主成分分析によるタンパク質主鎖立体構造表現法の提案

高橋勝利
産業技術総合研究所 生命情報科学研究センター

概要

タンパク質分子は非常に多数の原子からなる生体高分子である。特にタンパク質主鎖の折れ畳み構造はタンパク質を特徴付ける。最も精密な構造表現は原子座標を直接取り扱う事であるし、最も粗い表現は主鎖の局所的な構造（2次構造）を記号で表し、記号列として主鎖構造を記述する方法である。本論文では、数残基からなるタンパク質主鎖フラグメントを単位として、精密な表現から粗い表現まで自由に分解能を調整できる「タンパク質局所構造記述」法を提案する。また、主鎖フラグメントの構造からタンパク質主鎖全体の構造を記述する方法についても述べる。

Principal Component Analysis of Short Backbone Fragment Conformations in Globular Proteins and Its Use for Protein Backbone Conformational Description

TAKAHASHI Katsutoshi
Computational Biology Research Center,
National Institute of Advanced Industrial Science and Technology

Abstract

In this paper, we propose the new method which describes efficiently the protein backbone folding structures. The most precise conformational description is to use atomic coordinates. One of the most course descriptions of the protein local conformation is the secondary structural representation. We carried out the principal component analysis of the atomic corrdinates of the many short peptide backbone fragment conformations found in many globular protein entries of protein data bank, in order to find out the most significant conformational properties to describe backbone fragment conformations. We present the flexible manner to describe the peptide backbone conformation and the folding structures of the proteins using such significant conformational properties.

1　はじめに

タンパク質分子は非常に多数の原子からなる生体高分子である。特定のアミノ酸配列を持つタンパク質分子の立体構造特に主鎖折れ畳み構造は一義に決まると考えられており [1]、タンパク質を理解するためにはタンパク質の立体構造を扱う必要がある。現在精力的に、X線結晶解析法や核磁気共鳴法による原子レベルでのタンパク質立体構造決定が実施されており、その結果はプロテインデータバンク (PDB; Protein Data Bank) [2] に登録されている。

タンパク質の折れ畳み構造を取り扱うために、様々なレベルでの記述法が考えられている。最も精密なのはタンパク質を構成する全原子のＸＹＺ座標を用いることであるが一つのタンパク質分子には数千から数万の原子が含まれており、タンパク質の折れ畳みパターンや局所的な構造を議論するためにはあまりにも自由度が高すぎて逆に使いにくい。

一方で、タンパク質主鎖の局所的構造に着目して立体構造を記述することも盛んに行われている。有名なのがペプチド主鎖原子間の水素結合パターンに基づいて数残基からなる局所構造を分類し、文字列の形で構造を記述する方法 [3] である。しかしこの方法では原子レベルでの構造情報が完全に失われており、タンパク質の折れ畳みパターンを記述できるほどの分解能は有していない。

そこで本論文では、タンパク質主鎖の立体構造を任意の分解能で記述する新しい方法を提案する。この方法を使えば、２次構造レベルの局所構造記述からタンパク質主鎖の原子座標を使った精密な構造記述まで、様々な分解能で構造を記述することが可能

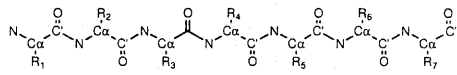

図 1: タンパク質主鎖フラグメント（７残基）を構成する主鎖原子

である。

2 タンパク質主鎖フラグメントのコンフォーメーション

タンパク質分子は２０種類からなる天然アミノ酸が鎖状に脱水縮重合して作られている。タンパク質分子内でもともと一つのアミノ酸に属していた原子団をアミノ酸残基と呼び、タンパク質を構成する単位と考える。アミノ酸残基の中でアミノ酸の種類を決定する部分を側鎖、それ以外の部分を主鎖と呼ぶ。生体内では、ゲノムＤＮＡ上の遺伝子にタンパク質のアミノ酸配列情報が納められており、アミノ酸配列が与えられればタンパク質の折れ畳み構造、特に主鎖の折れ畳み構造は一義的に決定される。

