Mol Divers (2010) 14:789–802
DOI 10.1007/s11030-010-9232-y

FULL-LENGTH PAPER

Prediction of carcinogenicity for diverse chemicals based on substructure grouping and SVM modeling

Kazutoshi Tanabe · Bono Lučić · Dragan Amić · Takio Kurita · Mikio Kaihara · Natsuo Onodera · Takahiro Suzuki

Received: 3 August 2009 / Accepted: 5 February 2010 / Published online: 26 February 2010

Abstract The Carcinogenicity Reliability Database (CRDB) was constructed by collecting experimental carcinogenicity data on about 1,500 chemicals from six sources, including IARC, and NTP databases, and then by ranking their reliabilities into six unified categories. A wide variety of 911 organic chemicals were selected from the database for QSAR modeling, and 1,504 kinds of different molecular descriptors were calculated, based on their 3D molecular structures as modeled by the Dragon software. Positive (carcinogenic) and negative (non-carcinogenic) chemicals containing various substructures were counted using atom and functional group count descriptors, and the statistical significance of ratios of positives to negatives was tested for those substructures. Very few were judged to be strongly related to carcinogenicity, among substructures known to be responsible for carcinogens as revealed from biomedical studies. In order to develop QSAR models for the prediction of the carcinogenicities of a wide variety of chemicals with a satisfactory performance level, the relationship between the carcinogenicity data with improved reliability and a subset of significant descriptors selected from 1,504 Dragon descriptors was analyzed with a support vector machine (SVM) method: the classification function (SVC) for weighted data in 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. The quality and stability of the models presented were tested by performing a dual cross–validation procedure. A single SVM model as the first step was developed for all the 911 chemicals using 250 selected descriptors, achieving an overall accuracy level, i.e., positive and negative correct estimate, of about 70%. In order to improve the accuracy of the final model, the 911 chemicals were classified into 20 mutually overlapping subgroups according to contained substructures, a specific SVM model was optimized for each subgroup, and the predicted carcinogenicities of the 911 chemicals were determined by the majorities of the outputs of the corresponding SVM models. The model developed on the basis of grouping of chemicals into 20 substructures predicts the carcinogenicities of a wide variety of chemicals with a satisfactory overall accuracy of approximately 80%.

Keywords Quantitative structure–activity relationship (QSAR) · Carcinogenicity prediction · Substructure grouping · Support vector machine (SVM) ·

Electronic supplementary material The online version of this article (doi:10.1007/s11030-010-9232-y) contains supplementary material, which is available to authorized users.

K. Tanabe (✉) · T. Kurita
Neuroscience Research Institute, National Institute of Advanced Industrial Science and Technology, Umezono 1-1-1, Tsukuba 305-8568, Japan
e-mail: k-tanabe@aist.go.jp

K. Tanabe · N. Onodera
Graduate School of Library, Information and Media Studies, University of Tsukuba, Kasuga 1-2, Tsukuba 305-8550, Japan

B. Lučić
NMR Center, The Rudjer Bošković Institute, P.O. Box 180, 10002 Zagreb, Croatia

D. Amić
Faculty of Agriculture, The Josip Juraj Strossmayer University, P.O. Box 719, 31107 Osijek, Croatia

M. Kaihara
Department of Chemical Engineering, Ichinoseki National College of Technology, Takanashi, Hagisho, Ichinoseki 021-8511, Japan

T. Suzuki
Natural Science Laboratory, Toyo University, Hakusan 5-28-20, Bunkyo-ku, Tokyo 112-8606, Japan

Support vector classification (SVC) · Molecular descriptors · Correlation coefficient · Cross-validation (CV)

Abbreviations

QSAR	Quantitative structure–activity relationship
PAH	Polycyclic aromatic hydrocarbons
PTC	Predictive Toxicology Challenge
EL	Ensemble learning
ML	Machine learning
NTP	US National Toxicology Program
FDA	US Food and Drug Administration
DB	Database
ANN	Artificial neural network
SVM	Support vector machine
IARC	International Agency for Research on Cancer
EU	European Union
EPA	US Environmental Protection Agency
ACGIH	American Conference of Governmental Industrial Hygienists
JSOH	Japan Society for Occupational Health
PRTR-MSDS	Pollutant Release and Transfer Register-Material Safety Data Sheet
CRDB	Carcinogenicity Reliability Database
CC	Correlation coefficient
SVC	Support vector classification function in the LIBSVM program
RBF	Radial basis kernel function
CV	Cross-validation
LOO	Leave-one-out
AUROC	Area under the receiver operating characteristic curve
MCC	Matthews correlation coefficient
OA	Overall accuracy
TP	True positive
TN	True negative
FP	False positive
FN	False negative

Introduction

Cancer is one of the leading causes of human death (around 13% of all deaths) worldwide, and deaths from cancer are projected to be continuously rising. Cancer arises from three main factors: physical (radiation and ultraviolet rays), chemical (carcinogen), and biological (bacteria, virus, and inheritance). Carcinogens contained in food, beverages, and tobacco are the most dominant among these factors [1,2]. There are only carcinogenicity data for a limited few among the hundreds of thousands of chemicals existing in the environment. Animal tests (rodent bioassays) for carcinogenicity are very laborious, time consuming, costly, and require many animals. Therefore, it is impossible to get carcinogenicity data on all unascertained chemicals via animal tests.

A reliable tool for predicting the carcinogenicity of chemicals that have not been tested experimentally would be highly desirable as a screening for animal tests. Quantitative structure–activity relationship (QSAR) approaches have been applied for the prediction of the carcinogenicity of various chemicals [3–11]. They succeeded in modeling congeneric chemicals such as aromatic amines [12–14], and polycyclic aromatic hydrocarbons (PAHs) [15–17], but failed in predicting the carcinogenicity of non-congeneric chemicals [18,19]. Several systems were presented at the Predictive Toxicology Challenge (PTC) 2000–2001 workshop [20]; however, the exercise revealed that all the proposed models poorly predicted the carcinogenicities of a wide variety of chemicals [21].

