A Method for Extracting Spatially Close Peptide Segments in Proteins

Zenmei OHKUBO*
zenmei@kuicr.kyoto-u.ac.jp

Minoru KANEHISA*
kanehisa@kuicr.kyoto-u.ac.jp

INSTITUTE FOR CHEMICAL RESEARCH
KYOTO UNIVERSITY
UJI, KYOTO 611 JAPAN

Abstract

In order to predict protein structures from their primary sequences, the understanding of long-range interactions is one of the most critical points. We are dealing with this problem by focusing on the pairs of peptide segments which are separated in the primary sequence but are close in the three-dimensional structure. The method is applied to a set of structure-resolved proteins to see if there are any significant features for association of local structures, such as secondary structure segments. The dataset consists of 88 nonhomologous proteins selected from the Brookhaven Protein Data Bank (PDB) using the superfamily classification of the Protein Information Resource (PIR). In the method, given the definition of the distance between two segments, spatially close segment-pairs are extracted for Ca segments of 4 or 7 residues long. The result shows that there are no preferred distances for association of two helical segments but there is a minimum of twenty intervening residues required for parallel helical segments.

1. INTRODUCTION

Toward understanding the relationship between the amino acid sequence and the three-dimensional structure of a protein, many researchers have investigated the sequence patterns of polypeptide segments and their three-dimensional structures. For example, Argos [1] compared the structures of

* 大久保善明、金久　實
京都大学化学研究所、〒611 京都府宇治市五ヶ庄

penta-peptides which had at least four identical residues at the same positions. Sternberg and Islam [2] compared the structures of peptides having more than twenty residues. Sander and Schneider [3] inspected the threshold of sequence similarity sufficient for structural homology and found the threshold depended strongly on the length of the sequence alignment. Matsuo and Kanehisa [4] converted amino acid sequences into symbol strings and made comparison among them to detect structural motifs.

These previous works dealt with single segments consisting of sequential residues and indicated that short segments with identical amino acid patterns could take different structures. This presents a problem when trying to effectively predict secondary structures of proteins from their sequences. To predict protein structures, information gathered from residues which are separated in the primary sequence but spatially close is indispensable. Alexandrov *et al.* [5] investigated several protein backbone fragments which were separated in the primary sequence. However, the fragments were not always spatially close in their work.

In this work, we focus on a pair of short segments which are separated in the primary sequence but spatially close to each other. We call it a segment-pair. The main purpose here is to develop a methodology to identify and characterize segment-pairs, and find examples of segment-pairs serving important biological functions.

2. METHODS

2.1. Selection of a Non-homologous Set of Proteins

The proteins are all taken from the April 1992 release of the Brookhaven Protein Data Bank (PDB) [6]. According to the procedure described below, we select a non-homologous set of proteins. The data set contains a total of 88 proteins, comprising 16,713 amino acid residues.

a. PDB-PIR Cross Reference

In order to make a non-redundant data set, we first create the PDB-PIR cross reference. The sequences of PDB entries are searched for similar sequences in the PIR release 33 with the FASTA program [7, 8, 9].

Thus, each of the PDB entries is matched with an identical or most similar PIR entry. The results are collected as the PDB-PIR cross reference.

b. Selection of the Representatives from Superfamilies

One representative is selected from each superfamily. In order to make the process as automatic as possible, prospective representatives of the superfamilies are first screened using the following criteria:

(i) if an entry contains the coordinates of only backbone or Cα atoms, it is excluded;
(ii) NMR-resolved entries are excluded;
(iii) if an entry lacks resolution or R-factor values, it is excluded;
(iv) entries with R-factor values more than 0.30 are excluded.

Then the remaining entries are assigned penalties based on six factors shown in **TABLE I**.

TABLE I. Factors and Penalties

Factor	Penalty
resolution	resolution value [Å]
R-factor	R-factor value
complex formation	0.20
mutant	0.10
chain break in the middle	0.10
chain break on the end	0.05

Lastly the entry with the lowest penalty, meaning most reliable in our definition, is selected from each superfamily except with its score larger than 3.0 (listed in **TABLE II.**).

TABLE II. The Data Set of Known Protein Structures Used

1YCC	2CDV	2CCYA	256BA	3B5C	2CPP
4FD1	1HIP	5RXN	2TRXA	2TRXB	1PCY
4FXN	8ADH	6LDH	7ICD	1GOX	1GD1O
8DFR	3GRS	2CYP	8CATA	1GP1A	1PHH
1FNR	3TMS	6AP1A	3CLA	3PFK	3AKD
1BP2	1SNC	1RNH	3RNT	7RSA	1LZ1
3LZM	6CPA	2SGA	4PTP	1CSEE	9PAP
1PSG	6TMN	3BLM	1ALD	2CTS	1CA2
1YPIA	2TS1	5PTI	2OVO	1CSE	1TABI
4SGBI	1HOE	5P21	1XY1A	4INSA	4INSB
1TNFA	3EBX	2MLTA	2I1B	2RHE	2MCG1
3HLAA	1MBC	2MHR	1UBQ	1CTF	1GCR
4CPV	1IFB	1MSBA	1RBP	1UTG	2LTNA
2LTNB	9WGAA	2LIV	8ABP	2GBP	2WRPR
2CRO	2RSPA	1HRHA	2GN5		

