From: AAAI Technical Report SS-99-01. Compilation copyright © 1999, AAAI (www.aaai.org). All rights reserved.

Overview of Different Artificial Intelligence Approaches Combined with a Deductive Logic-based Expert System for Predicting Chemical Toxicity

Ferenc Darvas[I], Ákos Papp[I], Alex Allerdyce[I], Emilio Benfenati[II], Giuseppina Gini[III], Miloň Tichý[IV], Nicholas Sobb[V] and Aida Citti[V]

I. ComGenex Inc., POB. 667/9., 1399 Budapest, Hungary (df@comgenex.hu, akospapp@comgenex.hu)
II. Laboratorio di Farmacologia e Tossicologia Ambientali, Istituto di Ricerche Farmacologiche "Mario Negri"
Via Eritrea 62, 20157 Milano, Italy (benfenati@irfmn.mnegri.it)
III. Department of Electronics, Politecnico di Milano, Piazza Leonardo da Vinci, 32, 20133 Milano, Italy (gini@elet.polimi.it)
IV. Predictive Toxicology Laboratory, Toxicology Analysis Group, National Institute of Public Health,
Strobarova 48, 100 42 Praha 10, Czech Republic
V. CompuDrug International Inc., 705 Grandview Drive, South San Francisco, CA 94080, USA
(nsobb@compudrug.com, aidacitti@compudrug.com)

Abstract

Using the knowledge base collected by the US Environmental Protection Agency, an expert system family (HazardExpert) has been developed in 1987. The paper focuses on the different artificial intelligent approaches which had been applied by the system during its 12 years experience, notably:

a.) the deductive logic of HazardExpert for predicting toxicity

b.) reasoning by analogy for improving the context-dependency of the metabolism engine of HazardExpert

c.) using neural network in combination of HazardExpert

The presentation compares the performance of the different released versions used at approximately 100 industrial, academic and governmental institutions in 15 countries.

HazardExpert — Overview

Using the knowledge base collected by the US Environmental Protection Agency, an expert system family (HazardExpert) has been developed in 1987. HazardExpert predicts the toxicity of a compound in seven toxicity classes, such as oncogenicity, mutagenicity, teratogenicity, irritation, sensitivity, immunotoxicity and neurotoxicity by identifying toxic fragments in the molecule and assigning expected toxicity based on the detected fragments. For predicting the toxic effect of the metabolites, the software generates their structures, then searches for the toxic fragments, and summarizes the results. For the prediction, the MetabolExpert engine is used. Besides the toxic effects, HazardExpert predicts physico-chemical and bioavailability data, plus the bioaccumulation of the compound. The overall predicted toxicity is taken to be equal to the highest relative toxicity among the toxic effects. Additionally, HazardExpert calculates the bioavailability from predicted pK_a and log$P$ values, furthermore bioaccumulation from the user-set values of dosage and duration. There are eight biosystems that can be handled by HazardExpert: mammals, plants, fishes, birds, microbes, algae, aquatic and soil invertebrates. Additional conditions, like duration of the administration and the dosage can also be set for the calculations.

HazardExpert — The Model

The toxicity of a molecule is highly dependent upon its structural elements. Certain molecule fragments are characteristic of hazardous compounds and therefore called toxic fragments. The toxicity prediction is based on rules comprising the chemical substructure of the molecule which is the essential part of the rule together with a list of substructures representing the required environment of the fragment (and are necessary for its clear-cut description), as well as a list of substructures which may not be present in the neighborhood of the fragment.

HazardExpert — The Concept

HazardExpert originally was the name of a research project, initiated by CompuDrug in Hungary in 1986, with the aim to model xenobiotics (foreign substances) in a living system or in the environment by expert system approach, as its earliest publication is in 1987 [i]. The project, which was finalized in 1987 with help of US EPA, resulted in a first-order logic based model of chemical toxicity in a compartmentalized system, like humans, plants or ecosystem, and in a series of computer programs which have been commercialized mostly under the same name, HazardExpert. A concise description of the underlying model, originally developed for a metabolic transformations, is given in [ii].

