化学物質発ガン性データベースおよび化学構造からの発ガン性予測システムの開発

田辺和俊*　栗田多喜夫**　西田健次*　鈴木孝弘***

Construction of a Chemical Carcinogenicity Database and a Carcinogenicity Prediction System from Chemical Structures

Kazutoshi TANABE*, Takio KURITA**, Kenji NISHIDA* and Takahiro SUZUKI***

Abstract

A database system which consists of experimental carcinogenicities and their reliabilities for a diverse range of chemicals, and a quantitative structure-activity relationship (QSAR) system for satisfactorily predicting carcinogenicities of a wide variety of chemicals have been constructed as a tool to present information on carcinogenicities of numerous chemicals existing in our daily life and the environment. The chemical carcinogenicity database was constructed by collecting experimental carcinogenicity data on about 1,500 chemicals from six sources including IARC and NTP databases. The carcinogenicity data were ranked into six unified categories on the basis of their reliabilities. A wide variety of about 900 organic chemicals were selected from the database for QSAR modeling, and molecular descriptors were calculated using the Dragon software. To construct the QSAR system for predicting carcinogenicities of diverse chemicals with a satisfactory performance level, the relationship between the carcinogenicity data with improved reliability and a subset of significant descriptors selected from the Dragon descriptors was analyzed utilizing a support vector machine (SVM) method. The classification function (SVC) for weighted data

* 産業技術総合研究所ヒューマンライフテクノロジー研究部門：〒 305-8568 茨城県つくば市梅園 1-1-1 つくば中央第 2
Human Technology Research Institute, National Institute of Advanced Industrial Science and Technology, AIST Central 2, 1-1-1 Umezono, Tsukuba, Ibaraki 305-8568 JAPAN
** 広島大学大学院工学研究院：〒 739-8527 広島県東広島市鏡山 1-7-1
Faculty of Engineering, Hiroshima University, 1-4-1 Kagamiyama, Higashi-Hiroshima, Hiroshima 739-8527 JAPAN
*** 東洋大学自然科学教室：〒 112-8606 東京都文京区白山 5-28-20
Natural Science Lab., Toyo University, 5-28-20 Hakusan, Bunkyo-ku, Tokyo 112-8606 JAPAN

in the LIBSVM program was used to classify chemicals into two carcinogenic categories (positive or negative), where weights were set depending on the reliabilities of the carcinogenicity data. Seven models were created and tested: a batch model, combination models of SVMs with a boosting, with a bagging, with a decision tree method, with a repêchage decision tree method, and serial and parallel combination models of SVMs for congeneric chemicals. The quality and stability of the models were tested by performing a dual cross-validation procedure. The parallel model developed on the basis of grouping of chemicals into twenty substructures predicts the carcinogenicities of a wide variety of chemicals with a satisfactory overall accuracy of approximately 80%. The predicting performance was higher than any of previously proposed models for diverse chemicals.

Key Words: carcinogenicity database; carcinogenicity prediction; molecular descriptors; quantitative structure-activity relationship (QSAR); support vector machine (SVM)

1. はじめに

人類は新規の化学物質を続々に合成し、我々の生活および健康の向上に役立ててきた。しかし一方で、化学物質の毒性が地球規模で深刻な問題をもたらしている。1962 年、Rachel Carson は有名な書物「Silent Spring」を出版し、その中で農薬を中心とする化学物質の毒性について人類に警告した。それにもかかわらず、有毒な化学物質に起因する様々な事故、事件が地球上のあらゆる地域で発生した。この原因は、大多数の化学物質がその毒性が未知なまま我々の周囲に存在しているからである。そのため、化学物質の毒性に関する情報の取得が喫緊の課題となっている。

ヒトに対する化学物質の毒性には、急性毒性、亜急性毒性、慢性毒性、生殖毒性、変異原性、催奇形性、発ガン性等、様々な種類があるが、中でも化学物質による発ガンが最も重大である。ガンは 1980 年以降、日本人の死因の第 1 位であり、近年では死亡者の 30 %(男 34%、女 26%) を占めている。ガンの発生には、物理的 (放射線、紫外線等)、化学的 (発ガン性化学物質)、生物的 (ウイルス、細菌、遺伝等) 等、様々な原因が上げられる。その内、飲食物や喫煙等により体内に取り込まれる発ガン性化学物質が最大原因であることが明らかになっている [Doll and Peto, 1981; Harvard Center for Cancer Prevention, 1996]。そのため、化学物質の発ガン性の情報を把握することが不可欠であるが、我々の周りには発ガン性不明の化学物質が氾濫しているのが現実である。環境中に存在する化学物質の種類は約十万種といわれているが、この内で発ガン性が判明している化学物質は数千種に過ぎず、99% 以上の化学物質は発ガン性が不明なまま我々の周囲に存在している。

発ガン性等の化学物質の毒性の評価には通常、動物を用いた試験が行われ、きわめて長い期間、莫大な費用と多数の動物が必要である。特に発ガン性試験は、ラットおよびマウス 100 匹以上に被験物質を 24 ケ月以上投与するため、費用、期間とも最大である。また、近年では動物愛護の観点から毒性試験が問題になっており、特に欧州では動物実験に対して厳しい法規制がとられている。したがって、発ガン性未知の膨大な数の化学物質の全て

について動物実験により発ガン性を評価することは現実には不可能である。

そこで、化学物質の発ガン性評価の動物実験に代わる手段として、コンピュータを用いる発ガン性の予測技術の確立が渇望されている。化学物質の毒性予測の原理は、類似の構造をもつ化学物質（同族体、congener）は類似の毒性を示すという定量的構造活性相関（Quantitative Structure-Activity Relationship, QSAR）[Hansch and Fujita, 1964] である。これを基に、類似の化学物質群について化学構造を反映する記述子（descriptor）を説明変数として毒性データの回帰式を決定すれば、毒性未知の化学物質の毒性を予測することが可能となる。化学物質の構造から毒性を予測することができれば、多数の動物を用いる毒性試験が不要になり、化学物質の安全管理の観点から大きな意義がある [Benfenati et al., 2009]。特に発ガン性は動物実験が最長の期間を要するので、QSAR による予測が最も効果的である。また、新規の化学物質を合成、製造する前にその毒性を事前に評価することも可能になり、新規化学物質の開発にとってもコスト削減に寄与するなど意義が高い。

このような観点から、化学物質の発ガン性を構造から予測する手法の研究開発が欧米では活発に行われている [Vracko, 2000; Passerini, 2003; Patlewicz et al., 2003; Benigni, 2004; Sun, 2004; Contrera et al., 2005; Crettaz and Benigni, 2005; Helguera et al., 2005; Benigni and Bossa, 2008: Guyton et al., 2009; Tan et al., 2009; Toropov et al., 2009a, 2009b; Venkatapathy et al., 2009]。芳香族アミン等の同族体については、発ガン性を比較的高い精度で予測できる場合もある [Braga et al., 1999; Vendrame et al., 1999; Benigni et al., 2000; Franke et al., 2001; Benigni et al., 2003b; Zhou et al., 2003]。しかし、我々の周囲に存在する発ガン性未知の化学物質の構造は多種多様であり、任意の構造の化学物質の発ガン性を十分な精度で予測できる手法は未だに存在しない [Benigni, 2003a; Benigni, 2005]。これまでに開発された多数の発ガン性予測手法の性能を共通のデータを用いて評価する公開テスト PTC（Predictive Toxicology Challenge）[Helma et al., 2000; Helma and Kramer, 2003] が実施されたが、予測精度の最高は 70% 程度であり、動物実験代替の予測手法としては満足できる性能でない。構造の多様な非同族体の化学物質群の発ガン性の予測手法の開発が最も重要な課題となっている。

非同族体の予測が困難な原因の 1 つは、発ガン性予測モデルを開発するために必要な信頼性の高いデータの不足である。動物実験による発ガン性データは幾つかの機関で収集・公開しているが、その中で Registry of Toxic Effects of Chemical Substances（RTECS）のデータベース（DB）は最大規模であり、現在、約 20 万件のデータが収録されている。しかし、この DB は様々な試験法で得られたデータが原論文から評価されずに収録されているため、データの信頼性の点で問題がある。一方、National Toxicology Program（NTP）、International Agency for Research on Cancer（IARC）、US Environmental Protection Agency（EPA）等の DB は、動物実験データの信頼性が評価され格付けされており、信頼性が高い。しかし、これらの DB に収録されている化学物質の数は数千種程度と少ない。また、データの信頼性を表す発ガン性の格付けに統一性がなく、混乱している。これらの理由から、過去数回行われた発ガン性予測の公開テストでも十分な数の化学物質が利用できなかった [Benigni and Bossa, 2008]。そのため、予測モデル開発やその性能検証に必要な発ガン性データを集積した大規模 DB の構築が求められている。

非同族体の予測が困難なもう 1 つの原因は、不特定の化学物質の発ガン性を構造情報のみから予測できるかという点である。Hansch-Fujita による QSAR は元来、毒性発現機構が類似する同族体を対象としたものであるが、現在までに発ガン機構が判明している化学物質は数十種程度にすぎない。しかも、化学物質の発ガン機構の解明は長期間の生化学的な研究を要し、機構が解明された化学物質の数が急激に増加することは考えにくい。さらに、「多段階発ガン説」[Nordling, 1953] に象徴されるように化学物質による発ガン機構はきわめて複雑であり、物質によって作用機構が異なること [Woo and Lai, 2003]、また、多くの発ガン性物質は単一の機構ではなく、複数の機構によりガンを引き起こすと考えられること等、化学物質の発ガン機構に関しては未解明の部分が多い。そのため、化学物質の発ガン性をその発現機構に基づいて予測する理論的研究はきわめて少ない [Benigni, 2010]。また、発ガン性について QSAR 解析が行われている同族体も数グループ、数十物質程度にすぎない。したがって、発ガン機構に基づいて広範囲の化学物質の発ガン性を予測できる手法を開発することは現状では不可能である。

このようなきわめて複雑な問題に対して、現在、有効と考えられるのは帰納的アプローチ、すなわち統計解析に基づく予測手法である。代表的な手法としては、多数の化学物質について構造を特徴づける記述子を説明変数とし、発ガン性データを目的変数として両者の関係を解析する重回帰分析がある。しかし、重回帰分析を行うためには目的変数と説明変数の間の関係式を予め仮定する必要があるが、化学物質の発ガン機構が殆ど解明されていない現在では、両変数の関係は全く不明であり、重回帰分析を用いた予測モデルでは高い精度は期待できない。

このような不明確な要素を含む相関関係に対して一つの対処策と考えられるのが人工ニューラルネットワーク（Artificial Neural Network, ANN）[Devillers, 1996a, b; Zupan and Gasteiger, 1999; Peterson, 2000; Ivanciuc, 2009a, b] である。ANN は重回帰分析と異なり、目的変数と説明変数の間の関係式を予め仮定する必要がなく、あらゆる相関関係の解析が可能である。発ガン性予測に ANN を適用した研究もあるが、その対象は同族体に限られている [Bahler et al., 2000; Basak et al., 2000; Hemmateenejad et al., 2005; Fjodorova et al., 2009]。しかも、ANN には局所解、過学習、計算時間等、多くの問題があることが指摘されている [Devillers, 1996b]。我々は ANN を用いて PTC の発ガン性データを解析したが、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 [Tanabe et al., 2005]。

そこで我々は、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（Support Vector Machine, SVM）[Chen et al., 2004a; Ivanciuc, 2007] を用いて非同族体の発ガン性予測の可能性を検討してきた。SVM は、ANN において深刻な局所解の問題がないことや、処理がきわめて高速なため大規模な問題にも簡単に実行できること等の利点がある。そのため、様々な問題を解決できると期待され、QSAR 分野でも多くの報告がある [Byvatov et al., 2003; Chen et al., 2004a; Helma et al., 2004; Xue et al., 2004; Yao et al., 2004; Jorissen and Gilson, 2005; Bhavani et al., 2006; Bruce et al., 2007; Doucet et al., 2007; Tang et al., 2007]。我々は SVM を用いて PTC のデータを解析し、非同族体の発ガン性予測に SVM が有効であることを実証した [Tanabe et al., 2008]。しかし、PTC のデータは物質数

が少ないため、広範囲の化学物質の発ガン性予測の有効性については明らかにできなかった。我々以外に SVM を発ガン性予測に適用した論文は幾つかあるが、どれも比較的少数の同族体に限られている [Ivanciuc, 2002; Luan et al., 2005; Bhavani et al., 2006; Massarelli et al, 2009; Tan et al., 2009]。そのため、大規模な非同族体の発ガン性データを用いて、広範囲の化学物質の発ガン性を高精度で予測する実用的なモデルの開発が求められている。

環境中に存在する十万種以上もの莫大な化学物質の発ガン性の情報を取得しなければならないという社会的要請に応えるためには、不特定の化学物質の発ガン性を高精度で予測できる手法を緊急に開発する必要がある。我々はこの社会的要請に応えるべく、多種多様な化学物質の発ガン性について信頼性の高いデータを集積した発ガン性 DB の構築、および発ガン性未知の化学物質についてその構造から発ガン性を高精度で予測するシステムの構築の 2 点を目的として研究してきた。本論文では我々のこれまでの研究成果を総合して報告する。

