

ニューラルネットワークによる 多種類の有機化合物の発ガン性の予測

田辺 和俊^{a*}, 大森 紀人^a, 小野 修一郎^a, 鈴木 孝弘^b, 松本 高利^c, 長嶋 雲兵^d, 上坂 博亨^e

^a 千葉工業大学社会システム科学部経営情報科学科, 〒 275-0016 習志野市津田沼 2-17-1

^b 東洋大学経済学部経済学科, 〒 112-8606 東京都文京区白山 5-28-20

^c 東北大学多元物質科学研究所, 〒 980-8577 仙台市青葉区片平 2-1-1

^d 産業技術総合研究所グリッド 研究センター, 〒 305-8568 つくば市梅園 1-1-1

^e 富山国際大学地域学部, 〒 930-1292 富山県上新川郡大山町東黒牧 65-1

*e-mail: tanabe@pf.it-chiba.ac.jp

(Received: August 27, 2004; Accepted for publication: June 20, 2005; Published on Web: August 31, 2005)

広範囲の化学物質について構造活性相関により化学構造から有害性を高い精度で予測するモデルを開発することを目指して、ニューラルネットワークを用いて発ガン性のデータを解析した。Predictive Toxicology Challenge 2000-2001 で公開された 454 種類の有機化合物についての記述子と発ガン性のデータを用いてニューラルネットワークの学習とテストを行った。分子軌道計算から求まる homo や lumo などの軌道エネルギー、分子表面積、log P など 37 種類の記述子について主成分分析を行い、得られた 10 個の主成分データをニューラルネットワークに入力した。Error-back-propagation 法によりニューラルネットワークを学習する際の過学習、over-fitting、局所解などの問題を解決するために、学習回数や中間層ユニット数などの条件を最適化した。モデルの作成に用いなかった化合物を用いて判別テストを行った結果、的中率 73.7 % の予測モデルを開発することができた。

キーワード：構造活性相関, ニューラルネットワーク, 発ガン性予測, 主成分分析, 過学習

1 緒言

ヒト、生物、環境に対する化学物質の毒性、蓄積性、濃縮性、分解性などの有害性を評価するために、定量的構造活性相関 (Quantitative Structure-Activity Relationship, QSAR) のモデルを用いた研究が行われ、化学構造から有害性を予測するモデルが開発されている [1–4]。しかし、それらの予測モデルの性能は未だ不十分である。例えば、米国の NIEHS による発ガン性の公開テストの結果では、CASE、TOPKAT、REPAD、COMPACT、FALS などのモデルはいずれも 60 % 以下の低い予測率であった [5, 6]。また、最近行われた Predictive Toxicology Challenge (PTC) 2000-2001 のコンテストでも応募モデルの予測性能はほとんど 60 %

以下であった [7–12]。

一般に QSAR モデルを用いる有害性予測には、予測精度、汎用性、処理時間という相反する 3 種類の問題があるが、これらをすべて解決するのは難問である。これらの問題に関係する選択肢は、構造 有害性の相関の解析にどのようなモデルを用いるかという点と、相関解析に有効な構造記述子をいかに選定するかという点である。

既存モデルの予測性の低さは、記述子として化合物分子内の原子、結合、部分構造の有無やそれらの個数など単純な情報しか用いていないことと、構造 有害性の相関解析に重回帰分析などの線形解析モデルを用いているためと考えられる。したがって、構造情報をより精密に反映する量子化学的記述子などを用い、非

線形解析モデルと組み合わせることにより、既存のモデルよりも精度が高い予測モデルを開発できる可能性がある。

そのような非線形解析モデルの一つとしてニューラルネットワークがあり、薬理活性の解析や予測などの分野で多く適用されてきている [13-15] が、有害性予測に応用した研究は少ない。そこで、我々は QSAR により化合物の構造情報のみから有害性を高い精度で予測するモデルを開発することを目指して、ニューラルネットワークを用いて構造と有害性の相関を解析する研究を行っている。

前報 [16] においては、有機塩素化合物の発ガン性データを解析し、ニューラルネットワークが高精度の予測に有効であることを見出した。すなわち、41 種類の有機塩素化合物について、分子軌道計算などから求まる 7 種類の記述子を用いてニューラルネットワークを学習し、leave-one-out test を行った結果、的中率 93 % で予測できることを見出した。しかし、そこで扱ったのはわずか 41 種類の有機塩素化合物であり、多種多様な化合物についてこのような高い的中率で予測できるかどうかは疑問が残る。

そこで本研究では、ニューラルネットワークを用いる我々のモデルが広範囲の化合物についてどの程度まで拡張できるか、さらに他のモデルに比べてどの程度高い的中率で予測できるかを検討するために、PTC のデータをニューラルネットワークで解析し、コンテストの参加者の結果と比較することにした。

ただし、コンテストの参加者は正解、すなわちテスト用化合物の発ガン性データを知らされない条件での中率を算出したのに対し、本研究では正解を知っている条件での中率を算出する。したがって、本研究の立場は、コンテストの参加者の結果の上に立って、さらによりよい結果を与えるようなモデルを得ることを目的とした。

2 方法

2.1 ニューラルネットワークの構造

ニューラルネットワークを用いてデータを解析する際には、ネットワークの構造および学習において注意すべき点が幾つかある。ネットワーク構造に関しては、入力層、中間層、出力層の各ユニット数と、中間層の層数がある。これらのユニット数や層数が多すぎるとパラメータが過剰になり、over-fitting [17] という状態

に陥り、予測的中率が低下する。したがって、最適なネットワーク構造を決定することが必要である。

本研究では、ニューラルネットワークは前報と同様に、入力層、中間層、出力層各 1 層から成る 3 層構造のニューラルネットワークを使用した。入力層に化合物の構造を特徴づける記述子を入力し、出力層には各化合物の発ガン性の有無を教師データとして入力した。したがって、入力層のユニット数は記述子の数とし、出力層のユニット数は 1 とした。中間層のユニット数は over-fitting の問題と密接に関連するので、下記のようにユニット数を種々変化させて誤差を調べて最適値を探した。

