J. Chem. Software, Vol. 8, No. 1, p. 17–26 (2002)

分子の構造活性相関解析のためのニューラルネットワークシミュレータ:Neco (NEural network simulator for structure-activityCOrrelation of molecules) の開発（７）
― ペリラルチン類の疎水性パラメータ logP の予測 ―

高橋 梨紗 [a], 細矢 治夫 [b], 福田 朋子 [c], 長嶋 雲兵 [c]*

[a] お茶の水女子大学人間文化研究科, 〒 112-8610 文京区大塚 2-1-1
[b] お茶の水女子大学理学部情報科学科, 〒 112-8610 文京区大塚 2-1-1
[c] 産業技術総合研究所先端情報計算センター, 〒 305-8561 つくば市東 1-1-1
**e-mail: u.nagashima@aist.go.jp*

(Received: January 10, 2001; Accepted for publication: October 10, 2001; Published on Web: December 7, 2001)

学習方法として中間層において自己組織化を行い、全体に教師付学習の両方を行う、自己組織化とパーセプトロンを融合した二ューラルネットワークシミュレータの開発をプラットフォーム非依存性を持つ Java 言語を用いて行った。本シミュレータは、自己組織化と教師付学習の双方の特徴を併せもつため、従来のニューラルネットワークよりも高精度な認識処理を実現し、かつ高速学習が可能となる。

このシミュレータを用いて、分子構造によって甘味や苦味の性質を示す 22 種類のペリラルチン類の疎水性パラメータ logP の予測を行った。入力パラメータは分子構造 STERIMOL パラメータ 5 種のパラメータと甘味／苦味の分類値を用いた。出力層のノード数を１つにすることで、 logP の値を連続した数値データとして予測できるようにした。絶対誤差が平均して 0.02 までの学習を 500 回程度で行うことができ、また未学習データに対しては平均して 0.3 程度の絶対誤差、最大でも 0.8 程度の絶対誤差で予測が可能であった。単純パーセプトロンの予測精度は、平均して 0.6 程度の絶対誤差であり、また最大の絶対誤差は 1.3 程度と大きく、本手法がより精度の高い予測を行っていることがわかった。

本手法は学習回数が単純パーセプトロンに比較して 1/5 - 1/10 程度少なく、高速学習が可能であった。

キーワード : Self-organized network, Neural network, Structure-Activity Correlation, Perillartine derivatives, Hydrophobic parameter

1　はじめに

ニューラルネットワーク法は、脳における神経細胞の信号伝達系をモデルにした情報処理法である。この方法では、ニューロンと呼ばれる多くのノードを網目のように結合させ、その結合の強さの形で情報の処理手順や量的な関係を習得させる。動作の特徴として、入力と出力の間の非線形の関係付けを行うことが知られている。このネットワークを用いることで、今までの決定論的なアルゴリズムでは不可能であった未知のデータに対する予測を行うことができる [1]。

このネットワークの結合方式は種々提案されており、それぞれに特色があり、優れている点、あるいは弱点が指摘されている。そこで、異種のネットワーク構造を組み合わせることにより、個々の長所を併せ持ったニューラルネットワークの構築が提案されている。

宮永らは、高速学習が可能な自己組織化モデルと高精度のパーセプトロンモデルを組み合わせることによ

り、高速で高精度なネットワークを実現した [2–4]。このモデルを本研究で開発中の Neco[5–11] に組み込み、性能を評価したところ、特に 2 値の分類問題に対し適用した際に、高精度高速学習が可能であることが既に報告されている [8]。

本稿では、本シミュレータをさらに改良し、1 値の連続する数値を予測する Fitting 問題に適用した結果について報告する。適用例として、分子構造によって甘味や苦味の性質を示す 22 種類のペリラルチン類の疎水性パラメータ logP の予測を行った [12]。疎水性パラメータ log P は、様々な分子の構造活性相関解析に重要な役割を果たすにもかかわらず、その実験的測定は非常に難しく、また理論的な導出は不可能であることが知られている。

また本シミュレータは、様々な計算機での使用を目的とし、プラットフォーム非依存性を持つ Java 言語を用いて実装した。

2　教師付学習を取り入れた自己組織化ニューラルネットワーク法概要

本研究で用いたニューラルネットワークは、入力層、1 層の中間層、出力層の 3 層からなる階層型ニューラルネットワークである。ここでは 1 値の連続する数値の予測を行うため、出力層のノード数は 1 つとした。概念図を Figure 1 に示す。