The original form of the HazardExpert model was related to logic programming, based on Kowalski's classic idea of using mathematical logic directly as a programming tool [iii]. Till 1991 Prolog [iv], an artificial intelligence language was used for writing the software, after a decade of successful application of the same language for developing chemical expert systems for calculating logP values [v], predicting carcinogenic activity [vi], or automatic interpretation of QSAR equations [vii] or predicting metabolites [viii]. The core of the model has been the "Biotransform graph", a graph modeling the composition of living system together with transport processes and metabolic pathways.

Transformations were formulated in the model and later in the software programs as "if... then" rules. An example for a (simplified) rule related to conditional toxic symptoms:

A secondary metabolite of a carboxylic acid is a hyppurate,
if
the carboxylic group is connected to an aromatic ring
and
the neighboring atom on the aromatic ring does not contain any substituent,

The current HazardExpert software versions use a simplified version of the biotransform graph, which contains only a single compartment. As a result, toxic symptoms corresponds to a vertex of a single graph, where otherwise only the metabolic transformations are displayed, like in the case of the metabolic transformation of Eugenol:

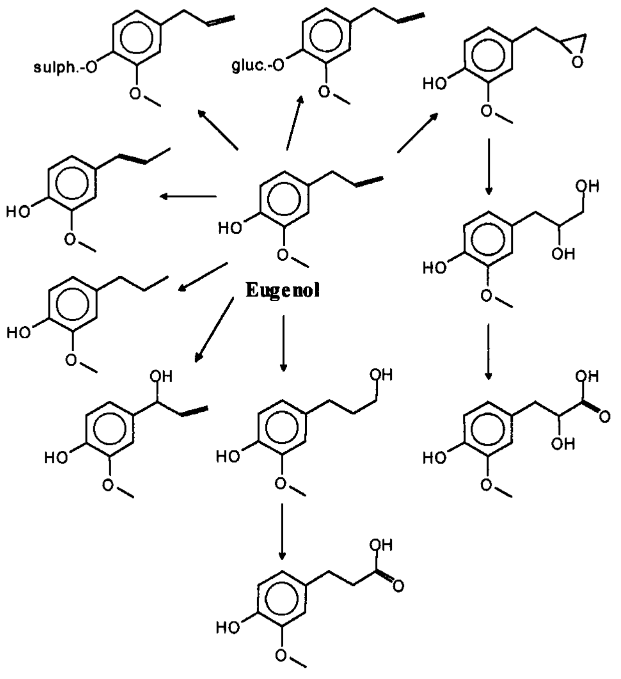


Figure 1

Metabolism of Eugenol in Human (Oral, 150 mg[ix])

While the underlying model of HazardExpert by this way is somewhat more complex than similar, subsequently developed systems, like DEREK (Deductive Estimation of Risk from Existing Knowledge), the enhanced complexity pays through enabling to develop a complex integrated system encompassing both the metabolic and toxicodynamic aspects.

Technical Background

All members of the HazardExpert family are expert systems with a common architecture. They are composed of one or several databases, Knowledge Base (KB), in the followings, and one or more prediction engine, which is producing the prediction results. The input of the expert systems are the structural formula(s)of the compound(s) to be predicted concerning the structure of their metabolites, their toxicity values, or their retrometabolites. The result of the calculation is displayed in a graphical tree structure, including the structure of the metabolites, or the structure of the retrometabolites, or, in some cases, a list of the structure of the metabolites together with the expected toxicity values or retention time.

The Rule System of HazardExpert

The KB's are composed of "if...then" rules, like the "hyppurate formation rule" given above. The "if" part of the rule is composed of a series of substructures, separated by one or more "And" or "Or" type logical connectors.

