Automatic Procedure to Extract Signature Pentapeptides from the Protein Sequence Database

Ikuo Uchiyama[1] , Atsushi Ogiwara[1], Zenmei Ohkubo[1]
Minoru Kanehisa[1]

Institute for Chemical Research, Kyoto University,
Uji, Kyoto 611, Japan
E-mail: uchiyama@kuicr.kyoto-u.ac.jp

Abstract

A method is described for extracting signature pentapeptides that are conserved and exclusively found in a group of homologous proteins. The BLAST algorithm is used to count the frequency of occurrences of pentapeptide patterns allowing limited substitutions, as well as to perform homology search. For those pentapeptides that appear in a given sequence we examine the frequency of occurrences of these pentapeptides and related ones in homologous sequences which are ordered according to the homology score. By comparing against the frequency in the entire database, we can extract uniquely conserved pentapeptides and at the same time perform a grouping of homologous sequences. Thus, our procedure can automatically identify, if any, pentapeptides that are strongly tied with the group. Possibility of using our pentapeptide word dictionary to infer protein function is discussed.

[1]内山郁夫、荻原淳、大久保善明、金久實　京都大学化学研究所　京都府宇治市五ケ庄

1 Introduction

Motif libraries are becoming an increasingly useful resource to assign biological functions to newly determined protein sequences. A motif can be defined as a sequence pattern which is strongly related to protein function so that it is conserved within a functional group of proteins and, in general, uniquely found in the group. Previously, we extracted uniquely conserved peptide words in each superfamily defined in the NBRF/PIR database, using frequency tables of short oligopeptides in the entire database [1]. Once a functional classification of protein sequences is given, this method can be used to extract motifs automatically. However, because of the rapid increase of the number of sequences in the database, classification of sequences itself becomes a serious problem. It is desirable that sequence classification and motif extraction are done simultaneously.

One of the simplest and fastest methods to classify sequences is to use homology search, where similar sequences distributed around a given sequence are combined to form a related group. However, it is difficult to judge whether one sequence belongs to the group or not, when its similarity score is not so high. In fact, in such a case motifs can be used to discriminate the noise level. If a sequence entry identified by the homology search has a pattern that is strongly conserved in the group, it is highly likely that the sequence belongs to this group. Therefore, entries whose homology scores are high enough and which share a common pattern can be considered as members of one group.

In this paper, we focus on, what we call, signature pentapeptide patterns which are uniquely conserved within a group of homologous proteins. There are a million possible pentapeptide patterns, and it is known that the same pentapeptide patterns can take entirely different 3-D structures [2]. However, the actual proteins do not use all possible pentapeptide patterns uniformly; there is a strong dependence on the evolutionary history. Thus, given the constraint of global homology we expect to see signature pentapeptides which may be able to characterize specific functions.

To automatically extract signature pentapeptides from the protein sequence database, we use the BLAST algorithm [3]. BLAST is one of the best known programs for homology search. It is especially suited at detecting locally conserved similarity patterns. We used this algorithm not only to perform homology search but also to count the frequency of occurrences

of each pentapeptide and to select signature pentapeptides that are unique and well conserved in entries whose homology scores are high enough.

2 Methods

BLAST initially extracts all W-mers from a query sequence, and next enumerates neighborhood words which have similarity scores of at least T when aligned with each of the W-mers from the query sequence. Using this word list, BLAST constructs a deterministic finite automaton and rapidly searches the database for these words. Once a word in the list is found, BLAST extends it by calculating the similarity score to seek for a maximal-scoring segment pair, in which the score is greater than a given threshold S, or the Poisson probability is less than a given threshold P.

Our purpose is to extract pentapeptides from a query sequence, which are uniquely conserved in the homology group consisting of entries found by the homology search. We choose the word size parameter W = 5 and using the BLAST algorithm, search for pentapeptide words which have similarity scores grater than or equal to T. Every hit is counted and later uniqueness and conservation are evaluated. In our previous paper, uniqueness U and conservation C are defined by the following equations:

$$U = \frac{Np}{Nt} \tag{1}$$

$$C = \frac{Np}{Ng} \tag{2}$$

where Np and Nt are the numbers of entries containing a given pattern p in a given group g and in the entire database, respectively, and Ng is the number of entries belonging to the group g. Here, since we allow each pattern to contain substitutions to a limited extent, we count the occurrences of a set of patterns P distributed around the pattern p, instead of the occurrence of the pattern p. Moreover, we define the group g as a set of entries which give homology score higher than threshold S. Therefore, to describe parameter dependency explicitly, above equations are rewritten as the following:

$$U = \frac{Np(P(\mathrm{T}), g(\mathrm{S}))}{Nt(P(\mathrm{T}))} \tag{3}$$

$$C = \frac{Np(P(\mathrm{T}), g(\mathrm{S}))}{Ng(g(\mathrm{S}))}. \tag{4}$$

Our purpose is to find out both a pattern set P and a group g, which give U and C values exceeding some thresholds with a suitable choice of parameters T and S. Since decreasing T or S increases the possibility to pick up noise, it can be thought that Np, Ng and Nt should be weighted by T and/or S. But in this work, we simply assumed no weights and summed up the frequencies.

The word score threshold T_0 and cutoff score S_0 used for homology search determine the minimal values of T and S, respectively. Here, $\mathrm{T}_0 = 24$ and $\mathrm{S}_0 = 45$ were used for homology search. For the similarity matrix, we used BLOSUM62 matrix [4] contained in the BLAST package. Calculation was done by CRAY YMP-2 in the Supercomputer Laboratory of the Institute for Chemical Research, Kyoto University.

3 Results

Figure 1 shows a sample output of our procedure when the SWISS-PROT database release 25.0 was searched with the sequence of human cytochrome b5 (SWISS-PROT code: CYB5_HUMAN) as a query. Each line represents an entry sorted by the homology score, and each column represents the occurrence of a pentapeptide word contained in the query sequence aligned in the order of N-terminus to C-terminus. An asterisk shows that the sequence has a pattern exactly matched with the pattern in the query, a dot shows that the sequence doesn't have the pattern, and a digit shows the difference of the word score from T_0 when the sequence has a similar pattern with the query sequence. For example, since we chose here $\mathrm{T}_0 = 24$, '4' means the sequence has a pentapeptide whose similarity score is 28. Note that the query sequence itself (top of the lines) contains patterns not matched since a pentapeptide whose self similarity score is less than T_0 is excluded at the first step of the BLAST algorithm. Seeing this figure, one can extract conserved patterns and construct a maximal group which contains the largest number of members. Here, cyotochrome b5, nitrate reductase, and sulfite oxidase are compiled within one group, in which the best conserved pentapeptides are EHPGG or HPGGE. This result is consistent well with (but slightly different

from) PROSITE [5], in which the consensus pattern is defined as 'F-[LIV]-x(2)-H-P-G-G' where H of position 5 is heme iron binding site. In our case, the pattern is defined as [EKSDN]-H-P-G-G or H-P-G-G-[EASPQV]. The 2 residues denoted by 'x' in PROSITE are so variable, then our procedure failed to detect the preceding region. The group defined here is the same as that of PROSITE 10.1 except that the additional entries NIA_CHLKE and ACO1_YEAST are included in our group. The former is missed by PROSITE 10.1 because this sequence is newly added to SWISS-PROT since release 25.0. The latter sequence is stearoyl-CoA desaturase of yeast, which has additional C-terminal region when compared with other stearoyl-CoA desaturase sequences. This region has weak similarity with cytochrome b5 domain including HPGG signature sequence. Thus, we think this region potentially came from cytochrome b5, although the authors of the original paper didn't mention it [6]. This sequence is missed by PROSITE because the first position of the pattern in PROSITE is substituted from F to Y.

Table 1 shows the number of entries ($Np(P(\mathrm{T}), g(\mathrm{S}))$) containing the pattern EHPGG when various values of homology score cutoff S and word score threshold T are used. Decreasing S increases the number of homologous sequences, and decreasing T increases the number of substitute patterns, thus increasing the number of entries containing these patterns. The right most column shows the group size ($Ng(g(\mathrm{S}))$), which is the total number of entries having scores greater than or equal to cutoff S, and the last line shows the total number of entries containing these patterns in the entire database ($Nt(P(\mathrm{T}))$). Using this table we can easily calculate uniqueness or conservation defined by the equations (3) and (4). In this case, when S = 80 and T = 24 are used, we can construct the largest group where the pentapeptide EHPGG are completely conserved ($C = 1$), as already shown in Figure 1. However, this pentapeptide cannot be unique in this group because another 13 entries have one of the same patterns in this pattern set, so that uniqueness is not so high ($U = 0.66$).

