Some results for the prediction of carcinogenicity using hybrid systems

Giuseppina Gini, Marco Lorenzini, Angela Vittore
DEI, Politecnico di Milano, Milano, Italy
gini@elet.polimi.it

Emilio Benfenati, Paola Grasso
Department of Environment and Health, Istituto di Ricerche Famacologiche "Mario Negri", Milano, Italy

Abstract

Until recently, problem solvers have typically used single-technique-based tools to build the solution. Also in the field of predictive toxicology, a few systems have been developed in that way, with positive preliminary results. One approach to deal with real complex systems is to use two or more techniques in order to combine their different strenghts and overcome each other's weakness to generate hybrid solutions. In this project we pointed out the needs of an improved system in toxicology prediction. An architecture able to satisfy these needs has been developed. The main tools we integrated are rules, ANN, graph search, and rule learning algorithms. We defined fragments responsible for carcinogenicity according to human experts, developing a module able to recognize these fragments into a given chemical. To each fragment a carcinogenicity category was associated. Furthermore, we developed an ANN, using molecular descriptors as input, to predict carcinogenicity as a real value. PCA was used to reduce the number of descriptors used by the ANN. Finally, we developed an automatic learning program to combine the results obtained from the two previous modules into a single predictive class of carcinogenicity to man. We tuned the system to maximize the predictive power of the system.

Introduction

The goal to predict carcinogenicity is a challenging one, in consideration of the social and economical importance of the problem. Chemicals are responsible for many tumors. However, the experimental tests on chemicals are year-longs (because carcinogenicity is a form of chronic toxicity), costly, and require the use of animals, with ethical problems.

Some recent reviews on the topic have been published [1-3]. So far the most popular programs have been expert systems (ES), as HazardExpert [4], CASE [5,6], TOPKAT [7,8], DEREK [9, 10], Oncologic [11].

More recently neural networks (ANN) have been used. In some cases the results were promising [12, 13], but in another one no generalization of the ANN appeared [14].

Another way is inductive logic programming [15]. Other challenges are on-going and this fact confirm the interest on the matter [16].

In the present study we tried a new approach, combining different systems into a hybrid architecture. We developed an ES able to recognize toxic residues prediciting a class of toxicity. Furthermore, we trained a ANN with molecular descriptors and obtained a second value of predicitive toxicity. Finally, we used an ILP to merge the information steming from the two sources.

Definition of the phenomenon to model: carcinogenicity

Cancer is not a single disease. Furthermore, each single cancer involves a complex sequence of events. The complexity of the phenomenon means that experimental data are not precise, and in some cases contradictory.

Carcinogens are listed in classes by several agencies. For instance, the International Agency on Research on Cancer (IARC) considers four classes: the first (class 1) contains the compounds which have been recognized as carcinogenic to man; the last (class 4) has compounds which are not carcinogenic, and the other compounds are splitted in classes of different degree on uncertainty: probably or possibly carcinogenic to man, (class 2A and 2B) and with a quite high uncertainty (class 3 - the most numerous one) [17].

A different approach has been introduced by Gold and colleagues [18]. Their database contains standardized results for carcinogenicity for more than 1200 chemicals; it reports the results for carcinogenicity on rat and mouse, expressed in term of the parameter TD50 which is the chronic dose rate which would give half of the animals tumors. This database refers only to animal, and this is another major difference between the IARC database and the Gold's one.

We used both kinds of classification: categorial and continuous.

On animals there are more data, and as a consequence this information is more detailed. For this reason we used also this information to recognize toxic fragments. For ANN we used the TD50 as output.

The above approach may be limited in its application to man, since it is strongly related to the activity in animal. We extended its applicability to man using rule learning, training it with the IARC classification.

The residue approach: definition and search

Many toxicologists consider the presence of given fragments in the molecule as an indication of potential carcinogenicity. Individuating all the fragments and modulating in a detailed way their activity is necessary. For instance, while aniline has a carcinogenic potential, p-phenylendiamine does not [19]. Similarly, 2-naphthylamine is a quite potent carcinogen (IARC class 1), but 1-naphthylamine has a very low if any activity (IARC class 3) [17]. These examples show that to simply rely on the presence of an aromatic amino group may be misleading.