Every transformation rule is composed of four elements (see also the example on Figure 2):

1. The substructure changed during the transformation *(active substructure).*
2. A list of substructures at least one of which must be present in the molecule for the transformation to occur *(positive conditions).*
3. A list of substructures whose presence prevents the transformation from occurring *(negative conditions).*

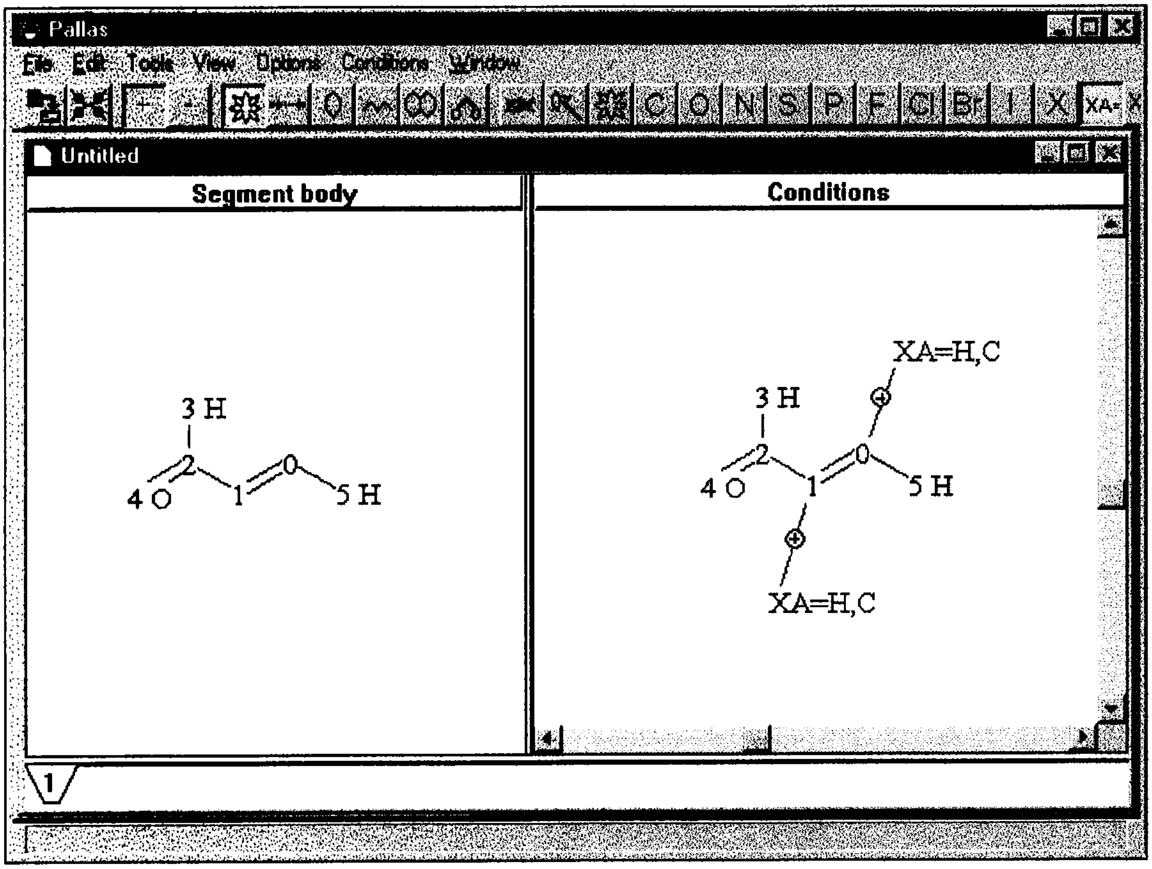


Figure 2

The Metabolism Prediction Rule System of HazardExpert

The Knowledge Base is composed of the rules, completed or not with further metabolic data. The system of *Transformation DB* is essentially restricted to the transformation rules, while the collection of the *Learned Tree* includes a series of other data related to the metabolic fate of a particular compound, like the excretion pathway, transformation or excretion percentage of the metabolites, the analytical methods used to identify the metabolites, e.g.

