AN APPROACH TO DETECTION OF PROTEIN STRUCTURAL MOTIFS USING AN ENCODING SCHEME OF BACKBONE CONFORMATIONS

H. MATSUDA, F. TANIGUCHI, A. HASHIMOTO
Department of Informatics and Mathematical Science,
Graduate School of Engineering Science,
Osaka University,
1-3 Machikaneyama, Toyonaka, Osaka, 560 Japan
e-mail: {matsuda, fumihiro, hasimoto}@ics.es.osaka-u.ac.jp

This paper presents an approach to detection of protein structural motifs. In our approach, first all protein backbone conformations are converted into character strings using an encoding scheme. Then we use the Smith-Waterman local alignment algorithm to detect common structural motifs. By comparing results with the PROSITE regular expression patterns, our method can detect several motifs which the PROSITE patterns fail to detect.

1 Introduction

The amount of protein sequence data has been rapidly increasing due to the progress of DNA sequencing technology. Using those sequence data, a number of methods for investigating the evolutionary process of taxa (biological entities such as genes, proteins, individuals, populations, species, or higher taxonomic units) have been proposed. For example, construction of phylogenetic trees from the sequence data of various organisms has been widely used to clarify the evolutionary histories of those organisms[1,2].

Previous studies on molecular evolution suggested that the evolutionary process of proteins is closely related to their structure and function. For example, when some mutations have a higher chance of causing deleterious effects on the function of the protein than other mutations, the majority of the former mutations will be eliminated from the population by purifying selection. The result will be a reduction in the rate of those mutations compared to that of other mutations.

Recent studies[3,4] revealed that many protein domains adopt the same fold structures even if they have statistically insignificant sequence similarity. In order to understand the relationship among the sequence, structure and function of proteins, it is necessary to infer how the relationship has been emerged in evolutionary process.

In this paper, we focus on locally common substructures, called structural motifs, among related proteins. These motifs are recognized as functionally

important sites because they are well conserved in a wide variety of distantly related proteins. We consider those motifs play a role of *building blocks* to establish a variety of functions of proteins.

In order to extract the motifs from proteins, we developed a method for exploring similar protein structure using an encoding scheme of backbone conformations of proteins[5]. In our method, first describe a protein structure as a sequence of vectors obtained by connecting the C_α atoms of consecutive amino acids in the protein, then quantizing each vector to twenty representative by fitting it into one of the normal vectors of the twenty faces in an icosahedron, so that a protein structure is represented by a string of twenty characters. Because the set is limited to twenty representatives, our notation makes it possible to utilize available string matching algorithms.

Previous researches for encoding protein structures into character strings are based on two-dimensional square lattice with two types (nonpolar and other) of amino acids[6], backbone dihedral angles[7,8] or based on the root mean square distance of every seven amino acid segment[9]. Our work relates to researches based on inter-C_α torsion angles[10] and it is an extension of the chain coding scheme in image processing[11,12]. Our method is a general method which is not restricted to describe protein structures but applicable to arbitrary line-drawing objects in three-dimensional space.

The later sections describe the detail of our encoding scheme and its application to detecting protein structural motifs.

2 Encoding scheme

We used three-dimension coordinate data of C_α atoms taken from the Protein Data Bank (hereafter, PDB) Release 72.0.

The method to encode a protein structure into a string is described below. Let n be the number of C_α atoms in a protein. Let p_{i-2}, p_{i-1}, p_i and p_{i+1} $(3 \leq i \leq n-1)$ be successive four C_α atoms in the protein. Place p_{i-2}, p_{i-1} and p_i in the Cartesian coordinates as follows (see Figure 1):

- p_i to the origin of the coordinates,
- p_{i-1} on the z-axis,
- p_{i-2} on the zx-plane.

