

多数の無相関混合正規分布によるタンパク質構造分布のモデル化

4 X-3

田中 秀俊, 白石 将, 川上かおり, 青山功, 佐藤 裕幸

新情報処理開発機構 並列応用三菱研究室

(三菱電機(株) 情報技術総合研究所内)

1 はじめに

20 種類のアミノ酸 (ACD..Y) からなる列を $R_n = \{r_1 \dots r_n | r_i \in \{A, C, D, \dots, Y\}\}$ とすると、各アミノ酸残基 r_i に対応する構造特徴 d_i の列として、タンパク質構造列 D_n が定義できる。例えば第 i 残基の構造特徴として主鎖二面角 3 種類 (ϕ_i, ψ_i, ω_i) を選ぶと、構造列は $D_n = \{d_1 \dots d_n | d_i = (\phi_i, \psi_i, \omega_i)\}$ である。この R_n の中の m 残基からなる部分残基列 R_m とそれに対応する部分構造列 D_m について、 $D_m = f_m(R_m)$ のような関係が、小さい m で成り立てば、タンパク質の立体構造予測に使うことができる。 $m=1$ すなわち 1 残基の関係 $D_1 = f_1(R_1)$ のうち、 D_1 に ω 以外の 2 種類の主鎖二面角を用いる場合の f_1 は、ラマチャンドランプロットとしてよく知られている。

残念ながら、タンパク質立体構造データベース PDB を少し眺めると、 $D_m = f_m(R_m)$ は確定的には成り立たないことがすぐ分かる。例えば $R_4 = \text{AAAA}$ の場合の D_4 は、 α 螺旋 (ϕ, ψ) = (-60, -40) 周辺をとることが多いものの、それ以外の全く違う構造もとりうる。よって、構造は例えば $E[D_m] = f_m^{(c)}(R_m)$ である確率が $p^{(c)}$ となる C 個の確率分布 $\{f_m^{(c)} | c \in \{0, 1, \dots, C-1\}\}$ の混合によって記述される必要がある。いくつの確率分布で記述するべきかについては、例えば最小記述長基準をもとに ϕ, ψ の分布を適正に分類する数を求める試みが既になされている [1]。

さらに、第 i 残基から m 個とった残基列 R_{mi} によって同じ場所の構造列 D_{mi} が決まると考えるよりは、遠くの他の残基も寄与すると考える方が自然である。例えば寄与が第 a 残基と第 b 残基の 2箇所なら、簡単のために m を同じとすると $E[D_m] = f_{(ma, mb)}^{(c)}(R_{ma}, R_{mb})$ のような形になる。この点、全体を考慮するスレッディング法や、HMM を用いて残基と構造の関係を記述した HMMSTR[2] は、遠くの残基を考慮していると言える。

本稿では、タンパク質の構造が離れた場所の残基種類にも依存することを前提として、その依存度を、複数の確率分布の混合によって構造が確率的に決まる

いう枠組に基づいて見積もった結果を報告する。確率分布では最も単純な、対角共分散成分のみの無相関な正規分布による混合分布を、立体構造データのインデックスに用いて、(以下、混合対角共分散ガウシアンインデックス (MDGI)¹ と称する) 配列から構造を予測する精度という観点から、上記の依存度を見積もった。

2 手順

まず、PDB データから切断鎖などの信頼性の低いデータを除去し、さらに似た構造は適当に間引いて、立体構造の出現頻度の偏りを緩和したデータセットを作成する。MDGI の計算は時間がかかるので、小さなセット (ベースセット) でインデックスの元を計算し、より大きなセット (プラスセット) を用いてインデックスのカバー範囲を拡大するという手順を踏む。本稿では、PDB2001.1.7 版標準拠の PDB-REPRDB[3] と PDB2000.7.1 版を用いて、RMSD 10 Å 以内、配列類似度 30% 以下の基準で代表的立体構造チェーン集合 (ベースセット, 1451 鎖 262331 残基) を作成した。また、配列類似度 50% 以下の基準でもチェーン集合 (プラスセット, 1959 鎖 341502 残基) を作成した。

次にベースセット、プラスセットそれぞれについて、MDGI 作成対象特微量を計算する。本稿で用いているのは、主鎖二面角 ϕ と、主鎖カルボキシル基の酸素原子の距離 OO_i である。ここで OO_i は、着目する残基 (第 0 残基) から C 末端側を + として i 残基離れた残基 (第 i 残基) との主鎖酸素原子間距離を表す。MDGI 作成対象特微量には、着目残基の周辺、数残基分の特微量を含める。この MDGI 作成対象特微量データセットを「構造窓」と称する。本稿では構造窓として $(OO_{-2}, OO_{-1}, OO_{+1}, OO_{+2})$ という窓と、対照用に ϕ の連続 4 残基分という窓 (どちらも $m=4$) を用いた。このような構造窓と周辺のアミノ酸配列との関係を調査するために、構造窓に着目残基周辺の配列 (「配列窓」と称する) を添付する。本稿では着目残基の前後 15 残基、合計 31 残基の配列窓を各構造窓に添付した。

続いて、ベースセットの構造窓について MDGI を作成する。作成には EM 法 [4] を用いた。MDGI は $f_m = \sum_{c=0}^{C-1} w^{(c)} f_m^{(c)}$ と表現される。 w は各正規分布 $f_m^{(c)}$ の

Modelling of Protein Structure Distribution using MDGI
Hidetoshi Tanaka, Masashi Shiraiishi, Kaori Kawakami, Isao Aoyama, Hiroyuki Sato
Real World Computing Partnership, Parallel Application Mitsubishi Laboratory, Mitsubishi Electric Corp.