The difference between a Learned Tree and a tree corresponding to a specific DB is demonstrated by Figure 3, where the Learned Tree of Eugenol is presented (vs. the tree depicted in Figure 1).

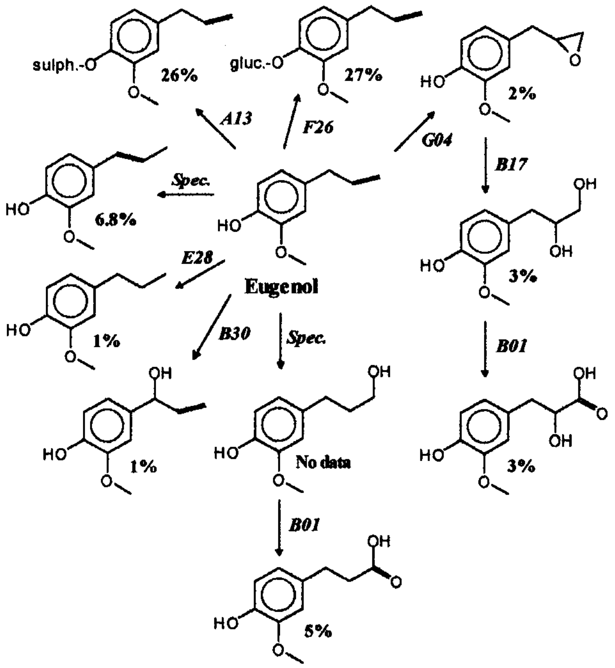


Figure 3

The Metabolism Prediction Engine. Using Reasoning by Analogy

In the MEX family, there are two kinds of prediction. One is made by the Basic transformation DB, and results a preliminary metabolic simulation called 'Frameless Generation'. The other is the 'Generation by Analogy', which is a species specific, quantitative metabolic prediction.

During *Frameless Generation*, the metabolism prediction engine tries to match the Basic transformations to the compound structure. By default, the resulting metabolites will be produced by the matched transformations automatically. Depending on the selected number of the metabolism levels, the program applies the Basic transformations to the metabolites, and stops only after the last level.

Generation by Analogy is an extended predictive tool, which is based on finding analogues in the DB of Learned Trees, and uses the metabolic transformations of the compound having the most similar metabolic fate in a selected species (human, rodent). It starts with an automatic Frameless Generation in one level, then produces a list of matched transformations. This list will be compared with first level of the metabolic trees in the DB of Learned Trees in the selected species, then the analogues will be listed in the order of similarity. The user can select from the list of similar compounds manually, or can choose automatic prediction in which the most similar compound will be selected. In the next step the metabolism prediction engine uses the transformation set of the learned tree of the selected analogue (including the specific transformations), so the resulted metabolic tree will be species specific, and the program calculates the excretion percentages of the metabolites using the conversion data in the learned tree. Finally the metabolites are ordered according to their relative importance, and prioritization is also expressed (the metabolites are considered as prevalent, dominant, important, unimportant or negligible derivatives).

Combining HazardExpert with Neural Network

Under the frame of an international project (COPERNICUS) we tried to combine the predictive ability of HazardExpert with that of an artificial neural network. For that purpose we had to synchronize the input/output structure of the different parts of the hybrid system. First of all, we had to decide what kind of toxic effect will be predicted by the combined system. Since the prediction of carcinogenicity is very important, we selected to deal with oncogenicity, as the most important toxicity category in the aspect of carcinogenicity.

In the hybrid system, the predicted oncogenicity of a compound is used as an analogue input of the Neural Network that is responsible for the final calculations. Together with the log*D* values at pH=2, 7.4 and 10 (which are also predicted by a rule based system, PrologD), plus special molecular descriptors (selected for represent the compound structure for the neural network) it can initialize responds on the output of the Neural Network, and result the predicted LD_{50} values.

