Bull. Inst. Chem. Res., Kyoto Univ., Vol.71, No.3, 1993

Characterization of Spatially Close Peptide Segments in Proteins

Zenmei OHKUBO* and Minoru KANEHISA*

Received August 10, 1993

A method of extracting pairs of peptide segments which are separated in the primary sequence but are close in the three-dimensional structure is developed. First, 88 nonhomologous proteins are selected as a representative data set from the Brookhaven Protein Data Bank (PDB) using the superfamily classification of the Protein Information Resource (PIR). Then the Cα segments of 4 or 7 residues long are examined. Given a measure of the distance between two segments and a cut-off value for the distance, spatially close segment-pairs are extracted from the data set. The occurrences of segment-pairs are investigated in relation with the secondary structure types and the number of residues intervening between the two segments. It is found that two β-sheet segments are arranged at fixed distances due to inter-segment hydrogen bonding. There are no preferred distances for association of two helical segments, but there is a minimum number of intervening residues required for parallel helical segments. In addition, a library of segment-pairs which correspond to functional motifs defined in PROSITE is constructed.

KEY WORDS: Protein data bank / Protein superfamily / Sequence motif / Structural motif / Active site / Cα segment

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 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.

* 大久保善明, 金久　實: Laboratory of Molecular Biology and Information III, Institute for Chemical Research, Kyoto University, Uji, Kyoto 611

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. A segment is defined by Cα atoms of four or seven sequential residues. The distance between two Cα segments is defined either by the center of mass of Cα positions or by the root mean square of all Cα pairs. If the distance is lower than a preset cut-off value, the pair of Cα segments is regarded as being close and, as a result, retained as a segment-pair.

All data are extracted from the Brookhaven Protein Data Bank, a database of protein structures. We are interested in such information as the number of intervening (or spacer) residues between two segments and the specific secondary structures that constitute segment-pairs. de Gennes[6] suggested the existence of a "magic number", the number of spacer residues among peptide segments that form an active site. This is quite an interesting proposition. Not only residues serving active sites, but those constructing specific structures possibly need a certain length of spacer for their relative positioning.

In addition to investigating statistically significant features in the formation of segment-pairs, an actual library of segment-pairs which correspond to functional motifs of PROSITE is constructed. This library can be used to search spatially close portions of sequences which have known 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)[7]. The entire PDB is not used, because it contains entries with various resolutions and some of them are dupulicates or minor modifications of the same proteins. According to the procedure described below, we select a non-homologous set of proteins, which is listed in **TABLE I.** The data set contains a total of 88 proteins, comprising 16,713 amino acid residues.

a. PDB-PIR Cross Reference

The Protein Information Resource (PIR) is a database of known protein sequences. Entries in a portion of PIR called PIR1 are grouped into superfamilies according to their evolutionary origins at the molecular level[8]. The assignment of superfamilies is determined by sequence similarity with statistical checks and experts' knowledge[9]. In general, entries in different superfamilies have different ancestral protein molecules. The PDB entries, unlike the PIR entries, are not classified according to their evolutionary origins.

In order to make a non-redundant data set, we first create the PDB-PIR cross reference. When a PDB entry consists of multiple polypeptide chains, they are divided into separate sequence entries. In the case of homomultimer proteins, only one subunit is used as an entry. Monomers in the same heteromultimers are considered as individual entries. Enzymes and their inhibitors are designated as separate entries. The sequences of these detached PDB entries are searched for similar sequences in the PIR1 release 33 with the FASTA program[10].

TABLE I. 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		

The above 88 proteins were extracted from the Brookhaven Protein Data Bank[7], release 60 (Apr. 1992). The entry name is the identification code used in the PDB followed by the chain identifier.

