DATABASE OF PATTERNS PROF_PAT, USED TO DETECT LOCAL SIMILARITIES

Bachinsky A.G., Grigorovich D.A., Naumochkin A.N., Nizolenko L.Ph., Yarigin A.A.

1Research Institute of Molecular Biology, Koltsovo, Russia
e-mail: bachin@vector.nsc.ru
2Institute of Cytology and Genetics SB RAS, Novosibirsk, Russia
e-mail: odip@bionet.nsc.ru
*Corresponding author

Keywords: protein families, patterns, motifs, similarity search, database

Resume

Motivation:
When analysing novel protein sequences, it is now essential to extend search strategies to include a range of 'secondary' databases. Pattern databases have become vital tools for identifying distant relationships in sequences, and hence for predicting protein function and structure. The main drawback of such methods is the relatively small representation of proteins in trial samples at the time of their construction. Therefore a negative result of an amino acid sequence comparison with such a databank forces a researcher to search for similarities in the original protein banks. We developed a database of patterns constructed for groups of related proteins with maximum representation of amino acid sequences of SWISS-PROT in the groups.

Results:
Software tools and a new method have been designed to construct patterns of protein families. By using such method, a databank of protein family patterns, PROF_PAT, is produced. This bank is based on SWISS-PROT (rl.38) and TrEMBL (rl.11), and contains patterns of more than 14,000 groups of related proteins in a format similar to that of the PROSITE. Motifs of patterns, which had the minimum level of probability to be found in random sequences, were selected. Flexible fast search program accompanies the bank. The researcher can specify a similarity matrix (the type PAM (PAM, BLOSUM and other). Variable levels of similarity can be set (permitting search strategies ranging from exact matches to increasing levels of "fuzziness").

Availability:
The Internet address for comparing sequences with the bank is: http://wwwmgs.bionet.nsc.ru/mgs/programs/prof_pat/. The local version of the bank and search programs (approximately 50 Mb) is available via ftp: ftp://ftp.bionet.nsc.ru/pub/biology/vector/prof_pat/, and ftp://ftp.ebi.ac.uk/pub/databases/prof_pat/. Another appropriate way for its external use is to mail amino acid sequences to bachin@vector.nsc.ru for comparison with PROF_PAT 1.3.

Introduction
Up to now, the main method of suggesting possible functions of the newly deciphered amino acid sequences has been to search them for similarity with sequences available in protein banks such as PIR (Barker et al., 1999), SWISS-PROT (Bairoch and Apweiler, 1999) and others. As these banks grow larger, such comparisons become more promising but at the same time more time-consuming. In addition, in the case of distant proteins the search for global similarity of complete sequences may fail to show a positive result, because the conservative blocks responsible for their special functions may prove to be relatively short and scattered all over the sequence. This may be why a number of works appeared in the last few years, aimed at the selection of sites in groups of related proteins. These sites are representative of a protein family as a whole, and both identify new proteins and refine structural and functional properties of those already known. Such databases as PROSITE (Hofmann, et al., 1999), BLOCKS (Henikoff and Henikoff, 1991, Henikoff, et al., 1999), PRINTS (Attwood et al., 1999) are among the most well known and accessible via Internet. There is also a number of other similar databases i.e. PFAM (Bateman, et al., 1999), SBASE (Murvai, et al., 1999), IDENTIFY (Nevill-Manning, et al. 1998).

We have devoted our efforts to develop a technique and construct patterns for the greatest possible number of proteins belonging to the SWISS-PROT+TrEMBL (Bachinsky et al., 1996, 1997). We are convinced that if a secondary bank is not really representative, it would not be widely used. It is because of negative results in the comparison of a sequence with this bank force the user to consult other banks or make direct comparisons of the sequence with large banks of sequences.
Methods and algorithms

The selection and concurrent alignment of related protein groups

All full-length sequences of prototype banks that had more than 30 amino acids in lengths were combined in one file. In order to select groups of related proteins, a special program was written based on FASTA 2.0 (Pearson, 1994). The sequences similar in the sense of FASTA form a primary set of related proteins. Pairwise similarity of the proteins belonging to a set was assessed by the program CLUSTALV (Higgins et al., 1992). Then, if not all pairs of proteins had 30% similarity, the set was divided into subsets so that all pairwise similarities were at least 30%. Thus, more than 14,000 subsets or groups were obtained, containing more than 100,000 sequences.

Proteins of every subset were aligned together. The files containing aligned sequences were supplemented with two fields: DE - description(s) of proteins forming the group, and KW - key words (mainly the union of values of field KW for proteins falling into the set). Patterns were constructed based on such aligned families.

The construction of patterns of protein families. We will regard the combination of motifs that represent relatively conservative intervals of positions of aligned proteins of the family as a pattern of a family of related proteins. The motifs of patterns are represented by ambiguous words of the type:


Thus, an initial pattern of a protein family is an ordered combination of non-overlapping motifs of the type: $rA_1A_2A_3...A_n$. Here $r$ is position number of an aligned group of proteins (the trial sample), where the motif begins, $A_j$ is a set of amino acids, located in $r+i-1$ position of the trial sample. For a passive position $A_j=X$: any amino acid is acceptable.

