THE APPLICATION OF FEL-EXPERT SYSTEM IN THE INTERPRETATION OF BORON COMPOUNDS TOXICITY

Oldřich ŠTROUF^a and Vladimír MAŘík^b

^a Institute of Inorganic Chemistry,

Czechoslovak Academy of Sciences. 160 00 Prague 6 and ^b Faculty of Electrical Engineering,

Czech Technical University. 160 00 Prague 6

Received May 4, 1989 Accepted July 7, 1989

Dedicated to Professor Otto Exner on the occasion of his 65th birthday.

The influence of substructural features on boron compounds toxicity (LD₅₀, mice, i.p.) has been studied by FEL-EXPERT system developed at the Czech Technical University of Prague. A set of 108 compounds containing one or two boron atoms in their molecule has been arbitrarily divided into three classes: the compounds with a high toxicity (LD₅₀ < 100 mg/kg), with a medium toxicity (100 mg/kg \leq LD₅₀ < 1 000 mg/kg) and with a low toxicity (LD₅₀ \geq 1 000 mg//kg). The compounds have been represented by 70 substructural fragments, 27 of them being "central substructures" containing boron atom(s). The inference net consists of 118 nodes (74 of the Bayesian type), 362 production rules and 74 context links. The total classification correctness has been 98%. As a case-study, the classification of *p*-tolylboronic acid (LD₅₀ = 520 mg/kg) and 4-carboxyphenylboronic acid (LD₅₀ = 3 838 mg/kg) has been discussed.

Boron compounds possess interesting biological properties, some of them being useful in medicine¹. Recently, a great deal of interest has been focused on boron compounds for neutron capture therapy of cancer²⁻⁴. This therapy is based on cumulation of compounds containing boron-10 in cancerous tissues followed by thermal neutron irradiation. The capture of a such low-energy (0.025 eV) neutron by ¹⁰B results in a nuclear reaction producing high-energy (2.4 MeV) particles (α and ⁷Li). The limited range of the particles of about 10 µm provides an exciting possibility to develop an extremely specific treatment of tumours on the molecular level. A condition for reaching this goal is the availability of specific stable compounds with a sufficiently high solubility in water and with a high content of ¹⁰B in their molecule. Last but not least, the compounds should not be toxic because they are applied in a robust dose, the actual concentration to be 20–50 µg of ¹⁰B in one gram of cancerous tissue during the irradiation.

This paper deals with trials to evaluate the toxicity from structure of compounds containing one or two boron atoms in their molecule by means of the expert system FEL-EXPERT (version 1.5) (refs^{5,6}). The goal is to verify the possibility of the use of this expert system for this purpose.

THEORETICAL

FEL-EXPERT SYSTEM

FEL-EXPERT, an empty rule-based expert system (shell) developed at the Czech Technical University of Prague, has the following characteristics^{5,6}:

a) Domain independence. The application area can be changed by replacing the knowledge base, with no modification required to the empty FEL-EXPERT system.

b) Machine independence. FEL-EXPERT is written in the standard Pascal programming language. It is conservative of memory, permitting its use on a personal computer.

c) Diagnostic character. FEL-EXPERT is a suitable tool for solving diagnostic tasks. A finite set of goal hypotheses is considered, and evaluated and re-evaluated during the consultation run.

d) Ability to handle uncertainty. Uncertainty of both knowledge and data is considered and accepted. A built-in model for handling uncertainty is based on ideas previously used by PROSPECTOR (ref.⁷).

e) Explanation capabilities. A wide spectrum of explaining abilities (including the answering of "What?" and "Why?" questions) makes possible very detailed, user friendly explanations of the decision making process and also of the actual model state of the case in hand.

Knowledge Representation

Basically, three types of knowledge representations in the knowledge base are used: production rules, logical functions and context links.

A) The production rules have the following form:

If {evidence E} THEN {hypothesis H} WITH {probability P_1 } ELSE {hypothesis H} WITH {probability P_2 }

where {evidence E} and {hypothesis H} are propositions, {probability P_1 }, {probability P_2 } are subjective uncertainty weights (not probabilities in an exact mathematical sense), called, as in ref.⁷, sufficiency and necessity weights, respectively. They can be expressed, in terms of PROSPECTOR, as subjective conditional probabilities P(H|E) and $P(H|\overline{E})$, respectively. Their values are given by the expert. The model for uncertainty handling requires assigning a prior probability to each proposition. This value is also given by the expert.

