1M-6

ニューラルネットワークによるタンパク質立体構造予測の試み

渡邉 和之　　三枝 亮　　橋本 周司

早稲田大学理工学部応用物理学科

1. はじめに

タンパク質の構造と機能には密接な関係があり，アミノ酸残基配列情報からタンパク質の立体構造を予測することはバイオインフォマティクスにおける中心的な問題の一つである．構造予測問題の一般的な手法として，物理的エネルギーの最小状態を探索する手法[1]やホモロジーモデリング法が挙げられる．さらに近年，タンパク質の二次構造やアミノ酸埋もれ度などの中間的特徴量を予測し，それを用いて最終的な立体構造を予測する手法の有用性が主張されている．我々は中間的特徴量としてアミノ酸残基間距離マップに着目しタンパク質立体構造予測を試みている．

提案手法では，まずニューラルネットワーク(NN)を用いて距離マップを予測する．タンパク質の構造上，残基配列の一部分のみから残基間距離を予測するのは困難であり，全残基配列を入力とする必要がある．我々は可変長の残基配列すべてを入力とするため，リカレント型NN(RNN)を 2 つ導入し，N・C 両末端から遷移させ，距離マップを予測する．次に，得られた距離マップから多次元尺度構成法(Multidimensional Scaling: MDS)[2]を用いて各アミノ酸残基の 3 次元座標値を求める．提案手法は，距離マップが 3 次元空間の関係を満足する度合いを予測の信頼度の尺度となることが特徴である．本稿では提案手法の概要と実験結果について報告する．

2. 手法

2.1 距離マップ予測

提案手法は図 1 に示すグラフィカルモデルで表現され，以下に示す入力層，状態層，出力層からなる．

(1) 入力層：残基配列情報であり，各要素 $\boldsymbol{I}_t \in \boldsymbol{R}^k$ は位置 t における残基を表す．タンパク質を構成するアミノ酸は 20 種類あるため，1 残基は one-hot encoding により 20bit で表現し，k=20 とする．

(2) 状態層：入力配列の両末端からの遷移情報を保持するところであり，Feedforward 配列($\boldsymbol{F}_t \in \boldsymbol{R}^n$)と Backward 配列($\boldsymbol{B}_t \in \boldsymbol{R}^n$)からなる．状態遷移則は以下の式で示される．

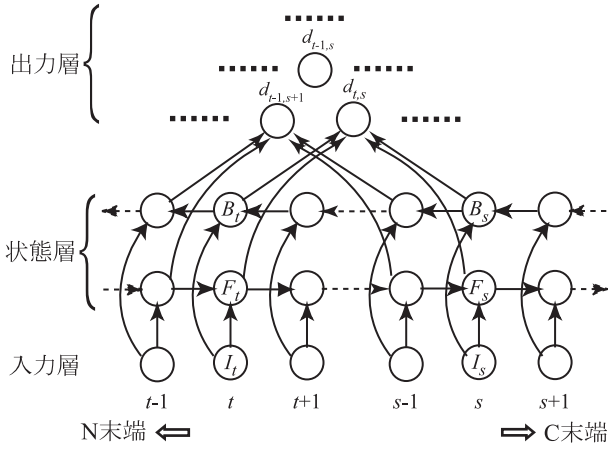

図1 グラフィカルモデル

$$\begin{aligned} F_t &= \phi(F_{t-1},\ I_t), \\ B_t &= \beta(B_{t+1},\ I_t). \end{aligned} \tag{1}$$

$\phi(.)$と$\beta(.)$は状態遷移関数である．状態 $\boldsymbol{F}_t$ は入力配列を N 末端から C 末端方向に遷移し，状態 $\boldsymbol{B}_t$ は C 末端から N 末端へ逆向きに遷移する．状態ペア($\boldsymbol{F}_t$, $\boldsymbol{B}_t$)は全入力配列を考慮した t 番目の残基の状態を表している．

(3) 出力層：各要素の出力 $d_{t,s}$ は t 番目と s 番目の残基間距離である．出力 $d_{t,s}$ は以下の式で計算される．

$$d_{t,s} = \eta(F_t,\ B_t,\ F_s,\ B_s). \tag{2}$$

$\eta(.)$は出力関数である．

以上の関数$\phi(.)$，$\beta(.)$，$\eta(.)$は NN の学習により獲得する．状態遷移関数$\phi(.)$と$\beta(.)$はジョルダン型 RNN N_ϕ ，N_βで構成し，出力関数 $\eta(.)$は階層型 NN N_ηで構成する．また，セル数を自由に定められる層のセル数に関し，以下の指標を用いる．

$n_h = N_\eta$ の中間層セル数；

$n_s = N_\phi$ と N_β の出力層セル数；

$n_{hs} = N_\phi$ と N_β の中間層セル数.