There could be several reasons for these failures, the most likely being that many of PTC models analyzed data of all the chemicals together. However, the performance of QSAR approaches may be improved by modeling not whole data together as PTC, but subgroups individually. Ensemble learning (EL) techniques such as boosting and bagging have recently been developed in the field of machine learning (ML) [22], and have already been applied to QSAR [23–26]. In such techniques, several subsets are randomly extracted from a whole data set, a model is individually trained for each subset, and the predicted value of a test sample is given by averaging the outputs of all the models. In addition, it has been stressed that QSAR analysis of carcinogenicity should be based on the mechanism of carcinogenesis [18], and biomedical studies have revealed that chemical carcinogens are classified into several groups, according to the mechanisms [27]. For example, aromatic amines and PAHs, which are target chemicals in QSAR studies, are classified as genetoxic carcinogens, while chloroethylenes, dioxins, and phthalate esters are epigenetic carcinogens. It is reasonable to assume that QSAR approaches satisfactorily predict congeneric chemicals, but unsatisfactorily non-congeneric chemicals. Therefore, the performance of a QSAR approach should be examined in predicting the carcinogenicities of a wide variety of chemicals by modeling not the whole data together as PTC, but chemical subgroups individually as EL.

The second reason is that several problems are found in the carcinogenicity data supplied at PTC [21]. The training data were taken from the NTP (US National Toxicology Program), which mainly consist of simple chemicals with low molecular weights, while the test data were taken from the FDA (US Food and Drug Administration), which mainly consist of drugs with complex structures and high molecular weights. The training and test data of PTC also contained many more negative (non-carcinogenic) than positive (carcinogencic) chemicals, and the imbalance of these numbers may be a cause of failure for the PTC models: in fact,

they especially predicted poorly positive chemicals. Furthermore, the carcinogenicity data of PTC were supplied without information on their reliabilities, while carcinogenicity data in the most recent databases (DBs) are ranked into different categories depending on their reliabilities. A QSAR analysis should be carried out using recently published carcinogenicity data and taking into account such reliability information.

The third reason for the failure is that most models presented at PTC were based on linear relationships between chemical structure and carcinogenicity. Linear approaches like the Hansch–Fujita equation may work well for congeneric chemicals. Artificial neural network (ANN) [28–31], a nonlinear technique, has been applied to limited cases of carcinogenicity modeling with some success [32–34]. However, it has significant disadvantages, including local minima, over-fitting, over-training, and long processing time [35]. We have applied ANN to PTC data, but failed to find an optimum model due to the local minimum problem [36]. Support vector machine (SVM) [37,38], a recently developed nonlinear technique, has several advantages over ANN; many articles have been published demonstrating its advantage in QSAR modeling [39–48]. We have applied SVM to PTC data, and found that an SVM approach better predicts the carcinogenicities of a wide variety of chemicals than any of previously proposed ones [49]. Until now, few studies have been devoted to SVM modeling of carcinogenicity, and even those studies were limited to small data sets of congeneric chemicals [50,51].

The primary purpose of this study is to construct a QSAR model for satisfactorily predicting the carcinogenicities of a wide variety of chemicals. For this purpose, a DB was constructed which consists of experimental carcinogenicities and their reliabilities for a diverse range of chemicals. SVM models were then applied to those data and individually optimized for chemical subgroups, and their performances were compared with those of existing models for non-congeneric chemicals.

Data and methods

Carcinogenicity data

Experimental carcinogenicity data for a diverse range of chemicals were collected from six DBs [52]: the IARC (International Agency for Research on Cancer), EU (European Union), EPA (US Environmental Protection Agency), NTP, ACGIH (American Conference of Governmental Industrial Hygienists), and JSOH (Japan Society for Occupational Health). In these DBs, experimental carcinogenicity data are compiled along with their reliabilities. The reliability represents the risk of cancer incidence based on animal tests, and is qualitatively classified into several ranks. Some confusion is found in reliability rankings of various DBs: numbers and descriptions of ranks differ from one DB to the other, as summarized in Table 1 of the Supplementary Material. There also are many cases for which identical chemicals are assigned to different ranks in different DBs. For example, formaldehyde is ranked as 1 in IARC, 3 in EU, B1 in EPA, R in NTP, A2 in ACGIH, and 2A in JSOH.

The criteria of PRTR-MSDS [53] were adopted to solve the confusion of reliability rankings. First, the original differing DB ranks were all translated into five ranks (I–V). Next, reliability ranks in various DBs were unified into five ranks (A–D, and F) according to the criteria [53]. In cases when two ranks or more were given to identical chemicals, the highest one was adopted; formaldehyde was ranked A. Negative chemicals were also collected from NTP, where experimental data via animal tests (male and female rats and mice) were accumulated. Chemicals showing at least one instance of negative carcinogenicity were ranked E. The Carcinogenicity Reliability Database (CRDB) was constructed which consists of experimental carcinogenicity data and their reliabilities for about 1,500 chemicals, as summarized in Table 2 of the Supplementary Material.

The CRDB contains not only a diverse range of pure chemicals, but also inorganic, generic, and mixture chemicals such as arsenic compounds, asbestos, cresol, gasoline, and tar. QSAR approaches cannot be applied to inorganic, generic, and mixture chemicals because their molecular structures are indefinite. In order to select chemicals appropriate for QSAR, the following were excluded:

(1) chemicals containing elements other than H, C, N, O, P, S, and X (halogen), such as metals, salts, and inorganic compounds,
(2) mixtures such as gasoline,
(3) polymers such as poly(vinyl alcohol), and
(4) chemicals whose molecular structures are unidentified.

Of the 911 chemicals thus selected, there are 409 positives (ranks A–C), and 502 negatives (ranks D–F). This ensured that positive and negative chemicals are well balanced.

