アミノ酸残基の分布に基づく タンパク質局所構造のモデル化

喜多川 哲, 松田 秀雄, 橋本 昭洋
E-mail: {s-kitagw, matsuda, hasimoto}@ics.es.osaka-u.ac.jp
大阪大学 大学院基礎工学研究科 情報数理系専攻
〒560-8531 大阪府豊中市待兼山町 1 番 3 号 TEL/FAX 06-850-6602

本稿では，本研究室で提案された，タンパク質の立体構造を表すコードを生成する手法を紹介し，これを用いて，アミノ酸配列からタンパク質の立体構造の予測を試みる．本研究では，アミノ酸レベルで局所的に共通なパターンを持つタンパク質数種類に限定した上で，連続するアミノ酸４残基の分布により，どの立体構造コードが割り当てられるのかを調べ，頻度の統計をとった．次にこれを基にし，同じ共通なパターンを持つタンパク質について，立体構造のコード予測を行った．

キーワード：タンパク質立体構造, チェインコード，立体構造予測

A MODELING METHOD OF LOCAL PROTEIN STRUCTURES BASED ON THE DISTRIBUTION OF AMINO ACID RESIDUES

Satoshi Kitagawa, Hideo Matsuda, Akihiro Hashimoto
E-mail: {s-kitagw, matsuda, hasimoto}@ics.es.osaka-u.ac.jp
Department of Infomatics and Mathematical Science,
Graduate School of Engineering Science,
Osaka University,
1-3 Machikaneyama-cho, Toyonaka, Osaka 560-8531 Japan TEL/FAX 06-850-6602

In this paper, we present an encoding method of tertiary protein structures, and using this codes we attempt to predict tertiary protein structures from amino acid residues. In our approach, first, we select proteins which have common local patterns in amino acid sequences as a data set. Second, we compute statistics of tertiary protein structure codes and four successive amino acid residues from the data set. Third, using the statistics we attempt to predict the tertiary structure from an amino acid sequence which has the same common local patterns.

keyword: tertiary protein structure, chain code, tertiary structure prediction

1 はじめに

タンパク質は生体内で様々な機能を果たしているが, その機能は立体構造により決定されると考えられている. タンパク質は多数のアミノ酸が一本鎖状にペプチド結合したポリペプチドであるが, 生体内では, それを構成する２０種のアミノ酸の相互の位置関係の影響で, 安定な状態になるように複雑に折りたたまれ, タンパク質の立体構造が決定される. この意味においては, 立体構造はそれを構成するアミノ酸だけの関数であると言える. タンパク質の立体構造は, 分子進化の過程でもよく保存されており, アミノ酸の配列においてはほとんど相同性がない場合でも, 立体構造は保存されている場合がある. そのため, 立体構造の比較は分子進化の過程を明らかにするのに有用であるが, 現在立体構造の判明しているタンパク質の数は, アミノ酸配列の判明しているタンパク質よりもはるかに少ないことから, アミノ酸配列から立体構造を予測したい要求がある. 以下では, すでに我々が提案したタンパク質の立体構造を文字列を使って表現する方法 [1] （以下, チェインコード）を用いて, タンパク質の同じ部位のアミノ酸配列とチェインコードの対応関係の統計をとり, アミノ酸配列からチェインコード列の予測を試みる. また, その結果を示す.

2 タンパク質の立体構造とその表現法

2.1 タンパク質の構造

タンパク質はアミノ酸のペプチド結合により, 図 1 のような構造をしている. タンパク質は一般に図 1 の N-末端（左側）から C-末端（右側）へと進む向きに記述される. 生体内では, タンパク質の立体構造は, アミノ酸配列上の各アミノ酸の相互の位置関係に依存して, 安定な状態になるように, 複雑に折り畳まれた状態（図２）で存在している.

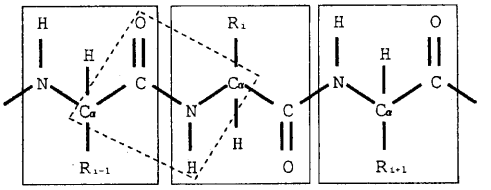

図 1: タンパク質の構造:R_{i-1}, R_i, R_{i+1} 等には, アミノ酸の種類に対応した原子団がつく. 破線で囲んだ部分の原子は同一平面上にある.