B) The logical function makes the expression of composed propositions possible. Three types of logical functions are considered in the FEL-EXPERT system: AND, OR, NOT. C) The context links are used in cases where, before one proposition can be investigated, the complete checking of the other one is necessary, see Fig. 1; the context link is expressed by a dashed line. The proposition E_3 is a context of E_1 , that means that before investigating E_1 the proposition E_3 has to be proved in a sufficient way (the context must be satisfied). In the opposite case E_1 is excluded from the investigation. To each context link two numbers, α_1 and α_2 , are assigned by the expert; they represent the range within which the probability (validity of the context) has to lie.

Both the syntax of the knowledge representation language and the control strategy are described in ref.⁸.

Actual Model

The actual model is composed of a set of all posterior node (proposition) probabilities. Starting the consultation it consists of the set of the prior probabilities. With any user's particular information the node probabilities change to posterior ones and thus the actual model becomes – step by step – tailored for the case in hand.

Uncertainty Processing

To provide a more detailed explanation of the knowledge contained in the inference net, a short description both the interpretation of the inference net parameters and the way of the information propagation in the inference net is given. The fragment of an inference net is considered as shown in Fig. 2. To each node a priori probability P(.) is assigned by the expert; two parameters $P_1 = P(H/E_i)$ and $P_2 = P(H/\overline{E}_i)$, respectively, are assigned to each oriented arc (\overline{E}_i represents the negation of E_i). The parameters P_1 and P_2 , subjectively given by the expert, can be interpreted like



FIG. 1FIG. 2Production rules and context links in theFragment of an inference net of the FEL-FEL-EXPERTEXPERT

conditional probabilities. $P(H/E_i)$ is the probability of conclusion H the evidence E_i being true, $P(H/\overline{E_i})$ is the probability of H the evidence E_i being categorically false.

By making use of the "classical" Bayesian formulas and by introducing the odds by the formula

$$O(\bullet) = \frac{P(\bullet)}{1 + P(\bullet)} \tag{1}$$

it is possible to write

$$O(H/E_{i}) = \frac{P(H/E_{i})}{P(\overline{H}/E_{i})} = \frac{P(E_{i}/H)}{P(E_{i}/\overline{H})} \frac{P(H)}{P(\overline{H})} = L_{i} \cdot O(H)$$
(2)

and

$$O(H/\bar{E}_{i}) = \bar{L}_{i} \cdot O(H) .$$
⁽³⁾

Analogically, as in the classical Bayesian theory, the measures L_i and \bar{L}_i are called sufficiency and necessity measures, respectively. Let's notice that there is the unique relation between L_i and $P_{1i} = P(H/E_i)$ and between \bar{L}_i and $P_{2i} = P(H/\bar{E}_i)$, respectively:

$$L_{i} = \frac{P(H/E_{i})}{1 - P(H/E_{i})} \cdot \frac{1 - P(H)}{P(H)},$$
(4)

$$\bar{L}_{i} = \frac{P(H/\bar{E}_{i})}{1 - P(H/\bar{E}_{i})} \cdot \frac{1 - P(H)}{P(H)}.$$
(5)

That's why it is possible to express the strength of a production rule in the form of the weights (probabilities) $P(H/E_i)$, $P(H/\overline{E}_i)$ or, adequatly, in the form of the measures L_i , \overline{L}_i . The advantage of L_i , \overline{L}_i lies in the fact that they express the strength of a rule in absolute values (the strength of $P(H/E_i)$, $P(H/\overline{E}_i)$ is relative to P(H)).

If the hypothesis H is supported by several evidencies $E_1, E_2, ..., E_n$ (there are more rules containing the same hypothesis on the right-hand side) and the independency of evidencies $E_1, E_2, ..., E_n$ is considered, it holds

$$O(H/E_1, E_2, ..., E_n) = L_1 \cdot L_2 \cdot ... L_n \cdot O(H), \qquad (6)$$

$$O(H/\bar{E}_1, \bar{E}_2, ..., \bar{E}_n) = \bar{L}_1 \cdot \bar{L}_2 \cdot ... \bar{L}_n \cdot O(H), \qquad (7)$$

respectively. Making use of formula (1), the probabilities $P(H|E_1, E_2, ..., E_n)$ and $P(H|\overline{E}_1, \overline{E}_2, ..., \overline{E}_n)$ may be computed.

