タンパク質の二次構造予測を行なうニューラルネットワークにおける汎化能力の向上

Comparison of generalization methods for multilayer neural network applied to predicting protein secondary structure

福村 健一† 中山 謙二‡ 平野 晃宏‡
ken-ichi fukumura Kenji Nakayama Akihiro Hirano

†金沢大学大学院 自然科学研究科 電子情報システム専攻
Division of Electronics and Computer Engineering
Graduate School of Natural Science and Technology, Kanazawa Univ.
‡ 金沢大学大学院 自然科学研究科 電子情報科学専攻
Division of Electrical Engineering and Computer Science
Graduate School of Natural Science and Technology, Kanazawa Univ.
E-mail: nakayama@t.kanazawa-u.ac.jp

あらまし

アミノ酸配列からタンパク質の二次構造を予測する階層形ニューラルネットワークにおいて、汎化能力の向上について検討を行った。学習データにも依存するが、過学習が起こりやすい。しかし，学習に要する時間が膨大であり、学習データを増やすことなく汎化能力を高めることが望ましい。学習データが限られているので、学習データの周りで領域を広くカバーする必要がある。これは、学習中に活性化関数の傾斜を緩やかに制御することにより可能である。そのため、学習係数の制御、学習データに小さな雑音を混入、学習中における重みの減衰制御等について検討を行った。その結果、これらの方法は全て汎化能力を高めることができた。中でも、重みの抑制制御がもっとも高い予測精度を実現した。

ABSTRACT

Generalization methods for learning of multilayer neural networks, applied to predict protein secondary structure using a multiple alignment of amino acid as the input data is investigated. Although depending on the training data set, over learning is easily occurred. However, the training needs a very long training time, therefore it desirable to achieve high generalization except for increasing the training data. Several view points, including a learning rate, a momentum term, the number of hidden units, adding noise to the input data, modular networks and weight decay, are taken into account. Since the training data are rather limited, it is necessary to extend a region around the training data. This can be done by relaxing sloop of activation functions. For this purpose, a small learning rate, adding random noise to the training data, weight decay in a learning process and so on are investigated. As a result, these techniques are useful. Among them, the weight decay method can provide the highest accuracy of prediction.

1 まえがき

人間の体は少なくとも 10 万個以上の異なったタンパク質の合成によって作られている。タンパク質はアミノ酸と呼ばれる物質がいくつもつながって作られる複雑な分子である。タンパク質鎖を生成するアミノ酸配列を一次構造と呼ぶ。アミノ酸配列はαヘリックスやβシートなどの二次構造を形成する。そして、いくつかの領域の中に二次構造が結合されることによってタンパク質の立体構造が形成される [1]。タンパク質の役割や立体構造を理解するうえで二次構造予測は有用である。従って、二次構造予測はゲノムサイエンスの重要なテーマの一つとなっている。今日多くのコンピュータによる予測ツールが開発されてきた。その中でもニューラルネットワーク法は最も精度の高い方法の一つである [2]-[5]。タンパク質立体構造はアミノ酸配列によって一意に決まる。予測では局所的なアミノ酸配列を入力とし、出力層

はその配列の中央のアミノ酸が置き換わる二次構造クラスを表す3つのユニットで構成する。入力データは、正答率の改善に非常に重要であるとされている生物の進化的な情報を付加することができるマルチプル配列アラインメントを用いて生成する [6]。

本稿では、予測精度を改善するためにニューラルネットワークの学習プロセスの最適化を行なう。ニューラルネットワークを用いた二次構造予測では過学習が重要な問題となる。よって、学習プロセスやパラメータの最適化による過学習の抑制と汎化能力の改善を行なう。

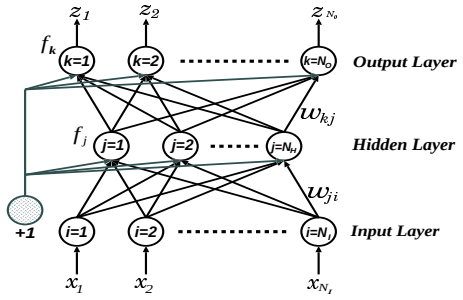

図 1: 階層形ニューラルネットワーク

2 ニューラルネットワークによる二次構造予測

2.1 データの準備

入力データはftp://ftp.ebi.ac.uk/pub/databases/hssp/から得られたHSSP(homology-derived structures of proteins) ファイルに記録されているデータ情報を使用した [7]。HSSP ファイルが含んでいる予測データに使用したい項目を以下に示す。

