

ニューラルネットワークによる有機塩素化合物の発ガン性の予測

田辺 和俊^{a*}, 松本 高利^b

^a 産業技術総合研究所計算科学研究部門, 〒 305-8568 つくば市梅園 1-1-1

^b 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

*e-mail: k-tanabe@aist.go.jp

(Received: January 10, 2002; Accepted for publication: February 12, 2002; Published on Web: March 22, 2002)

構造活性相関により化学物質の構造から有害性を高い精度で予測する手法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。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% の予測手法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, 有機塩素化合物

1 緒言

化学物質の生物に対する毒性、大気・水・土壌などの自然環境条件における分解性や蓄積性、および生物体内における蓄積性や濃縮性など、化学物質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) の手法を用いた研究が行われている [1]。また、化学物質の有害性を予測するシステムが開発されているが、性能はいずれも不十分である。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 [2–6]、TOPKAT [7–10]、REPAD [11]、COMPACT [12]、FALS [13–15] などのシステムはいずれも 60% 台の低い成績であった [16]。このように既存のシステムの性能が低い原因としては、化学構造と有害性データとの間に線形関係を仮定し、重回帰分析のような単純な手法で解析しているためと考えられる。すなわち、化学物質の構造と有害性と間の関係は線形関係のように単純なものではなく、きわめて複雑な関係と考えられるからである。したがって、既存のシステムに採用されている重回帰分析ではなく、もっと複雑な関係を強力に解析する手法を用いれば、既存のシステムよりも性能が高い予測手法を開発できる可能性がある。

そのような複雑な関係の解析に有効な手法として

ニューラルネットワークがあり、化学物質の有害性予測にニューラルネットワークを用いた研究はかなり多い [17–27]。しかし、それらの研究では多種多様の記述子を用いて芳香族化合物やハロゲン化合物の発ガン性データをニューラルネットワークで解析しているが、予測的中率はいずれも 80% 以下である。それに対し、我々は記述子として化学物質の構造中の結合数だけを用いて有機塩素化合物の発ガン性との相関をニューラルネットワークで解析し、83% の予測的中率を得た [28]。この方法では記述子が化学構造から簡単に求まるので、迅速な予測が可能である反面、このような単純な記述子ではこれ以上の予測的中率の向上は困難である。たとえば、ベンツピレンをはじめとする縮合多環芳香族炭化水素化合物 (PAH) ではベンゼン環の結合様式の違いによる多数の構造異性体があるが、このような構造の違いによる発ガン性の有無を前回の方法で説明することは不可能である。そこで前回のこの問題点を改良するために今回は新たな記述子を検討した。

2 方法

前報の結果と比較するために、発ガン性のデータは NTP (National Toxicology Program) のデータベース

[29] から抽出した同じ有機塩素化合物 41 種類のデータを用いた。ニューラルネットワークも Figure 1 に示すように、3 層ニューラルネットワークの入力層に化合物の記述子を入力し、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。

構造活性相関に用いられる記述子には、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、多種多様の記述子がある。今回は構造異性体の構造の違いを反映する記述子として Gibbs 標準生成自由エネルギー、イオン化ポテンシャル、LUMO エネルギー、HOMO-LUMO のエネルギー差、Connolly 体積、分子量、log P の 7 種類を検討した。記述子としては他にハードネスや電気陰性度なども考えられるが、今回は量子化学計算から直接得られる上記の記述子を検討した。これらの記述子は化合物の構造的・物理化学的・電子的・立体的性質を代表しており、構造異性体で値が異なるため、そのような異性体の発ガン性の有無を説明できる可能性がある。そこで、これらの記述子の内、log P と Gibbs 標準生成自由エネルギーは CS ChemDraw 5.0 で計算し、イオン化ポテンシャル、LUMO エネルギー、HOMO-LUMO のエネルギー差、

Connolly 体積は MOPAC93 revision2 で計算した。なお、その際にこれらの記述子の値は化合物の立体構造に依存するので、各化合物について CONFLEX を用いて最安定エネルギーの構造を求め、その構造について記述子の値を計算した。それらの記述子の計算値を Table 1 に示す。

