An Approach to Systematic Detection of Protein Structural Motifs

Yo Matsuo* and Minoru Kanehisa

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

**Present address: International Institute For Advanced Study of Social Information Science, Fujitsu Laboratories Ltd., 1-17-25 Shinkamata, Otaku, Tokyo 144, Japan.*

A procedure to detect similar local structures of proteins from Cα coodinates is presented. First, the conformations of peptide segments of seven residues long are approximated by a limited number of representatives, each of which is assigned a symbol. Thus, the entire conformation of a protein is represented by a symbol string. Then, the comparison of these symbol strings using sequence alignment technique gives pairs of similar local structures. These pairs are considered candidates of structural motifs. The application of the procedure to the analysis of 93 proteins gave 858 pairs of similar local structures, which inculded several well-known structural motifs such as nucleotide-binding βαβ motif and calcium-binding EF hand. The characterization of amino acid patterns of similar local structures given by the procedure should be useful for the development of protein structure prediction based on the acquisition of empirical rules from a large-scale database.

MATERIALS AND METHODS

Coordinate Data

We used coordinate data of 93 polypeptide chains taken from the Protein Data Bank[1], release 49 (July 1989). Resolutions of all those coordinate data were higher than 3.0Å, and the 93 chains showed no significant amino acid sequence homology. Only α-carbons were considered in this work. However, it is straightforward to extend our approach to consider all main chain atoms.

Conversion of Protein Coordinate Data into Symbol Strings

First, the conformation of short peptide segments are approximated by a limited number of *representatives*. The peptide segment length l is fixed. The *sampling interval* d_S is a parameter, which determines the degree of approximation of peptide segment conformations. In the present work, l and d_S were set at 7 residues and 2.01Å, respectively, which we empirically found to give a good result.

Suppose that the total of N_S peptide segments of l-residues are contained by all the polypeptide chains in a given database. Let us arrange N_S segments in an arbitrary order: $s_1, s_2, \ldots, s_i, \ldots, s_{N_S}$. Here, s_i is the i-th segment. Representatives, which should approximate all the N_S segments, are selected as follows. Below, r_p denotes the p-th representative.

(1) Set $r_1 = s_1$.

(2) Suppose that q elements have so far been selected: $r_1=s_1,\ldots,r_p=s_i,\ldots,r_q=s_j$ $(1<\ldots<i<\ldots j)$. Starting from s_{j+1}, find the first segment s_k $(j<k)$ which satisfies the following set of inequalities: $d(r_1,s_k)\geq d_S,\ldots,d(r_p,s_k)\geq d_S,\ldots,d(r_q,s_k)\geq d_S$. Then, $r_{q+1}=s_k$. Here, $d(r,s)$ is the r.m.s distance between equivalent atoms of segments r and s; the distance is calculated using the best-superposition algorithm by Kabsch[4,5].

(3) Repeat the operation (2) until all the N_S peptide segments are examined.

Suppose that N_T segments have been selected. Then, the conformation of each $s_i(1\leq i\leq N_S)$ is approximated by a representative r_p which has the minimum distance $d(r_p,s_i)$ among all $d(r_q,s_i)$ $(1\leq q\leq N_T)$. For each representative r_q $(1\leq q\leq N_T)$, an arbitrary symbol is

chosen. The same symbol is assigned to all the peptide segments which are approximated by r_q.

Now, the entire conformation of a polypeptide chain is repreesnted by the symbol string: $P_1 P_2 ... P_k ... P_{Li-l+1}$. Here, P_k is the symbol assigned to the l-residue segment spanning from k-th through (k+l-1)-th residues of the polypeptide chain. L_i is the number of residues of the polypeptide chain.

Detection of Similar Local Structures of Proteins

Similar local structures were detected through the comparison of symbol strings which were derived through the above procedure. To compare symbol strings, we used the Goad-Kanehisa algorithm[2]. The algorithm is a generalization of Needleman-Wunsch algorithm[3] and detects similar subsequences in the form of the best-alignment including gaps. The r.m.s distance between two representatives $d(r_i,r_j)$ defined by Kabsch's algorithm[4,5] was taken as the similarity score for the two symbols A and B assigned to r_i and r_j, respectively. The gap penalty was taken to be the same as the maximum of the r.m.s. distances between representatives.

