Carlo Bertinetto[a], Celia Duce[a], Alessio Micheli[b],
Roberto Solaro[a], Maria Rosaria Tiné[a]
[a]Dipartimento di Chimica e Chimica industriale
Università di Pisa
[b]Dipartimento di Informatica
Università di Pisa
mrt@dcci.unipi.it

TOXICITY PREDICTION BY A CHEMINFORMATICS APPROACH

In this paper the advantages of predicting toxicity through QSAR analysis by direct and adaptive treatment of the molecular structure are discussed. This approach indeed allows for retaining the whole structural information and thus tackling the issue in a more flexible way. It is particularly suitable in new problems for which little or no background information is available.

The tens of thousands of substances produced every year by industrial and agricultural activities give increasing importance to the evaluation of their hazards to the life and health of humans and other living species. This highly challenging task confronts national and international regulatory agencies including the U.S. Environmental Protection Agency (U.S. EPA), Canadian Ministry of the Environment, and the European Union [1-3]. Time and funding constraints do not allow for feasibly performing toxicity tests on all new and existing chemicals released into the environment. It is therefore necessary to integrate experimental data with predictive methods able to provide additional information and maximize efficiency by prioritizing those chemicals that are most suitable for empirical assessment. Since it was observed that the properties of compounds depend on their structure, quantitative structure-activity relationships (QSARs) have been employed in elucidating the specific mechanisms underlying toxic effects. Nowadays, QSARs have been recognized by the regulatory authorities as scientifically credible tools for the prediction of acute toxicity when few or no empirical data are available [1, 4, 5]; they may hence greatly contribute to achieving the goal of abandoning animal tests.

Basically, a QSAR aims at finding a function F that relates an appropriate representation of the molecular structure to the biological activity (or any other target property). More in detail, F can be decomposed into a *feature representation* function f and a *mapping* function g. The choice of the functions *f* and *g* is the discriminating aspect among the different approaches, with a major role of the issues related to the *f* function. As already outlined [6], the function *f* entails both the representation of the molecular structure and the subsequent *encoding* of the structure into a set of numerical descriptors. Different approaches have been used to realize the *f* function. Standard QSAR approaches use molecular properties or structural molecular descriptors to encode the molecules (*f* function) and an either linear or nonlinear regression model to compute the output value (*g* function). Traditional approaches to prediction of substances toxicity are based on the standard QSAR assumption.

A great number of models have been developed in the past decades, ranging from the first simple Meyers-Overton rule at the beginning of the 20th century to very sophisticated ones that employ many descriptors and can account for different Mechanisms Of toxic Action (MOA) [7-9]. These models can be classified into two categories:

a) QSARs that use the structure information only indirectly, while physicochemical properties are used as input variable(s) to predict the biological effect of interest;

b) QSARs that directly take as input only the information reflecting structural features (structural molecular descriptors), such as number of occurrences of specific molecular groups/atoms and how they are connected inside the molecule (topological indices and matrices).

The models gathered into the C-QSAR computer package developed by Hansch and his group at Pomona College within the Medicinal Chemistry Project and distributed by BioByte Corp. [10] belong to the first category. Another set of rules of thumb was then implemented in the 3.1 version of the M-CASE program created by Klopman *et al.* [11]. Physicochemical information does not only allow for correlating the correct toxicity, but also its corresponding MOA. For instance, K_{ow} (octanol/water partition coefficient), K_a (acidity constant) and E_{LUMO} (energy of the lowest occupied molecular orbital) are often used to discriminate among polar narcosis, oxidative uncoupling and (direct or metabolically induced) electrophilic mechanisms of toxic action [12-15]. A major problem associated to these techniques, giving rise to the need for large databases such as the one contained in the C-QSAR software, is that the preparation of the needed input frequently requires either collection of experimental data or computation of the associated physicochemical parameters. For example, K_{ow} or a relevant incremental derivative, which is one of the most widely used input variables in this type of QSARs, is often calculated through appropriate software packages, such as MedChem [16]. This *de facto* corresponds to using (more or less) the structural information as direct input in other structure-activity relationships, which are themselves subject to various restrictions and can generate serious additional errors [17]. These QSARs are performing well on very narrow classes of compounds comprising molecules having similar basic skeleton and only one MOA, or more MOAs that are known to correlate to specific descriptors (local models). The best accuracy is expected in such specific tasks. The larger challenge is to simultaneously handle a variety of more complex substances with unknown MOAs (non-local models). The search for a more comprehensive solution to this issue led to the development of the models belonging to the second category, of which a few examples are mentioned below. These methods are based on simple molecular descriptors derived solely from the chemical structure, do not use the octanol/water partition coefficient, do not involve any rules of thumb, and do not require a substance classification for the handling of the information or generating the results.

For the formal description of relationships between the target property and the chosen molecular representation (*g* function), the most frequently used methods are multiple linear regression (MLR) and partial least squares (PLS). Other widely employed statistical and machine learning techniques include discriminant analysis, multivariate analysis, adaptive least squares, support vector machines (SVM), k-nearest neighbour (kNN) and genetic algorithms [18, 19]. Because the relationship between structural descriptors and toxicity is usually nonlinear, neural networks (NNs) of different types are also frequently and successfully employed for the realization of the *g* function. [18-24]. NNs are universal approximators able to learn nonlinear relationships between a proper representation of a chemical structure and a given target property from a set of examples.