Much more fragments are necessary. We studied this topic for all the aromatic compounds with at least a nitrogen linked to the aromatic ring (Ar-N compounds).

Carcinogenicity of aromatic compounds

Ar-N compounds contain a large number of chemicals, many of them carcinogens. In order to define the fragments responsible for carcinogenicity, we used several bibliographic sources [17-20].

The Ar-N group is divisible into 10 chemical classes with different mechanisms of bioactivity, further splitted into some subclasses. Subclasses are defined by the following criteria:

- presence of the same atom or substituent or chemical structure in a fixed position relative to the the Ar-N bond;
- affinity of chemicals in terms of TD50 values;
- toxicological, target tissue and/or IARC class.

The list of the knowledge base is shown in Table 1.

Table 1. Ar-N compounds divided into classes and subclasses.

1) PRIMARY AMINES
a- Monocyclic aromatic primary amines
b- Pentaatomic heteroaromatic primary amines
c- Hexaatomic heteroaromatic primary amines
d- Biphenyl primary amines
e- Di- and triphenylmethane amines and analogues
f- 4- and 4,4'-Stilbenes
g- 2-aminofluorene and analogues
h- Condensed polycyclic primary aromatic amines 1
i- Condensed polycyclic primary aromatic amines 2
2) NITROCOMPOUNDS
a- Monocyclic aromatic nitro compounds
b- 2-nitro-5-furyl
c- Thio- and azo-pentaatomic nitro compounds
d- Condensed polycyclic nitro compounds 1
e- Condensed polycyclic nitro compounds 2
f- Miscellaneous nitro compounds
3) AZOCOMPOUNDS
a- Dibenzo azo compounds
b- 1-naphtho azo compounds
c- 2-naphtho azo compounds
4) HYDRAZINES
a- Hydrazines 1
b- Hydrazines 2
5) SECONDARY AMINES
a- Aromatic secondary aliphatic amines
b- Diphenyl secondary amines
c- Carbazole
d- Solfonic secondary amines
e- Purines
6) AMIDES
a- Monocyclic aromatic amides
b- Biphenyl amides
c- 2-acetylaminofluorene derivatives
d- Pentaatomic heteroaromatic amides
e- Hexaatomic heteroaromatic amides
7) TERTIARY AMINES
a- Monocyclic aromatic tertiary amines
b- Di- and triphenylmethane tertiary amines
c- N,N-dihydroxyethyl tertiary amines
d- Nitrogen mustards
e- Pentaatomic heterocyclic tertiary amines
8) C-NITROSOCOMPOUNDS
9) N-NITROSOCOMPOUNDS
10) ISOCYANATES

We implemented this knwledge as a two level structure, eventually with negative conditions for each layer.

- *FIRST SEARCH LEVEL*: search of the characterizing element of the subclass (also named "body" of the residue); this element is composed by the class general discriminant structure of nitrogen fragment and the general aromatics structures bonded to that group (e.g. a mono or bicycle).
- *FIRST INHIBITION LEVEL*: it has been originated to solve the problem of chemical groups that, even if related to the structure of the subclass, are not carcinogens or anyway toxicologically not similar. Another problem is the exclusion of groups of compounds that in fact belong to other subclasses. For example implementing Monocyclic Aromatic Primary Amine we had to exclud the Biphenylic Amines, that belong to another specific subclass of Primary amine.
- *SECOND INHIBITION LEVEL***:** the second search level does not discriminate between compounds not cancerogenic or not yet toxicologically defined and carcinogenic ones that belong to the same structural subclass: this second inhibition level is useful to exclude a specific compound. For example in the subclass of Monocyclic Aromatic Nitro Compound, 2,4-dinitrophenol has to be excluded because not cancerogenic.

This two-layer structure allows an easy introduction of new subclasses or even classes of compounds, and exclusion of compounds or entire groups of them; it is also possible to modify only a portion of the structure.