これらの記述子はその値が -0.9 ~ 0.9 の範囲に収まるように以下の式で規格化した後、ニューラルネットワークに入力した。

$$\text{規格化値} = [(x - x_{min}) / (x_{max} - x_{min})] \times 1.8 - 0.9$$

ここで x 、 x_{min} 、 x_{max} は記述子の規格化前の値、41 種類の化合物中でのその記述子の最小値と最大値である。出力層には発ガン性ありの化合物に 1.0、なしの化合物に 0.0 の値を教師データとして入力した。ニューラルネットワークのユニット数は入力層 7、中間層 3、出力層 1、バイアスなし、収束条件は 0.1、学習定数は $=5.0$ 、 $=0.5$ とし、エラーバックプロパゲーション法で学習を行った。中間層のユニット数については、2 の時はニューラルネットワークが収束せず、また、4 以上では予測的中率が低下したので、今回は中間層のユニット数は 3 を用いた。

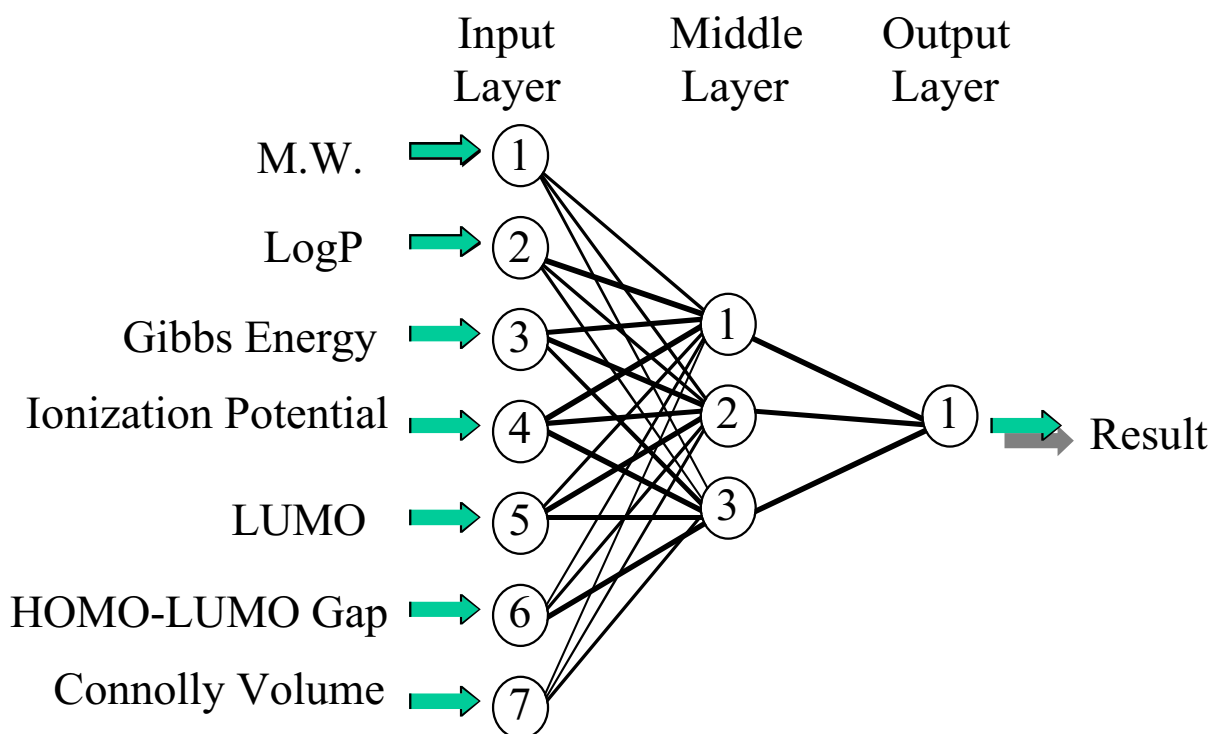


Figure 1. Structure of the neural network. Thickness of a connected line shows the absolute value of the weight.

