COMPUTATIONAL MODELS FOR TOXICITY OF ALLELOCHEMICALS

Lo Piparo E.[1], Fratev F.[1], Benfenati E.[1], Lemke F.[2,3], Gini G.[2]

[1] *Istituto di Ricerche Farmacologiche "Mario Negri" Milano, Italy*
[2] *Dipartimento di Elettronica e Informazione, Politecnico of Milano, Italy*
[3] *Knowledge Miner Software Berlin, Germany*

INTRODUCTION

The test on animals are the principal methods used today to identify toxic chemicals and to limit the incidence of diseases but, due to the cost and the time needed for testing, the number of chemicals introduced in our life is higher than the number of experimental studies done. Moreover, ethical consideration and public pressure push to reduce tests on animals [1]. Thus, more convenient and efficient methods to predict biological activity from structural information is demanded. The use of validation models is encouraged by the European Union and USA to cover the many knowledge gaps [2]. The poor knowledge of chemical toxicity is much worse for natural compounds because the regulation forces the chemical industries to produce experimental data on synthetic chemicals, which will be put into the environment.

QSARs (Quantitative Structure-Activity Relationships) are based on the interaction of biology, chemistry, and statistics to build-up mathematical models for the prediction of property such as toxicity. The basic assumption of QSAR is that there is a quantitative relation between the molecular structure of compounds and their biological, chemical and physical properties.

The first step in formulating QSARs is to build-up a data set that must reflect one or more well-defined toxic endpoints. Subsequently, geometric optimization that finds energy minimum state of the structure must be performed, because many physical and chemical properties of molecules can be calculated only if the structures have been optimized before. So it is possible to calculate the descriptors that mathematically characterize the molecules and to use them to build-up models. Finally, QSAR models should be validated to ensure that they are capable of making sufficiently accurate predictions of toxicity for new compounds not used yet to generate the models.

METHOD

We built up a data set by selecting synthetic pesticides with structures similar to that of allelochemicals. In particular, all compounds contain an aromatic and/or heterocycle ring with N and O. We collected for them toxicity values for different species (rat, daphnia) and different sources: The Pesticide Manual [3], the U.S. EPA Ecotox database [4], and the CIRCA website [5].

Molecular optimization was done using semi-empirical methods. Two different approaches to generate data of molecular descriptors were applied:

- We calculated physicochemical descriptors such as logP and, using Codessa software [6], descriptors such as constitutional (number and type of atoms, bonds and functional groups), geometrical (molecular surface area and volume, moment of inertia, shadow area, projections and gravitational indices), topological (molecular connectivity indices, related to the degree of branching in the compounds), electrostatic (partial atomic charges and others depending on the possibility to form hydrogen bonds).
- A new approach is computing Comparative Molecular Field Analysis (CoMFA) descriptors. CoMFA [7] measures the steric and electrostatic interaction energy between an atomic probe (carbon atom, positively or negatively charged atoms, lipophilic probes …) and each molecule at points of a grid surrounding it. These fields are used as point descriptors of the 3D molecular structure and physic-chemical behavior of the molecule. Moreover, a graphical analysis allows

a simple interaction of the field and the visualization of the regions where the probe interacts most strongly with the target either by attraction or repulsion.

We also used GRID and Docking methods to compare and increase the information obtained from CoMFA descriptors. Docking [8] is a computational methods used to find the best matching between a biological macromolecule and a ligand, while Grid [9] can detect the interaction between target protein and different probes such as alkyl hydroxyl, phenolic hydroxyl, ether oxygen, ketone oxygen, carboxy and phosphate groups. When the crystallographic structure of the target protein is not available, homology model (HM) is constructed using the amino-acid sequence of the receptor and searching for a template for modeling it. An appropriate template is a protein of known structure that shares a minimum of 20-25% homology in amino-acid sequence. Then a model of the receptor is built by threading it onto the template and optimizing it.