2. 発ガン性データベースの開発

我々が構築した発ガン性 DB は、データの信頼性が高いとされる IARC、European Union（EU）、EPA、NTP、American Conference of Governmental Industrial Hygienists（ACGIH）、Japan Society for Occupational Health（JSOH）の 6 種の DB [JETOC, 2007] から発ガン性の動物実験データを収集した。それぞれの DB では、Table 1 に示すように、化学物質の発ガン性は動物実験の信頼度により幾つかのランクに分類されているが、DB によりランクの数が異なっている。また、その信頼度に関する表現が異なり、表現の違いが理解しにくい。

さらに、同一の化学物質でも DB によって異なるランク付けがされている物質が多数存在する。例えば、IARC で発ガン陽性のランク 1、2A、2B、または EU で 1、2、3 と分類されているが、他の DB では陰性のランクに分類されている化学物質が 40 種以上もある。その逆に、IARC で発ガン陰性のランク 3 と分類されているが、他の DB で陽性ランクに分類されている化学物質が 70 種以上もある。代表例を Table 2 に示す。

このような各種発ガン性 DB における信頼性ランクの不統一を解決するために、種々の DB における信頼性を総合的に評価し、ランク付けの統一を試みた。そのためにまず、Table 1 に示す PRTR-MSDS の基準 [Urano, 2001] を採用し、各 DB のランクを I ～ V の 5 段階に格付けした。次に、Table 3 に示すように、各 DB のランクを総合的に評価して、A ～ F の 6 段階に統一的に格付けした。その際、同一の化学物質について異なるランク付けがなされている場合は、信頼性が最も高い IARC と EU を重視して格付けした。ただし、以上の DB だけでは発ガン陰性のデータが不足するので、NTP の DB（他の DB には陽性のデータのみが集積されているので）から動物試験で陰性の物質を収集し、ランク E に格付けした。以上の手順により発ガン性 DB に収録できた総物質数は 1,512 種である。Table 3 には各ランク別に身近に存在する化学物質の例を幾つか示した。この内容を見ると、我々の周囲に存在する日用品の中にも発ガン陽性の化学物質を用いたものがあることが分かる。

Table 1. Reliability ranks and their explanations in various carcinogenicity databases

UR	IARC		EU		EPA		NTP		ACGIH		JSOH	
I	1:	Carcinogenic to humans	1:	Known as a human carcinogen	A:	Carcinogenic to humans confirmedly	K:	Known as a human carcinogen	A1:	Carcinogenic to humans confirmedly	1:	Carcinogenic to humans
II	2A:	Probably carcinogenic to humans	2:	Should be regarded as if a human carcinogen	B1:	Probably carcinogenic to humans	R:	Reasonably anticipated as a human carcinogen	A2:	Carcinogenic to humans suspectedly	2A:	Probably carcinogenic to humans
	2B:	Possibly carcinogenic to humans			B2:	Carcinogenic to animals, but unknown to humans			A3:	Carcinogenic to animals, but unknown to humans	2B:	Possibly carcinogenic to humans
III			3:	Possibly carcinogenic to humans	C:	Possibly carcinogenic to humans						
IV	3:	Not classifiable as a human carcinogen			D:	Not classifiable as a human carcinogen			A4:	Not classifiable as a human carcinogen		
V	4:	Probably not carcinogenic to humans			E:	Not carcinogenic to humans confirmedly			A5:	Not suspected as a human carcinogen		

UR : Unified rank.

Table 2. Example chemicals assigned to different carcinogenicity ranks in different databases

Chemical name	IARC	EU	EPA	NTP	ACGIH	JSOH
1,3-Butadiene	2A	1	B2	K	A2	1
N,N-Dimethylaniline	3	3			A4	2B
Formaldehyde	1	3	B1	R	A2	2A
Methyl methacrylate	3		E		A4	
Naphthalene	2B	3	D		A4	
Trichloroethylene	2A	2		R	A5	2B

Table 3. Chemicals and their carcinogenicity ranks accumulated in the carcinogenicity database

Carcinogenicity	Rank	Criteria	NC	Example chemical
Positive	A	I in Any of DBs	167	Alcohol, Arsenic Inorganic Compounds, Asbestos, Benzene, Benzidine, Formaldehyde, 2-Naphthylamine, Tar, 2,3,7,8-Tetrachlorodibenzo-p-dioxin (TCDD)
	B	II in IARC or EU, or II in Two DBs	407	Acetaldehyde, Acrylamide, Benzo(a)pyrene, Chloroform, DDT and Associated Compounds, p-Dichlorobenzene, Di(2-ethylhexyl) Phthalate, Gasoline, Naphthalene, Polychlorinated Biphenyls (PCB), Tetrachloroethylene, Trichloroethylene
	C	II in Only One DB, or III in Any DBs	186	Cresol, Polycyclic Aromatic Hydrocarbons (PAH)
Negative	D	IV in Any DBs	631	Acetone, Caffeine, Camphor, CFC-11, Ethanol, Ethylene, Ethylene Glycol, Ozone, Phenol, Poly(vinyl alcohol), Toluene, Xylene
	E	At Least One (-) in NTP	117	Bisphenol A, Dibenzo-p-dioxin
	F	V in Any DBs	4	ε-Caprolactam
Total			1512	

NC : Number of chemicals

3. 化学構造からの発ガン性予測システムの開発

3.1 方 法

3.1.1 発ガン性データ

上記の発ガン性 DB には純有機物以外の様々な物質（金属、金属塩、混合物等）も収録されているが、これらについては構造が不明で、記述子が計算できないため、QSAR による解析を行うことができない。そこで、QSAR に不適な物質として、

① H, C, N, O, F, Si, P, S, Cl, Br および I 以外の原子を含むもの、
② キシレン異性体等の混合物、
③ ポリビニルアルコール等の高分子、
④ アスベスト等の化学構造が特定できないもの、

等は除いて、Table 4 に示す構造が明確な有機化学物質 911 種を抽出し、予測モデルの構築に用いた。

PTC のデータは発ガン陰性の化学物質が過多であったため、コンテスト参加者の殆どのモデルでは陰性物質は高精度で予測できたが、陽性物質は殆ど予測できなかった。これに対して、本研究のデータは物質数が発ガン陽性 409、陰性 502 とバランスがとれているので、これらのデータを用いて構築したモデルは陽性・陰性の化学物質の発ガン性をバランスよく予測すると期待できる。また、各ランクの物質は炭素数および分子量が広い範囲に及んでおり、多種多様な化学物質の発ガン性を予測できると期待される。

Table 4. Organic chemicals used for constructing prediction models

			Number of C atoms			Molecular weight			Target	Weight
Carcinogenicity	Rank	NC	Min	Max	Mean	Min	Max	Mean	value	value
Positive	A	29	1	26	10.0	30.0	371.6	206.4	1.0	1.00
	B	264	0	32	9.5	32.1	840.9	206.6	1.0	0.50
	C	116	1	33	8.7	46.1	959.1	215.5	1.0	0.25
	Total	409	0	33	9.3	30.0	959.1	209.1		
Negative	D	396	0	42	10.1	28.1	788.7	210.7	-1.0	0.25
	E	102	1	39	10.9	56.1	644.9	237.8	-1.0	0.50
	F	4	5	10	6.5	100.1	272.8	181.6	-1.0	1.00
	Total	502	0	42	10.2	28.1	788.7	215.9		
Total		911	0	42	9.8	28.1	959.1	212.9		

NC : Number of chemicals

3.1.2 記述子データ

以上の化学物質について、Corina プログラム [Gasteiger et al, 1996; Oellien et al., 2000] を用いて平面構造から立体構造を生成し、構造を最適化した。次に、記述子作成プログラム Dragon 5.4 [Todeschini and Consonni, 2006] を用いて、Table 5 に示す 1,504 種の記述子を作成し、QSAR 解析に用いた。ただし、これらの記述子は物質数に比べて明らかに過多である。一般に、学習モデルの説明変数として、予測に有効な変数の他に有効でない変数も加えると、学習時の誤差は減少するが、予測時の誤差は逆に増大し、いわゆる過学習状

態に陥る。そのため、予測に有効かつ不可欠な説明変数をスクリーニングする必要がある。

統計解析における変数選択には、段階的増減選択法（stepwise forward or backward selection）、焼きなまし法（simulated annealing）、モンテカルロ法、遺伝的（進化的）アルゴリズム（genetic or evolutionary algorithm）、粒子群最適化法（particle swarm optimization）、蟻コロニー最適化法（artificial ant colony system）等、様々な方法が提案されている [Barak et al, 2009]。しかし、本研究の予測モデルの最適化では、記述子選択の操作をきわめて多数回行わねばならないため、長い計算時間を要するこれらの方法は利用できない。そこで、変数選択法としては若干厳密性に欠けるが、迅速性を優先して、発ガン性データと各記述子との単相関係数を計算し、相関の高い記述子から採用する個数を変えながら予測精度を調べて最適数を探索する方法を用いることにした。

Table 5. Types, numbers, and examples of Dragon descriptors used

Type of descriptors	ND	Example
Constitutional descriptors	46	MW(molecular weight)
Topological descriptors	105	BAC(Balaban centric index), W(Wiener W index)
Walk and path counts	44	CID(Randic ID number), TPC(total path count), TWC(total walk count)
Connectivity indices	32	X0(connectivity index chi-0), X1(Randic connectivity index)
Information indices	27	IAC(total information index of atomic composition), Uindex(Balaban U index)
2D autocorrelations	96	MATS1e(Moran autocorrelation-lag 1/weighted by atomic Sanderson electronegativities)
Edge adjacency indices	105	EPS0(edge connectivity index of order 0)
Burden eigenvalues	64	BEHm1(highest eigenvalue n. 1 of Burden matrix / weighted by atomic masses)
Topological charge indices	21	GGI1(topological charge index of order 1), JGT(global topological charge index)
Eigenvalue-based indices	43	VED1(eigenvector coefficient sum from distance matrix)
Randic molecular profiles	41	DP01(molecular profile no. 01), SP01(shape profile no. 01)
Geometrical descriptors	62	AROM(aromaticity index), J3D(3D-Balaban index), W3D(3D-Wiener index)
RDF descriptors	150	RDF010u(Radial Distribution Function - 1.0 / unweighted)
3D-Morse descriptors	160	Mor01u(3D-Morse - signal 01 / unweighted)
WHIM descriptors	99	L1u(1st component size directional WHIM index / unweighted)
Gateway descriptors	197	ITH(total information content on the leverage equality)
Functional group counts	101	nArCO(number of ketones (aromatic)), nCar(number of aromatic C(sp2))
Atom-centred fragments	85	C-024(R–CH–R), Cl-086(Cl attached to C1(sp3))
Molecular properties	26	ALOGP(Ghose-Crippen octanol-water partition coeff.)
Total	1504	

ND : Number of descriptors

3.1.3　SVM によるモデル化

発ガン性データと記述子データとの相関を解析する SVM のソフトウェアは LIBSVM ver.2.89 [Chang and Lin, 2009a] を用いた。発ガン性データが陽性・陰性の 2 群のため、LIBSVM の 2 群分類を行う SVC（support vector classification）機能を用いた。その際、発ガン性データには信頼性のランクが付与されているので、SVC での発ガン性データの目標値と重みを Table 4 に示すように設定し、重み付きデータに対する 2 群分類機能 [Chang and Lin, 2009b] を用いた。

SVM は ANN と同じく非線形解析手法であり、SVM ではカーネルと呼ぶ非線形写像関数を用いて線形分割できるように変換することで、ANN と比較して飛躍的な高速処理が可能になる。ANN に対する SVM の最大の利点は、局所解問題を回避できることである。

一般に、データを 2 群に分類する問題の場合、分割線は無数に引くことができ、ANN において局所解が無数に存在することはこの分割線に対応する。一方、SVM では、2 群のデータの丁度中間を通るように分割線を決めるため、それが唯一の解となり、局所解問題は発生しない。このように SVM では解は一義的に決まるが、ANN と同様、過学習の問題がある。したがって、モデルの最適化が必要であり、LIBSVM では Table 6 に示すように最適化すべきパラメータが多数ある。中でも g(gamma) と c(cost) の設定が重要であり、これらと記述子数の計 3 個のパラメータについては最適設定が不可欠である。

そこで、以下の Dual Cross-Validation Test により、SVM の学習・テストと最適化を同時に行った。すなわち、

① 解析対象の化学物質を 10 群に分割する、
② その内の 9 群を学習用とし、この群について Leave-One-Out を用いてパラメータ g(gamma) と c(cost) および記述子数を最適化する、
③ その最適モデルを用いて、テスト用物質の発ガン性を予測する、
④ 以上の手順を学習用とテスト用物質を入れ替えながら 10 回繰り返し、全ての物質について発ガン性を予測する、

という手順である。

2 群分類モデルの性能評価には様々な指標が用いられるが、ここでは次式で計算される総合正解率（Overall Accuracy; OA）を用いた。

$$OA = \frac{TP + TN}{TP + TN + FP + FN}$$

ここで TP は true positive、TN は true negative、FP は false positive、FN は false negative の物質数である。

Table 6. Parameters to be adjusted in SVM

Symbol		Meaning	Default
d	degree	set degree in kernel function	3
g	gamma	set gamma in kernel function	1/k
r	coef0	set coef0 in kernel function	0
c	cost	set the parameter C of C-SVC, epsilon-SVR, and nu-SVR	1
n	nu	set the parameter nu of nu-SVC, one-class SVM, and nu-SVR	0.5
p	epsilon	set the epsilon in loss function of epsilon-SVR	0.1
m	cachesize	set cache memory size in MB	100
e	epsilon	set tolerance of termination criterion	0.001
h	shrinking	whether to use the shrinking heuristics, 0 or 1	1
b	probability_estimates	whether to train a SVC or SVR model for probability estimates, 0 or 1	0
wi	weight	set the parameter C of class i to weight*C, for C-SVC	1

k in the g option : the number of attributes in the input data.