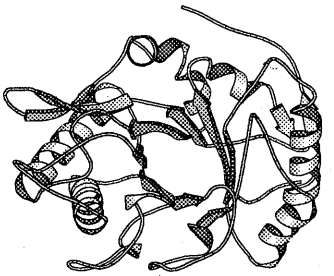

図 2: タンパク質の立体構造の一例

2.2 立体構造のモデル化

各アミノ酸は C_α 原子と呼ばれる, 構造の中心となる炭素原子を持つ. 図 1 で, 破線で囲んだ部分の C 原子と N 原子の間の結合は, 二重結合に近いものとなっており. 破線で囲んだ部分は三次元的に同じ平面上に位置する. この結合が安定なため, 隣合うアミノ酸の C_α 原子間の距離は約 3.8 Å とほぼ一定になっている. 従って, タンパク質の立体構造を, 骨組みだけで大まかに表現すると, 各アミノ酸の C_α 原子を頂点とした折れ線で表現することができる.

次に, アミノ酸配列上の連続する 4 つのアミノ酸中心炭素原子 C_αに注目する. これらの座標を N-末端から C-末端へ順に p_{i-2}, p_{i-1}, p_i, p_{i+1} とする. この時, 3 点 p_{i-2}, p_{i-1}, p_i で決

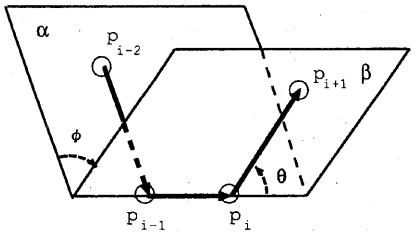

図 3: θ と ϕ の定義:N-末端側から C-末端側へ連続している 4 つのアミノ酸の C_α 原子の三次元座標をそれぞれ p_{i-2}, p_{i-1}, p_i, p_{i+1} とする.

定される平面 α と 3 点 p_{i-1}, p_i, p_{i+1} で決定される平面 β とがなす角をねじれ角 (torsion angle) という. 但し, α を β に重ねる時, 右ねじの進む方向がベクトル $\overrightarrow{p_{i-1}p_i}$ の向きに一致する場合のねじれ角が正であるとする. 本稿ではこの角度を ϕ で表すことにする. また, 2 つのベクトル $\overrightarrow{p_{i-1}p_i}$ と$\overrightarrow{p_ip_{i+1}}$がなす角を$\theta$ で表す. (図 3 参照)

このようにして得られる 1 組の (θ, ϕ) は折れ線の進む向き, すなわちベクトルを与える.

3 チェインコード

3.1 立体構造のコード化手法

前節の内容をふまえ, 本研究室では, タンパク質の立体構造を表すコード化手法を開発した [1]. 基本的な考え方は, となりあうアミノ酸の中心炭素原子間の距離がほぼ一定であるため, タンパク質の構造が (θ, ϕ) 列のみで表現できることによる. (θ, ϕ) を球座標を表す方向ベクトルとみなし, 適当に量子化することにより, 量子化ベクトル列が得られる. さらに, それぞれの量子化ベクトルに文字を割り当てることにより, 量子化ベクトル列から文字列を得ることができる.

図 3 における p_{i-2} p_{i-1} p_i の, 各 C_α 原子について, 次の方法にしたがって, これらを座標上に配置する.

1. p_i を原点に配置

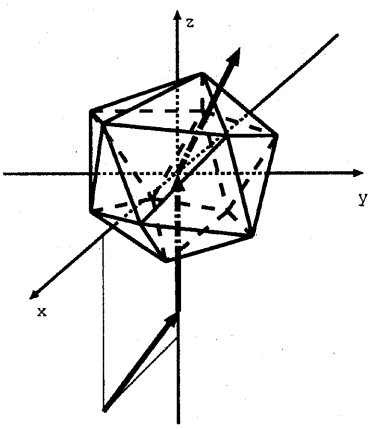

図 4: 重心を原点に一致するように配置された二十面体: 3 つの矢印は回転並進移動をした後の p_{i-2}, p_{i-1}, p_i, p_{i+1}を順に結んだもの.