Until now we have considered the user's answer to be categorically yes (E) or no (\overline{E}) . But the expert system has the possibility to accept the uncertain user's informa-

tion, too. In such a case the user provides the probability P(E/E'), where E' is the observation relevant to E. In the extreme situations: The user is sure in E being true (then P(E/E') = 1) or the user is sure in E being false (P(E/E') = 0). If the user really does not know the answer, he cannot provide a new piece of information. Then the initial probability of the proposition remains P(E/E') = P(E) and no information propagation is to be executed.

To remove the probability inconsistency in the model for uncertainty handling, the following interpolation formula for recomputing the probabilities is used:

$$P(H/E'_{i}) = P(H/\bar{E}_{i}) \cdot P(\bar{E}_{i}/E'_{i}) + P(H/E_{i}) \cdot P(E_{i}/E'_{i}) =$$

= $P(H/\bar{E}_{i}) + [P(H/E_{i}) - P(H/\bar{E}_{i})] \cdot P(E_{i}/E'_{i}) .$ (8)

The probabilities $P(H/E_i)$, $P(H/\overline{E}_i)$ are subjective sufficiency and necessity weights P_{1i} , P_{2i} ; $P(E_i/E'_i)$ expresses the uncertainty in the user's answer.

To each $P(H|E'_i)$ corresponds the (so called) effective measure L'_i

$$L'_{\rm i} = \frac{O(H/E'_{\rm i})}{O(H)}.$$
 (9)

To combine the influence of several rules on the same hypothesis H, a modified formula (6) is used

$$O(H/E'_{1}, E'_{2}, ..., E'_{n}) = L'_{1} \cdot L'_{2} \dots L'_{n} \cdot O(H) .$$
⁽¹⁰⁾

From this theoretical review follows that the graphical expression of the inference net displays all the influences among all the propositions as well as the strengths of these links in absolute measures L and \overline{L} . The measure L is within the interval $\langle 1; \infty \rangle$; the higher this value the higher is the positive influence of E on H. Analogically, the value of $\overline{L} \in \langle 0; 1 \rangle$ expresses the negative influence. In the case of L = 1 (or $\overline{L} = 1$) there is no positive (or negative) influence, respectively.

KNOWLEDGE BASE

Boron Compounds, Their Structure and Toxicity

In this introductory study only compounds with one or two boron atoms in the molecule are considered. They are represented by their structures, because chemical structure can be generally accounted for as an extremely integrated and comprised information about the compound under study.

The estimation of toxicity represents a time and money consuming experiment. The interpretation of the toxicity in terms of structural information does not represent a trivial task. Also in the case of the boron compounds, the measure of influence of structural feature on the final toxicity is unknown. Moreover, the influence does not behave in an additive way. Therefore, the exact rules for the correlation have not been available in the case under study.

These facts lead to the conclusion that the expert system approach is the most suitable way to solve this toxicity estimation problem. Theoretically, there are other methods to solve such diagnostic tasks, namely within the statistical-decision approach or syntactical pattern recognition. For the statistical decision-making there were not enough examples at disposal. The syntactical pattern recognition theory provides only very strict tools without any possibility to express the uncertain knowledge (heuristics). On the other hand, the diagnostic expert system approach has satisfied all demands on uncertainty processing, flexibility and modularity.

Formulation of the Task

The task which has been chosen can be expressed in the following way: To develop a knowledge base enabling to estimate the toxicity of boron compounds by making use of the structure description. As a result, two knowledge bases BASE1 and BASE2 have been developed. For the construction of these bases 48 and 108 compounds have been considered, respectively.