3.2 各種モデルの予測結果

3.2.1 単一の SVM による一括予測

まず、単一の SVM を用いて全物質 911 種を一括して解析するモデルを検討した。変数選択の例として、発ガン性データとの相関の高い記述子の数を変えながら正解率を調べた

結果を Fig. 1 に示す。記述子の数を 50 個から増していくと正解率は向上するが、250 個以上に記述子が増えると正解率は減少し、過学習状態に陥る。250 個の記述子を用いた時に正解率の最高値 68.8% が得られたが、その時の TP、FP、TN、FN の内訳を Fig. 2 のように示す。

この正解率は PTC テストでの最高値と同程度であり、非同族体を一括して解析するモデルの最高精度はこの程度であると思われる。しかし、68.8% という予測精度は動物実験代替の予測手法としては満足できる性能ではない。しかも、発ガン陽性の物質を陰性と誤判定する FN の比率が 37% と高いことは致命的であり、もっと高性能の予測手法を開発する必要がある。

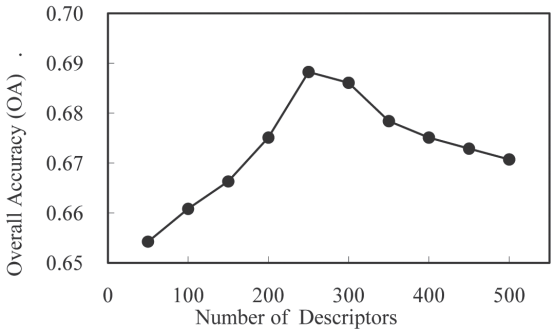

Fig. 1　The dependence of the overall accuracy on the number of descriptors in the batch model

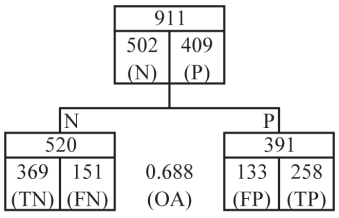

Fig. 2　Result of the batch model for overall chemicals

3.2.2　SVM とアンサンブル学習（ブースティング）の組み合わせによる予測

上記の一括モデルの成績は発ガン性予測手法としては採用できないレベルであるが、化学物質の発ガン機構の複雑さを考慮すると、単一の SVM ではこれ以上の精度の向上は期待できない。そこで次に、多数の SVM を組み合わせたアンサンブル学習を検討した。この方法は機械学習の分野で進展しており [Ivanciuc, 2009]、QSAR の分野でも既に適用例がある [Svetnik et al., 2005; Fukunishi et al., 2008; Langham and Jain, 2008; Liu et al., 2008]。アンサンブル学習とは、予測精度が低い弱学習器 (weak learner) をランダムに複数構築し、それらを組み合わせて高精度の予測モデルを構成する方法である。

アンサンブル学習にはブースティングやバギング等、幾つかの手法が提案されているが、まず、ブースティングの中で最もよく知られている AdaBoost [Freund and Schapire, 1997] を検討した。この方法は、誤分類率に応じて（adaptive）重みを変えるブースティングであり、以下のアルゴリズムで学習を行う。

① N 個の全データに最初は均等な重み 1/N を割り当てる。

② 全データを用いて 1 台目の学習器 h_1 を学習し、不正解率 ε_1 から次式により信頼度 β_1 を求める。

$$\beta_i = \frac{1}{2}\ln\left(\frac{1-\varepsilon_i}{\varepsilon_i}\right)$$

③ 1台目の学習器が正解したデータは重みを exp(-β_1) 倍し、不正解のデータは重みを exp(β_1) 倍する。
④ 2台目以降の学習器 h_i について同様の重み付きの学習を繰り返す。
⑤ 以上の方法で M 台の学習器を作り、信頼度付き多数決で判別器としての正解率を計算する。

$$y = \frac{1}{M}\sum_{i=1}^{M}\beta_i h_i(x)$$

以上のアルゴリズムに従って SVM を用いて 10 台の学習器を作成し、発ガン性データの学習とテストを行った結果を Table 7 に示す。予想に反して、学習器が増えるほどその正解率が徐々に低下すると共に、全体の総合正解率も徐々に低下した。この原因は、AdaBoost では上の③に示すように、正解データの重みを減少させ、不正解データの重みを増加させることで全体正解率の向上を企図する。しかし、一括モデルの結果が示すように、単一の SVM で全データを高い正解率で予測することは困難であり、不正解データはいくら重みを増やしても不正解のままだからである。

Table 7. Result of the combination model of SVM and AdaBoost

Cycle No.	ε	TP	FP	TN	FN	OA
1	0.344	259	163	339	150	0.656
2	0.351	248	155	347	161	0.653
3	0.361	239	150	352	170	0.648
4	0.363	246	159	343	163	0.646
5	0.367	220	136	366	189	0.643
6	0.367	222	139	363	187	0.642
7	0.370	233	152	350	176	0.640
8	0.368	222	141	361	187	0.639
9	0.373	209	130	372	200	0.638
10	0.378	205	127	375	204	0.637

ε : Error rate

3.2.3 SVM とアンサンブル学習（バギング）の組み合わせによる予測

AdaBoost において用いられる学習器は全データを単一の SVM で学習するため、発ガン性データの場合にはそれらを多数組み合わせても全体の正解率は向上しない。一方、アンサンブル学習の中でブースティングと並んで知られているのがバギング [Breiman, 1994] である。この方法では、N 個のデータから重複を許してデータ K 個をサンプリングして数台の学習器を作り、AdaBoost と同様、多数決で正解率を計算する。この方法は比較実験ではブースティングに対して劣ることが多いとされている。しかし、発ガン性予測では、データ全体でなく部分集合を学習するという利点を生かせば、AdaBoost より高成績が期待できる。そこで、このバギングの修正法として、全データを幾つかのグループに分け、グループごとに学習する方法を検討した。

Fig. 3 に示すように、まず全データを SVM で学習・予測しながら、発ガン性の実測値を用いて陽性と陰性の誤判定（FP、FN）の物質を枝刈り（pruning）する。この操作を繰

り返すと、3 回目の SVM(G1) で陽性 211 物質、陰性 321 物質、計 532 物質の第 1 群が生成され、このグループでは 94.9 % という高い正解率が得られた。次に、ここまでに枝刈りされた残りの 379 物質について同様の操作を行うと、2 回目 (G2) で陽性 187 物質、陰性 7 物質、計 194 物質の第 2 群が生成され、このグループでは 98.5 % という高い正解率が得られた。残りの 185 物質は第 3 群を形成し、このグループでは 97.8 % の正解率となった。

このようにして、911 種の全物質が 532、194、185 物質の 3 群に分けられ、それぞれ 90% 以上の高率で予測できることは分かったが、全物質をこれら 3 群に振り分ける方法がない。そこで、全物質についてこれら 3 個の SVM で予測した結果を親 SVM に入力して発ガン性を予測するモデルを構築した。その結果は Fig. 3 の左下に示すように、TP が 268、FP が 141、TN が 361、FN が 141 で、全体の正解率は 69.0 % となったが、この成績は以上の方法と同程度であり、満足できる正解率ではない。

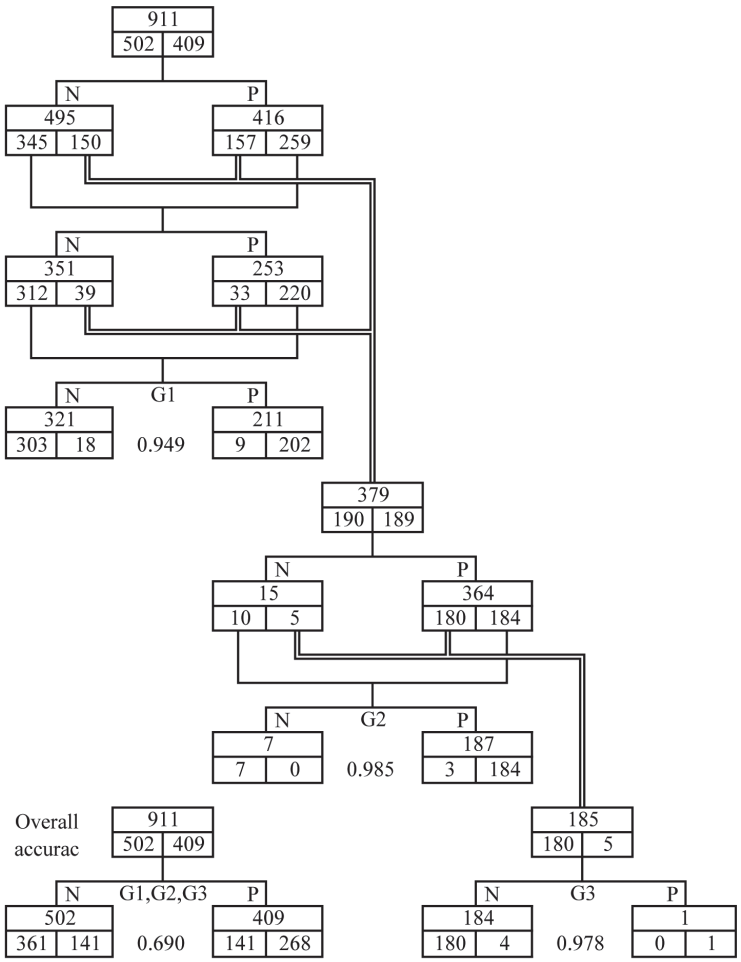

Fig. 3　Result of the combination model of SVM and ensemble learning (revised bagging)

3.2.4 SVM と決定木の組み合わせによる予測

以上の結果から、AdaBoost のように全データを単一の SVM で学習するモデルでは高い正解率は期待できないが、バギングのように全データを幾つかのグループに分け、グループごとに個別の SVM で学習するモデルの方が高い正解率が期待できる。そこで、一括モデルによる予測結果について、陽性・陰性に判定されたグループをさらに個別に SVM で学習し、その操作を続けて多数の SVM を決定木状に組み合わせるモデルを検討した。

その結果を Fig. 4 に示す。最下段まで伸びず途中で止まった枝は、レコード数が少なくなり、SVM による解析が不可能になった場合である。この方法による最終結果は、最下段だけでなく途中で止まった枝も含めて、TP が 297、FP が 174、TN が 328、FN が 112 で、総合正解率は 68.6 % となった。これまでの方法と比較して FN の比率は低下したが、正解率は同程度であり、やはり満足できる結果ではない。

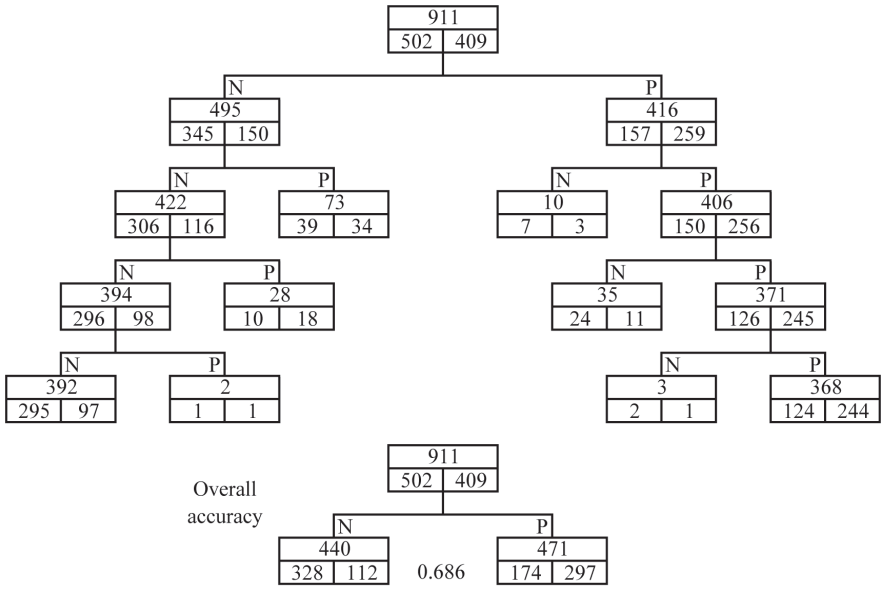

Fig. 4 Result of the combination model of SVM and decision tree

3.2.5 敗者復活を取り入れた決定木と SVM の組み合わせによる予測

本研究で解析している広範囲の化学物質の発ガン性データは陽性・陰性の 2 群の分離度がきわめて悪く、両群のデータがかなり重なっている。このような場合には、少数のデータの追加や削除等の変更でも、陽性・陰性の境界付近にある物質の予測値が影響を受け、正解率が大きく変化する可能性がある。しかし、Fig. 4 のような決定木では、一旦、陽性あるいは陰性と判定された物質は、それ以後、修正される余地がない。

そこで次に、このような状況に柔軟に対処して予測モデルを頑健なものとするために、誤判定（FP や FN）されたグループを再解析にかけ、いわば敗者復活を取り入れた SVM を決定木状に組み合わせたモデルを検討した。すなわち、Fig. 4 の 2 段目の SVM で陽性・