In order to obtain a good QSAR prediction, it is of fundamental importance to have empirical data of high quality [25]. A database that satisfies this requirement [23] is the *Tetrahymena* toxicity database, Tetratox, which contains population growth inhibition concentrations (IGC_{50}) of the freshwater ciliate *Tetrahymena Pyriformis* [26]. This protozoan is attractive for its fast growth rate and inexpensive assays and for its significance in the estimation of the impact of toxicants in aquatic environments [27-29]. The measurement procedure [12] has been carefully established over the years and is now widely recognized as a standard; the database is constantly growing and currently includes more than 2,500 bioassays tested on approximately 1,400 organic compounds [26].

Several QSAR studies were performed in the past on the Tetratox database with models of either type. Considering only those that employ neural networks for the realization of the *g* function, an almost up-to-date list of NN QSAR models for *Tetrahymena* is contained in ref. 23. Data sets of small size (< 300 compounds), containing homogeneous molecules with known MOAs are successfully described by models employing descriptors related to the toxic effect of the studied compounds. Among these, the best performances were achieved by Melagraki and co-workers [19], who employed a Radial Basis Function (RBF) NN over 180 substituted phenols represented through five descriptors: logarithm of the octanol/water partition coefficient ($\log K_{ow}$), acidity constant (pK_a), the number of hydrogen bond donors (N_{Hdon}), the energies of the highest occupied and lowest unoccupied molecular orbital (E_{HOMO} and E_{LUMO}, respectively). The toxicity of these phenols involves four different MOAs [30,31], polar narcotics, weak acid respiratory uncouplers, pro-electrophiles and soft electrophiles, which are highly correlated to the chosen descriptors [32]. In particular, K_{ow}, expressing hydrophobicity, provides a measure of the degree of penetration of the toxic substance into the cell tissue and its concentration at the different sites of action [33]. The ability to act as oxidative uncouplers is associated with specific pK_a values [13]. Hydrogen bonding, evaluated by N_{Hdon}, can discriminate between polar and nonpolar narcotics [34] and may also play a role in fixating the toxicant in the course of bioreactive interactions with endogenous macromolecules. E_{HOMO}, characterizing the ionisation potential and thus the ability of the molecule to donate electrons to reaction partners [35], can be used to model the metabolic activation required by pro-electrophiles in the case of oxidative pathways. E_{LUMO}, which quantifies the electron affinity of the chemical [35], has also been demonstrated to discriminate various MOAs [31], as it may reflect the tendency of phenols to directly attack electron-rich sites of endogenous macromolecules, as well as their ability to undergo metabolic activation following 1-electron reduction [36]. The correlation obtained by Melagraki showed a squared correlation coefficient (R^2) of 0.94 for the training set, 0.88 for an external test set of 41 molecules and 0.72 for the leave-one-out cross validation.

For larger and more heterogeneous data sets, the above-mentioned methods give way to non-local models that use structural features as input. Among these, the model that achieves the best performance is the one by Niculescu *et al.* [17] that used 33 descriptors derived from structural features, such as molecular weight and occurrence of specific atoms or molecular fragments, to fit a set of 750 diverse organic compounds through a Probabilistic Neural Network (PNN). The R^2 value for the training set and for the external validation set of 75 molecules was 0.93 and 0.89, respectively. The largest investigated data set was the one used by Kahn *et al.* [23], who selected a set of six descriptors through Heuristic Back-Propagation Neural Networks (hBNN) for 1371 compounds, partitioned into a training, test and external validation set of 610, 304 and 457 molecules, respectively. The obtained correlations showed R^2 values of respectively 0.83, 0.82 and 0.79 for each set.

These models achieved good predictions and are often able to simultaneously treat different chemical classes and mechanisms of toxic action. However, because the used regression methods take fixed-size numerical vectors as input, all these models require the definition of suitable molecular descriptors in order to reduce molecules to vectors of the same dimension. The choice of the number and types of numerical descriptors used to represent chemical compounds of a particular data set is a delicate and cumbersome task as the most suitable descriptors are strictly dependent on the target property and the structural characteristics of the considered compounds. In particular, descriptors summarize the background knowledge on the specific task at hand. This fact is especially significant in toxicological prediction because the latter involves the study of a wide range of chemical classes, as well as different target properties (LD_{50}, IC_{50}, IGC_{50}…), measured on different species with different methodologies. The need for molecular descriptors characterizes the limit of applicability for these methods, as a new set of molecular descriptors must be calculated and selected every time a new property or compound type is investigated [6].

In this framework, our goal is to make a first step for the critical evaluation of toxicity prediction by means of a non-standard QSPR (Quantitative Structure-Property Relationship) approach developed by our research group for diverse fields during the last years. This approach is based on Recursive Neural Network (RNN) methods, which belong to the area of machine-learning models developed to directly handle structured data. The main advantage of the RNN approach stems from the automatic generation of descriptors by learning the numerical encoding (function *f*) of the input structure together with the regression function *g*. A second important point concerns the treatment of molecules as variable-size structures, consisting of hierarchical sets of labelled vertexes connected by edges belonging to subclasses of graphs, such as rooted trees.

Labelled structures are highly abstract and graphical tools that can represent a molecule at different levels of detail, such as atoms, bonds, or chemical groups. A natural representation of a molecule is made possible by reproducing its 2D structure in the input graph. This feature allows for applying the model to different prediction problems without the need for lengthy calculations of descriptors or a great amount of background knowledge. This model is particularly suitable for all the new tasks in which it is useful to retain the whole structural information and there is little or no available background knowledge. These aspects often characterize toxicity evaluation problems.