- 二次構造に由来する既知構造のタンパク質の配列
- ターゲットのタンパク質と構造的に相同性があると考えられるアラインメントされた配列
- 保存の度合を示す重み

二次構造と溶媒露出度に関する DSSP(Dictionary fo Secondary Structure assignment of Profile) データベースによると、二次構造状態は８つの種類に区別される [8]。それら８つの種類は、次の３つのクラスにグループ分けされる。H(αヘリックスなど)、E(βシートなど)、L(その他) である。HSSP は、DSSP の二次構造アラインメントを取り入れているので、HSSP の分類はDSSP と同様になる。ニューラルネットワークは、テスト配列の各アミノ酸に対応する二次構造状態の種類 (H,E,L) を予測する。

2.2 データの操作

本稿では図 1に示すような階層形ニューラルネットワークを使用する [6]。

入力と出力データの操作を図 2に示す。入力層はアミノ酸配列パターンから成る。WTKC... はアミノ酸を指す。入力窓の幅 w は 9 とする。よって、c 番目のアミノ酸の二次構造を予測するために $c-4$ から $c+4$ 番目の配列のアミノ酸情報を用いる。マルチプルアラインメントから生成されるプロファイルは、アミノ酸を表すために使用される。マルチプルアラインメントは、HSSP ファイルから得ることができる。アミノ酸は 20 種類あり、更に予測に有用な 2 種類の情報が付加される。従って、1 つのアミノ酸には 22 個の入力ユニットが使用され、一度の予測で 9 個のアミノ酸を使用するので入力ユニットは $(20+2)*9$ 個となる。出力層は 3 個 (H,E,L) となる。教師データは (H,E,L)=(100,010,001) に設定する。教師データは 9 個のアミノ酸配列中の中央のアミノ酸の二次構造を表す。

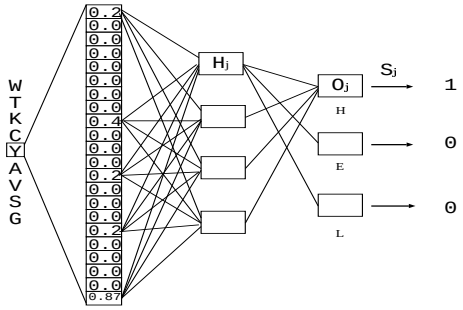

図 2: 入力と出力データの操作

2.3 予測精度の測定法

以下に示す最も広く使用されている評価方法を採用する。

$$Q_3 = 100 * \sum_{i=1}^{3} c_i/N \qquad (1)$$

ここで c_H,c_E,c_L は H、E、L のそれぞれの状態で正しく予測されたアミノ酸の数、N は全体のアミノ酸の総数である。

3 学習プロセスの最適化

3.1 学習とテストに用いるデータ

HSSP ファイルから得られる 25%以下の類似性を持つ 35 個のタンパク質を学習とテストに使用した [6]。アミノ酸の総数は 5964 個である。ランダムに 596 個のアミノ酸をテストに、残りを学習に使用した。

3.2 学習回数の影響

隠れユニット数を 5 個、学習係数 η を 0.001、慣性項なしでシミュレーションを行なった。学習曲線と学習データの正答率、テストデータの正答率をを図 3、4、5 に示す。出力誤差は単調に減少し、それに比例して学習データの正答率は増加している。しかし、テストデータの正答率はある点を境に徐々に減少している。これが過学習である。正答率が減少し始める点を見つけることは非常に困難である。