Here a position vector of p_{i+1} is determined on this coordinates. Then place an icosahedron in the Cartesian coordinates so that the gravity center of the icosahedron is set to the origin of the coordinates. Label each normal vector of twenty faces of the icosahedron twenty letters of the English alphabet as

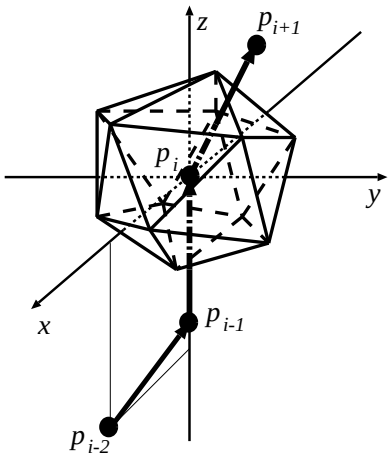


Figure 1: The placement of a protein backbone chain with an icosahedron in the Cartesian coordinates.

shown in Figure 2. `B`, `J`, `O`, `U`, `X` and `Z` are omitted from the alphabet so that it agrees with the one letter code of amino acids.

Using the icosahedron, the position vector of p_{i+1} is quantized by the nearest normal vector in the coordinates, then the position of p_{i+1} is represented by the letter of the normal vector. By this notation, every backbone chain of protein can be denoted to a string of twenty letters.

As a practical matter, there exists ambiguousness for the placement of an icosahedron in the coordinates. We set it so that every amino acid in α-helices is encoded into the same letter (`W` in Figure 2). From the analogy to an encoding method of diagrams in image processing[11,12], we named our notation as *three-dimensional chain code* (hereafter, 3D chain code). Same as the original chain code, the 3D chain code is independent on its position, moving and rotation in any coordinates. Thus the comparison of protein backbone structures can be carried out by some string matching algorithm. Moreover, the method is easily extended to pairwise and multiple alignment.

In the 3D chain code, the letters corresponding to the first two amino acids p_1 and p_2 and the last amino acid p_n cannot be determined since the placements of these points in the coordinates are not defined. However the letter corresponding to p_2 can be decided by restricting the placement of p_1, p_2 and p_3 on the zx-plane (p_1 on the z-axis, p_2 at the origin and p_3 on the zx-plane). Thus the length of the 3D chain code in a protein becomes $m-2$

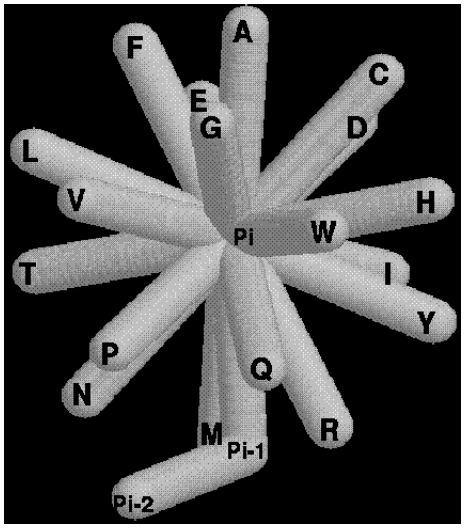


Figure 2: The 3D chain code with the normal vectors of an icosahedron.

where m is the number of amino acids in the protein.

Figure 3 shows the distribution of 3D chain code letters in PDB Release 72.0. As described above, the letter `W` corresponds to α-helix. The number of occurrences of `W` amounts to 34.7% of all occurrences. The letters `E` and `D` correspond to β-strand. These amount to 17.6% and 12.0%, respectively.

3 Application to searching structural motifs

The PROSITE database[13] contains a large number of entries on motifs. They are mainly classified by using regular expressions which accept amino acid sequence patterns specific to motifs. However even if homology between two amino acid sequences is not so high (such as around 30% of amino acid identities), their tertiary structures are sometimes very similar (within 2.0 Å in the root mean square distance by best-fit superposition method[14]). In this case, it is difficult to detect the candidates of motifs by an analysis only based on sequence similarity. Actually in some case the regular expressions of PROSITE cannot detect motifs.