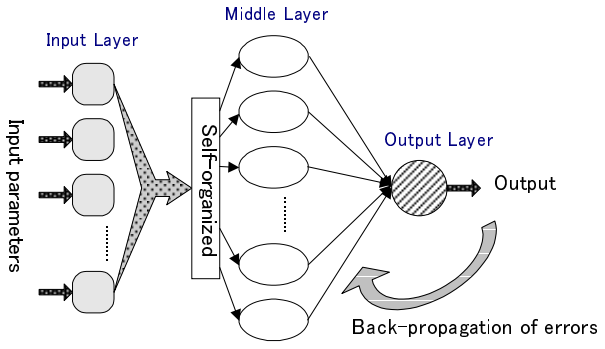

Figure 1. Model of this neural network

学習方法に、中間層でマハラノビスの距離を用いた自己組織化を行い、中間層と出力層間の重みの決定にデルタ規則による教師付学習を行っている [2–4, 8]。

本研究では、本手法の予測精度を向上させるために、更にこの学習方法の改良を行った。その改良点を次に説明する。

2.1　教師付学習前の自己組織化

改良以前の学習方法は、入力層から入力パラメータが入力され、その入力データの特徴にもとづいて、中間層を構築する。この時、中間層は重みつきメンバ数、平均ベクトル、分散行列の 3 つの内部情報によって、その特徴を表現する。その後、中間層の出力値と中間層－出力層間の重みにより、出力値を算出し、デルタ規則による学習により、中間層－出力層間の重みと、中間層が保持する内部情報を学習させる。

しかし、この方法は自己組織化により決定された中間層の内部情報を、教師付学習によって更新させるために、正確な自己組織化が行われていないという問題があることがわかった。

そこで、Figure 2 に示すように、より正確な自己組織化を行うために、教師付学習を行う前に自己組織化を完全に行い、ネットワーク構造を最初に決定した後に中間層の内部情報を学習によって更新するように改良した。これは「教える前に、考えさせる」ということに例えることができる。

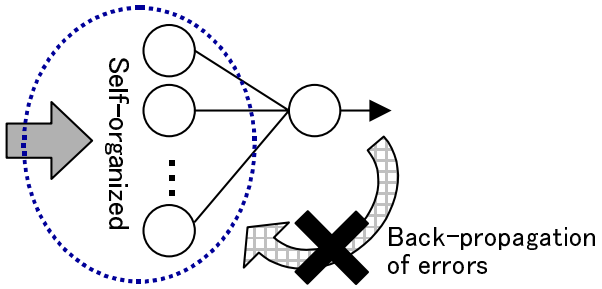

Figure 2. Self-organized before learning

2.2　中間層の内部表現

2.1 で説明したように、本手法の中間層のノードは、重みつきメンバ数、平均ベクトル、分散行列の 3 つの内部情報を保持する。これらの情報が、それを保持する中間層ノードの特徴を表現し、これをもとに計算した入力データとの距離が中間層の出力値となる。

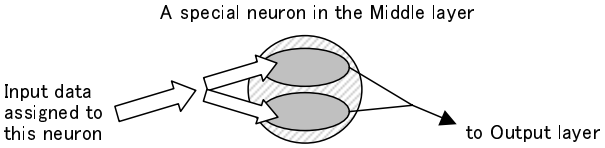

Figure 3. Flow of numerical data through a special neuron in the Middle Layer

そこで、中間層の特徴をよりよく表現することができるネットワークを作成するために、1 つの中間層ノードに 2 組の内部情報をもたせることとした。言い換えると、比較的特徴の一致している 2 つの中間層ノードをそれぞれの内部表現はそのままで、ネットワークの重みを全く同じとした。概念図を Figure 3 に示す。

これは、実質的に中間層ノード数を増加させ、かつ学習によって決められる変数の数を減らすことになる。結果的に中間層ノード数の増加がニューラルネットワークの自由度を増加させ、変数の減少が、学習速度の向上と精度の向上をもたらすこととなっている。

3　性能評価 －ペリラルチン類の疎水性パラメータ logP の予測

本手法の適用例として、実験による測定が困難な問題への応用を考え、ここではペリラルチン類の疎水性パラメータ logP（（1 －オクタノール／水の分配係数の対数値 ）の予測を行った結果について報告する。

3.1　入力データとパラメータ

ペリラルチンはシソ糖とも呼ばれる化合物であり、植物のシソに含まれるペリラルアルデヒドをオキシム化することによって得られる甘味物質である。構造式を Figure 4 に示す。

ショ糖は代表的な甘味量として使われているが、消費量の増大に伴い、肥満、心臓病、高血圧、糖尿病等の成人病が問題となってきている。この問題を克服するために、低カロリーの人工甘味料の開発が近年盛んに行われており、新しい人工甘味料をデザインするためには、甘味化合物の構造－味質相関についての知見が大変有用な情報となっている。

