COMET: the approach of a project in evaluating toxicity

E. Benfenati, S. Pelagatti, P. Grasso, G. Gini^

Istituto di Ricerche Farmacologiche "Mario Negri", Milan, Italy
benfenati@irfmn.mnegri.it
^Dipartimento di Elettronica e Informazione, Politecnico di Milano, Milan, Italy
gini@elet.polimi.it

Abstract

The European Commission has funded COMET (Computerized Molecular Evaluation of Toxicity), a project to evaluate the possibility to predict a series of toxic and ecotoxic endpoints. This research will focus on the relationship between chemicals and their toxic and ecotoxic effects investigated through updated computerized approaches. The major objective of COMET is to extract as much as possible information from toxicity and ecotoxicity databases and suitable molecular descriptors, using advanced computer approaches such as fuzzy logic, ANN, EA, rule learning algorithms and AI.

Introduction

COMET is a three-year long project funded by the European Commission, which started in 1998. Five groups are involved in the project, as listed in Annex 1.
There are several points involved in the success of a model for toxicity prediction in ecotoxicology and toxicology. The major ones are:
1. the availability of a reliable and sufficiently large data base with toxicological data;
2. the treatment of the chemical information;
3. the software approach to build up the model.
The major objective of COMET is to evaluate the influence of these points and in particular of advanced software systems. This is quite difficult since these points are interrelated. For instance, the use of a software able to extract information even in presence of noise may in principle allow to use a data set wider, with a less stringent control of its reliability.
To address the key objective, we will study several topics. There are several "minor", related objectives:

- to check the availability of appropriate toxicological databases (for simplicity sake we use the term toxicology in its wider sense, including ecotoxicology);
- to check the reproducility of the data;
- to compare different softwares to calculate molecular descriptors;
- to evaluate the uncertainty of the different descriptors;
- to consider two approaches to deal with the chemical information: global molecular descriptors or fragments;
- to compare different software approaches, considering numerical systems, such as multivariate linear models (the classical Quantitative Structure Activity Relationship - QSAR - approach), advanced non linear systems (ANN, fuzzy logic), symbolic systems (AI);
- to evaluate the integrability of the different approaches and systems;
- to evaluate the integrability of toxicological data with chemical information to predict a different toxicological endpoint;
- to evaluate the usefulness of these predictive tools in real cases, such as in the case of plant protection products companies.

The above mentioned objectives reflet a number of activities; the resulting picture should allow to evaluate the feasibility and reliability of a given QSAR model depending on the data set, the chemical information and the informatic system.

Criteria for the selection of the areas to be studied

In order to better compare the modelling performances as a function of 1) the toxicological endpoint, 2) the chemical information and 3) the software approaches, we preferred to limit the variability of the molecule data set. Thus, as far as possible, we looked for a single or limited number of datasets of molecules characterized with many, hopefully all, the toxicity values to be modeled. Aquatic toxicity, toxicity on terrestrial organisms, acute toxicity on rat, mutagenicity, carcinogenicity, reproductive and teratogenic effects, and the extrapolation of toxicity to humans are the main endpoints which will be addressed within the project. The availability of toxicity data required a compromise between the number of compounds and the number of toxicities, due to a limited overlapping of the databases.
Pesticides have been selected in consideration of the availability of toxicity and ecotoxicity data and of the industrial interest for these compounds. Aquatic toxicity was initially considered. For carcinogenicity, preliminary studies were done on a different molecule data set, because carcinogenicity data were not found for most of the compounds selected for aquatic toxicity. Results for carcinogenicity are presented elsewhere (Gini, 1999).

