European Journal of Medicinal Chemistry 44 (2009) 3658–3664

Contents lists available at ScienceDirect

European Journal of Medicinal Chemistry

journal homepage: http://www.elsevier.com/locate/ejmech

Original article

Development of QSAR models for predicting hepatocarcinogenic toxicity of chemicals

Ilaria Massarelli [a], Marcello Imbriani [b], Alessio Coi [c], Marilena Saraceno [c], Niccolò Carli [d], Anna Maria Bianucci [c,*]

[a] Istituto Nazionale per la Scienza e Tecnologia dei Materiali, Via Giusti 9, 50121 Firenze, Italy
[b] Fondazione S. Maugeri, IRCCS, Via S. Maugeri, 4, 27100 Pavia, Italy
[c] Dipartimento di Scienze Farmaceutiche, Università di Pisa, Via Bonanno 6, 56126 Pisa, Italy
[d] Via Buonarroti 117/b, 55059 Viareggio (Lucca), Italy

ARTICLE INFO

Article history:
Received 25 July 2008
Received in revised form
29 January 2009
Accepted 12 February 2009
Available online 20 February 2009

Keywords:
Carcinogenic potency database
WEKA
Quantitative structure–activity relationship
Hepatocarcinogenic
Sphere-exclusion

ABSTRACT

A dataset comprising 55 chemicals with hepatocarcinogenic potency indices was collected from the Carcinogenic Potency Database with the aim of developing QSAR models enabling prediction of the above unwanted property for New Chemical Entities. The dataset was rationally split into training and test sets by means of a sphere-exclusion type algorithm. Among the many algorithms explored to search regression models, only a Support Vector Machine (SVM) method led to a QSAR model, which was proved to pass rigorous validation criteria, in accordance with the OECD guidelines. The proposed model is capable to explain the hepatocarcinogenic toxicity and could be exploited for predicting this property for chemicals at the early stage of their development, so optimizing resources and reducing animal testing.

1. Introduction

One of the most costly problem, when working in the early steps of discovery of new potential drugs, is related to the failure of candidates due to poor absorption, distribution, metabolism, elimination or toxicity (ADMET) properties. Recent studies attribute to ADMET problems over 60% of failures of drug candidates in development. The prediction of ADMET properties of a compound still represents a big challenge nowadays. In this perspective, several research groups working on Quantitative Structure–Activity relationship (QSAR) are emerging for the development of robust models capable to predict ahead of time these properties in order to prioritize compounds with greatest chance of success during drug discovery. Furthermore, these kind of approaches are also regarded with great interest as possible alternative methods to animal testing by a number of institutions, committees and centres of excellence like the European Centre for the Validation of Alternative Methods (ECVAM) [1]. These organizations concentrate their efforts on development and validation of alternative test methods that *refine*, *reduce* or *replace* animal usage, a principle that is comprised in the 'Three Rs' concept proposed by Russell and Burch [2].

The Organization for Economic Co-operation and Development (OECD) Group on (Q)SARs recently published (February 2007) a 'Guidance Document on the Validation of (Q)SAR Models' with the aim of providing guidance on how specific (Q)SAR models can be evaluated with respect to the OECD principles [3].

The guidelines described by OECD were considered during the development of the QSAR models described in this work. The selected end-point considered in this study is a well defined carcinogenicity index reported in the Carcinogenic Potency Database (CPDB) [4].

Nowadays, a lot of work has been done, in the field of mutagenicity and carcinogenicity, about predictions of these malignant toxicities. Unfortunately such an effort lead to a limited success. That is probably due to the fact that these end-points are very hard to be defined. Moreover the great variety of carcinogenicity mechanisms contributes to increase prediction failures. Many carcinogens are mutagenic, as they form covalent bonds with DNA, hence mutagens can be predicted by identifying electrophilic functional groups. In some cases metabolic transformations, that usually act by detoxifying the exogenous molecules, active the chemicals and produce a potential carcinogen (e.g. the N-oxidation of aromatic

* Corresponding author. Tel.: +390502219564; fax: +390502219605.
E-mail address: bianucci@dcci.unipi.it (Anna Maria Bianucci).

0223-5234/$ – see front matter © 2009 Elsevier Masson SAS. All rights reserved.
doi:10.1016/j.ejmech.2009.02.014

amines). Other carcinogens act as promoters, stimulating cell proliferation. Promoters are much more difficult to predict, as they have heterogeneous structures with diverse activities, and many are species specific.

Several QSAR works were performed in recent years with the aim of explaining carcinogenesis activity, shown both by focused chemical classes of compounds and by noncongeneric chemicals [5–11].

In this work, the homogeneity of the dataset was maximized by choosing a unique type of carcinogenic toxicity. In particular, only toxicity data referring to chemicals, which show hepatocarcinogenic potential, were chosen. Furthermore, only data referring to a unique specie (mouse) and sex (female) were taken into account during the selection of the initial dataset. Such kind of studies may be exploited to reduce animal testing in the field of carcinogenicity, by generating hazard data useful for, at least, preliminary risk assessment from exposure to chemicals and could be used within a battery of models for the prediction of this type of toxicity.