陰性と判定されたグループ同士をまとめて再解析する方法である。その結果、Fig. 5 に示すように、再解析を 4 回繰り返すと、それ以上の再解析ができない状態になった。このモデルによる最終結果は、TP が 311、FP が 165、TN が 337、FN が 98 となり、FN の比率がかなり低下し、総合正解率も 72.0 % となった。これまでの方法の中では最高の予測成績が得られたものの、動物実験代替の発ガン性予測手法としてはやはりまだ満足できる性能ではない。

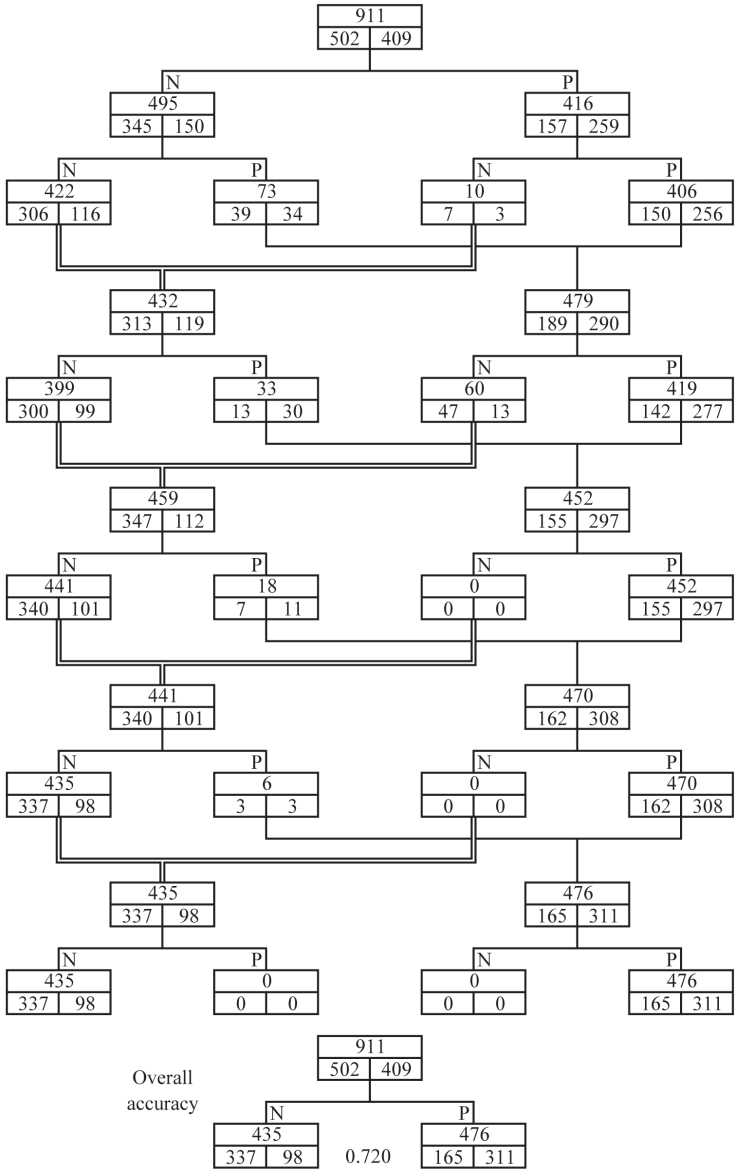

Fig. 5　Result of the combination model of SVM and repêchage decision tree

3.2.6 同族体に対する SVM の直列組み合わせによる予測

これまでの 5 種類のモデルでは満足できる結果が得られなかったので、より高性能の予測モデルを開発するために、これらのモデルの低成績の原因を考察した。アンサンブル学習（バギング）や決定木のようにデータ全体を分割して解析しても、単一の SVM で解析する一括モデルと大差ない結果が得られたことは、これまでのモデルにおける機械的なデータ分割法に改良の余地があると考えられる。すなわち、これらのモデルのようにデータを機械的に分割するのではなく、解析対象の化学物質群を発ガン機構に関連させて分類し、個別に解析することにより予測精度の向上が期待される。

事実、発ガン機構と化学物質の部分構造との間に密接な関連があることが明らかになっている。元来、共通の部分構造を有する同族体を対象とする Hansch-Fujita の QSAR が芳香族アミン等、一部の同族体で高精度の発ガン性予測式を導出できるのはそのためである。したがって、発ガン機構に基づいて全物質を分類し、個別に解析できれば高精度の予測モデルを構築できる可能性がある。しかし、これまでに発ガン機構が判明している化学物質は僅少であるため、それらの部分構造に基づいて今回の解析対象の全化学物質を分類することは不可能である。

そこで、同じ部分構造を含む同族体では発ガン機構が同じであると仮定し、発ガン性を高精度で予測できるような同族体を探索し、それぞれ個別の SVM で予測した結果を組み合わせて全物質の発ガン性を予測するモデルを検討した。そのためにまず、Dragon 記述子の内で各種の部分構造を有する化学物質の個数をカウントする Functional Group Counts を用いて、種々の同族体の物質数を集計した。その結果を Table 8 に示す。一般に、化学物質の部分構造は共存可能（非排他的）なので、多くの物質は幾つもの同族体に属している。Table 8 における各同族体の物質数の合計 1,490 は今回の解析対象物質数 911 の 1.6 倍である。このことは、各化学物質は平均して 1.6 個の部分構造を有することを示している。ただし、Table 8 の芳香族炭化水素とハロ炭化水素の 2 群のみは排他的に設定したが、前者はベンゼン、ベンゾピレン等、後者は塩化ビニル、テトラクロロエチレン等のいずれも発ガン性があるとされる物質群に対して、予測モデルの精度が高くなることを企図したためである。

次に、Table 8 に示した部分構造の中から発ガン性を高精度で予測できるような同族体を選定した。その際、同族体の選定条件として、正解率と頑健性の 2 点を設定した。一般にどのような予測モデルでも、データ数が少ないとモデルの正解率は向上するが、頑健性（robustness）が低下し、逆にデータ数が多いとモデルの頑健性は向上するが、精度が低下する。そこで、正解率 80% 以上と物質数 50 ～ 150 程度を目標に同族体の候補を探索した。正解率の目標を 80% とした理由は、非同族体に対する既存の発ガン性予測モデルの最高精度が 70% 程度であるからである。

このようにして Table 9 に示す 23 種の同族体を選定した。次の問題は、これらの同族体をどのように組み合わせるかである。化学物質の部分構造は非排他的なので、同族体を直列（樹状）に組み合わせるか、並列に組み合わせるかで、予測モデルの最終的な正解率が異なってくると考えられる。そこでまず、直列組み合わせを検討した。すなわち、Fig. 6 に示すように、まず Step 1 において、選定した 23 種の同族体ごとに個別に SVM 解析

を行い、その中から正解率最高の ArHC を選出したところ、これを含む同族体 54 物質について 87.0% の正解率を得た。次に Step 2 において、残りの 857 物質について同じ操作を繰り返し、Ketone を含む同族体 58 物質について正解率 86.2% を得た。以上の操作を全物質がなくなるまで繰り返した。その結果、Fig. 6 に示すように、17 段の直列分岐を繰り返すことで、解析対象の全化学物質 911 種がどれかの SVM で予測できるモデルを構築できた。最終段の Others も含む 18 種の同族体全体の予測結果は図の右下に示すように、TP = 328、FP = 95、TN = 407、FN = 81 となり、80.7% の総合正解率が得られた。

この 80.7% という正解率はこれまでの 5 種のモデルの成績より大幅に向上しており、このモデルで採用した同族体ごとに個別に解析する方式が有効であることを示している。また、過去に開発された発ガン性予測モデルの性能（約 70%）をはるかに凌駕するものであり、さらに誤判定 FN の比率もきわめて低いので、動物実験代替の発ガン性予測手法としては満足できる性能である。

しかし、このモデルには頑健性の点で問題がある。すなわち、Fig. 6 に示す 18 種の同族体の内で、Step 8 の Amide の 28 物質、Step 10 の AliAmine の 26 物質等のように物質数の少ないものが幾つか存在する。これらの同族体の物質数は Table 8 に示す物質数（Amide は 47 物質、AliAmine は 61 物質）と比べて大幅に少ないが、これはこれらの同族体はそれ以前の段階で分岐された物質が多いためである。これらの同族体の物質数は、上記の部分構造選定基準の内、同族体の物質数 50 ～ 150 程度には該当していない（物質数 50 以下の同族体は他にも多数存在）が、モデル全体の予測性能向上のために正解率 80% 以上の基準を優先したので、選出せざるを得なかった。このような物質数が少ない同族体は若干のデータの追加削除により正解率が大きく変動し、モデル全体の頑健性を低下させる可能性がある。

Table 8. Atom and functional group count descriptors, numbers of chemicals containing those atoms or functional groups, and statistical significance for positive/negative ratios

Atom and Functional Group Count Descriptors	NT	NP	NN	SS	Explanation
n135-Triazines	6	3	3		number of 1,3,5-triazines
nAB	530	226	304		number of aromatic bonds
nArC=N	8	2	6		number of imines (aromatic)
nArCHO	2	1	1		number of aldehydes (aromatic)
nArCL	90	39	51		number of chlorine atoms on aromatic ring
nArCO	27	11	16		number of ketones (aromatic)
nArCONH2	4	1	3		number of primary amides (aromatic)
nArCONHR	7	3	4		number of secondary amides (aromatic)
nArCOOH	10	0	10	−−	number of carboxylic acids (aromatic)
nArCOOR	17	5	12		number of esters (aromatic)
nArNCO	4	2	2		number of isocyanates (aromatic)
nArNH2	99	48	51		number of primary amines (aromatic)
nArNHR	16	5	11		number of secondary amines (aromatic)
nArNNOx	2	1	1		number of N-nitroso groups (aromatic)
nArNO	2	0	2		number of nitroso groups (aromatic)
nArNO2	86	40	46		number of nitro groups (aromatic)
nArNR2	24	11	13		number of tertiary amines (aromatic)
nArOCON	8	3	5		number of (thio-) carbamates (aromatic)
nArOH	66	25	41		number of aromatic hydroxyls (aromatic)
nArOR	80	40	40		number of ethers (aromatic)
nArX	98	43	55		number of X on aromatic ring
nAT	911	409	502		number of atoms
nAziridines	12	4	8		number of aziridines

Table 8. (continued) Atom and functional group count descriptors, numbers of chemicals containing those atoms or functional groups, and statistical significance for positive/negative ratios

Atom and Functional Group Count Descriptors	NT	NP	NN	SS	Explanation
nBM	798	354	444		number of multiple bonds
nBnz	476	202	274		number of benzene-like rings
nBO	911	409	502		number of non-H bonds
nBR	29	14	15		number of bromine atoms
nBT	911	409	502		number of bonds
nC	908	408	500		number of carbon atoms
nC(=N)N2	6	1	5		number of guanidine derivatives
nC=N-N<	11	5	6		number of hydrazones
nCar	530	226	304		number of aromatic C(sp2)
nCb-	475	201	274		number of substituted benzene C(sp2)
nCbH	467	196	271		number of unsubstituted benzene C(sp2)
nCconj	190	70	120	− −	number of non-aromatic conjugated C(sp2)
nCconjX	13	5	8		number of X on exo-conjugated C
nCH2RX	56	38	18	+ +	number of CH2RX
nCHR2X	9	5	4		number of CHR2X
nCHRX2	15	9	6		number of CHRX2
nCIC	665	291	374		number of rings
nCIR	665	291	374		number of circuits
nCL	231	121	110	+ +	number of chlorine atoms
nCONN	47	24	23		number of urea (-thio) derivatives
nCp	466	201	265		number of terminal primary C(sp3)
nCq	44	16	28		number of total quaternary C(sp3)
nCR2X2	2	1	1		number of CR2X2
nCR3X	2	0	2		number of CR3X
nCrq	30	12	18		number of ring quaternary C(sp3)
nCrs	145	75	70		number of ring secondary C(sp3)
nCrt	76	37	39		number of ring tertiary C(sp3)
nCRX3	27	12	15		number of CRX3
nCs	283	132	151		number of total secondary C(sp3)
nCt	103	46	57		number of total tertiary C(sp3)
nCXr	16	12	4	+ +	number of X on ring C(sp3)
nCXr=	14	7	7		number of X on ring C(sp2)
nDB	577	249	328		number of double bonds
nF	22	7	15		number of fluorine atoms
nFuranes	24	11	13		number of furanes
nH	886	396	490		number of hydrogen atoms
nHAcc	764	333	431		number of acceptor atoms for H-bonds (N,O,F)
nHBonds	104	38	66		number of intramolecular H-bonds (with N,O,F)
nHDon	408	171	237		number of donor atoms for H-bonds (N and O)
nI	3	1	2		number of iodine atoms
nImidazoles	20	10	10		number of imidazoles
nIsoxazoles	2	0	2		number of isoxazoles
nN	491	227	264		number of nitrogen atoms
nN(CO)2	16	8	8		number of imides (-thio)
nN+	111	53	58		number of positively charged N
nN=C-N<	4	2	2		number of amidine derivatives
nN=N	22	9	13		number of N azo-derivatives
nN-N	19	10	9		number of N hydrazines
nNq	4	1	3		number of quaternary N
nO	610	256	354		number of oxygen atoms
nO(C=O)2	4	0	4		number of anhydrides (-thio)
nOHp	42	13	29		number of primary alcohols
nOHs	34	12	22		number of secondary alcohols
nOHt	22	9	13		number of tertiary alcohols
nOxiranes	29	17	12		number of oxiranes
nOxolanes	6	1	5		number of oxolanes
nP	55	11	44	− −	number of phosphorous atoms
nP(=O)O2R	7	0	7	− −	number of phosphonates (thio-)
nPO4	36	7	29	− −	number of phosphates/thiophosphates
nPyrazines	6	2	4		number of pyrazines
nPyridines	32	17	15		number of pyridines
nPyrimidines	7	1	6		number of pyrimidines
nPyrroles	8	5	3		number of pyrroles
nPyrrolidines	10	5	5		number of pyrrolidines