It must be stressed that the use of hierarchical labelled structures as class of data introduces both constraints and flexibility to the molecular representation. In particular, the choice of fragments, i.e. the level of detail by which chemical groups are represented in the structures determines at the same time the level of chemical information, the fragment sampling in the data set, the structure size and complexity (see *Molecular Representation* section). An effective representation seeks a good balance among these often conflicting issues [37].

The RNN model has already been successfully applied to the prediction of various physicochemical properties of different classes of chemical compounds, ranging from simple molecules to polymers. Previous works dealt with the boiling points of linear and branched alkanes [38, 39], the pharmacological activity of a series of substituted benzodiazepines [38-40] and 8-azaadenine derivates [41], the free energy of solvation of monofunctional compounds [6], the glass transition temperature of (meth)acrylic polymers and copolymers [37, 42-46] and the melting point of pyridinium bromides [37, 47]. In each case, the flexibility introduced by our approach allowed the use of the same computational method to study a variety of properties and different molecular structures just by tuning the level of structural detail to the characteristics of the investigated molecular data set.

This paper reports our first step in the use of the RNN-QSPR technique, as a predictive model based only on the graph molecular structure, in toxicological studies. To this end, we employed the data set from the mentioned work by Melagraki [19].

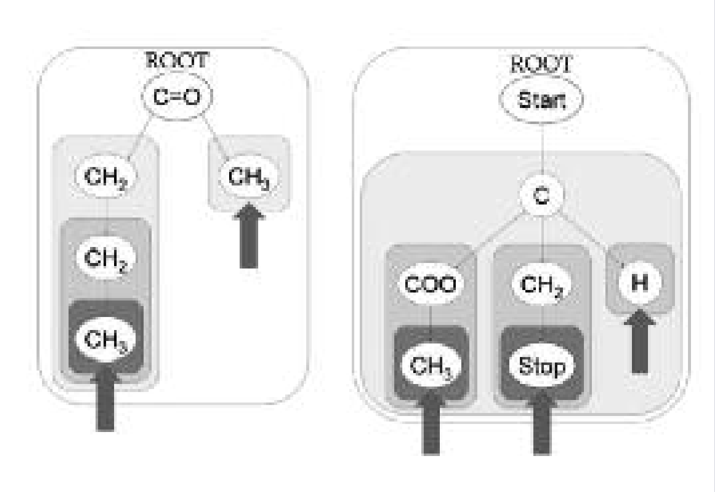

Fig. 1 - Examples of unfolding of the encoding process across a tree structure. Each box includes the sub-trees progressively encoded by the recursive neural network. The encoding process begins from the leaves, as indicated by the thicker arrows at the bottom, and proceeds step-by-step up to the root. The code computed for the root is thus considered the code for the entire molecule

Method

We here briefly explain the main characteristics of the RNN model, which is given a complete description in ref. [6, 38-41]. What mainly characterizes RNNs is their ability to directly deal with labelled hierarchical structured representation of molecules, in particular in the form of rooted trees, a subclass of DPAGs (Directed Positional Acyclic Graphs). Trees have a variable size and are a much more rich and flexible vehicle of information than the flat vectors of descriptors employed in traditional QSAR approaches.

Another major characteristic of RNNs is their ability to adaptively encode the input structures by learning from the given structure-property training examples. In order to achieve this goal, the RNN recursively encodes each structure through a bottom-up approach that mimics their morphology, see Fig. 1. For each vertex of the input structure, the model computes a numerical code by using information of both the vertex label and, recursively, the code of the sub-graphs descending from the current vertex. This process computes a code for the whole molecular structure. The code is then mapped to the output property value.

The learning algorithm allows the model for tuning the free parameters of the neural network functions on the basis of the training examples and by this process the RNN models a direct and adaptive relationship between molecular structures and target properties. Generally, the learning process becomes more effective as the number of training examples increases.

The use of an adaptive encoding for structures avoids the need for computing/measuring or selecting *ad hoc* molecular descriptors, as in traditional methods, and the need for the prior definition of a similarity measure for the structured data. It can also be interpreted as an automatic way to discover by learning the specific structural descriptors for the particular task to be solved.

Molecular representation

Chemical compounds are represented as labelled rooted ordered trees by a 2-D graph that can easily be obtained from its structural formula. Each molecule is partitioned into defined atomic groups: each group corresponds to a vertex of the tree and each bond between them corresponds to an edge, as depicted in Figs. 2-4. The criteria guiding this fragmentation are basic notions of chemical reactivity and structural sampling. When dealing with structured domains, the latter consists in the coverage of the input space, i.e. occurrence of the different fragments/components in their possible topologies in the given data set. The sampling issue suggests choosing the smallest number of atomic groups able to build the greatest number of molecules in a reasonably compact form. An appropriate set of rules was defined in order to have a unique correspondence between each molecule and its chemical tree. All fragments were rated according to a priority scale that was used to determine the tree root and the total order on each vertex subtree

