# Searching for Common Sequence Patterns among Distantly Related Proteins

Mikita Suyama, Takaaki Nishioka, and Jun'ichi Oda

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 gaps in the sequences. First, this program generates all possible amino acid patterns composed of up to 5 amino acids. Sequences containing the amino acid residues in the same order to a generated pattern are selected as subsequences, where the differences in the distances between two consecutive amino acids are neglected. Then, the motifs are extracted from the subsequences under the conditions where the all four distances between the five amino acids are fixed. In this stage, motifs with gaps in the subsequence are also found by relaxing one of the four fixed distances. Statistical significance for a motif obtained is calculated based on the amino acid composition of the sequences under consideration. When the GAPE program is applied to 64 ATP-(AMP-forming)-related sequences, motifs extracted with low expectation of occurrence contain some of the amino acid residues chemically proved to be involved in the ligand recognition.

*Keywords*: amino acid sequence / motif / statistical significance / molecular evolution / enzyme reactions / database / WWW

There are many short sequence patterns, often called motifs, among distantly related proteins. These motifs have been derived from a common ancestor and often directly correspond to the functionally important sites, because of the resistance against the mutation on the sites. Then the motifs facilitate to detect very distant relationships that have been obliterated in whole amino acid sequences. Such patterns are useful for the prediction of protein function of uncharacterized sequences such as those determined by genome sequencing projects. As the number of amino acid sequences determined has rapidly increased, it has become clear that automated procedures to find motifs would be useful.

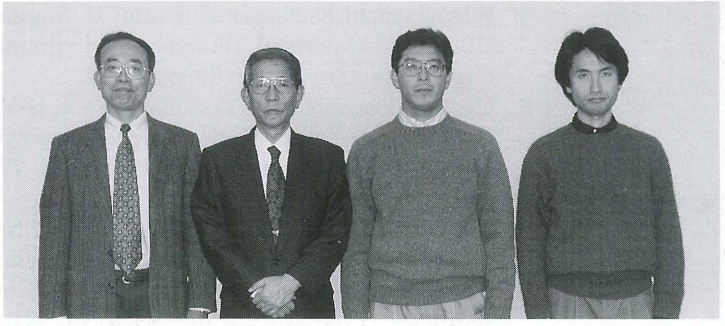
Two methods have been applied to discover sequence motifs, alignment method and pattern-generating method. The first method has such disadvantage that it is hard to make multiple sequence alignment unless the sequences have global similarities. The second method intended for automated search, generates all 3-amino acid patterns including intervening residues of fixed length. Motifs with conservative substitutions are identified on the bases of amino acid patterns common to a majority of the sequences.

About one-tenth of the motifs found in protein sequences, however, contain intervening residues of flexible length called "gap". Any methods so far developed do not explicitly deal with gaps in local

## MOLECULAR BIOFUNCTION —Functional Molecular Conversion—

*Scope of research*

*Our research aims are to analyze structure-function relationships of biocatalysts in combination with organic chemistry, structural biology and computer science, and to apply biocatalysts to stereospecific organic synthesis. Major subjects are the design and preparation of monoclonal antibodies catalyzing chemiluminescence, the reaction mechanisms of glutathione synthetase from E. coli with static, cryogenic, and time-resolved X-ray crystallography, the mechanism of action of lipase-activating protein, crystallographic analysis of asparagine synthetase and γ-L-glutamyl-L-cystein synthetase, and molecular evolution of enzymes and metabolic pathways*



NISHIOKA ODA HIRATAKE KATO

**Professor**
ODA,Jun'ichi (D Agr)
**Associate Professor**
NISHIOKA, Takaaki(D Agr)
**Instructors**
KATO,Hiroaki (D Agr)
HIRATAKE, Jun(D Agr)
TANAKA, Takuji (D Agr)
**Students**
NAKATSU,Toru(DC)
SHIBATA,Hiroyuki(DC)
AOYAGI, Amane(DC)
KATO,Makoto(DC)
SAWA,Kuniaki(DC)
YAMASHITA,Atsuko(DC)
IMAEDA,Yasuhiro(MC)
MATSUDA,Keiko(MC)
TANOUE,Shintaro(MC)
HISADA, Hiromoto(MC)
KITAMURA, Yukiji(RS)

patterns because of combinatorial problems.

We have developed a motif search algorithm GAPE which explicitly deals with gaps in the sequences [1]. GAPE does not compare the sequences each other to find common patterns, but generate all possible patterns to screen out the subsequences. Initially, all possible 3-amino acid patterns (8,000 patterns) are generated. For each pattern, subsequences of length $\leq W_{max}$ that match with a pattern in the order of the three amino acids are searched, and if $\geq s_{min}$ subsequences are detected, they are stored in an array together with the pattern. These 3-amino acid order patterns selected by the above procedure are then added with one more amino acid residue to find 4-amino acid order patterns. The 4-amino acid order patterns satisfying the above criteria of $s_{min}$ and $W_{max}$ are subjected to next extension of the patterns. This procedure is repeated up to 5-amino acid order patterns. In the next step, the subsequences having the same 5-amino acid order pattern are compared with each other with respect to the four distances, $d_1$, $d_2$, $d_3$, and $d_4$, between the five amino acids. When all the corresponding distances are equivalent in $\geq s_{min}$ subsequences, the pattern with the distances is accepted as a rigid motif. Flexible motif which has gaps in subsequences is obtained by relaxing one of the four distances.

