Integrating rules and neural nets for carcinogenicity prediction

Giuseppina Gini,
Marco Lorenzini

DEI, Politecnico di Milano

Milano, Italy
gini@elet.polimi.it

Emilio Benfenati,
Raffaella Brambilla,
Luca Malvè
Istituto di Ricerche Famacologiche
"Mario Negri"
Milano, Italy
benfenati@marionegri.it

Abstract
One approach to deal with real complex systems is to use more techniques in order to combine their different strengths 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 and ANN. We defined chemical structures of fragments responsible for carcinogenicity according to human experts, developing a module able to recognize these fragments into a chemical. Furthermore, we developed an ANN, using molecular descriptors as inputs to predict carcinogenicity as a numerical value. Finally, we developed an automatic learning program to combine the results into a classifications of carcinogenicity to man.

1. Introduction

Chemicals are responsible for many tumors, and industry is required to take into account carcinogenicity of the chemicals used and produced. However, the experimental tests on chemicals last for years, are costly and require the use of animals, with the consequent ethical problems. Considering the importance of the goal, it is mandatory to develop predictive systems [1, 2, 3, 4].

Carcinogens are listed in classes by international agencies. The International Agency on Research on Cancer (IARC) considers four classes: (1) contains the compounds recognized as carcinogenic to man, (4) the compounds which are not carcinogenic; in the middle the other compounds are split in classes as probably or possibly carcinogenic (2A and B), or unknown (3). This classification combines the experimental evidences with the amount of epidemiological knowledge available.

A different approach has been introduced by Gold [5], with the construction of a data set that contains standardized results for carcinogenicity for more than 1200 chemicals expressed in term of TD50 (the chronic dose rate which would give half of the animals tumors within standard experiment time).

In the present study we tried a new approach, combining different systems into a hybrid architecture. We recognized toxic residues predicting a class of toxicity, we used ANN to predict TD50, finally we used a symbolic rule induction program to merge the information from the two sources.

2. Carcinogenicity and its Prediction

On an operational basis, an agent can be considered carcinogenic when an alteration in the frequency or intensity of exposure to this agent is followed by a change in the frequency of occurrence of one or more type of cancer. The great majority of factors involved in the carcinogenic process are of exogenous origin. They are therefore defined "environmental carcinogens", but this term should not be confused with the term "environmental pollutants". For instance, a great variety of dietary constituents might be significantly associated with cancer induction. The exposure to chemicals represents anyway one of the main causative agent of cancer.

2. 1. Mechanisms of action

Many different kinds of tumors are described, and toxicologists have studied the relationship between a given structure and the carcinogenic effect, comparing similar structures. This is not an easy task, because the tumor is an effect mediated by many mechanisms, often antagonistic. For instance, nitrosoamines are known carcinogens, but their effect are shown in different tissues, for different compounds. Furthermore, to consider effects produced by similar compounds, we should define what is similar.

2. 2. Data and databases

Information on cancer studies are available in several databases.

- *IARC database.* It is a complete database on the evaluation of carcinogenicity, produced by the International Agency for Resesarch on Cancer (IARC).

- *CPDB (Carcinogenic Potency Database.* It is a widely used resource on the results of long term animal cancer tests for 1298 chemicals. It is available from L. Gold, University of Berkeley.
- *NTP chemical health and safety data* Information on effects on health and safety of chemicals are collected in this databank.
- *RTECS (Registry of Toxic Effects of Chemical Substances).* It provides toxicity information on over 132480 substances. It is created by the U.S. National Institute of Occupational Safety and Health. The database is available on-line (url: http.cdc.gov/niosh/rtecs.html.
- *IRIS (Integrated Risk Information System.* It contains EPA information on over 600 chemicals. It is part of the integrated system TOXNET, available online.
- *CCRIS (Chemical Carcinogenic Information System).* Sponsored by the National Cancer Institute (NCI), CCRIS contains scientifically evaluated data on over 7000 chemicals. It is part of TOXNET,.
- *IRPTC-PC Chamicals Databank.* It is produced by the United Nations Environment Programme.
- *IUCLID (International Uniform Chemical Information Database).* It has been developed from the European Union and is an ORACLE database.

2. 3. Definition of the data set under study

Depending on the goal of the research, the choice of the data to use varies. All the previously reported databases have been considered for selecting the output of this project.

IARC database provides for several hundreds of substances a classification of carcinogenicity, which has been judged useful for the implementation of the last module of the hybrid system, in order to pass from an evaluation of carcinogenicity towards animals to one towards humans.