Thus, each of the PDB entries is matched with an identical or most similar PIR1 entry. If there are multiple PIR1 entries with scores very near to the highest one, the local sequence alignment is carried out between the PDB sequence and each of the PIR1 sequences. The most similar entry is chosen considering the alignment results and the information about entries written in the PIR1, such as TITLE, SOURCE, and REFERENCE. The results are collected as the PDB-PIR cross reference.

b. Selection of the Representatives from Superfamilies

In order to make a reliable, non-redundant data set, a representative is selected from each superfamily. This is usually a cumbersome process requiring the help of human experts. 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.

TABLE II. 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

Then the remaining entries are assigned penalties based on six factors shown in **TABLE II.** Resolution and R-factors are most important, but we prefer entries without complex formation, with identical sequence, and without any missing coordinates. The entry with the lowest penalty, meaning most reliable in our definition, is selected from each superfamily.

In case all entries in a superfamily have penalties larger than 3.0, no representative is selected. Most of the entries with penalties larger than 3.0 have resolution values larger than 2.5 and R-factor values larger than 0.20, which indicate that the coordinates may be unreliable.

As the result of this computerized selection, the 88 proteins shown in **TABLE I** are retained.

2.2. Selection of Segment-Pairs

A segment is composed of several sequential amino acid residues in a protein sequence. To simplify the collection and management of data, only the coordinates of Cα atoms are considered. This treatment is not an over-simplification, because the positions of atoms in side-chains can be reconstructed from those of Cα atoms[11]. A segment-pair is a pair of Cα segments which are spatially close but linearly apart on the sequence. 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 daibj be the Euclidean distance between Cα atoms ai and bj ($1 \leq i,\ j \leq L$) and dc_Ac_B be the Euclidean distance between the center of mass of Segment A and that of Segment B. The following Dc and Drms are used to calculate distances between Segments A and B:

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

$$D_C = d_{CACB} \qquad (1)$$

(ii) the root mean square distance

$$\mathrm{Drm} = \sqrt{\frac{1}{L^2}\sum_{j=1}^{L}\sum_{i=1}^{L} d^2_{a_ib_j}} \ . \qquad (2)$$

Both Dc and Drms are likely to be proportional to the separation of Segments A and B, but Drms may be more sensitive than Dc to the shapes of the two segments in the three-dimensional structure. We calculate both Dc and Drms to see if there are any such effects.

b. Collection of Segment-Paris from the Data Set

The length of a segment L is fixed at 4 or 7 in this study. A four or seven residue segment corresponds to one or two turns of an α-helix.

As illustrated in **Fig. 2** the calculations of Dc and Drms are carried out between all possible combinations of two segments in a sequence. The combinations are made in a way such that the two segments are separated by at least L residues in each sequence. Segment-pairs are collected if the distance (Dc or Drms) is below a given cut-off value. The initial cut-off value is set at 16.0Å.

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

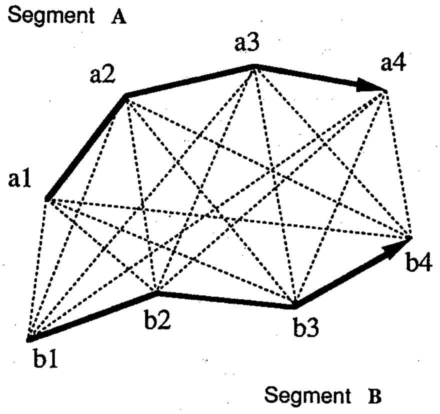

Fig. 1. An illustration of a segment-pair. Segments A and B, each with four residues long, in a protein sequence are represented by solid lines. Cα atoms in Segments A and B are numbered a1 ~ a4 and b1 ~ b4 in the direction of N-terminus to C-terminus. The dotted lines indicate inter-Cα distances used in equations (1) and (2).