2. p_{i-1} を z 座標が負となる様な z 軸上に配置

3. p_{i-2} を x 座標が正となる様な zx 平面上に配置

こうした時の, ベクトル $\overrightarrow{p_ip_{i+1}}$ の方向が, 球座標表現 (θ, ϕ) となる.

次に, この座標系に, 重心が原点に一致する様に正二十面体を配置する (図 4). 正二十面体は球に一番近い正多面体であることから, 正二十面体を量子化に用いている. 配置された正二十面体の法線ベクトルの中でベクトル $\overrightarrow{p_ip_{i+1}}$ の方向に最も近いものが, このベクトルに対応する量子化ベクトルとなる.

実際に正二十面体を配置するのには自由度があるが, 立体構造が局所的に規則的な構造 [1]をとる場合に, それらができるだけ同じ文字にコード化されるように, 正二十面体を回転し, 各面の法線ベクトルを決定した. また, 実際のコード化はタンパク質立体構造データベース (PDB)[2] のデータを使用している.

[1]局所的に規則的な構造は二次構造と呼ばれ, 主なものに, らせん状の α ヘリックス, まっすぐに伸びた β ストランドがある.

```
          (A)                (C)
2bb2    WVGYEQANCKGEQFVF   EEEEQCFHKDCCEEEE
1blbA   WVGYEQANCKGEQFVF   EEEEQHEHKDHCEEEE
4gcr    WMLYERPNYQGHQYFL   EEEEQCFHKDCCEEEE
1gcs    WMLYERPNYQGHQYFL   EEEEQCFHKDHCEEEE
2gcr    WMLYEQPNFTGCQYFL   EEEEQCFHKDCCEEEE
1prs    AILYQNDGFAGDQIEV   EEEEEHEHKDHGEEEE
```

図 5: Greek Key モチーフ部分のアミノ酸配列 (A) と チェインコード列 (C) の例

4 立体構造予測手法

4.1 これまでの手法

立体構造予測はこれまでにもさまざまな手法が試みられている．以前から行われてきたのが，立体構造全体を予測するよりもまず，どのタンパク質にも見ることができる，局所的に規則的な構造を予測する手法である．２０種の各アミノ酸残基が，局所的に規則的な構造をどの程度とりやすいかを示す傾向性指数を求め，指数の高い残基が連なる部分を中心に，局所構造を予測する手法がある [3]．また，連続する３つのアミノ酸を用いて，各アミノ酸間の共有結合の周りの回転を表す内部回転角 (二面角ともいう) の情報から，立体構造全体を予測する手法 [4] や，最近では，溶解性，水素結合性などを考慮した立体構造上のエネルギーの特徴を表す 3Dprofile を用いる方法 [5] も提案されている．

4.2 チェインコード列とアミノ酸配列の対応関係

チェインコードを用いると，アミノ酸配列上では類似性が見られないものの，立体構造上では類似性が見られるものが存在することが容易にわかる（図 5）．このことから，複数のアミノ酸配列に対し，ほぼ似たチェインコード列が生成されると想像でき，アミノ酸配列から，チェインコード列を予測することを試みる．

まず，予備実験として，チェインコード１文字を生成するには，連続する４つのアミノ酸 (4-tuple) が必要であることから，この，オーバーラップを許す 4-tuple の分布と，生成されるチェインコードの種類の一般性を調べてみた．

```
35 AAAA CEEEEFHHHHHHHHHHHHHHHHHHHHHHIKKI
27 SGSL ADDDEEEEEFFFFHHIIIIIIKNNNNP
27 ALLA EEEHHHHHHHHHHHHHHHHHHHHHHP
26 SVTV DEEEEEEEEEEEEEEEEEEEEEEEQ
26 SSSV CCCCCEEEEEEEEEEEEEFHHHHII
25 VEEL AAACEEGHHHHHHHHHHHHHHHHH
25 VAAA DEEEFFHHHHHHHHHHHHHHHHH
25 LGLP ACDDEEEEEEEEEEEEGHHIIKPPP
25 LAAA CEEEEEFHHHHHHHHHHHHHHHHQ
25 GGGA CCCDDDEEEEFFGHHHHHIIKKKNN
25 ALKA EEFFHHHHHHHHHHHHHHHHHHN
25 AAGA AAAAEEEEEFFHHHHIKKKKQQQQQ
24 VASG CCCDEEEEEEEEEHHHHHHHHHI
24 VAAG CEEEEEEFHHHHHHHHHHHQQQR
24 SSVT CEEEEEEEEEEEEEEEEHHHHHP
24 IEEL CEEGHHHHHHHHHHHHHHHHHH
24 AAGV CCEEHHHHHHKKKQQQQQQQQQQ
```