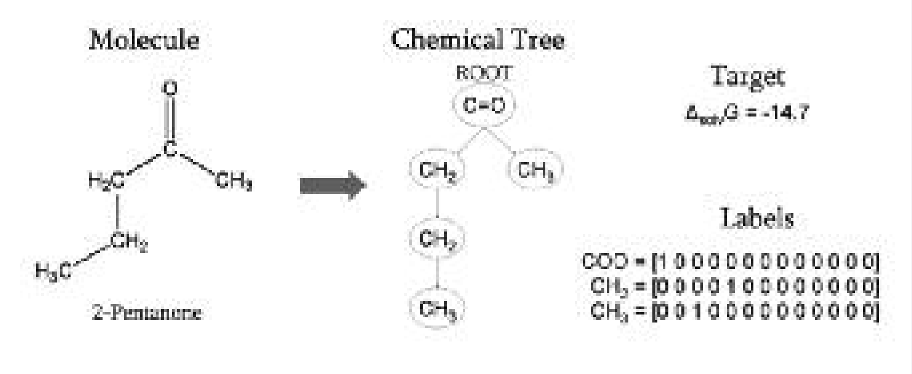

Fig. 2 - Tree representation of 2-pentanone. The fragments in this example have labels that are all orthogonal to each other

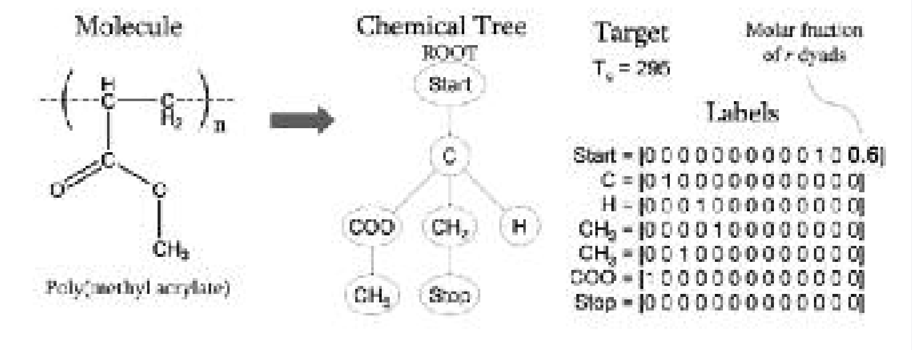
Molecule
H
C
C
H_2
n
C
O
O
CH_3
Poly(methyl acrylate)
Chemical Tree
ROOT
Start
C
COO
CH_2
H
CH_3
Stop
Target
T_g = 295
Molar fraction of r dyads
Labels
Start = [0 0 0 0 0 0 0 0 0 0 1 0 **0.6**]
C = [0 1 0 0 0 0 0 0 0 0 0 0 0]
H = [0 0 0 1 0 0 0 0 0 0 0 0 0]
CH_2 = [0 0 0 0 1 0 0 0 0 0 0 0 0]
CH_3 = [0 0 1 0 0 0 0 0 0 0 0 0 0]
COO = [1 0 0 0 0 0 0 0 0 0 0 0 0]
Stop = [0 0 0 0 0 0 0 0 0 0 0 0 0]

Fig. 3 - Tree representation of poly(methyl acrylate). The Start label contains information about average macromolecular characteristics, here highlighted in bold

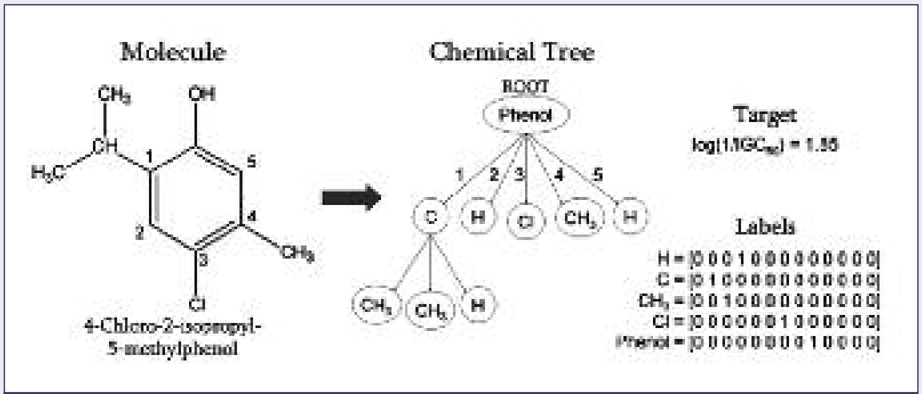
Molecule
CH_3
OH
CH
H_3C
1
5
4
CH_3
2
3
Cl
4-Chloro-2-isopropyl-
5-methylphenol
Chemical Tree
ROOT
Phenol
1
2
3
4
5
C
H
Cl
CH_3
H
CH_3
CH_3
H
Target
$\log(1/IGC_{50}) = 1.55$
Labels
H = |0 0 0 1 0 0 0 0 0 0 0 0 0 0|
C = |0 1 0 0 0 0 0 0 0 0 0 0 0 0|
CH_3 = |0 0 1 0 0 0 0 0 0 0 0 0 0 0|
Cl = |0 0 0 0 0 0 1 0 0 0 0 0 0 0|
Phenol = |0 0 0 0 0 0 0 0 0 1 0 0 0 0|

Fig. 4 - Tree representation of 4-chloro-2-isopropyl-5-methylphenol. The children of Phenol, which is the tree root, are ordered according to the position of the substituents on the phenyl ring, as indicated by the numbers 1-5 (not to be confused with the position numbers of the IUPAC nomenclature)

[6]. Each vertex is assigned a label, which is a tuple of variables categorically distinguishing the symbol of the atomic group. Despite being conventionally defined, a label can convey chemical information through orthogonality or similarity to other labels, see Fig. 2.