図１に７残基からなるタンパク質主鎖フラグメントを構成する主鎖原子を示した。この図から判るように主鎖原子は１残基あたり４個となり、７残基の場合、主鎖フラグメントは２８個の原子から構成されることになる。

従って７残基フラグメントの立体構造を記述するための自由度数は $7 \times 4 \times 3 - 6 = 78$ となる。しかし主鎖フラグメントを構成する原子は共有結合で結ばれているし様々な立体的な制約がかかっている。このため主鎖フラグメントのコンフォーメーションを記述するのに必要な自由度はもっと少なくなる。様々なタンパク質内の多数の主鎖フラグメントの原子座標値を主成分分析 (PCA; Principal Component Analysis) することで、コンフォーメーション記述に重要な変数（主成分座標）を抽出することができる [4]。

図２に主鎖フラグメントの原子座標空間から主成分座標空間への変換について簡単にまとめた。この方法ではまず原子座標の張る多次元空間内で様々な主鎖フラグメントがとる分布を考え、分布の広がりの大きい方向（主成分座標）がフラグメント間のコンフォーメーションの違いを記述するのに重要であると考える点がポイントである。

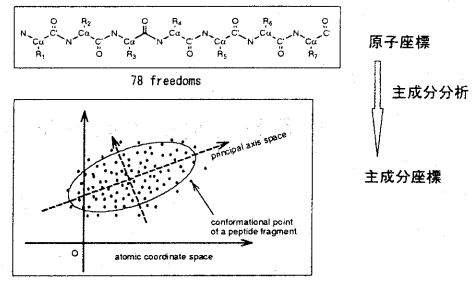

図 2: PCA による原子座標空間から主成分座標空間への変換

2.1 主鎖フラグメント原子座標の主成分分析

原子数 n_a 個のフラグメントのコンフォーメーションが $3n_a$ 次元の原子座標空間、$x_i = m_i^{1/2} r_i M^{1/2}$, $(i = 1, 2, \ldots, 3n_a)$, 内の一点として表せるものと考える。但し、$m_{3j-2}, m_{3j-1}, m_{3j}, r_{3j-2}, r_{3j-1}, r_{3j}$ は j 番目の原子の質量とカーテジアン座標であるとし、M はフラグメントを構成する原子の質量の総和である。

ここで、データベースから抽出した n 個のコンフォーメーションを考え、k 番目のコンフォーメーションに対応する i 番目の座標を x_{ki} とし、n 個のコンフォーメーションの分布を考える。（ここで分布の中心が原点になるように座標変換しておくことにする。）次に、分布の広がりを特徴付ける直交基底ベクトル（主軸ベクトル）$\boldsymbol{f}_1, \boldsymbol{f}_2, \ldots, \boldsymbol{f}_{3n_a}$ を考える。ここで、l 番目の主軸ベクトル $\boldsymbol{f}_l$ の i 番目の成分を f_{il} とすると、主成分（主軸）座標系における各点の座標は

$$y_{kl} = \sum_i x_{ki} f_{il} \tag{1}$$

で与えられる。第一主軸ベクトル $\boldsymbol{f}_1$ は分布の平均二乗偏差を最大にするベクトルであり、

$$\frac{1}{n} \sum_k y_{kl}^2 = \sum_{i,i'} C_{ii'} f_{il} f_{i'l} \tag{2}$$

を満たす様に決定できる。ここで $l = 1$ であり、$C_{ii'}$ は

$$C_{ii'} = \frac{1}{n} \sum_k x_{ki} x_{ki'} \tag{3}$$

で定義される分散共分散行列 C の要素である。$\boldsymbol{f}_l$ は単位ベクトルであり、$\sum_i f_{il}^2 = 1$ を満たすため、式 (2) は

$$\sum_{i'} C_{ii'} f_{i'l} = \lambda_l f_{il} \tag{4}$$

で表される固有値問題に帰着することができる。固有値 λ_l は主軸ベクトル f_l 方向に分布を射影した際の平均二乗偏差を表しており、l 番目の主成分によって表現できる二乗偏差の割合 $\lambda_l/tr(C)$ はフラグメントのコンフォーメーションを表すための重要性を表す指標として用いることができる。