Descriptors

The 3D geometries for all these 911 chemicals were created from 2D structures using the Corina program [54,55]. The 1,504 molecular descriptors listed in Table 1 were calculated using the Dragon software (professional version 5.4) [56] by inputting 3D structures. Chemical names, CAS numbers, carcinogenicity reliability ranks, and Dragon descriptors for the 911 chemicals are available in Table 3 of the Supplementary Material. These descriptors belong to various classes, but there are too many descriptors to model the 911 chemicals; therefore, effective descriptors have to be selected for QSAR analysis. Variable selection is an important process in any statistical analysis, including QSAR. Various methods have

Table 1 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.)
	1504	

ND number of descriptors

been proposed: stepwise forward or backward selection, simulated annealing, the Monte Carlo method, genetic or evolutionary algorithms, modified particle swarm optimization, and artificial ant colony system. In this study, correlation coefficients (CCs) between carcinogenicities and descriptors were calculated, and descriptors with higher (in absolute value) CCs were selected. This method may not necessarily provide the set of most effective descriptors, but it had to be adopted due to its rapidness since this process was repeated many times in optimizing SVM models.

SVM modeling

The support vector classification (SVC) function in the LIBSVM program (version 2.89) [57] was used to classify the carcinogenicity of chemicals into two categories (positive or negative), and the radial basis kernel function (RBF) was employed. Target values of carcinogenicity in SVM were set as 1.0 and −1.0 for positive and negative chemicals, respectively. Weights of the carcinogenicity data were set as 1.0 for the ranks A and F, 0.5 for B and E, and 0.25 for C and D, and then SVC for weighted data [58] was applied.

Three parameters have to be determined to optimize SVM models: the number of descriptors, gamma (g), and cost (c). A single cross-validation (CV) has often been employed in many QSAR studies: whole data sets are divided into training and test sets, with the former used for optimizing the model and the latter for evaluating its performance. Predicted values in such a single CV are obtained only for the test set, not for the whole set. The following dual CV procedure was adopted in this study since it was necessary to predict the carcinogenicity for all the chemicals:

(1) all chemicals were divided into 10 subsets (A to J),
(2) first, the subset A was assigned to the test set, and the training set was created by combining the remaining nine subsets,
(3) the SVM model was optimized by applying the leave-one-out (LOO) CV to the training set,
(4) the carcinogenicity of the test set A was predicted by applying the optimized model obtained in the previous step to the test set A,
(5) next, the subset B was assigned to the test set by exchanging the subsets A and B,
(6) the SVM model was optimized, and the carcinogenicity of the test set B was predicted,
(7) the carcinogenicities of all the chemicals were predicted by repeating the above procedures to the remaining subsets (C–J), and,
(8) the model performance was evaluated by counting the correctly predicted chemicals.

Various measures to evaluate the performance of a QSAR model have been employed in the literature: accuracy, complexity, precision, recall, area under the receiver operating

Table 2 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 descriptor	NT	NP	NN	SS	Explanation
nAB	530	226	304		Number of aromatic bonds
nArCL	90	39	51		Number of chlorine atoms on aromatic ring
nArCO	27	11	16		Number of ketones (aromatic)
nArCOOH	10	0	10	−−	Number of carboxylic acids (aromatic)
nArCOOR	17	5	12		Number of esters (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)
nArOH	66	25	41		Number of aromatic hydroxyls
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
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-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
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)
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

Table 2 continued

Atom and functional group count descriptor	NT	NP	NN	SS	Explanation
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)
nImidazoles	20	10	10		Number of imidazoles
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=N	22	9	13		Number of N azo-derivatives
nN-N	19	10	9		Number of N hydrazines
nO	610	256	354		Number of oxygen atoms
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
nP	55	11	44	−−	Number of phosphorous atoms
nPO4	36	7	29	−−	Number of phosphates/thiophosphates
nPyridines	32	17	15		Number of pyridines
nPyrrolidines	10	5	5		Number of pyrrolidines
nR=CHX	10	7	3		Number of R=CHX
nR=Cp	59	34	25	++	Number of terminal primary C(sp2)
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
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
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)
nRCO	36	16	20		Number of ketones (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)
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)

Table 2 continued

Atom and functional group count descriptor	NT	NP	NN	SS	Explanation
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
nROR	80	42	38		Number of ethers (aliphatic)
nRSR	18	3	15	−−	Number of sulfides
nRX	172	100	72	++	Number of halogen atoms (aliphatic)
nS	110	41	69		Number of sulfur atoms
nSK	911	409	502		Number of non-H atoms
nSO2N	16	5	11		Number of sulfonamides (thio-/dithio-)
nTB	19	8	11		Number of triple bonds
nX	261	138	123	++	Number of halogen atoms
ArHC[a]	54	18	36		
XHC[b]	85	55	30	++	

NT number of chemicals (groups whose members are below 10 are omitted with some exceptions), *NP* number of positives, *NN* number of negatives, *SS* statistical significance for P/N ratio, ++ positive rich, −− negative rich. [a]Aromatic hydrocarbons counted as nCar > 0 and nN = nO = nP = nS = nX = 0; [b] 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

characteristic curve (AUROC), sensitivity, specificity, and the Matthews correlation coefficient (MCC). Two measures were used in this study, overall accuracy (OA) and AUROC, calculated as

$$\mathrm{OA} = \frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}}; \tag{1}$$

$$\mathrm{AUROC} = \frac{1}{2}\left(\frac{\mathrm{TP}}{\mathrm{TP+FN}} + \frac{\mathrm{TN}}{\mathrm{TN+FP}}\right), \tag{2}$$

where TP, TN, FP, and FN are the numbers of true positive, true negative, false positive, and false negative chemicals, respectively.

Results and discussion

Correlation between carcinogenicity and substructure

It is well known that some substructures are responsible for carcinogenesis, as described above. It is interesting to examine whether the results obtained here correspond to such knowledge of carcinogens, that is, whether those substructures are contained in significantly more positive chemicals than negatives, or extremely not contained in negative chemicals. For this purpose, the numbers of positives and negatives containing various substructures were counted by using atom and functional group count descriptors obtained with the Dragon program. It was then statistically judged whether the ratio of the number of positives to negatives for each substructure significantly differs from the whole ensemble ratio (P/N = 409/502) [1]. If the P/N ratio for some substructure was judged to be significantly higher than the whole ensemble ratio, it was regarded as strongly related to carcinogenesis, and called positive rich. If the ratio was lower, the substructure was regarded as strongly unrelated to carcinogenesis, and called negative rich. The result is summarized in Table 2, where the substructures judged as positive and negative rich are marked by ++ and −−, respectively, in the column SS. As is well known, chemical structures are multifunctional, and substructures are non-exclusive. Therefore, it should be noticed that chemicals containing any of the substructures listed in this table also contain some of the other substructures.