図 6: 左：データセット中に出現したアミノ酸 4-tuple の回数，中：アミノ酸 4-tuple，右：4-tuple 出現ごとの，それぞれに対応するチェインコード

データとして用いたのは，データの偏りを防ぐため，アミノ酸配列レベルで類似性の高いものを含まないタンパク質立体構造データベースの 1532 エントリ [6] である．結果の一部を図６に示す．

これより，アミノ酸 4-tuple に対し，まったくばらばらなチェインコードが対応した訳ではないが，図 6 の上から２番目の'SGSL' のように，N と F など，まったく方向の異なるベクトルを示す可能性があることがわかる．また，ここでは割愛しているが，データセット中に一回しか現れない 4-tuple があまりに多いこともわかった．従ってこの情報を用いてチェインコードを予測するのは，統計的信頼性に乏しいこともあり，非常に限定しにくい．

もちろん，4-tuple だけで，立体構造が決定する訳ではなく，4-tuple の周辺の鎖に沿った残基や，鎖に沿っては遠いが空間的に近くなる残基の影響が考えられるため，より強力で一般的な，アミノ酸配列からチェインコードを生成する規則を見つけることが理想的ではあるが，これについてはさらなる議論の余地がある．

本研究では，一般性はなくなるが，予め特徴のあるタンパク質だけに限定することにし，

アミノ酸配列レベルで局所的な共通パターン（モチーフと言われる）を含むタンパク質に注目することにする．あるモチーフを含むタンパク質に対して，アミノ酸 4-tuple からチェインコードを予測する手法を試みた．

4.3 試みた手法

以下のようなステップを実行し，チェインコード予測を行う．

1. あるモチーフをもつタンパク質に注目し，このモチーフについて，PROSITE [7]（モチーフデータベース）に記述のあるタンパク質立体構造データベース (PDB) のエントリをデータセットとする．

2. そのデータセットの各々のエントリに対し，オーバーラップを許したすべてのアミノ酸残基 4-tuple と，その 4-tuple に対応するチェインコードの組合せの統計をとる．

3. 立体構造を予測したいアミノ酸配列の先頭の 4-tuple に対し，すでに出来上がった統計の情報の中に，その 4-tuple の出現があれば，それに対応する複数のチェインコードを候補として挙げておく．これを，アミノ酸配列の最後の 4-tuple まで繰り返し，候補集合の列をつくる．

4. 最も確からしくなるように，チェインコードの前後関係を考慮に入れながら，候補集合の列から１つずつチェインコードを決定し，それをつなげて予測結果とする．

統計の情報を用いて，チェインコードのその位置における生起確率を求め，また，その位置の前後のチェインコードから，注目すべきチェインコードへの遷移確率をも考慮に入れることになる．

5 実行結果

5.1 herix–turn–herix モチーフ

PROSITE に記述されている herix-turn-herix モチーフを持つ２８個のエントリ中，２７個のエントリを頻度統計情報として用い，残りの１つを予測に用いて，予測されるコードと実際のコードの比較を行った．

残基番号	頻度統計	予測コード	実際のコード
64	EEEE	E	E
65	EEEQ	E	E
66	AACCEEKQ	E	A
67	EEEW	E	E
68	HHHI	H	H
69	IIII	I	I
. .			
170	HHHH	H	H
171	HHHH	H	H
172	HHHH	H	H
173	EEEEHHHH	H	H
174	HHHH	H	H
175	FKKK	K	K
176	DEEE	E	E
177	PPPP	P	P

herix-turn-herix モチーフを持つ
タンパク質の予測結果の一部

上段の例では，66 残基目の予測が難しいと思われる．チェインコードの遷移を考えると，E や H が連続することが多く，66 残基目は前後関係から E を予測することになる．下段の例では，173 残基目は間違いなく H を予測できる．

5.2 Greek key モチーフ

これも herix–turn–herix モチーフと同様に，PROSITE に記述されている Greek key モチーフを持つ１１個のエントリから１０個のエントリを頻度統計情報として用い，残りの１つを予測に用いて，予測されるコードと実際のコードの比較を行った．