One of the main issues, to be faced when developing QSAR models, is represented by the validation step. In view of submitting the model obtained to a rigorous validation check, the whole available dataset of known molecules was split into training (TR) and test (TS) sets, according to a protocol which ensured optimal sampling both in the domain of molecular structure and molecular properties. Molecules are properly represented as points in a multi-dimensional space defined by molecular descriptors. Points which represent both TR and TS set molecules have to be evenly distributed within the whole descriptor space, and each point of the TS set has to be close to at least one point of the TR set. This approach ensures that the similarity principle is applied when predicting the properties of the TS set.

In this work, the rational splitting of the whole dataset into TR/TS set pairs was obtained by using a sphere-exclusion type algorithm, optimized in our lab and described elsewhere [12]. It ranks similarities among molecules before proceeding to the selection step. Similarities are calculated in terms of Euclidean distances computed on the basis of the molecular descriptors that will be subsequently exploited for model development. Molecular descriptors were computed by the CODESSA program [13], as described later in more detail.

It may be worth to point out here that the rational splitting of the whole dataset not only ensures that the TS exploited for external validation of the model contains molecules that fall into the chemical space defined by the TR set, but also defines the applicability domain of the model developed on the TR itself. As the final step, the WEKA program [14] (Waikato Environment for Knowledge Analysis) was used in order to develop the QSAR models, searched by means of many different regression algorithms. The best obtained models were then submitted to rigorous validation analysis based on several statistical parameters described in detail later on. Among them, only one model turned out to possess a very high predictive power.

2. Theoretical aspects

2.1. Rational TR/TS sets splitting

Rational splitting of the available dataset into training and test Set (TR/TS) pairs is required in order to obtain QSAR models endowed with high predictive power. Such a splitting should be performed so that points representing both TR and TS sets are properly distributed within the whole descriptor space defined by the entire dataset. The descriptor space may be defined as the multi-dimensional space where each one of the coordinate axes is associated to a molecular descriptor. Each molecule in the initial dataset is represented as a point in such a space. In this frame each point, representing a molecule of the TS set, should be close to at least one point of the TR set. This approach ensures that the similarity principle can be employed for the activity prediction of the TS set.

An optimized implementation of a sphere-exclusion type algorithm [15–18] previously obtained in our lab [12] was used in this work for rationally splitting the whole dataset into different TR/TS set pairs. A number of molecular descriptors computed for each molecule by using the CODESSA program [13] are handled by the algorithm which normalized them and subsequently calculated the Euclidean distances between all pairs of the dataset molecules, in the multi-dimensional descriptor space. Descriptors were normalized according to the following formula:

$$X_{ij}^{\mathrm{n}} = \frac{X_{ij} - X_{j,\min}}{X_{j,\max} - X_{j,\min}}$$

where X_{ij} and X_{ij}^{n} are the non-normalized and normalized j-th ($j = 1,\ldots, K$) descriptor values for compound i ($i = 1,\ldots, N$), correspondingly, and $X_{j,\min}$ and $X_{j,\max}$ are the minimum and maximum values for j-th descriptor. Thus, for descriptors, $\min X_{ij}^{\mathrm{n}} = 0$ and, $\max X_{ij}^{\mathrm{n}} = 1$.

By using different similarity thresholds, it is possible to select several TR/TS set pairs which are subsequently exploited for developing QSAR models.

It may be worth to point out here that each TR set defines a specific applicability domain (AD) where the model (developed on it) is expected to possess high predictive power. The criterion, underlying the rational molecule selection described above for TR/TS set splitting, must also be exploited in order to check if new chemical entities (NCE) or drugs, that are to be subjected to the QSAR model for property predictions, are comprised in the AD defined by the model itself. Only in this case property predictions for new molecules will be reliable.

2.2. Machine learning

A collection of machine learning algorithms for data mining tasks contained within the WEKA program package [14] was used for the selection of an optimal subset among all the calculated molecular descriptors and, subsequently, for the search of the best-performing algorithm during QSAR modeling.

As what concerns the group of descriptors to be used, it is worth to recall here that is should have a limited size, especially if multiple linear regression (MLR) is used. In this case, the ratio between the number of descriptor exploited and the available known molecules should be about 1:5. MLR calculates QSAR equations by performing standard multivariable regression calculations with multiple variables in a single equation. When using MLR, it is assumed that the variables belong to an orthogonal set, which is difficult to achieve in practice; nevertheless a poor correlation between variables is the condition ensuring the achievement of powerful predictive models. In this perspective, the number of independent variables initially considered should not be higher than one-fifth the number of known compounds in the training sets [19]. A higher ratio often leads to over-correlated equations, that in turn gives rise to poorly reliable predictions.

2.3. Statistical analysis and model validation

The criteria used for validating the obtained models rely on several statistical parameters that have been proved to ensure rigorous model validation. A detailed description is reported and discussed elsewhere [20]. Here it may be worth to only recall that conditions which have to be satisfied for the TR set are: $R^2 > 0.6$,