¹Mixture of Diagonal covariant Gaussian Index

寄与度、 C は混合数である。これら C 個の正規分布 $f_m^{(c)}$ のどれにプラスセットの各構造窓が帰属するかを、 $w^{(c)} f_m^{(c)}$ が最大になる c によって決める。ここで求めた帰属正規分布が、後述の正解正規分布となる。本稿では混合数を $C=50$ とした (MDGI-50)。これは確率分布を 50 個の無相関正規分布の和で近似することを狙っている。

一方、各正規分布に帰属するデータの相対位置毎の残基分布を得るために、各構造窓に添付された配列窓に着目する。正規分布 $f_m^{(c)}$ に帰属するプラスセットの配列窓内の、第 i 残基の位置における各アミノ酸 a の頻度を $\hat{q}(c, i, a)$ と表す。この \hat{q} の位置に関する平均 $\mu(c, a)$ と分散 $\sigma^2(c, a)$ を求め、 $q(c, i, a) = (\hat{q}(c, i, a) - \mu(c, a)) / \sigma(c, a)$ によるプロファイル $q(c, i, a)$ を作成する。

このプロファイル $q(c, i, a)$ と各正規分布の寄与度 $w^{(c)}$ により、残基 a が相対位置 i にあったときの中で、最大の wq をとる c を a, i について網羅的に求めることができる。そのような正規分布 $f_m^{(c)}$ の中心を、残基 a 相対位置 i からの立体構造特微量（構造窓）の予測値とする。また、残基が相対位置 i から S 個、 a_0, \dots, a_{S-1} と並んでいた場合に、 $\sum_{s=0}^{S-1} w^{(c)} q(c, i+s, a_s)$ の最大をとる c によって、同様に着目位置の予測値を得ることができる。これらの予測値と、正解正規分布 $f_m^{(*)}$ の中心との距離が $\sqrt{0.1} \text{ \AA}$ 以内を正解とし、相対位置毎、並び幅 S （配列予測窓と称する）毎に正解率を求めた。

3 結果

OO の相対位置毎正解率を図 1 に、 ϕ の相対位置毎正解率を図 2 に示す。どちらの図も、横軸は相対位置 i 、縦軸は正解率を表している。配列予測窓の幅 $S=1, 2, 3$ における正解率を、それぞれ折れ線 1, 2, 3 で表示している。

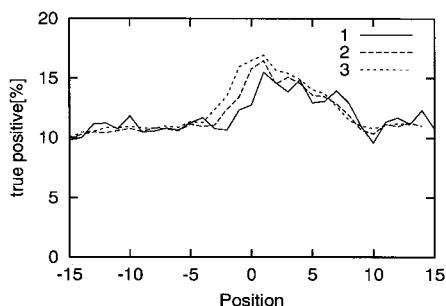


図 1: MDGI-50 による OO の予測精度

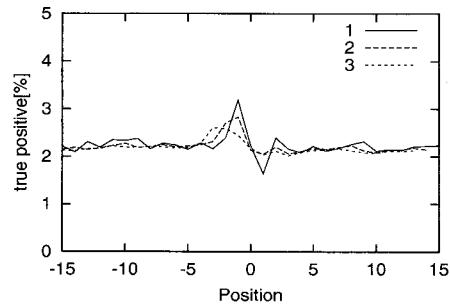


図 2: MDGI-50 による ϕ の予測精度

4 考察と課題

二面角 ϕ と主鎖の隣接残基の酸素間距離 OO_{+1} とは相関が高い。この結果では正解率に 1% 程度の幅しかない ϕ に対し、OO は着目残基の C 末端側 (+ 側) に +5 ~ 7% の大きな傾向が見られる。立体構造の性質上、着目残基を中心に対称性が期待されたが、この結果、特に OO についてははっきりとした非対称性を示しており興味深い。また、残基が遠くなると OO では 10% 程度、 ϕ では 2% 程度の正解率となる。これは、MDGI-50 によって実質的に OO はおよそ 10 分類、 ϕ は 50 分類されているために、遠い残基でもこの程度正解率が得られると考えられる。本稿では触れなかったが、混合数、構造窓、配列予測窓の各パラメータ、および予測用の評価式に関しては他の組合せでも実験を行い、同様の非対称性が確認できている。OO 以外の構造特微量についても同様の解析を行い、各構造特微量の分布を明らかにすることにより、タンパク質に特化した統計的ポテンシャル関数の設計を試みる。

参考文献

- [1] Dowe,D.L., Allison,L., Dix,T.I., Hunter,L., Wallace,C.S., and Edgoose,T. Circular Clustering of Protein Dihedral Angles by Minimum Message Length. Pacific Symposium on Biocomputing. 242–255, (1996).
- [2] Bystroff,C., Thorsson,V. and Baker,D. HMMSTR: a Hidden Markov Model for Local Sequence-Structure Correlations in Proteins. J. Mol. Biol. 301, 173–190. (2000).
- [3] Noguchi,T., Onizuka,K., Akiyama,Y., and Saito,M. PDB-REPRDB: A Database of Representative Protein Chains in PDB (Protein Data Bank). Proc. of the 5th Int. Conf. on Intelligent Systems for Molecular Biology. 214–217, (1997).
- [4] 渡辺, 山口. EM アルゴリズムと不完全データの諸問題. 多賀出版, (2000).