Table 1. Seven descriptors, output values of neural networks, and predicted and experimental carcinogenicity

Molecular Name	MW	LogP	GFE	IP	LUMO	HLGap	CV	Output	Pred	Exp
Aldrin	364.91	3.50	222.6	9.81	-0.316	9.492	264.08	0.9997	+	+
Allyl chloride	76.53	1.63	50.3	10.47	0.668	11.141	70.55	1.0000	+	+
Benzyl chloride	126.59	2.59	108.5	9.70	0.075	9.773	113.63	0.9874	+	+
Carbon tetrachloride	153.82	2.43	-87.3	12.38	-1.117	11.261	84.74	0.9989	+	+
Chlordane	409.78	4.20	71.4	10.06	-0.528	9.53	275.94	0.9997	+	+
9-Chloro-10-chloromethyl-anthracene	261.15	5.23	348.4	8.28	-1.292	6.99	207.34	0.9997	+	+
Chloroethane	64.51	1.30	-46.0	11.15	1.498	12.651	61.15	1.0000	+	+
Chloroform	119.38	1.71	-80.7	11.77	-0.303	11.468	71.00	0.9908	+	+
7-(Chloromethyl)benza(a)-anthracene	276.76	5.70	500.6	8.29	-1.041	7.253	233.16	0.9997	+	+
9-Chloromethyl-10-methyl-anthracene	240.73	5.12	368.7	8.10	-1.086	7.018	215.16	0.9997	+	+
3-Chloro-2-methylpropene	90.55	1.73	50.2	10.10	0.622	10.72	87.04	0.9981	+	+
p,p'-DDD	320.04	5.77	220.0	9.56	-0.269	9.286	249.35	0.9708	+	+
p,p'-DDE	318.02	5.81	288.0	9.34	-0.480	8.862	247.29	0.9997	+	+
1,4-Dichlorobenzene	147.00	3.11	78.6	9.52	-0.216	9.308	112.57	0.0000	-	+
1,2-Dichloroethane	98.96	1.59	-57.9	11.42	0.685	12.102	75.41	0.9991	+	+
Dichloromethane	84.93	1.08	-66.3	11.39	0.595	11.985	55.75	0.8398	+	+
1,2-Dichloropropane	112.99	2.00	-51.9	11.38	0.684	12.067	94.32	0.9432	+	+
1,3-Dichloropropane	112.99	1.80	-49.5	11.37	1.020	12.392	94.67	0.9080	+	+
Heptachlor	373.32	3.59	121.0	9.98	-0.481	9.501	251.08	0.9997	+	+
Hexachlorobenzene	284.78	5.37	-7.7	9.91	-1.040	8.872	158.07	0.2534	-	+
Hexachloroethane	236.74	3.21	-99.9	12.18	-0.967	11.215	136.99	1.0000	+	+
Methylallyl chloride	90.55	2.00	42.5	9.49	0.792	10.28	89.79	0.9924	+	+
Mirex	545.54	6.33	161.0	11.19	-0.268	10.923	342.85	0.9996	+	+
Pentachloroethane	202.29	2.71	-93.3	11.87	-0.681	11.188	121.64	1.0000	+	+
1,1,1,2-Tetrachloroethane	167.85	2.40	-78.9	11.79	-0.485	11.308	105.35	0.9996	+	+
1,1,2,2-Tetrachloroethane	167.85	2.22	-86.6	11.66	-0.074	11.581	108.03	1.0000	+	+
Tetrachloroethylene	165.83	2.42	-18.6	9.90	-0.437	9.464	94.18	1.0000	+	+
1,1,2-Trichloroethane	133.40	1.90	-72.3	11.57	0.170	11.737	91.87	0.9930	+	+
Trichloroethylene	131.39	2.25	1.8	9.96	-0.061	9.895	83.82	0.9963	+	+
1,2,3-Trichloropropane	147.43	2.29	-63.9	11.44	0.760	12.202	119.67	0.9979	+	+
Vinyl chloride	62.50	1.66	41.9	10.21	0.856	11.066	52.90	0.7704	+	+
Chlorobenzene	112.56	2.54	100.1	9.56	0.155	9.716	96.60	0.1442	-	-
1-Chlorobutane	92.57	2.20	-29.1	11.13	1.511	12.644	96.38	0.8568	+	-
DDT	354.49	6.33	213.3	9.59	-0.518	9.069	265.76	0.0000	-	-
1,2-Dichlorobenzene	147.00	3.11	78.6	9.60	-0.142	9.46	112.35	0.0000	-	-
1,1-Dichloroethane	98.96	1.43	-60.3	11.42	0.582	12.004	75.07	0.1897	-	-
Hexachlorocyclopentadiene	272.77	2.02	-27.9	9.61	-1.411	8.194	167.05	0.0000	-	-
Lindane	290.83	4.20	-86.0	11.36	-0.151	11.212	203.40	0.0186	-	-
Pentachlorobenzene	250.34	4.81	13.9	9.79	-0.890	8.896	149.09	0.0987	-	-
1,1,1-Trichloroethane	133.40	2.10	-67.0	11.99	-0.265	11.727	88.66	0.0041	-	-
Vinylidene chloride	96.94	1.83	21.4	10.19	0.379	10.569	68.32	0.1498	-	-