Several substructures were judged to be significantly positive or negative rich as compared with the whole ensemble, but there are very few positive rich groups. Aromatic amine (ArNH2, ArNHR, and ArNR2) and aromatic nitro (ArNO2) groups have been classified as strong carcinogens as stated above, and QSAR approaches on these congeneric chemicals have succeeded in modeling the carcinogenicity. However, these groups were not judged to be positive rich in this study. Aromatic nitroso (ArNO) and aromatic N-nitroso (ArNNOx) groups are also known as quite strong carcinogens, but they were not judged to be positive rich because there are only two

[1] According to the statistics theory, if the ratio of positives in a group is greater than $p_0 + p_1$, the group is judged as significantly positive rich at the significance level of 0.05 as compared with the whole ensemble, where p_0 is the ratio of positives in the whole ensemble, given in this case by $p_0 = 409/911 = 0.449$, and p_1 is given by $p_1 = 1.96 * [(409/911) * (502/911)/n]^{1/2} = 0.975/n^{1/2}$ where n is the size of the group.

chemicals, too few to be statistically judged. Only aliphatic N-nitroso (RNNOx) group was judged to be clearly positive rich. Aromatic chlorine (ArCL) and aromatic halogen (ArX) groups were not judged to be positive rich, while their aliphatic relatives (RCL and RX) were positive rich. Aromatic hydrocarbons (ArHC), and halohydrocarbons (XHC) were introduced as exclusive groups to examine strong carcinogens such as benzene, benzo(a)pyrene, chloroform, and chloroethylenes. Their numbers were counted as $nCar > 0$ and $nN = nO = nP = nS = nX = 0$ for ArHC, and $nX > 0$ and $nN = nO = nP = nS = 0$ for XHC. XHC was judged to be positive rich, but ArHC was not, irrespective of strong carcinogens such as benzene and benzo(a)pyrene.

It is summarized from these results that

(1) most of substructures which are known to be responsible for carcinogenesis are not only contained in positive chemicals, but also in negative chemicals,
(2) very few substructures (aliphatic N-nitroso and aliphatic halogen groups) are contained in significantly many more positive chemicals,
(3) these go against our knowledge of carcinogenic structures as revealed from biomedical studies, and therefore,
(4) many factors beyond responsible substructures may contribute to the carcinogenesis of chemicals.

Batch model on overall chemicals

A single SVM was applied to modeling all the 911 chemicals together, making it the first model for predicting the carcinogenicity across a wide variety of chemicals. CCs between carcinogenicities and descriptors were calculated for all the chemicals, and the model was optimized by choosing the number of descriptors that gave the highest performance. The accuracy changes slightly with the number of descriptors, as shown in Fig. 1, and the highest performance, OA = 0.688 and AUROC = 0.683 with TP = 258, TN = 369, FP = 133, and FN = 151, was obtained when 250 descriptors were used. The accuracy is comparable to the highest one in the PTC contest, which may be an upper limit in predicting the carcinogenicity of non-congeneric chemicals with a single QSAR model. The performance of this batch model (68–69%) is quite low, and the rate of FN in particular is quite high. Therefore, this model cannot be accepted for predicting the carcinogenicity of chemicals as a screening for animal tests.

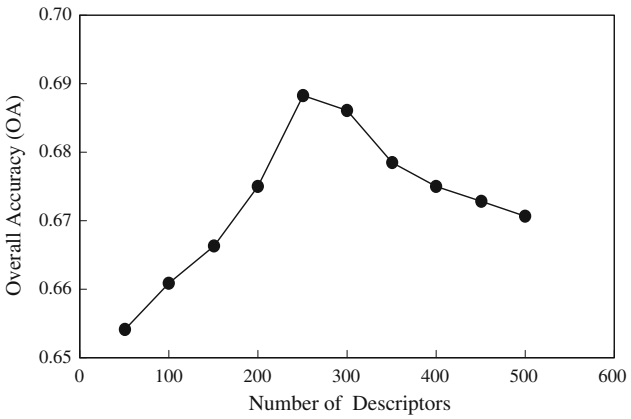

Fig. 1 The result of the batch model on overall chemicals

Substructure grouping model

Next, SVM models were created and trained on individual subgroups according to the idea of ensemble learning (EL), with the hope of obtaining improved prediction performance. Three points must be considered for selecting subgroups. The first is the choice of what types of subgroups to adopt. Training subsets in many EL studies are usually selected by random extraction from overall data sets. However, in this study, candidate subgroups were selected from the substructures listed in Table 2 for the following reasons. Mechanisms of chemical carcinogenesis are strongly related to molecular structures, and models on congeneric chemicals containing the same substructures might possibly present a satisfactory performance, as stated above. It was also stressed in the literature that QSAR studies on carcinogenicity should be based on the mechanisms of carcinogenesis [18]. Unfortunately, there are too few carcinogens whose biological mechanisms have been elucidated [22]; there are too few data to be applied to modeling the carcinogenicity of a wide variety of chemicals. Therefore, it was assumed that cancer will be caused due to the same mechanism in chemicals containing the same substructures, and on the basis of these reasons, candidate subgroups were selected from the substructures listed in Table 2.

The second choice is which substructures should be adopted among the candidates listed. There are two to over 900 chemicals containing various substructures listed in Table 2. The generality of a model conflicts with its performance in any ML method: the more training data there are, the lower the performance is, and vice versa. Therefore, it is necessary to determine the target performance and the data size for selecting substructures. In this study, first the target performance of the model was set at 80% or higher prior to the generality, because the highest accuracy of existing ones on non-congeneric chemicals is around 70%, as stated above. Next, the criterion on the member size of substructures to adopt was determined so as to achieve this performance. Preliminary trials on member size showed that if a substructure group with fewer than 50 members is trained, the model will be less robust, while if a group with more than 200 members is trained, the performance will be lower than 80%. Therefore, substructures with 50–200 members were

selected (with some exceptions) as a criterion for selecting candidates.

