1W－6

Time-of-reduction-recovery 予測による ニューラルネットワークの酵素分類への適用

久原泰雄　清水謙多郎　土井淳多*
東京工芸大学女子短期大学部　東京大学応用生命工学専攻　千葉工業大学情報工学科

1　はじめに

隠れ層のユニット数を増減させる3層型逆伝播法ニューラルネットワークを使用して酵素機能を予測した．ネットワークはアミノ酸配列を入力とし，酵素機能を予測する．本研究では，学習性能に応じて増減するユニット数を監視し，最適なネットワークを抽出する Time-of-reduction-recovery 予測を提案した．これまで，隠れ層のユニット数を変化させて最適サイズを決定する手法が考案されてきたが，本手法によって得た最適なネットワークによって，より高い予測性能を達成することができた．

2　方法

酵素データ　酵素は EC 番号によって6クラスに分類される．表1に各クラスの本研究で使用した学習用酵素数とデータベース内に含まれる酵素数を示した．クラス毎の学習用およびテスト用データの酵素数はデータベースの数と同じ比率にした．学習用酵素数の合計は 100，テスト用は 50 である．テスト用酵素データは学習データに対するホモロジーが異なる複数のグループを使用した．

ネットワーク構成　本モデルは逆伝播法の3層ネットワークであり，出力ユニットの誤差が許容値を越えると隠れ層のユニット数が1つ増加する．また，ある一定時間ユニットの増加が発生しないとユニットが一定数（2個）減少し，再びユニットの増加が始まる．従って，隠れ層のユニット数は絶えず増減する．ネットワークは酵素の全アミノ酸配列を入力とし，酵素機能を予測する [2,3]（図1参照）．入力層は 220 (20 × 11) ユニットを持ち，タンパク質

表 1: 酵素機能と使用データ

EC 番号	酵素クラス	train	database
1	Oxidoreductase	25	1600
2	Transferase	28	1763
3	Hydrolase	31	1923
4	Lyase	10	646
5	Isomerase	2	147
6	Ligase	4	276
	合計	100	6355

を構成する 20 種類のアミノ酸が N 末端から C 末端まですべて入力されるが，一度に入力されるのはウィンドウ内の 11 のアミノ酸だけである．このウィンドウが主鎖上を走査して，全アミノ酸が入力される．出力層は4ユニット持ち，各ユニットは酵素クラスを表す．EC 番号 4,5,6 のクラスは絶対数が少ないため，1つのグループにした (表1参照)．ウィンドウの中央のアミノ酸に対して4つの出力ユニットが出力値を持つ．1つの酵素に含まれる全アミノ酸に対して，各出力ユニットの出力値の合計値に基づいて酵素クラスを予測する．

3　結果と考察

ユニット増減型のネットワークを隠れ層のユニット数が 10,30,50,70 であるネットワークと比較した．

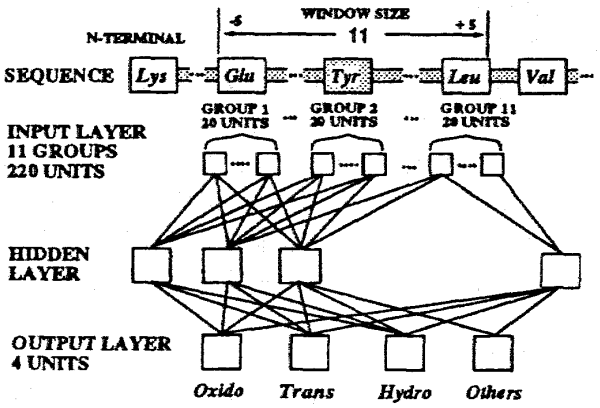

図 1: ネットワーク構成

*"Enzyme Classification Using Neural Networks of Time-of-reduction-recovery Estimation", Yasuo Kuhara, Tokyo Institute of Polytechnics, 2184 Iiyama, Atsugi, Kanagawa 243-02, Japan, Kentaro Shimizu, Department of Biotechnology, The University of Tokyo, and Junta Doi, Department of Computer Science Chiba Institute of Technology

テストデータとして学習データに対するホモロジーが10から30%の酵素を使用した．固定ユニット型では50ユニットのネットワークが最も高い性能を示した．図2(a)に3000イタレーションまでの50ユニット固定型とユニット増減型の予測正答率を示した．固定型では，2800イタレーションで最大値58%，1040イタレーションで最小値36%の正答率であるが，どの時点のネットワークが高い予測性能を持つかを判断することは困難である．1000から1700イタレーション間での平均の正答率は48.6%である．