This kind of structure-based representation has the advantage of generality, as it can adequately represent any sort of chemical compound [44]. A clear example of this flexibility is given by its implementation for the description of polymers: the 2D graph of the repeating unit is partitioned and treated like a molecule, with the difference that its ends are capped by two fictitious groups, named *Start* (which is the root vertex) and *Stop*, see Fig. 3. The former does not serve only representational purposes, as it can also be used to convey information on average macromolecular characteristics. In our experiments, *Start* accounted for the main chain stereoregularity recorded as molar fraction of *r* dyads [42]. Copolymers were represented in a similar manner, with the *Start* group connecting two repeating units instead of one and containing information concerning their respective molar fraction [45, 46].

Different representations were also devised for compounds, both molecular and polymeric, containing cyclic moieties [37, 44, 47]. They span from a compact description to a more detailed and general one that makes use of standard chemical representation systems, such as SMILES and InChI [48-50].

In the current work we described the 221 phenols in the data set with 27 atomic fragments: Phenol, C, CH_2, CH_3, CH_2-CH_2 (two consecutive CH_2), $C_{(sp2)}$ (forming double bonds), H (either aliphatic or aromatic), $H_{(acid)}$, $H_{(aldehyde)}$, C=O, COO, O, OH, $N_{(amine)}$, $NH_{2(amine)}$, $NH_{(amide)}$, $NH_{2(amide)}$, NO, NO_2, $N_{(sp2)}$, CN, F, Cl, Br, I, Phenyl, Cyclopentyl. Acids and aldehydes were represented as COO-$H_{(acid)}$ and C=O-$H_{(aldehyde)}$, respectively, because they are not frequent in the data set and can in this way be trained also through their analogies with the COO and C=O groups. The tree root was always placed on the *Phenol* group and its children, i.e. the substituents on the phenyl ring, were ordered according to their position on the ring, see the example in Fig. 4. The direction of the ordering (i.e. clockwise or counter-clockwise) is the one that assigns the lowest position to the group with highest priority, in analogy with IUPAC rules. In the case of two or more groups with identical priority, the direction is determined by the subsequent group in the priority scale, until the ambiguity is solved.

Although based on a 2D graph, the tree-structured molecular representation can also be extended to describe 3D properties through definition of appropriate rules, e.g. the children ordering can be exploited to indicate chirality in analogy with Fisher's projections [6].

Experiments and results

In the context of the discussion about the proposed approach for the adaptive processing of chemical structures by RNN, as mentioned in the introduction, we addressed a first step in its critical evaluation for toxicology applications through the following computational problem. We used a data set taken from Melagraki [19, 32], consisting of 221 phenols and their corresponding toxicity data to the ciliate *Tetrahymena pyriformis* in terms of $\log(1/IGC_{50})$, with IGC_{50} indicated in mmol/l. These compounds comprise four different modes of toxic action: polar narcotics, weak acid respiratory uncouplers, pro-electrophiles and soft electrophiles. This data set was selected because it contains structurally similar compounds that can be easily represented by the rules previously determined and assessed on other data sets [6, 37]. Its size seemed a reasonable compromise between the necessity of many training compounds and the need of a data set that is easy to compile and to analyse. The data source, the Tetratox database, guarantees that toxicity figures are very reliable and that the set can easily be expanded for further and more complete works. The total data set of 221 compounds was split into a training set and an external test set of 180 and 41 molecules, respectively; the partition is the same as in ref. [19].

The connection weights of the RNN model are initialized at random because of the use of a stochastic gradient-based technique to solve a least mean square problem. Consequently, different outcomes can be achieved during the training of the network by starting from diverse initial conditions. In order to have significant appraisal of the results, in each experiment sixteen trials were carried out for the RNN simulation and the results were averaged over the different trials. It is worth noting that a naive approach based on the selection of the best outcome over the various trials can lead to an unsatisfactory and unreliable estimation of the model performance. Moreover, this practice discards potentially useful information on the model behaviour, which is stored in the discarded regression estimates. The use of a basic ensemble method avoids these problems while offering an improved regression estimate.

Learning was stopped when the maximum error for each compound of the training set was below a preset threshold value. Care must be taken in the selection of this value, in order to model an accurate relationship according to the suitable tolerance for the noisiness and uncertainty of the data. A general rule suggests choosing a training error tolerance equal to or near the experimental uncertainty, whenever such information is known. In our experiments the threshold was set at 1.00 unit of log(1/IGC50) ([IGC50] = mmol/l). Preliminary steps were performed to asses the soundness of this threshold value, also with respect to the outliers. We basically observed that changing the fitting level does not allow for addressing the issue of the analysis reported in the following, while the set threshold value allows us to achieve a fitting comparable to the literature results. In particular, for this task, variations on the model hyper-parameters do not seem to significantly affect the basic results.

