CABIOS APPLICATIONS NOTE

Vol. 10 no. 4 1994
Pages 453–454

Fast databank searching with a reduced amino-acid alphabet

Claudine Landès[1] and Jean-Loup Risler

Abstract

Fast sequence databanks search algorithms generally make use of hash tables and look for exactly matching words. An increased sensitivity—at the expense of a decreased selectivity—can be attained in the case of proteins by using a reduced amino acid alphabet. We propose here an alphabet reduced to 10 symbols, that we used in modified versions of the FASTP and SCAN programs. An application to the aminoacyl-tRNA synthetases shows that this technique may be useful in detecting distant relationships between proteins.

Introduction

The need for efficient algorithms aimed at rapidly scanning protein or nucleic acid databanks has long been recognized. For example, the algorithm of Wilbur and Lipman (1983), published 10 years ago, was quickly incorporated in the program WORDSEARCH of the GCG package (Devereux *et al.*, 1984). The well-known FASTP and FASTA programs appeared soon after (Lipman and Pearson, 1985; Pearson and Lipman, 1988). In each case, the speed of the search is due to the fact that only *exact matching* is looked for, that is, a look-up table (Dumas and Ninio, 1982) containing the positions of each word of length *k* in the query sequence is generated first, the process being repeated for each sequence in the databank. In the SCAN program of the PSQ package (Dayhoff *et al.*, 1983) or in the PRELATE system (Collins and Coulson, 1987), a list of all the words of a given length occuring in the whole databank together with a pointer to their position, is pre-constructed. In order to improve sensitivity without impairing efficiency, the more recent BLAST program (Altschul *et al.*, 1990) generates neighbourhood words from the query sequence.

From their very nature, these algorithms present the drawback that they look for exactly matching words. Hence they may lack sensitivity. However, a compromise between rapidity and sensitivity can be attained by using a *reduced amino-acid alphabet*, that is, considering similar amino acids as strictly identical.

Introduction of a reduced amino-acid alphabet in SCAN and FASTP

The problem is now to define such a reduced alphabet. From the consideration of the chemical properties of amino-acid side-chains, Dayhoff *et al.* (1972) and Miyata *et al.* (1979) came to the optimistic conclusion that the amino-acid alphabet could be reduced to six letters. A more rigorous study of the PAM250 matrix led Collins and Coulson (1987) to conclude that 15 symbols was a minimum. On the basis of a quantitative comparison of six different similarity matrices (Risler *et al.*, 1988), we propose here a minimal 'consensus' reduced alphabet, which lies midway between the extremes cited above. This alphabet is reduced to 10 symbols as follows: (A=S=T), C, (D=E=N), (F=Y), G, H, (I=L=M=V), (K=Q=R), P, W. These groupings are close to those proposed by Santibanez and Rohde (1987), the differences being their grouping A with (I, L, M, V) and Q with (D, E, N). The five amino acids that are left alone have been shown to be barely exchangeable in proteins sharing homologous structures (Risler et al., 1988).

The usefulness of this reduced alphabet was checked by incorporating it in modified versions of the SCAN and FASTP programs. These modifications were straightforward. In SCAN, the amino acids that belonged to the same group were given the same numerical value in the hash code table. In this way, two tripeptides differing by the replacement of F by Y, for example, receive the same hash value. In FASTP, the modification consisted simply in rewriting the sequences in the new alphabet. In both cases, it is clear that the use of the reduced alphabet increases the background noise (Collins and Coulson, 1987) and, therefore, the query sequences should not be too long. Thus the modified FASTP program should be effective in searching for peptides rather than for entire protein sequences (note that for short query sequences, FASTP and FASTA should give essentially similar results).

Applications

The modified SCAN program is useful for rapidly checking potential 'consensus' sequences. For example, the tetrapeptide HIGH (and its variations) is found in the so-called Class I aminoacyl-tRNA synthetases (aaRS). When used with the reduced alphabet, SCAN shows that

Centre de Génétique Moléculaire du CNRS 91198 Gif sur Yvette Cedex France

[1] *To whom reprint requests should be sent*

Table I. Results of databank searches using the original, unmodified FASTP program (normal alphabet), the modified FASTP version (reduced alphabet), FASTA and BLAST. The search with BLAST was performed on the NCBI server. Databanks used: MIPS protein release 32 for FASTP and FASTA, non-redundant cumulative protein database at NCBI for BLAST. The query sequence was a peptide of 49 residues belonging to arginyl-tRNA synthetase (ArgRS) from *E. coli*. The figure shows which aminoacyl-tRNA synthetases have been found in the top 200 scores, together with their rank.

