タンパク質のCα鎖セグメント・パッキングの分類

P11

A New Method for Finding Packing Patterns of Protein Cα Segments

大久保　善明　　金久　實
OHKUBO Zenmei　　Minoru Kanehisa
京都大学・化学研究所
Institute for Chemical Research, Kyoto University
E-mail: zenmei@kuicr.kyoto-u.ac.jp

We have developed a new method for finding packing patterns of Cα segments. The method allows detecting in PDB (Protein Data Bank) several Cα segments which are spatially local but are apart on a sequence from one another. The data of those segments ("segment-sets") is an expression of the long distance interactions between amino acid residues and will be a great help to prediction of protein structures from their sequences.

The method consists of the following steps:

1. PDB entries are classified under PIR (Protein Identification Resource) super-families. From each of the families, a single representative PDB entry is selected mainly based on its high resolution and good R-value. 88 representatives are determined and used as a data set ("refined data set").
2. A distance metric, D, is defined between Cα segments in order to extract spatially local segment-sets from the refined data set.
3. An extention of Kabsch's method (1979) is employed to find out several segment-sets which have similar packing patterns to that of a query segment-set.
4. The statistics of the residual length between any two segments in one segment-set is also investigated.

タンパク質の高次構造を推定するためには、配列上は離れているにも拘わらず近接する残基の情報が必要となる。そこで我々は、ＰＤＢのなかから配列上はローカルではないが空間的にはローカルなCαセグメントの組を抽出・解析して、幾つかのタンパク質に共通に存在するパッキング・パターンを探している。まず、PDBのエントリーをPIRのスーパーファミリーで分類し、スーパーファミリー毎に１エントリーを選択することにより、構造に偏りのないデータセットを作った。次に、Cαセグメント間の距離Dを定義し、前記データセットよりDが一定範囲内の値を取るセグメントの組を抽出した。これらセグメントの組のなかから、r.m.s.d.の値に基づき類似のパッキング・パターンを持つセグメントの組同士を求めた。さらに、各セグメントの組から、セグメント間の配列上の距離（残基数）の統計をとってその結果を考察した。

I. データセット

まず、タンパク質の立体構造データベースであるPDB(rel. 60)のエントリーをチェイン毎に別々のエントリーとした。その各エントリーをタンパク質の配列データベースであるPIR(rel. 33)に対してFASTAで検索に掛け、対応するエントリーを決定した。それらをまとめてPDBのチェイン毎のエントリーとPIR のエントリーとのクロス・リファレンスとした。PIRのエントリーはスーパーファミリーに分類されているものが多いので、このリファレンスを用いてPDBのエントリーをPIRのスーパーファミリーに分類した。次に、各スーパーファミリーより代表の１エントリーを決定した。決定に際しては、X線結晶解析によるもので側鎖の座標まで与えられているエントリーを候補とし、その中から解像度・R因子・修飾の有無を考慮して決めた。解像度等の良いエントリーの無いスーパーファミリーからは代表を選ばなかった。その結果、88個の代表が得られ、これらをデータセットとした(Fig. 1)。

<Fig. 1>

PDB - PIR Cross Reference

1	155C	CCPC50	cytochrome c
1	1C2RA	CCRF2C	cytochrome c
1	1C2RB	CCRF2C	cytochrome c
.	.	.	.
.	.	.	.
2324	2TMVP	VCTMVU	tobacco mosaic virus coat protein
2357	2STV	VCTNS	satellite tobacco necrosis virus
2358	4SBVA	VCBW	southern bean mosaic virus
2420	2GN5	DDBPF1	class I filamentous phage

Select one PDB entry from each superfamily

The refined data set (88 proteins)

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	3ADK	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

II. セグメント間の距離Dの定義

１つのタンパク質分子内で近接してパッキングされているCα鎖セグメントの組を調べるに当たり、パラメターとしてセグメント長L とセグメント数N を考えた。そして、予めL, Nを一定値に固定しておき、セグメント同士が全体として近接している指標として セグメント間の距離Dを以下のように定義した：

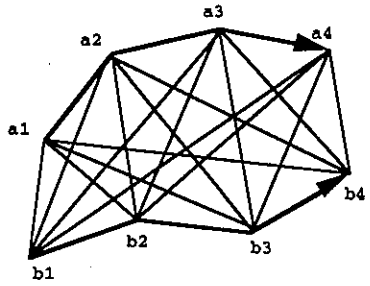
$$D = \sqrt{\frac{1}{{}_NC_2 \cdot L^2} \sum_{k=1}^{N-1} \sum_{l=k+1}^{N} \sum_{i=1}^{L} \sum_{j=1}^{L} d_{k_i l_j}^2}$$

但し、セグメントにN末側より順に1, 2, .., N と番号を付した。同様に、セグメント内のCα原子をN末側より1, .., L とした。d_{kilj} はセグメントkの原子 i とセグメントlの原子 j との距離である。つまり、Dは セグメント間のすべての組み合わせのCα原子間距離の自乗平均距離である。