Molecule selection

Structures and toxicity data of pesticides have been obtained from "the Pesticide Manual eleventh edition". It contains data on 759 compounds.
The molecule selection has been done on the basis of the presence of aquatic toxicity data. LC50 on rainbow trout (*Onchorynkus mykiss*) and daphnia (*Daphnia magna*) were the most common endpoints. About 200 molecules presented these two aquatic toxicity data. However, we eliminated pesticides for which the toxicities are referred to mixtures of diastereoiomers, because most of our molecular descriptors can distinguish among diastereoisomers; we kept data referred to a single diastereoisomer. We maintained pesticides with one chiral centre, even if mixtures. Polymers have not been considered. A set of 164 pesticides has been finally obtained.

Availability and reproducibility of the experimental data

We are interested in assessing the uncertainty associated with the individual areas involved in the definition of the model: the toxicity data, the description of the chemical information and the software aproach.
Aquatic toxicity data from three different databanks (HSDB, RTECS, ECDIN) have been collected for the comparison with data adopted (The Pesticide Manual). Almost a complete lack of data has been verified in RTECS and ECDIN (data for less than the 10% of the 164 molecules). Data for 39 molecules have been instead found in HSDB. A comparison between data from the Pesticide Manual and from HSDB database was therefore done. A poor correlation has been found. Some discrepancies (of more than one order of magnitude) were found. This is not unexpected in the case of toxicity data (van den Heuvel et al., 1990, Gold et al. 1984). A comparison made between two editions of the Pesticide Manual (issued in 1983 - PEST 1983 - and in 1997 - PEST 1997) and HSDB database has been considered. The variability found between these two editions and the bigger variability between HDSB and PEST 1997 than between HSDB and PEST 1983 indicates the presence of a time-related source of variability.
A check of the availability of data for the other selected endpoints has been done, considering the Pesticide Manual and the other sources available (the 11 databases contained in the TOMES Plus system from Micromedex Inc., USA).
Data availability results good (data for the 70-80% of the 164 molecules have been found) for toxicity on birds, acute and chronic toxicity on rat, and the acceptable daily intake.
Problems arose instead for mutagenicity, reproductive and teratogenic effects, since for the first endpoint data coming from each source are no more than 20% of the totality, and for the other two are even less or absent, respectively.

Descriptors

In order to have a general unbiased view of the different descriptors which can be involved in the modelling process, we wanted to have a large number of them to be used with the different successive programs. This approach is different from another one, in which specific descriptors are selected, for instance on the basis of the literature. However, in our case we may have the problem of noise, redundancy and casual correlations. A careful use of the variables and eventually a variable selection should be done.
Preliminary molecular modelling has been done using HyperChem to generate three-dimensional representations of the compounds. The three-dimensional structures have been refined with the PM3 Hamiltonian, a semiempirical method for energy minimisation of the geometry. Accurate three-dimensional representations of structures were necessary for the generation of descriptors dependent on geometry.
Most of the descriptors have been calculated by CODESSA 2.2.1: in particular
1. constitutional descriptors, depending on the number and type of atoms, bonds and functional groups, 38 descriptors (18 as discrete values);
2. geometrical descriptors, which give molecular surface area and volume, moments of inertia, shadow area projections and gravitational indices, 12 descriptors;
3. topological descriptors, which are molecular connectivity indices related to the degree of branching in the compounds, 38 descriptors;
4. electrostatic descriptors, such as partial atomic charges and others depending on the possibility of some sites in the molecule to form hydrogen bonds, 77 descriptors (3 as discrete values).
Quantum-chemicals descriptors, i.e. total energy of the molecule, HOMO and LUMO energies, ionisation potentials, heat of formation etc., have been calculated using MOPAC (with the PM3 Hamiltonian).
A class of descriptors largely used for QSAR studies, the logD, has been calculated by Pallas 2.1. These physico-chemicals descriptors are the expression of the lipophilicity of the molecule at various pH.
A total set of about 160 descriptors has been built.