To cope with this problem, we applied our encoding scheme to detection of motifs. We used the `SSEARCH` and `LALIGN` programs in the version 2.0 of the `FASTA` program package[16]. These programs are based on the Smith-Waterman

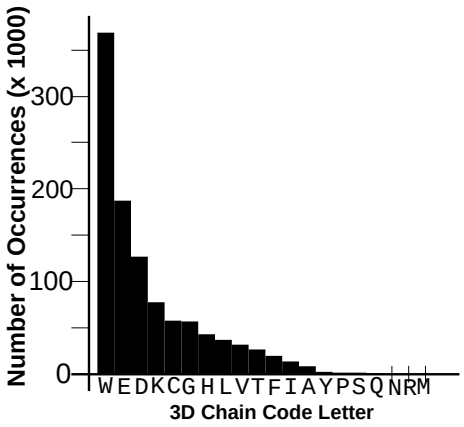


Figure 3: The distribution of 3D chain code letters in PDB Release 72.0

local alignment algorithm[15] and thus suitable for finding local common subsequences.

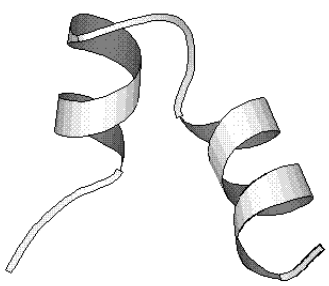
A similarity score s_{ij} between any two letters c_i and c_j of the 3D chain code ($1 \leq i, j \leq 20$) is calculated as:

$$s_{ij} = 10 \cos \theta_{ij}$$

where θ_{ij} denotes the angle between two normal vectors corresponding to the 3D chain code letters c_i and c_j (see Figure 2). The penalties for an opening gap and an extension gap are set to -30 and -2, respectively. The opening gap penalty is much higher than the default value (-12) since we intend to find longer common subsequences.

3.1 Helix-turn-helix

Helix-turn-helix motif is specific to the DNA-binding domain of many bacterial transcription regulation proteins. In the PROSITE database, this motif is classified into subfamilies based on sequence similarities. To extract of a pattern for the motif in the 3D chain code, we picked up two regulation proteins; *E. coli* lactose operon repressor LacI (PDB id 1lccA) and *E. coli* catabolite gene activator (PDB id 3gapA). Then we performed the `LALIGN` (local alignment) program with the 3D chain code scoring matrix to find common sequence patterns between them. As the result, we obtained very similar backbone confor-



(a) Ribbon model of the extracted helix-turn-helix domain from 3gapA (drawn by MOLSCRIPT[17]).

(b) Superposition of the two extracted domains of 3gapA (dark) and 1lccA (light).

Figure 4: Helix-turn-helix DNA-binding domains extracted from *E. coli* lactose repressor (PDB id 1lccA) and *E. coli* catabolite gene activator (PDB id 3gapA).

mations and confirmed that these correspond to helix-turn-helix DNA-binding domains.

The extracted chain-coded conformations and amino acid sequences of the helix-turn-helix domains are as follows:

```
          3D chain code          Amino acid
1lccA TWWWWWWWKCLKWWWWWWG LYDVAEYAGVSYQTVSRVV
3gapA KWWWWWWGKCFKWWWWWWW RQEIGQIVGCSRETVGRIL
```

In this case, although the identity of amino acids is very low ($5/19 = 26.3\%$), the superposition of these two domains is well overlapped (see Figure 4) and the root mean square distance between them is 0.57 Å.