$q^2 > 0.5$; where R^2 is the correlation coefficient of the regression line between the predicted vs. experimental pTD_{50} and q^2 is the leave-one-out (LOO) cross-validated correlation coefficient. While the conditions that have to be satisfied for the TS set are: $R^2 > 0.6$, $0.85 < k_0 < 1.15$, $(R^2 - R_0^2)/R^2 < 0.1$; where R^2 is the correlation coefficient of the regression line between the predicted vs. experimental pTD_{50}, R_0^2 is the correlation coefficient of the same regression line forced through the origin and k_0 is the slope of this line. It means that the regression line correlating the x and y values should be as close as possible to the bisector of the axes as in the 'ideal' QSAR model. Moreover, in addition to the above mentioned parameters, originally suggested by Golbraikh and Tropsha [21], other validation techniques, mentioned in the OECD report [3] on QSAR, were considered. In particular, such a document assesses that more realistic estimate of the predictive ability (than using q^2 LOO) is obtained removing more than one compound at each step, according to the so-called leave-many-out cross-validation (LMO-CV) procedure. In LMO-CV, the dataset is split into a number of blocks (cancellation groups) defined by the user. At each step, all the compounds belonging to a block are left out from the derivation of the model. Rules for selecting the group of compounds for the test set at each step must be adopted in order to leave out each compound only one time. It is straightforward that the LOO method is equivalent to a LMO method with a number of cancellation groups equal to the number of compounds. By introducing a larger perturbation in the data set, the predictive ability estimated by LMO is more realistic than the one estimated by LOO. In this work, a 10-fold cross-validation was used. The relevant values of

Table 1
The 55 selected molecules constituting the available dataset, with their CAS number, their code assigned by CPDB (chemcode) and their pTD_{50} values. The 9 molecules of the TS set are evidenced in bold, while the remaining belong to the TR set.

	Name	CAS number	Chemcode	pTD50
1	2,3,7,8-tetrachlorodibenzo-p-dioxin	1746-01-6	tcd	8.87
2	kepone	143-50-0	kep	5.66
3	p,p′-dde	72-55-9	pde	4.49
4	2,4,5-trimethylaniline	137-17-7	5ma	4.34
5	1,2,3-trichloropropane	96-18-4	tc1	4.04
6	4,4′-thiodianiline	139-65-1	tda	3.81
7	hydrazobenzene	122-66-7	hzb	3.79
8	chloroprene	126-99-8	clr	3.69
9	2,4-diaminotoluene	95-80-7	tod	3.66
10	5-nitroacenaphthene	602-87-9	nac	3.64
11	**1,1,2,2-tetrachloroethane**	**79-34-5**	**4te**	**3.63**
12	pentachloroethane	76-01-7	5ce	3.39
13	michler's ketone	90-94-8	mke	3.38
14	**1,1,2-trichloroethane**	**79-00-5**	**2te**	**3.38**
15	chloroform	67-66-3	chf	3.36
16	ethylene thiourea	96-45-7	eth	3.34
17	sulfallate	95-06-7	veg	3.32
18	nitrofen	1836-75-5	nif	3.20
19	**5,5-diphenylhydantoin**	**57-41-0**	**dph**	**3.18**
20	**1,5-naphthalenediamine**	**2243-62-1**	**nda**	**3.14**
21	tetrachloroethylene	127-18-4	tre	3.10
22	1,3-butadiene	106-99-0	bde	3.09
23	bromodichloromethane	75-27-4	bcm	3.06
24	trifluralin, technical grade	1582-09-8	trf	2.96
25	hc blue no. 1	2784-94-3	hb1	2.95
26	tetrafluoroethylene	116-14-3	tfe	2.91
27	4,4′-oxydianiline	101-80-4	4ox	2.90
28	chlorobenzilate	510-15-6	chb	2.86
29	**5-chloro-o-toluidine**	**95-79-4**	**5ct**	**2.77**
30	**5-nitro-o-toluidine**	**99-55-8**	**nto**	**2.74**
31	hexachloroethane	67-72-1	hce	2.68
32	p-cresidine	120-71-8	pcr	2.64
33	trichloroethylene (without epichlorohydrin)	79-01-6	tcw	2.47
34	6-nitrobenzimidazole	94-52-0	nbi	2.46
35	cupferron	135-20-6	cup	2.44
36	1,1,1,2-tetrachloroethane	630-20-6	4ce	2.42
37	1,4-dichlorobenzene	106-46-7	4cb	2.33
38	1-amino-2,4-dibromoanthraquinone	81-49-2	abq	2.31
39	2-aminoanthraquinone	117-79-3	aaq	2.18
40	**trichloroethylene**	**79-01-6**	**tce**	**2.16**
41	sulfisoxazole	127-69-5	sxx	2.00
42	p,p′-ethyl-ddd	72-56-0	edd	1.99
43	2,6-dichloro-p-phenylenediamine	609-20-1	2dp	1.94
44	5-nitro-o-anisidine	99-59-2	nan	1.93
45	**4-chloro-m-phenylenediamine**	**5131-60-2**	**cmd**	**1.90**
46	1,4-dioxane	123-91-1	dio	1.88
47	salicylazosulfapyridine	599-79-1	sal	1.87
48	**4-chloro-o-phenylenediamine**	**95-83-0**	**cpd**	**1.73**
49	chloramben	133-90-4	cam	1.61
50	methylene chloride	75-09-2	myc	1.58
51	cinnamyl anthranilate	87-29-6	cna	1.53
52	2,5-dithiobiurea	142-46-1	dta	1.42
53	tetrahydrofuran	109-99-9	thf	1.23
54	d-limonene	5989-27-5	lmn	1.11
55	chloroethane	75-00-3	cle	0.93