2.2 発ガン性データ

本研究でニューラルネットワークの学習とテストに用いた記述子および発ガン性のデータは、PTC 2000-2001 のコンテスト [7] において公開されたデータである。このコンテストでは、約 500 種類の化合物のデータを用いて発ガン性を予測する web 上及びワークショップでのコンテストが行われ、多数の参加者によるモデルの予測結果が公表された。したがって、このコンテストの発ガン性データは、広範囲の化合物に対する我々のモデルの適用性を検証し、他のモデルの性能と比較する点できわめて有効なデータである。

この PTC コンテストは、予測モデル構築用とテスト用の化合物群について記述子を公募する Data Engineering 段階、モデル構築用化合物群の記述子のデータを用いて参加者が予測モデルを構築する Model Construction 段階、及びその予測モデルを用いて発ガン性の試験結果を伏せたテスト化合物群に対してその発ガン性を予測する Model Evaluation 段階の 3 段階で行われた。

発ガン性データとしては、モデル構築用化合物には NTP による試験結果、テスト用化合物には FDA の試験結果のデータが使用され、それぞれ male と female の rat と mouse、計 4 種類についてデータが公開された。本研究ではその中から参加者の予測的中率が最も低かった male rat のデータについて検討した。その male rat についてモデル構築用、テスト用の発ガン性、非発ガン性化合物数の内訳を Table 1 に示す。

Table 1. Numbers of compounds

Data Set	Carcinogen	Non-carcinogen	Total
Model construction	107	180	287
Model evaluation	45	122	167
Total	152	302	454

2.3 記述子データ

コンテストの Data Engineering 段階では、これらの化合物群に対する構造情報を主催者が公開し、それらにもとづいて計算される記述子を一般から公募した。その結果、7種類の記述子群、総数 7000 個以上の記述子が利用可能となった。7種類の記述子群の内訳は、Leuven (官能基やその間の距離等)、Dragon (原子、結合、部分構造の個数等)、tReymers (量子化学計算による記述子等)、BCI (BCI 社の記述子群)、VINITI (ヘテロ原子や電子等の記述子群)、Helma (量子化学計算による記述子群)、Kramer (線形記述子) である。

しかし、これら 7000 個以上の記述子の値を本研究の化合物群に対して適用して詳細に調べると、分子内に含まれる原子、結合、部分構造の有無やそれらの個数を示す記述子等の中には、ほとんどの化合物についてそれらの出現頻度が 0 となるものが多数あった。また、出現頻度が 0 でなくてもきわめて低く、発ガン性の有無との相関を見るためには明らかに有効でないと思われる記述子も多々存在していた。

そこで本研究では、発ガン性予測に有効と思われる記述子として、tReymers と Helma の記述子群から 37 個の記述子を選定した。それらの記述子は分子軌道法などで計算される量子化学的記述子であり、すべての化合物に対して 0 でない値をもっている。それらの記述子の内訳を Table 2 に示す。

QSAR による予測モデルを構築する際、最も重要な因子は入力記述子の選択である。Table 2 に示すように、今回用いた PTC の記述子の中には homo、lumo、log P のように tReymers と Helma で重複していたり、相関が高いと思われる記述子が幾つかある。しかし、それらの記述子は Table 3 の相関係数に示すように、tReymers と Helma でその値が異なっている。

Table 2. Descriptors used in this study

Descriptor Set	Descriptors	R ^a
tReymers	SLOGP	-0.01
	SMR	-0.03
	ROTBOND	-0.06
	FLEX	0.02
	VOLUME	-0.09
	SURF_A	-0.09
	TPSA	-0.08
	HBD	-0.06
	HBA	-0.04
	LUMO	0.03
	HOMO	-0.01
	DIPOLE	-0.15
	LOGD_2	-0.05
LOGD_7	-0.02	
LOGD_10	0.00	
Helma	logP	0.02
	dipole	-0.08
	electronic energy	0.06
	electronegativity	-0.04
	heat of formation	0.17
	homo	-0.06
	homo-lumo	-0.04
	hybrid dipole	-0.01
	ionization potential	0.06
	lumo	0.00
	largest interatomic distance	-0.07
	molecular weight	0.02
	point-chg. dipole	-0.09
	total energy	0.05
	/POLARIZA	-0.04
	/DELTAHF	0.19
	/STABIL	0.03
/STRAIN	-0.11	
total_access	-0.07	
non_polar_access	-0.04	
polar_access	-0.04	
perc_nonpolar	0.04	

^aR : Correlation coefficient with carcinogenicity

Table 3. Correlation coefficients between similar descriptors

Descriptor A	Descriptor B	R ^a
tReymers SLOGP	Helma logP	0.94
tReymers SMR	Helma /POLARIZA	1.00
tReymers SURF_A	Helma total_access	0.94
tReymers LUMO	Helma lumo	0.14
tReymers HOMO	Helma homo	0.76
tReymers DIPOLE	Helma dipole	0.76
Helma heat of formation	Helma /DELTAHF	0.90
Helma homo	Helma ionization potential	-1.00

^aR: Correlation coefficient between descriptors A and B

これはそれらの記述子を計算した時の化合物の立体配座(構造最適化の計算方法の違いによる)などが異なるためと思われるが、我々としてどちらの値を用いるべきかの選択にあたっての基準が見当たらない。そこで、本研究ではこのような重複している記述子から独立かつ有効な情報量を抽出するために、主成分分析を行い、その結果の累積寄与率から判断した主成分の値をニューラルネットワークの入力層に入力した。

ただし、主成分分析の結果をニューラルネットワークの入力に使うと、記述子の感度分析が行えなくなるなどの問題が生じる。しかし、今回用いた 37 個の記述子の間にはかなり線形の相関が高い記述子の組がいくつかあり、また学習用のデータについて自由度(すなわち、データの数とパラメータ数の比)を高く保つため、入力変数の数を絞り込む必要があった。そこで、本研究では主成分分析を行うことにしたが、その有効性を確認するために、主成分分析を行わなかった結果と比較した。

一方、出力層の教師データには発ガン性化合物に 0.9、非発ガン性化合物に 0.1 の値を入力した。このように、教師データを 0.0 ~ 1.0 でなく、0.1 ~ 0.9 に規格化したのは、ニューラルネットワークのユニット(ニューロン)において数値変換に使われている sigmoid 関数の特性を考慮したためである。また、テスト化合物の予測的中率の判定においては、出力層の値が 0.5 以上なら発ガン性あり、0.5 以下なら発ガン性なしと判定した。