2.2. Selection of Segment-Pairs

A segment-pair is a pair of Cα segments which are spatially close but linearly apart on the sequence. To simplify the collection and management of data, only the coordinates of Cα atoms are considered because the positions of atoms in side-chains can be reconstructed from those of Cα atoms [10]. We select segment-pairs from the data set of the 88 proteins as follows.

a. Measure of Distance Between Two Segments

In order to select segment-pairs, the distance between two segments needs be defined. In **Fig. 1** Cα atoms in Segments A and B are designated a1~a4 and b1~b4, respectively, where the numbering starts from the N-terminus and the number of residues L in a segment is four. Let d_{aibj} be the Euclidean distance between Cα atoms ai and bj ($1 \leq i, j \leq L$) and d_{cAcB} be the Euclidean distance between the center of mass of Segment A and that of Segment B.

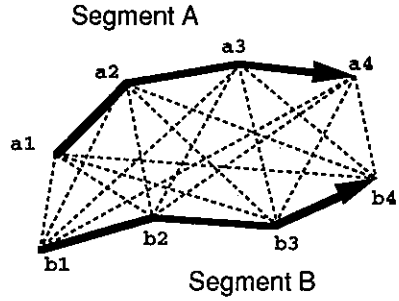

Fig. 1. An illustration of a segment-pair.

The following D_c and D_{rms} are used to calculate distances between Segments A and B:

(i) the distance between the two segments' centers of mass

$$D_c = d_{c_A c_B}, \tag{1}$$

(ii) the root mean square distance

$$D_{rms} = \sqrt{\frac{1}{L^2}\sum_{j=1}^{L}\sum_{i=1}^{L} d_{aibj}^2}. \tag{2}$$

b. Collection of Segment-Pairs from the Data Set

The length of a segment L is fixed at 4 or 7 in this study, corresponding to one or two turns of an α helix.

The calculations of Dc and Drms are carried out between all possible combinations of two segments in a sequence. Segment-pairs are collected if the Dc- or Drms-distance is below a given cut-off value (16.0Å in this study).

It is often the case that adjacent, overlapping segment-pairs are collected by the above procedure. For example, the same segment on a sequence is close to several segments at overlapping positions. In such cases, the segment-pair with the smallest Dc (or Drms) value is retained and others are excluded.

c. Characterization of Segment-Pairs

The selected segment-pairs are associated with four types of data: the number of intervening residues between Segments A and B ("NIR"), secondary structures, Dc or Drms value, and d_{NC}, an index of relative chain direction.

The secondary structures are computed from the coordinates using the DSSP program [11]. Four classes are considered here: 'E' (β-strand), 'H' (helix), 'T' (turn and bend), and 'X' (others). Thus, segment-pairs are classified into ten groups: 'EE', 'EH', 'ET', 'EX', 'HH', 'HT', 'HX', 'TT', 'TX', 'XX'.

The index of relative chain direction d_{NC} is calculated from the first and last Cα atoms of two segments in a segment-pair:

$$d_{NC} = d_{a1bL}+d_{aLb1}-d_{a1b1}-d_{aLbL} \quad . \qquad (3)$$

This parameter indicates orientations, such as vertical ($d_{NC}\sim 0$), parallel ($d_{NC}>0$), and anti-parallel ($d_{NC}<0$), of segment-pairs.

3. RESULTS AND DISCUSSION

3.1. Non-homologous Set of Proteins

We have developed a computerized procedure for selecting a reliable data set of PDB entries. This method of selection is likely to meet the same standard of reliability as that of an expert. Non-experts can easily repeat this computerized selection procedure and arrive at the same results.

NMR-resolved entries are excluded before the selection, because the comparison between the reliability of NMR-resolved entries and that of X-ray-

resolved ones is difficult. It will soon become necessary to modify our method to include NMR-resolved entries, as well as to improve current criteria of selection.

3.2. Distance of Segment-Pairs

Both Dc and Drms have been defined and used as measures of the distance between two segments. Similar patterns are seen in the plots of frequencies against Dc and Drms values, although the absolute values of Dc and Drms are different. When the longer segment length, L = 7, is used there seems no significant difference in the result in comparison to the result with L = 4.

The absolute value of Drms is nearly proportional to the segment length L, while that of Dc is more or less constant. When L is set at much larger values, Drms may no longer be proportional to L and the difference of using Dc and Drms values may become perceptible.