In the task formulated above, the structural information on the compound has formed the data, the estimation of the toxicity level has been the goal. Each chemical structure has been represented by a set of appropriate substructures. As a central substructure the boron atom and its surrounding atoms is first defined (examples see Table I). The remaining part of the structure has been coded by such substructures which preserve the structural information. The central substructural descriptors have formed the content of the leaf nodes of the inference net. After a structural analysis of the 108 boron compounds, the total number of 70 substructural descriptors (27 of them being of the central type) have been established and used in a verbal form in the BASE2:

The toxicity of the studied boron compounds has been expressed in lethal doses LD_{50} estimated on mice for the intraperitoneal application. The values have been extracted from Kliegel's monograph¹. Two classes of toxicity ($LD_{50} < 500 \text{ mg/kg}$ and $LD_{50} \ge 500 \text{ mg/kg}$) have been chosen to form the goal hypotheses in the BASE1, three classes ($LD_{50} < 100 \text{ mg/kg}$, $100 \text{ mg/kg} \le LD_{50} < 1000 \text{ mg/kg}$ and $LD_{50} \ge 1000 \text{ mg/kg}$, $100 \text{ mg/kg} \le LD_{50} < 1000 \text{ mg/kg}$ and $LD_{50} \ge 1000 \text{ mg/kg}$) have been defined in the BASE2. The classes have been defined in a formal way and not with respect to a pharmacological categorization. From practical point of view, only the slightly and medium toxic compounds belonging to the classes H_3 and H_2 can be considered as candidates for neutron capture therapy because the total therapeutic dose is applied in grams of the compound.

Boron	Compounds	Toxicity
-------	-----------	----------

•

TABLE I

Examples of central substructural descriptors

No. of]	Representation of descriptors ^a
descriptor	Graphical	Verbal
1	$\begin{bmatrix} \mathbf{F} \\ \mathbf{F} \\ \mathbf{F} \\ \mathbf{B} \\ \mathbf{F} \end{bmatrix}^{-1}$	tetrafluoroborate anion
2	0_0 B 0	on boron are three O
3	C 	on boron are two H, one C and one N (not in a cycle)
4	C HBH N	on boron are two H, one C and one N (in a cycle)
5	OO BC	on boron are two O and one C (in a cycle)
6	H 	on boron are three H and one N (not in a cycle)
7	H HBH	on boron are three H and one N (in a cycle)
8	`C´ CC	on boron in a cycle -B-N-C-C-N- is one C (in a cycle)
9	0 (-) 0 C 0 C C C	on boron in a cycle -B-O-C-C-C-C- are two O

.

(Continued) Representation of descriptors^a No. of descriptor Graphical Verbal I 10 on boron are two H, one N and one I -H Ń : : B 20 on boron is one amino group H-N-H Ĥ в on boron is one N with three C 21 -C (not in a cycle) 22 B-OH on boron is a hydroxylic group 27 В-СООН on boron is a carboxylic group

^a For technical reasons, the consultations in this study were performed using the verbal representation of the substructures.

The heuristic rules expressing the dependency (relation) among the structural descriptors and the classes of toxicity have been constructed in the following way: The descriptors have been considered as features and linear classifiers⁹ with binary inputs have been constructed. The final weights of these classifiers have been accounted for as $P(H/E_i)$ and $P(H/\bar{E}_i)$ and used to compute L, \bar{L} by means of the formulae (4) and (5), respectively. The illustrative examples of the heuristic rules (including the weights L, \bar{L} and their interpretations) are given in the case-study section in detail.

RESULTS

The knowledge bases, BASE1 and BASE2, have been developed. Only the BASE2 is discussed here because the tutorial solutions by means of reduced BASE1 are practically the same¹⁰.

Collect. Czech. Chem. Commun. (Vol. 55) (1990)

TABLE I

The inference net consists of 118 nodes (74 of them being of Bayesian type), 362 production rules and 74 context links. Moreover, 70 nodes are askable ones; they contain the verbal description of the substructure. The logical nodes are unaskable and they are used as inner or top ones. The context links ensure the common sense ordering of questions on the structure fragments and eliminate the redundant questions.

The process of the knowledge base improvement has been based rather on iterative changes of necessity and sufficiency measures than on significant changes of the knowledge base architecture. Moreover, the training process has been negatively influenced by rather inhomogeneous population of structural types and inaccuracy in the data. The iterative process of the knowledge base improvement has been performed by making use of the methodology described in ref.¹¹. The initial correctness of the decision making has been less than 50%, whereas at the end of the iterative process it has reached 98%.

