1O09

化学物質発ガン性データベースの開発及び サポートベクターマシンによる発ガン性予測

○田辺 和俊[1]、鈴木 孝弘[2]、貝原 巳樹雄[3]、小野寺 夏生[1]
[1]筑波大学大学院図書館情報メディア研究科(〒305-8550 つくば市春日 1-2)
[2]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)
[3]一関工業高等専門学校物質化学工学科(〒021-8511 一関市萩荘字高梨)

【緒言】

ガンは日本人の死因の第 1 位であるが、その原因には物理的（放射線、紫外線等）や生物的（ウイルス、ピロリ菌等）要因以外の化学物質による発ガンが最も大きい。しかし、環境中に存在する化学物質の内で発ガン性が判明しているものは僅少であり、動物愛護の観点から動物実験のスクリーニングとして、化学物質の発ガン性を構造から予測する手法が注目されている。現在、構造の類似した同族体については発ガン強度を高い精度で予測できる場合もあるが、任意の構造の化学物質の発ガン性を高精度で予測することは困難である。

この原因の 1 つは、非同族体の発ガン性を統一的に予測するモデルを開発するために必要な信頼性の高い発ガン性データの不足である。動物実験に基づく発ガン性データは、IARC、NTP 等の機関で収集、公開しているが、発ガン性のランクに統一性がなく混乱している。

もう 1 つは、非同族体の発ガン性を構造のみから予測できるかという問題である。Hansch-Fujita による QSAR は元来、同族体を対象としたものであるが、化学物質の発ガン機構は複雑であり、しかも化学物質により発ガン機構への関与の仕方が異なるため、非同族体の発ガン性を構造のみから統一的に予測することは簡単ではない。我々はこのようなきわめて複雑な予測問題に有効なニューラルネットワーク（ANN）を用いて非同族体の発ガン性予測を検討したが、多数の局所解の存在のために最適解が得られず、予測精度を確定できないという問題があった [1]。

そこで本研究では、各種発ガン性 DB に収録されている多数の化学物質についての発ガン性を総合評価して大規模な発ガン性 DB を構築し、近年非線形解析法として注目されているサポートベクターマシン（SVM）[2] を用いてこの発ガン性データを解析することにより、非同族体の発ガン性予測の可能性を検討した。

【発ガン性データベースの開発】

動物実験による発ガン性の実測データは、信頼性が高い IARC、EU、EPA、NTP、ACGIH、JSOH の 6 種の DB [3] から収集した。しかし、それぞれの DB により発ガン性のランクやその信頼度に関する表現が混乱している [4]。そこで、これらの発ガン性のランクを統一するために、PRTR-MSDS のランク付け [5] を参考に統一基準を設定し、これに基づき表 1 のように発ガン性の信頼度を A(+++)、B(++)、C(+)、D(+/ー)、E(ー)の 5 段階にランク付けした。総物質数は 1508 である。

表 1　発ガン性データの統一ランク

ランク	A	B	C	D	E
発ガン性の信頼度	+++	++	+	+/ー	ー
IARC	1	2A,2B		3	
EU	1	2	3		
EPA	A	**B1,B2**	B1,B2,C	D	
NTP	A	**B**	B		ー
ACGIH	A1	**A2,A3**	A2,A3	A4	
JSOH	1	**2A,2B**	2A,2B		
物質数	167	407	186	631	117
設定値	0.9	0.7	0.5	0.3	0.1

注：ランク B における太字は複数の機関で格付けされている場合を表す。

【SVM による発ガン性予測】

以上の発ガン性データを有する化学物質の中から、QSAR 解析が可能な化学物質として、化学構造が不明確なもの、金属原子等を含むもの、混合物、高分子等を除く 908 種の化学物質について、平面構造から立体構造を生成し、構造を最適化した後、CAChe Project-Leader（富士通）に入力して、表 2 に示す 66 種の記述子を作成した。

SVM による解析には LIBSVM [6] を用い、発ガン性データは表 1 に示した値に設定し、SVM による回帰（SVR）を行った。最適条件の探索や予測精度の評価のために、全ての化学物質をランクごとに分子量順にソートしてセット I とセット II に交互に振り分け、cross-validation-test を行った。

SVM では ANN と異なり局所解の問題はなく、解は一義的に決まる。ただし、SVM ではカーネルの選択の問題があり、問題によって最適のカーネルを選択する必要がある。LIBSVM では linear、polynomial、radial basis function（RBF）及び sigmoid の 4 種類のカーネルが用意されているので、これらを試行して最適のカーネルを探索した。

また、ANN と同様、SVM でも過学習の問題があり、パラメータの最適設定が必要である。LIBSVM の SVR では 2 個のパラメータ g（gamma）と c（cost）が調節可能であり、これらのパラメータの数値を大きくすると過学習状態に陥る。そこで、これら 2 個のパラメータをグリッドサーチして検証セットの予測誤差が最小になる最適条件を探索した。

Cross-validation-test によりモデルの最適化を行った結果、得られた予測値のクロス集計表を表 3 に示す。設定値と同一のランクに予測できた物質の割合は全物質中のわずか 29%に過ぎず、また動物実験のスクリーニングという視点からの的中率、すなわち予測値が 0.4 以上になった物質の内で実際に発ガン陽性の物質の割合も 66%に過ぎない。しかし、誤判定の内で特に問題なるのは発ガン陽性を陰性と判定する false negative であるが、A～C ランクの物質の内で発ガン陰性と予測された物質は 1%以下である。

以上の結果から、今回の発ガン性データに対する SVM モデルの予測性能は充分ではないが、有効な記述子を追加することで、実用的な予測レベルまで向上すると考えられる。SVM は、予測性能は ANN より若干劣るものの、局所解の問題がなく最適解が一義的に決まり、ANN より圧倒的に短い時間で処理可能という利点がある。SVM を用いて大規模な発ガン性データを解析することにより、不特定の化学物質の発ガン性を高い精度で予測するシステムの開発が期待できる。

表 2　解析に用いた記述子

分類	記述子	個数
構造的	各種の原子・結合・原子団の個数、環の大きさ・個数等	41
電子的	電荷、双極子モーメント、配座・立体エネルギー、HOMO・LUMO のエネルギー等	11
立体的	分子容、表面積、分子屈折、カッパ形状指数等	6
その他	分子結合指数、logP、分子量等	8

表 3　発ガン性予測結果（%）のクロス集計表

	設定値				
予測値	0.9	0.7	0.5	0.3	0.1
1.0～0.8	15	1	0	0	0
0.8～0.6	45	28	11	6	4
0.6～0.4	35	63	69	67	58
0.4～0.2	5	8	19	25	32
0.2～0.0	0	0	1	2	6
計	100	100	100	100	100

参考文献

[1] 田辺和俊, 大森紀人, 小野修一郎, 鈴木孝弘, 松本高利, 長嶋雲兵, 上坂博亨, *J. Comp. Chem. Jpn.*, **4**, 89-100 (2005)

[2] V. N. Vapnik, *The Natute of Statistical Learning Theory*, Springer (1995)

[3] 日本化学物質安全・情報センター、化学物質の発がん性評価とその分類基準 (2007)

[4] 田辺和俊, 大森紀人, 小野修一郎, 松本高利, 長嶋雲兵, 上坂博亨, 鈴木孝弘, 情報知識学会誌, **16**, 63-84 (2006)

[5] 浦野紘平, PRTR-MSDS 対象化学物質の毒性ランクと物性情報, 化学工業日報社 (2001)

[6] C. C. Chang, C. J. Lin, http://www.csie.ntu.edu.tw/~cjlin/libsvm/