2.4 ニューラルネットワークの学習

本研究では、ニューラルネットワークのソフトウェアは前報と同じく富士通の NEUROSIM/L ver.3 を用い、学習は error-back-propagation 法で行った。主成分分析には EXCEL 多変量解析 Ver.3.0 (エスミ) を用い、学習経過時の誤差および正解率表示のプログラムは C で作成した。また、error-back-propagation のアルゴリズムにおける重みの修正量を計算する式に現れる学習定数の値は NEUROSIM/L のデフォルトである learning rate () = 5.0、momentum term () = 0.4 を用いた。ニューラルネットワークの学習の収束条件としては、すべての学習データについて出力値と学習データ値との差が 0.4 以下、すなわち、発ガン性化合物では出力値が 0.5 以上、非発ガン性化合物では出力値が 0.5 以下になれば学習が収束したと見なした。

ニューラルネットワークの学習に関しては、過学習

(over-training または over-learning) 及び局所解 (local minimum) という問題がある。過学習とは、error-back-propagation 法により学習用化合物群に対してネットワークの学習を過剰に行うと、テスト化合物群に対する予測的中率が低下する状態に陥る現象である。すなわち、学習用化合物についてネットワークの学習を繰り返すと、学習用化合物については一般に誤差はどんどん小さくなる。しかし、このようにして得られたネットワークがテスト化合物に対して最大の予測的中率を与えるかどうかは疑問である。

一般に、データ解析モデルにはモデルの誤差とデータの誤差(雑音)があり、学習データに完全にフィットしたモデルが必ずしも最善とはいえないからである。したがって、ニューラルネットワークの予測性能を最大にするためには、最適の学習回数を決定する必要がある。そこで、本研究では学習回数の最適値を求めるために、以下の方法を検討した。まず、モデル構築用の発ガン性及び非発ガン性化合物それぞれを分子量の小さい順にソートし、学習用と検証用に機械的に振り分けた。そして、学習用化合物群のみを用いてニューラルネットワークを学習しながら、検証用化合物群に対する平均誤差の変化を観測した。

また、局所解 (local minimum) とは、error-back-propagation 法でニューラルネットワークの学習を行うと、ネットワークの重みとしきい値の初期値によって異なる解に収束する現象である。すなわち、error-back-propagation 法では、与えられた重みとしきい値の初期値から steepest descent 法により誤差の少ない状態を探索するので、重みとしきい値の初期値を種々変えつつ学習を行い、誤差が最小になる解を探索する必要がある。本研究で用いた NEUROSIM/L では、重みとしきい値の初期値は乱数で発生させているので、乱数発生 random seed を種々変えて学習を行い、検証用化合物群に対する誤差が最も小さくなる場合を探した。

3 結果と考察

3.1 入力記述子の選定

上記のように、37 個の記述子に対して情報量を集約するために 474 種類の全化合物について主成分分析を行った。その結果、主成分分析を行わず、37 個の記述子を入力に用いた場合の予測率は 59.9% となったが、主成分分析を行った場合の最終的な予測率は、下記の

ように 73.7% となり、主成分分析の有効性が確認された。そこで以下では主成分分析を行った場合の結果を示す。

主成分分析で得られた累積寄与率を Table 4 に示す。この結果から、主成分 10 個で累積寄与率は 93 % 以上に達し、37 個の記述子の情報が 10 個の主成分でほぼ集約できることが分かった。そこで、ニューラルネットワークの入力層は 10 ユニットとし、この 10 個の主成分に変換した記述子を入力した。

3.2 学習回数 の 決定

上記のように、ニューラルネットワークの学習において深刻な過学習の問題を回避し、最適の学習回数を決めるために、学習経過時における学習用、検証用、テスト用の各化合物群の誤差の変化を調べた。その結果の 1 例として、中間層ユニット数 2、重み初期値発生 random seed が 9199 の場合の学習用、検証用、テスト用各化合物群の平均誤差の変化図を Figure 1 に示す。

この図から、error-back-propagation 法による学習を

Table 4. Result of principal component analysis

No.	Eigenvalue	Proportion (%)	Cumulative proportion (%)
1	0.2284	36.0084	36.0084
2	0.1356	21.3767	57.3851
3	0.0819	12.9115	70.2966
4	0.0469	7.3974	77.6941
5	0.0300	4.7265	82.4206
6	0.0227	3.5843	86.0049
7	0.0135	2.1282	88.1331
8	0.0115	1.8156	89.9487
9	0.0103	1.6201	91.5688
10	0.0096	1.5073	93.0761
11	0.0073	1.1503	94.2264
12	0.0060	0.9441	95.1705
13	0.0044	0.6990	95.8696
14	0.0037	0.5794	96.4490
15	0.0032	0.5081	96.9571

繰り返すと、学習用化合物の誤差は限りなく減少していくのに対し、検証用化合物の誤差は最初は減少していくが、ある回数において極小を示し、その後は増加に転じていることがわかる。この検証用化合物の誤差が増加した状況はニューラルネットワークの過学習を示しており、この誤差が極小を示す回数が最適の学習回数であると考えられる。本研究ではこの方法により、ニューラルネットワークの過学習を解決し、最適の学習回数を決定した。

3.3 中間層ユニット数 の 決定

上記のように、中間層ユニット数についても、過学習と同様、over-fitting の問題を回避するために、検証用化合物の誤差を調べることで最適値を決定した。中間層ユニット数を 2 から 20 まで変えた時の検証用化合物の誤差の変化を Figure 2 に示す。しかし、中間層のユニット数を増やせばパラメータも増えるので、誤差が最小になる中間層ユニット数が最適値であるとはいえない。