L = 4, N = 2 の下図の場合では、セグメント間に引いた16本の線分の１本当たりの自乗平均距離である。

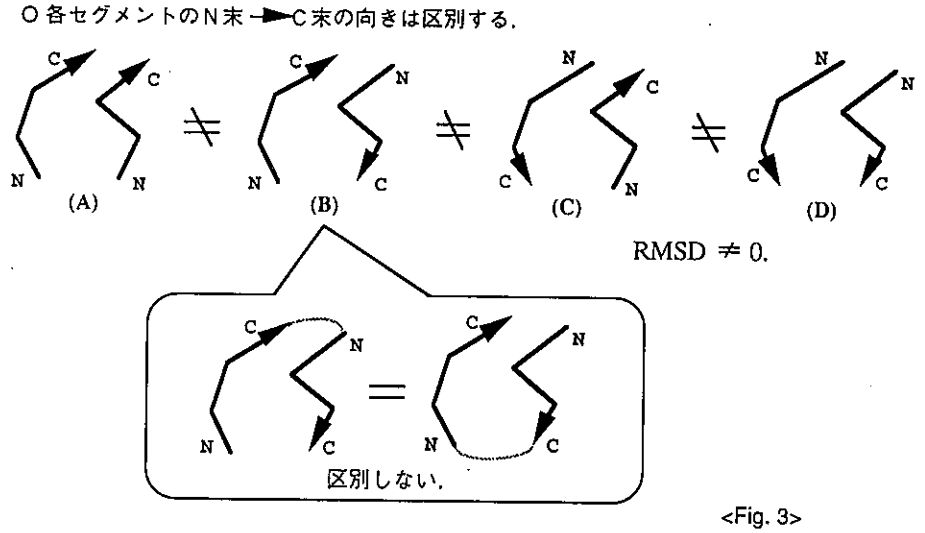
< Fig. 2 >

$$D = \sqrt{\frac{1}{4^2}\sum_{j=1}^{4}\sum_{i=1}^{4} d^2_{aibj}}$$



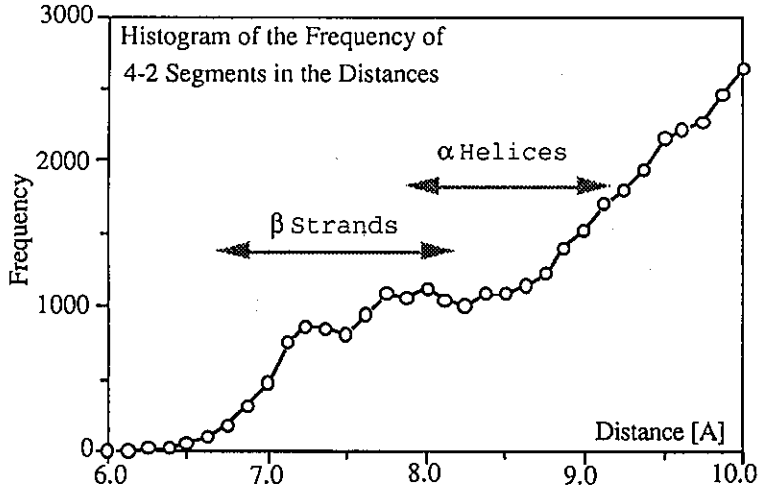
III. r.m.s.d.の計算法

セグメントの組同士の構造を比較するには、Kabschの方法(1978)を用いた。即ち、比較する２つの構造の重心を一致させた条件の下で、対応する原子間距離のr.m.s.d.の最小値を構造の類似度の指標とした。この際、各セグメントの向きは区別したが、配列上での順序は区別しなかった(Fig. 3)。これにより、配列上でのセグメントの並びにとらわれずに空間内で類似の構造を持つセグメントの組を抽出できることになる。



<Fig. 3>

L = 4, N = 2 の場合に、0.125幅に区切ったDに対するセグメントの組の頻度のプロットを Fig. 4 に示す。Dの値によって 2 次構造の異なるセグメントの組がおおよそ分離されている。



<Fig. 4>

類似のセグメントの組とr.m.s.d.の値とは、以下のような形式で表示するようにし、またステレオ図の表示も出来るようにした(Fig. 5)。

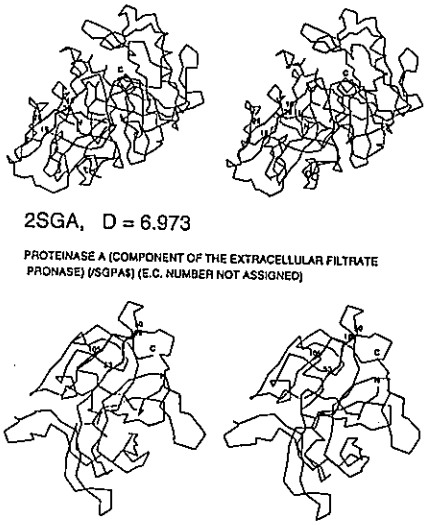
```
3cla.31-34.192-195  1fnr.74-77.94-97       0.4971
3cla.31-34.192-195  1hoe.21-24.53-56       0.4938
3cla.31-34.192-195  1psg.16-19.26-29       0.4763
3cla.31-34.192-195  2sga.50-53.105-108     0.4937
3cla.31-34.192-195  3grs.249-252.264-267   0.4799
3cla.31-34.192-195  5rxn.3-6.49-52         0.4966
3cla.31-34.192-195  9pap.109-112.207-210   0.4252
```

1PSG, D = 6.986

PEPSINOGEN

2SGA, D = 6.973

PROTEINASE A (COMPONENT OF THE EXTRACELLULAR FILTRATE PRONASE) (SGPA5) (E.C. NUMBER NOT ASSIGNED)



<Fig. 5>

* 本研究は文部省科学研究費重点領域研究「ゲノム情報」の助成を受けています。また京都大学化学研究所スーパーコンピュータラボラトリより計算機のCPU timeの提供を受けました。