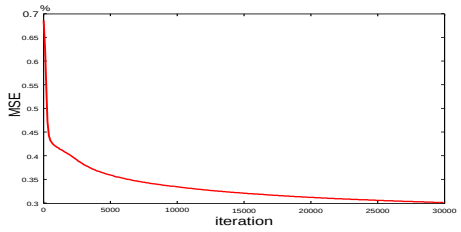

図 3: 隠れユニット 5 個、学習係数 0.001 での出力誤差 MSE

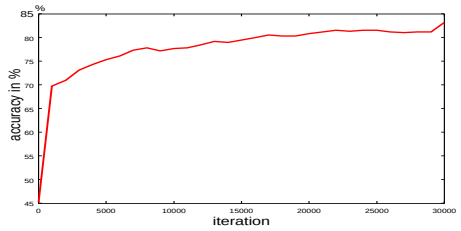

図 4: 隠れユニット 5 個、学習係数 0.001 での学習データの正答率

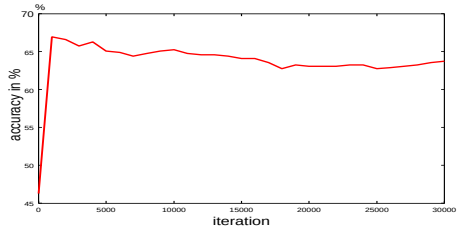

図 5: 隠れユニット 5 個、学習係数 0.001 でのテストデータの正答率

3.3 隠れユニット数の影響

隠れユニット数を 1 から 10 個まで変更してその影響を調べる。出力誤差と全体の正答率を図 6、7に示す。出力誤差は隠れユニット数が増える毎に減少している。しかし、テストデータの正答率Q_3はユニット数 10 個の場合よりも 5 個、3 個の場合の方が良くなっている。

3.4 慣性項の影響

慣性項を加えてシミュレーションを行なった。図 8、9 に出力誤差とテストデータの正答率を示す。慣性項の比重は 0.9 とする。この場合でも過学習が起こっている。

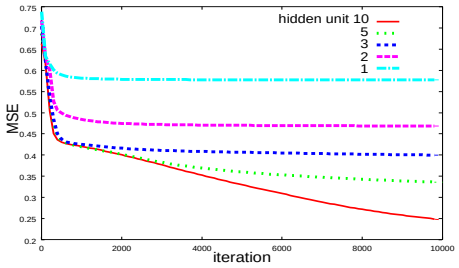

図 6: 隠れユニット数の違いによる出力誤差

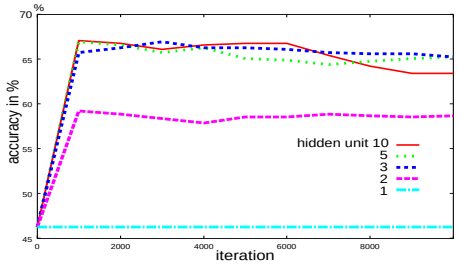

図 7: 隠れユニット数の違いによるテストデータの正答率

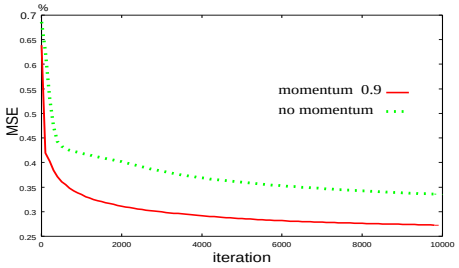

図 8: 慣性項の有無における出力誤差

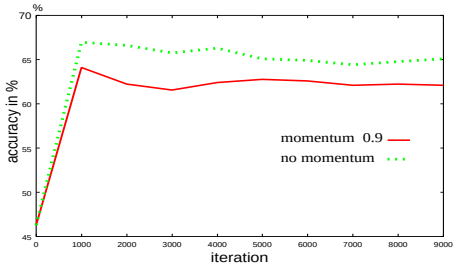

図 9: 慣性項の有無におけるテストデータの正答率

3.5 学習係数の影響

学習係数ηを制御する。出力誤差とテストデータの正答率を図 10、11 に示す。学習係数$\eta = 0.00001$ を用いた場合、出力誤差は非常に緩やかに減少し、誤差が大きい。しかし、正答率は学習回数 35000 回以降は非常に安定している。

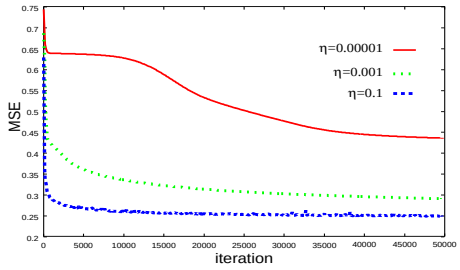

図 10: 学習係数 η の違いにおける出力誤差

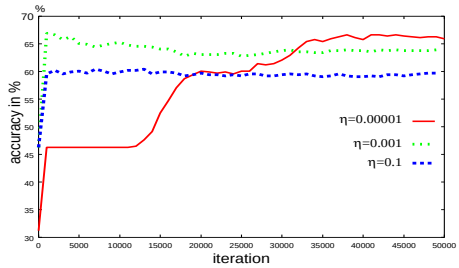

図 11: 学習係数 η の違いにおけるテストデータの正答率

3.6 ランダムノイズの付加

学習係数を小さい値に設定すると学習が遅くなる。そこで、学習係数を小さくせずに正答率を改善するために入力データに小さな乱数を加える。入力データは 0 から 1 の値で分布しているので、乱数は-0.1〜0.1 の間で生成する。学習する毎に異なった乱数を加える。学習曲線とテストデータの正答率を図 12、13 に示す。入力データに乱数 (margin) を加えることにより、学習曲線はうまく減少しないがテストデータの正答率は良くなっている。