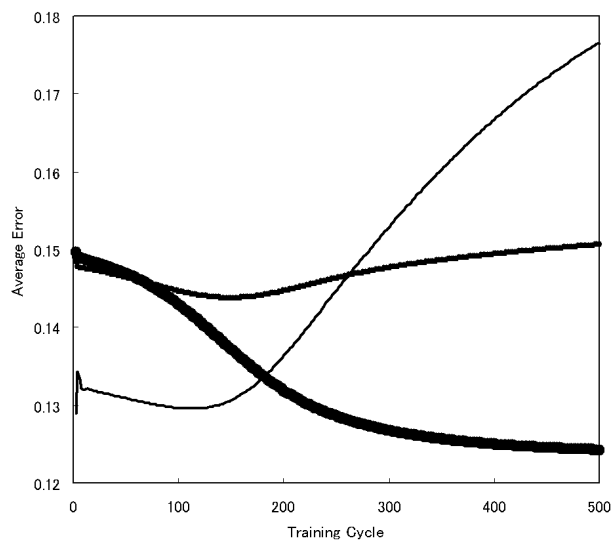


Figure 1. Variations of average errors of training (bold line), validation (thick line) and test (thin line) sets in training cycles

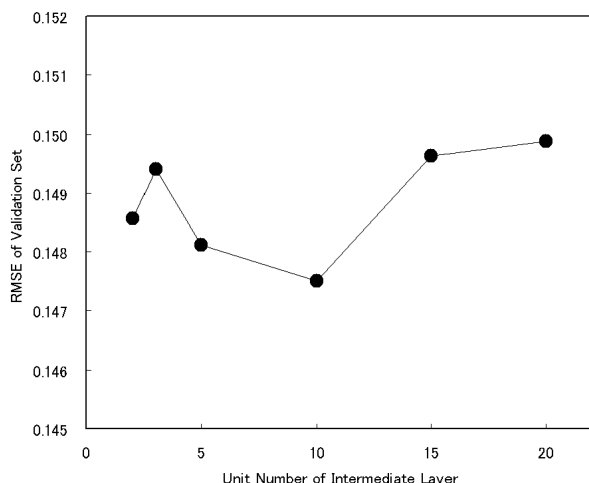


Figure 2. Dependence of average error of validation set on numbers of intermediate layer units

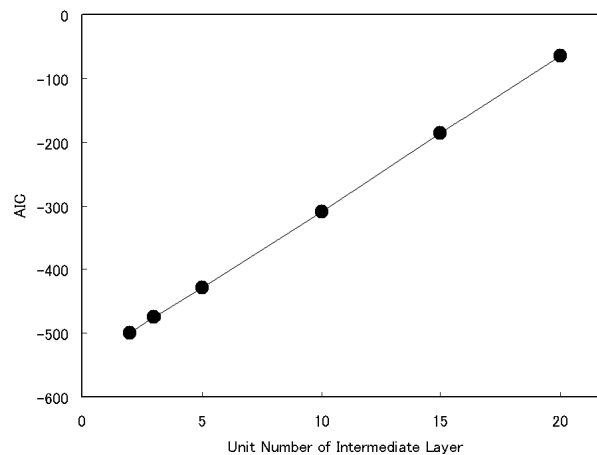


Figure 3. Dependence of AIC on numbers of intermediate layer units

Table 5. Results of our model^a

Random Seed	Cycle ^d	RMSE ^b			CCR ^c of Training Set			CCR ^c of Validation Set			CCR ^c of Test Set		
		Tr ^e	Va ^e	Te ^e	+ ^f	- ^f	Overall	+ ^f	- ^f	Overall	+ ^f	- ^f	Overall
9199	146	0.1058	0.1436	0.1306	0.113	1.000	0.671	0.056	0.978	0.632	0.044	0.992	0.737
9116	199	0.1058	0.1439	0.1296	0.151	0.967	0.664	0.056	0.956	0.618	0.000	0.992	0.725
16931	250	0.1072	0.1439	0.1336	0.132	1.000	0.678	0.111	0.967	0.646	0.044	0.992	0.737
22801	179	0.1055	0.1440	0.1305	0.094	1.000	0.664	0.093	0.989	0.653	0.044	0.992	0.737
10111	2604	0.1058	0.1440	0.1747	0.491	0.856	0.720	0.389	0.789	0.639	0.489	0.623	0.587
22157	277	0.1055	0.1441	0.1339	0.075	1.000	0.657	0.037	0.967	0.618	0.067	0.992	0.743
18382	380	0.1082	0.1441	0.1303	0.094	1.000	0.664	0.037	0.978	0.625	0.000	0.992	0.725
2417	441	0.1053	0.1441	0.1292	0.132	0.956	0.650	0.093	0.956	0.632	0.000	0.992	0.725
4418	154	0.1068	0.1441	0.1315	0.094	0.989	0.657	0.037	0.978	0.625	0.000	1.000	0.731
29762	258	0.1138	0.1441	0.1301	0.094	1.000	0.664	0.130	0.967	0.653	0.000	0.975	0.713

^aUnit number of intermediate layer=2

^bRMSE : Root mean square error between output and teaching data

^cCCR : Correct classification rate

^dOptimum training cycle where RMSE of the validation set was minimum.

^eTr: training set, Va: validation set, Te: test set

^f+ : carcinogenic active, - : carcinogenic inactive

この問題に対して栗田 [18] は AIC (Akaike Information Criteria) 理論 [19] を適用して中間層ユニット数の最適値を決定する方法を提案している。本研究の 3 層構造のニューラルネットワークの場合には、AIC は次式のように定義される :

$$AIC = Nk \log(\sigma^2) + 2K$$

ここで、N は学習データの数、k は出力層のユニット数、 σ^2 は出力値と教師データとの誤差の分散、K はニューラルネットワークに含まれるパラメータの数である。今の場合に上記の誤差の結果を用いて AIC を計算し、中間層ユニット数に対してプロットすると Figure 3 のようになった。この結果から AIC が最小にな

る中間層ユニット数の最適値を 2 に決定した。

3.4 重み初期値の決定

上記のように、局所解の問題を回避するために、重みとしきい値の初期値発生のための random seed を 1 から 30000 まで変えながら、検証用化合物に対する誤差を調べた。一例として中間層ユニット数 2 の場合の検証用化合物の誤差の最小位から 10 位までの結果を Table 5 に示す。この結果から、最適解は重み初期値の random seed が 9199 の場合であると判定し、この