Before evaluating the experimental results and comparing them with the reference work, we need to make some considerations about the purpose and conditions of our study. The current investigation was not aimed at a state-of-the art prediction, but instead at exploring the possibility of deriving toxicity relationship by direct treatment of the chemical structure. This is an open challenge and the present RNN application to this particular problem only constitutes an early forward step. Therefore we employed only a generic, mainly untuned representation of the molecular structures. In principle, this representation allows for getting rid of *ad hoc* background knowledge, while retaining the flexibility to model molecules with a graphical tool. Such general approach may not be the most suitable for very specific tasks. It is important to point out that the RNN model is not constrained to the exclusive use of molecular structure, which can be complemented with information of other type inserted into the vertex labels, as was done by inserting stereoregularity and composition in the representation of polymers. However, we have chosen not to exploit this possibility because we are more interested in applicative generality rather than accurate performance on this particular problem.

On the other hand, the purpose of Melagraki was to apply and refine the RBF technique on a molecular representation already known to be appropriate for the investigated compounds and property. He made use of five descriptors taken from the literature that intrinsically contain a significant amount of *a priori* information, including the correlation with the corresponding MOA [32]. Moreover, two descriptors are even experimentally measured properties instead of purely structure-derived parameters. The training-test split was determined through the Kennard-Stone algorithm [51], which is based on the relative distance between pair of compounds in terms of some metric defined on the input variables. These inputs (numerical descriptors) are of a different type than those used in our method (hierarchical structures), therefore the test set may adequately map the variable space for the RBFNN model but not for the RNN one.

In the first experiment we applied the described RNN method on the whole data set without further adjustments. The training set was fitted with very good accuracy, showing a mean absolute error (MAE) of 0.17 and a squared correlation coefficient (R^2) of 0.92. These values are comparable to the best models available in the literature. The test set prediction showed instead a MAE = 0.34 and an R^2 = 0.60, a performance not on par with that of the other referenced methods. The calculated vs experimental points are plotted in Fig. 5.

The reported results point at a high complexity of the problem with respect to molecular structures. For instance, there are various groups of compounds in the data set with similar structure but very different target values. Examples are given for instance by 1,3,5-trihydroxybenzene and 1,2,3-trihydroxybenzene, having $\log(1/IGC_{50})$ of -1.26 and 0.85, respectively. Consider that the total target range is from -1.50 to 2.71 log units. More training examples for these types of molecules are required to correctly model their relationship. In other words, a few points with widely spaced apart target values characterize some regions of the input structure space for which the sampling level is insufficient. In this case, the sampling issue does not significantly affect descriptor-based methods since the input space used by Melagraki, i.e. a five-dimensioned vector, is small. In the case of our structure-based method, a major improvement of the prediction can be expected by expanding the data set to increase the sampling level.

A group of molecules that can be singled out for showing greater errors consists in those with very low target values. Compounds in the test set with $\log(1/IGC_{50})<0$ have a MAE of 0.43, which even reaches 0.63 for targets lower than -0.20. We have to stress that

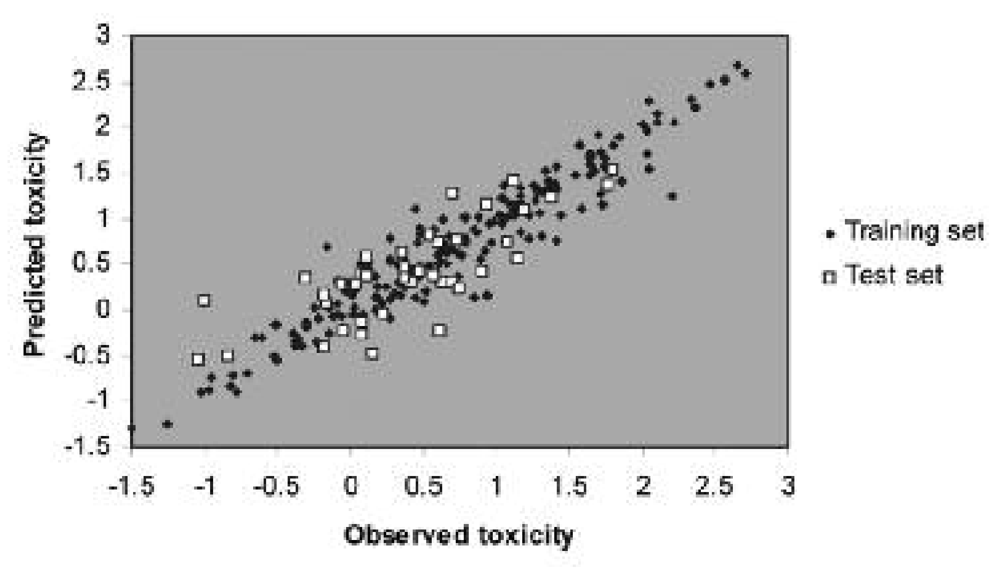

Fig. 5 - Predicted vs observed toxicity using the RNN model. Numbers correspond to $\log(1/IGC_{50})$, with $[IGC_{50}]$ = mmol/l. Test compounds with very low toxicity show higher deviations from the correct value

these values correspond to IGC50>2 mmol/l, that is to compounds of low toxicity. In any case, we tried to overcome this issue by running a second experiment in which the training molecules with target lower than -0.20 were sampled twice. The overall results were almost identical to the previous experiment, but the test compounds with negative targets showed a slight but significant improvement (MAE=0.38). A similar behaviour, though not clearly attributable to resampling, was observed at the other end of the target range: the MAE of molecules in the test set with log(1/IGC$_{50}$)>1 went from 0.31 in the first experiment to 0.28 in the second one. These observations show that resampling can be held as a useful tool for recalibration of the results even if it did not generate a much better average outcome.