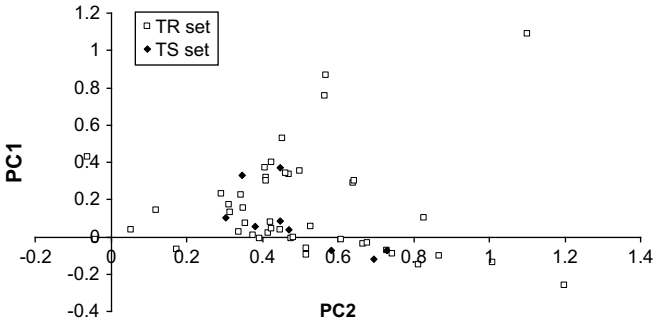

Fig. 1. Score plot for training and test sets molecules obtained after Principal Component Analysis. In the plot PC1 vs. PC2 is reported.

$R^2(R^2_{\_10fold})$ were reported in addition to the other statistical parameters considered in our previous papers. Moreover an additional technique was used to check the robustness of the QSAR model, in accordance to what is suggested by OECD, i.e. the so-called y-scrambling or response permutation test.

Such a test enables identifying models based on chance correlation, *i.e.* models where the independent variables are randomly correlated to the response variables. Y-scrambling is performed by calculating the quality of the model (usually R^2 or, better, q^2) randomly modifying the sequence of the response vector ***y***, *i.e.* by assigning to each compound a response randomly selected from the true set of responses. If the original model has no chance correlation, there is a significant difference in the quality of the original model and the ones obtained with random responses.

3. Experimental section

3.1. Dataset collection

The Carcinogenic Potency Database (CPDB) [4] is a unique and widely used international resource of results from 6153 chronic, long-term animal cancer tests on 1485 chemicals. CPDB provides a standardized and easily accessible database with qualitative and quantitative analyses of both positive and negative experiments that have been published in the general literature through 1997 and by the National Cancer Institute/National Toxicology Program through 1998. For each experiment, information is included on species, strain, and sex of test animal; features of experimental protocol such as route of administration, duration of dosing, dose level(s) in mg/Kg body weight/day, and duration of experiment; target organ, tumor type, and tumor incidence; carcinogenic potency (TD_{50}) and its statistical significance; shape of the dose-response, author's opinion as to carcinogenicity, and literature citation.

Among the several carcinogenic toxicity kinds reported in the CPDB, results taken from studies, referring to hepatocarcinogenic toxicity performed on mouse, were selected for developing the model described here. Furthermore, taking into account that TD_{50} values are reported as mg/Kg, in order to avoid artefacts, possibly coming from differences in weight between male and female mouses, a subsequent selection was made, by only considering individual of the same sex. Because of that only data referred to female (more numerous subgroup) were selected (Table 1). The whole dataset comprises 55 molecules. In order to prevent the influence of the molecular weight of the compounds during the QSAR models development, the TD_{50} values, originally expressed as mg/kg, were transformed into the corresponding μM/Kg values and submitted to the −log function, thus obtaining corresponding pTD_{50} indices (Table 1). Hence high value of pTD_{50} means high hepatocarcinogenic toxicity.

Theoretical 3D starting structures of the molecules were downloaded from ChemIDPlus [22] by supplying the relative CAS number (available from CPDB). Structures were then optimized by means of the semi-empirical quantum mechanics method implemented in the program MOPAC [23], where the AM1 Hamiltonian was used. Molecular properties enabling calculation of quantum-chemical and thermodynamic descriptors were also obtained by MOPAC. The optimized structures were subsequently loaded in the CODESSA program [13] which enabled computing 267 molecular descriptors (the ones that were found to be computable for all the molecules).

3.2. WEKA computations

The WEKA program package, version 3.5.6 was used for the selection of the best-performing algorithm and an optimal set of molecular descriptors. WEKA is a JAVA software from the University of Waikato, New Zealand [14] with an open source issued under the GNU General Public License. In this study, many algorithms available in WEKA were initially trained by using WEKA's default settings. After analysis of preliminary results, the algorithms that turned out to perform at the best (LibSVM [24–28], MultiLayerPerceptron [29] and SMOregression [30,31]) were re-trained by more finely tuning the parameter values in order to maximize the algorithm performance.

4. Results and discussion

3D structures of molecules comprised in the whole dataset were downloaded from ChemIDPlus [22] and subjected to quantum-chemical and thermodynamic calculations. CODESSA molecular descriptors were calculated for all the molecules and 267 of them were found to be shared by all the molecules. The selected descriptors were exported and subjected to an algorithm that parsed and normalized them. Subsequently similarities between pairs of molecules, among the selected 55 ones, were calculated in terms of Euclidean distances in the 267-dimensional descriptor space. A value of 2 was chosen for the similarity threshold, thus generating a TR/TS set pair consisting in 46 and 9 molecules, respectively (Table 1).

Table 2
Statistical parameters referred to the preliminary QSAR models obtained by applying MLR and some of the other regression algorithms of WEKA.