場合のテスト用化合物の的中率を本法の予測的中率とした。

3.5 結果の比較

以上のような方法で、検証用化合物について最適化したモデルに対して、テスト用化合物のデータを適用して求めた我々のモデルの予測的中率の結果を PTC の参加者の結果 [7] と比較して Table 6 に示す。このように、記述子を詳細に検討し、ニューラルネットワークを注意深く最適化した本研究のモデルは、参加者のモデルの結果のうちで公表された結果のどれよりも高い予測的中率が得られることが分かった。

ただし、上記のように、本研究の的中率は、テスト用化合物について最適化したモデルで得られたのではないという点では、コンテストの参加者と同じではあるが、参加者は正解を知らない条件で解析しているので、比較は公平ではない。参加者のモデルの結果のうちで公表されていない結果が本研究の的中率を超える可能性があることは否定できない。

本研究の的中率が高くなった理由としては、記述子の選択とニューラルネットワークという非線形解析モデルの利用の 2 点が考えられる。参加者の予測モデルの大半では、前記のように非常に多数の記述子と比較的単純な線形解析モデルが用いられているため、構造情報のみから多種類の化合物の発ガン性を予測するという問題では、このような機械的な解析モデルでは良好な成績を得にくいと考えられる。

すなわち、上記のようにコンテストの主催者から公開された 7000 種類以上にのぼる記述子の大半は、化合物中の種々の原子、結合、部分構造の有無、個数などを示す記述子であり、ほとんど 0 の値をもつ sparse な記述子である。このような記述子を用いて比較的単純な線形モデルで機械的に用いて解析しても良好な成績を得ることは困難であろう。

Table 6. Correct classification rates of various models

No	Model name	Carcinogen	Non-carcinogen	Overall
1	This work	0.044	0.992	0.737
2	ANI1	0.100	0.870	0.664
3	ANI2	0.077	0.941	0.710
4	ANI3	0.159	0.860	0.672
5	ANU1	0.611	0.402	0.458
6	ANU2	0.350	0.653	0.572
7	ANU3	0.572	0.300	0.373
8	BAUS	0.312	0.773	0.649
9	BAUN	0.251	0.771	0.631
10	BAUP	0.352	0.712	0.616
11	GONS	0.251	0.850	0.689
12	JAN0	0.363	0.671	0.589
13	JAN4	0.240	0.732	0.601
14	KWA1	0.169	0.909	0.711
15	LEU1	0.900	0.071	0.293
16	LEU2	0.098	0.832	0.635
17	LISI	0.271	0.740	0.615
18	LUC1	0.503	0.501	0.502
19	LUC2	0.210	0.801	0.643
20	LUC3	0.210	0.771	0.621
21	PLE1	0.253	0.633	0.531
22	PLE2	0.141	0.781	0.609
23	PLE3	0.391	0.463	0.443
24	SMU1	0.350	0.653	0.572
25	SMU2	0.210	0.740	0.598
26	SMU3	0.523	0.483	0.494
27	SUT	0.523	0.531	0.529
28	VINI	0.411	0.663	0.596
29	WAI1	0.652	0.412	0.476
30	WAI2	0.332	0.694	0.597
31	WAI3	0.442	0.552	0.522

それに対し、本研究では記述子の内容を吟味し、このような sparse な記述子は排除し、Table 2 に示す記述子のみを用いて解析を行い、参加者より良好な結果を得た。このような記述子選択の重要性については、今回のコンテスト以前に開発されていた既存のモデル、たとえば CASE [20–24]、TOPKAT [25–28]、REPAD [29]、COMPACT [30]、FALS [31–33] などの性能が十分でない原因として、原子、結合、部分構造の数などの sparse な記述子が用いられていることから支持される。

3.6 発ガン性データの問題

PTC コンテスト参加者の内で予測的中率が最も高いのは KWA1 の 0.711 で、それに次ぐのは ANI2 の 0.710 である。PTC コンテストの Model Evaluation の段階では、テスト用化合物について各モデルで得られた結果が集約され、主催者により ROC 分析が行われ

評価された。我々のモデルの成績を ROC 分析した結果を Figure 4 に示す。PTC コンテストでは、US NIEHS と EPA の専門家が理解のしやすさとその毒性研究における有用性の観点から各レポートを評価した。その結果、全モデルのうちで KWA1 が最良モデルと評価された [7, 8]。

しかし、我々のモデルおよび上記の 2 モデルでは、テスト化合物のうちで非発ガン性化合物の的中率はいずれも好成績であるのに対し、発ガン性化合物の的中率はきわめて低い。このような観点から、Table 6 および Figure 4 の結果を分析すると、我々とコンテスト参加者の成績は大まかに 3 つのグループに分けられる。

第 1 グループは、我々及び ANI、KWA であり、上記のように発ガン予測率はきわめて低いが、非発ガン予測率が高く、両者を合わせた予測率も高いグループである。第 2 グループは、ANU、LEU、WAI であり、第 1 グループとは逆に、発ガン予測率は高いが、非発ガン予測率がきわめて低く、両者を合わせた予測率もかなり低いグループである。第 3 グループは、残りのモデルで、発ガン予測率も非発ガン予測率も両者を合わせた予測率も中程度のグループである。このような分類は、各モデルの記述子の選択あるいは相関解析法の違いを反映していると考えられる。