Each fragment is associated with a category expressing the level of toxicity. The system reports the highest level obtained and the residue responsible for the activity; in other words, if more than a toxic residue is present, the program select the most active. For the definition of the five “carcinogenicity levels”, three parameters have been considered:

1. the TD50 of the molecule;
2. the level of carcinogenicity ascribed to the fragment contained in the molecule (this has been defined taking into account the quantitative information available, averaging the evaluation for each fragment on all the molecules containing the substructure);
3. the classification or the evidence of carcinogenicity given by the IARC, IRIS, HSDB, NTP, RTECS [17, 18] databases.

All chemical structures are represented by graphs. The COSMIC format has been chosen to describe the molecule. This decision allows us to use atom hybridization instead of information on atomic bonds, with two positive consequences:

1. All bonds are equals. The chemical information is hidden in the nodes and so the implementation of the search algorithm is easier.
2. Hydrogens are left out. The molecular graph is smaller.

Molecules and chemical structures are represented by adjacency lists. This paradigm is a development of the adjacency matrix where the rows are replaced by linked lists, one for each vertex of the graph. For any given list, *i*, the nodes in the list contain the vertices that are adjacent to

vertex *i*. Notice that all the chemical information, atomic number and hybridization, are hidden in the nodes, while the links are indifferentiated.

Search

The search of a fragment in a molecule can be formalized, in graph theory, as a *subgraph isomorphism problem* in which we have to find *all* the isomorphisms between a given graph and subgraphs, where a graph $\mathbf{G}_a$ is *isomorphic* to a subgraph of a graph $\mathbf{G}_b$ if and only if there is a 1 to 1 correspondence between the node sets of this subgraph and of $\mathbf{G}_a$ that preserves adjacency. This problem is, in general, NP-Complete.

Our search has been divided into two parts: the first search level is performed by finding all the possible isomorphisms between the considered sub-structure and the molecule. Note that looking for all the isomorphisms is necessary to check all the possible spatial configurations of the residue in the molecule. This target is reached using the Ullmann's algorithm, modified to manage hydrogens and wildcards.

Afrter finding a first level sub-structure, the second part of our search procedure checks if positive and negative conditions are true. In other words, for each isomorphism found, we pass to consider the next structure levels. This second part of the search procedure is based on two important hypotheses:

- there is only one instance of the sub-structure belonging to the second search level that is linked to a first level isomorphism.
- moreover we need to find only one isomorphism of the inhibition levels to infer that the exclusion is necessary.

For this level we used a backtracking technique which performs an atom-by-atom search. This simple algorithm is sufficient because the sub-structures associated with the inhibition levels and the second search level are very simple and composed by few atoms.

If the backtracking finds a second level structure and no inhibition, we have found one instance of the residue in the molecule.

Summing up, the search of residues is:

1. Search a new isomorphism of residue first level structure and molecule (Ullmann's algorithm);
2. IF there is no other isomorphism THEN GOTO 4;
3. FOR each isomorphism found THEN
   - 3.1. IF (Check first level inhibition = TRUE) THEN GOTO 1;
   - 3.2. Search second level structure;
   - 3.3. FOR each second level structure found
     - 3.3.1. IF (Check second level inhibition = FALSE) THEN one instance of a residue is fuond;
   - 3.4. GOTO 1;
4. END;

The ANN

Molecular descriptors used as input

From the chemicals included in the Gold's database 104 molecules presenting an aromatic ring and a nitrogen linked to the aromatic ring have been chosen to train the ANN. We used molecular descriptors as input for the ANN.

Molecular descriptors represent different structural attributes of the moleculesTheir use constitues a different approach from that of using substructures. In the case of substructures the problem is that the rest of the molecule is not considered, but some general parameters may influence biological activities.

Different kinds of molecular descriptors are reported in the QSAR literature. We tried to use many of these parameters. The programs VAMP 6.1 (Oxford Molecular Limited, England) has been used for the quantum-chemical and thermodynamic calculations, HazardExpert 3.0 (Compudrug, Budapest) for logD calculation, TSAR version 3.0 (Oxford Molecular Limited, England) for all the other descriptors.

