# Detecting common amino acid sequence patterns with a gap

Mikita Suyama Takaaki Nishioka Jun'ichi Oda
suyama@kuicr.kyoto-u.ac.jp nishioka@kuicr.kyoto-u.ac.jp

Institute for Chemical Research, Kyoto University
Uji, Kyoto 611, Japan

**Abstract**

*We have developed a program GAPE (Gap Allowing Pattern Explorer) to extract amino acid sequence motifs conserved among distantly related proteins. The GAPE program is designed to allow a gap in the sequences. When the program is applied to some ligand-related consensus sequences, motifs extracted with low expectation of occurrence contain some of the amino acid residues chemically proved to be involved in the ligand recognition.*

## 1 Introduction

As the number of amino acid sequences determined has rapidly increased, it has become clear that automated procedures to find motifs would be useful. A number of attempts to extract motifs automatically have been proposed. It was difficult, however, to deal with gaps in searching for the motifs because of combinatorial problems. Any methods so far developed practically extract local patterns with no gap. On the other hand, the real motifs such as those collected in the PROSITE database [1] often contain gaps.

In the present study we have developed a motif search algorithm that explicitly deals with gaps in a set of sequences. The utility of the program is shown by applying it to several ligand-related consensus sequences [2].

## 2 Methods

To find the sequence motifs conserved among distantly related sequences, we have developed a program, GAPE (Gap Allowing Pattern Explorer), that searches all the subsequences having any common 5-amino acid pattern allowing a gap. The GAPE program is an extension of the MOTIF program devised by Smith *et al.* [3]. Recently another program, ASSET, which is also an extension of the MOTIF program, have been devised by Neuwald and Green [4]. In both the MOTIF and the ASSET programs, gaps are not allowed within the patterns.

There are two main steps in motif extraction by GAPE. In the first step, a "branch-and-bound method" is applied to search for subsequences that contain a pattern in the order of the amino acids. In the next step, the subsequences are selected by the distances of the amino acids matched with the pattern. When all four distances between the five amino acids match with the patterns are the same in another subsequences, we define the pattern as "rigid motif". Motifs with a gap in the subsequences are also extracted by relaxing one of the four distances. We call these patterns "flexible motifs".

To evaluate statistical significance of the motifs, we devised a method to calculate the expected value of occurrence of each motif *a priori* from the amino acid composition of the sequences under consideration.

須山幹太、西岡孝明、小田順一： 京都大学化学研究所, 〒611 宇治市五ケ庄

# 3 Results and Discussion

## 3.1 Motifs in ligand-related proteins

There were 67 consensus sequences extracted from enzymes recognizing pyridoxal phosphate (PLP). When GAPE was applied to the consensus sequences, the numbers of rigid and flexible motifs with expected value $E < 0.05$ were 146 and 77, respectively. One of the flexible motifs is shown in Fig. 1. The lysine residue in the flexible motif is known to form Schiff base with the ligand.

```
Motif: S-(x0)-K-(x5,6)-G-(x1)-R-(x1)-G
E = 2.2457×10-02

2.6.1.1   02  2  243  x+xxxIx*NGx SKox-SMTGWR*G Yxx:xx=II!xM
2.6.1.5   02  1  246  AGLPALVSNSF SKIF-SLYGERVG GLSVMCEDAEAA
2.6.1.9   01  5  213  xypnlvxlrTx SKaf-glAGlRxG xxxaxxxxxxxx
4.4.1.14  01 16  277  nkdLvHIvySL SKD-mGlPGFRvG IiYS@NDxVVxc
2.6.1.2   01  3  312  xqqeLaSFhSv SKGYmGECGfRGG YvEvvnmdAxVq
                                  SK      G R G
```

Fig. 1. A motif extracted from PLP-related consensus sequences. The first and second columns represent the EC number and ID number of the consensus sequence for the enzyme. The third column represents the number of amino acid sequences to construct the consensus sequence. The number following the consensus sequence is the position of the first amino acid in the motif.

When GAPE was applied to 14 tetrahydrofolate-related consensus sequences, motifs containing the residues that participate directly in the recognition of the ligand were also extracted.

## 3.2 Biological meaning of the motifs

The motifs obtained from the ligand-related consensus sequences in the present study were experimentally found to be located in the ligand recognition sites of the enzymes. These motifs imply that the enzymes sharing the motif have been evolved from a common ancestor and diverged to specific reactions for each enzyme retaining the ligand specificity. The flexible motifs are found around the functionally important sites. The GAPE program is applicable to not only the consensus but also their composite amino acid sequences and useful to detect such patterns in distantly related sequences.

# Acknowledgement

The authors thank Ikuo Uchiyama and Atsushi Ogiwara for helpful discussions. Computation time was provided by Supercomputer Laboratory, Institute for Chemical Research, Kyoto University. This work is supported by a Grant-in-Aid for Scientific Research on Priority Areas, 'Genome Informatics', from the Ministry of Education, Science and Culture of Japan, and by a Fellowship of the Japan Society for the Promotion of Science for Japanese Junior Scientists to M. S.

# References

[1] Bairoch, A., "PROSITE: a dictionary of sites and patterns in proteins, its current status," *Nucleic Acids Res.*, **21**, 3097-3103, 1993.

[2] Suyama, M., Nishioka, T. & Oda, J., "Extraction of the ligand-related motifs in enzymes," in *Proceedings of Genome Informatics Workshop IV*, eds. Takagi, T., Imai, H., Miyano, S., Mitaku, S. & Kanehisa, M., Universal Academy Press, Tokyo, 245-254, 1993.

[3] Smith, H. O., Annau, T. M. & Chandrasegaran, S., "Finding sequence motifs in groups of functionally related proteins," *Proc. Natl. Acad. Sci. USA*, **87**, 826-830, 1990.

[4] Neuwald, A. F. & Green, P., "Detecting patterns in protein sequences," *J. Mol. Biol.*, **239**, 698-712, 1994.