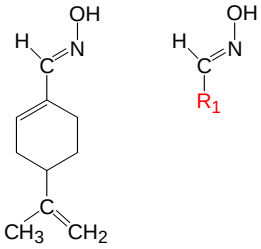

Figure 4. Structures of perillartine (left) and its derivatives (right)

そこで、非線形動作の予測が可能であるニューラルネットワークである本手法を用いて、ペリラルチン類の疎水性の因子を説明するための疎水性パラメータ logP の予測を行った。

入力データとして、ペリラルチン誘導体の甘味データと苦味データをそれぞれ 11 種類ずつ、計 22 種類を用いた。

入力パラメータは、分子の形状を表わす STERIMOL パラメータ 5 種（Figure 5　L, W_u, W_d, W_l, W_r）と、甘味／苦味の分類値の計 6 種類を用いた [12]。

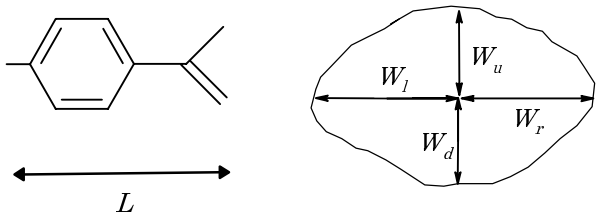

Figure 5. Details of STERIMOL parameters

本研究で用いた 22 種類のペルラルチン誘導体の構造を Figure 6 に示した。また本計算に用いた入力データ（入力パラメータと教師データ）を Table 1 に示した。

3.2　計算結果

学習は、300 回の自己組織化の後に、教師付学習を行った。自己組織化の段階で、本手法の中間層ニューロン数は 6 となった。これは入力データを 6 種類に分類したことを意味する。教師付き学習は、累積 2 乗誤差をこのネットワークの評価関数として用い、0.001 まで収束させたところ、500 ～ 1000 回程度で学習を終了することができた。