Conclusions

The development of QSAR techniques based on the direct and adaptive treatment of molecular structure has a great potential in toxicological studies. The challenges posed nowadays in this field point to the formulation of general methods able to deal with continuously arising issues featuring unknown properties of newly discovered chemicals. The structured approach allows for retaining the whole structural information in order to better tackle predictive problems for which background knowledge is scarce or absent. It may still be less suitable for other more specific tasks, in which descriptor-based methods exploiting *a priori* information for local models still provide the state-of-the-art performances.

In the present implementation, we applied a direct structured representation by means of our RNN model on an already thoroughly investigated data set. In the described sampling scenario that characterizes this task, the power of the model to fit the data is not yet adequate to achieve the same predictive performances obtained by reference methods [19] on the external test set. However, our results must be evaluated by considering the different purpose and conditions of our study. Our molecular representation stemmed from a much more general hypothesis that ignores specific knowledge concerning this particular problem, such as MOA information and target-related physicochemical properties. This is a choice rather than an obligation, because the flexibility of our model does allow for including such knowledge in the input data, probably improving the results. It was observed that correlating the target property to the structure alone gives rise to an especially complex task, as highlighted by the series of compounds with similar structure and very different toxicity value. Additionally, some regions of the input space could not be trained properly due to lack of sampling, whereas descriptor-based methods avoid this problem because their input space for this task (where the background knowledge is summarized by five descriptors) is of a much lower dimension.

The present investigation is thus meant to be an exploratory effort, paving the way for further application of our model to the toxicological field. Major improvements are expected by increasing the number of molecules in the data set in order to provide more adequate sampling. Full exploitation of available databases, including Tetratox, will provide the necessary data. The appropriate definition of our representation rules could also allow for including 3D features (e.g. chirality), which are known to often have a great influence on toxicity, in our 2D description. Finally, the second experiment showed that partial resampling of the data set, even if not always effectively improving the overall outcome, could be useful for recalibration.

References

[1] C.M. Auer *et al., Environ. Health. Perspect.,* 1990, **87,** 183.

[2] D.R.J. Moore *et al.,* Ecological risk assessments of priority substances in Canada: identification and resolution of difficult issues, in F.J. Dwyer *et al.* (Eds.), Environmental Toxicology and Risk Assessment, Modeling and Risk Assessment, 6th Ed., STP 1317, American Society for Testing and Materials, Philadelphia, PA, 1997, pp. 130-147.

[3] A.G. Van Haelst, B.G. Hansen, *Environ. Toxicol. Chem.,* 2000, **19,** 2372.

[4] J.D. Walker, *J. Mol. Struct. (Theochem),* 2003, **622,** 167.

[5] White paper, COM (2003) Directive of the European Parliament and of the Council: Amending council directive 67/548/EEC in order to adapt it to regulation (EC) of the European Parliament and of the Council concerning the registration, evaluation, authorisation and restriction of chemicals. 2003/0256/COD, 2003/0257/COD (October 29, 2003).

[6] L. Bernazzani *et al., J. Chem. Inf. Model.,* 2006, **46,** 2030.

[7] R.L. Lipnick, *TIPS,* 1986, **81,** 161.

[8] R.L. Lipnick, *TIPS,* 1989, **10,** 265.

[9] S.P. Bradbury *et al.,* Polar narcosis in aquatic organisms, in U.M. Cowgill, L.R. Williams (Eds.), Aquatic Toxicology and Hazard Assessment, 12th Ed., STP 1027, American Society for Testing and Materials, Philadelphia, PA, 1989, pp. 59-73.

[10] BioByte Corp., 201 West 4th St, Suite 204, Claremont, CA 91711, C-QSAR. A general approach to the organization of quantitative structure-activity relationships in chemistry and biology, http://www.biobyte.com/bb/prod/qsarman2k.pdf.

[11] G. Klopman *et al., Environ. Toxicol. Chem.,* 2000, **19,** 441.

[12] T.W. Schultz, *Toxicol. Mech. Meth.,* 1997, **7,** 289.

[13] G. Schüürmann *et al., Environ. Toxicol. Chem.,* 1996, **15,** 1702.

[14] G. Schüürmann *et al., Aquat. Toxicol.,* 1997, **38,** 277.

[15] T.W. Schultz, M.T.D. Cronin, *Environ. Toxicol. Chem.,* 1997, **16,** 357.

[16] A. Leo, D. Weininger, MedChem, Release 3.53, Pomona

Acknowledgements: The financial support by Fondazione Cassa di Risparmio di Pisa 2006/172 Project is gratefully acknowledged.

medicinal Chemistry project, Pomona College, Claremont, CA.

[17] S.P. Niculescu *et al., Arch. Environ. Contamin. Toxicol.,* 2000, **39,** 289.

[18] U. Bukard, Methods for data analysis, in J. Gasteiger, T. Engel (Eds.), Chemoinformatics, Wiley VCH, Weinheim, 2003, pp. 439-485.

[19] G. Melagraki *et al., J. Mol. Model.,* 2006, **12,** 297.

[20] K.L.E. Kaiser, *J. Mol. Struct. (Theochem),* 2003, **622,** 85.

[21] J. Gasteiger, Handbook of chemoinformatics: from data to knowledge, Vol. 3, Wiley VCH, Weinheim, 2003.

