

ニューラルネットワークによる タンパク質三次構造予測の試み

渡辺 和之[†] 三枝 亮[†] ピトヨ ハルトノ[‡] 橋本 周司[†]

早稲田大学理工学部[†] 早稲田大学WABOT-HOUSE研究所[‡]

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、アミノ酸配列情報からタンパク質の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。従来、ニューラルネットワークを用いてタンパク質の二次構造を予測する研究¹⁾が行われていたが、タンパク質の三次構造予測する試みは数少ない。

そこで我々は階層型ニューラルネットワークを用いてタンパク質三次構造を予測する手法を検討している。タンパク質の三次構造がアミノ酸の一次配列のみにより決定されることから、アミノ酸配列とタンパク質の構造の関係にある種の関数としてとらえ、ニューラルネットワークの学習を行うことにより三次構造を予測することが可能であると考えられる。本稿ではその試みについて報告する。

2. 手法

図 1 に示すような三層階層型ニューラルネットワークを用いる。中間層、出力層の応答関数としてはシグモイド関数を用いる。

入力アミノ酸配列情報である。生体を構成するアミノ酸は 20 種類存在するため、1 つのアミノ酸座に対して 20 個の入力セルを用意する。その座のアミノ酸の種類に対応するセルを 1 とし、他の 19 個のセルは 0 にする。したがって、アミノ酸残基数が N の場合、入力は $20 \times N$ 桁のビット列で表現する。

出力はタンパク質を構成する各々のアミノ酸残基の三次元極座標 (r, θ, ϕ) で表現する。各 (r, θ, ϕ) は隣接する残基から見た相対的な極座標値を用いる。したがって、出力層で立体構造は $3 \times N$ 次元の実数配列で表現される。また、応答関数にシグモイド関数を用いているため、 $(0, 1)$ の範囲で正規化を行う。

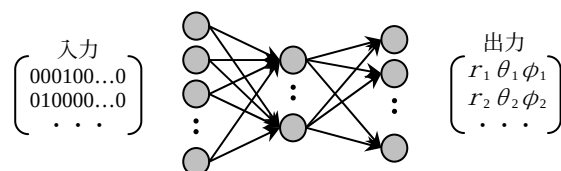


図 1 ネットワークの構成

実際にネットワークを構成する際は学習・試験に用いるデータの中で残基数が最大のものに合わせて入力層、出力層を用意する。それにより残基数が少ないデータの場合、入力層の余ったセルには全て 0 を入れることで学習には関与させないようにする。短い残基列の場合は、アミノ酸配列をシフトすることによりサンプルを増やすことができるので、最大残基数 n の場合、残基数 m のタンパク質からはサンプルを $n-m+1$ 個構築できる。学習に用いるデータとして PDB(Protein Data Bank)にある実測データを使用し、これを学習用、試験用に分割し、Back Propagation 法で学習を行った。

3. 評価実験

3.1 出力誤差の評価

本研究では学習・試験に用いるデータセットとして、アミノ酸残基数が(a)18, (b)13から18, (c)21から60の3種を用いた。それぞれに含まれるタンパク質数は12個, 53個, 720個であり、作成したサンプル数は12, 159, 12458となる。サンプルを試験用、学習用に分割する際、(a)は試験用1個, 学習用11個に分け、全ての分割の仕方における結果の平均をとった。(b), (c)はそれぞれ8分割, 100分割し、そのうち1つを試験用にし、残りの全てを学習用とした。それぞれ中間層のセルを1個, 10個, 100個, 1000個の4通りで実験を行った。学習の評価関数は教師信号との平均2乗誤差を用いた。学習が収束した後の教師信号との誤差の標準偏差と、試験用データのRMS値を図2に示す。RMS値とは、隣接残基からの相対座標ではなく、絶対座標に変換した立体構造のずれの標準偏差である。

The Trial of Protein Tertiary Structure Prediction with Neural Networks

[†] Kazuyuki Watanabe, Ryo Saegusa and Shuji Hashimoto
Dept. of Science and Engineering, Waseda University