The final version of the knowledge base (in connection with the FEL-EXPERT shell) can give the classification decision on toxicity only for the compounds, the structure of which can be expressed as a composition of the involved substructures. The number of questions which should be answered by the user to obtain the toxicity classification has varied from 2 to 7. The results of a consultation are expressed as actual probabilities of the toxicity classes (see Fig. 3).

The misclassifications have occured only in the cases of following three amine-boranes:

1. Amino-tris(2-ethoxyphenyl)borane, $LD_{50} = 1500 \text{ mg/kg}$

 $P(H_1) = 0.000$, $P(H_2) = 0.940$, $P(H_3) = 0.914$,

2. *p*-Toluidine-tris(2-propoxyphenyl)borane, $LD_{50} = 87 \text{ mg/kg}$ The LD_{50} values range between 43 and 174 mg/kg.

 $P(H_1) = 0.981$, $P(H_2) = 0.998$, $P(H_3) = 0.000$,



FIG. 3

Posterior probabilities of the FEL-EXPERT classification of the BASE2. Probability for the class H_1 (.....); H_2 (.....); H_3 (.....);

3. Pyridine-tris(2-propoxyphenyl)borane, $LD_{50} = 145 \text{ mg/kg}$ The LD_{50} values range between 90 and 232 mg/kg.

 $P(H_1) = 0.992$, $P(H_2) = 0.932$, $P(H_3) = 0.000$.

It is evident that in all these cases: (i) the LD_{50} values are not far from the boundary between "correct" and "incorrect" classes, (ii) the differencies between the probabilities of the respective classes do not exceed the value of 0.06 in favour of the "incorrect" class and (iii) the probability of the third (remaining) class differs significantly from the probabilities of the two classes under consideration.

Nevertheless, the final results of the consultations can be accounted for as a promising example. The resulting knowledge base can be used as a valuable starting knowledge source for further study of the influence of structural features on the toxicity.

CASE STUDY - EXAMPLE OF BORONIC ACIDS

Fig. 4 shows a fragment of the inference net relevant to 4-substituted phenylboronic acids Ia, $R = CH_3$ and Ib, R = COOH.







The classification pathway into the classes of highly toxic (H_1) , medium toxic (H_2) and slightly toxic (H_3) compounds is illustrated for these two boronic acids which differ dramatically in their toxicities.

1) If we only know that the compound contains the substructure No. 5 (Table I) then (a) this evidence (E_1) directly supports the class of medium toxic compounds (H_2) with the strength of the support L = 9 or slightly toxic compounds (H_3) with a somewhat lower strength L = 1.74. (b) The posterior probability of E_2 becomes higher – this fact would influence the ordering of next questions (the question on E_2 would be preferred).

2) If the substructure No. 22 (Table I) is also present, then the direct support is given to all the classes, but it differs with respect to the corresponding *L*-measures. $L(H_2/E_2) = 9 > L(H_1/E_2) = 8.46 > L(H_3/E_2) = 3.95$. In this stage of the dialogue, it is possible to conclude that all three classes of toxicity are possible, but the most probable seems the class H_2 (strong supports from both evidencies E_1 , E_2 are accumulated in the node H_2).

3) If the user specifies, in the following step, the cycle being a phenyl group, it is possible to deduce that (a) phenyl itself does not possess a discriminatory ability; compounds containing a phenyl in their molecule can have very different toxicities, and (b) at least in the BASE2 cases, slightly preferential influence for the class H_1 can be observed.

4) Let's suppose, the phenylboronic acids under study are substituted in the para--position of their phenyl group. Therefore, it can be concluded that if this fact holds (a) it slightly supports both the classes H_2 and H_3 with the measures L = 1.86 and L = 1.42, respectively, (b) the probabilities of the nodes E_3 and E_6 are rapidly growing (to force their investigation in the next steps of consultation). If the evidence E_4 does not hold, (a) the probabilities of the classes H_2 and H_3 are slightly quashed ($\overline{L} = 0.82$ and $\overline{L} = 0.91$, respectively), (b) the probability of the nodes E_3 and E_6 are

TABLE II

Common d	Posterior probability of		
Compouna —	H ₁	<i>H</i> ₂	H ₃
Ia	0.003	0.965	0.022
Ib	0.000	0.040	0.994

The classification of p-tolylboronic acid and 4-carboxyphenylboronic acid into toxicity classes of the BASE2

slightly changing (to ensure only a small preference to them in the next steps of consultation).