The third choice is how to combine SVM models that have been trained individually on selected substructures and obtain the predicted carcinogenicities of all the chemicals. As is well known, chemical structures are multifunctional, and substructures listed in Table 2 are non-exclusive: on an average, one chemical contains about two substructures, as described later. There are two models, parallel and serial, to combine models for subgroups. A parallel model, which is analogous to the boosting or bagging technique in EL, was adopted in this study: SVM models were created for selected substructures and trained in parallel, and the carcinogenicity of a test chemical was determined by averaging the output values of models corresponding to substructures contained in the test chemical.

Many trials were carried out to create effective subgroups, by combining or decomposing candidate substructures; ultimately, the 20 subgroups listed in Table 3 were found to satisfy the above conditions: all the chemicals treated here contain any of these substructures, and the overall performance for all subgroups reached the target value of 80%. This table also lists many substructures that were tested but not adopted due to low performance. The total number of chemicals containing the adopted substructures (1,488) is 1.6 times larger than the total number of all chemicals treated here (911), meaning that one chemical contains an average of 1.6 substructures. Among the 911 chemicals, 543 chemicals contain only one substructure, while four chemicals contain at most six substructures, indicating the degree of multi-functionality of chemical structures.

In order to evaluate the performance of this model across all the chemicals, output values of corresponding SVMs were collected for individual chemicals, and the predicted carcinogenicity of each chemical was determined by the majority of those outputs. When the number of models giving true estimate is equal to the number of those giving false one, TP and FN were counted as 0.5 for a carcinogenic chemical, and TN and FP were counted as 0.5 for a non-carcinogenic chemical. Consequently, the overall performance on all the chemicals reached the target 80% (OA = 0.796 and AUROC = 0.793), with TP = 313.5, TN = 411.5, FP = 90.5, and FN = 95.5. This predictability is higher than any of previously proposed models for non-congeneric chemicals, and the rate of FN is especially low: therefore, the present model can be accepted as a screening for animal tests. In order to examine the diversities of chemicals containing the adopted substructures, minimal, maximal, mean, and standard deviation values of the numbers of carbon atoms, molecular weights, and CCs between descriptors were calculated, which are listed in Table 4 of the Supplementary Material. It is understood from these data that each subgroup covers a diverse range of chemicals. Therefore, it is expected that this model will present a satisfactory predictability for chemicals of a wide variety of structures.

Comparison of performances of models

The highest reported accuracy of previously developed models for non-congeneric chemicals was about 70% in the PTC contest [20,21]. It is important to consider the reasons why the present model outperforms any of existing ones. The primary reason is that the present model is based on SVMs optimized not for all the chemicals together, but separately for individual subgroups; this is confirmed by the numbers of effective descriptors used in SVM modeling. In the batch model, where all the chemicals were analyzed together with a single SVM as in most PTC models, there were very few effective descriptors. The highest (in absolute value) CC between descriptors and carcinogenicities in the batch model is –0.182 for the descriptor G2 (gravitational index (bond-restricted)), while the highest CC in the substructure grouping model is 0.603 for FDI (folding degree index) in the aromatic hydroxyl derivatives group, as summarized in Table 5 of the Supplementary Material. In addition, there are only 110 descriptors with CCs above 0.1 (in absolute value) in the batch model, while in the substructure grouping model, there are, on an average, 709 descriptors with CCs above 0.1, and 59 descriptors with CCs above 0.3, as shown in Fig. 2.

These differences can be understood from the above data of the CCs between descriptors and carcinogenicity. For example, in the substructure grouping model, the descriptor CIC1 (complementary information content (neighborhood symmetry of 1-order)) negatively contributes to carcinogenicity in aliphatic ethers (CC = − 0.354), amides and carbamates (CC = −0.228), and urea derivatives (CC = − 0.173), but positively in aromatic hydroxyl derivatives (CC = 0.334), aromatic hydrocarbons (CC = 0.280), and ketones (CC = 0.266). Consequently, its contribution to all chemicals in the batch model is quite low (CC = –0.025). This is the case for many other descriptors as well, which suggests that some chemical factors expressed with these descriptors are related to carcinogenesis in a very different manner depending on substructures. Therefore, it is reasonable to conclude that separate modeling on individual substructures is quite effective, and yields a high prediction performance. Such an approach corroborates the recommendation that carcinogenicity modeling should be based on the mechanisms of chemical carcinogenesis [18].

These results may also suggest that the separate modeling of individual subgroups adopted here, similar to EL, performs better than the modeling of randomly extracted subgroups used in many EL studies, because there might not be effective descriptors with so high CC in the latter model as those of the former model based on chemical substructures. The parallel analysis of substructures presented here is effective

for QSAR modeling, especially for predicting the carcinogenicity of a wide variety of chemicals, because of the strong correlation between molecular structures and carcinogenic mechanisms, as described above.

The substructure grouping model developed in this study reached an overall performance of 80%, which is acceptable for predicting the carcinogenicity of a wide variety of chemicals as a screening method for animal tests. However, there could still be room to improve the performance and stability of the model. In particular, aromatic nitro, nitroso, and N-nitroso group showed the lowest accuracy (OA = 70.8%) among the adopted substructures listed in Table 3, which causes a decline in overall performance. Other N-containing groups including amides and carbamates (OA = 74.6%), aromatic amines and imines (OA = 73.0%), and N hydrazines and azo-derivatives (OA = 73.2%) also performed poorly. Many attempts were carried out to raise the performance for these N-containing groups; however, the improvement could not be achieved, as shown in the un-adopted substructures of Table 3.

The reason for the low performance is unclear; however, some factors beyond the Dragon descriptors may contribute to the carcinogenicity of these N-containing groups. The most likely among them is an electronic factor: electron densities of nitrogen atoms in these molecules are easily affected by other coexisting functional groups, especially in aromatic compounds, and consequently, the reversal of carcinogenicity occurs. Many types of Dragon descriptors listed in Table 1 were used in this study, but descriptors reflecting such electronic factors are not included. This may cause the low performances of these subgroups; therefore, the introduction of electronic descriptors including quantum-chemical quantities will possibly improve the performance of the model presented here for a wide variety of chemicals.