3.7 重みの抑制制御

シグモイド関数の傾きを緩やかにすることによって学習データの周りで同じデータに分類される範囲を広め、汎化能力の向上を図る。データの分離と重みの抑圧の両方を考慮するために Weight Decay 法を用いる [9]。コストファンクションを以下に示す。

$$Q = E^2 + \frac{\lambda}{2}\{\sum_{i=0}^{I}\sum_{j=0}^{J} w_{ij}^2 + \sum_{j=0}^{J}\sum_{k=0}^{K} w_{jk}^2\} \qquad (2)$$

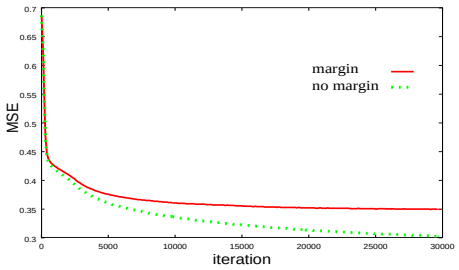

図 12: 入力データに乱数を付加した場合 (margin) の出力誤差

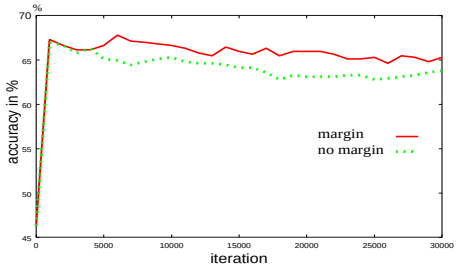

図 13: 入力データに乱数を付加した場合 (margin) のテストデータの正答率

その時の更新式を以下に示す。

$$w_{ij} = w_{ij} - \eta\frac{\partial E^2}{\partial w_{ij}} - \eta\lambda w_{ij} \qquad (3)$$

$$w_{jk} = w_{jk} - \eta\frac{\partial E^2}{\partial w_{jk}} - \eta\lambda w_{jk} \qquad (4)$$

Q はコストファンクション、E^2は二乗誤差、w_{ij}, w_{jk}は入力-隠れ層間の重み、隠れ-出力層間の重み、ηは学習係数、λはペナルティ項の影響を制御するパラメータである。$\eta = 0.001$ とし$\lambda = 0.001, 0.0001, 0.00001$ と変えたときの学習曲線とテストデータの正答率の比較をそれぞれ図 14、15 に示す。λが大きくなると学習曲線は減少しなくなっていく。正答率は$\lambda = 0.001, 0.00001$ のときは減少しているが、$\lambda = 0.0001$ のときは高くなっている。

次に$\lambda = 0.0001$ とし、$\eta = 0.01, 0.001, 0.0001, 0.00001$ と変えたときの学習曲線、テストデータの正答率の比較をそれぞれ図 16、17に示す。ηが小さくなると学習曲線の収束速度が遅くなっている。正答率は学習回数 35000 回以降はどのηの値の場合も非常に安定している。

3.8 ネットワークの分離

各二次構造状態 (H,E,L) を 3 個のネットワークを用いて別々に学習する。各ネットワークの出力の最大値を最終的な予測二次構造とする。H、E、L それぞれを予測した各ネットワークの誤差曲線とテストデータの正