[‡] Pitoyo Hartono

WABOT-HOUSE Laboratory, Waseda University

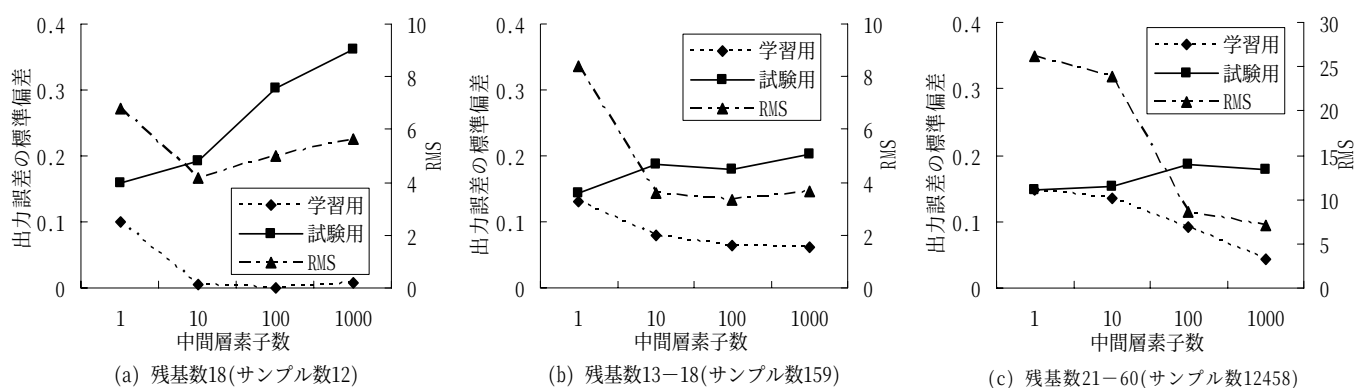


図2 学習完了後の誤差

図2の各々のデータセットの結果において、中間層素子数を増やすと学習用データの誤差が減少するが、汎化能力が劣化して試験用データの誤差が大きくなっている。一方、中間層素子数を増やすとRMS値は小さくなっている。この結果から、中間層素子数が少ないと局所的な関係しか学習できず、中間層素子数が多いと局所的な誤差は多少増えるが、大域的な関係がある程度学習できているといえる。

予測システムとしては、あらゆる残基数のタンパク質の構造予測が行えることが好ましい。しかし、図2(a)~(c)の比較から、残基数の多いタンパク質を扱うと、RMS値を下げるためにはより多くの中間層素子数が必要になる。その際には学習にかかる計算量が膨大になるという問題が生じるが、一度学習が終了すれば構造予測は高速に行えるため学習に要する時間は大きな問題ではない。

3.2 予測構造の評価

図3と図4に残基数13-18のデータセットにおける中間層素子数10と100の場合の誤差が最小と最大の予測構造と実際の構造の比較を示す。

誤差の大きい図3(b)と図4(b)は不規則で複雑な構造をしている。それに対し、誤差の小さい図3(a)と図4(a)を見ると、どちらも2本の残基列が平行に並ぶβシートの構造が見られる。本研究で用いた階層型ニューラルネットワークは二次構造予測において一般的に用いられていることから、二次構造のような規則的なパターンの部分は、他の不規則な部分に比べ正確な予測ができていと考えられる。また、どちらもU字型の構造をとっていることから、学習用データセットにU字型の構造を持つものが多いことも原因の一つであると推測できる。

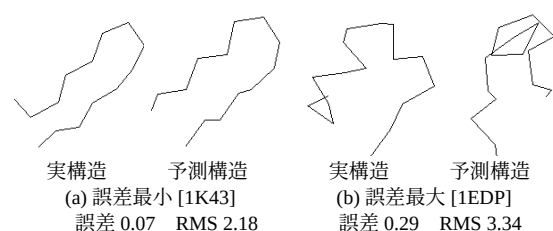


図3 残基数13-18・中間層素子数10の構造比較

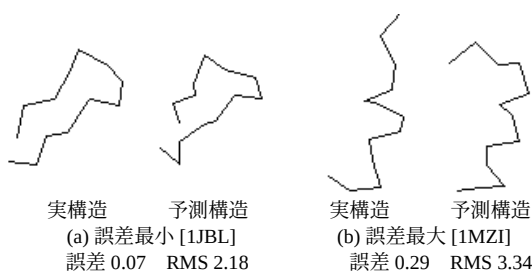


図4 残基数13-18・中間層素子数100の構造比較

4. まとめ

階層型ニューラルネットワークを用いてタンパク質三次構造予測を試みた。結果の良いもので誤差が0.15程度と予測精度は必ずしも良好とはいえないが、出力結果をab initio法のエネルギー最小化計算²⁾の初期値として用いて極小解に陥ることを避けることができていると考えている。

参考文献

- 1) L. Howard Holley and Martin Karplus. Protein secondary structure prediction with a neural network. Proc. National Academy of Sciences USA, Vol.86, 152-156, 1989.
- 2) 廣安知之, 三木光範, 小椋信弥, 青井桂子, 吉田武志, 岡本祐幸. 遺伝的交叉を用いた並列SAによるタンパク質立体構造のエネルギー最小化. 情報処理学会 SACSIS'03 講演論文集, pp.341-348, 2003.