三次元曲線の記号表現法とその蛋白質立体構造表現への応用

1N−6

鬼塚健太郎[1]，石川幹人[1]，浅井潔[2]

1: (財) 新世代コンピュータ技術開発機構　2: 通商産業省 工業技術院 電子技術総合研究所

1 はじめに

蛋白質の立体構造は 3 次元のデータであるため、構造同士の検索、比較が大変難しい。これまで、蛋白質の局所構造を解析分類し、この分類を用いて構造を記述することはなされてきたが [1][2][3]、大域的な構造をこれによって記述することは困難であるため、大域構造の分類などはできなかった。今回考案した方法は、これを克服し、大域構造と局所構造を同じ形式で記号あるいは言語で記述できるようにし、さらに、局所構造と大域構造との関係も形式言語における一種の文法として記述することができるものであり、これによって、蛋白質の大域構造と局所構造の関係 (垂直関係) と局所構造同士、あるいは大域構造同士の関係 (水平関係) の両者を法則として記述できることになり、構造予測では威力を発揮することになる。

また、この方法は、蛋白質の立体構造解析にとどまらず、一般的に、空間の曲線を記号記述することに応用できる。たとえば、フラクタル構造をもった空間曲線などを記述するのに利用できる訳である。

2 構造記述方法

今回考察した蛋白質の部分構造を、配列上連続するアミノ酸残基の空間配位によって記号記述する方法について説明する。

最初に、部分構造の数値表現方法について説明する。ここで考える部分構造は、蛋白質の配列上で連続する n 個のアミノ酸残基の C^{α} 原子の空間配置に基づくものである。

まず、この部分構造から構造を良く表現していると思われる 3 個のベクトルを抽出し、次に、このベクトルの長さや、ベクトル同士の角度で、部分構造を数値表現し、この数値を離散化して文字列表現する。

配列上連続する残基の空間配位で重要と考えられる空間ベクトルとして今回は、以下の 3 種類を考えた。

1. Wavelet 変換による 2 個のベクトル
   近年信号処理の分野で広く使われ始めている Wavelet 変換 [4] を応用し、考えている n 残基の座標にアミノ末端 (以後 N 末端) 側から順番に幅 n の Wavelet を掛け合わせて、それらの和をとったものである。

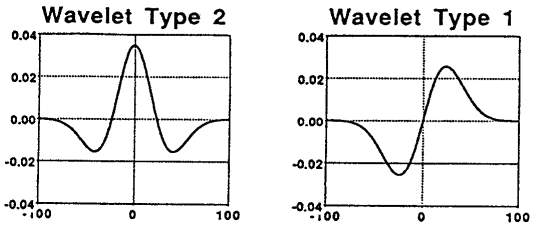

図 1: Wavelet

Symbolic Representation Method for Three-Dimensional Curves and its Application for Protein Structure Analysis
Kentaro ONIZUKA[1], Masato ISHIKAWA[1], Kiyoshi ASAI[2]
1: Institute for New Generation Computer Technology(ICOT)　2: Electrotechnical Laboratory (ETL)

(a) 前進ベクトル
Type 1 の Wavelet(図 1) を用いると、n 残基の N 末端側の残基の大まかな位置とカルボキシル末端 (以後 C 末端) 側の残基の大まかな位置を結ぶベクトルが得られ、このベクトルの大きさから、考えている n 残基の局所構造がどれくらい空間的に長く分布しているかがわかる (図 2)。これを前進ベクトルと呼ぶことにする。

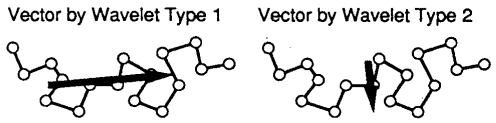

図 2: Wavelet によるベクトル

(b) 折れ曲がりベクトル
Type 2 の Wavelet を (図 1) 用いると、n 残基の配列上の両端に位置する残基の平均の位置から中央付近の残基の平均の位置を結ぶベクトルが得られ、このベクトルの大きさから、考えている n 残基の局所構造がどれくらい曲がっているかが分かる (図 2)。これを折れ曲がりベクトルと呼ぶことにする。