Then we executed the `SSEARCH` similarity search program with the 3D chain code scoring matrix between the domain of 1lccA and all chain-coded conformations of PDB Release 72.0. Figure 5 shows the result of `SSEARCH` sorted by the descending order of S-W score (the similarity score of Smith-Waterman local alignment). The first two proteins are helix-turn-helix DNA-binding domains of other bacterial transcription regulation proteins (1lcdA: another PDB entry for *E. coli* lactose operon repressor, 1trt: *E. coli* tetracycline repressor). The latter two lines are fragments of repressor proteins of bacteriophages (1pra: bacteriophage 434, 4cro: bacteriophage lambda). According to the description of the SWISS-PROT entries (Accession numbers P16117 and P03040, respec-

PDBid	S-W score	3D Chain Code	RMSD	Amino Acid	PROSITE pattern
1lcdA	91	TWWWWWWWKCFKWWWWWWG	0.33	LYDVAEYAGVSYQTVSRVV	HTH_LACI
1trt	88	KWWWWWWWKCFKWWWWWGG	0.73	TTRKLAQKLGIEQPTLYWH	HTH_TETR
1pra	86	TWWWWWWWKCLKWWWWWWW	0.41	QAELAQKVGTTQQSIEQLE	no
4croA	86	TWWWWWWWKCLKWWWWWWW	0.58	GQTKTAKDLGVYQSAINKA	no

Figure 5: Similarity search result for helix-turn-helix DNA-binding domains between *E. coli* lactose repressor (PDB id 1lccA) and PDB using 3D chain coding scheme.

tively), these are also helix-turn-helix DNA-binding domains. But there do not exist motif patterns on these domains in the PROSITE database.

3.2 EF-hand

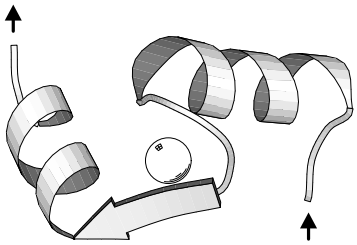
Many calcium-binding proteins belong to the same evolutionary family and share a type of calcium-binding domain known as the EF-hand. We did the same approach to the detection of helix-turn-helix motifs. First we picked up two proteins known to have the EF-hand calcium-binding domains; *Paramecium tetraurelia* calmodulin (PDB id 1clm) and *Triakis semifasciata* (leopard shark) parvalbumin alpha (PDB id 5pal), then we extracted chain-coded backbone conformations of EF-hand domains by using `LALIGN`. Second similarity search between one of these conformation (the domain of 1clm) and all entries of PDB 72.0.

The extracted chain-coded conformations and amino acid sequences of the EF-hand domains are as follows:

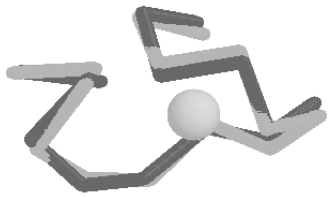
```
         3D chain code                          Amino acid
1clm WWWWWWWWWWWDWKHKCEETWWWWWWWWW EELIEAFKVFDRDGNGLISAAELRHVMTNL
5pal WWWWWWWWWWHDWKIKAKETWWWWWHWWW AQVKEVFEILDKDQSGFIEEEELKGVLKGF
```

In this case, the identity of amino acids is also relatively low ($9/30 = 30.0\%$) but the superposition of these two domains is well overlapped (see Figure 6) and the root mean square distance of these domains is 1.72 Å.

Figure 7 shows the result of `SSEARCH`. Although the difference among each S-W score is small, the values of RMSD (the root mean square distance by best superposition) ranges from 0.41 to 1.07 which seems to be due to quantization error by fitting to the normal vectors of the icosahedron. However all domains detected by this search have the pattern of the EF-hand motif described by the regular expression `D-x-[DNS]-{ILVFYW}-[DENSTG]-[DNQGHRK]-{GP}--[LIVMC]-[DENQSTAGC]-x(2)-[DE]-[LIVMFYW]` in the PROSITE database.



(a) Ribbon model of the extracted EF-hand domain of 1clm (drawn by MOLSCRIPT).



(b) Superposition of the two extracted domains of 1clm (dark) and 5pal (light).

Figure 6: EF-hand calcium-binding domains extracted from *P. tetraurelia* calmodulin (PDB id 1clm) and *Triakis semifasciata* (leopard shark) parvalbumin alpha (PDB id 5pal).