3.3. Number of Intervening Residues

Two parallel α-helices seem to require their intervening sequence at least 20 residues long for proper positioning. If the definition of the "magic number" suggested by de Gennes [12] is generalized into the "minimum length of spacer sequence that has a segment-pair at both ends", the magic number of two parallel α-helices is likely to be 20 (**Fig. 2**). We do not observe any other magic numbers in antiparallel HH-segment-pairs nor EE-segment-pairs.

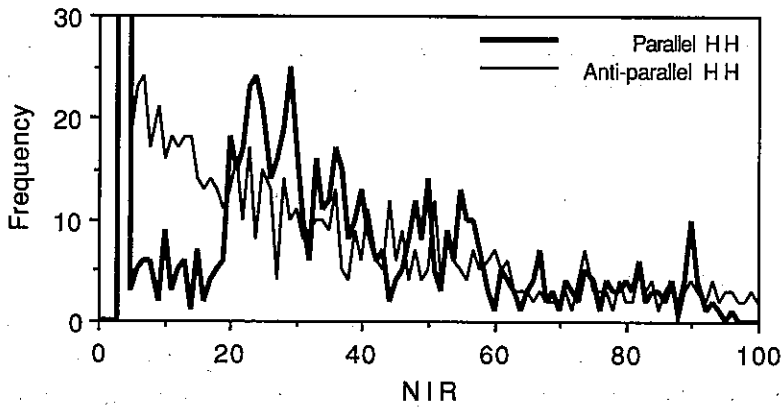

Fig. 2. The observed frequencies of HH segment-pairs plotted against the number of intervening residues (NIR). The distance was Drms-measured and the segment length L was 4. Note that there is a sharp increase around NIR of 20 for parallel HH-segment-pairs.

3.4. Segment-Pairs Having PROSITE Motifs

We have compared our collection of segment-pairs with PROSITE [13] and constructed a library of segment-pairs containing biologically important sequence motifs. It is sometimes observed that a single segment-pair has two different motifs. Segment-pairs having two specific motifs are found in phospholipase A2, uteroglobin, carboxypeptidase A, and papain. As more information is added and our procedure to extract segment-pairs is refined, we can construct a useful library for searching portions of amino acid sequences which have biologically important roles and are spatially close to each other.

3.5. ...and Perspective

We plan to make modifications of the current method and perform more extensive analysis. For example, we would like to link two or more segment-pairs which share the same segment(s) into a "segment-complex". It is a complex consisting of more than two segments and may contain structural motifs among them.

In order to find the structural motifs, segment-pairs and segment-complexes need be classified not according to their secondary structures as in section 2.2.c but according to their three-dimensional local structures. We are trying to develop a method for such classification. We hope, eventually, to find any relationships between structural motifs and sequence patterns of segments, which can be utilized as constraints on possible topologies when predicting three-dimensional protein structures.

3.6. Conclusion

The Following are the observations we have at present:

(i) β-strands are arranged at fixed distances;
(ii) there are no preferred distances for association of helical strands;
(iii) parallel helical strands require at least 20 residues separating them;
(iv) anti-parallel helical strands and parallel and anti-parallel β-strands do not have such a "magic number".

ACKNOWLEDGMENTS: This work was supported in part by the grant-in-aid for scientific research on the priority areas "Genome Informatics" from the Ministry of Education, Science and Culture. Computing resources were provided by the Supercomputer Laboratory, Institute for Chemical Research, Kyoto University.

REFERENCES

[1] Argos,P. (1987) *J. Mol. Biol.* **197**, 331-348.

[2] Sternberg,M. and Islam,S. (1990) *Protein Engng* **4**, 125-131.

[3] Sander,C. and Schneider,R. (1991) *Proteins* **9**, 56-68.

[4] Matsuo,Y. and Kanehisa,M. (1993) *CABIOS* **9**, 153-159.

[5] Alexandrov,N., Takahashi,K., and Go,N. (1992) *J. Mol. Biol.* **224**, 5-9.

[6] Bernstein,F.C., Koetzle,T.F., Williams,G.J.D., Meyer,E.F., Brice,M.D., Rodgers,J.R., Kennard,O, Shimanouchi,T and Tasumi,M. (1977) *J. Mol. Biol.* **112**, 535-542.

[7] Dayhoff,M.O., Barker.W.C., and Hunt,L.T. (1983) *Methods Enzymol.* **91**, 524-545.

[8] Barker,W.C., George,D.G., and Hunt,L.T. (1990) *Methods Enzymol.* **183**, 31-49.

[9] Pearson,W.R. and Lipman,D.J. (1988) *Proc. Natl. Acad. Sci.* **84**, 4355-43586.

[10] Levitt,M. (1992) *J. Mol. Biol.* **226**, 507-533.

[11] Kabsch,W. and Sander,C. (1983) *Biopolymers* **22**, 2577-2637.

[12] de Gennes,P. (1992) *Science* **256**, 495-497.

[13] Bairoch,A. (1992) *Nucleic Acids Res.* **20**, 2013-2018.