ユニット増減型では，ユニット数が60前後で絶えず緩やかに変化し，正答率も50%前後で振動している．ユニット減少の後にユニット数が回復しているが，正答率は減少直後よりも高い値になっている．隠れユニット数の変化によって，極小値を脱出し，過学習の回避が可能となる．例として1140と1640イタレーション付近は各々ユニット数が59，61であるが，正答率の動きを観察すると，ユニットが減少した後，ある一定の成熟時間を経た後で各々50.0，56.0%という高い正答率が得られている．そこで，ユニット減少後，200イタレーションの成熟時間を経たネットワークの正答率の平均は53.0%となり，160イタレーションの場合は55.0%となる．成熟時間が160, 180, 200イタレーションであるネットワークの総平均正答率は，53.0%となる．

3.1 Time-of-reduction-recovery 予測

ユニットの増加は学習データに対してネットワークの容量が不足している場合に効果的であり，ユニット間の結合に振動を与え，極小値からの脱出および過学習の回避にも寄与する．ユニットの減少は学習データに対してネットワークが大きすぎる場合に，適切な大きさに収縮する効果があり，学習データに依存した結合を持つユニットを削除して，より大きな振動を与えるため，過学習の回避の点でさらに効果的である．

この例では，ユニット減少後，160から200イタレーションの成熟時間を経たネットワークが最も良い正答率を持つ．表2に，1000から1700イタレーション間におけるTime-of-reduction-recovery予測と固定型ネットワークの正答率の平均値を示した．本手法によって4.4%高い正答率が可能となった．

4 結論

ネットワークの最適サイズを決定するには試行錯誤が必要であったが，ユニットの増減方法や計算負荷に改善の余地があるものの，Time-of-reduction-recovery予測によって性能の良いネットワークを抽出することが可能となった．

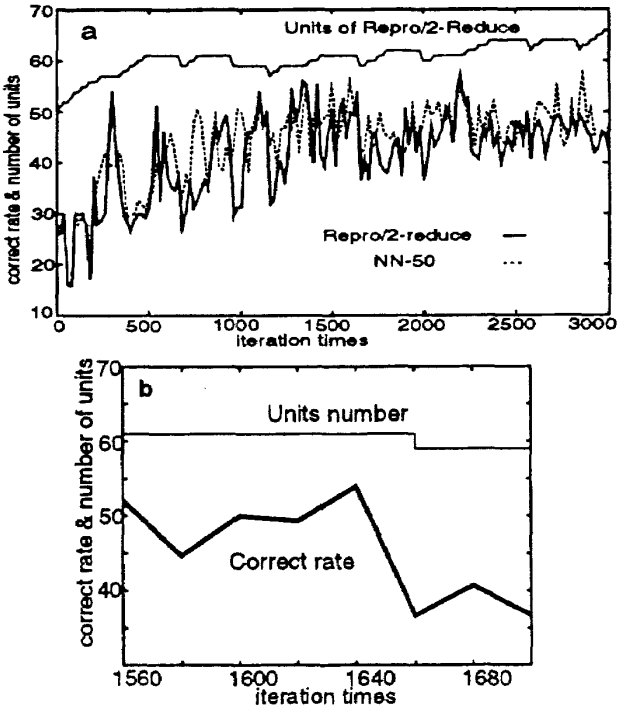

図 2: ユニット増減型と50ユニット固定型の正答率(a)．1640イタレーション付近の拡大図(b)．

Table 2: ネットワークごとの正答率

ホモロジー	10-20%			20-30%		
	最大	平均	最小	最大	平均	最小
50(固定)	44.0	36.5	28.0	58.0	48.6	36.0
70(固定)	40.0	34.2	26.0	60.0	44.8	26.0
本研究	37.5			53.0		

参考文献

1. E. B. Bartlett. Dynamic Node Architecture Learning: An Information Theoretic Approach, *Neural Networks*, 7(1):129-140, 1994

2. Y. Kuhara, K. Shimizu, and J. Doi, Chymotrypsin Active Site Estimation Using Neural Networks, *Proceedings of 13th Israeli Symposium on AI and CV*, 57-60, 1997.

3. Y. Kuhara,K. Shimizu and J. Doi, Neural Networks Trained by Hydrolase And Location of Chymotrypsin Catalytic Triad, *Proceedings of the 1997 Miami Nature Biotechnology Winter Symposium*, 64, 1997.