階層型のN_ηには Back Propagation 法で学習し，

The Trial of Protein Structure Prediction by Neural Networks

Kazuyuki Watanabe, Ryo Saegusa and Shuji Hashimoto
Department of Applied Physics, School of Science and Engineering, Waseda University

リカレント型の N_ϕ と N_β には Back Propagation Through Time 法を用いて学習する．

2.2 距離マップから 3 次元空間への配置

NN により得られた距離マップから MDS により各残基の 3 次元座標値を求める．MDS とは複数の対象の距離情報から多次元空間へ対象をマッピングする手法である．一般に対象が N 個あった場合，距離関係を完全に満たすには空間は N-1 次元必要であるが，本手法では第 1 主成分から第 3 主成分までを 3 次元座標値とする．

図 2 に提案手法の流れを示す．もし NN により予測した距離マップの精度が良い場合，この距離関係は 3 次元空間の関係を満足するため，3 次元空間への配置が矛盾無く行えるといえる．このため，NN の出力の距離マップ(図 2, b)と MDS の出力(図 2, c)から幾何学的に計算した距離マップ(図 2, d)が一致する．すなわち，NN の予測精度を距離マップの類似度(図 2, e)で計ることができる．ここでは，距離マップ b と d の平均二乗誤差の平方根を予測の非信頼度としている．

3. 実験

実験のためのデータとして，残基数が 11 から 20 のサンプル 79 個を用い，学習用・試験用に 64 個と 15 個にランダムに分割した．各階層のセル数は経験的に $(n_h, n_s, n_{hs}) = (20, 20, 20)$ とした．最終的な予測構造の精度の指標として，予測構造と正解構造からそれぞれ幾何学的に計算した距離マップの標準偏差を用い，これを予測誤差とする．

図 3 に試験用サンプル 15 個に関する予測の誤差と非信頼度の関係を示す．相関係数は 0.69 であり，誤差と非信頼度にある程度相関があることが判る．図 4 に予測構造の一例を示す．一般手法の例としてタンパク質構造予測コンテスト CASP6 (2004) NF 部門で上位の成績を収めた予測サーバーROKKY により，エネルギー探索手法[1]を主に用いた結果を挙げた．残基数の範囲が非常に限られてはいるが，誤差の平均値でエネルギー探索手法を上回る結果を得た．また，一般手法では予測の誤差が大きいかどうかは正解構造と比較して初めてわかるが，提案手法では予測の非信頼度をもとにある程度見当をつけることが可能である．

4. まとめ

NN を用いて残基間距離マップを予測する手法を提案した．距離マップを予測する際，可変長である残基配列を入力とするため，RNN を双方向に遷移させる方法を導入した．さらに得られた距離マップから MDS により 3 次元座標値を得た．また，予測の非信頼度を導入し，実験により予測構造の誤差と非信頼度の相関を確認した．

今後，残基数の範囲を拡大して実験を行うこと，ならびに学習の高速化を検討する予定である．

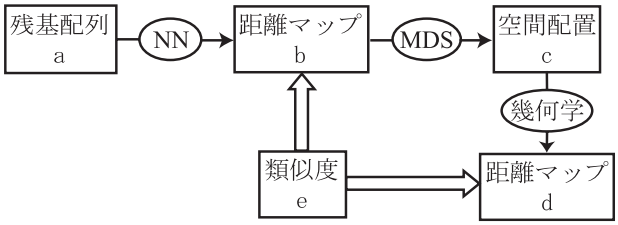

図2 提案手法の流れ

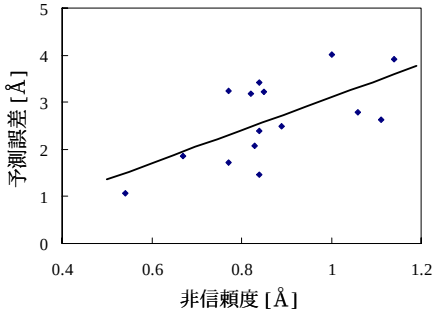

図3 誤差と非信頼度

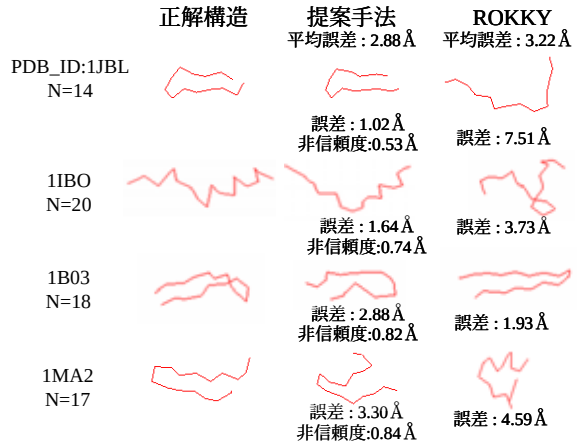

図4 構造比較

参考文献

[1] Y. Fujitsuka, S. Takada, Z. A. Luthey-Schulten, and P. G. Wolynes (2004), “Optimizing Physical Energy Functions for Protein Folding,” Proteins 54, 88-103.

[2] W. S. Torgerson, “Theory and methods of scaling,” New York, Wiley, 1958