Conformation influence

The variability of the molecular descriptors depends on several aspects: 1) the conformation of the molecule, 2) the minimization of the energy of the structure, and 3) the software used.
Studies on a representative subset of molecules (about 10% of the population) have been done using plausible conformers with low energy, obtained by successive rotations of fragments of the molecule. The geometry optimisation has been done using the PM3 Hamiltonian; descriptors have been calculated as previously described.
Results vary from molecule to molecule. Electrostatic

descriptors are the most dependent on changes of conformation (average variation: 20%). Constitutional and topological descriptors are invariant to conformational changes.
Geometric descriptors (particularly the moments of inertia) show a certain dependency to conformation, but variations are not important (less than 10%). As the differences in geometric descriptors depend directly on the shape of the molecule, molecules with more freedom degrees are more sensitive to conformational variation.
The intermolecular descriptor variance has been calculated and referred to the intramolecular variance. This analysis has shown that intermolecular variations are always significant, more important than the intramolecular ones. This means that uncertainty due to conformation changes is not a major problem.

Minimisation method influence

The same molecule subset used for the conformational analysis has been also used for studies about the influence on molecular descriptors of the choice of the geometry minimisation method.
Semiempirical methods, such as PM3, CNDO, AM1, have been used as well as molecular mechanics force field (+mm).
Differences in descriptors up to 70% occur between semi-empirical and force field optimisation procedures. Some descriptors have shown significant differences (up to 100% in one case) in the same class of methods (for instance using AM1 instead of PM3).
This is an important aspect to be considered if QSAR results from different studies have to be compared.

Solvatation influence

All the above described studies have been done considering the molecules under vacuum. However, there is the possibility to keep into account solvent effects. We evaluated the case of water. Structures solved by water molecules have been minimised using the PM3 Hamiltonian. Then descriptors have been calculated. Differences in the results occur for the same descriptors that show variance related to the conformation (i.e. electrostatic and geometric ones), but the variation percentages in the case of solvatation are inferior to those due to conformation.
This study has however to be completed, since solute-solvent interactions are not limited to structural effects: electrostatic interactions, modification of electron density and molecular orbitals etc have to be considered.

Cluster analysis

A hierarchical clustering has been performed on the dataset using the SCAN software. This exploratory analysis allows to find clusters of objects in high dimensional space based on inter-object distances. Clusters are defined by an agglomerative algorithm. The number of clusters depends on the similarity level selected. Reliable results have not been achieved, because of the diversity of the molecules and the presence of many redundant or noisy descriptors. Probably a variable selection approach is needed.
A simple classification, based on the presence of particular functional groups, such as organophosphates, carbamates, ureas, etc., has been performed in order to generate more homogeneous subgroups of molecules and to help the interpretation of modelling results.

Regression studies

The data set so obtained, with the uncertainty of the toxicities and decriptors, has been distributed to all the partners for successive modeling (Grauel et al., 1999). Furthermore, preliminary regression analysis has been performed using SCAN software. In particular the PCR algorithm has been used as it allows to analyse underdetermined data sets which have fewer observations than predictors or the predictors highly correlated. PCR is a non-least squares regression that models the response variable as a linear combination of the principal components with higher variances.
The analysis has been performed on all the 164 molecules and on the subset of organophosphorus pesticides (OP - 27 molecules). The regression model has been validated using the leave-one-out procedure.
Major results are shown in Table 1. A preliminary analysis has been done on the total set of molecules giving R2 = 0.53 and R2cv = 0.40 for trout, and R2 = 0.56 and R2cv = 0.28 for daphnia. A further analysis has been done on the subset of OP: regression model for predicting the toxicity towards rainbow trout, using all descriptors (about 150), gave poor results. Better results have been achieved for the activity towards daphnia: R2 = 0.83 and = R2cv = 0.67. Selecting variables (in order to reduce noise, redundancy and inner correlation) the regression model for trout (R2 = 0.97, R2cv = 0.87) has been better than that for daphnia (R2 =0.89, R2cv =0.83).
The variable selection method used for these models is based on a preliminary choice of some descriptors using PCA; the set of descriptors has been divided into six subclasses: geometrical, topological, electrostatic, constitutional, quantum-chemicals and logD. A PCA analysis, including the output datum, has been performed for each subclass and the descriptor nearest to the output has been selected. So, for trout, a subset of 6 descriptors has been built (R2 = 0.62, R2cv = 0.40). A further selection has been performed inserting a descriptor one at a time for each subclass and checking if it would improve the predictive power of the regression model. This procedure has given the subset of 25 descriptors for the best regression model found. Descriptors, which have shown strong dependence on conformational variations, are not present in great number in this subset.
The same procedure has been performed in order to find a good predictive regression model for the daphnia.