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[12]. Four classes are considered here: 'e' (β-strand), 'h' (3/10, α-, and π-helix), 't' (turn and bend), and 'x' (others). The DSSP program made classifications on the residue level and we expanded to the segment level. When L is 4, the segment classification is performed using the following criterion:

(i) if a segment has three (five, when L is 7) or more residues assigned 'e', the segment is assigned 'E';

(ii) if a segment has three (six, when L is 7) or more residues assigned 'h', the segment is assigned 'H';

(iii) if a segment has two (also two, when L is 7) or more residues assigned 't', the segment is assigned 'T';

(iv) if a segment is not assigned 'E', 'H', or 'T', it is assigned 'X'.

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:

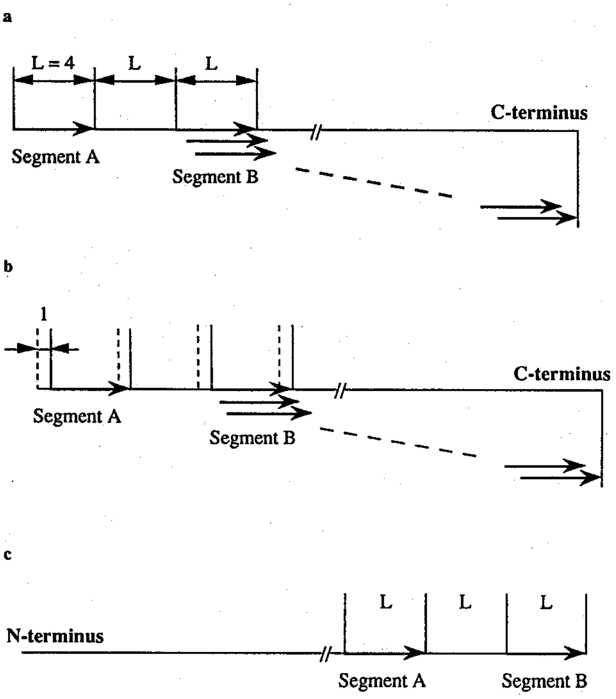

Fig. 2. The collection of segment-pairs from a protein sequence. The long horizontal line indicates the whole sequence of a protein. **(a)** Segment A is at residues 1～L. Segment B is shifted from residues (2L+1)～3L to the last four residues. For each pair of A and B segments, Dc and Drms are calculated, and the pair is retained as a segment-pair if the distance is less than the cut-off value. **(b)** Segment A is moved to residues 2～(L+1). Segment B is generated from residues (2L+2)～(3L+1) to the last segment, calculating Dc and Drms as in step **(a)**. **(c)** The shifts and calculations are repeated until Segment A reaches its right-most position (i.e. residues 3L～(2L+1) counted from the C-terminus).

$$d_{NC} = d_{albL} + d_{aLbl} - d_{albl} - d_{aLbL}. \tag{3}$$

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

3. RESULTS

3.1. Distance of Segment-Pairs

Fig. 3 shows the number of segment-pairs observed when L = 4 at various Dc (or Drms) values for different secondary structural groups, 'EE', 'EH', and 'HH'. EE-segment-pairs have very sharp peaks in both plots, while EH- and HH-segment-pairs are spread out over wide ranges. The peaks of EE-segment-pairs at 5.0Å in **Fig. 3(a)** and at 7.2Å in **Fig. 3(b)** reflect that most EE-segments are portions of β-sheets with hydrogen bondings among them. These structures appear to be restrained far more than other kinds of segment-pairs.