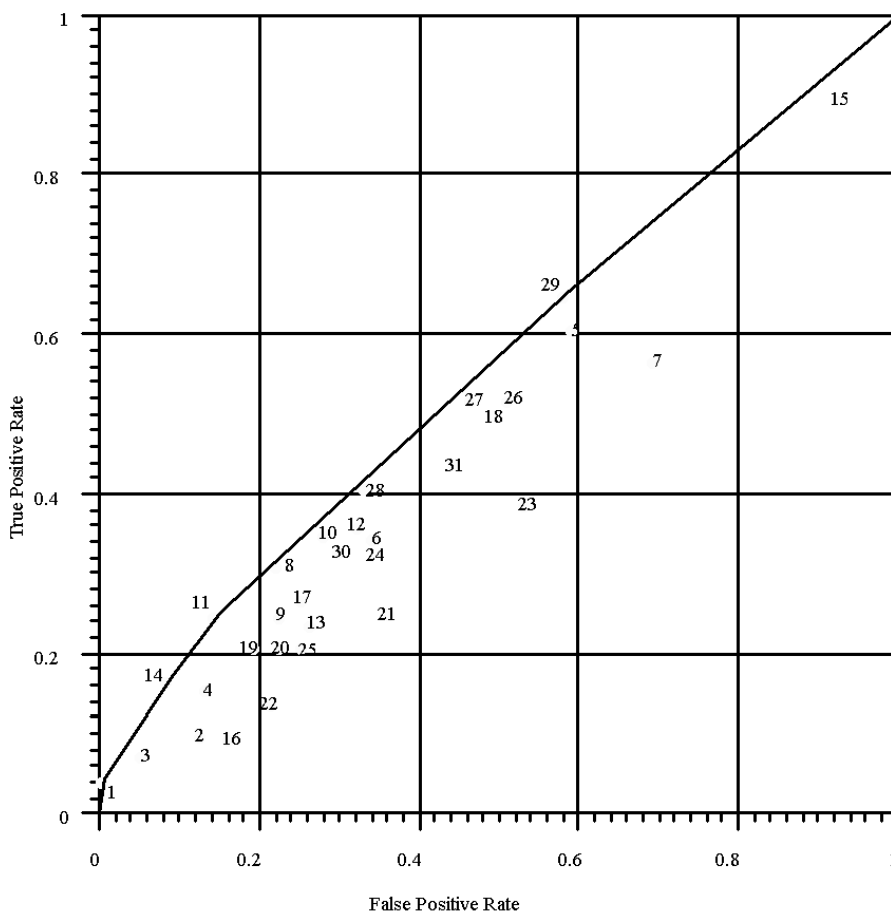


Figure 4. ROC points and convex hull of various models (Numbers in the figure correspond to those in Table 6)

我々のモデルの場合、学習用、検証用、テスト用各化合物に対するニューラルネットワークの出力値の分布を Figure 5 に示す。この図のテスト用化合物の出力値の分布から、非発ガン性化合物は好成績で予測できるのに対し、発ガン性化合物がほとんど予測できていないことが分かる。また、PTC コンテストのその他の大半のモデルでも発ガン性化合物より非発ガン性化合物的中率が高い。

この結果は実用的にはきわめて問題である。すなわち、冒頭に記したように、QSAR モデルによる有害性予測を動物試験の前段階のスクリーニングと位置付けるならば、発ガン性化合物を非発ガンと判定して見逃すことは好ましくないからである。すなわち、非発ガン性化合物よりも発ガン性化合物の方をできるだけ高い精度で予測することが実用上からは望ましいからである。

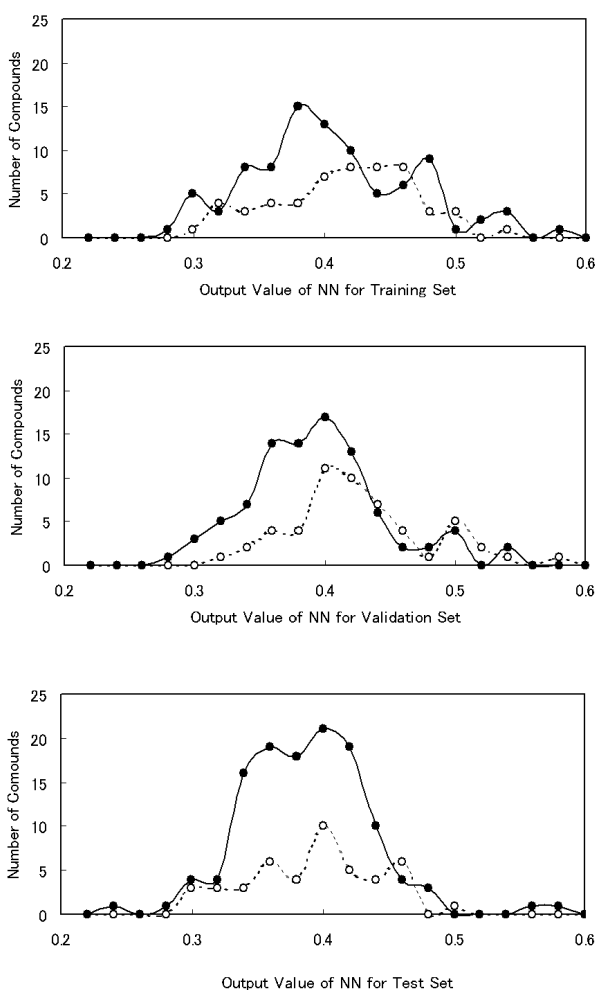


Figure 5. Histogram of output values of neural network for carcinogen (○) and non-carcinogen (●) of training, validation and test sets

このように、PTC コンテストのデータを用いた我々のモデルを含む大半のモデルが、非発ガン性は高成績で、しかし発ガン性は低成績でしか予測できなかった原因としては、Table 1 に示すように、モデル構築用化合物の数が発ガン性 107 種類、非発ガン性 180 種類と不均衡なために、モデルの学習が非発ガン性化合物に偏って行われたためと考えられる。

また、Table 6 に見られるように、PTC コンテストの応募者の中には発ガン性化合物的中率が非発ガン性化合物より高くなったモデルも幾つかある。しかし、それらのモデルによる発ガン性、非発ガン性、両者を合計した的中率はいずれも 50% 前後であり、QSAR による発ガン性予測モデルとしてはきわめて不十分な成績である。この原因としては、NTP のモデル構築用化合物群が構造的に広範囲なバルクの化学物質を含み、かつ比較的小さな分子が多いのに対し、FDA のテスト化合物群は医薬品に偏り、化学構造の面でかなり異なっている点が指摘されている [9, 10]。

3.7 記述子の問題

PTC コンテストで公開されたデータに関しては、発ガン性データ以外に、記述子についても問題がある。QSAR に用いられる記述子に関しては、化合物の構造的、物理化学的、電子的、立体的、トポロジカル的性質など多種多様な記述子があり、その中で今回用いた 37 種類の記述子が発ガン性予測に対して最適であるとは結論できない。