RTECS contains information on the tumorigenic effects of substances, but these represent just the results of single researches conducted, not a synthesis of the results achieved by all studies. Therefore it does not represent in this case a basic tool, but it has been used in some cases with a confirmatory purpose.

IRIS database presents interesting information on the EPA classification and on the epidemiological evidences on which it is based; the problem for using it was that data were available for not a very large set of compounds.

NTP database contains a lot of data on toxicity, but the only synthetical evaluation is the definition of a hazard class for each compound, which seems more properly to indicate a risk relative to its delivery from the producing place.

Gold's database contains data for more than 1200 chemicals; it reports for each substance the results for carcinogenicity, expressed in term of the parameter TD50 which is the chronic dose rate which would give half of the animals tumors within some standard experiment time - the "standard lifespan" for the species. The Carcinogenic Potency Database by Gold et al. includes, for each chemical, results of experiments coming from the literature and from NTP studies for rat and mouse. For each experiment the numerical index TD50 and an evaluation of the statistical significance of the result is provided. Not a single TD50 value is reported for each substance and animal, but the series of eventual values reported for each substance is ordered from the most potent and significant TD50 to the least. The huge amount of data contained in the Gold's database and the quantitative values represented two important advantages for this project. This database was therefore adopted as the basic tool to select the output parameter for the neural network.

In the present study, we chosen for each chemical under study the lowest (i.e. most potent) TD50 with good significance ($p<0.01$). For a purpose of homogenicity all data choosen refer to experiments on the mouse.

3. Features Describing A Molecule

Molecular descriptors are often used for toxicity prediction [1]. Different kinds of descriptors have been considered in the literature. They can be grouped into seven main classes of descriptors: physico-chemical, constitutional, topological, geometric, electrostatic, quantum-chemical, thermodynamic.

Physico-chemical descriptors - include parameters like molecular weight, log D (the hydrophobic degree of a molecule at different pH), molar refractivity.

Constitutional descriptors - simply reflects the molecular composition of the compound (i.e. number of atoms, bonds, rings), without using the geometry or electronic structure of the molecule.

Topological descriptors - describe the atomic connectivity in the molecule. Wiener, Randic, Balaban, connectivity and shape indices are representative of this kind of indices.

Geometric descriptors - derive from calculations based on the 3D structure of the molecule. Moments of inertia, molecular volume, surface area are examples of these descriptors.

Electrostatic descriptors - reflect the charge distribution of the molecule. The electrotopological sum and

charged partial surface area are examples of these descriptors.

Quantum-chemical descriptors - result from semi-empirical calculations and describe electronic properties (i.e. HOMO, the highest occupied molecular orbital and LUMO, the lowest unoccupied molecular orbital), the configuration of the energy of the molecule and its polarizability.

Thermodynamic descriptors - represent thermo-dynamic properties of the molecule, like heat of formation, entropy and enthalpy.

3.1. Computation of descriptors

In the present study the Program NEMESIS has been used to draw the molecular structures and derive the molecular format CSSR. The CSSR format has been successively converted into the ARC format by the program VAMP (version 6.1), which has then been introduced into the program TSAR.

The following descriptors have been computed.

Physico-chemical: molecular weight; molar refractivity; logD at pH 2, 7.4, 10.

Geometric: surface area; molecular volume; moments of inertia and principal axes of inertia.

Electrostatic: electrotopological sum.

Topological: Wiener, Randic, Balaban indices; shape indices (kappa and kappa alfa indices, flexibility, ellipsoidal volume); connectivity indices (ChiV indices).

Quantum-chemical: HOMO, LUMO; dipole moment; total energy; polarizability.

Thermodynamic: heat of formation.

Programs Used : VAMP version 6.1, Oxford Molecular Limited, England, for the quantum-mechanical calculations.; HAZARD EXPERT version 3.0, Compudrug, Budapest, for logD calculation.; TSAR version 3.0, Oxford Molecular Limited, England, for the calculation of all other descriptors.

3.2. PCA to reduce the number of descriptors

Descriptors selection was necessary in order to avoid an excessive time for training the network. Descriptors were selected in order to obtain the most information and the least correlation between input variables. Principal component analysis (PCA) has been used, since it permits to:

- evaluate correlation and relevance of variables
- see the multivariate information characterising the objects in a two dimensional orthogonal space
- synthetising data eliminating noise

Principal components are derived from the calculation of eigenvalues and eigenvectors of the covariance matrix of data.