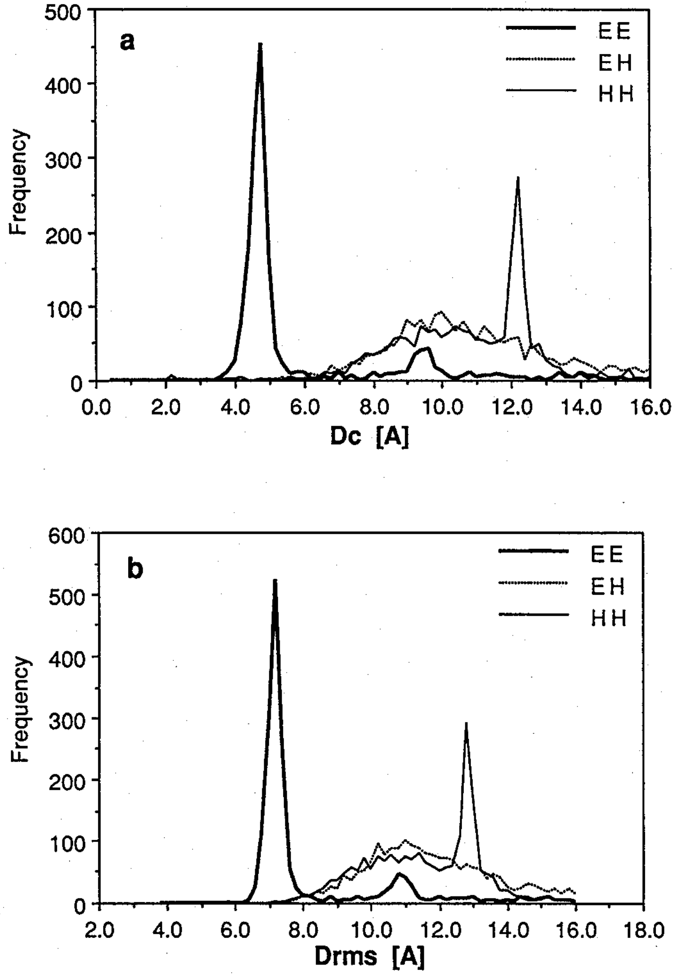

Fig. 3 The observed frequencies of segment-pairs in 'EE', 'EH', and 'HH' groups plotted against the distance in **(a)** Dc and in **(b)** Drms. The segment length was: L = 4. 'EE', thick solid line; 'EH', dotted line; and 'HH', thin solid line.

Typical distance between parallel β-sheets is 5.0Å and that of anti-parallel β-sheets is a little less than 5.0Å. We expected that EE-segment pairs would be double-peaked corresponding to the two distances of parallel and anti-parallel sheets. However, this was not the case. This is probably due to the fact that twists and bends in segments influence their Dc or Drms values more than the difference of ideal forms of parallel and anti-parallel β-sheets. EE-segment-pairs also have tiny peaks at 9.6Å in **Fig. 3 (a)** and at 10.8Å in **Fig. 3 (b)**. It turned out that these tiny peaks contain segment-pairs composed of two selected segments with a third segment in between. Such segment-pairs were sometimes selected, especially from β-sheet structures.

HH-segment-pairs have sharp, but lower peaks at 12.2Å in **Fig. 3 (a)** and at 12.8Å in

Fig. 3(b). These lower peaks contain segment-pairs composed of two segments in the same α-helices. That is, there are 660 HH-segment-pairs between 11.5Å and 12.5Å of Dc in **Fig. 3(a)**. Among them, 580 are segment-pairs in which two segments are four residues apart in the primary sequence and belong to the same helix. In **Fig. 3(b)**, there are 664 HH-segment-pairs between 12.5Å and 13.5Å of Drms, and 599 of them are segment-pairs in which two segments are four residues apart and are on the same helix.

Therefore, the peaks observed in **Fig. 3** are either trivial ones or artefacts. We examined other combinations of secondary structures, as well as all structures combined. However, we did not observe any other peaks when the distribution is plotted against the distance Dc or Drms.

3.2. Number of Intervening Residues

Fig. 4 shows the distributions of segment-pairs plotted against the number of intervening residues, or the NIR value, of parallel ($d_{NC}>0$) segment-pairs (L = 4, Dc-measured). It is noteworthy that there is a sharp increase around the NIR of 20 for parallel HH-segment-pairs, from a low plateau below 20. The highest peak at the smallest NIR values corresponds to the segments on the same helix. This implies that two parallel α-helices require their intervening sequence at least 20 residues long for proper positioning (**Fig. 5**). We did not observe the magic number[6] as de Gennes suggested. However, if the definition of magic number 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.