[22] J. Zupan, J. Gasteiger, Neural networks in chemistry and drug design, Wiley VCH, Weinheim, 1999.

[23] I. Kahn *et al., Journal of Chemical Information and Modeling,* 2007, **47,** 2271.

[24] A.K. Debnath, Quantitative structure-activity relationship (QSAR): a versatile tool in drug design, in A.K. Ghose, V.N. Viswanadhan (Eds.), Combinatorial library design and evaluation: principles, software tools, and applications in drug discovery, Marcel Dekker, New York, 2001, pp. 73-129.

[25] S.P. Bradbury, *Toxicol. Lett.,* 1995, **79,** 229.

[26] T.W. Schultz, Tetrahymena in aquatic toxicology: QSARs and ecological hazard assessment. Proceedings, International Workshop on a Protozoan Test Protocol with Tetrahymena in Aquatic Toxicity Testing, German Federal Environmental Agency, Berlin, Germany, April 1996, pp. 31-65.

[27] S.D. Dimitrov *et al., J. Mol. Struct. (Theochem),* 2000, **622,** 63.

[28] J.R. Seward *et al., Chemosphere,* 2002, **47,** 93.

[29] I. Kahn *et al., ATLA-Altern. Laborat. Anim.,* 2007, **35,** 15.

[30] R. Garg *et al., Crit. Rev. Toxicol.,* 2001, **31,** 223.

[31] T.W. Schultz *et al.,* Identification of mechanisms of toxic action of phenols to Tetrahymena pyriformis from molecular descriptors, in F. Chen, G. Schüürmann (Eds.), Quantitative Structure-Activity Relationships in Environmental Sciences, VII, SETAC Press, Pensacola, FL, 1997, pp. 329-342.

[32] A.O. Aptula *et al., Quant. Struct.-Act. Relat.,* 2002, **21,** 12.

[33] S.P. Bradbury *et al., Environ. Toxicol. Chem.,* 2003, **22,** 1789.

[34] J.C. Dearden *et al., Quant. Struct.-Act. Relat.,* 2000, **14,** 427.

[35] G. Schüürmann, Ecotoxic modes of action of chemical substances, in G. Schüürmann, B. Markert (Eds.), Ecotoxicology, John Wiley and Spektrum Akademischer Verlag, New York, NY, 1998, pp. 665-749.

[36] H. Schmitt *et al., Chem. Res. Toxicol.,* 2000, **13,** 441.

[37] C. Bertinetto *et al., J. Mol. Graph. Model.,* 2009, **27,** 797.

[38] A. Micheli *et al., Stud. Fuzz. Soft Comp.,* 2003, **120** (Soft Comp. Appr. Chem.), 265.

[39] A.M. Bianucci *et al., Appl. Int. J.,* 2000, **12,** 117.

[40] A. Micheli *et al., J. Chem. Inf. Comput. Sci.,* 2001, **41,** 202.

[41] A. Micheli *et al.,* Design of new biologically active molecules by recursive neural networks, Proc. Int. Joint Conference on Neural Networks, 2001, **4,** 2732.

[42] C. Duce *et al., J. Math. Chem.,* 2009, **46,** 729.

[43] C. Duce *et al., Macromol. Rapid Commun.,* 2006, **27,** 712.

[44] C. Bertinetto *et al., Polymer,* 2007, **48,** 7121.

[45] C. Bertinetto *et al.,* Modelling Structure-Property Relationship for Copolymers by Structured Representation of Repeating Units, in G. Maroulis, T.E. Simos (Eds.), Advances in Computational Science: Lectures presented at the International Conference on Computational Methods in Sciences and Engineering 2008 (ICCMSE 2008), AIP Conference Proceedings 1108, in press.

[46] A. Micheli *et al.,* Recursive Neural Networks for Cheminformatics: QSPR for Polymeric Compounds (towards Biomaterial Design), in F. Masulli *et al.* (Eds.), Knowledge-Based Intelligent Engineering Systems - Frontiers in Artificial Intelligence and Applications, IOS Press, Amsterdam (NL), 2009, vol. 196, pp. 37-52.

[47] R. Bini *et al., Green Chem.,* 2008, **10,** 306.

[48] D. Weininger, *J. Chem. Inf. Comput. Sci.,* 1988, **28,** 31.

[49] D. Weininger *et al., J. Chem. Inf. Comp. Sci.,* 1989, **29,** 97.

[50] S.E. Stein *et al.,* The IUPAC Chemical Identifier - Technical manual, www.iupac.org/inchi/

[51] R.W. Kennard, L.A. Stone, *Technometrics,* 1969, **11,** 137.

RIASSUNTO

Un approccio chemioinformatico alla predizione della tossicità basato sull'apprendimento adattivo di rappresentazioni molecolari strutturate

In questo lavoro viene presentato un nuovo metodo QSAR basato sull'uso di una rete neurale ricorsiva (RNN) per la predizione "in silico" dei valori di tossicità di sostanze. Il metodo proposto utilizza una rappresentazione gerarchica strutturata delle molecole, in particolare nella forma di albero ordinato etichettato, molto più generale e flessibile dei tradizionali vettori di descrittori molecolari numerici e che permette di mantenere tutta l'informazione strutturale. Le rappresentazioni ad albero vengono trattate dalla RNN tramite un metodo di apprendimento adattivo che realizza contemporaneamente sia la funzione di codifica sia quella di mapping sulla proprietà target.