TR set statistics			TS set statistics				
R^2	q^2	$R^2_{\_10fold}$	R^2	R^2_0	$(R^2 - R^2_0)/R^2$	k	Method
0.58	0.16	0.10	0.20	−1.12	6.63	0.9861	MLR
0.45	0.21	0.09	0.08	−23.85	292.93	0.9249	Libsvm
0.71	0.40	0.18	0.12	−5.29	43.39	0.7057	Multylayerperceptron
0.58	0.25	0.20	0.19	−1.06	6.63	0.9925	PLSClassifier
0.55	0.29	0.19	0.32	−2.65	9.24	0.9213	SMOreg
0.82	0.18	0.16	0.15	−0.75	5.92	1.0221	Additive regression

Table 3
Statistical parameters referred to the proposed QSAR model.

TR set statistics			TS set statistics			
R^2	q^2	$R^2_{-10fold}$	R^2	R^2_0	$(R^2 - R^2_0)/R^2$	k
0.919	0.580	0.600	0.707	0.696	0.015	1.011

In order to determine the proper combination of attributes (the CODESSA molecular descriptors) to be used in MLR and other regression analysis, the CfsSubsetEval attribute evaluator [32] of WEKA was employed, which selected an optimal subset among all possible subsets of attributes. Using a 10-fold cross-validation on the 46 molecules of the TR set, and also using an arbitrary threshold for attribute significance, different subsets of attributes were selected. Using a threshold of ≥70%, the number of selected attributes was 11 while using a threshold of ≥80% the number of attributes was limited to 5.

The number of molecular descriptors (eleven), selected by using the 70% threshold appear to be too high in consideration of the limit suggested by MLR requirements; because of that, only the 5 attributes selected by applying the 80% threshold were considered.

This choice is also in agreement with the principle of parsimony. It suggests that, in order to obtain a robust model that is simple to be interpreted as well, a limited number of descriptors should be used. It suggests that, among results being more or less equal, the simplest model should be chosen [33].

The selected 5 molecular descriptors are listed below: (a) *number of Cl atoms*, a very simple constitutional descriptor endowed with intuitive meaning; (b) *LUMO energy*, a quantum-chemical descriptor related to the energy of the Lowest Unoccupied Molecular Orbital of the molecule; (c) *FNSA-3*, a quantum-chemical descriptor obtained as the ratio between weighted atomic charge partial negative surface area (PNSA) and total molecular surface area (TMSA); (d) *highest normal mode vibrational frequency*; and (e) *highest normal mode vibration transition dipole* . These last two descriptors refer to quantum mechanical molecular rotational–vibrational energies.

The analysis of the descriptors, identified as the most suitable ones, can give some interesting suggestions about their diagnostic power in predicting hepatocarcinogenic toxicity of chemicals, even though most of them are not easily related to simple molecular features.

For example, the descriptor accounting for the *number of Cl atoms* is in highly good agreement with data reported in literature, in which often organochlorine compounds are reported to be agents potentially able to induce many types of tumors, as colorectal, breast, ovarian and other cancers [34–36].

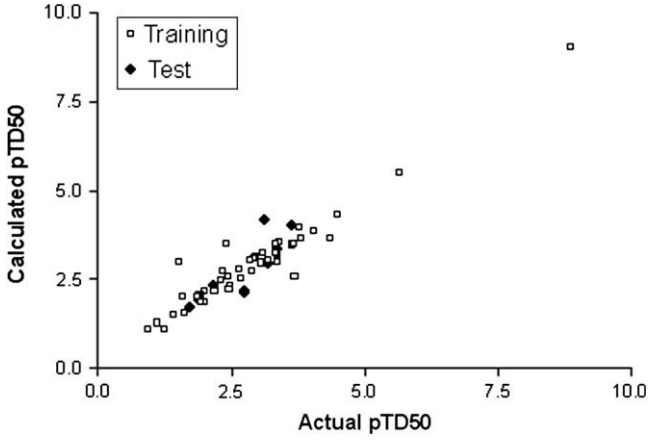

Fig. 2. Actual and calculated pTD_{50}s for the TR and TS sets molecules.

As for the *LUMO energy*, it has to be recalled that the reactivity of a molecule is related to its gap between HOMO (Lowest Unoccupied Molecular Orbital of the molecule) and LUMO energy levels. A small HOMO–LUMO energy difference indicates that the molecule is kinetically labile. Furthermore the LUMO energy is highly correlated to the onset potential of a molecule during reduction reactions and consequently to its capability to accept electrons and produce radical anions. The smaller is the LUMO energy the higher is the capability of producing radical anions. Indeed, among the molecules belonging to the dataset reported in this study, the ones

Table 4
Actual and calculated pTD_{50}s for the TR set and TS set molecules.