We also examined anti-parallel HH-segment-pairs and both parallel and antiparallel

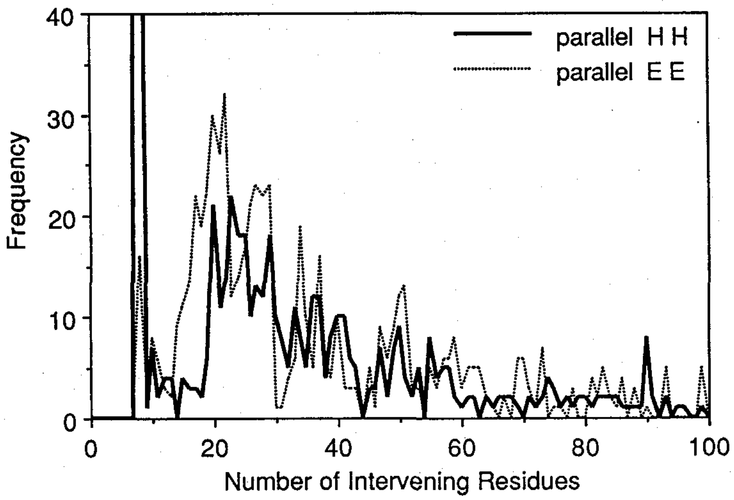

Fig. 4 The observed frequencies of parallel ($d_{NC} > 0$) segment-pairs plotted against the number of intervening residues (NIR). The distance was Dc-measured and the segment length was: L = 4. 'EE', dotted line; and 'HH', solid line.

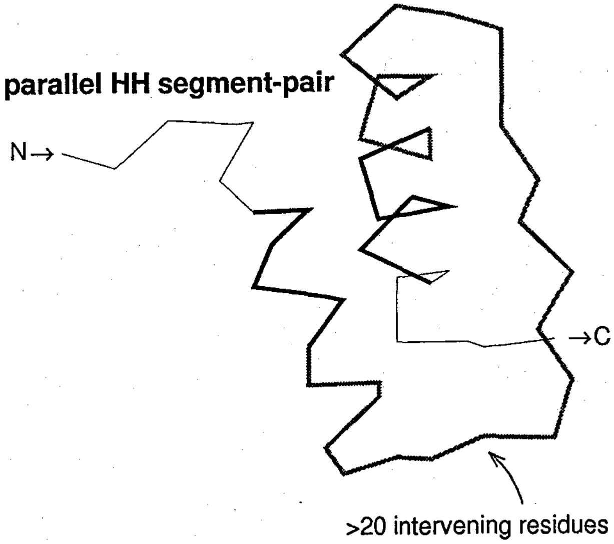

Fig. 5 An example of two parallel α-helical segments (bold lines) and the intervening sequence (shaded line).

EE-segment-pairs. However, we did not observe any other magic number.

3.3. Segment-Pairs Having PROSITE Motifs

PROSITE[13] is a database of biologically significant sites and patterns in protein sequences, which are experimentally known and computationally refined. We have compared our collection of segment-pairs with PROSITE 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 (1BP2 in the PDB identifier), uteroglobin (1UTG), carboxypeptidase A (6CPA), and papain (9PAP). Segment-pairs of 1BP2 and 9PAP have active sites on both segments. The stereo view of the segment-pair in carboxypeptidase A is shown in **Fig. 6**.

In a sense, we added three-dimensional annotation to PROSITE, which is often used to predict functional properties from the amino acid sequence. 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.