Table 8. (continued) Atom and functional group count descriptors, numbers of chemicals containing those atoms or functional groups, and statistical significance for positive/negative ratios

Atom and Functional Group Count Descriptors	NT	NP	NN	SS	Explanation
nR#C-	5	5	0	++	number of non-terminal C(sp)
nR#CH/X	5	5	0	++	number of terminal C(sp)
nR=CHX	10	7	3		number of R=CHX
nR=Cp	59	34	25	++	number of terminal primary C(sp2)
nR=CRX	7	5	2		number of R=CRX
nR=Cs	133	55	78		number of aliphatic secondary C(sp2)
nR=Ct	58	21	37		number of aliphatic tertiary C(sp2)
nR=CX2	10	7	3		number of R=CX2
nR03	49	21	28		number of 3-membered rings
nR04	7	4	3		number of 4-membered rings
nR05	173	81	92		number of 5-membered rings
nR06	592	256	336		number of 6-membered rings
nR07	21	6	15		number of 7-membered rings
nR08	21	12	9		number of 8-membered rings
nR09	104	52	52		number of 9-membered rings
nR10	172	74	98		number of 10-membered rings
nR11	27	9	18		number of 11-membered rings
nR12	31	14	17		number of 12-membered rings
nRC=N	1	0	1		number of imines (aliphatic)
nRCHO	15	8	7		number of aldehydes (aliphatic)
nRCL	149	86	63	++	number of chlorine atoms (aliphatic)
nRCN	12	2	10	−−	number of nitriles (aliphatic)
nRCNO	3	1	2		number of oximes (aliphatic)
nRCO	36	16	20		number of ketones (aliphatic)
nRCONH2	5	3	2		number of primary amides (aliphatic)
nRCONHR	18	7	11		number of secondary amides (aliphatic)
nRCONR2	14	4	10		number of tertiary amides (aliphatic)
nRCOOH	36	14	22		number of carboxylic acids (aliphatic)
nRCOOR	69	22	47	−−	number of esters (aliphatic)
nRNH2	18	6	12		number of primary amines (aliphatic)
nRNHO	2	1	1		number of hydroxylamines (aliphatic)
nRNHR	16	4	12		number of secondary amines (aliphatic)
nRNNOx	28	22	6	++	number of N-nitroso groups (aliphatic)
nRNO2	6	3	3		number of nitro groups (aliphatic)
nRNR2	31	12	19		number of tertiary amines (aliphatic)
nROCON	16	9	7		number of (thio-) carbamates (aliphatic)
nROH	134	44	90	−−	number of hydroxyl groups (aliphatic)
nROR	80	42	38		number of ethers (aliphatic)
nRSR	18	3	15	−−	number of sulfides
nRSSR	2	0	2		number of disulfides
nRX	172	100	72	++	number of halogen atoms (aliphatic)
nS	110	41	69		number of sulfur atoms
nS(=O)2	4	1	3		number of sulfones
nSK	911	409	502		number of non-H atoms
nSO	3	0	3		number of sulfoxides
nSO2	3	2	1		number of sulfites (thio-/dithio-)
nSO2N	16	5	11		number of sulfonamides (thio-/dithio-)
nSO3	8	7	1	++	number of sulfonates (thio-/dithio-)
nSO4	3	3	0		number of sulfates (thio-/dithio-)
nTB	19	8	11		number of triple bonds
nThiazoles	6	3	3		number of thiazoles
nTriazoles	4	2	2		number of triazoles
nX	261	138	123	++	number of halogen atoms
ArHC	54	18	36		
XHC	85	55	30	++	

NT : Number of chemicals. NP : Number of positives. NN : Number of negatives.
SS : Statistic significance for P/N ratio. ++ : positive rich, −− : negative rich.
According to the statistics theory, if the ratio of positives in a group is greater than p0+p1, the group is judged as significantly positive rich at the significance level of 0.05 as compared with the whole ensemble, where p0 is the ratio of positives in the whole ensemble, given in this case by p0=409/911=0.449, and p1 is given by p1=1.96*[(409/911)*(502/911)/n]1/2 =0.975/n1/2 where n is the size of the group.
ArHC : Aromatic hydrocarbons counted as nCar>0 and nN=nO=nP=nS=nX=0.
XHC : Halohydrocarbons counted as nX>0 and nN=nO=nP=nS=0.
Note that many chemicals belonging to groups except ArHC and XHC also contain other functional groups.

Table 9. Selected substructure groups and their conditions for Dragon descriptors

Substructure Group	Condition for Dragon Descriptors
Aldehydes	nArCHO, nRCHO > 0
Amides	nArCONH2, nArCONHR, nArCONR2, nRCONH2, nRCONHR, nRCONR2 > 0
Amines	nArNH2, nArNHR, nArNR2, nRNH2, nRNHR, nRNR2 > 0
Aromatic Hydrocarbons	nN=nO=nP=nS=nX=0 & nCar > 0
Carbamates	nArOCON, nROCON > 0
Carboxylic Acids	nArCOOH, nRCOOH > 0
Chlorine Compounds	nArCL, nRCL > 0
Esters	nArCOOR, nRCOOR > 0
Ethers	nArOR, nROR > 0
Halogen Compounds	nArX, nRX > 0
Halohydrocarbons	nN=nO=nP=nS=0 & nX > 0
Hydroxyl Derivatives	nArOH, nROH > 0
Imines	nArC=N, nRC=N > 0
Ketones	nArCO, nRCO > 0
N Azo-derivatives	nN=N > 0
N Hydrazines	nN-N > 0
N-containing Heteroaromatics	n135-Triazines, nImidazoles, nIsoxazoles, nPyrazines, nPyridines, nPyrimidines, nPyrroles, nPyrrolidines, nThiazoles, nTriazoles > 0
Nitro Compounds	nArNO2, nRNO2 > 0
Nitroso Compounds	nArNO, nRNO > 0
N-Nitroso Compounds	nArNNOx, nRNNOx > 0
Phosphorous Compounds	nP > 0
Sulfur Compounds	nS > 0
Urea Derivatives	nCONN > 0

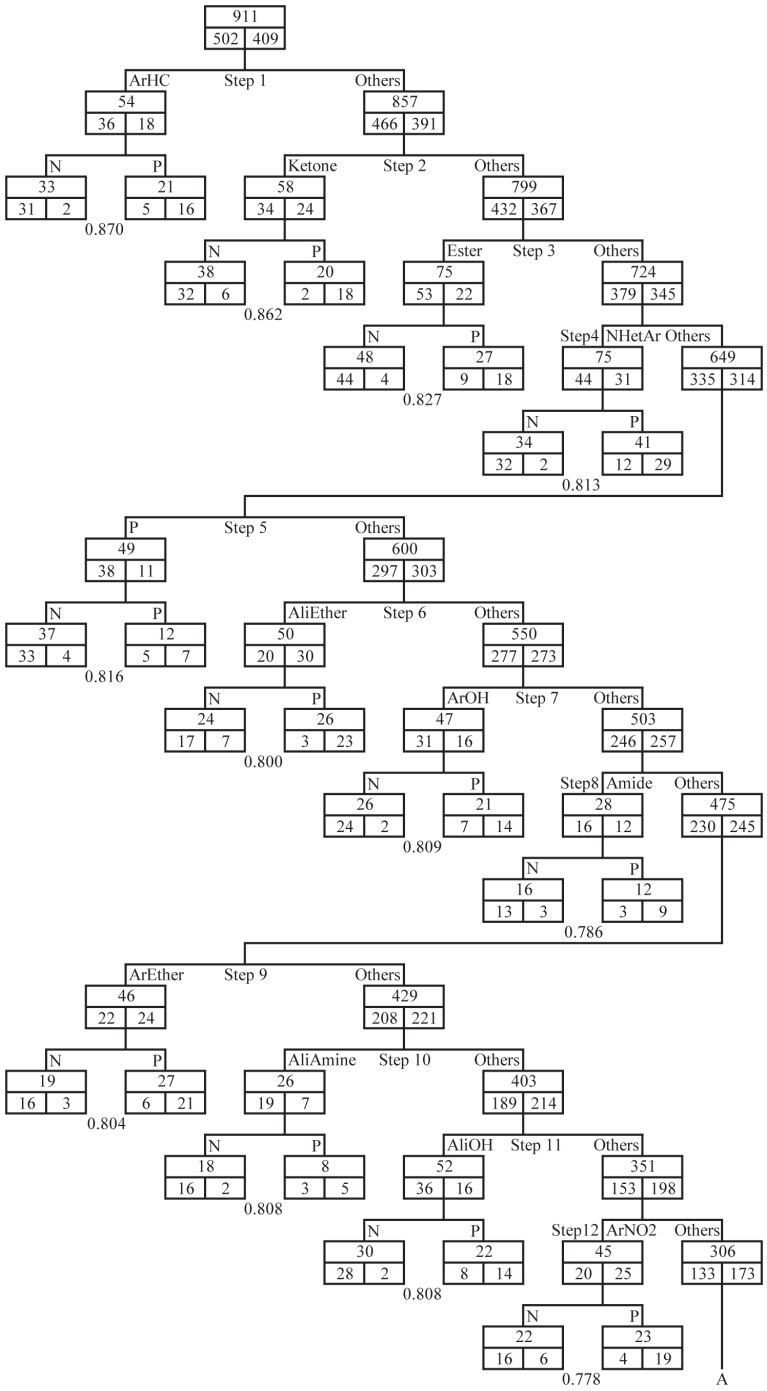

Fig. 6 Result of the serial combination model of SVMs for congeneric chemicals

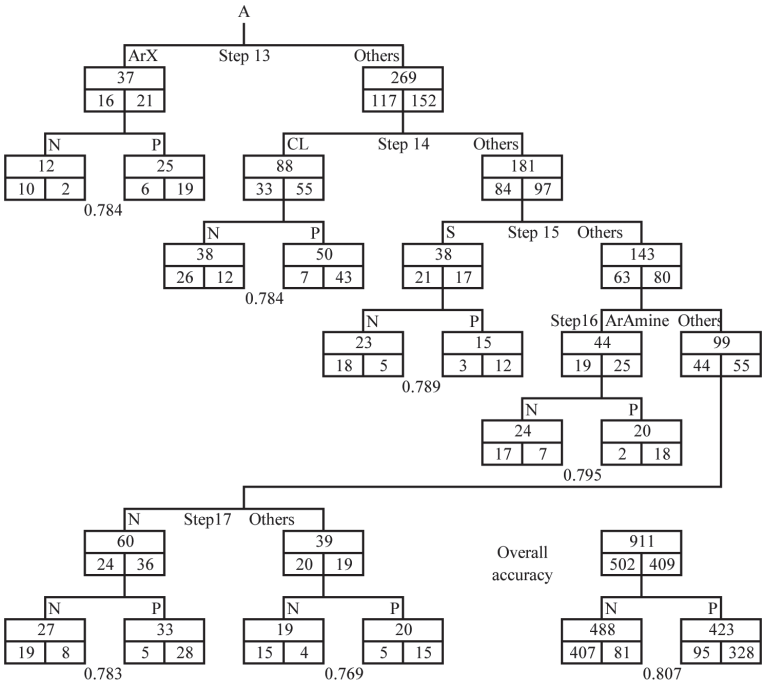

Fig. 6 (continued) Result of the serial combination model of SVMs for congeneric chemicals

3.2.7 同族体に対する SVM の並列組み合わせによる予測

同族体を直列に組み合わせたモデルは高い正解率が得られたものの、モデルの頑健性の点で問題があることが判明した。そこで次に、同族体の並列組み合わせ法を検討し、正解率と頑健性を調べた。この並列組み合わせ法は、予測したい化学物質が含む同族体に対応する SVM の予測結果の多数決で陽性・陰性を判定する方法である。したがって、どのような同族体を選定するかが重要になる。そこで、直列組み合わせ法の場合と同じ選定基準を設け、Table 9 に示す 23 種の部分構造について分割や融合を多数回試行した。その結果、Table 10 に示す 20 種の同族体を選定すれば、ここで解析対象とした全物質が含まれるモデルを構築することができた。幾つかの同族体ではかなり低い正解率になったものの、全同族体での平均正解率は 79.8% となり、目標の 80% にほぼ到達できた。Table 10 には不採用の同族体についての正解率も示した。

ただし、この正解率は各同族体についての個別の SVM の正解率を平均したものであり、化学物質ごとの予測値から算出した正解率ではない。そこで、Table 10 の予測結果から各化学物質ごとに該当する同族体の予測値の多数決により陽性・陰性を判定し、全物質に対する正解率を算出した。その際、該当同族体の数が偶数で、それらの予測結果が陽性・陰性同数の場合は正解数を 0.5 とカウントした。その結果、全物質 911 種に対する予測結果は TP = 313.5、FP = 90.5、TN = 411.5、FN = 95.5 となり、正解率 79.6% という結果が得られた。この正解率は、既存の予測モデルの最高精度よりはるかに高く、また、誤判定 FN

