of the cumulative score is between 0 and 8. The C-value can be calculated for each domain type present in a BLAST output, and the domain-type with the highest C-value is assigned to the query. As a starting point, we set the threshold values T_{NSD} and T_{AVS} to the lowest corresponding value within the domain group (which is equivalent to choosing the minimal C-value as a threshold). In this manner, no false positives and false negatives were found in 1365 out of the 1515 domain groups. In the remaining 150 groups An additional C-threshold was determined by optimizing the Matthews coefficient (Matthews, 1975). Table 1 shows the statistics for selected domain groups.

Summarizing, the C-score seems to be a simple and robust predictor of domain similarities that compares quite favourably with more complicated methods in terms of CPU time and update costs. As BLAST runs are routinely used in most sequencing projects, we think that this score can be of use in genome sequence evaluation.

Table 1.

Domain type	*C*-value[1]	Threshold			Protein sequence[2] versus domain sequence[3] database comparison			
		NSD	AVS	*C*	*N*	True positives	False positives	False negatives
Fibronectin type III	7.22 ± 0.32 (6.40–7.91)	5	54.05	6.65	350	346	0	4
EGF-like, laminin-type	7.91 ± 0.21 (7.02–8.00)	242	67.23	7.02	47	47	0	0
Protein kinase	7.07 ± 0.52 (3.70–7.96)	13	54.00	5.47	1642	1641	0	1
Kringle domain	7.44 ± 0.32 (7.01–8.00)	172	69.69	7.01	61	61	0	0
Annexin-repeat	7.80 ± 0.38 (6.31–8.00)	196	67.38	6.31	70	70	0	0
Sushi-domain	7.38 ± 0.40 (6.11–7.99)	19	44.07	6.11[4]	164	164	3	0
Trypsin-like	7.21 ± 0.52 (6.03–7.98)	9	48.00	6.03[4]	443	443	3	0
ANK-repeat	7.28 ± 0.33 (6.28–7.89)	5	46.43	6.28[4]	157	157	4	0
Globin	7.06 ± 0.60 (3.72–7.86)	13	54.20	5.80[4]	866	860	4	6
ABC transporters	7.16 ± 0.54 (5.26–7.99)	174	57.38	5.26[4]	769	769	6	0
EGF-like	6.92 ± 0.85 (4.00–8.00)	52	35.00	5.70	469	444	21	25
Immunoglobulin	7.14 ± 0.43 (5.93–7.89)	4	41.90	6.30	2584	2530	61	54

[1]Given as average ± standard deviation (minimum–maximum). [2]Protein sequences were taken from a non-redundant protein sequence database formed from Swiss-Prot 37 (Bairoch and Apweiler, 2000) and PIR 59 (Barker *et al.*, 2000) and compared to SBASE-A. [3]Domain sequences were taken from the SBASE-A domain sequence library version 7 (Murvai *et al.*, 2000), the BLAST parameters are given under Figure 1. [4]Threshold values selected based on the Matthews correlation coefficient (Matthews, 1975). In the other cases *C* is the minimum value of the group.

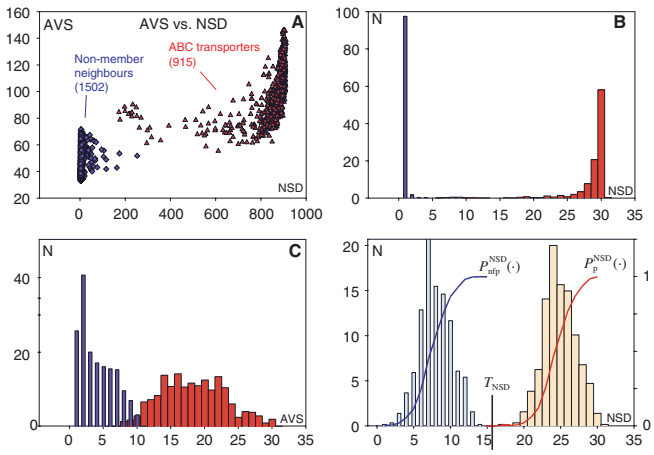

Fig. 1. (a) Average blast similarity score (AVS) versus number of significant BLAST similarities (NSD) plot for ABC transporter domains (red) from SBASE 7.0 (Murvai *et al.*, 2000). Each dot represents a domain sequence, and the AVS and NSD values are computed from the BLAST similarities between the sequence and other members of the group. Blue signs represent domain sequences that are not ABC transporters but still produce a significant BLAST similarity score with some of the ABC transporters. The database versus database comparison was carried out with BLAST (Altschul *et al.*, 1990), using a significance threshold of 0.8 and a minimum score of 32. (b) Histogram of NSD values. (c) Histogram of AVS values. (d) Schematic representation of a histogram (see text for details). T_{NSD} is a threshold value; if the NSD value of a given sequence is above this threshold, the sequence receives a threshold premium of 2, plus the respective P_p^{NSD}(NSD) value. A similar procedure is followed using the AVS values. *C* is calculated only if at least one of the threshold values is reached.

References

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

Atwood,T.K. (2000) The role of pattern databases in sequence analysis. *Briefings in Bioinformatics*, **1**, 45–59.

Attwood,T.K., Croning,M.D., Flower,D.R., Lewis,A.P., Mabey,J.E., Scordis,P., Selley,J.N. and Wright,W. (2000) Prints-S: the database formerly known as prints. *Nucleic Acids Res.*, **28**, 225–227.

Bairoch,A. and Apweiler,R. (2000) The Swiss-Prot protein sequence database and its supplement trembl in 2000. *Nucleic Acids Res.*, **28**, 45–48.

Barker,W.C., Garavelli,J.S., Huang,H., McGarvey,P.B., Orcutt,B.C., Srinivasarao,G.Y., Xiao,C., Yeh,L.S., Ledley,R.S., Janda,J.F., Pfeiffer,F., Mewes,H.W., Tsugita,A. and Wu,C. (2000) The protein information resource (Pir). *Nucleic Acids Res.*, **28**, 41–44.

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

Higgins,C.F., Gallagher,M.P., Mimmack,M.L. and Pearce,S.R. (1988) A family of closely related Atp-binding subunits from prokaryotic and eukaryotic cells. *Bioessays*, **8**, 111–116.

Matthews,B.W. (1975) Comparison of the predicted and observed secondary structure of T4 phage lysozyme. *Biochim. Biophys. Acta*, **405**, 442–451.

Murvai,J., Vlahovicek,K., Barta,E., Parthasarathy,S., Hegyi,H., Pfeiffer,F. and Pongor,S. (1999) The domain-server: direct prediction of protein domain-homologies from blast search. *Bioinformatics*, **15**, 343–344.

Murvai,J., Vlahovicek,K., Barta,E., Cataletto,B. and Pongor,S. (2000) The Sbase protein domain library, release 7.0: a collection of annotated protein sequence segments. *Nucleic Acids Res.*, **28**, 260–2.