MW: molecular weight, LogP: octanol-water partition coefficient, GFE: Gibbs standard free energy of formation (kJ mol^{-1}), IP: ionization potential (eV), LUMO: LUMO energy (eV), HLGap: HOMO-LUMO energy gap (eV), CV: Connolly volume (m^{-30}), Output: output value of neural network, Pred: predicted carcinogenicity, Exp: experimental carcinogenicity.

3 結果と考察

この方法による発ガン性予測の性能を求めるために leave-one-out test を行った。すなわち、全 41 種類の化合物の内 40 種類を学習データとしてニューラルネッ

トワークに学習させ、収束後のニューラルネットワークに残りの 1 つを未学習データとして入力して出力層の結果から発ガン性の予測値を求め、実測値と比較するという操作を全ての化合物について繰り返した。その際、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。その結果は Table 1 に示すように、全化合物 41 種類の中で 3 種類を除く 38 種類について実測値に一致し、本法による予測的中率は 93% となった。

この予測的中率は既存の予測システムの性能を凌ぐものであり、今回採用した 7 種類の記述子とニューラルネットワークを組み合わせた予測手法の威力を実証している。それに対し、既存の予測システムでは前記のように単純な重回帰分析が用いられるので、このような解析手法の違いが的中率の違いの理由と考えられる。ただし、既存の予測システムの性能テストに用いられた化合物は本研究で用いた有機塩素化合物のように限定したものでなく、非常に広範囲の化合物である。したがって対象の化合物が異なるので単純な比較はできないが、対象の化合物群ごと今回のような方法を開発すれば、広範囲の化合物についても的中率の高い予測システムが開発できると考えられる。

本研究の予測的中率が既存のシステムより高くなったもう 1 つの理由は記述子である。既存のシステム、たとえば CASE [2-6]、TOPKAT [7-10]、COMPACT [12]、FALS [13-15] では結合や部分構造の数や有無のようなきわめて単純な記述子が用いられている。しかし、上記のようにこのような単純な記述子では限界があることは明白である。

今回用いた記述子の数値は Table 1 に示すようにそれぞれの記述子単独では発ガン性の有無との相関は認めにくいので、7 種類の記述子が関連しあって発ガン性の有無を説明していると考えられる。Figure 1 には学習収束後のニューラルネットワークについてネットワークの重みの絶対値の大きいものを太線で、絶対値の小さい重みを細線で示すが、どの記述子も発ガン性の説明には不可欠であることが分かる。実際に記述子を 6 種類に減らして解析したが、90% 以上の予測的中率は得られなかった。したがって、今回用いた 7 種類の記述子が必要最小限の記述子であると考えられる。

また、これらのことから、今回用いた 7 種類の記述子により 40 種類の化合物のデータを学習させる際の過学習の可能性は低いものと考えられる。多変量解析の場合に統計的に有効なモデルを組み立てるにはデータ数の 3 分の 1 以下の未知数を用いるべきとされている。今回ではデータ数 40 に対し、未知数であるネットワークの結合数は 24 であり、過学習の可能性が考えられる。しかし、一般に過学習の場合は未知数を減らすと予測率が向上するとされている。今回は上記のようにそのような結果は得られなかったので、過学習