2. 回転ベクトル
蛋白質の 2 次構造で特徴的な helix などの螺旋構造を捕らえるために、連続する 3 個の残基 A, B, C について角度 $\angle ABC$ の大きさを持ち、外積 $\vec{AB} \times \vec{BC}$ の方向を持つベクトルを考え、これを考えている n 残基で、ずらしながら計算し、足し合わせたものである (図 3)。helix の部分では、このベクトルの長さは一般に最大値をとる。また、これによって、考えている n 残基が、全体としてどう螺旋を描いているか (例えば、全体として右巻きか左巻きかなど) が分かる。これを回転ベクトルと呼ぶことにする。

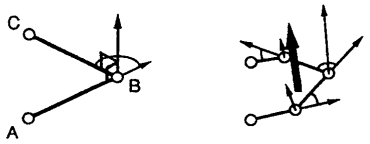

図 3: 回転ベクトル

これらのベクトルを用いて、部分構造を数値化するには上述の 3 個のベクトルの長さと、前進ベクトルと折れ曲がりベクトルの間の角度、回転ベクトルと折れ曲がりベクトル、前進ベクトルとの角度の 6 個の数値、および、前進ベクトルと、折れ曲がりベクトルの外積で得られるベクトルと、回転ベクトルとの内積の値の正負で得

られるバリティの7個のパラメータを用いる。この6個の数値と、バリティによって、ベクトルの相対関係は完全に記述できる。また、この6個の数値、およびバリティは、空間のアフィン変換の自由度を持たないため、元のデータの回転並行移動に関して、客観的な量である。また、ここで数値的に記述された構造の表現からは原理的に元の立体構造が完全に復元できる。

次に、蛋白質の立体構造を配列に沿って、端から1残基ずつずらしながら適当なn残基の部分構造を上述の数値表現を計算して、全体構造を数値化する。数値のばらつきを調べて、6個の数値それぞれを離散化する。これは、数値の平均値と標準偏差の値をもとに十分細かく分割した。この分割数はベクトルの長さについては細かくとり、角度は歪んだ構造でもトポロジカルに似ていることなどを考えれば数個とれば十分であると考えた。

そこで、こうして分類された構造を、似た構造が似た名前を持つように記号化する。ベクトルの長さについては、上述の離散値に対して、アルファベットの r, l, m, n など流音鼻音を除く子音の文字を当てる(除いた子音はその他のことを表現するためにとっておく)。ベクトルの長さの数値の小さいものから順に子音を "x k g j d t s f p b w" などと当てると、発音の類似性と値の近さが分かり易い。

また、角度については、アルファベットの母音にあたる文字を適当に当てる。これも発音の類似性から、角度の小さい順に、"i e a o u"などとあてることにすると、値の近さが分かり易い。

このように離散値にたいして発音可能な文字を当てることで、部分構造は、子音＋母音＋子音＋母音＋子音＋母音 で表現し、またバリティが負の場合は前に in をつける。このように記号化された部分構造は発音可能な単語となる。例えば、"insididi jatata fadusa intasede intotoda fadedo tesotu" などとなる。

実際に構造上特記すべきなのは、前進ベクトル、あるいは折れ曲がりベクトルが極大値を取っているような場所や、N末端、C末端である。そこでそのような部分構造のみを考えて他は捨てることにする。これでも、部分構造同士は十分に重なり合うので、立体構造復元のための情報は保存されている。重なり具合の分類は、後続の部分構造との重なりの幅が部分構造の幅の何分の一であるかに応じて、4種類程度に分け、それを表すために、語の語尾に重なりの大きいものから順番に "m n ng" を付ける(何も付けないものもある)。すると、蛋白質の立体構造は、例えば、ヘモグロビンのA鎖を65残基の構造で表現すると、"insididim jatatan fadusam inteseden intotodam fadedom tesotu" となる。