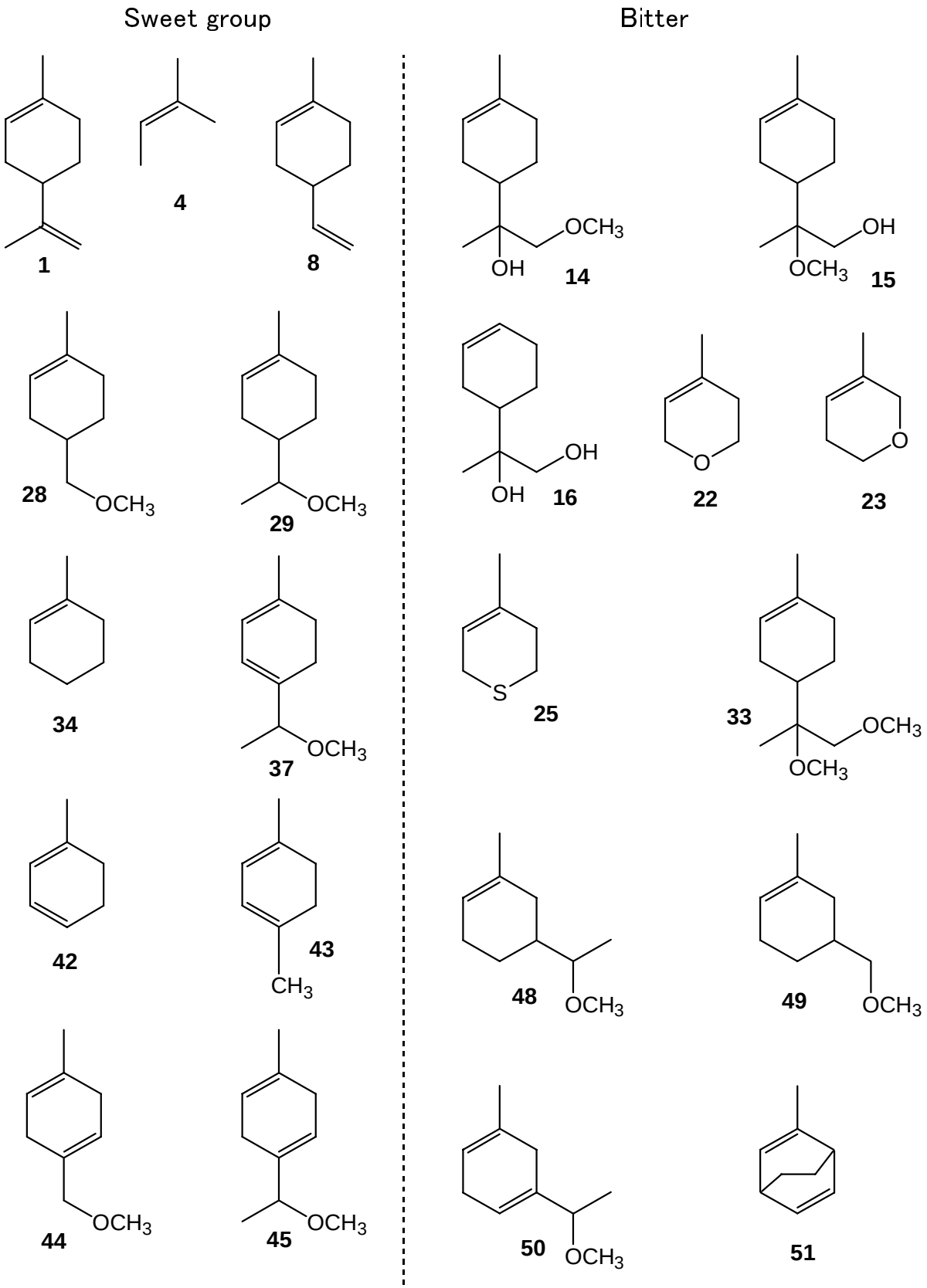

Figure 6. Sweet/bitter structures of perillartine derivatives and related compounds

Table 1. Input and supervised parameters for perillartine derivatives

No.[*1]	Input parameters						Supervised parameter
	Sweet/Bitter[*2]	L	W_l	W_u	W_r	W_d	logP
1	1	8.52	3.13	2.85	3.42	1.99	2.58
4	1	5.10	3.13	1.91	2.94	1.9	0.87
8	1	8.69	3.19	2.84	3.42	1.99	2.28
28	1	9.36	3.14	2.94	3.41	1.98	1.10
29	1	9.36	3.14	3.26	3.56	2.10	1.40
34	1	6.06	3.09	2.08	3.01	1.71	1.48
37	1	8.87	3.3	2.63	3.07	2.52	1.10
42	1	6.29	3.09	2.63	3.07	2.52	1.48
43	1	7.10	3.09	1.91	3.41	1.91	0.78
44	1	9.01	3.09	2.2	3.41	2.02	0.80
45	1	9.01	3.08	2.52	3.43	2.53	1.10
14	0	10.67	3.33	4.11	3.56	2.10	-0.10
15	0	9.36	3.04	3.76	3.62	2.22	-0.10
16	0	9.37	3.14	3.56	3.56	2.10	-0.92
22	0	5.51	3.05	2.53	3.41	1.97	-0.72
23	0	6.15	3.16	2.67	3.01	1.72	-0.72
25	0	6.05	3.25	2.62	3.43	2.03	0.34
33	0	10.67	3.51	4.08	3.63	2.22	0.72
48	0	7.98	3.12	3.42	5.96	2.00	1.40
49	0	7.68	3.09	2.32	5.84	1.96	0.80
50	0	7.68	3.09	2.43	5.89	2.57	1.10
51	0	5.88	2.72	2.95	3.92	3.85	1.90

*1 Structure and STERIMOL parameters of perillartine derivatives given in Figures 5, 6.
*2 1 and 0 represent sweet and bitter respectively.

改良前の方法を用いると、本論文の例の場合中間層の数が８となり、学習回数は 700-2000 となった。学習に要する計算時間は、本論文に示す改良により約 1/2 となっている。

広く用いられている単純パーセプトロンとの比較では、パーセプトロンと自己組織化ニューラルネットワークではネットワークの構成原理が異なるため、速度の比較には学習精度がほぼ同等な値となるネットワーク構造での性能比較を行うことにした。すなわち、同様の学習を入力層ニューロン数 7、中間層ニューロン数 12、出力層ニューロン数１の単純パーセプトロンで行ったところ、同程度の精度まで収束させるのに 6000 回程度の学習回数を必要とした。なお、入力層ニューロン数が入力パラメータ数より１つ多いのは、常に１の信号を出力するバイアスニューロンを加えたためである。

Figure 7 に本手法と単純パーセプトロンモデルの収束曲線を示した。本手法の中間層での処理法及び自己組織化の回数を考慮しても、単純パーセプトロンでは入力層と中間層に重みの教師付学習を行っているため、本手法は単純パーセプトロンと比較してかなり高速な学習を行っていると言える。