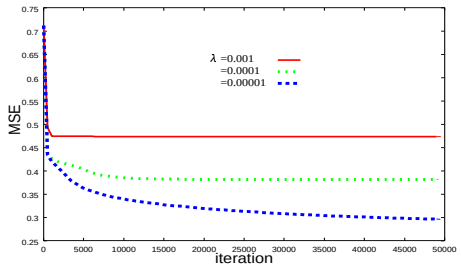

図 14: Weight Decay でλを変えたときの出力誤差の比較

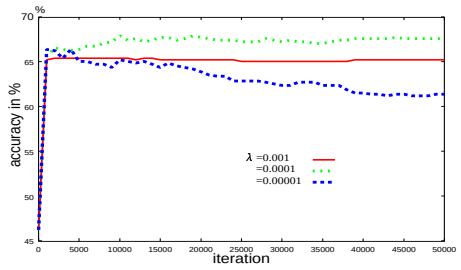

図 15: Weight Decay でλを変えたときのテストデータの正答率の比較

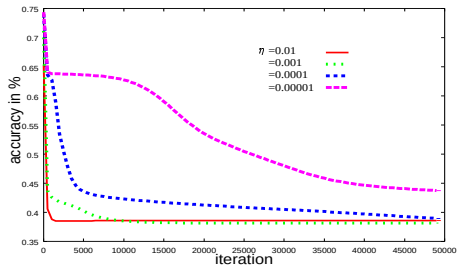

図 16: Weight Decay でηを変えたときの出力誤差の比較

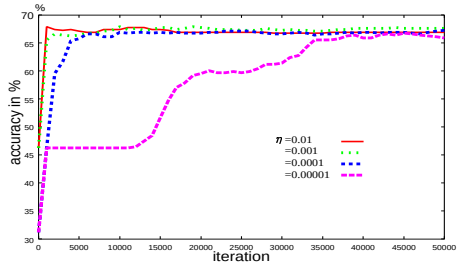

図 17: Weight Decay でηを変えたときのテストデータの正答率の比較

答率、最終的な予測二次構造のテストデータの正答率 (network separation) をそれぞれ図 18、19、20に示す。比較のため同じ条件で予測を行なったネットワークを分離しない場合の正答率 (no separation) を図 20に示す。誤差曲線、各ネットワークの正答率ともにこれまでより良い値に収束しているが、最終的な予測二次構造の正答率はネットワーク分離を用いない場合の方が高くなっている。

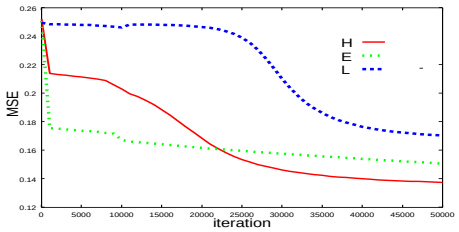

図 18: 3 個のネットワークそれぞれの誤差曲線

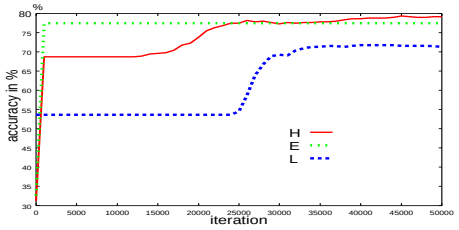

図 19: 3 個のネットワークそれぞれのテストデータの正答率

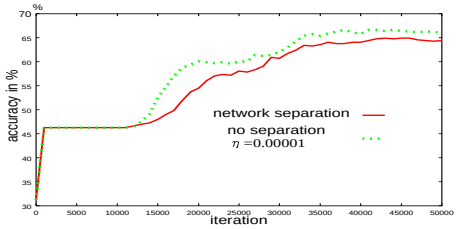

図 20: 3 個ネットワークによる最終的なテストデータの正答率

次に学習済みのネットワークを通して誤差の大きさによりそれぞれ 2 個のネットワークにデータを振り分け別々に学習を行なった。最終的な予測二次構造は 2 個のネットワーク出力の最大値を採用した。誤差 0〜0.6 のデータを 1 個目のネットワークに、誤差 0.2〜1 のデータを 2 個目のネットワークに振り分けた。振り分けられたデータ数はそれぞれ 3985 個、3546 個となった。MSE はそれぞれ 0.0519、0.0624 と良い値になったが、全体の正答率は 35.7%と非常に低くなった。