Molecule ID	Actual pTD50	Calculated pTD50	|Error|
1	8.867	9.026	0.159
2	5.662	5.502	0.16
3	4.492	4.332	0.16
4	4.344	3.677	0.667
5	4.041	3.88	0.16
6	3.809	3.649	0.161
7	3.793	3.953	0.16
8	3.689	2.577	1.112
9	3.66	3.5	0.16
10	3.638	3.478	0.16
12	3.39	3.536	0.146
13	3.383	3.224	0.16
15	3.361	2.98	0.381
16	3.341	3.501	0.16
17	3.32	3.239	0.082
18	3.203	3.043	0.16
21	3.1	3.261	0.161
22	3.091	2.931	0.16
23	3.056	2.96	0.097
24	2.96	3.12	0.16
25	2.953	3.098	0.145
26	2.907	3.067	0.16
27	2.9	2.74	0.16
28	2.862	3.022	0.16
31	2.683	2.522	0.16
32	2.639	2.798	0.159
33	2.47	2.31	0.16
34	2.458	2.237	0.221
35	2.439	2.564	0.124
36	2.419	3.488	1.069
37	2.328	2.723	0.395
38	2.313	2.474	0.16
39	2.185	2.164	0.021
41	2.004	1.844	0.16
42	1.988	2.148	0.16
43	1.941	1.88	0.06
44	1.93	2.09	0.16
46	1.883	2.043	0.16
47	1.866	2.026	0.16
49	1.607	1.541	0.066
50	1.581	2.022	0.441
51	1.529	3.015	1.486
52	1.423	1.481	0.058
53	1.233	1.072	0.16
54	1.113	1.273	0.16
55	0.93	1.089	0.159
		$\overline{E}$	0.244
11	3.63	4.00	0.37
14	3.38	3.37	0.01
19	3.18	2.93	0.25
20	3.14	4.19	1.05
29	2.77	2.19	0.58
30	2.74	2.10	0.64
40	2.16	2.31	0.15
45	1.90	1.92	0.02
48	1.73	1.71	0.03
		$\overline{E}$	0.34

showing lower values of LUMO energy are generally associated with higher values of pTD_{50}.

With regard to the remaining descriptors, it is quite difficult to find a direct relationship with known features generally related to the hepatocarcinogenic potential of molecules.

In order to better visualise the data, the dataset was subjected to Principal Component Analysis (PCA) [37].

In general PCA is a technique able to reduce the number of variables describing the data set.

The example mentioned below may help in clarifying the above concepts. When analyzing a data set described by 100 variables and 100 observations (data objects), one must look into 100 mean values, 100 variances, and $[(100 \times 100) - 100]/2$ covariances, for a total of 5150 statistics to be studied, as a proper representation of the underlying multivariate normal population sampled. PCA is a method of simplifying this task. The key to the problem is that much of the variability in the data set is not independent, i.e., significant covariation does exist among the variables. If we could extract two variables that captured most of the independent variability in the entire data set, from all variables under consideration, a simple binary scatter diagram would reveal most of the information in the data. Accordingly, data reduction is the primary objective to extract a few uncorrelated variables that may capture most of the variability in the data set, while preserving the orthogonality of these new optimal reference axes/variables (i.e., principal components). The 1st principal component captures the maximum variation in the data set. The 2nd principal component has the next most variation, and so on.

PCA was performed in this study on the ingredient data set (55 molecules $\times$ 5 attributes), just to visualise the data. The first principle component explained 63% of the variance; this component was mainly related to *highest normal mode vibration transition dipole* descriptor on one side of the scale and to the Number of Cl atoms on the other. The second PC explained 33% of the variance and this component was mainly related to the *FNSA-3* descriptor on one side of the scale and to *LUMO energy* on the other. If we plot the first principal component as a function of the second principal component (Fig. 1), we will be able to better visualise the data and study most variations in the data set.

The search of good QSAR models started by applying to the TR set, which comprises 46 instances (molecules) and 5 attributes (the selected CODESSA molecular descriptors), MLR and some of the other regression methods available in WEKA and using, at first, with default parameter values.

The algorithm performances were estimated on TR (R^2, q^2 and $R^2_{-10fold}$) and TS (R^2, R^2_0, $(R^2 - R^2_0)/R^2$, k) sets, according to what is mentioned in Section 2.3. The results of such a preliminary analysis are reported in Table 2.

The first group of models, obtained by applying various algorithm with default parameters, didn't satisfy the validation criteria imposed that are quite strict. Further investigations were performed by changing the values of parameters, which enabled tuning the process of model development. The MultiLayerPerceptron [29], SMOregression [30,31] and LibSVM [24–28] methods, that had shown to give better results, were exploited in a more sophisticated search.

MultiLayerPerceptron (MLP) is a basic class of feed-forward neural networks capable to approximate generic classes of functions, including continuous and integrable ones. Among the MLP features, ability to learn and generalize, smaller TR set requirements, fast operation, ease of implementation are the most relevant. Because of that MLP are of large use in QSAR. Any type of MLP neural network consists of three types of layers: an input layer, an output layer and one or more hidden layers. The addition of hidden layers revived the perceptron by extending its ability to solve problems which are non-linear, non-smooth, or contain many variables. An appropriate structure would help to achieve high accuracy of models. In the study presented here, several attempts were performed to optimize the statistical parameters (performance) on TR and TS sets but no models were found, which passed all the chosen validation criteria (data not shown).

The SMOregression implements Smola and Scholkopf's sequential minimal optimization algorithm for training a support vector regression model. This implementation globally replaces all missing values and transforms nominal attributes into binary ones. It also normalizes all attributes by default. Several attempts were made trying to optimize the performance on TR and TS sets but, also in this case, no models were found, which passed all the chosen validation criteria (data not shown).

Support Vector Machine (SVM) is one of the most recent development in machine learning modeling. This method is based on the theory of risk minimization and is able to find an optimal separation hyperplane in a multi-dimensional space to perform classification or regression tasks. It has also shown great promise in QSAR studies due to its ability to interpret the non-linear relationships between molecular structure and bioactivities.