The loading plot graph permits to identify the role of each variable towards the two principal components considered and their direct or inverse correlations. The couple of loadings associated to each variable towards two selected components constitutes their coordinates in the space described by the two components. Variables with loadings closer to the unity (in absolute value) have an higher influence on one component respect loadings near 0. Variables are represented by arrows in the loading plot. Closer to each other are the arrows in the graph and more correlated are the variables. Furthermore, variables in opposite direction have also a high but inverse correlation.

In this study, the first principal component of PCA analysis accounts for the 62% of the total variability, results mainly associated with the descriptor total energy (tenergy in the graph) and with a pool of correlated descriptors (on the left of the graph). The further selection of descriptors, reduced this redundant information without eliminating important elements for the explanation of the system. Dipole moment, HOMO, balaban, and logD at the different pH are associated instead with the second principal component, which accounts for the 8% of the total variability. In particular logD at different pH results an informative parameter as it explains a variation of the system not already related to other kind of descriptors, even if in this case the variation accounted by the second component is small.

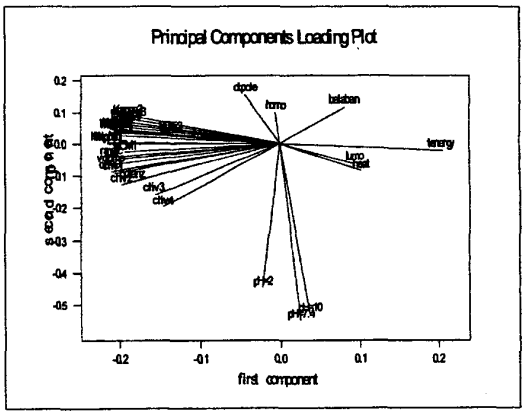


Figure 1 : Loading plot of all the descriptors on the first two components.

The scores plot permits to analyze the behavior of objects towards the components and their similarities. Once defined two components of interest, the coordinates of each object are the two scores on the two components selected. Clusters of objects and outliers

can be noted with the aid of this graphical representation.

In the present study, the population of individual molecules has a higher density towards the origin of the cartesian plane; however it seems that there are not two different populations. Considering the correlation matrix obtained by PCA, descriptors with pairwise correlation exceeding 0.9 were investigated, and one of these two descriptors was removed from the pool. The set of descriptors was reduced from 34 descriptors to a final number of 13. These numbers were considered more suitable for the performance of the network, considering that the objects were 104.

Besides the criterion of eliminating correlated descriptors, the first criterion of selection was the importance of the variable on the first four components (which overall accounted for 83.5% of the total variance of the system, giving more weight to the four principal components in a decreasing order). A third criterion was the manteinance of a descriptors pool well representing the different ways of parametrizing a molecule, i.e: the physico-chemical, electronic, topological, quantum-chemical information.

4. The Neural Approach For Prediction

Backpropagation neural networks [6] has been adopted in this study to implement the quantitative prediction of carcinogenicity; for details see [7].

From the Gold's database 104 molecules presenting an aromatic ring and a nitrogen linked to the aromatic ring have been chosen. We computed molecular descriptors and selected the best as described in Section 3

The output is toxicity expressed as

Log (MW*1000/TD50)

where MW is the molecular weight.

For validation the N/2-fold-crossvalidation has been used. Results of 10000 iterations of the BPNN, using different number of internal neurons, gave as best values for 4 neurons: R^2_{cv} 0.69.

The presence of outliers in the set has been investigated; 12 molecules were identified as outliers and removed. Results after outliers removal showed clear improvement in the R^2_{cv} which became 0.824. The majority (9 out of 12) of the outliers is molecules for which the experimental results were not statistically significant and an arbitrary 10^{31} value was given by Gold.

The major experimental evidences for these molecules tend to non carcinogenicity. The ANN model presented therefore a lower prediction for non carcinogenic compounds. Carcinogenic compounds were instead correctly predicted, thus assuring the capacity of the network of avoiding false negatives.

5. Toxicity Rules as Residues Finding

The presence of some particular fragments in the molecule is often related to the occurrence of the carcinogenic effect. For instance, many nitrosoamines have been found carcinogenic. From these observations, a few software have been developed, trying to encode the carcinogenic effect within chemical fragments, to be searched in the chemical compound to be evaluated [1]. We used the residue approach for aromatic compounds with at least a nitrogen linked to the aromatic ring (Ar-N compounds). We studied these particular chemicals because it has been found that these compounds may present problems with the residue approach. This was shown with a version of HazardExpert (HE) [8]. HE is an expert system predicting toxicity on the basis of a list of fragments responsible for at least one of seven toxic endpoints. Each fragment is associated to numbers (from 0 to 100) expressing the level of toxicity. The system reports the highest level obtained for each toxic endpoint and the residue responsible for the activity; if more than a toxic residue is present, the program select for each endpoint the most active. HE gives as output histograms where, in the case of oncogenicity, the IARC classes have been indicated; thus it is easy to compare the predicted activity with the activity reported by the IARC.