さて、このように、適当な残基数nについて調べた蛋白質の立体構造を文として表現できるようになったので、次に、残基数nを変えて幾つかの系統で、文として表現することができる。例えば、倍々(ここでは、2^n+1)に増えるように、3残基、5残基、9残基、17残基、33残基、65残基、129残基、257残基でそれぞれ文にする。こうすると、例えば、ヘモグロビンのA鎖の場合、構造を以下のように表現できる。

残基数	構造の記号表現
3	intatatam intadadam injasasam injafafam injafafam indasasam ...
5	insejudom indafasan injasafin injasafin injasafem injasafin ...
9	injatosim injajapim jajapim jegepim injadaping indatatim ...
17	indajapin dadafim datasim indadofim indojoping injadasin ...
33	tatafim tatapim insajafin indaditom dadodom datosen insajapin ...
65	insididim jatatan fadusam inteseden intotodam fadedom tesotu
129	dadadom datatam datata

このようにすることで、大域的な構造と局所的な構造を幾つかの系統で文として表現することができる。ここで、例えば、3残基で観察したものをもっとも細かい構造を表していることから、violet 文、以下、blue 文、magenta 文、green 文、yellow 文、orange 文、red 文, infra-red 文などと命名すると分かり易い。

勿論、蛋白質1個の全体構造の概要も蛋白質全体について6個の数値で表現し、それを単語にして、例えば、"グロビン族の蛋白質は、detato, 或いは dadeta 構造である。"などと表現できることになる。

このような構造記述方法によって、似た構造をもつ蛋白質はこの構造記述言語で似た文章によって記述されることになり、構造の類似性をとらえやすくなる。また、単語が各色ごとにどれくらいあるかを調べることで、実際の蛋白質の部分構造の構成要素を調べることができる。現在までの研究で95種類の蛋白質の立体構造から、violet 文では、97単語あり、blue 文では、1230単語ある。

3 構造の文法記述

上述のように蛋白質の立体構造を文で表現すると、単語と単語との関係によって、水平関係は、同じ色の文の2単語のつながりとして考えられるから、これを考えれば、violet 文で "injasasam injasasa" という語句が良く見うけられるということなどが、文法として記述されることになる。実際これは helix 構造に対応している。

また、垂直関係は、異なる色の単語の間の関係を調べれば良く、helix の場所などは、"blue injasafi :- violet injasasam injasasam injasasa." と言った記述が可能である。勿論、大規模な部分構造はそれより局所的な構造では任意性があるから、これを、"blue injasafi :- violet injasafam injasafam injasafa; violet injasasam injasasam injasasa; injasafam injasasam injasafa."などと表現できる。

こうして、水平関係、垂直関係を全て形式文法の形で表現でき、これを構造の制約とすることができる訳である。

4 おわりに

今回の方法により、これまで出来なかった蛋白質の立体構造の大域的構造から局所的構造までの階層的記述が可能となり、また、構造の制約を文法として取り出すことができることになる。

また、この方法を一般化して、空間曲線一般に応用すれば、ある種の空間曲線の性質を、文法としてとらえられることになり、パターン認識や、パターンの記号処理(文字認識など)に威力を発揮するものと考えられる。

今後の展望としては、今回の方法によって記述された蛋白質の立体構造と、アミノ酸配列との関係を、何らかの方法で、調べることによって、蛋白質の構造予測、また、構造からの配列予測を行なう方法を研究することである。

参考文献

[1] Chou, P.Y. and G. D. Fasman, 1974, 'Prediction of protein conformation', *Biochemistry 13*, 222-244.

[2] Fasman, G.D.(Ed), 1989, *Prediction of Protein Structure and the Principles of Protein Conformation*, New York: Plenum Publishing Corporation.

[3] Von Heijne, G., 1987, *Sequence Analysis in Molecular Biology*, San Diego: Academic Press, Inc.

[4] Combes, J.M., A. Grossmann and Ph. Tchamitchian (Eds), 1987, *Wavelets*, Berlin and New York: Springer-Verlag.