3 手法

3.1 主成分座標系から主鎖フラグメント原子座標系への変換

タンパク質主鎖フラグメントの原子座標の主成分分析を実施することで、コンフォーメーションの分布の主軸を求めることができる。この主軸（主成分）ベクトルが張るより少ない次元の主成分空間でフラグメントのコンフォーメーションを議論することにより、取り扱う自由度を格段に減らすことが可能である。原子座標系と主成分座標系の間の変換は単純な直交基底変換となるため逆変換が可能であるし、少数の主成分座標のみを用いその他の主成分に関するばらつきを無視することで、任意の個数の主成分座標からフラグメントを構成する原子の座標を再現することができる。

m 個の主成分座標、$y_1, y_2, \ldots, y_m$ からフラグメントの $3n_a$ 個のカーテジアン座標値、$x_1, x_2, \ldots, x_{3n_a}$ は、

$$x_i = \sqrt{\frac{M}{m}}_i \sum_{l=1}^{m} f_{il} y_l + \bar{x}_i \qquad (5)$$

で求めることができる。ここで、$\bar{x}_i$ はフラグメントのコンフォーメーション分布の中心座標、つまり n 個のフラグメントの平均原子座標値を表している。

$m < 3n_a$ 個の主成分座標のみを用いてタンパク質主鎖フラグメントのコンフォーメーションを記述し、そこからフラグメントの原子座標を再現した場合、当然情報が落ちている分誤差を生じることになる。m 個の主成分座標値から再現される原子座標と正しい原子座標間の平均的な誤差（平均二乗誤差の平方根; R.M.S. Error; Root Mean Square Error）は

$$\sqrt{tr(C) - \sum_{l=1}^{m} \lambda_l} \qquad (6)$$

で表すことができる。

4 実験

１９９６年５月バージョンのＰＤＢ（プロテインデータバンク）から代表的な４５７本のポリペプチド鎖を選び、これらのポリペプチド鎖の原子座標から重なりを許して７残基の主鎖フラグメントを抽出し、主鎖フラグメントデータベースを作成した。フラグメントの数は１２００４３個であった。以下このフラグメントデータベースを用いて解析を行った。

表 1: 主鎖フラグメント立体構造の主成分分析結果

l	λ_l	$\lambda_l/tr(C)$ (%)	$\sum_{j=1}^{l} \lambda_j/tr(C)$ (%)
1	1.95200	36.8	36.8
2	1.15949	21.9	58.7
3	0.60250	11.4	70.0
4	0.28809	5.4	75.5
5	0.22260	4.2	79.7
6	0.13531	2.6	82.2
7	0.11961	2.3	84.5
8	0.10413	2.0	86.5
9	0.08690	1.6	88.1

4.1 主鎖フラグメント原子座標の PCA

１２００４３個の７残基主鎖フラグメント原子座標の主成分分析を行って７残基の主鎖フラグメントのコンフォーメーションを記述する７８個の自由度を、フラグメントデータベースに登録された主鎖フラグメントコンフォーメーションのばらつきを良く説明できる主成分に変換した。その結果を表１に示した。その結果、高々第一主成分のみを用いるだけでデータベース中のフラグメントコンフォーメーションのばらつきの約３７％を説明できることが明らかである。この結果から、７８個の自由度のうち極少数のみを用いるだけで効率的に主鎖フラグメントのコンフォーメーションを記述できることが判明した。

4.2 主成分座標からの主鎖フラグメント原子座標再現性

主鎖フラグメント原子座標の主成分分析の結果に基づいて、任意の数の主成分座標を使ってフラグメントの原子座標を再現した場合の R.M.S.Error の値を計算してプロットしたのが図３である。

このグラフから、平均約１オングストロームの R.M.S.Error で原子座標を再現するために必要な主成分の数はたかだか５つであることが判る。このようにデータベースに登録されている様々なコンフォーメーションを記述するためには全原子座標を用いる必要は全く無く、構造記述に必要な自由度数は非常に少ないことが判明した。

次に、２次構造に代表される主鎖折れ畳み構造の局所的性質によって、構造記述性がどの程度変化するのかを調べる目的で、ＰＤＢに登録されている牛すい臓トリプシン阻害ペプチド (BPTI; Bovine Pancreatic