Out of 456 molecules, for 62 (14%) the prediction was correct, instead for the others there was a range of errors between -3 and +2 IARC units. The most represented class of errors were those characterized by -1 (143 compounds, about 31%) and +1 unit (143 compounds, about 31%) followed by those defined by -2 (74 compounds, 16%), -3 units (17 compounds, about 4%) and +2 IARC units (17 compounds, about 4%). 348 (76%) chemicals were considered satisfactorily predicted by HE, including both correct prediction and prediction with errors of + or - 1 IARC unit, 108 (24%) were not.

The worst results, with a delta of -3, were further evaluated. Within these chemicals there are aromatic amines (AA). CompuDrug developed a new version of HE, version 3.0, which avoids this problem: with HE 3.0 all AA are correctly recognized according to the presence of the presumed toxic fragment.

These examples show that to simply rely on the presence of an aromatic amino group may be misleading. HE version 2, did not contain this capacity of discrimination. A possibility to improve the

approach is to better detail the residues in order to follow in a closer way the variegate activities of compounds containing the same residue.
This is the reason why we selected the compounds used here and in the ANN part (same chemicals). The compounds with the Ar-N group were divided into 10 chemical classes further split into subclasses. In this way a greater detail of the structures has been done.
The structure for implementing this knowledge is a two-level structure. For each subclass we defined a first level structure, which identifies the chemical fragment common to each residue belonging to the subclass. A second level of structures further specify each residue. Two correspondent inhibition levels have been introduced to detail situations where the fragment found has not to be considered.
- *First level*: identifies the structure of the nitrogen fragment characterizing the class and the aromatics structures bonded to that group.
- *First inhibition level*: it solves the problem of compounds that, even if related to the structure of the subclass, are not carcinogens or have been ascribed to another subclass.
- *Second level*: the second search level permits the identification of a specific compound or small groups of compounds that refer to the same subclass but differ for some specific elements bound to the nitrogen group and/or to the aromatic structure, and suspected to be involved in the carcinogenicity process.
- *Second inhibition level*: this second inhibition level is useful to exclude a specific compound or a small group of compounds.
Each fragment is associated with a category expressing the level of toxicity. Our system reports the highest level obtained and the residue responsible.
Since graph theory is used to represent the structures, the search of a fragment in a molecule is a *subgraph isomorphism problem.* A graph is *isomorphic* to a subgraph of a graph G_β if and only if there is a one-to-one correspondence between the node sets of this subgraph and those of G_α that preserves adjacency. The computational complexity of this problem is, in general, NP-Complete. The search has been divided into two parts: the first search is performed by finding all the possible isomorphisms between the structure considered and the molecule, with the Ullmann's algorithm [9], modified to manage hydrogens and wildcards. After finding a first level structure, the second part of the search procedure checks positive and negative conditions, using a backtracking technique. If a second level structure and no inhibition are found, a residue is found. Each fragment is associated with a category expressing the level of toxicity as a combination of TD50, the level of carcinogenicity ascribed to the fragment, and the classification given by the recognized databases, as IARC.

6. Combining the two Modules: Hybrid System

Above we have seen two completely different approaches to predict carcinogenicity. Table 1 shows the main characteristics and differences of the two approaches.

Table 1: The approaches we used to predict toxicity

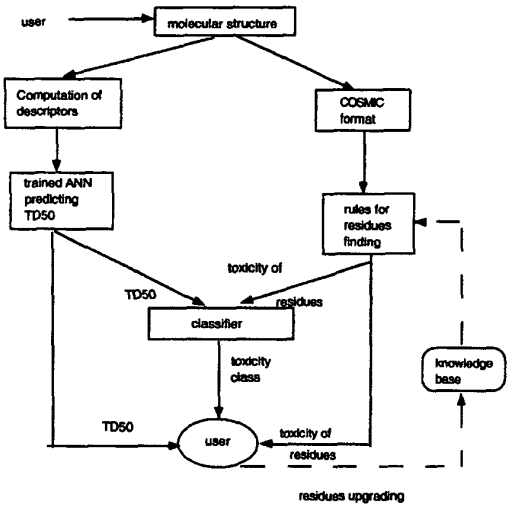
Approach	A	B
Databases used	Gold's	IARC, HSDB, etc
Carcinogenicity value	continuous	class
Chemical description	molecular descriptors	residues
software	ANN	graph theory

It is likely that these two approaches afford somewhere different predictions. The problem is how to integrate the results. The results we obtained from the two parts of the prediction have been combined with a third module dedicated to the classification. Given the output of the residue research, and the expected TD50, we wanted to extrapolate a combined prediction of the human carcinogenicity.
We split some classes of the IARC classification to define 5 classes, from lower to higher risks, based on TD50 and the evidence of the found residue. We tried C4.5, CART, and OC1 with leave-one-out; results are with an accuracy of more than 85%.In Figure 2 we see the architecture of the prototype.