The following 34 descriptors have been calculated: molecular weight; molecular volume, logD at pH 2, 7.4 and 10; three principal moments of inertia and three principal axes of inertia; Wiener, Randic and Balaban topological indices; three Kappa and three Kappa alfa shape indices, flexibility index; five ChiV connectivity indices; ellipsoidal volume and electrotopological sum; HOMO; LUMO; dipole moment; total energy; polarizability, heat of formation.

A selection was necessary in order to avoid an excessive time for training the network. The criterion adopted was that of obtaining the most information and the least correlation between input variables**.** Principal component analysis (PCA), one of the main techniques for the multivariate analysis of data, has been used.

The main differences within the set of 104 molecules resulted explained by the descriptor total energy and by a pool of descriptors including the topological, geometric and electrostatic. Dipole moment, the topological index of Balaban, the quantum-chemical HOMO and LUMO descriptors and the logD parameters explained another source of differences between the molecules. Considering these results of PCA and removing corrrelating descriptors using the correlation matrix obtained by PCA, a final set of 13 descriptors has been selected. The reduced set is the following: molecular weight, HOMO, LUMO, dipole moment, polarizability, Balaban, ChiV3 and flexibility indices, logD at pH 2 and pH 10, third principal axis of inertia, ellipsoidal volume, electrotopological sum. It is interesting to note that descriptors of different nature have been selected after PCA, meaning that no single category of descriptor was a source of complete information. We note that logD was obtained using an ES, and in this sense the ANN includes as input a value obtained from a symbolic program.

Output for ANN

The parameter TD50 created by Gold has been adopted for the output. The output has been derived from a transformation of the TD50 according to the following formula:

$$\text{output} = \text{Log}\,(MW*1000/TD50)$$

This transformation has been adopted in order to have a more continuous output space.

ANN experiments

Data pretreatment was needed in order to have a homogeneous range of variance of descriptors. Data were scaled between 0 and 1. The validation set was scaledon the basis of the scaling of the training set.

All the simulations were performed using MBP v 1.1 [21]. The working parameters were the following:

- weight initialized tor YPROP: $K_a = 0.7, K_d = 007$

The algorithm stops itself when it encounters one of the followings:

- gradient lower than 10^{-6} [for too low values no improvement of Mean Square Error (MSE) happens]
- MSE equal to 0;
- maximum calculated difference between calculated and desired output equal to 0;
- maximum number of iterations reached.

Each network has been trained starting from 100 points random in the space, in order to minimize the probability of converging towards local minima. For the validation step the Leave-Two-Out approach was adopted.

MSE and R^2_{cv} resulting from 10000 iterations of the back-propagation ANN, using different numbers of internal neurons, showed that best results were obtained using four or seven neurons: R^2_{cv} was in both cases 0.69. The presence of outliers in the set, i.e. of observations which are so distant from the others to suspect that a different mechanism underlies them, has been supposed and investigated.

For the removal of the outliers a conservative approach has been adopted removing just the molecules which presented an error in validation higher than 0.2 in both the two best models. 12 molecules were identified as outliers and removed from the set. Results obtained after outliers removal showed clear improvement in the R^2_{cv} which became 0.82 (with 4 internal neurons). The major part (9 out of 12) of the outliers are molecules for which the experimental results for carcinogenicity were not statistically significant and an arbitrary value of 10^{31} was given in the Gold database. The major experimental evidences for these molecules tend to non carcinogenicity. The developed ANN presented therefore a lower prediction for non carcinogenic compounds. These compounds, however, are subject to major experimental uncertainties in the database we used. Carcinogenic compounds were instead correctly predicted, thus assuring the capacity of the network of avoiding false negatives. Other ANN architectures have been tried. Counter propagaton on the complete set gave only 0.61 for R, and 0,72 after outliers removal.

Combining the two information: hybrid system

There has been a considerable amount of research in integrating connectionist and symbolic processing. While such an approach has clear advantages, it also encounters serious difficulties and challenges. The hybrid approach is premised on the complementarity of the two paradigms and aims at their synergistic combination in systems comprising both neural and symbolic components. However, it is still under discussion which engineering methodology to apply for effectively developing hybrid systems.