GAPE was applied to extract flexible motifs in the sequences of ATP-(AMP-forming)-related enzymes which use ATP as a substrate and form AMP as a product. Their amino acid sequences were systematically collected on the database LIGAND (Ligand Chemical Database for Enzyme Reactions) [2], in which 157 sequences of 37 enzymes were registered. After removing the closely related sequences, 64 sequences were left as the representative sequences.

Under the search condition of $W_{max}$ = 20, $s_{min}$ = 3 and $r_{max}$ = 3, 308 rigid and 287 flexible motifs were obtained with an expected value of occurrence of a motif less than 5 % ($E$ < 0.05). The motif with the lowest $E$ value was K-M-S-(x1)-S-(x2)-N, which occur in the sequences of 7 aminoacyl-tRNA synthetases class I. Tetrapeptide H-I-G-H, which is known to be conserved among aminoacyl-tRNA synthetases class I, was also found as a flexible motif P-(x1)-A-(x3,4)-H-(x1)-G-H In the three-dimensional structure of tyrosine—tRNA ligase, this region has been shown to be a part of the adenylate binding site (Brick et al., 1989). Most of the motifs obtained are variations of these two motifs. All of the sequences of aminoacyl-tRNA synthetases class I contain some of the variations of the two motifs.

One of the flexible motifs found among aminoacyl-tRNA synthetases class II is G-(x2)-P-(x2,3)-G-(x3)-G-(x2)-R (Figure 1). The enzymes containing this motif are lysine—tRNA ligase (EC 6.1.1.6), aspartate—tRNA ligase (EC 6.1.1.12), asparagine—tRNA ligase (EC 6.1.1.22), and aspartate—ammonia ligase (EC 6.3.1.1). This is the only motif with $E$ < 0.05 that occurs in the sequence of aspartate—ammonia ligase. The similarity of aspartate—ammonia ligase to

```
S17011   CNALEY GLPP-TGGWGCGIDR LAMFL
SYRTDT   IDSFRF GAPPH-AGGGIGLER VTMLF
S23761   LDALKY GTPPH-AGLAFGLDR LTMLL
SYBYDM   LNAFDM GTPPH-AGFAIGFDR MCAMI
SYECNT   RDLRRY GTVPH-SGFGLGFER LIAYV
AJECNA   HQALLR GEMPQTIGGGIGQSR LTMLL
Motif          G  P  G   G  R
```

**Figure 1**. Motifs extracted from ATP-(AMP-forming)-related sequences. The first column represents PIR entry codes for each sequence. Enzyme and the residue number of the first amino acid in the motif are as follows: S17011 = lysine — tRNA ligase, 543; SYRTDT = aspartate— tRNA ligase, 462; S23761 = aspartate— tRNA ligase, 143; SYBYDM = aspartate—tRNA ligase, 594; SYECNT = asparagine—tRNA ligase, 427; AJECNA = aspartate—ammonia ligase, 285.

aminoacyl-tRNA synthetases class II was first reported by Gatti and Tzagoloff (1991). For aspartate—ammonia ligase from *Escherichia coli*, mutation of the arginine residue in the motif confirmed that the residue is crucial for its activity (Hinchman *et al.*, 1992). This motif provides another support for the possibility that these two enzymes evolved from a common ancestral enzyme.

The motifs obtained among the ligand-related sequences by GAPE are well correlated with the ligand recognition sites of the sequences. These motifs imply that the enzymes sharing the motifs have been evolved from a common ancestor recognizing the ligand, though none of global sequence similarity is detected. The ancestral enzymes would have been diverged to specific reactions for each enzyme retaining the ligand specificity.

Amino acid sequences for the motif search have to be classified and systematically collected according to protein function such as ligand specificity or reaction type. For this purpose, we have constructed LIGAND. It is actually composed of two databases; ENZYME and COMPOUND. ENZYME contains the EC numbers, names, chemical equations catalyzed, substrates, products, cofactors, inhibitors of 3,489 enzymes. COMPOUND, a database of 5,118 chemical substances appeared in ENZYME, contains their chemical names, chemical structures (in a connection table format and in an image), and CAS registry number. ENZYME has a link to the databases of DNA base sequences, amino acid sequences, 3-D structures, and inheritance diseases and has a link to metabolic pathways. Each chemical substance in the COMPOUND has a link to ENZYME which enables to collect a set of enzymes and their sequences related to a substance desired. The WWW version of LIGAND is served on GenomeNet (http://www.genome.ad.jp).

## References

1. Suyama M, Nishioka T. and Oda J. *Protein Engineering, in press* (1995).
2. Suyama M, Ogiwara A, Nishioka T. and Oda J *Comput. Appl. Biosci.*, **9**, 9-15 (1993) .