4 Discussion

Although we show only one example here, we plan to repeatedly apply our procedure to the entire database in order to automatically classify sequences of the database and extract informative pentapeptides strongly tied with

the groups. The advantage of automating this process is clear, since newly determined sequences are rapidly increasing, and more frequent updates of motif library will be required. The BLAST algorithm is rapid enough for this purpose. Moreover, an automatic procedure without any prior knowledge may detect new motifs which are shared in several families, or potential functional sites.

Our procedure fails to construct any group when no pentapeptide signatures are found. As shown in the previous section, to satisfy both conditions of uniqueness and conservation is quite difficult. When we allow more substitutions to satisfy conservation, in not a few cases these patterns fail to satisfy uniqueness. One might suspect that a pentapeptide word has too little information to determine the function of the sequence. From the information theoretical point of view, the BLOSUM62 score 25 used in this work is identical to 11.5 bits of information. This information is obviously too little to distinguish a true hit from a random hit. According to Altschul [7], searching a database containing 4,000,000 residues with a query sequence of 250 residues requires 30 bits information to discriminate from the noise level. However, an actual sequence is not so random especially around a functionally important site. Since we are screening pentapeptides within sets of homologous sequences which are potentially related, less information may be sufficient to discriminate from the noise.

Of course this information needs not be confined within one pentapeptide. In fact, many patterns defined in PROSITE are fairly long, including 'x' (any amino acid) characters. Since protein function depends on the 3-D structural context, amino acid residues which are apart in the sequence can be cooperatively related. Therefore, longer peptide patterns or other descriptions of patterns should be considered in order to define motifs more widely. In principle, our procedure can be expanded to use a longer word size W, but it requires enormous memory. Another way is to define a motif as a combination of patterns uniquely conserved in a given group. In this case, one pattern needs not fulfill the criterion of uniqueness.

In spite of these improvements to be made, it is still interesting to see how many signature pentapeptides can be found in the current database according to our present procedure. We will also apply our method to sequences in the Protein Data Bank by taking each of them as a query in order to examine structural features of or structural relationships between signature pentapeptides.

5 Acknowledgements

This work was supported by a grant-in-aid for scientific research on the priority area 'Genome Informatics' from the Ministry of Education, Science and Culture, Japan. The computation time was provided by the Supercomputer Laboratory, the Institute for Chemical Research, Kyoto University.

References

[1] Ogiwara, A., Uchiyama, I., Seto, Y. & Kanehisa, M. (1992). "Construction of a dictionary of sequence motifs that characterize groups of related proteins". *Protein Engineering* **5**, 479-488.

[2] Kabsch, W. & Sander, C. (1984). "On the use of sequence homologies to predict protein structure: identical pentapeptides can have completely different conformations". *Proc. Natl. Acad. Sci. USA* **81**, 1075-1078.

[3] Altschul, S. F., Gish, W., Miller. W., Myers. E. W. & Lipman, D. J. (1990). "Basic local alignment search tool". *J. Mol. Biol.* **215**, 403-410.

[4] Henikoff, S. & Henikoff, J. G. (1992). "Amino acid substitution matrices from protein blocks". *Proc. Natl. Acad. Sci. USA* **89**, 10915-10919.

[5] Bairoch, A. (1992). "PROSITE: a dictionary of sites and patterns in proteins". *Nucleic. Acids. Res.* **20**, 2013-2018.

[6] Stukey, J. E., McDonough, V. M. & Martin, C. E. (1990). "The OLE1 gene of *Saccharomyces cerevisiae* encodes the $\Delta 9$ fatty acid desaturase and can be functionally replaced by the rat stearoyl-CoA desaturase gene". *J. Biol. Chem.* **265**, 20144-20149.

[7] Altschul, S. A. (1991). "Amino acid substitution matrices from an information theoretic perspective". *J. Mol. Biol.* **219**, 555-565.