The results we obtain from the two parts of the prediction should now be combined. The general target is to integrate both the components in order to maximize the predictive power of the system.

In a previous study we described a different hybrid system, in which the program was able to recognize the chemical class of a given compound and then to apply the toxicity rules defined for that chemical class [22]. Also in that case the inputs to the system were the molecular descriptors. The classification module assigned a chemical class to the compound, enabling the system to call only the appropriate model. These two modules gave as final result: a number, representing predicted toxicity, a class (active/non-active), and the explanation.

In the present study we changed the architecture of the program: indeed, the output of the two (different) modules, based on substructures and on ANN, were used within a third module giving the final prediction.

The target classification is the one proposed by IARC. From the studied molecules, 67 have a IARC classification. 43 of them are in class 3 (no definitive risk assessed) and 0 of class 4 (no risk). It was considered impossible to directly use the IARC classification, and we decided to split some semiquantitativeclasses according to the following criteria:

- to obtain 5 classes, 1 to 5, from lower to higher risks, based on the TD50 values;
- to check the presence of each residue in the molecules under study;
- to give to each residue a toxicity class obtained as the mean of the toxicity of the molecules where it was found;
- to assign to the molecule the maximum toxicity obtained from the residues and ANN module.

We have used different tree construction programs on our data set. The first is C4.5 which makes use of the maximization of the entropy gain, and build hyper-rectangular in the attributes space. The second one was CART which builds binary trees. A recent evolution of CART is OC1 [23] available on the Internet. It uses a random perturbation of parameters to escape from local minima. We tested the three programs on our data set, and obtained similar performances using the leave-one-out method, as illustrated in Table 2.

Table 2. Prediction obtained with the rule induction systems
(accuracy %)

	C4.5	CART	OC1
Training	93.3	88.5	90.2
Validation	81.9	85.5	82.8

Discussion and conclusions

Our study wants to contribute to the understanding of the possibilities to predict carcinogenicity.

The present study represents an example of a hybrid system, combining ANN and a system based on residue recognition. The strenght of the ANN may in theory be in its capacity to find the link between input and output, even in the case of unknown relationship.

So far ANN has been used in limited cases for carcinogenicity prediction. Ghoshal [12] used ANN for 9-nitroanthracenes and some heterocyclic compounds, with a correlation index r of 0.877 or 0.919, after removing the outliers. Vracko [13] obtained an r of 0.74-0.76 after removing the outliers, for a set of aromatic compounds, belonging to different chemical classes. Our study gave results comparable with those by Ghoshal and Vracko.

On the other hand, Benigni and Richard [14] had poor results, in a study using 280 compounds of various kind. However, Benigni and Richards tested their net on a much more heterogeneous population of molecules.

Our research confirms the feasibility of an ANN for carcinogenicity for chemicals. A valuable characteristic of our ANN is that it seems to correctly predict carcinogenic compounds, while it is less accurate in the prediction of non active compounds. The latter compounds, however, are subject to major experimental uncertainties.

It is likely that ANN alone cannot solve problems linked with carcinogenicity prediction. There are cases in which molecular descriptors, or at least those used by us, were not able to disciminate certain compounds. In this case an approach based on the residues can simply discriminate between the two chemicals. This is a clear example of the possibilities offered by hybrid systems.

Aknowledgements. We acknowledge the financial support of the European Commission (ERB-CP94 1029 until 1998 and ENV4-CT97-0508 since 1998) and the NATO (CRG 971505), since 1998.