この時の学習精度は、文献値と予測値の近似式 $y = 0.993x + 0.0003$、相関係数 0.999、誤差の標準偏差 0.005、最大誤差 0.08、平均誤差 0.02 であった。Figure 8 に示すように高精度な学習を行っていることがわかる。

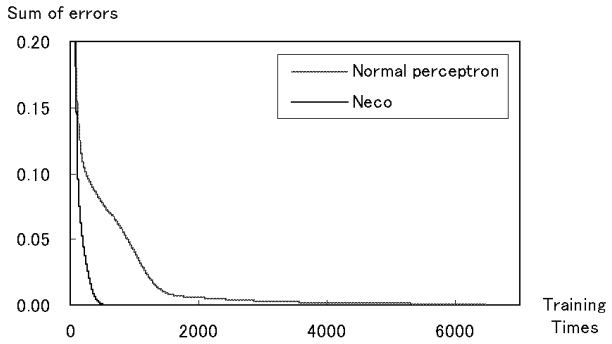

Figure 7. Error curves for training (normal perceptron and Neco)

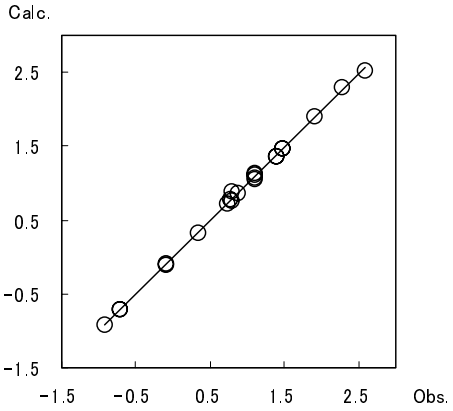

Figure 8. Relationship between observed and calculated logP values

次に中間層ノードの解析を行った結果について示す。Figure 9 に示すように、自己組織化により、6 つのニューロンからなる中間層は、ペリラルチン誘導体の置換基部分の構造が似通ったものが、きれいに分類され、かつ、甘味データ、苦味データも混ざることなく分類されていることがわかった。本手法は、この入力データの特徴を強く認識する分類によって、精度のよい学習と予測を行うことができるのである。

次に未学習データに対する予測精度について記す。予測精度は、上記の学習で使用した 22 種類のデータから 1 つを除いた 21 種類のデータを用いてネットワークの学習を行い、ここで作成したネットワークを用いて、除いたデータの予測を行うということを、全てのデータに対して行うことで確認した。

全データを用いた解析では、4 番と 25 番と 51 番の予測精度が両者のネットワークで悪くなり、単純パーセプトロンと本手法に大きな差は無かった。これらの３つは、他の分子に比べ類似度が低い（つまり独立性が高い）もので、これらをのぞいた学習セットで学習させた単純パーセプトロンでも、これらの予測精度はきわめて悪い。

これらのデータは、51 番のように Figure 9 の中間層ノードの分類において、Neuron No.6 のように１つのデータでノードを構成するものであり、また４番と２５番のようにそれぞれ No.2 と No.5 に分類されたとはいえ、それぞれの分類の中で異質なものである。これは、分類の結果をもとに予測を行っている本法の特徴を示していると言える。また、当然のことであるが、予測データがネットワークの中間層のどのノードとも低い類似度、つまりどこにも分類されないと判断された場合のデータの予測精度は低い。

学習の際に類似度が高いものがない場合、その予測精度はきわめて悪くなるが、これは教師付き学習によるニューラルネットワークすべてについていえることである。全データを用いた解析では、予測精度に大きな差が出なかったため、これら３つのデータの予測を除いた 19 のデータを用いた結果を Figure 10 に示す。これら３つを含む４つ以上の分子を除いた場合、本手法と単純パーセプトロンとの予測精度には３つを除いた場合とほぼ同等な差が見られた。これ以外のデータに関しては、本法はペリラルチン類の疎水性パラメータ logP の未学習データ予測に関して± 0.5 の誤差で予測可能であった。

Figure 10 左に示すように、予測精度は、文献値と予測値の近似式 $y = 0.953x + 0.058$、相関係数 0.935、誤差の標準偏差 0.224、最大誤差 0.82、平均誤差 0.27 であった。この結果より本手法は、未学習データに対して高精度な予測が可能であることがわかった。

Figure 10 右に示すように前述と同じ構造をもつ単純パーセプトロンにおいて、同様の実験を行った結果は、文献値と予測値の近似式 $y = 0.8298x + 0.1512$、相関係数 0.7460、誤差の標準偏差 0.3640、最大誤差 1.34、平均誤差 0.60 であった。本手法は単純パーセプトロンと比較し、未学習データに対して精度のよい予測結果を示すことがわかった。