の可能性は低いと考えられる。

ただし、今回用いた記述子に関しては、化合物の構造的・物理化学的・電子的・立体的・トポロジカルの性質など多種多様な記述子があり、その中で今回用いた 7 種類の記述子が最適であるとは断言できない。多種多様な記述子の中から最適の組み合わせを抽出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 [30, 31]。

今回用いた記述子を求めるためには ChemDraw や MOPAC による計算が必要であり、前報で用いた結合数を記述子とする予測ほどの迅速な予測は困難である。特に内部回転異性体が存在する化合物では、今回用いたイオン化ポテンシャル、LUMO エネルギー、HOMO-LUMO のエネルギー差、Connolly 体積などの記述子の値が回転異性体間で異なるので、最安定の回転異性体を探索する操作が必要になる。しかし、内部回転異性体が多数存在する化合物について最安定の異性体を探索するにはかなり時間がかかる場合があり、この点も今後の課題である。

我々は今後、(1) 発ガン性の有無だけでなく定量的な予測を行うこと、(2) 発ガン性以外の種々の有害性の予測を行うこと、(3) 有機塩素化合物以外の広範囲の化合物の予測を行うこと、などを目指している。(1) については発ガンの強さ (TD) を教師データとしてニューラルネットワークで解析することにより定量的な予測が可能である。(2) については我々は既に変異原性と生分解性の予測を検討した [32, 33] が、化学物質の有害性にはその他に様々なものがある。(3) が今後の最大の課題である。汎用性の高い予測システムを開発するためにはあらゆる化合物を 1 個のニューラルネットワークで解析することになり、当然、予測的中率の低下が予想される。的中率を向上させるためには化合物を分類し、それぞれにニューラルネットワークを構築することになるが、その場合の分類方法が課題である。

4 結論

ニューラルネットワークを用いて有機塩素化合物の発ガン性のデータを解析した。記述子として分子量、log P、Gibbs エネルギー、イオン化ポテンシャル、LUMO エネルギー、HOMO-LUMO のエネルギー差、Connolly 体積の 7 種類を取り上げ、41 種類の化合物について分子軌道法のプログラムを用いてこれらの記述子を計算した。3 層構造のニューラルネットワークの

入力層にこれらの記述子を、出力層にはNTPの発ガン性データを教師データとして入力して、エラーバックプロパゲーション法で学習を行った。leave-one-out testを行った結果、予測的中率は93%となり、既存のシステムより高性能の予測手法を開発することができた。