Another study has been performed on this subset of OP, using all not-constant WHIM descriptors, a set of 169 molecular descriptors. Although these descriptors have shown a better fitting power, they gave the same predictive capability as those from CODESSA. A preliminary study has shown that not significant improvement of the R2cv has been obtained selecting WHIM variables. This is probably due to the minor information expressed by these descriptors, compared to those expresses by CODESSA descriptors.

Table 1: *Regression coefficients for trout toxicity, using different descriptors and sets of molecules*

Regression coefficients	CODESSA all descriptors	CODESSA selected descriptors	WHIM all descriptors
164 molecules	R^2=0.53 R^2_{cv} =0.40		
27 molecules (OP)	R^2=0.38 R^2_{cv} =0.27	R^2=0.97 R^2_{cv} =0.87	R^2=0.80 R^2_{cv} =0.28

Future work

Future activites will be in all the areas involved in the project. To increase our data set on aquatic toxicity we will consider the database AQUIRE and another one done by the ECETOC.

Other toxicological endpoints will be added, such as toxicity on birds (as an indicator of toxicity for terrestrial organisms, which is one of the main fields to be targeted within the project), acute and chronic toxicity on rat, mutagenicity, reproductive and teratogenic effects and the extrapolation of chronic toxicity to humans.

Further studies will be on the use of residues, variable selections, and advanced software programs, which will be compared with regression methods.

Acknowledgments. We acknowledge the financial contribution of the European Commision (ENV4 CT97-0508) and, partially, of NATO (CRG 971505). We thank Prof. A. Katritzky, University of Florida, for giving us the opportunity to employ CODESSA and Dr. G. Macario, Medical Economix Italia, for the use of Micromedex.

References

Gini, G.; Lorenzini, M.; Vittore, A.; et al. 1999. “Some results for the prediction of carcinogenicity using hybrid systems” AAAI 1999 Spring Symposium Predictive Toxicology of Chemicals: Experiences and Impact of Artificial Intelligence Tools. March 22-24, 1999, Stanford, CA, USA

Gold, L. S.; Sawyer, C. B.; Magaw, et al. 1984, A Carcinogenicity Potency Database of the Standardized Results of Animal Bioassays, *Environmental Health Perspectives*, 58:9-319.

Grauel, A.; Ludwig, L. A.; and Berk, F. ” Computational Intelligence and Predictive Toxicology”, AAAI 1999 Spring Symposium Predictive Toxicology of Chemicals: Experiences and Impact of Artificial Intelligence Tools. March 22-24, 1999, Stanford, CA, USA

van den Heuvel, M. J.; Clark, D. G.; Fielder, R. J.; et al. 1990. The International Validation of a Fixed-dose Procedure as an Alternative to the Classical LD50 Test. *Food Chem. Toxicology*, 28:469-482.

Annex 1.

COMET: scientific partners

Dr. Emilio Benfenati, Coordinator, Istituto "Mario Negri" Milan, Italy

Prof. Adolf Grauel, University of Paderborn, Soest, Germany

Prof. Ramon Carbó-Dorca, University of Girona, Spain

Prof. Giuseppina C. Gini, Politecnico di Milano, Italy

Dr. Paola Ciocca: Società Italiana per i Prodotti Chimici e per l’Agricoltura (SIPCAM), Pero, Italy