5) The specification of the substituent R in the *p*-substituted phenylboronic acids I let be the final step in the interpretation of toxicity in terms of structural features (substructures). If a tolyl group is present, then the class H_2 is dramatically supported (L = 49), the class H_1 is supported, too (L = 6.83). If there is no tolyl group in the molecule, the evaluation of both the classes H_2 and H_1 becomes slightly lower $(\overline{L} = 0.82 \text{ and } \overline{L} = 0.8$, respectively). On the other hand, if a carboxyphenyl group is present, this fact very dramatically supports the class H_3 (L = 210).

If all the rules (influences) in the fragment of the inference net (Fig. 4) are combined with the facts (substructures present in the molecules of *Ia* and *Ib*), they are obtained the resulting probabilities of the three classes as shows Table II. These results are in full concordance with the experimentally found toxicity for both the compounds (for *Ia* $LD_{50} = 520 \text{ mg/kg}$ and for *Ib* $LD_{50} = 3838 \text{ mg/kg}$) (ref.¹).

CONCLUSIONS

It is evident that the inference of net can lead to valuable information on the influence of substructural features in terms of their discriminatory power expressed in a quantitative way (L, \overline{L}) . Moreover, this influence has been "intelligently" combined with other positive or negative influences, thus the trivial additivity being excluded.

The real knowledge base (and, thus, the inference net) usually contains several hundreds or thousands rules (including logical and context links). Therefore, the presented simple example of the two boronic acids can serve only as an illustration of the interpretation way in the inference net. It has been demonstrated, that the dialogue between the user and the expert system results in: (a) the classification of a set of chemical compounds and (b) the interpretation of the classification process. The latter result can be of particularly high gnoseologic value.

Finally, it should be underlined, that the generality of the obtained interpretation is dramatically dependent on the quality of knowledge base and it thus can represent, in a given moment, only a model extracted from available information. As discussed elsewhere in detail¹¹, for such an extraction "artificial intelligence" approach, advantageously the expert system approach, can be useful. Here the feasibility of the FEL-EXPERT system for these purposes has been demonstrated. The neutron capture therapy being considered, the expert system may be used to search for nontoxic candidates among structurally diverse boron compounds.

REFERENCES

- 1. Kliegel W.: Bor in Biologie, Medizin und Pharmazie. Springer-Verlag, Berlin 1980.
- 2. Proc. 1st Int. Symp. Neutron Capture Therapy (R. G. Fairchild and G. L. Brownell, Eds). Brookhaven Natl. Lab., Upton 1983.

- 3. Proc. 2nd Int. Symp. Neutron Capture Therapy. In: Neutron Capture Therapy (H. Hatanaka, Ed.). Nishimura, Niigata 1986.
- 4. Proc. 3rd Int. Symp. Neutron Capture Therapy. In: Strahlenther. Onkol. 165, No. 2/3 (1989).
- Mařík V., Zdráhal Z., Mendez R. A.: A Skeleton for Expert Systems: MEXEXP 1.0. Techn. Rep. AHR-84-29, C.I.E.A. del Instituto Politécnico Nacional, Mexico City 1984.
- 6. Mařík V., Zdráhal Z.: The FEL-EXPERT and MIFELEX Empty Expert Systems and Their Applications. Working Paper WP-86-61, IIASA, Laxenburg, Austria 1986, p. 47.
- 7. Duda R. O., Hart P. E., Nilsson N. J.: Subjective Bayesian Methods for Rule-Based Inference Systems. Report TN 124, SRI International, Stanford 1976.
- Mařík V., Zdráhal Z. in: Metody umělé inteligence a expertní systémy II (V. Mařík and Z. Zdráhal, Eds), p. 175. Czech Technical University, Prague 1985.
- 9. Štrouf O. in: Chemical Pattern Recognition, p. 13. Research Studies Press, Letchworth 1986.
- 10. Miller V.: Thesis. Czech Technical University, Prague 1984.
- 11. Mařík V., Zdráhal Z.: Expertní systémy. ÚISK, Prague 1987.

Translated by the author (O. Š.).