References

1. E. Benfenati and G. Gini, "Computational predictive programs (expert systems) in toxicology", *Toxicology*, 119:213-225, 1997.
2. J.C. Dearden, M.D. Barratt, R. Benigni, et al., "The development and validation of expert systems for predicting toxicity", *ATLA*, 25:223-252, 1997.
3. R.D. Combes and P. Judson, "The use of artificial intelligence systems for predicting toxicity", *Pestic. Sci.*, 45:179-194, 1995.
4. HazardExpert is supplied by CompuDrug Chemistry Ltd, H-1395 Budapest.
5. H.S. Rosenkranz and G. Klopman, "New structural concepts for predicting carcinogenicity in rodents: an artificial intelligence approach", *Teratog. Carcinog. Mutag.*, 10:73-88, 1990.
6. G. Klopman and H.S. Rosenkranz, "Approaches to SAR in carcinogenesis and mutagenesis. Prediction of carcinogenicity/mutagenicity using MULTI-CASE", *Mutat. Res.*, 305:33-46, 1994.
7. K. Enslein, V.K. Gombar and B.W. Blake, "Use of SAR in computer-assisted prediction of carcinogenicity and mutagenicity of chemicals by the TOPKAT program", *Mutation Res.*, 305:47-61, 1994.
8. V.K. Gombar, K. Enslein and B.W. Blake, "Carcinogenicity of azathioprine: an S-AR investigation", *Mutation Res.*, 302:7-12, 1993.
9. D.M. Sanderson and C.G. Earnshaw, "Computer prediction of possible toxic action from chemical structure; the DEREK system", *Hum. Exp. Toxicol.*, 10:261-273, 1991.
10. J.E. Rindings,. M.D. Barratt, R. Cary, et al, "Computer prediction of possible toxic action from chemical structure; an update of the DEREK system", *Toxicology*, 106:267-279, 1996.
11. Y-T. Woo, D. Lai, M. Argus and J. Arcos, "Development of structure-activity relationship rules for predicting carcinogenic potential of chemicals", *Toxicology Letters*, 79:219-228, 1995.
12. N. Ghoshal, S.N. Mukhopadhyay, T.K. Ghoshal and B. Achari, "Quantitative structure-activity relationship studies of aromatic and heteroaromatic nitro compounds using neural network", *Bioorganic & Medicinal Chemistry Letters*, 3:329-332, 1993.
13. M. Vracko, "A study of structure-carcinogenic potency relationship with artificial neural networks. The using of descriptors related to geometrical and electronic structures", *J. Chem. Inf. Comput. Sci.*, 37:1037-1043, 1997.
14. R. Benigni and A.M. Richard, "QSARS of mutagens and carcinogens: two case studies illustrating problems in the construction of models for noncongeneric chemicals", *Mutation Res.*, 371:29-46, 1996.
15. Srinivasan, A., Muggleton, S. H., Sternberg, M.J.E., King, R. D., Theories for mutagenicity: a study in first-irder and feature-based induction, *Artificial Intelligence* , 85:277-299, 1996.
16. Srinivasan, A., King, R. D., Muggleton, S. H., Sternberg, M.J.E., The predictive toxicology evaluation challenge, Proc.IJCAI 1997, 4-9 , 1997.
17. World Health Organization - International Agency for Research on Cancer (1987). IARC monographs on the evaluation of carcinogenic risks to humans. Supplement 7.
18. L. S. Gold, C. B. Sawyer, et al., "A carcinogenicity potency data base of the standardised results of animall bioassays", *Env Health Perspect*, 58:9-319, 1984.,
19. J. Ashby and R.W. Tennant, "Definitive relationships among chemical structure, carcinogenicity and mutagenicity for 301 chemicals tested by the U.S. NTP", *Mutation Res.*, 257:229-306, 1991.
20. World Health Organization-International Agency for Research on Cancer. IARC monographs on the evaluation of carcinogenic risks to humans, Vols 1, 3, 8, 9, 12, 13, 16, 19, 24, 26, 27, 30, 31, 33, 36, 39, 40, 46, 48, 50, 51, 56, 57.
21. D. Anguita, Matrix Back Propagation v 1.1: user's manual. 1993.
22. G. Gini, V. Testaguzza, E. Benfenati, R. Todeschini, "Hybrid toxicology expert system: architecture and implementation of multi-domain hybrid expert system for toxicology", *Chemometric and Intelligent Lab System*, 43:135-145, 1998.
23. Murthy, S. Kasif and S. Salzberg, "A system for induction of obliques decision trees", *J. of Artificial Intelligence Research*, .2:1-32, 1994.