4. DISCUSSION

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

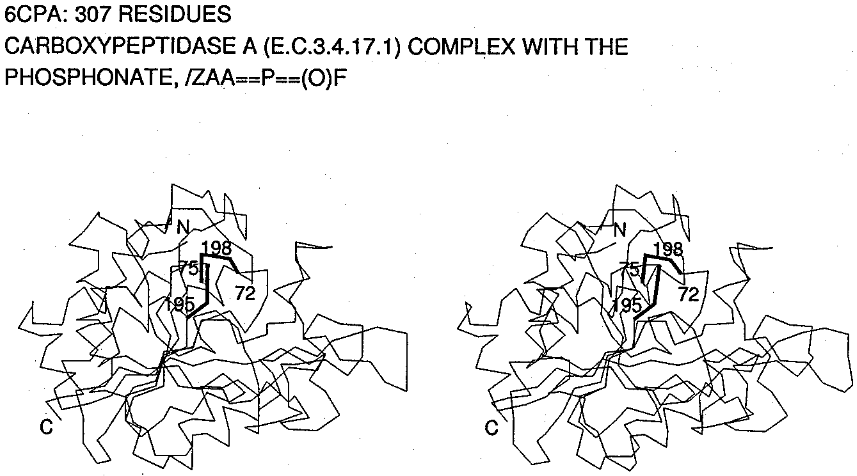
6CPA: 307 RESIDUES
CARBOXYPEPTIDASE A (E.C.3.4.17.1) COMPLEX WITH THE
PHOSPHONATE, /ZAA==P==(O)F
N
198
75
195
72
C
N
198
75
195
72
C

Fig. 6 Stereo view of functionally important segment-pairs in carboxypeptidase A (6CPA in the PDB identifier). Segment-pairs are shown in bold lines and the numbers represent the residue numbers on both ends.

an expert. Non-experts can easily repeat this computerized selection procedure and arrive at the same results. In the present study the procedure yielded a non-redundant and reliable data set of 88 proteins. We compared this with an expert's manual selection and found that eight proteins were different, which bear little significance in view of reliability (T. Noguchi, personal communication).

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. However, the number of NMR-resolved entries in the PDB is growing rapidly. They will soon be a major group among the entries. It will become necessary to modify our method to include NMR-resolved entries, as well as to improve current criteria of selection.

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 (**Figs. 3**), althouth the absolute values of Dc and Drms are different. When the longer segment length, L = 7, was used there seemed no significant difference in the result (data not shown) in comparison to the result with L = 4. The absolute value of Drms was nearly proportional to the segment length L, while that of Dc was more or less constant. When L is set at much larger values, Drms may no longer be proportional to L because it is more affected by the three-dimensional configuration of Cα atoms, and the difference of using Dc and Drms values may become perceptible.

A method for excluding redundant segment-pairs has been introduced in section **2.2.b.** This method is somewhat arbitrary and still requires further refinement. We plan to make modifications of the current method and perform more extensive analysis.

In conclusion we found the following:

(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., *J. Mol. Biol.* **197**, 331-348 (1987).
(2) Sternberg,M. and Islam,S., *Protein Engng* **4**, 125-131(1990).
(3) Sander,C. and Schneider,R., *Proteins* **9**, 56-68(1991).
(4) Matsuo,Y. and Kanehisa,M., *CABIOS* **9**, 153-159(1993).
(5) Alexandrov,N., Takahashi,K., and Go,N., *J. Mol. Biol.* **224**, 5-9(1992).
(6) de Gennes,P., *Science* **256**, 495-497(1992).
(7) 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., *J. Mol. Biol.* **112**, 535-542(1977).
(8) Dayhoff,M.O., Barker.W.C., and Hunt,L.T., *Methods Enzymol.* **91**, 524-545(1983).
(9) Barker,W.C., George,D.G., and Hunt,L.T., *Methods Enzymol.* **183**, 31-49(1990).
(10) Pearson,W.R. and Lipman,D.J., *Proc. Natl. Acad. Sci.* **84**, 4355-43586(1988).
(11) Levitt,M., *J. Mol. Biol.* **226**, 507-533(1992).
(12) Kabsch,W. and Sander,C., *Biopolymers* **22**, 2577-2637(1983).
(13) Bairoch,A., *Nucleic Acids Res.* **20**, 2013-2018(1992).