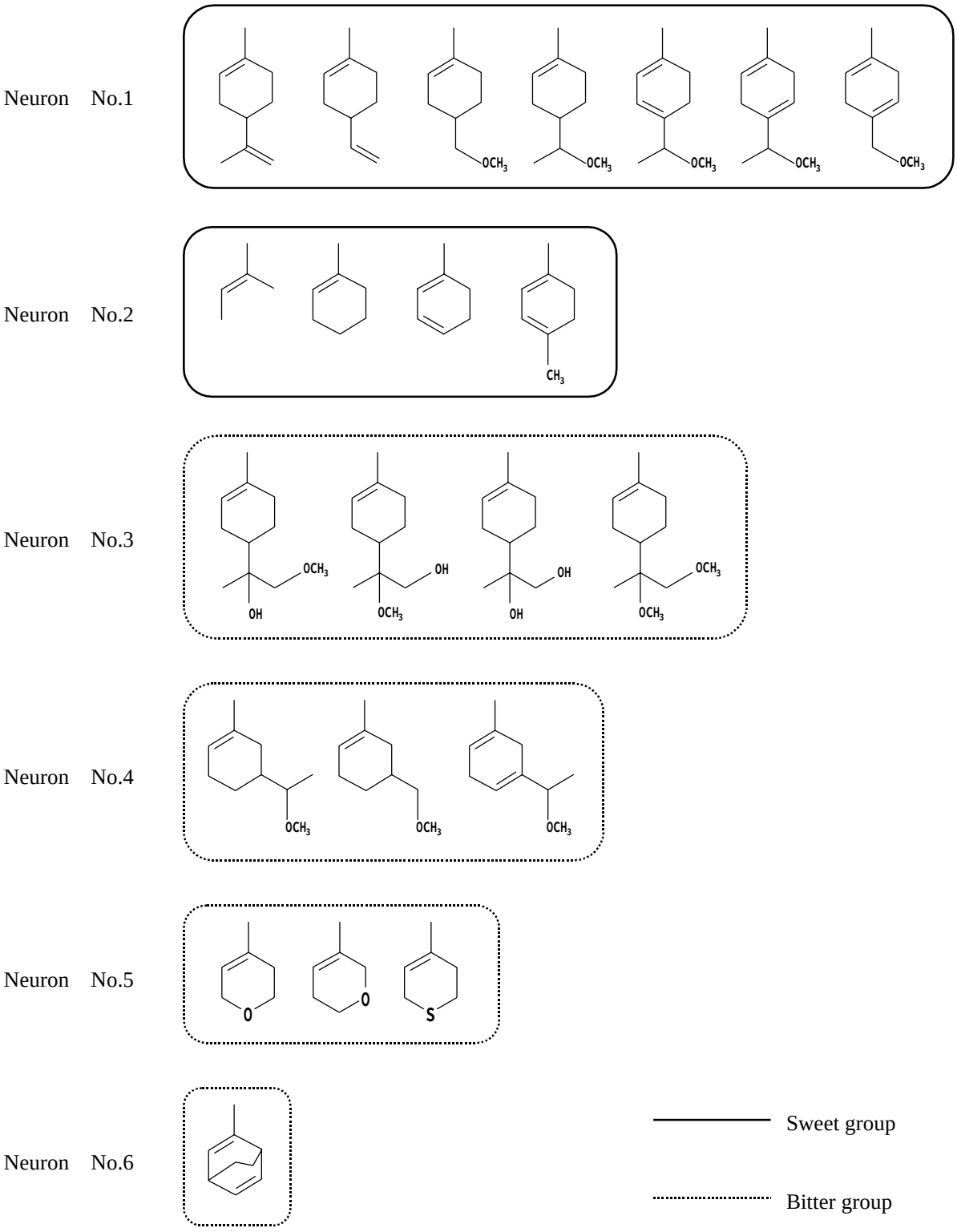

Figure 9. Analysis of neurons in Middle layer

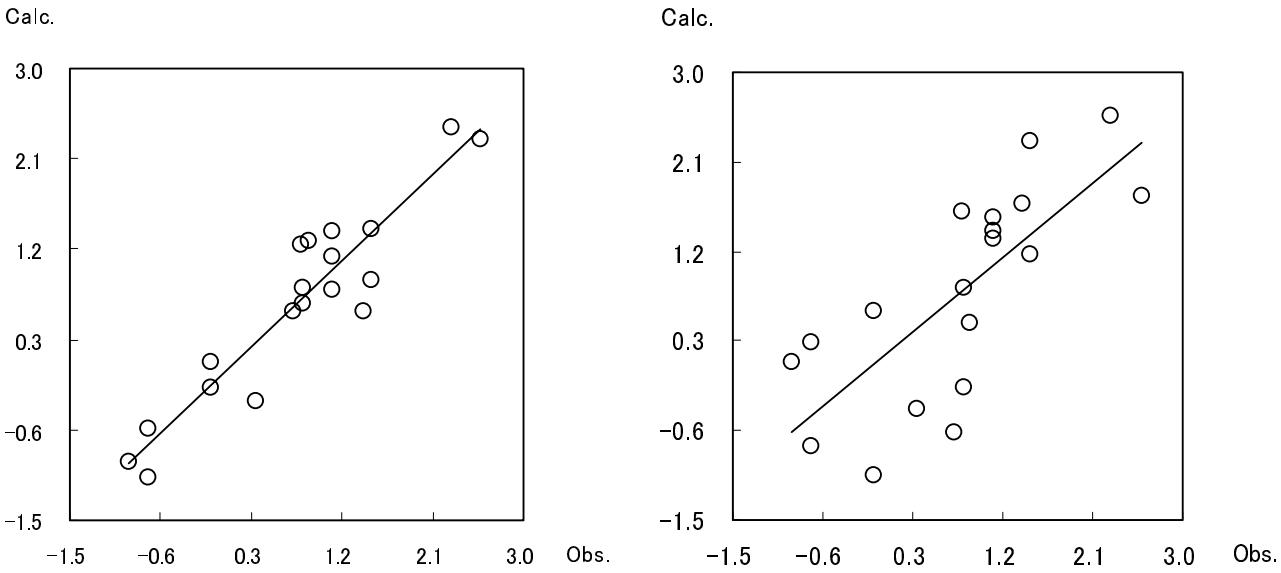

Figure 10. Relationship between observed and calculated logP values（Neco at right, Normal perceptron at left）

4　まとめ

学習方法として中間層において自己組織化を行い、全体に教師付学習の両方を行う、自己組織化とパーセプトロンを融合したニューラルネットワークモデルの改良を行った。教師付学習以前に自己組織化を行い、ネットワークを決定することで、より入力データの特徴を掴むネットワークの作成に成功した。マハラノビスの距離による自己組織化により、入力データの正確な分類を行っているため、分類に応じた最適な中間層のノード数で学習できることがわかった。更に、ネットワークの出力値は中間層ノードの分類の状態を強く反映するため、外挿予測に対しても従来の単純パーセプトロンより高い性能を持つことがわかった。また、中間層ニューロン内部に２つの内部情報を持たせたことによって、ネットワークの表現力が向上し、未学習データに対する予測精度の向上に貢献した。また中間層ノード内に１つの内部情報しかもたない時と比較して、少ない数の中間層での学習が可能であり、更に効率のよいネットワークを構築することができた。