3.3 Greek key

The beta and gamma crystallins (the dominant structural components of the eye lens) form a family of related proteins. Structurally, they consist of two similar domains which are each composed of two similar motifs with the two domains connected by a short connecting peptide. Each motif, which is about forty amino acid residues long, is folded in a distinctive 'Greek key' pattern. The key structure is composed of a β-sheet where two anti-parallel β-strands are connected.

We picked up a gamma-b crystallin (PDB id 4gcr) and a beta crystallin b2 (PDB id 1blbB) from *Bos taurus* (calf) eye lens and extracted Greek Key domains using `LALIGN` as follows:

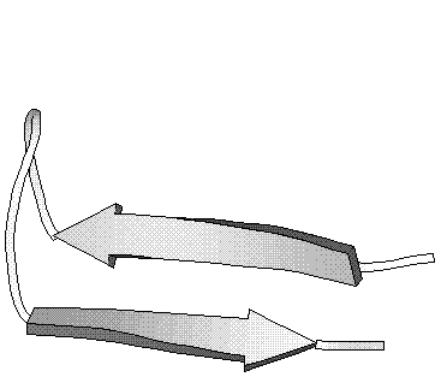
```
            3D chain code      Amino acid
4gcr    DEEEKGLWTCGVDEED  ITFYEDRGFQGHCYEC
1blbB   DEEEKGLGTDVWDDEE  IIIFEQENFQGHSHEL
```

As mentioned in Section 2, the 3D chain code letter for β-strand does not converge to one letter as α-helix but usually takes either `E` or `D`. Thus the identity of chain code in Greek Key decreases compared to those of previous examples. However the chain code identity is still larger than the amino acid identity. For example in the above case, the identity of the chain code is $10/16 = 62.5\%$ whereas that of amino acids is $7/16 = 43.8\%$. The superposition of these two domains is as shown in Figure 8 and the root mean square distance of these domains is 0.58 Å.

PDBid	S-W score	3D Chain Code	RMSD
1top	129	WWWWWWWWWWWDWKHKCEETWWWWWWWWWW	0.61
1ctr	129	WWWWWWWWWWWDWKHKCEETWWWWWWWWWW	0.65
3cln	127	WWWWWWWWWWWDWKHKCEDTWWWWWWWWWW	0.26
1osa	127	WWWWWWWWWWWDWKHKCEEKWWWWWWWWWW	0.64
1cll	127	WWWWWWWWWWWDWKYKCEETWWWWWWWWWW	0.65
1cdlC	127	WWWWWWWWWWWDWKHKCEEKWWWWWWWWWW	0.42
2bbmA	126	HWWWWWWWWWWDWKHKCEEKWWWWWWWWWW	0.89
5tnc	125	WWWWWWWWWWWDWKYKAEETWWWWWWWWWW	0.68
4cln	125	WWWWWWWWWWWDWKHKAEEKWWWWWWWWWW	0.66
2scpB	125	WWWWWWWWWWWDWKYKCEEKWWWWWWWWWW	1.07
1cdlA	125	WWWWWWWWWWWDWKYKCEEKWWWWWWWWWW	0.41
2sas	125	WWWWWWWWWWWIWKHKHEETWWWWWWWWWW	0.62
1cdmA	124	WWWWWWWWWWHDWKHKCEDKWWWWWWWWWW	0.61
1mysB	123	WWWWWWWWWWHDWKHKCECTWWWWWWWWWW	0.97