For each pair of polypeptide chains, 10 pairs of local structures were given by the Goad-Kanehisa algorithm. Therefore, from the comparison of the 93 chains, the total of 43,710 (= 10 x (93x92/2 +93)) pairs of local structures were given (we also considered the comparison of a chain with itself). Since we were interested in the detection of super-secondary-level structural similarities, we neglected those pairs of structures which had less than 25 residues to avoid detecting such small structures as single secondary structures. In some cases, not so similar local structures were paired because their string alignment score was better than other possible pairs of local structures. To avoid such cases, we calculated the r.m.s. distance between the detected pair of two local structures using Kabsch algorithm[4,5]. Equivalent atoms of the structures, including gaps if necessary, were determined by the alignment given by Goad-Kanehisa algorithm. Then, we neglected those pairs whose r.m.s. distances were larger than 4.0Å.

RESULTS

Conversion of Protein Coordinate Data into Symbol Strings

The 93 proteins in the database contained 15,320 peptide segments of 7 residues. Under the sampling interval d_S = 2.01Å, 37 representatives were selected. The 37 representatives were assigned symbols: 1, 2, 3, 4, 5, 6, 7, 8, 9, A, B, C, D, E, F, G, H, I, J, K, L, M, N, O, P, Q, R, S, T, U, V, W, X, Y, Z, @ and #. The mean error in the approximation by representatives was 1.28Å per peptide segment, and the standard deviation was 0.30Å.

Detection of Similar Local Structures of Proteins

The comparison of the symbol strings of the 93 polypeptide chains gave 858 pairs of similar local structures. The 858 pairs included βαβ-motif, EF-hand, β-hairpins, helix bundles, fragments of β/α-barrels, single helices with turns at its terminals, etc.

For example, calcium-binding EF-hand motif[6] appeared in the following five pairs: {1CPV(38-70), 2LZM(37-71)}, {1CPV(44-71), 3ICB(47-75)}, {1CPV(76-106), 1CPV(37-67)}, {1CPV(77-108), 2LZM(37-69)} and {1CPV(81-108), 3ICB(45-72)}. Their r.m.s. distances were 3.67Å, 1.77Å, 2.18Å, 2.67Å and 3.82Å, respectively.

Another example was βαβ-motif. The following three pairs contained βαβ-motif with additional α-helix at their C-terminal end: {4ADH(192-236),6LDH(20-65)}, {4ADH(192-236),1ABP(35-78)}, and {1GD1O(1-47),6LDH(20-65)}. Their r.m.s. distances were 2.02Å, 2.78Å and 3.14Å, respectively. Except for 1ABP(35-78), the structures were nucleotide-binding regions of dehydrogenases. The amino acid pattern of the nucleotide-binding βαβ-motif has widely been discussed by many researchers[7,8,9]. We also examined the amino acid pattern common to the structures in the above three pairs by constructing the multiple alignment of the structures (Figure 1). Each residue position in the multiple alignment was characterized according to the physico-chemical classification of amino acids by Taylor[10]. A pattern was derived which was comparable with that reported by other authors[7,8].

DISCUSSION

The systematic comparison of protein structures has been a difficult task. We utilized sequence comparison technique. To utilize sequence comparison technique for the purpose of structure comparison, we converted protein coordinate data into symbol strings (cf Taylor & Orengo[11]). The conversion was based on the approximation of conformations of short peptide segments by a limited number of *representatives*.

Representatives were selected from the whole set of peptide segments in a database through the unique algorithm devised by us. If the purpose was to find peptide segments whose conformations occur frequently in a given database, other clustering algorithms would be more suitable[12,13]. However, our purpose was to select peptide segments which could approximate all the peptide segments within a certatin upper limit of error set in advance (= sampling interval, ds). Therefore, representatives should cover the whole conformational distribution of peptide segments as *uniformly* as possible, and should be selected independently of the frequency of peptide segment conformations (Figure 2). Our algorithm is by definition guaranteed to give a set of representatives which can approximate any peptide segments within a certain error (ds).

There would be a number of possible sets of representatives depending on particular way of ordering peptide segments (s_1, s_2, ..., s_{Ns}) (see Materials and Methods). In any case, however, any set of representatives selected by our algorithm can approximate all the peptide segments within a given d_s.

Once three-dimensional coordinate data are converted into one-dimensional symbol strings, it is possible to apply a wide-range of techniques developed for sequence analysis. For example, Needleman-Wunsch algorithm[3] can be applied to the global alignment of protein structures if the structures show global homology. In the present study, we focused on the detection of local similarity between protein structures which show no overall similarity. For this purpose, Goad-Kanehisa algorithm[2] was suitable.

Through an exhaustive search of the 93 protein structures by the symbol string comparison approach, 858 pairs of local strucutres were detected. We did not examined all the pairs fully, and did not report any new structural motif in the present paper. However, well-known structural motifs such as βαβ-motif and EF-hand were successfully detected. This shows our approach should be useful for the search of a database for protein structural similarities, especially when the database is large.