本手法を、1 値の連続する数値の予測問題に適用したところ、この問題に対しても適用することができた。適用例として、ペリラルチン類の疎水性パラメータ logP の予測を行ったところ、高速に高精度な学習を行うことが示された。また未学習データの予測に対しても高精度な予測を行うことができた。更に単純パーセプトロンと比較したところ、本手法の方がパーセプトロンモデルよりも高速に学習し、高精度な予測を示すことがわかった。これら結果より、実験的に測定が困難な非線形的な問題に対し、本手法を有効に活用することができると考えられる。

また Java 言語を用いて本シミュレータを作成した。そのため、プラットフォームに依存しないシミュレータとなった。

本研究を行うにあたり、お世話になりました宮崎大学教授 青山智夫博士、旭硝子 山本博志氏に深く感謝を致します。

参考文献

[1] 市川紘, 階層型ニューラルネットワーク　非線型問題解析への応用, 共立出版 (1993).

[2] 宮永喜一，奥村伸二，栃内香次, 自己組織化クラスタリングの汎化性と適応能力について, 電子情報通信学会論文誌, **J75-A**, 1207-1215 (1992).

[3] 宮永喜一，奥村伸二，栃内香次, 自己組織化と教師によるネットワークの高速・高精度学習について, 電子情報通信学会論文誌, **J78-A**, 1475-1484 (1995).

[4] Y. Miyanaga, R. Islam, K. Tochinai, Nonlinear spectrum estimation using a modified self-

organized network, *IPSJ SIG Notes*, **95-HPC-55-9**, 65-72 (1995).

[5] 井須 芳美, 長嶋 雲兵, 細矢 治夫, 青山 智夫, 分子の構造活性相関解析のためのニューラルネットワークシミュレータ：Neco の開発, *J. Chem. Software*, **2**, 76-95 (1994).