The primary challenge in applying SVM modeling methods to a given domain lies in the selection of the kernel and its parameters. Kernels allow SVMs, which are linear machines, to transform the feature space and behave as non-linear models. The parameters of the kernel determine the shape of the separating margin used to classify a set of features (variables).

In this work, several attempts were made with SVM in order to optimize the process of parameter selection ("tuning"). This 'tuning' process implies that parameters are changed by fixed step-sizes; the performance of each set of parameters is measured on the TR and TS sets. After several attempts where parameters were finely tuned in applying LibSVM methods [24–28], available in the WEKA package, a model was found, which passed all the validation criteria chosen (fully-validated model). In particular, such a model was obtained using the epsilon–SVR algorithm, available among the SVMType algorithms in WEKA, with a polynomial kernel type.

Table 5
Statistical parameters referred to the 20 permuted TR sets.

Permuted TR	R^2	q^2	$R^2_{-10fold}$	Permuted TR	R^2	q^2	$R^2_{-10fold}$
y1	0.386	0.003	0.006	y11	0.763	0.000	0.000
y2	0.728	0.013	0.029	y12	0.329	0.001	0.001
y3	0.380	0.002	0.004	y13	0.469	0.015	0.019
y4	0.380	0.002	0.000	y14	0.501	0.001	0.000
y5	0.405	0.028	0.026	y15	0.854	0.002	0.027
y6	0.870	0.055	0.008	y16	0.855	0.001	0.033
y7	0.759	0.004	0.000	y17	0.466	0.009	0.001
y8	0.338	0.002	0.001	y18	0.295	0.006	0.004
y9	0.574	0.000	0.000	y19	0.440	0.047	0.003
y10	0.322	0.000	0.001	y20	0.379	0.005	0.006

Table 6
Comparison between the statistical parameters referred to the proposed model and the mean values obtained from y-scrambling.

R^2	q^2	$R^2_{-10fold}$	
0.919	0.580	0.600	proposed model
0.525	0.010	0.008	y-scrambling
0.394	0.570	0.592	|error|

The statistical parameters related to the above model are shown in Table 3. The actual vs. calculated values of pTD_{50} for the molecules of TR and TS sets are reported in Fig. 2. The actual and the calculated values of pTD_{50} for the test set molecules are also reported in Table 4, together with the value of average absolute error ($\overline{E}$) calculated in order to estimate the accuracy of the model.

Moreover the y-scrambling test was performed on the TR set in order to further check its robustness (see Section 2.3). This procedure involves fitting several models, twenty in our case, on the same dependent variables (X block) but on a permuted response. If a strong correlation remains between the descriptors characterizing the model and the randomized response, then the significance of the proposed QSAR model should be regarded as suspect. In our case, the proposed model performs much better than any of its permuted ones (Table 5). This is a clear proof that the model is not affected by any chance correlation, i.e. the probability of obtaining similar or better models using random numbers is null, and the model is likely to depict true relationships. In Table 6 the comparison between the statistical parameters referred to the proposed model and the mean values obtained from y-scrambling is reported.

5. Conclusion

The development of QSAR models supplies a very helpful tool in drug discovery and in the field of ADMET predictions. Understanding the relationships between the structural features of a series of compounds and their biological activity/toxicity allows optimizing features responsible for the wanted/unwanted properties. That is of great relevance, since QSAR models allow to quantitatively predict biological properties or toxicity profiles of newly designed compounds before their synthesis. It allows removing undesirable molecules at early stages of their development, thus preventing non-optimal use of resources.

Furthermore, *in silico* approaches are regarded with great interest as possible alternative methods to animal testing by several organizations that concentrate their efforts on development and validation of alternative test methods that *refine*, *reduce* or *replace* animal usage.

In this paper, QSAR models were developed on a dataset of 55 molecules reported in the CPDB to behave as hepatocarcinogenic agents with different potency. Particular care was put on rationally splitting the dataset into TR and TS sets before the development of the models, by using an algorithm based on the sphere-exclusion approach, so that TR/TS sets were optimally selected from the whole available data set of known molecules.

Some of the regression methods available in WEKA capable of handling numeric attributes (molecular descriptors) were used for the search of an optimal QSAR model. At first, an attribute evaluator available within WEKA was employed to determine the proper combination of attributes for the subsequent regression processes. Five of the 267 attributes were selected as the optimal subset, by means of 10-fold cross-validation. They are: the *number of Cl atoms*, the *LUMO energy*, *FNSA-3* that is the ratio between weighted atomic charge PNSA (partial negative surface area) and total molecular surface area, the *highest normal mode vibrational frequency* and the *highest normal mode vibration transition dipole*. Finally, an optimal model was found with the epsilon–SVR type of SVM available in WEKA using a polynomial kernel type. Fine tuning of the parameter values allowed to obtain a fully-validated QSAR model, able to satisfy the rigorous statistical criteria imposed by OECD guidelines on QSAR modeling.

Hence, we can conclude that the model proposed here could supply a simple but highly effective tool enabling the prediction of potential hepatocarcinogenicity of NCEs at the early stage of their development. The only condition to be fulfilled, in order to make reliable predictions on new molecules, is that they belong to the applicability domain described by the TR set.