The stability of the above model could also be improved. There are not necessarily enough numbers of chemicals containing the adopted substructures to create robust models, as seen in Table 3. For example, there are only 34 chemicals containing aliphatic nitro, nitroso, and N-nitroso group, and

Table 3 Result of the substructure grouping model

	Condition for Dragon descriptors	G	C	ND	NC	NP	NN	TP	TN	OA	AUC
Adopted substructure group											
Amides and carbamates	nRCONH2, nArCONH2, nRCONHR, nArCONHR, nRCONR2, nArCONR2, nROCON, nArOCON > 0	0.2	20	60	71	30	41	22	31	0.746	0.745
Amines and imines (Aliphatic)	nRNH2, nRNHR, nRNR2, nRC=N > 0	0.02	50	20	62	20	42	11	39	0.806	0.739
Amines and imines (Aromatic)	nArNH2, nArNHR, nArNR2, nArC=N > 0	0.5	10	100	141	64	77	35	68	0.730	0.715
Carboxylic acids	nRCOOH, nArCOOH > 0	0.01	20	40	46	14	32	8	28	0.783	0.723
Esters	nRCOOR, nArCOOR > 0	0.01	50	50	85	26	59	19	54	0.859	0.823
Ethers (Aliphatic)	nROR > 0	0.01	50	55	80	42	38	36	31	0.838	0.836
Ethers (Aromatic)	nArOR > 0	0.05	50	65	80	40	40	35	28	0.788	0.788
Hydroxyl derivatives (Aliphatic)	nROH > 0	0.05	10	120	134	44	90	26	79	0.784	0.734
Hydroxyl derivatives (Aromatic)	nArOH > 0	0.1	10	60	66	25	41	18	34	0.788	0.775
Ketones	nRCO, nArCO > 0	0.5	5	50	58	24	34	18	32	0.862	0.846
N-containing heteroaromatics	n135-Triazines, nImidazoles, nIsoxazoles, nPyrazines, nPyridines, nPyrimidines, nPyrroles, nPyrrolidines, nThiazoles, nTriazoles > 0	0.02	50	75	88	37	51	34	41	0.852	0.861

Table 3 continued

	Condition for Dragon descriptors	G	C	ND	NC	NP	NN	TP	TN	OA	AUC
N hydrazines and N azo-derivatives	nN-N, nN=N > 0	0.02	2	35	41	19	22	15	15	0.732	0.736
Nitro, nitroso and N-nitroso compounds (Aliphatic)	nRNNOx, nRNO, nRNO2 > 0	0.5	5	30	34	25	9	20	6	0.765	0.733
Nitro, nitroso and N-nitroso compounds (Aromatic)	nArNNOx, nArNO, nArNO2 > 0	0.05	50	75	89	41	48	31	32	0.708	0.711
Phosphorous compounds	nP > 0	0.1	20	50	55	11	44	6	41	0.855	0.739
Sulfur compounds	nS > 0	0.2	2	85	110	41	69	26	62	0.800	0.766
Urea derivatives	nCONN > 0	0.2	20	40	47	24	23	21	19	0.851	0.851
Aromatic hydrocarbons	nN = nO = nP = nS = nX = 0 and nCar > 0	0.5	10	35	54	18	36	16	31	0.870	0.875
Halohydrocarbons	nN = nO = nP = nS = 0 and nX > 0	0.2	5	75	85	55	30	53	18	0.835	0.782
Others		0.1	50	55	62	26	36	20	29	0.790	0.787
Total					1, 488	626	862	470	718	0.798	0.792
Un-adopted substructure group											
Aldehydes and ketones	nRCHO, nArCHO, nRCO, nArCO > 0	0.1	50	65	74	33	41	24	30	0.730	0.729
Amides	nRNH2, nRNHR, nRNR2, nArNH2, nArNHR, nArNR2 > 0	0.2	10	100	47	18	29	10	10	0.617	0.500
Amides, carbamates and urea derivatives	nRCONH2, nArCONH2, nRCONHR, nArCONHR, nRCONR2, nArCONR2, nROCON, nArOCON, nCONN > 0	0.05	20	105	117	54	63	41	46	0.744	0.745
Amines	nRNH2, nRNHR, nRNR2, nArNH2, nArNHR, nArNR2 > 0	0.2	5	245	193	82	111	40	91	0.679	0.654
Aliphatic amines	nRNH2, nRNHR, nRNR2 > 0	0.05	50	25	61	20	41	9	37	0.754	0.676
Aromatic amines	nArNH2, nArNHR, nArNR2 > 0	0.5	50	100	134	63	71	38	56	0.701	0.696
Chlorine compounds	nCL > 0	0.05	50	310	231	121	110	88	77	0.714	0.714
Aliphatic chlorine compounds	nRCL > 0	0.1	20	185	149	86	63	75	43	0.792	0.777
Aromatic chlorine compounds	nArCL > 0	0.5	5	50	90	39	51	24	35	0.656	0.651
Ethers	nROR, nArOR > 0	0.05	50	235	148	72	76	54	49	0.696	0.697
Aliphatic halogen compounds	nRX > 0	0.1	50	210	172	100	72	82	48	0.756	0.743
Aromatic halogen compounds	nArX > 0	0.1	10	100	98	43	55	28	38	0.673	0.671
Hydroxyl derivatives	nROH, nArOH > 0	0.2	5	245	184	61	123	30	111	0.766	0.697
Nitro Compounds	nRNO2, nArNO2 > 0	0.05	100	80	92	43	49	32	32	0.696	0.699

Table 3 continued

	Condition for Dragon descriptors	G	C	ND	NC	NP	NN	TP	TN	OA	AUC
Aromatic nitro compounds	nArNO2 > 0	0.1	50	125	86	40	46	28	31	0.686	0.687
N-nitroso, nitroso and nitro compounds	nRNNOx, nRNO, nRNO2, nArNNOx, nArNO, nArNO2 > 0	0.01	100	190	123	66	57	55	28	0.675	0.662