Comparison of amino acid sequences with patterns

The searches for exact matching between amino acid sequences' fragments and pattern motifs

The main algorithm for comparing an amino acid sequence with the pattern database uses the modification of finite automaton of Aho-Corasic (Aho and Corasic, 1975), constructed based on a set of samples, which are to be searched for in the input text. The automaton is presented as an oriented tree-like graph, where nodes are states of the automaton and arcs are admissible transitions from some states to the others, marked with symbols from the alphabet $S$ of the amino acids' designation. The automaton works in cycles. In every cycle one more symbol of a text is read, which determines the automaton's transition from the current state into a new one. The automaton's behaviour is characterised by three functions: function of transitions $G(s,a)$; rejections' function $F(s)$ and output function $O(s)$. The values of these functions are calculated once when constructing the automaton based on a given set of samples. In Fig. 1. the functions of the automaton constructed on the set of samples $R=\{r_1, r_2, r_3, r_4, r_5\} = \{HE, SHE, HIS, HER, HERS\}$. are illustrated.

When constructing the automaton in every motif, 4 neighbouring positions are chosen (the core of a motif), having minimum value of the product $P_j$ and containing no passive positions. Then this core is converted into exactly determined words of length 4 that act as samples in constructing automaton. If coincidence of a current fragment of an input sequence and one of the automaton samples is observed, comparison is performed (up to the first non coincidence) of all the other motif positions from the list of the output function, and the corresponding fragments of the sequence (the stage of extending the core). According to the results of this stage, the final decision is made on whether there is similarity or not.
The search for distant similarity

To reveal a distant similarity, the algorithm of comparison is modified. The user specifies the matrix of similarity of amino acid residues (e.g., using the one from families PAM, BLOSUM, etc.) and $D$ - the level of similarity within the limits of motif. For all states of the automaton, the function of rejection is set to zero. Besides, a sequence as a whole does not input to the automaton, but specially processed words.

The comparison of patterns with the parent banks

To examine the recognising ability of the patterns and exclude certain motifs, which are non-specific for a given family, all patterns were compared with all the proteins of the SWISS-PROT+TrEMBL. In the routine comparison between patterns and the banks, only exact similarities of two or more motifs per a pattern were registered, i.e. the cases when fragments of amino acid sequences belong to the motifs. The similarity is regarded as 'positive' one, if at least one of the two following conditions is met. 1. Query sequence belongs to the trial sample. 2. All words of one of the DE fields of the pattern (the names of the proteins forming the family) are present in the field DE (protein name) of the sequence. The similarity is considered 'conditionally positive' (UNKNOWN), if at least one of the DE or KW words of the pattern coincides with one of the words determined in fields DE and/or KW of the sequence. Thus, proteins are defined as conditionally related if they possess some common function (e.g., hydrolases, dehydrogenases, oxidoreductases, etc.) or some specific features of their structure (for instance, transmembrane segments). All other cases of similarity are regarded as false positive. As a result of comparison with the bank, a pattern bank entry is created, similar in its structure to entries of PROSITE.

Implementation and results

In version 1.3. presented here, the total number of motifs in more than 13,000 patterns is over 200,000 with specificities varying from one expected false positive prediction in $10^8$ tests and higher. The total combined length of patterns is about 2,000,000 positions.

To find a distant similarity, a very fast flexible comparison procedure is employed, using the modified algorithm of Aho-Corasic (Aho and Corasic, 1975), various matrices of similarity/distance for amino acid residues, the predetermined grade of similarity between a fragment of an amino acid sequence and a pattern motif. Patterns identify nearly 130 thousand of amino acid sequences of SWISS-PROT+TrEMBL as having shown 'positive' or 'conditionally-positive' similarity. In the latter case, the similar sequences, not included into the trial samples, are usually identified.

Almost all sequences of the trial samples are recognised by all motifs of the corresponding patterns. Certain violations of this rule are due only to the presence of non-standard symbols in the particular sequences of the trial samples that have fallen into the intervals of positions represented by pattern motifs.

A number of cases of false-positive similarity may be divided into two classes. Sometimes it is a really chance similarity. However, sometimes two or more pattern motifs show similarity to the fragments of a certain sequence; the order of the fragments' locations often correspond to that of the motifs' locations, which increases even more the certainty that the similarity is not random. In most cases, false-positive similarity is revealed with sequences described only as products of some genes, and this information is not included into descriptions of patterns.

All patterns were searched in 1480 new sequences of TrEMBL more recent version described as ORFs. 419 were undoubtedly identified. Some other comparisons of the PROF_PAT and other secondary banks show that PROF_PAT exceeds the most popular banks PROSITE, PRINTS, and BLOCKS under such index as number of patterns and motifs. The more pattern motifs show similarity to the sites of a query sequence, the higher would be the likelihood that the amino acid sequence is related to proteins of the trial sample (Henikoff and Henikoff, 1991), especially if the motifs' order coincides with that of the sample proteins.

Thus, we have constructed a bank of patterns for protein families, representing about 2/3 of the full-length protein sequences of the bank SWISS-PROT, release 38 and TrEMBL, release 11. The fast flexible search program for close and distant similarity provides comparisons of amino acid sequences of interest with the bank of patterns in the interactive mode. The PROF_PAT technology update has been developed and tested, so the new versions of PROF_PAT will be created following each new versions of SWISS-PROT+TrEMBL.

References