References

[1] R. Combes, ATLA 30 (2002) 1–3.
[2] W.M.S. Russell, R.L. Burch, The Principles of Humane Experimental Technique, http://altweb.jhsph.edu/publications/humane_exp/foreward.htm.
[3] Guidance document on the validation of (quantitative) structure–activity relationship [(Q)SAR] models. OECD 2007.
[4] http://potency.berkeley.edu/.
[5] A.M. Helguera, M.N. Cordeiro, M.A. Perez, R.D. Combes, M.P. Gonzales, Bioorg. Med. Chem. 16 (2008) 3395–3407.
[6] A.M. Helguera, M.P. Gonzales, M.N. Cordeiro, M.A. Cabrera Perez, Chem. Res. Toxicol. 21 (2008) 633–642.
[7] L.G. Valerio Jr., K.B. Arvidson, R.F. Chanderbhan, J.F. Contrera, Toxicol. Appl. Pharmacol. 222 (2007) 1–16.
[8] O. Deeb, B. Hemmateenejad, A. Jaber, R. Garduno-Juarez, R. Miri, Chemosphere 67 (2007) 2122–2130.
[9] R. Benigni, C. Bossa, Ann. Ist Super. Sanità 42 (2006) 118–126.
[10] R.W. Tennant, J. Spalding, S. Stasiewicz, J. Ashby, Mutagenesis 5 (1990) 3–14.
[11] D.W. Bristol, J.T. Wachsman, A. Greenwell, Environ. Health Perspect. 104 (1996) 1001–1010.
[12] A. Coi, F.L. Fiamingo, O. Livi, V. Calderone, A. Martelli, I. Massarelli, A.M. Bianucci, Bioorg. Med. Chem. 17 (2009) 319–325.
[13] A.R. Katritzky, V.S. Lobanov, M. Karelson, CODESSA: Reference Manual; Version 2, University of Florida, 1994.
[14] I.H. Witten, F. Eibe, Data Mining: Practical Machine Learning Tools and Techniques, second ed. Morgan Kaufmann, San Francisco, 2005.
[15] B.D. Hudson, R.M. Hyde, E. Rahr, J. Wood, Quant. Struct.–Act. Relat. 15 (1996) 285–289.
[16] M. Snarey, N.K. Terrett, P. Willett, D.J. Wilton, J. Mol. Graph. Model. 15 (1997) 372–385.
[17] R. Nilakantan, N. Bauman, K.S. Haraki, J. Comput. Aided Mol. Des. 11 (1997) 447–452.
[18] R.D. Clark, J. Chem. Inf. Comput. Sci. 37 (1997) 1181–1188.
[19] J.G. Topliss, R.P. Edwards, J. Med. Chem. 22 (1979) 1238–1244.
[20] I. Massarelli, A. Coi, D. Pietra, F. Ahmad Nofal, G. Biagi, I. Giorgi, M. Leonardi, F. Fiamingo, A.M. Bianucci, Eur. J. Med. Chem. 43 (2008) 114–121.
[21] A. Golbraikh, A. Tropsha, J. Mol. Graphics Model. 20 (2002) 269–276.
[22] http://chem.sis.nlm.nih.gov/chemidplus/ProxyServlet?objectHandle = Search&actionHandle = clear&nextPage = chemidheavy.jsp.
[23] M.J.S. Dewar, E.G. Zoebisch, E.F. Healy, J.J.P. Stewart, J. Am. Chem. Soc. 107 (1985) 3902–3909.
[24] Yasser EL-Manzalawy, WLSVM (2005) http://www.cs.iastate.edu/~yasser/wlsvm/.
[25] Chih-Chung Chang, Chih-Jen Lin, LIBSVM – A Library for Support Vector Machines (2001).http://www.csie.ntu.edu.tw/~cjlin/libsvm/.
[26] A. Niazi, S. Jameh-Bozorghi, D. Nori-Shargh, J. Hazard Mater. 151 (2008) 603–609.
[27] Shi, X. Liu, J. Appl. Polym. Sci. 101 (2006) 285–289.
[28] Li, H. Liu, X. Yao, M. Liu, Z. Hu, B. Fan, Anal. Chim. Acta 581 (2007) 333–342.
[29] X. Zhong, B.Z. Wang, H. Wang, Int. J. Infrared and Millimeter Waves 22 (2001) 1267–1276.
[30] A.J. Smola, B. Schoelkopf, A tutorial on support vector regression, NeuroCOLT2 Technical Report Series (1998).
[31] S.K. Shevade, S.S. Keerthi, C. Bhattacharyya, K.R.K. Murthy, Improvements to SMO Algorithm for SVM Regression, Control Division Department of Mechanical and Production Engineering, National University of Singapore, 1999.
[32] M.A. Hall, Correlation-based Feature Subset Selection for Machine Learning. Hamilton, New Zealand, 1998.
[33] D.M. Hawkins, J. Chem. Inf. Comput. Sci. 44 (2004) 1–12.
[34] R.E. Tarone, Environ. Health Perspect. 116 (2008) A374.
[35] V. Mathur, P.J. John, I. Soni, P. Bhatnagar, Adv. Exp. Med. Biol. 617 (2008) 387–394.
[36] M. Howsam, J.O. Grimalt, E. Guinó, M. Navarro, J. Martí-Ragué, M.A. Peinado, G. Capellá, V. Moreno, Environ. Health Perspect. 112 (2004) 1460–1466.
[37] J.W. Gardner, P.N. Bartlett, Sens. Actuators B 18–19 (1994) 211–220.