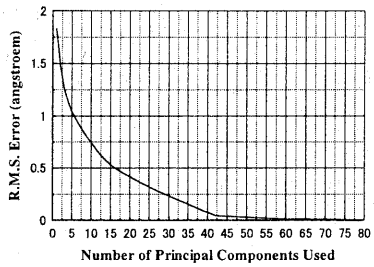

図 3: フラグメントの原子座標を再現する場合の平均的な誤差 (R.M.S.Error) と用いる主成分座標の数との関係

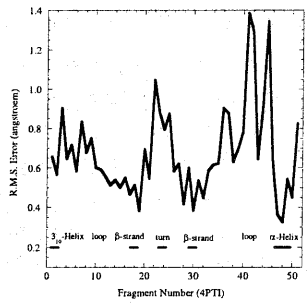

図 4: BPTI 主鎖構造に含まれる 7 残基フラグメントと 5 つの主成分座標のみを使って再現したフラグメント構造との誤差

Trypsin Inhibitor, PDB entry=4pti, 58 残基）の X 線構造に含まれる（重なりを許した）7 残基フラグメントの主成分座標値を計算し、構造記述性への寄与の高い 5 つの主成分座標値から再現したフラグメント原子座標と X 線構造の間の R.M.S.Error を計算した（図 4)。

この結果から、主成分座標を使って記述できる主鎖フラグメントコンフォーメーションの精度は 2 次構造によって変化し、ヘリックス構造や β ストランド構造などの規則的な構造の場合精度高く記述できるが、ターンやループなどといった不規則性の高い構造では精度が低くなる傾向が明らかである。

しかし、5 つの主成分を用いた場合誤差はせいぜい 1.4 オングストローム程度であった。この精度は BPTI の全折れ畳み構造を記述するのに十分である。詳しくは述べないが、重なりを許して定義した 7 残基フラグメントの構造を 5 つの主成分のみを使って

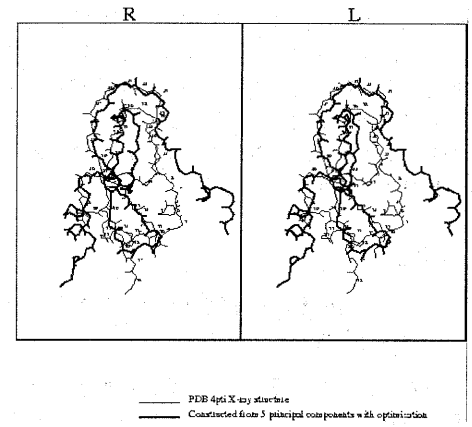

図 5: 重なりを許して定義した 7 残基フラグメントの構造を 5 つの主成分のみで記述し、全体の折れ畳み構造を再現した結果。太線が X 線構造であり細線が 5 つの主成分のみから再現された構造である。

記述し、そこから全体の折れ畳み構造を再現することが可能である（図 5)。この図に示した例ではターンやループ部分での曲がり方が十分に再現できなかった関係で全主鎖原子座標の R.M.S.Error は 5.6 オングストロームとなってしまったが、格段に少ない自由度のみを使って十分な精度で構造記述が行えることが判る。

5 まとめ

本論文中、全原子座標を用いるのではなく構造データベースに登録された多数のフラグメント構造の共通性、非共通性を統計解析し構造記述に重要な自由度を少数選び出して、主鎖フラグメントの立体構造及びタンパク質の折れ畳み構造を記述する新しい手法を提案した。この手法によって任意の精度でタンパク質主鎖の局所構造、折れ畳み構造記述することが可能であり、タンパク質立体構造予測をはじめとした様々な新規アプリケーションへの道が開けるものと期待できる。

参考文献

[1] C.B. Anfinsen: *Science*, Vol. 181, pp. 223 (1973).
[2] F.C. Bernstein, T.F. Koetzle, et. al: *J. Mol. Biol*, Vol. 112, pp. 535 (1977).
[3] W. Kabsh and C. Sander: *Biopolymers*, Vol. 22, pp. 2577 (1983).
[4] K. Takahashi and N. Go: *Biophysical Chem.*, Vol. 47, pp. 163 (1993).