特に、今回 PTC で公開された記述子に関しては、

- ・上記の sparse な記述子が多い、

という点の他に、

- ・意味が不明確な記述子が多い、
- ・発ガン性の有無と相関の高い有効な記述子が少ない、
- ・記述子を計算した時の立体配座が不明、

などの問題点が指摘されている。

QSAR による発ガン性予測に関してはこれまで数多くの研究が報告されている [34-44] が、それらの多くは同族体、いわゆる congener を対象とするものである。このような場合は発ガンの機構がかなり類似している可能性があり、そのため発ガン性データと相関の高い記述子を見出すことが容易である。

それに対して、本研究で扱ったような不特定の化合物群、いわゆる non-congener では、化合物の構造は多

様であり、発ガンの機構も複雑であり、それらの発ガン性を統一的に説明できる記述子を見出すことはきわめて困難な問題である。事実、今回の発ガン性データと記述子の値との相関係数は Table 2 に示すようになり、発ガン性と単独で相関の高い記述子はほとんど皆無であった。

生物に対する化学物質の種々の有害性の内で、生物体内における蓄積性などのように、non-congener においても相関の高い log P などの記述子を用いた単純なモデルで予測できる場合もある。しかし、発ガン性は種々の有害性の中でも有効なモデルを見出すことがきわめて困難な部類に入ると考えられる。多種多様な化合物群の発ガン性を統一的に予測できるようなモデルを開発する問題は多くの研究者が取り組んでいる最先端の研究課題である [45–49]。

3.8 今後の課題

有機化合物以外の一般の化合物も含めて汎用性の高い有害性予測モデルを開発する場合、3通りの方法が考えられる。第1は全ての化合物を1個の予測モデルで解析する方法であり、当然、予測的中率の低下が予想される。第2は化合物を分類し、それぞれ別個の予測モデルで解析する方法である。しかし、この場合には化合物の分類方法が課題である。第3は幾つかの同族化合物群(例えば発ガン性の場合、芳香族化合物、ハロゲン化合物、含窒素化合物など)にはそれぞれ別個のモデルを構築し、さらにそれ以外の化合物をまとめて1個の予測モデルで解析する方法である。しかし、これらのうちのどの方法が最適であるかは今後の研究課題である

我々は今後、発ガン性を含む種々の有害性について、有害性の実測データの厳選、有効な記述子データの開発、及び予測モデルの最適化を行うことにより、高精度、高汎用性の有害性予測モデルを構築し、公開したいと考えている。

4 結論

3層構造のニューラルネットワークを用いて Predictive Toxicology Challenge 2000-2001 のコンテストで公開された多種類の有機化合物の発ガン性データを解析し、参加者の結果と比較した。入力記述子、中間層ユニット数、学習回数、重みの初期値など、ニューラル

ネットワークの構造と学習における種々の問題点に細心の注意を払いながら、学習とテストを行った。その結果、既存のモデルより成績のよい予測モデルを開発することができた。しかし、多種多様な有機化合物の発ガン性の予測はきわめて困難な問題であり、今後、高精度の有害性予測モデルを構築するためには、有効な記述子を見いだすことや、発ガン性データの厳選などが必要であることが分かった。

本研究は科学研究費補助金基盤研究(A)(1) 14209022により行われた。また、情報量基準について助言いただいた産業技術総合研究所脳神経情報研究部門の栗田多喜夫博士に感謝します。

参考文献

- [1] F. Chen, G. Shuurman ed., *Quantitative Structure-Activity Relationships in Environmental Sciences-VII*, SETAC (1997).
- [2] 松尾昌季, *QSAR(定量的構造活性相関)手法を用いた化学物質の手計算による毒性予測*, Life-Science Information Center (1999).
- [3] R. Benigni, *Quantitative Structure-Activity Relationship (QSAR) Models of Mutagens and Carcinogens*, CRC Press (2003).
- [4] J. D. Walker ed., *QSARs for Pollution, Toxicity Screening, Risk Assessment, and Web Applications*, SETAC (2003).
- [5] R. W. Tennant, J. Spalding, S. Stasiewicz, J. Ashby, *Mutagenesis*, **5**, 3 (1990).
- [6] R. Benigni, *11th European Symposium on Quantitative Structure-Activity Relationship* (1997), p. 293.
- [7] <http://www.informatik.uni-freiburg.de/~ml/ptc/>
- [8] C. Helma, S. Kramer, *Bioinformatics*, **19**, 1179 (2003).
- [9] H. Toivonen, A. Srinivasa, R. D. King, S. Kramer, C. Helma, *Bioinformatics*, **19**, 1183 (2003).
- [10] R. Benigni, A. Giuliani, *Bioinformatics*, **19**, 1194 (2003).