の比率もかなり低く、動物実験代替の発ガン性予測手法としては十分な性能である。直列組み合わせ法の正解率 80.7% よりは若干劣るものの、同族体中で物質数が最少なのは脂肪族 N- ニトロソ・ニトロソ・ニトログループの 34 物質であり、また、同族体の平均物質数は、直列モデルの 50.6 物質に対して、この並列モデルでは 74.4 物質である。したがって、頑健性の点ではこの並列モデルがはるかに優れている。

この同族体並列組み合わせモデルがどの程度広範囲の化学物質の発ガン性を高精度で予測可能であるかについて検討した結果を Table 11 に示す。この表には、このモデルで採用した 20 種の同族体グループに該当する物質群の広がりを示すものとして、炭素数と分子量および記述子間の相関係数の最小値、最大値、平均が示されている。いずれのグループにおいても広範囲の化学物質が含まれていることから、この並列モデルが広範囲の化学物質の発ガン性を高精度で予測可能であると推測される。したがって、今後、発ガン性の

Table 10. Result of the parallel combination model of SVMs for congeneric chemicals

Adopted Substructure Group	NC	NP	NN	TP	TN	OA
Amides & Carbamates	71	30	41	22	31	0.746
Amines & Imines (Aliphatic)	62	20	42	11	39	0.806
Amines & Imines (Aromatic)	141	64	77	35	68	0.730
Carboxylic Acids	46	14	32	8	28	0.783
Esters	85	26	59	19	54	0.859
Ethers (Aliphatic)	80	42	38	36	31	0.838
Ethers (Aromatic)	80	40	40	35	28	0.788
Hydroxyl Derivatives (Aliphatic)	134	44	90	26	79	0.784
Hydroxyl Derivatives (Aromatic)	66	25	41	18	34	0.788
Ketones	58	24	34	18	32	0.862
N-containing Heteroaromatics	88	37	51	34	41	0.852
N Hydrazines & N Azo-derivatives	41	19	22	15	15	0.732
Nitro, Nitroso & N-Nitroso Compounds (Aliphatic)	34	25	9	20	6	0.765
Nitro, Nitroso & N-Nitroso Compounds (Aromatic)	89	41	48	31	32	0.708
Phosphorous Compounds	55	11	44	6	41	0.855
Sulfur Compounds	110	41	69	26	62	0.800
Urea Derivatives	47	24	23	21	19	0.851
Aromatic Hydrocarbons	54	18	36	16	31	0.870
Halohydrocarbons	85	55	30	53	18	0.835
Others	62	26	36	20	29	0.790
Total	1488	626	862	470	718	0.798
Unadopted Substructure Group	NC	NP	NN	TP	TN	AC
Aldehydes & Ketones	74	33	41	24	30	0.730
Amides	47	18	29	10	10	0.617
Amides, Carbamates & Urea Derivatives	117	54	63	41	46	0.744
Amines	193	82	111	40	91	0.679
Aliphatic Amines	61	20	41	9	37	0.754
Aromatic Amines	134	63	71	38	56	0.701
Chlorine Compounds	231	121	110	88	77	0.714
Aliphatic Chlorine Compounds	149	86	63	75	43	0.792
Aromatic Chlorine Compounds	90	39	51	24	35	0.656
Ethers	148	72	76	54	49	0.696
Aliphatic Halogen Compounds	172	100	72	82	48	0.756
Aromatic Halogen Compounds	98	43	55	28	38	0.673
Hydroxyl Derivatives	184	61	123	30	111	0.766
Nitro Compounds	92	43	49	32	32	0.696
Aromatic Nitro Compounds	86	40	46	28	31	0.686
N-Nitroso, Nitroso & Nitro Compounds	123	66	57	55	28	0.675

NC : Number of chemicals. NP : Number of positive chemicals. NN : Number of negative chemicals.
TP : Number of true positive chemicals. TN : Number of true negative chemicals. OA : Overall accuracy.

データが既知の化学物質の数が増えて直列組み合わせ法の頑健性が向上する可能性はあるが、正解率と頑健性の両者を考慮すると、現状ではこの並列組み合わせ法が最適の予測法であると結論される。

Table 11. Minimal, maximal, and mean values of numbers of carbon atoms, molecular weights, and correlation coefficients between descriptors to illustrate diversities of chemicals containg adopted substructure groups

Adopted Substructure Group	NC	Number of C atoms			Molecular Weight			Correlation Coefficient		
		Min	Max	Mean	Min	Max	Mean	Min	Max	Mean
Amides & Carbamates	71	2	32	11.2	59.1	629.5	245.7	0.108	1.000	0.750
Amines & Imines (Aliphatic)	62	1	33	12.5	43.1	608.8	261.3	0.252	1.000	0.745
Amines & Imines (Aromatic)	141	2	42	11.7	84.1	840.9	224.1	0.505	1.000	0.834
Carboxylic Acids	46	2	24	9.4	72.1	454.5	224.8	0.214	1.000	0.687
Esters	85	3	42	15.0	72.1	788.7	289.0	0.099	0.999	0.718
Ethers (Aliphatic)	80	2	33	10.9	44.1	656.7	243.6	-0.036	1.000	0.704
Ethers (Aromatic)	80	5	33	15.2	123.2	959.1	311.4	0.274	0.998	0.777
Hydroxyl Derivatives (Aliphatic)	134	1	35	11.0	46.1	656.7	242.4	0.198	1.000	0.733
Hydroxyl Derivatives (Aromatic)	66	6	32	14.4	94.1	840.9	275.3	0.231	1.000	0.771
Ketones	58	3	35	16.3	58.1	646.7	301.0	0.409	0.999	0.867
N-containing Heteroaromatics	88	2	33	10.9	79.1	608.8	230.6	0.234	1.000	0.736
N Hydrazines & N Azo-derivatives	41	0	32	13.2	32.1	840.9	248.7	-0.496	0.998	0.711
Nitro, Nitroso & N-Nitroso Compounds (Aliphatic)	34	1	10	4.8	61.1	313.7	145.6	0.561	0.994	0.867
Nitro, Nitroso & N-Nitroso Compounds (Aromatic)	89	3	24	10.1	123.1	487.5	231.1	0.484	1.000	0.803
Phosphorous Compounds	55	2	24	8.9	110.1	697.6	284.5	0.353	1.000	0.664
Sulfur Compounds	110	1	32	9.0	60.1	840.9	250.4	-0.117	0.992	0.594
Urea Derivatives	47	1	15	8.1	76.1	376.7	211.2	0.450	0.995	0.778
Aromatic Hydrocarbons	54	6	28	17.2	78.1	352.4	219.0	0.178	1.000	0.735
Halohydrocarbons	85	1	18	4.1	46.1	943.1	187.8	0.089	1.000	0.666
Others	62	0	24	5.5	28.1	413.1	129.0	-0.141	1.000	0.572

NC : Number of chemicals

3.3 考察

本研究では、多種多様な化学物質の発ガン性を構造情報のみから高い精度で予測できる方法を求めて、SVM をアンサンブル学習や決定木と組み合わせる方法、および同族体毎の個別 SVM を組み合わせる方法を検討した。以上の 7 モデルについて得られた正解率をまとめて Table 12 に示す。まず、全物質を一括して解析したモデルの成績がこのように低くなった原因を考察する。同族体 SVM の並列組み合わせ法で採用した 20 種の同族体について、発ガン性データとの相関の高い記述子の相関係数を Table 13 に示す。各同族体について最も相関の高い記述子は太字で示されている。

全物質 911 種を一括解析したモデルの相関係数は最上行の All Chemicals の欄に示されているが、どの記述子についても相関係数は低い。中でも記述子 CIC1 のように、発ガン性にプラスに寄与する同族体とマイナスに寄与する同族体が共存しているために、全物質を一括して解析すると、寄与が相殺して発ガン性との相関が低くなり、予測成績が低くなっ

たと考えられる。Fig. 7 に示す、発ガン性データと記述子との相関係数の分布からも、同族体を組み合わせるモデルでは、相関係数が高く、発ガン性に大きく寄与する記述子が多いのに対し、一括モデルでは相関係数が低い領域に集中しており、発ガン性の予測成績が低くなったことが理解できる。

この一括モデルに対して、全物質を幾つかのグループに分割し、グループ学習するモデルを幾つか検討したが、SVM をアンサンブル学習法（バギング）や決定木と組み合わせたモデルでは一括モデルと同程度の予測成績しか得られなかった。このことは、アンサンブル学習法や決定木のように機械的に分割する方法では、分割されたグループについての発ガン性との相関が全物質についての相関と大差ないために、同程度の予測成績になったと考えられる。これに対して、同族体を組み合わせるモデルで高い予測成績が得られたことは、Table 13 に示すように各同族体ごとに発ガン性予測に有効な記述子が存在していることから理解できる。

しかし、本研究で発ガン性予測に最も有効であると結論できた並列モデルについて、問題点がないわけではない。Table 10 に示した並列モデルの同族体毎の正解率の結果を見ると、芳香族 N- ニトロソ・ニトロソ・ニトロのグループの正解率（70.8%）が他のグループに比べて著しく低い点が目につく。このグループの低成績が並列モデルの総合正解率の低下に大きく寄与しているので、このグループの正解率を上げるために、このグループの分割や他の色々なグループとの融合を試みたが、正解率は向上しなかった。また、このグループ以外でも、芳香族アミン・イミン（正解率 73.0%）、N ヒドラジン・N アゾ誘導体（73.2%）、アミド・カーバメート（74.6%）等のグループや、不採用のアミド（61.7%）、N- ニトロソ・ニトロソ・ニトロ（67.5%）、アミン（67.9%）、芳香族ニトロ（68.6%）、芳香族アミン（70.1%）等、すべて窒素原子を含む同族体群ではいずれも正解率が低い。また、含窒素化合物以外でも、カルボン酸、エーテル、水酸基等の含酸素同族体も正解率が目標の 80% より低い。

これらのグループの正解率が低い原因としては、芳香族化合物については芳香環についた他の置換基の影響が考えられる。すなわち、芳香族アミンや芳香族ニトロ等の芳香族化合物では、芳香環についた他の置換基によりアミノ基やニトロ基の電子状態が大きく変化し、それが発ガン機構に影響して、発ガン陽性・陰性の逆転を生じる可能性がある。しかし、ここで用いた Dragon 記述子にはそのような電子的効果を表す記述子が含まれていないため、低い正解率になったと推測される。含酸素化合物についても同様の状況の可能性が考えられる。

このような共存グループの影響は Table 8 の結果からも見ることができる。この表には、本研究で解析対象とした全物質 911 種について記述子発生プログラム Dragon を用い、種々の部分構造を含む発ガン陽性・陰性別の物質数を集計した結果が NP および NN の欄に示されている。この陽性／陰性の物質数の比率について統計的に有意かどうか、すなわち、発ガン陽性または陰性物質が統計的に多いと判定できるかどうかの結果が SS の欄に示されている。この欄で + + 記号があるものは有意水準 95% で陽性物質が多いと判定された部分構造であり、– – 記号があるものは陰性物質が多いと判定された部分構造である。この結果を見ると、発ガン陽性物質優位と判定された部分構造がかなり少ないことは意外である。また、従来、発ガン原因構造とされてきた部分構造（例えば芳香族アミン、芳香族ニ

トロ等）でも陰性物質がかなり多く、統計的には発ガン陽性物質優位と判定されなかったものが多いことも意外である。このことは、芳香族アミン、芳香族ニトロ化合物等、従来、発ガン原因構造と考えられてきた同族体でも、共存する他の部分構造の影響で、発ガン陰性の物質が多いのではないかと推測される。

したがって、並列モデルにおける低成績グループの正解率を向上させるための対策としては、ここで用いたDragon記述子の外に、このような共存グループの電子的効果を表現する様々な記述子の導入が考えられる。例えば、電子密度等の量子化学的記述子は種々の部分構造の電子状態を表現できるので、これらの記述子の導入により含窒素芳香族化合物の正解率の向上が期待される。また、含酸素化合物群についても正解率向上の可能性があるので、このような量子化学的記述子を含む様々な記述子を導入して解析を行うことは今後の課題である。

本研究で発ガン性予測手法として最適であると結論した並列組み合わせモデルについても、データ数を増やしてモデルの頑健性をさらに向上させる必要はある。そのためには最低限の動物実験を行うことが不可欠であるが、ここで開発した予測モデルを用いて未知物質の発ガン性を予測することにより、動物実験データの効率的な取得が可能になる。環境中に存在する発ガン性未知の化学物質は十万種以上といわれており、この莫大な化学物質について発ガン性の情報を取得することが喫緊の課題である。そのような社会的要請に応えるべく、広範囲の化学物質の発ガン性を統一的に予測できる手法を緊急に開発するためには、国際協力のもとに化学物質の発ガン性データを飛躍的に増加させることが必要である。

Table 12. Summary of the overall accuracies for seven models treated in this study

Model		TP	FP	TN	FN	OA
Batch model	3.4.1	258	133	369	151	0.688
Combination model of SVM and AdaBoost	3.4.2	259	163	339	150	0.656
Combination model of SVM and revised bagging	3.4.3	268	141	361	141	0.690
Combination model of SVM and decision tree	3.4.4	297	174	328	112	0.686
Combination model of SVM and repêchage decision tree	3.4.5	311	165	337	98	0.720
Serial combination model of SVMs for congeners	3.4.6	328	95	407	81	0.807
Parallel combination model of SVMs for congeners	3.4.7	313.5	90.5	411.5	95.5	0.796