PDBid	Amino Acid	PROSITE pattern
1top	EELANCFRIFDKNADGFIDIEELGEILRAT	EF-hand
1ctr	AEFKEAFSLFDKDGDGTITTKELGTVMRSL	EF-hand
3cln	EEIREAFRVFDKDGNGYISAAELRHVMTNL	EF-hand
1osa	EELIEAFKVFDRDGNGLISAAELRHVMTNL	EF-hand
1cll	AEFKEAFSLFDKDGDGTITTKELGTVMRSL	EF-hand
1cdlC	EEIREAFRVFDKDGNGYISAAELRHVMTNL	EF-hand
2bbmA	EEIREAFRVFDKDGNGFISAAELRHVMTNL	EF-hand
5tnc	EELEDCFRIFDKNADGFIDIEELGEILRAT	EF-hand
4cln	AEFKEAFSLFDKDGDGTITTKELGTVMRSL	EF-hand
2scpB	TMAPASFDAIDTNNDGLLSLEEFVIAGSDF	EF-hand
1cdlA	EEIREAFRVFDKDGNGYISAAELRHVMTNL	EF-hand
2sas	NRIPFLFKGMDVSGDGIVDLEEFQNYCKNF	EF-hand
1cdmA	AEFKEAFSLFDKDGDGTITTKELGTVMRSL	EF-hand
1mysB	QDFKEAFTVIDQNRDGIIDKDDLRETFAAM	EF-hand

Figure 7: Similarity search result for EF-hand calcium-binding domains between *P. tetraurelia* calmodulin (PDB id 1clm) and PDB using 3D chain coding scheme.



(a) Ribbon model of the extracted Greek key domain of 4gcr (drawn by MOLSCRIPT).

(b) Superposition of the two extracted domains of 4gcr (dark) and 1blbB (light).

Figure 8: Greek key domains extracted from *Bos taurus* gamma-b crystallin (PDB id 4gcr) and *Bos taurus* beta crystallin b2 (PDB id 1blbB).

Figure 9 shows the result of `SSEARCH` (upper 5 lines) and `LALIGN` (lower 4 lines). There are usually four Greek Key motifs in proteins which have this motif. However in a few cases, the PROSITE regular expression pattern for Greek Key (`[LIVMFYWA]-x-{DEHRKSTP}-[FY]-[DEQHKY]-x(3)-[FY]-x-G--x(4)-[LIVMFCST]`) may fail to detect one of four motifs. Thus we tried to detect all four motifs of beta-b2-crystallin in *Bos taurus* lens (PDB id 2bb2) by using `LALIGN` with chain-coded conformations between 4gcr and 2bb2. All motifs including the second one which the PROSITE pattern cannot detect were successfully detected.

4 Analysis of structure evolution

As mentioned in Sect. 3, our encoding scheme can be easily applied to available alignment programs for amino acid sequences by replacing its scoring matrix. Similarly, in principle, it can be applied to available phylogenetic tree construction programs (e.g. our developed system based on the maximum likelihood method [18]). If we construct a tree with our encoding scheme, however, the distances between any two nodes of the tree do not present the evolutionary

PDBid	S-W score	3D Chain Code	RMSD	Amino Acid	PROSITE pattern
1gcs	112	DEEEKGLWTCGGDEED	0.23	ITFYEDRGFQGHCYEC	GreekKey
2gcr	111	DEEEKGLGTCGVEEED	0.39	WMLYEQPNFTGCQYFL	GreekKey
1prr	104	EEDEKGFWTCGLEEED	0.64	ITVFYNEDFQGKQVDL	GreekKey
1prs	101	EEEEEGFWTCGLEEED	0.77	ITVFYNEDFQGKQVDL	GreekKey
1blbC	100	CEDEKGFWLCGGDDEE	0.72	IIIFEQENFQGHSHEL	GreekKey
2bb2-1	108	DEEEKGFWTCGVEEEE	0.27	IIIFEQENFQGHSHEL	GreekKey
2bb2-2	106	DEEEKGLWTCGGEDEE	0.60	WVGYEQANCKGEQFVF	no
2bb2-3	110	DEEEKGLWLCGVEEED	0.33	ITLYENPNFTGKKMEV	GreekKey
2bb2-4	110	DEEEKGLWTCGVEEEE	0.56	WVGYQYPGYRGLQYLL	GreekKey

Figure 9: Similarity search result for Greek Key structure between *Bos taurus* gamma-b crystallin (PDB id 4gcr) and PDB using 3D chain coding scheme.

distances between protein conformations since the change of protein structure highly depends on its function. We will describe structure evolution with our encoding scheme and will analyze frequencies of changes of the code.