Figure 1: Distribution of the pentapeptides in the homologous sequences of human cytochrome b5 (CYB5_HUMAN)

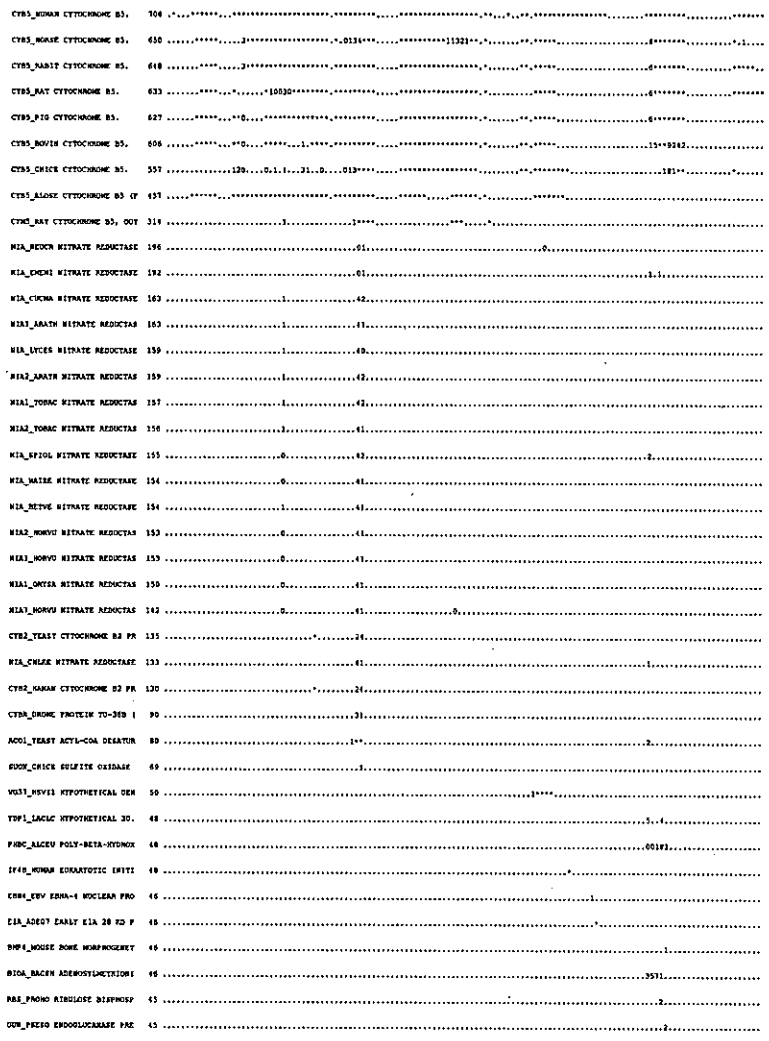

Table 1: Count of the occurrence of the pattern "EHPGG"

S \ T	24	25	26	27	28	29	30	EXACT	Ng
708	1	1	1	1	1	1	1	1	1
650	2	2	2	2	2	1	1	1	2
648	3	3	3	3	3	2	2	2	3
633	4	4	4	4	4	3	3	3	4
627	5	5	5	5	5	4	4	4	5
606	6	6	6	6	6	5	5	5	6
557	7	7	7	7	7	6	6	6	7
457	8	8	8	8	8	7	7	7	8
314	9	9	9	9	9	8	8	8	9
163	11	11	11	11	11	8	8	8	11
159	13	13	13	13	13	8	8	8	13
157	14	14	14	14	14	8	8	8	14
156	15	15	15	15	15	8	8	8	15
155	16	16	16	16	16	8	8	8	16
154	18	18	18	18	18	8	8	8	18
153	20	20	20	20	20	8	8	8	20
150	21	21	21	21	21	8	8	8	21
142	22	22	22	22	22	8	8	8	22
135	23	23	23	22	22	8	8	8	23
133	24	24	24	23	23	8	8	8	24
130	25	25	25	23	23	8	8	8	25
125	26	25	25	23	23	8	8	8	26
121	27	25	25	23	23	8	8	8	27
90	28	26	26	24	23	8	8	8	28
80	29	27	27	25	24	9	9	9	29
69	29	27	27	25	24	9	9	9	30
50	29	27	27	25	24	9	9	9	31
48	29	27	27	25	24	9	9	9	34
47	29	27	27	25	24	9	9	9	35
46	29	27	27	25	24	9	9	9	39
45	29	27	27	25	24	9	9	9	42
Nt	44	37	37	34	30	11	11	11	