4 最適化に用いた各方式の比較

予測精度を改善するために文献 [6] ではマルチモデルニューラルネットワーク (MNN) が提案されている。MNN ではいくつかのネットワークを並列にしてデータを入力し、各ネットワークからの出力の多数決で予測二次構造を決定する。予測精度は 1 個のネットワークを用いた場合よりも約 7%改善され、66%に達している。本稿では最適化された 1 個のネットワークを用いて同程度の正答率を達成した。文献 [6] で示されている 1 個のネットワークを用いた場合の予測精度と本稿で最適化した各方式の予測精度の比較を図 21に示す [6]。(A) は文献における 1 個のネットワーク、(B) は$\eta = 0.00001$ のときの 1 個のネットワーク、(C) は入力にランダムノイズを付加した場合 ($\eta = 0.001$)、(D) は Weight Decay 法を用いた場合 ($\eta = 0.001, \lambda = 0.0001$)、(E) は各二次構造で学習するネットワークを分離した場合である。(A) を基準として (B)(C)(D) は約 58%から 66%に予測精度が改善されている。(E) は (A) に比べて精度は改善されているが他の方式に比べ精度は低くなっている。

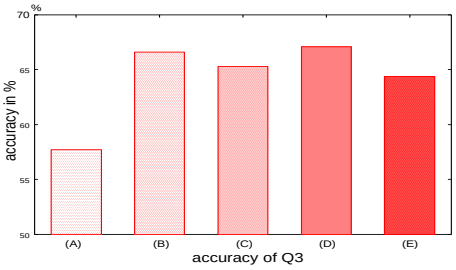

図 21: (A) 文献 [6](B)$\eta = 0.00001$ (C) ランダムノイズを付加 ($\eta = 0.001$)、(D)Weight Decay 法 ($\eta = 0.001, \lambda = 0.0001$)、(E) ネットワークを分離した場合の予測精度の比較

MNN では 5 個のネットワークを独立に学習させ、予測二次構造は多数決によって決定されている。この過程は平均化によって汎化能力が改善されることと同じである。従って、(B)(C)(D) の方式を用いることによって 1 個のネットワークでも同じ効果を得ることができる。

5 結論

タンパク質二次構造予測においてニューラルネットワークの学習は容易に過学習に陥る。過学習を避けることが予測精度を改善するための有益な方法の一つである。本稿では汎化能力を高めるためにいくつかの項目を調査した。隠れユニットの数は十分な正答率を得るために 3 個以上必要であるが少ない方が良い。学習データの周りで領域を広くカバーするために学習中に活性化関数の傾斜を緩やかに制御することを検討した。学習係数を非常に小さい値にすることで汎化能力が向上した。また、学習データに小さな雑音を混入、学習中における重みの減衰制御を行なうことで汎化能力が高まることを確認した。それらの学習の最適化によって 1 個のネットワークで 5 個のネットワークを必要とする MNN と同程度の約 66%の予測精度を達成した。特に、重みの減衰制御はわずかながら他の方式よりも高い予測精度を実現した。

参考文献

[1] D.M.Mount,"Bioinformatics:Sequence and Genome Analysis",Cold spring Habor Laboratory Press,New York 2001.

[2] TM Yi and S.Lander,"Protein secondary structure prediction using neirest-neighbor methode",J Mol Biol,232,pp.117-1129,1993

[3] AA Salamov and VV Solovey,"Prediction of protein secondary structure by combining neirest-neighbor algorithms and multiple sequence alignment",J Mol Biol,247,pp.11-15,1995.

[4] R.Burkhard and S.Chris,"Prediction of protein secondary structure at better than 70% accuracy ",Acadimic Press Lomited,pp.587-599,April 1993

[5] D.G.Kneller,F.E.Cohen and R.Langridge,"Improvements in protein secondary structure prediction by enhanced neural networks",J Mol Biol,214,pp.171-182,1990

[6] H.Zhu,I.Yoshihara and K.Yamamori,"Prediction of protein secondary structure by multi-modal neural networks",IEEE&INNS,Proc.IJCNN2002,pp280-285,May.2002

[7] C.Sander and R.Schneider,"Database of homology-derived structures and the structural meaning of sequence alignment",Protein Struct.Funct.Genet,9,pp.56-68,1991

[8] W.Kabsch and C.Sander,"Dictionary of protein secondary structure:Pattern recognition of Hydrogen bonded and geometrical features",Biopolymers,22,pp.2577-2637,1983

[9] Hanson,S.J. and Pratt,L.Y.,"Comparing biases for minimal network construction with back-propagation " In D.S.Touretzky, ed. Advances in Neural Information Processing Systems 1, pp.177-185, Morgan kaufmann.1989

[10] K.Nakayama, A.Hirano and K.Fukumura, "On generalization of multilayer neural network applied to predicting protein secondary structure", Proc. IEEE&INNS, IJCNN2004, Budapest, pp.1209-1213, July 2004.