5 Conclusion

We have developed a method to detect structural motifs by encoding protein backbone conformations into character strings. The encoding scheme is based on a general encoding scheme for line drawing objects and can give similarity score among each letter of the alphabet so that available string manipulation algorithms can be used for protein structure analysis. By using our method, protein sequences and their tertiary structures can be processed in the same manner except scoring matrices. We hope we will analyze the evolution process of protein structure using our encoding scheme and apply it to reconstruct phylogenetic relationships of protein structures.

Acknowledgement

This work was supported in part by a Grant-in-Aid (08283103) for Scientific Research on Priority Areas from The Ministry of Education, Science, Sports and Culture in Japan.

References

1. M. Nei, *Molecular Evolutionary Genetics* Chap.11 (1987).

2. D.L. Swofford and G.J. Olsen, "Phylogeny Reconstruction," in *Molecular Systematics,* ed. D.M. Hillis and C. Moritz, 411–501 (1990).
3. C. Chothia, "Proteins. One Thousand Families for the Molecular Biologist," *Nature*, **357**, 543–544 (1992).
4. C.A. Orengo, D.T. Jones and J.M.Thornton, "Protein Superfamilies and Domain Superfolds," *Nature*, **372**, 631–634 (1994).
5. H. Matsuda, F. Taniguchi and A. Hashimoto, "A Notation of Amino Acid Conformations for Exploring Similar Protein Structure," *Proc. 1st Pacific Symp. Biocomp.*, 732–733 (1996).
6. K.F. Lau and K.A. Dill, "Theory for Protein Mutability and Biogenesis," *Proc. Natl. Acad. Sci. USA*, **87**, 638–642 (1990).
7. M.J. Rooman, J.-P.A. Kocher and S.J. Wodak, "Prediction of Protein Backbone Conformation based on Seven Structure Assignments: I Influence of Local Interactions," *J. Mol. Biol.*, **221**, 961-979 (1991).
8. R.T. Miller, R.J. Douthart, and A.K. Dunker, "An Alphabet of Amino Acid Conformations in Protein," *Proc. HICSS*, **1**, 689–698 (1993).
9. Y. Matsuo and M. Kanehisa, "An Approach to Systematic Detection of Protein Structure Motifs," *Comp. Appl. Bio. Sci.*, **9**(2), 153–159 (1993).
10. M. Levitt and J. Greer, "Automatic Identification of Secondary Structure in Globular Proteins," *J. Mol. Biol.*, **114**, 181–239 (1977).
11. H. Freeman, "On the Encoding of Arbitrary Geometric Configurations," *IRE Trans. Electronic Computer*, **EC-10**(2), 260–268 (1961).
12. H. Freeman, "Computer Processing of Line-Drawing Images," ACM Comp. Surv., **6**(1), 57–97 (1974).
13. A. Bairoch, P. Bucher and K. Hofmann, "The PROSITE database, its status in 1995," *Nucl. Acids Res.*, **24**(1), 189–196 (1996).
14. W. Kabsch, "A Solution for the Best Rotation to Relate Two Sets of Vectors," *Acta. Cryst.*, **A32**, 922–923 (1976).
15. T.F. Smith and M.F. Waterman, "Identification of common molecular subsequences," *J. Mol. Biol.*, **147**, 195–197 (1981).
16. W.R. Pearson, "Rapid and Sensitive Sequence Comparison with FASTP and FASTA", *Methods in Enzymology*, **183**, 63–98 (1990).
17. P.J. Kraulis, "MOLSCRIPT: A Program to Produce Both Detailed and Schematic Plots of Protein Structuresm," *J. Appl. Cryst.*, **24**, 946–950 (1991).
18. H. Matsuda, "Protein Phylogenetic Inference Using Maximum Likelihood with a Genetic Algorithm," *Proc. 1st Pacific Symp. Biocomp.*, 512–523 (1996).