Table 13. Correlation coefficients between carcinogenicities and selected descriptors used

Substructures	G2	SPAM	HOMT	MAXDP	G(O..Cl)	MATS2v	CIC1	FDI	N-072
All Chemicals	**-0.182**	0.144	0.162	-0.123	-0.006	0.014	-0.025	0.164	-0.026
Amides & Carbamates	-0.296	**0.382**	0.172	-0.160	0.012	-0.075	-0.228	0.273	-0.112
Amines & Imines (Aliphatic)	-0.181	0.144	**0.416**	0.068	0.234	-0.018	-0.030	0.352	-0.019
Amines & Imines (Aromatic)	-0.155	0.131	0.109	**-0.333**	-0.078	-0.083	0.160	0.206	-0.181
Carboxylic Acids	-0.164	0.379	0.243	-0.115	**0.487**	-0.074	-0.040	0.406	-0.017
Esters	-0.042	0.003	0.047	0.111	-0.153	**-0.371**	0.003	0.193	0.031
Ethers (Aliphatic)	-0.113	0.146	0.095	-0.019	-0.008	-0.149	**-0.354**	-0.023	0.148
Ethers (Aromatic)	-0.298	0.153	**0.437**	-0.015	0.011	0.000	-0.164	0.398	-0.116
Hydroxyl Derivatives (Aliphatic)	-0.073	0.123	0.246	0.080	0.185	0.050	0.044	**0.395**	-0.018
Hydroxyl Derivatives (Aromatic)	-0.025	0.120	0.280	0.196	0.062	0.197	0.334	**0.603**	-0.210
Ketones	-0.357	-0.021	0.489	0.181	-0.070	0.236	0.266	**0.570**	-0.223
N Hydrazines & N Azo-derivatives	-0.128	0.038	0.009	-0.473	-0.214	-0.312	-0.050	-0.176	**-0.527**
N-containing Heteroaromatics	-0.236	0.208	0.428	-0.052	-0.115	-0.009	0.104	0.388	-0.030
Nitro, Nitroso & N-Nitroso Compounds (Aliphatic)	-0.181	0.155	-0.159	0.072	0.098	0.028	0.207	0.197	0.265
Nitro, Nitroso & N-Nitroso Compounds (Aromatic)	-0.247	0.126	0.365	0.122	-0.013	0.161	0.115	0.398	0.023
Phosphorous Compounds	-0.031	0.018	0.183	0.070	0.019	0.234	0.010	0.026	0.102
Sulfur Compounds	-0.253	0.361	0.179	-0.031	-0.017	-0.002	-0.083	0.362	-0.118
Urea Derivatives	-0.406	0.237	0.175	0.097	0.339	-0.024	-0.173	0.351	-0.169
Aromatic Hydrocarbons	0.141	-0.023	-0.048	0.136	nu	0.000	0.280	-0.182	nu
Halohydrocarbons	0.071	-0.208	0.026	-0.203	nu	0.189	0.169	-0.182	nu
Others	-0.074	0.037	-0.090	0.030	0.100	-0.106	-0.162	-0.109	0.025

Substructures	HOMA	nRCOOH	nCH2RX	Hyper tens-80	GATS1m	IC1	F-083	BELm8
All Chemicals	0.175	-0.006	0.128	-0.076	-0.004	-0.096	-0.056	-0.076
Amides & Carbamates	0.086	-0.052	0.016	0.004	0.177	0.042	nu	-0.171
Amines & Imines (Aliphatic)	0.367	0.073	0.321	0.079	-0.070	0.092	nu	0.096
Amines & Imines (Aromatic)	0.212	0.024	0.058	0.093	0.010	-0.189	0.060	0.070
Carboxylic Acids	0.324	0.424	0.327	0.065	-0.058	0.052	nu	-0.008
Esters	0.126	0.227	-0.143	0.036	0.101	0.138	-0.075	0.124
Ethers (Aliphatic)	-0.005	-0.175	-0.025	-0.008	-0.020	0.065	-0.123	-0.075
Ethers (Aromatic)	0.416	0.078	0.082	-0.075	0.159	0.184	0.072	-0.044
Hydroxyl Derivatives (Aliphatic)	0.287	0.120	0.226	0.094	0.065	0.107	nu	0.110
Hydroxyl Derivatives (Aromatic)	0.586	0.082	nu	0.070	0.146	-0.005	nu	0.287
Ketones	0.534	-0.186	-0.203	0.178	-0.211	-0.055	nu	0.195
N Hydrazines & N Azo-derivatives	-0.013	-0.027	nu	-0.278	0.058	-0.376	nu	-0.387
N-containing Heteroaromatics	**0.432**	-0.237	nu	-0.174	0.123	-0.002	nu	0.123
Nitro, Nitroso & N-Nitroso Compounds (Aliphatic)	-0.183	**-0.450**	0.181	-0.105	-0.113	-0.147	nu	0.049
Nitro, Nitroso & N-Nitroso Compounds (Aromatic)	**0.401**	nu	nu	0.078	-0.022	-0.026	0.108	0.100
Phosphorous Compounds	0.127	-0.074	**0.461**	-0.400	-0.026	-0.164	nu	-0.291
Sulfur Compounds	0.381	-0.102	0.199	**-0.419**	0.188	-0.136	0.083	-0.240
Urea Derivatives	0.225	nu	0.365	-0.199	**-0.500**	0.056	-0.133	-0.061
Aromatic Hydrocarbons	-0.178	nu	nu	-0.332	nu	**-0.472**	nu	0.030
Halohydrocarbons	0.050	nu	0.188	nu	-0.037	0.001	**-0.352**	-0.112
Others	-0.195	nu	0.049	-0.123	0.073	0.054	0.102	**-0.322**

Bold number : the highest correlation coefficient in that group. nu : the unused descriptor due to zero count.

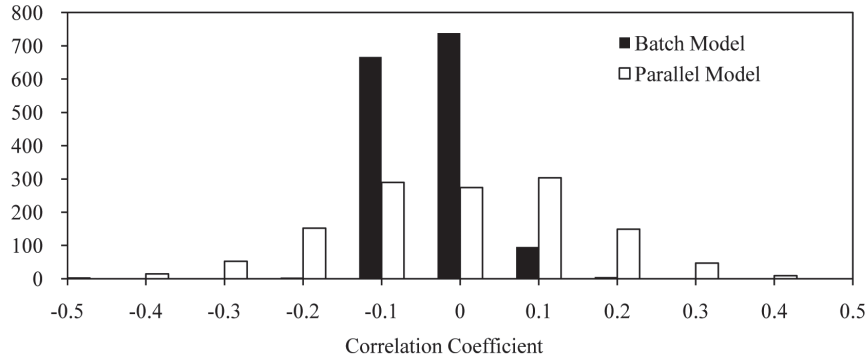

Fig. 7 Histogram of correlation coefficients between carcinogenicities and descriptors in the batch model and in the parallel model

4. 結 論

本研究では、環境中に存在する莫大な数の化学物質について発ガン性に関する情報を提供する手段として、信頼性の高い動物実験データを集積した発ガン性 DB の構築、および動物実験代替の発ガン性予測手法として、多種多様な化学物質についてその化学構造から発ガン性を高精度で予測するシステムの構築の 2 点を目的として検討した。

発ガン性の動物実験データは、信頼性が高い IARC、EU、EPA、NTP、ACGIH、JSOH の 6 種の DB から収集し、データの信頼性ランクを統一するために、種々の DB のデータを総合的に評価して、総物質数 1,512 種について信頼性を 6 段階にランク付けした発ガン性 DB を構築した。

化学構造から発ガン性を高精度で予測するシステムを開発するために、この DB から構造が明確な有機物質 911 種を抽出し、Dragon プログラムを用いて記述子を算出した。これらの記述子から発ガン性を予測するモデルとして、LIBSVM プログラムを用いて SVM を様々な解析技術と組み合わせた種々のモデルを検討した。それらは、単一の SVM による一括予測、SVM とブースティングの組み合わせによる予測、SVM とバギングの組み合わせによる予測、SVM と決定木の組み合わせによる予測、敗者復活を取り入れた決定木と SVM の組み合わせによる予測、同族体に対する SVM の直列組み合わせによる予測、および同族体に対する SVM の並列組み合わせによる予測、の計 7 モデルである。Dual Cross-Validation Test により、各モデルの学習とテストと最適化を行い、正解率と頑健性の 2 点から各モデルの性能を総合的に評価した。その結果、同族体に対する SVM の並列組み合わせによる予測モデルが既存モデルを凌駕する正解率 79.6％を示し、広範囲の化学物質の発ガン性を高精度で予測するモデルとして最適であると結論した。

謝　辞

本研究において協力いただいたた Bono Lučić（The Rudjer Bošković Institute, Croatia）, Dragan Amić（The Josip Juraj Strossmayer University, Croatia）、貝原巳樹雄（一関工業高等専門学校物質化学工学科）、小野寺夏生（前筑波大学大学院図書館情報メディア研究科）の各氏に感謝します。

参考文献

Bahler, D.; Stone, B.; Wellington, C.; Bristol, D. (2000) Symbolic, Neural, and Bayesian Machine Learning Models for Predicting Carcinogenicity of Chemical Compounds. *J. Chem. Inf. Comput. Sci.,* **40**, 906–914.

Basak, S.C.; Grunwald, G.D.; Gute, B.D.; Balasubramanian, K.; Optiz, D. (2000) Use of Statistical and Neural Net Approaches in Predicting Toxicity of Chemicals. *J. Chem. Inf. Comput. Sci.,* **40**, 885–890.

Barak, S.; Deep, K.; Katiyar, V.K.; Katiyar, C.K. (2009) A State of Art Review on Application of Nature Inspired Optimization Algorithms in Protein-Ligand Docking. *Ind. J. Biomech.,* Special Issue, pp.219–224

Benfenati, E.; Benigni, R.; DeMarini, D.M.; Helma, C.; Kirkland, D.; Martin, T.M.; Mazzatorta, P,.; Ouédraogo-Arras, G.; Richard, A.M.; Schilter, B.; Schoonen, W.G.E.J.; Snyder, R.D.; Yang, C. (2009) Predictive Models for Carcinogenicity and Mutagenicity: Frameworks, State-of-the-art, and Perspectives. *J. Environ. Sci. Health, Part C*, **27**, 57–90.

Benigni, R.; Giuliani, A.; Franke, R.; Gruska, A. (2000) Quantitative Structure–activity Relationships of Mutagenic and Carcinogenic Aromatic Amines. *Chem. Rev*., **100**, 3697–3714.

Benigni, R. (2003a) SARs and QSARs of Mutagens and Carcinogens: Understanding Action Mechanisms and Improving Risk Assessment. In: Benigni, R. (ed.) Quantitative Structure-activity Relationship (QSAR) Models of Mutagens and Carcinogens. CRC Press, Boca Raton, pp 259–282.

Benigni, R.; Giuliani, A.; Gruska, A.; Franke, R. (2003b) QSARs for the Mutagenicity and Carcinogenicity of the Aromatic Amines. In: Benigni, R. (ed.) Quantitative Structure-activity Relationship (QSAR) Models of Mutagens and Carcinogens. CRC Press, Boca Raton, pp 125–144.

Benigni, R. (2004) Prediction of Human Health Endpoints: Mutagenicity and Carcinogenicity. In: Cronin, M.T.D.; Livingstone, D.J. (eds.) Predicting Chemical Toxicity and Fate, CRC Press, Boca Raton, pp 173–192.

Benigni, R. (2005) Structure-activity Relationship Studies of Chemical Mutagens and Carcinogens: Mechanistic Investigations and Prediction Approaches. *Chem. Rev*., **105**, 1767–1800.

Benigni, R.; Bossa, C. (2008) Predictivity of QSAR. *J. Chem. Inf. Model.,* **48**, 971–980.

Benigni, R. (2010) The Benigni/Bossa Rulebase for Mutagenicity and Carcinogenicity–A Module of Toxtree. European Commission Report, EUR 23241 EN, pp.1–75.

Bhavani, S.; Ngargadde, A.; Thawani, A.; Sridhar, V.; Chandra, N. (2006) Substructure-based Support Vector Machine Classifiers for Prediction of Adverse Effects in Diverse Classes of Drugs. *J. Chem. Inf. Model.*, **46**, 2478–2486.

Braga, R.S.; Barone, P.M.V.B.; Galvao, D.S. (1999) Identifying Carcinogenic Activity of Methylated Polycyclic Aromatic Hydrocarbons (PAHs). *J. Mol. Struct.*, **464**, 257–266.

Breiman, L. (1994) Bagging Predictors. Technical Report 421, Statistics Dept., Univ. California, Berkley.

Bruce, C.L.; Melville, J.L.; Pickett, S.D.; Hirst, J.D. (2007) Contemporary QSAR Classifiers Compared. *J. Chem. Inf. Model.*, **47**, 219–227.

Byvatov. E.; Fechner, U.; Sadowski, J.; Schneider, G. (2003) Comparison of Support Vector Machine and Artificial Neural Network Systems for Drug/nondrug Classification. *J. Chem. Inf. Comput. Sci.*, **43**, 1882–1889.

Chang, C.C.; Lin, C.J. (2009a) LIBSVM-A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/ (accessed May 25, 2009).

Chang, C.C.; Lin, C.J. (2009b) LIBSVM-A Library for Support Vector Machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/#14 (accessed May 25, 2009).