参考文献

- [1] 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).
- [2] G. Klopman, *J. Am. Chem. Soc.*, **106**, 7315 (1984).
- [3] G. Klopman, A. N. Kalos, H. S. Rosenkrantz, *Mol. Toxicol.*, **1**, 61 (1987).
- [4] H. S. Rosenkrantz, G. Klopman, *Mutagenesis*, **5**, 333, 425 (1990).
- [5] G. Klopman, *Quant. Struct.-Act. Relat.*, **11**, 176 (1992).
- [6] G. Klopman, H. S. Rosenkrantz, *Mutation Res.*, **305**, 33 (1994).
- [7] K. Enslein, P. N. Craig, *J. Environ. Pathol. Toxicol.*, **2**, 115 (1978).
- [8] K. Enslein, P. N. Craig, *J. Toxicol. Environ. Health*, **10**, 521 (1982).
- [9] K. Enslein, T. R. Lander, M. E. Tomb, W. G. Landis, *Teratogenesis Carcinogenesis Mutagenesis*, **3**, 503 (1983).
- [10] K. Enslein, V. K. Gombar, B. W. Blake, *Mutation Res.*, **305**, 47 (1994).
- [11] R. Benigni, *Mutagenesis*, **6**, 423 (1991).
- [12] D. F. V. Lewis, C. Ioannides, D. V. Parke, *Mutagenesis*, **5**, 433 (1990).
- [13] I. Moriguchi, S. Hirono, Q. Liu, Y. Matsushita, T. Nakagawa, *Chem. Pharm. Bull.*, **38**, 3373 (1990).
- [14] I. Moriguchi, S. Hirono, Y. Matsushita, Q. Liu, I. Nakagome, *Chem. Pharm. Bull.*, **40**, 930 (1992).
- [15] I. Moriguchi, S. Hirono, Q. Liu, I. Nakagome, *Quant. Struct. Act. Relat.*, **11**, 325 (1992).
- [16] R. Benigni, *Eur. Symp. Quant. Struct.-Act. Relat.*, *11th* (1997), p.293.
- [17] D. Villemin, D. Cherqaoui, A. Mesbah, *J. Chem. Inf. Comput. Sci.*, **34**, 1288 (1994).
- [18] X-H. Song, M. Xiao, R-Q. Yu, *Comput. Chem.*, **18**, 391 (1994).
- [19] S. Hatric, P. Zahradnik, *J. Chem. Inf. Comput. Sci.*, **36**, 992 (1996).
- [20] M. Vracko, *J. Chem. Inf. Comput. Sci.*, **37**, 1037 (1997).
- [21] G. Gini, M. Lorenzini, *J. Chem. Inf. Comput. Sci.*, **39**, 1076 (1999).
- [22] R. Vendrame, R. S. Braga, Y. Takahata, D. S. Galvao, *J. Chem. Inf. Comput. Sci.*, **39**, 1094 (1999).
- [23] M. J-Heravi, F. Parastar, *J. Chem. Inf. Comput. Sci.*, **40**, 147 (2000).
- [24] S. C. Basak, G. D. Grunwald, B. D. Gute, K. Balasubramanian, D. Opitz, *J. Chem. Inf. Comput. Sci.*, **40**, 885 (2000).
- [25] D. Bahler, B. Stone, C. Wellington, D. W. Bristol, *J. Chem. Inf. Comput. Sci.*, **40**, 906 (2000).
- [26] F. R. Burden, M. G. Ford, D. C. Whitley, D. A. Winkler, *J. Chem. Inf. Comput. Sci.*, **40**, 1423 (2000).
- [27] F. R. Burden, D. A. Winkler, *Chem. Res. Toxicol.*, **13**, 436 (2000).
- [28] 松本高利, 田辺和俊, *JCPE Journal*, **11(1)**, 29 (1999).
- [29] <http://ntp-server.niehs.nih.gov/>
- [30] 岡田孝, 第24回情報化学討論会講演要旨集 (2001), p.13.
- [31] 鈴木孝弘, 黒田泰史, 第2回グリーン・サステイ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 (2001), p.121.
- [32] 松本高利, 田辺和俊, 久保隆, 浦野紘平, *化学とソフトウェア*, **22**, 127 (2000).
- [33] 松本高利, 田辺和俊, *化学とソフトウェア*, **22**, 55 (2000).

Prediction of Carcinogenicity of Chlorine-containing Organic Compounds by Neural Network

Kazutoshi TANABE^{a*} and Takatoshi MATSUMOTO^b

^aNational Institute of Advanced Industrial Science and Technology
1-1-1 Umezono, Tsukuba, Ibaraki 305-8568, Japan

^bIMRAM, Tohoku University, Aoba-ku, Sendai, Miyagi 980-8577, Japan

**e-mail: k-tanabe@aist.go.jp*

A neural network was applied to the prediction of the carcinogenicity of 41 kinds of organic chlorine-containing compounds. Seven kinds of structural and quantum-chemical descriptors: molecular weight, log P, Gibbs free energy, ionization potential, LUMO energy, HOMO-LUMO energy gap, and Connolly volume were determined. These descriptors were entered into the input layer of a three-layered neural network, and carcinogenicity data from the NTP database were entered into the output layer as teaching data. The network was trained with an error-back-propagation method, and a leave-one-out test showed a correct classification rate of 93%.

Keywords: Structure-activity relationship, Neural network, Carcinogenicity prediction, Chlorine-containing organic compounds