Codessa and logP descriptors were used for Group Method of Data Handling Neural Network modeling, while on CoMFA descriptors we applied PLS:

- Creating reliable models from a limited set of molecular descriptors (Codessa and logP descriptors in our case) is an ill-posed task characterized by a high-dimensional descriptor space, noisy and uncertain toxicity. One most promising data-driven modeling technology is based on Statistical Learning Networks. Adding a key principle – that of induction – to these Learning Algorithms defines the Group Method of Data Handling (GMDH) Neural Networks [10]. They are able to self-organize an optimal complex model composed of a set of self-selected relevant descriptor variables starting from a completely unknown and not predefined model structure. In this way it is possible to systematically self-organize not overfitted linear or non-linear models, which are also available in analytical form, making this knowledge extraction technology well suited for QSAR modeling and model combining.
- Alternatively, whit CoMFA descriptors we used Partial Least Squares (PLS). The general idea of PLS is to try to extract the most relevant and often hidden information to build-up a robust model.

RESULTS

Rat toxicity prediction model:

- Different methods provides different perspectives, so Docking, GRID and CoMFA were integrated into a new approach for toxicity prediction (LD50) for rat. The comparison of these methods can provide important information for understanding toxicity and for describing the mechanism of action. The method was tested on a set of 73 chemicals. CYP1A2 has an important role in the rat metabolism of the examined chemicals and its homology modeling was done using crystallographic coordinates of rabbit CYP2C5. We found the significant residues and positions of oxidation in the binding site by docking analysis. GRID examination showed that the docked ligand position overlaps the favored GRID binding areas. The overall approach allowed deriving a model with high predictive ability ($R^2 = 0.92$).
- A linear, a non-linear, and a combined model was created for the same compounds, using Codessa and logP descriptors and the GMDH implementation of the KnowledgeMiner software. The linear model is composed of 6 descriptors and shows a $R^2 = 0.68$, while the non-linear model selected 4 relevant descriptors and has an accuracy of $R^2 = 0.77$. A combined model built on the two individual models increases accuracy to $R^2 = 0.79$.

Daphnia toxicity prediction model:

Again the GMDH Neural Network was used to generate a linear QSAR model on a data set of 86 compounds and 164 Codessa descriptors and logP. The reported overall model performance on training and test data is $R^2 = 0.70$.

CONCLUSION

We developed independent models to predict toxicity, using different approaches. 3D models, using CoMFA, GRID and Docking approaches, gave good results, which also indicate important likely molecular toxic metabolism processes. The use of chemical descriptors and GMDH vice versa is more related on automatic extraction of knowledge, independent on mechanism. Even if it gives lower performances, it can be more general, and applicable to more heterogeneous structures.

ACKNOWLEDGMENTS

The research described in this abstract was performed as part of the project "FATEALLCHEM", "Fate and Toxicity of Allelochemicals (natural plant toxins) in Relation to Environment and Consumer". The project was carried out with financial support from the Commission of the European Communities under the Work program Quality of Life, contract no. QLK5-CT-2001-01967. We acknowledge Prof. Katritzky and Prof. Karelson for the use of Codessa and Dr. Irini Doytchinova for Chem-X software.

REFERENCES

[1] Omen, G.S. (1995) Assessing the Risk Assessment Paradigm. *Toxicology 102*, 23-28.

[2] John D. Walker, QSARs for Toxicity Screening: Current Practices, chapter 5 QSARs for Pollution Prevention, Toxicity Screening, Risk Assessment, and Web Applications (2003), Edited by D. Walker, Published by SETAC.

[3] The Pesticide Manual, Eleventh Edition, Edited by C D S Tomlin 1997.

[4] http://www.epa.gov/ecotox/

[5] http://forum.europa.eu.int

[6] Katritzky, A. R., Karelsol, M., Lobanov, V. S., CODESSA: COmprehensive DEscriptors for Structural and Statistical Analysis, version 2.2.1. *Reference Manual*. University of Florida, Gainesville, Florida, U.S.A.

[7] Cramer III, R. D., Patterson, D. E. and Bunce, J. D. (1988) Effect of shape on binding of steroids to carrier proteins. *J. Am. Chem. Soc. 110*, 5959-5967.

[8] Morris, G. M., Goodsell, D. S., Halliday, R.S., Huey, R., Hart, W. E., Belew, R. K., and Olson, A. J. (1998) Automated Docking Using a Lamarckian Genetic Algorithm and Empirical Binding Free Energy Function. *J. Comput. Chem. 19*, 1639-1662.

[9] GRID, Goodford, P.J. Molecular Discovery Ltd, University of Oxford, England, SGI.

[10] Mueller, J.-A., Lemke, F.: Self-Organising Data Mining. Extracting Knowledge From Data, Hamburg BoD, 2000.