Chen, N.; Lu, W.; Yang, J.; Li, G. (eds.) (2004a) Support Vector Machine in Chemistry. World Scientific, Singapore.

Chen, N.; Lu, W.; Yang, J.; Li, G. (2004b) SVM Applied to Structure-activity Relationships. In: Chen, N.; Yang, L. (eds.) Support Vector Machine in Chemistry. World Scientific, Singapore, pp 186–219.

Contrera, J.F.; MacLaughlin, P.; Hall, L.H.; Kier, L.B. (2005) QSAR Modeling of Carcinogenic Risk Using Discriminant Analysis and Topological Molecular Descriptors. *Curr. Drug Discov. Tech.*, **2**, 55–67.

Crettaz, P.; Benigni, R. (2005) Prediction of the Rodent Carcinogenicity of 60 Pesticides by the DEREKfw Expert System. *J. Chem. Inf. Comput. Sci.*, **45**, 1864–1873.

Devillers, J. (1996a) Neural Networks in QSAR and Drug Design, Academic Press, San Diego.

Devillers, J. (1996b) Strengths and Weaknesses of the Back-propagation Neural Network in QSAR and QSPR Studies. In: Devillers, J. (ed.) Neural Networks in QSAR and Drug Design, Academic Press, London, pp 1–46.

Doll, R.; Peto, R. (1981) The Causes of Cancer: Quantitative Estimates of Avoidable Risks of Cancer in the United States Today. *J. Natl. Cancer Inst.*, **66**, 1192–1309.

Doucet, J-P.; Barbault, F.; Xia, H.; Panaye, A.; Fan, B. (2007) Nonlinear SVM Approaches to QSPR/QSAR Studies and Drug Design. *Curr. Comp. Aid. Drug Design*, **3**, 263–289.

Fjodorova, N.; Vračko, M.; Tušar, M.; Jezierska, A.; Novič, M.; Kühne, R.; Schüürmann, G. (2009) Quantitative and Qualitative Models for Carcinogenicity Prediction for Non-

congeneric Chemicals Using CP ANN Method for Regulatory Uses. *Mol. Div*., published online: 15 August 2009.

Franke, R.; Gruska, A.; Giuliani, A.; Benigni, R. (2001) Prediction of Rodent Carcinogenicity of Aromatic Amines: A Quantitative Structure-activity Relationships Model. *Carcinogenisis*, **22**, 1561–1571.

Freund, Y.; Schapire, R.E. (1997) A Decision-theoretic Generalization of On-line Learning and an Application to Boosting. *J. Comput. System Sci*., **55**, 119–139.

Fukunishi, H.; Teramoto, R.; Shimada, J. (2008) Hidden Active Information in a Random Compound Library: Extraction Using a Pseudo-structure-activity Relationship Model. *J. Chem. Inf. Model*., **48**, 575–582.

Gasteiger, J.; Sadowski, J.; Schuur, J.; Selzer, P.; Steinhauer, L.; Steinhauer, V. (1996) Chemical Information in 3D Space. *J. Chem. Inf. Comput. Sci.,* **36**, 1030–1037.

Guyton, K.Z.; Kyle, A.D.; Aubrecht, J.; Cogliano, V.J.; Eastmond, D.A.; Jackson, M.; Keshava, N.; Sandy, M.S.; Sonawane, B,; Zhang, L.; Waters, M.D.; Smith, M.T. (2009) Improving Prediction of Chemical Carcinogenicity by Considering Multiple Mechanisms and Applying Toxicogenomic Approaches. 1. *Mutat. Res*., **681**, 230–240.

Hansch, C.; Fujita, T. (1964) A Method for the Correlation of Biological Activity and Chemical Structure. *J. Am. Chem. Soc.*, **86**, 1616–1626

Harvard Center for Cancer Prevention (1996) Harvard Report on Cancer Prevention. Volume 1: Causes of Human Cancer. *Cancer Causes Control,* **7**, S3–S59.

Helguera, A.M.; Perez, M.C.A.; Combes, R.D.; Gonzalez, M.P. (2005) The Prediction of Carcinogenicity from Molecular Structure. *Curr. Comp. Aid. Drug Des.,* **1**, 237–255.

Helma, C.; Kramer, S. (2003) A Survey of the Predictive Toxicology Challenge 2000-2001. *Bioinformatics*, **19**, 1179–1182.

Helma, C.; King, R.D.; Kramer, S.; Srinivasan, A. (2000) The Predictive Toxicology Challenge (PTC) for 2000-2001. http://www.informatik.uni-freiburg.de/~ml/ptc/ (accessed May 1, 2009).

Helma, C.; Cramer, T.; Kramer, S.; De Raedt, L. (2004) Data Mining and Machine Learning Techniques for the Identification of Mutagenicity Inducing Substructures and Structure Activity Relationships of Noncongeneric Compounds. *J. Chem. Inf. Comput. Sci*., **44**, 1402–1411.

Hemmateenejad, B.; Safarpour, M.; Miri, R.; Nesari, N. (2005) Toward an Optimal Procedure for PC-ANN Model Building: Prediction of the Carcinogenic Activity of a Large Set of Drugs. *J. Chem. Inf. Model.*, **45**, 190–199.

Ivanciuc, O. (2002) Support Vector Machine Classification of the Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons. Internet Electron. *J. Mol. Design*, **1**, 203–218.

Ivanciuc, O. (2007) Applications of Support Vector Machines in Chemistry. *Rev. Comput. Chem*., **23**, 291–400.

Ivanciuc, O. (2009a) Drug Design with Artificial Neural Networks. In: Meyers, R.A. (ed.)

Encyclopedia of Complexity and System Science, Springer-Verlag, New York.

Ivanciuc, O. (2009b) Drug Design with Machine Learning. In: Meyers, R.A. (ed.) Encyclopedia of Complexity and System Science, Springer-Verlag, New York.

JETOC (2007) Estimation and Classification Criteria of Carcinogenicity of Chemical Substances. Japan Chemical Industry Ecology-Toxicology and Information Center, Tokyo, pp 21–23.

Jorissen, R.N.; Gilson, M.K. (2005) Virtual Screening of Molecular Databases Using a Support Vector Machine. *J. Chem. Inf. Comput. Sci*., **45**, 549–561.

Langham, J.J.; Jain, A.N. (2008) Accurate and Interpretable Computational Modeling of Chemical Mutagenicity. *J. Chem. Inf. Model*., **48**, 1833–1839.

Liu, T-Y.; Li, G-Z.; Yang, J.Y.; Yang, M.Q. (2008) Feature Selection for the Imbalanced QSAR Problems by Using EasyEnsemble. *Internatl. J. Comput. Biol. Drug Design*, **1**, 334–346.

Luan, F.; Zhang, R.; Zhao, C.; Yao, X.; Liu, M.; Hu, Z.; Fan, B. (2005) Classification of the Carcinogenicity of N-nitroso Compounds Based on Support Vector Machines and Linear Discriminant Analysis. *Chem. Res. Toxicol*., **18**, 198–203.

Massarelli, I.; Imbriani, M.; Coi, A.; Saraceno, M.; Carli, N.; Bianucci, A.M. (2009) Development of QSAR Models for Predicting Hepatocarcinogenic Toxicity of Chemicals. *Eur. J. Med. Chem*., **44**, 3658–64.

Nordling, C.O. (1953) A New Theory on Cancer-inducing Mechanism. *Brit. J. Cancer,* **7**, 68–72.

Oellien, F.; Nicklaus, M.C. (2000) Online SMILES Translator and Structure File Generator. http://cactus.nci.nih.gov/services/translate/ (accessed July 17, 2009).

Passerini, L. (2003) QSARs for Individual Classes of Chemical Mutagens and Carcinogens. In: Benigni, R. (ed.) Quantitative Structure–activity Relationship (QSAR) Models of Mutagens and Carcinogens. CRC Press, Boca Raton, pp 81–123.

Patlewicz, G.; Rodford, R.; Walker, J.D. (2003) Quantitative Structure-activity Relationships for Predicting Mutagenicity and Carcinogenicity. *Environ. Toxicol. Chem*., **22**, 1885–1893.

Peterson, K.L. (2000) Artificial Neural Networks and their Use in Chemistry. In: Lipkowitz, K.B.; Boyd, D.B. (eds.) Reviews in Computational Chemistry, Volume 16, Wiley-VCH, New York, pp 53–140.

Sun, H. (2004) Prediction of Chemical Carcinogenicity from Molecular Structure. *J. Chem. Inf. Comput. Sci*., **44**, 1506–1514.

Svetnik, V.; Wang, T.; Tong, C.; Liaw, A.; Sheridan, R.P.; Song, Q. (2005) Boosting: An Ensemble Learning Tool for Compound Classification and QSAR Modeling. *J. Chem. Inf. Comput. Sci.,* **45**, 786–799.

Tan, N.X.; Rao, H.B.; Li, Z.R.; Li, X.Y. (2009) Prediction of Chemical Carcinogenicity by Machine Learning Approaches. *SAR QSAR Envir. Res.,* **20**, 27–75.

Tanabe, K.; Ohmori, N.; Ono, S.; Suzuki, T.; Matsumoto, T.; Nagashima, U.; Uesaka, H. (2005) Neural Network Prediction of Carcinogenicity of Diverse Organic Compounds. *J. Comput.*

Chem. Jpn., **4**, 89–100.

Tanabe, K.; Suzuki, T.; Kaihara, M.; Onodera, N. (2008) Prediction of Carcinogenicity of Noncongeneric Chemical Substances by a Support Vector Machine. *J. Comput. Chem. Jpn*., **7**, 93–102.

Tang, L-J.; Zhou, Y-P.; Jiang, J-H.; Zou, H-Y.; Wu, H-L.; Shen, G-L.; Yu, R-Q. (2007) Radial Basis Function Network-based Transform for a Nonlinear Support Vector Machine as Optimized by a Particle Swarm Optimization Algorithm with Application to QSAR Studies. *J. Chem. Inf. Model.,* **47**, 1438–1445.

Todeschini, R.; Consonni, V. (2006) DRAGON Professional 5.4 Program, TALETE srl, Milano, Italy, (http://www. talete.mi.it/dragon.htm).

Toropov, A.A.; Toropova, A.P.; Benfenati, E.; Manganaro, A. (2009a) QSAR Modelling of Carcinogenicity by Balance of Correlations. *Mol. Div*., **13**, 367–373.

Toropov, A.A.; Toropova, A.P.; Benfenati, E. (2009b) Additive SMILES–based Carcinogenicity Models: Probabilistic Principles in the Search for Robust Predictions. *Int. J. Mol. Sci*., **10**, 3106–3127.

Urano, K. (2001) Toxicity Ranks and Physical Property Information for PRTR–MSDS Chemical Substances, Chapter 2, Rank of Carcinogenicity. Kagaku Kogyo Nippo, Tokyo, pp 21–23.

Vendrame, R.; Braga, R.S.; Takahata, Y.; Galvao, D.S. (1999) Structure-activity Relationships of Carcinogenic Activity of Polycyclic Aromatic Hydrocarbons Using Calculated Molecular Descriptors with Principal Component Analysis and Neural Network Methods. *J. Chem. Inf. Comput. Sci*., **39**, 1094–1104.

Venkatapathy, R.; Wang, C.Y.; Bruce, R.M.; Moudgal, C. (2009) Development of Quantitative Structure–activity Relationship (QSAR) Models to Predict the Carcinogenic Potency of Chemicals I. Alternative Toxicity Measures as an Estimator of Carcinogenic Potency. *Toxicol. Appl. Pharmacol*., **234**, 209–221.

Vracko, M. (2000) A Study of Structure-carcinogenicity Relationship for 86 Compounds from NTP Database Using Topological Indexes as Descriptors. *SAR QSAR Environ. Res*., **11**, 103–115.

Woo, Y-T.; Lai, D.Y. (2003) Mechanisms of Action of Chemical Carcinogens and their Role in Structure-activity Relationship (SAR) Analysis and Risk Assessment. In: Benigni, R. (ed.) Quantitative Structure-activity Relationship (QSAR) Models of Mutagens and Carcinogens. CRC Press, Boca Raton, pp 41–80.

Xue, Y.; Li, Z.R.; Yap, C.W.; Sun, L.Z.; Chen, X.; Chen, Y.Z. (2004) Effect of Molecular Descriptor Feature Selection in Support Vector Machine Classification of Pharmacokinetic and Toxicological Properties of Chemical Agents. *J. Chem. Inf. Comput. Sci*., **44**, 1630–1638.

Yao, X.J.; Panaye, A.; Doucet, J.P.; Zhang, R.S.; Chen, H.F.; Liu, M.C.; Hu, Z.D.; Fan, B.T. (2004) Comparative Study of QSAR/QSPR Correlations Using Support Vector Machines,

Radial Basis Function Neural Networks, and Multiple Linear Regression. *J. Chem. Inf. Comput. Sci.*, **44**, 1257–1266.

Zhou, Z.; Dai, Q.; Gu, T.A. (2003) QSAR Model of PAHs Carcinogenesis Based on Thermodynamic Stabilities of Bioactive Sites. *J. Chem. Inf. Comput. Sci.*, **43**, 615–621.

Zupan, J.; Gasteiger, J. (1999) Quantitative Structure-activity Relationships. In: Zupan, J.; Gasteiger, J. (eds.) Neural Networks in Chemistry and Drug Design, Second Edition, Wiley-VCH, Weinheim, pp 219–242.