[6] 井須芳美, 長嶋 雲兵, 細矢 治夫, 大島 茂, 坂本 曜子, 青山 智夫, 分子の構造活性相関解析のためのニューラルネットワークシミュレータ：Neco の開発 (2) －多環式芳香族炭化水素（PAH）の ^{13}C-NMR ケミカルシフトとその発癌性－, *J. Chem. Software*, **3**, 1-10 (1996).

[7] Isu, Y., Nagashima, U., Aoyama, T., Hosoya, H., Development of Neural Network Simulator for Structure-Activity Correlation of Molecules (Neco), *J. Chem. Info. Comp. Sci.*, **36**, 286-293 (1996).

[8] 藤谷康子, 小野寺光永, 井須芳美, 長嶋雲兵, 細矢治夫, 青山智夫, 分子の構造活性相関解析のためのニューラルネットワークシミュレータ：Neco の開発 (3)　組み合わせモデルとパーセプトロンの性能比較, *J. Chem. Software*, **4**, 19-32 (1998).

[9] 田島 澄恵, 松本 高利, 長嶋 雲兵, 細矢 治夫, 青山 智夫, 分子の構造活性相関解析のためのニューラルネットワークシミュレータ: Neco(NEural network simulator for structure-activity COrrelation of molecules) の開発 (４) ―ペリラルチン類の甘味・苦味分類― , *J. Chem. Software*, **6**, 115-126 (2000).

[10] 福田 朋子, 田島 澄恵, 斎藤 久登, 長嶋 雲兵, 細矢 治夫, 青山 智夫, パーセプトロン型ニューラルネットワークと多次元 Ck 級補間法を用いた樹脂被覆肥料の溶出誘導時間および 80%溶出時間の推定― 分子の構造活性相関解析のためのニューラルネットワークシミュレータ: Neco(NEural network simulator for structure-activity COrrelation of molecules) の開発 (5) ―, *J. Chem. Software*, **7**, 115-128 (2001).

[11] 福田 朋子, 田島 澄恵, 松本 高利, 長嶋 雲兵, 細矢 治夫, 青山 智夫, 分子の構造活性相関解析のためのニューラルネットワークシミュレータ: Neco(NEural network simulator for structure-activity COrrelation of molecules) の開発 (6) ―機械構造用 Cr-Mo 鋼、Ni 鋼、Ni-Cr 鋼および Ni-Cr-Mo 鋼の力学的性質の推定― , *J. Chem. Software*, **7**, 179-190 (2001).

[12] 宮下芳勝, 佐々木愼一, ケモメトリックス　化学パターン認識と多変量解析, 共立出版 (1995).

Development of a Neural Network Simulator for Structure-activity Correlation of Molecules: Neco (7) - Hydrophobic Parameter (logP) Prediction of Perillartine Derivatives -

Risa TAKAHASHI[a], Haruo HOSOYA[b], Tomoko FUKUDA[c] and Umpei NAGASHIMA[c]*

[a]Department of Human Culture and Sciences, Graduate School of Ochanomizu University
2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan
[b]Faculty of Sciences, Ochanomizu University
2-1-1 Otsuka, Bunkyo-ku, Tokyo 112-8610, Japan
[c]National Institute for Advanced Industrial Science and Technology
1-1-1 Higashi, Tsukuba, Ibaraki 305-8562, Japan
**e-mail: u.nagashima@aist.go.jp*

We developed a neural network simulator for structure-activity correlation of molecules: Neco. A self-organized network model for high-speed learning was included in Neco, a perceptron type with three layers. In the hidden layer the neurons are self-organized by using Mahalanobis generalized distance.

This report proposes an improved training algorithm to the network. A self-organizing module decides the number of neurons in the hidden layer, at first. Then, a neuron in the hidden layer has two informations which describe a characteristic of the neuron. In this way, the network can evaluate stochastic characteristics from input data better.

Using this simulator, the hydrophobic parameter, logP, of perillartine derivatives was predicted. We used for inputs a set of six parameters: five STERIMOL (L, W_l, W_u, W_r, and W_d) and the sweet/bitter activity. The 22 sampled data are used for training. Our neural network can accurately predict hydrophobic parameter, logP. Compared with a normal perceptron network, the learning ability of our network is somewhat higher and its convergence speed is greatly much larger.

This simulator doesn't depend on the machine environment because it codes by the Java programming language.

Keywords: Self-organized network, Neural network, Structure-Activity Correlation, Perillartine derivatives, Hydrophobic parameter