FASTP reduced alphabet	FASTP normal alphabet	FASTA	BLAST
1 ArgRS—*E.coli*	1 ArgRS—*E.coli*	1 ArgRS—*E.coli*	1 ArgRS—*E.coli*
2 LeuRS—*N.crassa* mito.	3 LeuRS—*N.crassa* mito.	3 LeuRS—*N.crassa* mito.	2 LeuRS—*N crassa* mito.
2 LeuRS—*E.coli*	5 LeuRS—*E.coli*	7 ValRS—*S. cerevisiae*	3 LeuRS—*E.coli*
4 CysRS—*E.coli*	7 CysRS—*E.coli*	8 LeuRS—*E.coli*	5 CysRS—*E.coli*
7 LeuRS—*S.cerevi.* mito.	9 LeuRS—*S.cerevi.* mito.	10 CysRS—*E.coli*	6 ValRS—*S.cerevi.* mito.
10 MetRS—*E.coli*	12 ValRS—*S.cerevisiae*	12 LeuRS—*S.cerevi* mito.	
10 ValRS—*B.stearother.*			
11 ValRS—*S.cerevisiae*			
13 ValRS—*N.crassa*			
15 MetRS—*B.stearother.*			

the peptides HIGH, HLGH, HMGH and HVGH are indeed found in the synthetases, but also in a number of unrelated proteins. Hence this peptide is not a signature sequence specific of the aaRS. On the contrary, the peptide GDSGGP is considered as highly specific of the serine proteases. A search in the MIPS databank resulted in 196 hits: 194 were GDSGGP in serine proteases, 1 was GESGGP-conservative substitution of D by E- in a serine proteinase and 1 was GEAGGP in a non-related protein lacking the active series. We have also compared the results given by the unmodified and modified versions of FASTP by searching the MIPS databank against a peptide of 49 residues belonging to arginyl-tRNA synthetase (ArgRS). This peptide (from ^{115}Ile to ^{163}Gly) is situated at the beginning of the nucleotide binding domain and encompasses the HVGH peptide. As shown in Table I, the unmodified FASTP detected six aaRS while the modified version found 10 aaRS, including the MetRS that ranked tenth and was not found before. The modified FASTP, as expected, is more sensitive. Note also that, as shown in Table I, neither FASTA nor BLAST detected the similarity with MetRS. Incidentally, this result suggests that ArgRS is a member of the Cys-, Ile-, Leu-, Met-, ValRS subfamily (see Landès *et al.*, 1992).

Acknowledgements

This work was funded by grants from the CNRS and from CM2AO. One of us (C.L.) is recipient of a fellowship from Biocem, Orsan, Rhône-Poulenc and Roussel-Uclaf.

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.

Collins,J.F. and Coulson,A.F.W. (1987) Molecular sequence comparison and alignment. In Bishop,M.J. and Rawlings,C.J. (eds), *Nucleic Acid and Protein Sequence Analysis*. IRL Press, Oxford, pp. 323–358.

Dayhoff,M.O., Eck,R.V. and Park,C.M. (1972) A model of evolutionary change in proteins. In Dayhoff,M.O. (ed), *Atlas of Protein Sequence and Structure*, **5**, 89–99.

Dayhoff,M.O., Barker,W.C. and Hunt,L.T. (1983) Establishing homologies in protein sequences. *Methods Enzymol.*, **91**, 524–545.

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

Landès,C., Hénaut,A. and Risler,J.L. (1992) A comparison of several similarity indices used in the classification of protein sequences. *Nucleic Acids Res.* **20**, 3631–3637.

Lipman,D.J. and Pearson,W.R. (1985) Rapid and sensitive protein similarity searches. *Science*, **227**, 1435–1440.

Miyata,T., Miyazawa,S. and Yasunaga,T. (1979) Two types of amino acid substitutions in protein evolution. *J. Mol. Evol.*, **12**, 219–236.

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

Risler,J.L., Delorme,M.O., Delacroix,H. and Henaut,A. (1988) Amino acid substitutions in structurally related proteins. *J. Mol. Biol.*, **204**, 1019–1029.

Santibanez,M. and Rohde,K. (1987) A multiple alignment program for protein sequences. *Comput. Applic. Biosci.*, **3**, 111–114.

Wilbur,W.J. and Lipman,D.J. (1983) Rapid similarity searches of nucleic acid and protein data banks. *Proc. Natl. Acad. Sci. USA.*, **80**, 726–730.