The combined system was developed to be able to predict the toxicity of compounds in environmental situations. The compound can express its toxic effects through its metabolites (degradation products). To take into account the effect of the metabolites, another rule based system for prediction of photodegradation has been introduced into the hybrid system. The generation of the structures of the degradation products is made by a special transformation rule database of MetabolExpert. The generated structures then carried back to the input of HazardExpert to complete the prediction of the overall toxicity of the parent compound, and this value will be transferred to the input of the neural network.

The structure of the hybrid system can be seen on Figure 4.

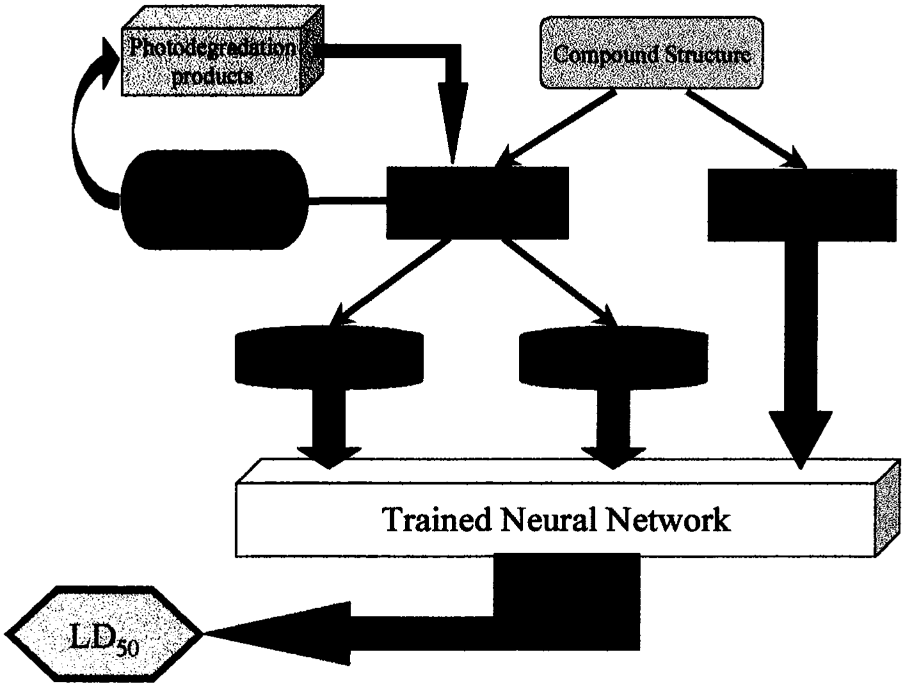


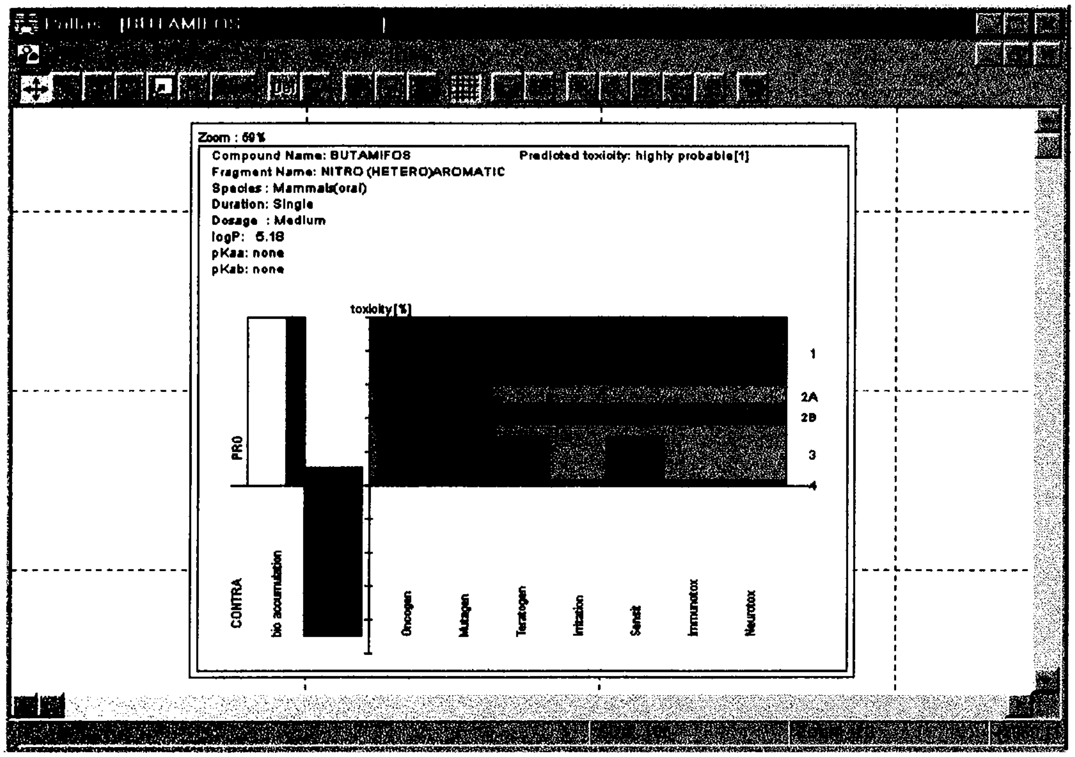
Figure 4

Prediction of Toxicity. An Example

An example for toxicity prediction (Butamifos) can be seen in Figure 5. The system predicted highly probable overall toxicity, since both the oncogenic and the mutagenic effects are over the 60% limit.

Acknowledgments

Here we would like to thank for the research supports received between 1994 and 1997 from The Commission of the European Communities under the frame of a COPERNICUS project, named "Development of Second Generation Expert Systems for Environmental Toxicology". We also