- [11] V. G. Blinova, D. A. Dobrynin, V. K. Finn, S. O. Kuznetsov, E. S. Pankratova, *Bioinformatics*, **19**, 1201 (2003).
- [12] T. Okada, *Bioinformatics*, **19**, 1208 (2003).
- [13] J. Zupan, J. Gasteiger, *Neural Networks for Chemists*, VCH (1993).
田辺和俊, 長塚義隆 (訳), 化学者のためのニューラルネットワーク入門, 丸善 (1996).
- [14] J. Devillers, *Neural Networks in QSAR and Drug Design*, Academic Press (1996).
- [15] K. L. Peterson, Artificial Neural Networks and Their Use in Chemistry, in *Reviews in Computational Chemistry Volume 16*, ed. by K. B. Lipkowitz, D. B. Boyd, Wiley-VCH (2000).
- [16] 田辺和俊, 松本高利, *J. Comput. Chem. Jpn.*, **1**, 23 (2002).
- [17] 過学習 (over-training または over-learning) と over-fitting を区別しない場合が多いが、ここでは文献 [10] に従い、学習の過剰繰り返しを過学習、中間層ユニット数などの設定過剰を over-fitting と定義する。
- [18] 栗田多喜夫, 電子情報通信学会技術報告, **PRU89-16**, 17 (1989).
- [19] H. Akaike, *IEEE Trans. Automatic Control*, **AC-19**, **6**, 716 (1974).
- [20] G. Klopman, *J. Am. Chem. Soc.*, **106**, 7315 (1984).
- [21] G. Klopman, A. N. Kalos, H. S. Rosenkrantz, *Mol. Toxicol.*, **1**, 61 (1987).
- [22] H. S. Rosenkrantz, G. Klopman, *Mutagenesis*, **5**, 333 (1990).
H. S. Rosenkrantz, G. Klopman, *Mutagenesis*, **5**, 425 (1990).
- [23] G. Klopman, *Quant. Struct. Act. Relat.*, **11**, 176 (1992).
- [24] G. Klopman, H. S. Rosenkrantz, *Mutation Res.*, **305**, 33 (1994).
- [25] K. Enslein, P. N. Craig, *J. Environ. Pathol. Toxicol.*, **2**, 115 (1978).
- [26] K. Enslein, P. N. Craig, *J. Toxicol. Environ. Health*, **10**, 521 (1982).
- [27] K. Enslein, T. R. Lander, M. E. Tomb, W. G. Landis, *Teratogenesis Carcinogenesis Mutagenesis*, **3**, 503 (1983).
- [28] K. Enslein, V. K. Gombar, B. W. Blake, *Mutation Res.*, **305**, 47 (1994).
- [29] R. Benigni, *Mutagenesis*, **6**, 423 (1991).
- [30] D. F. V. Lewis, C. Ioannides, D. V. Parke, *Mutagenesis*, **5**, 433 (1990).
- [31] I. Moriguchi, S. Hirono, Q. Liu, Y. Matsushita, T. Nakagawa, *Chem. Pharm. Bull.*, **38**, 3373 (1990).
- [32] I. Moriguchi, S. Hirono, Y. Matsushita, Q. Liu, I. Nakagome, *Chem. Pharm. Bull.*, **40**, 930 (1992).
- [33] I. Moriguchi, S. Hirono, Q. Liu, I. Nakagome, *Quant. Struct. Act. Relat.*, **11**, 325 (1992).
- [34] D. Villemin, D. Cherqaoui, A. Mesbah, *J. Chem. Inf. Comput. Sci.*, **34**, 1288 (1994).
- [35] X-H. Song, M. Xiao, R-Q. Yu, *Comput. Chem.*, **18**, 391 (1994).
- [36] S. Hatrik, P. Zahradnik, *J. Chem. Inf. Comput. Sci.*, **36**, 992 (1996).
- [37] M. Vracko, *J. Chem. Inf. Comput. Sci.*, **37**, 1037 (1997).
- [38] G. Gini, M. Lorenzini, *J. Chem. Inf. Comput. Sci.*, **39**, 1076 (1999).
- [39] R. Vendrame, R. S. Braga, Y. Takahata, D. S. Galvao, *J. Chem. Inf. Comput. Sci.*, **39**, 1094 (1999).
- [40] M. J-Heravi, F. Parastar, *J. Chem. Inf. Comput. Sci.*, **40**, 147 (2000).
- [41] S. C. Basak, G. D. Grunwald, B. D. Gute, K. Balasubramanian, D. Opitz, *J. Chem. Inf. Comput. Sci.*, **40**, 885 (2000).
- [42] D. Bahler, B. Stone, C. Wellington, D. W. Bristol, *J. Chem. Inf. Comput. Sci.*, **40**, 906 (2000).

- [43] F. R. Burden, M. G. Ford, D. C. Whitley, D. A. Winkler, *J. Chem. Inf. Comput. Sci.*, **40**, 1423 (2000).
- [44] F. R. Burden, D. A. Winkler, *Chem. Res. Toxicol.*, **13**, 436 (2000).
- [45] 岡田孝, 第 24 回情報化学討論会講演要旨集, p.13 (2001)
- [46] 鈴木孝弘, 黒田泰史, 第 2 回グリーンサステナブルケミストリーシンポジウム 2001 よこはま講演要旨集, p.121 (2001)
- [47] Z. Zhou, Q. Dai, T. Gu, *J. Chem. Inf. Comput. Sci.*, **43**, 615 (2003).
- [48] H. Sun, *J. Chem. Inf. Comput. Sci.*, **44**, 748 (2004).
- [49] H. Sun, *J. Chem. Inf. Comput. Sci.*, **44**, 1506 (2004).

Neural Network Prediction of Carcinogenicity of Diverse Organic Compounds

Kazutoshi TANABE^{a*}, Norihito OHMORI^a, Shuichiro ONO^a, Takahiro SUZUKI^b,
Takatoshi MATSUMOTO^c, Umpei NAGASHIMA^d and Hiroyuki UESAKA^e

^aDepartment of Management Information Science, Chiba Institute of Technology
Tsudanuma 2-17-1, Narashino, Chiba 275-0016, Japan

^bFaculty of Economics, Toyo University
Hakusan 5-28-20, Bunkyo-ku, Tokyo 112-8606, Japan

^cInstitute of Multidisciplinary Research for Advanced Materials, Tohoku University
Katahira 2-1-1, Aoba, Sendai, Miyagi 980-8577, Japan

^dGrid Research Center, National Institute of Advanced Industrial Science and Technology
Umezono 1-1-1, Tsukuba, Ibaraki 305-8568, Japan

^eDepartment of Regional Science, Toyama University of International Studies
Higashikuromaki 65-1, Ohyama, Kamishinkawa, Toyama 930-1292, Japan

*e-mail: tanabe@pf.it-chiba.ac.jp

A three-layered neural network model to predict the hazards of a variety of compounds based on a quantitative structure-activity relationship was developed. The inputs were 10 principal components from 37 kinds of molecular descriptors calculated with MO programs. For the output the data used in the Predictive Toxicology Challenge (PTC) 2000-2001 contest were employed, containing 454 compounds with the carcinogenic activity of male rats. The total database of 454 compounds was split into training (144 compounds), validation (143) and test (167) sets. To solve the problems such as over-training, over-fitting and local minimum in training the neural network with the error-back-propagation algorithm, various conditions of the network such as the training cycles and neuron numbers of the intermediate layer were optimized. The optimum model showed a correct classification rate close to 74 %, higher than any of the PTC contestants.

Keywords: Quantitative structure-activity relationship, Neural network, Carcinogenicity prediction, Principal component analysis, Over-training