108 番目は P と W の生起確率は等しいが，107 番目が E で，かつコード化に用いた正二十面体の E 面が P 面の隣であるため,P の方が確率が高くなり P を予測した．

頻度の少ない 4-tuple が存在した時，かつその 4-tuple に対応するチェインコードの前後の

コードにEやHが無い時（50番目や105番目）は予測は難しい．EやH以外の前後関係の情報をもっと知る必要があると思われる．

以上の2つの例では，大まかにはアミノ酸配列からチェインコード列を予測できたと言える．

残基番号	頻度統計	予測コード	実際のコード
43	EEE	E	E
44	QQQ	Q	Q
45	CC	C	C
46	FF	F	F
47	HH	H	H
48	KK	K	K
49	DD	D	D
50	CH	?	H
51	CC	C	C
52	EE	E	E
53	EE	E	E
. .			
101	EEEE	E	E
102	EEEE	E	E
103	EEEE	E	E
104	PPP	P	P
105	ACI	?	I
106	FKK	K	K
107	EEE	E	E
108	PPWW	P	P
109	CCCC	C	C

Greek key モチーフを持つタンパク質の予測結果の一部

6 考察と今後の展望

本研究では，タンパク質の立体構造を表すチェインコードを用いて，アミノ酸配列と，それから生成されるチェインコードの対応をとり，これをもとにしてアミノ酸配列からチェインコードの予測を試みた．上で挙げた例ではうまく予測できていると言えるが，実際は，統計をとるためのデータセットの数が少な過ぎるため，herix–turn–herix モチーフの場合におけるアミノ酸4-tupleの分布は2005通りで，Greek key モチーフ場合においては651通りしか得られていない（可能性としては20^4の場合がある）．

立体構造予測をする対象を，タンパク質一般でなく，特定のモチーフを持つもののみに限定したのだが，それに伴ってアミノ酸4-tupleの分布範囲が狭まり，これをもとにするとチェインコードを予測するには不十分だと考えられる．従って，統計に現れないアミノ酸4-tupleが，予測するアミノ酸配列にたくさん存在すれば，まったく予測できないことになる．従って，統計に現れないアミノ酸4-tupleをどう評価するか，を考案すべきである．また，予測の精度を上げるために，チェインコードの前後関係，つまりコードの遷移確率を詳しく調べる必要があるとも思われる．今後は，これらの問題点について取り組む予定である．

謝辞

本研究は一部，文部省科学研究費補助金重点領域研究「ゲノムサイエンス」
（課題番号 08283103）によっている．

参考文献

[1] Matsuda, H., Taniguchi, F. and Hashimoto, A., "A Notation of Amino Acid Conformations for Exploring Similar Protein Structure", In *Proc. of 1st Pacific Symposium on Biocomputing*, pp.732–733, 1996.

[2] Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer Jr., E. E. Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., Tasumi, M. " The Protein Data Bank: a computer-based archival file for macromolecular structures" *J. Mol. Biol.*, Vol. 112, No.3, pp. 535–542, 1977.

[3] Garnier, J., Osguthorpe, D.J., Robson, B., " Analysis of the accuracy and implication of simple method for predicting the secondary structure of globular proteins", *J. Mol. Biol.*, vol.120, pp.97–120, 1978.

[4] Wu, T. T. and Kabat, E. A., "An attempt to evaluate the influence of neighboring amino acids (n–1) and (n+1) on backbone conformation of amino acid (n) in proteins" *J. Mol. Biol.*, vol.75 pp.13–31, 1973.

[5] Matsuo, Y., Nakamura, H., Nishikawa K., "Detection of protein 3D–1D compatibility characterized by the evaluation of side-chain packing and electrostatic interactions", *J. Biochem.*, vol.118, pp.137–148, 1995.

[6] Noguchi, T., Onizuka, K., Akiyama, Y., Saito, M., "http://pdap1.trc.rwcp.or.jp/papia-cgi/pdbreprdb_table.pl?83" ,1998.

[7] Bairoch, A., Bucher, P., Hofmann, K., "The PROSITE database, its status in 1997" *Nucleic Acids Research*, vol.25, pp.217–221, 1997.