thank to Prof. Jure Zupan and his colleges (at the National Institute of Chemistry, Ljubljana, Slovenia) for the development of the neural network part of the combined system in the above mentioned project.



Zoom : 69%

Compound Name: BUTAMIFOS

Predicted toxicity: highly probable[1]

Fragment Name: NITRO (HETERO)AROMATIC

Species : Mammals(oral)

Duration: Single

Dosage : Medium

logP: 5.18

pKaa: none

pKab: none

Figure 5

References

i F. Darvas, *J. Mol. Chem.*, **6**, 80-85 (1988)

ii F. Darvas, in: *QSAR in Environmental Toxicology - II,* Ed. K.L.E. Kaiser, D. Riedel Publishing Company, **1987**, 71-81

iii R. Kowalski, *Logic for Problem Solving.* North-Holland, New York, **1979**.

iv W.F. Clocksin, C.S. Mellish, *Programming in Prolog,* Springer Verlag, Berlin, **1981**.

v F. Darvas, I. Erdõs, and Gy. Teglás, in *QSAR in Drug Design and Toxicology,* Eds. D. Hadzi and B. Jerman-Blazic, Elsevier, Amsterdam, **1987**, 70

vi F. Darvas, F. Futó, E. Cholnoky, in *Proc. First Int'l Conf. The J. Neumann Society,* Ed. Gy. Kovács, J. Neumann Society, Budapest, **1979**, 216

vii F. Darvas, I. Futó, P. Szeredi, in *Proc. Symp. Chem. Struct-Biol. Act., Quant. Approaches,* Ed. R. Franke, Akademie Verlag, Berlin, **1978**, 251

viii F. Darvas, J. Mol. Graph., **6**, 80, (1988)

ix I.U. Fischer, G.E. von Unruh, and H.J. Dengler, *Xenobiotica,* **20**, 209, (1990).