The development of DNA cloning and sequencing technology brought about the rapid increase of protein sequence data. The rate of increase will further be accelerated by new technology developments under the Human Genome Project. In order to make full use of the vast amount of sequence data, it becomes more and more important to develop computational methods to derive information of biological interest from the amino acid sequence. The information on protein structure is useful for a better understanding of the function of protein. In the present work, we characterized amino acid pattern of βαβ-motif as an example, and obtained the result comparable with that reported by other authors[7,8]. The systematic characterization of amino acid patterns of similar local structures detected by our method should be useful for the development of structure prediction from sequence.

ACKNOWLEDGMENTS

We thank Kenta Nakai, Motohisa Oobatake, and Atsushi Ogiwara for useful discussions. This work was partly supported by a grant-in-aid from the Ministry of Education, Science and Culture of Japan.

REFERENCES

1. Bernstein,F.C., Koetzle,T.F., Williams,G.J.D., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O., Shimanouchi,T. & Tasumi,M. *J.Mol.Biol.*112: 535-542, 1977.
2. Goad,W.B. & Kanehisa,M.I. *Nucleic Acids Res.*10: 247-263, 1982.
3. Needleman,S.B. & Wunsch,C.D. *J.Mol.Biol.* 48: 443-453, 1970.
4. Kabsch,W. *Acta Cryst.*A32: 922-923, 1976.

5. Kabsch,W. *Acta Cryst.*A34: 827-828, 1978.
6. Tufty, R. M. & Kretsinger, R. H. *Science*, 187: 167-169, 1975.
7. Wierenga, R.K. & Hol,W.G.J. *Nature(London)*. 302: 842-844, 1983.
8. Wierenga,R.K., Terpstra,P. & Hol,W.G.J. *J. Mol.Biol.* 187: 101-107, 1986.
9. Taylor, W. R. *Protein Eng.* 2: 77-86, 1988.
10. Taylor,W.R. *J.Theor.Biol.* 119: 205-218, 1986.
11. Taylor,W.R. & Orengo,C.A. *J.Mol.Biol.* 208: 1-22, 1989.
12. Unger, R., Harel, D., Wherland, S. & Sussman. J.L. *Proteins*, 5: 355-373, 1989.
13. Rooman, M. J., Rodriguez, J.R. & Wodak, S. J. *J.Mol.Biol.* 213: 327-336, 1990.
14. Kabsch,W. & Sander,C. *Biopolymers*, 22: 2577-2637, 1983.

Figure 1. Multiple alignment of 4 βαβ-containing structures

```
(a) Symbol strings
                              +          +          +          +
1ABP  ( 35- 78):    ##ZW-GD5CNNNCCCCNN16L-GWVXV#@@-WR9YCCCNNN
6LDH  ( 20- 65):    ##ZWVWF45NNNNNNNNN168BAUVX##Z@-WGF5NNNNNN
4ADH  (192-236):    ##ZWVWF45NNCCCCCNN16AW-@VX##Z@-WG95CCCNNN
1GD1O(  1- 47):     ##ZWVWF45NNNNCNNCN2A92OXZ@VXZ@GURD5NNNNNN

(b) Secondary structure assingments by DSSP[14]
                              +          +          +          +
1ABP  ( 35- 78):  EEE  -SHHHHHHHHHHHHHT -      B  -S SS TTHHHHHH
6LDH  ( 20- 65):SSEEEEE  SHHHHHHHHHHHTTT  SEEEEE -S HHHHHHHHHHH
4ADH  (192-236):T  EEEE  SHHHHHHHHHHHHHS -  EEEE -S GGGHHHHHHHT
1GD1O(  1- 47): EEEEEE  SHHHHHHHHHHTT SSEEEEEEE SS HHHHHHHHHEE

(c) Amino acid sequences
                              +          +          +          +
1ABP  ( 35- 78):VIKIAVP-DGEKTLNAIDSLAASG-AKGFVICT-PDPKLGSAIVAKA
6LDH  ( 20- 65):YNKITVVGVGAVGMACAISILMKDLADEVALVD-VMEDKLKGEMMDL
4ADH  (192-236):GSTCAVFGLGGVGLSVIMGCKAAGA-ARIIGVD-INKDKFAKAKEVG
1GD1O(  1- 47):AVKVGINGFGRIGRNVFRAALKNPDIEVVAVNDLTDANTLAHLLKYD
```

Figure 2. Selection of representative peptide segments