7. Discussion and Conclusions

The main results of our study are two: a general and a specific one. On a general point of view we showed an architecture combining the information arising from very different sources, in a system which gives a good result, also in validation. On a specific point of view we merged the information contained in the residues with that contained in the whole molecule. Indeed, it is well known that some residues have been associated to the carcinogenic effect, but it is also known that some general features of the molecule can modify the physico-chemical property of the compound, changing completely the final toxic effect. In may cases the actual mechanisms of the toxic effect are not known,
and it may be useful to take advantage of all possible approaches, combining human knowledge (when

Figure. 2 - Architecture of the hybrid system



available, for instance coded in residues) and ways to extract automatically the knowledge, such as in the case of ANN. It is interesting to note that ANN have not been able to extract all knowledge, but this is hardly achievable, at least using as information the molecular descriptors. Indeed, there are two compounds, p-aminoanisole and o-aminoanisole, in our data sets which are very similar, on a point of view of the chemical descriptors, but they show very different toxicity. ANN cannot distinguish them [7]. However, the residue approach can easily recognize them. These two compounds are not isolated cases. For instance, while aniline has a carcinogenic potential, p-aminoaniline does not. Similarly, 2-naphthylamine is a quite potent carcinogen (IARC class 1), and 1-naphthylamine has a very low if any activity (IARC class 3).

It is quite easy, with our approach, to include results from other separate models, and integrated in a more powerful tool. In this sense our method is a pilot one, and should be improved using separate models as inputs of the hybrid system. In any case, an interesting part is that in this general philosophy all models can work predicting the toxic activity on the simple basis of the chemical structure, without the need of the real compound, and of laboratory experiments. In this way the advantage is that the method works also if the compound has not been tested and not even prepared. Indeed, the preparation and the laboratory experiments may be long and expensive.

Our research confirms the feasibility of an ANN for carcinogenicity prediction. It is likely that ANN alone cannot solve all the problems linked with carcinogenicity prediction. In this case an approach based on the residues can help. An advantage of our architecture is that the output is not a binary classification as in several programs predicting toxicity. Moreover, we do not need biological data to predict carcinogenicity.

Acknowledgements

We acknowledge the financial contribution of the European Commission, projects COMET (ENV4-CT97-0508) and IMAGETOX (HPRN-CT-1999-00015).

8. References

[1] E. Benfenati and G. Gini, "Computational predictive programs (expert systems) in toxicology", *Toxicology*, vol.119, 1997, pp 213-225.

[2] HazardExpert, version 3.0. Compudrug Chemistry Ltd, Budapest, Hungary.

[3] 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.*, vol.371, 1996, pp.29-46.

[4] E. Benfenati, S. Pelagatti, P. Grasso, G. Gini. COMET: the approach of a project in evaluating toxicity. In: *Predictive Toxicology of Chemicals: Experiences and Impact of AI Tools. AAAI 1999 Spring Symposium Series*; Gini, G. C.; Katritzky, A. R., Eds.; AAAI Press, Menlo Park, CA, 1999; pp 40-43.

[5] http://potency.berkeley.edu/cpdb.html

[6] Anguita, Matrix Back Propagation v. 1.1 User Manual, 1993.

[7] G. Gini, et al. "Predictive Carcinogenicity: a Model for Aromatic Compounds, with Nitrogen-containing Substituents, Based on Molecular Descriptors Using an Artificial Neural Network", *J of Chemical Information and Computer Sciences*, Vol. 39, 1999, pp 1076-1080.

[8] E. Benfenati, M. Tichy, L. Malvè, P. Grasso, G. Gini, "Expert systems for toxicity prediction based on fragment recognition: evaluation of a commercial system and improved approaches". American Chemical Society Meeting, Las Vegas, Nevada, USA, September 8-12, 1997. Abstracts of paper, *COMP 136*

[9] J. R. Ullmann, "An algorithm for subgraph isomorphisms", *Journal ACM*, Vol. 23 - 1, 1976.