G gamma parameter in optimized SVM model, *C* cost parameter in optimized SVM model, *ND* number of descriptors in optimized SVM model, *NC* number of chemicals, *NP* number of positives, *NN* number of negatives, *TP* number of true positives, *TN* number of true negatives, *OA* overall accuracy, *AUC* area under the receiver operating characteristic curve (AUROC)

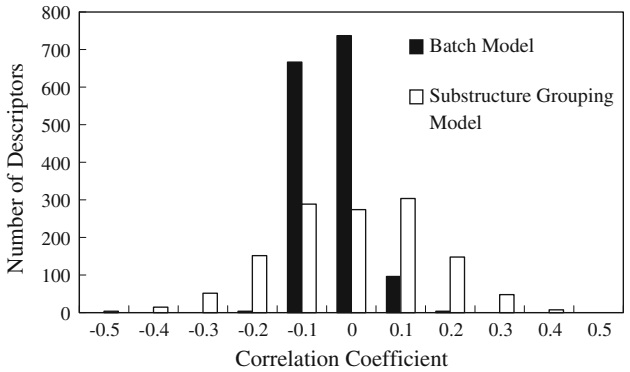

Fig. 2 Histogram of correlation coefficients between carcinogenicities and descriptors in batch model and substructure grouping model. For the latter model, counts averaged for all subgroups are shown

this is the case for other substructures as well. In order to construct more robust models, experimental carcinogenicity data must be collected on many more chemicals utilizing QSAR approaches like the one presented above.

Several articles concerning QSAR analysis of chemical carcinogenicity have been published very recently [59–64], demonstrating the importance of the subject. The practice of a screen for animal tests is very important, and can be achieved utilizing QSAR models to predict the carcinogenicity of various chemicals existing in the environment that are not experimentally tested. For that purpose, the compilation of experimental carcinogenicity data and the development of a carcinogenicity prediction model for arbitrary chemicals must be promoted with international cooperation [65].

Conclusions

In order to construct a QSAR model to predict the carcinogenicity of a wide variety of chemicals with a higher predictability than any existing models, the relationship between carcinogenicities and selected descriptors was analyzed with SVM models. The carcinogenicity data were collected from six sources, and their reliabilities were ranked into six unified categories. About 1,500 types of molecular descriptors of 911 organic chemicals were calculated with the Dragon software. The positive and negative chemicals containing various substructures were counted by using atom and functional group count descriptors. Very few substructures were statistically judged as strongly related with carcinogenicity, in contrast to our knowledge of chemical structures of carcinogens. Based on the ensemble learning technique, the 911 chemicals were classified into 20 subgroups according to contained substructures, and SVM models were optimized for each subgroup. Consequently, this model was found to predict the carcinogenicity of a wide variety of chemicals with an overall performance of 80%.

Acknowledgements This study was supported in part (study done in this article by K.T.) out of grants awarded by the Ministry of Education, Science, Sports and Culture of Japan, Grant-in-Aid for 14209022, from the Conflex Corporation, and in part (study done by B.L. and D. A.) awarded by the Ministry of Science, Education and Sports of the Republic of Croatia.

References

1. 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
2. Harvard Center for Cancer Prevention (1996) Harvard report on cancer prevention. Volume 1: Causes of human cancer. Cancer Causes Control 7:S3–S59. doi:10.1007/BF02352719
3. 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. doi:10.1080/10629360008039117
4. 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
5. Patlewicz G, Rodford R, Walker JD (2003) Quantitative structure–activity relationships for predicting mutagenicity and carcinogenicity. Environ Toxicol Chem 22:1885–1893. doi:10.1897/01-461
6. Benigni R (2004) Prediction of human health endpoints: mutagenicity and carcinogenicity. In: Cronin MTD, Livingstone DJ (eds) Predicting chemical toxicity and fate. CRC Press, Boca Raton, pp 173–192
7. Sun H (2004) Prediction of chemical carcinogenicity from molecular structure. J Chem Inf Comput Sci 44:1506–1514. doi:10.1021/ci049917y

8. 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. doi:10.1021/ci050150z
9. Helguera AM, Perez MCA, Combes RD, Gonzalez MP (2005) The prediction of carcinogenicity from molecular structure. Curr Comp Aid Drug Des 1:237–255
10. Contrera JF, MacLaughlin P, Hall LH, Kier LB (2005) QSAR modeling of carcinogenic risk using discriminant analysis and topological molecular descriptors. Curr Drug Discov Tech 2: 55–67. doi:10.2174/1570163054064684
11. Benigni R, Bossa C (2008) Predictivity of QSAR. J Chem Inf Model 48:971–980. doi:10.1021/ci8000088
12. Benigni R, Giuliani A, Franke R, Gruska A (2000) Quantitative structure–activity relationships of mutagenic and carcinogenic aromatic amines. Chem Rev 100:3697–3714. doi:10.1021/cr9901079
13. 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
14. Benigni R, Giuliani A, Gruska A, Franke R (2003) 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
15. Vendrame R, Braga RS, Takahata Y, Galvao DS (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. doi:10.1021/ci990326v
16. Braga RS, Barone PMVB, Galvao DS (1999) Identifying carcinogenic activity of methylated polycyclic aromatic hydrocarbons (PAHs). J Mol Struct 464:257–266. doi:10.1016/S0166--1280(98)00557-0
17. Zhou Z, Dai Q, Gu TA (2003) QSAR model of PAHs carcinogenesis based on thermodynamic stabilities of bioactive sites. J Chem Inf Comput Sci 43:615–621. doi:10.1021/ci0256135
18. Benigni R (2003) 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
19. Benigni R (2005) Structure–activity relationship studies of chemical mutagens and carcinogens: Mechanistic investigations and prediction approaches. Chem Rev 105:1767–1800. doi:10.1021/cr030049y
20. Helma C, King RD, 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)
21. Helma C, Kramer S (2003) A survey of the predictive toxicology challenge 2000–2001. Bioinformatics 19:1179–1182
22. Ivanciuc O (2009) Drug design with machine learning. In: Meyers RA (ed) Encyclopedia of complexity and system science. Springer-Verlag, New York
23. Svetnik V, Wang T, Tong C, Liaw A, Sheridan RP, Song Q (2005) Boosting: an ensemble learning tool for compound classification and QSAR modeling. J Chem Inf Comput Sci 45:786–799. doi:10.1021/ci0500379
24. 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. doi:10.1021/ci7003384
25. Langham JJ, Jain AN (2008) Accurate and interpretable computational modeling of chemical mutagenicity. J Chem Inf Model 48:1833–1839. doi:10.1021/ci800094a
26. Liu T-Y, Li G-Z, Yang JY, Yang MQ (2008) Feature selection for the imbalanced QSAR problems by using EasyEnsemble. Int J Comput Biol Drug Design 1:334–346. doi:10.1504/IJCBDD.2008.022206
27. Woo Y-T, Lai DY (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
28. Devillers J (1996) Neural networks in QSAR and drug design. Academic Press, San Diego
29. Zupan J, Gasteiger J (1999) Quantitative structure–activity relationships. In: Zupan J, Gasteiger J (eds) Neural networks in chemistry and drug design, 2nd edn. Weinheim, Wiley-VCH, pp 219–242
30. Peterson KL (2000) Artificial neural networks and their use in chemistry. In: Lipkowitz KB, Boyd DB (eds) Reviews in computational chemistry . Wiley-VCH, New York pp 53–140
31. Ivanciuc O (2009) Drug design with artificial neural networks. In: Meyers RA (ed) Encyclopedia of complexity and system science. Springer-Verlag, New York
32. Basak SC, Grunwald GD, Gute BD, 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. doi:10.1021/ci9901136
33. 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. doi:10.1021/ci990116i
34. 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. doi:10.1021/ci049766z
35. Devillers J (1996) 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
36. 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. doi:10.2477/jccj.4.89
37. Chen N, Lu W, Yang J, Li G (eds) (2004) Support vector machine in chemistry. World Scientific, Singapore
38. Ivanciuc O (2007) Applications of support vector machines in chemistry. Rev Comput Chem 23:291–400. doi:10.1002/9780470116449.ch6
39. 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. doi:10.1021/ci0341161
40. Yao XJ, Panaye A, Doucet JP, Zhang RS, Chen HF, Liu MC, Hu ZD, 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. doi:10.1021/ci049965i
41. 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. doi:10.1021/ci034254q
42. Xue Y, Li ZR, Yap CW, Sun LZ, Chen X, Chen YZ (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. doi:10.1021/ci049869h
43. Chen N, Lu W, Yang J, Li G (2004) SVM applied to structure–activity relationships. In: Chen N, Lu W, Yang J, Li G (eds)

Support vector machine in chemistry. World Scientific, Singapore pp 186–219
44. Jorissen RN, Gilson MK (2005) Virtual screening of molecular databases using a support vector machine. J Chem Inf Comput Sci 45:549–561. doi:10.1021/ci049641u
45. 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. doi:10.1021/ci0601281
46. Bruce CL, Melville JL, Pickett SD, Hirst JD (2007) Contemporary QSAR classifiers compared. J Chem Inf Model 47:219–227. doi:10.1021/ci600332j
47. 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. doi:10.1021/ci700047x
48. 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. doi:10.2174/157340907782799372
49. 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. doi:10.2477/jccj.H1921
50. Ivanciuc O (2002) Support vector machine classification of the carcinogenic activity of polycyclic aromatic hydrocarbons. Internet Electron J Mol Design 1:203–218
51. 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. doi:10.1021/tx049782q
52. Japan Chemical Industry Ecology–Toxicology and Information Center (2007) Estimation and classification criteria of carcinogenicity of chemical substances. JETOC, Tokyo, pp 21–23
53. Urano K (2001) Toxicity ranks and physical property information for PRTR–MSDS chemical substances, Chap 2. In: Rank of carcinogenicity. Kagaku Kogyo Nippo, Tokyo, pp 21–23
54. 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. doi:10.1021/ci960343+
55. Oellien F, Nicklaus MC. (2009) Online SMILES Translator and Structure File Generator: http://cactus.nci.nih.gov/services/translate/ (accessed July 17, 2009)
56. Todeschini R, Consonni V (2006) DRAGON Professional 5.4 program, TALETE srl, Milano, Italy, (http://www.talete.mi.it/dragon.htm)
57. Chang CC, Lin CJ (2009) LIBSVM–A library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvm/ (accessed May 25, 2009)
58. Chang CC, Lin CJ (2009) LIBSVM–A library for support vector machines. http://www.csie.ntu.edu.tw/~cjlin/libsvmtools/#14 (accessed May 25, 2009)
59. Toropov AA, Toropova AP, Benfenati E, Manganaro A (2009) QSAR modelling of carcinogenicity by balance of correlations. Mol Div 13:367–373. doi:10.1007/s11030-009-9113-4
60. 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 Divers. doi:10.1007/s11030-009-9190-4
61. Toropov AA, Toropova AP, Benfenati E (2009) Additive SMILES-based carcinogenicity models: probabilistic principles in the search for robust predictions. Int J Mol Sci 10:3106–3127. doi:10.3390/ijms10073106
62. Tan NX, Rao HB, Li ZR, Li XY (2009) Prediction of chemical carcinogenicity by machine learning approaches. SAR QSAR Environ Res 20:27–75. doi:10.1080/10629360902724085
63. Venkatapathy R, Wang CY, Bruce RM, 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. doi:10.1016/j.taap.2008.09.028
64. Guyton KZ, Kyle AD, Aubrecht J, Cogliano VJ, Eastmond DA, Jackson M, Keshava N, Sandy MS, Sonawane B, Zhang L, Waters MD, Smith MT (2009) Improving prediction of chemical carcinogenicity by considering multiple mechanisms and applying toxicogenomic approaches. 1. Mutat Res 681:230–240
65. Benfenati E, Benigni R, De Marini DM, Helma C, Kirkland D, Martin TM, Mazzatorta P, Ouédraogo-Arras G, Richard AM, Schilter B, Schoonen WGEJ, Snyder RD, 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. doi:10.1080/10590500902885593