Rule Induction for Systems Predicting Biological Activity

Philip N. Judson

Heather Lea, Bland Hill, Norwood, Harrogate HG3 1TE, U.K.

Received June 30, 1993*

Knowledge-based expert systems are now in practical use, giving advice about the potential biological activities of substances. Systems depending on the automatic generation of rules for their knowledge bases had the disadvantage that rules were not easily comprehensible to human users, making them difficult to verify. REX and DEREK link rule generation and the application of rules via a knowledge-base language that is fully comprehensible to human users, so that scientists can edit rules and incorporate knowledge coming from diverse sources.

INTRODUCTION

Having been first hailed as the clear leaders for the development of artificial intelligence, and then found wanting, knowledge-based expert systems are now recognized as reliable and effective tools within their limitiations. This paper discusses their use for the prediction of pharmacological, or other biological, activity of substances.

Throughout the paper the term "knowledge-based system" (KBS) means a system that uses existing human knowledge to make decisions or predictions about new situations. A trivial illustration of this approach would be basing the prediction that a helicopter would crash if its rotor blades stopped on the observation that this is what always has happened in the past and that it is intuitively reasonable to expect it, rather than on the solving of an equation for aerodynamic lift for the case where blade motion is zero.

The above example makes the assumption that it is the motion of the rotor blades that keeps a helicopter in the air, rather than some other property. But even simple assumptions can lead to wrong conclusions. Motion is not the only requirement, and a slightly different rule that said that helicopters always stay in the air if their rotor blades are in motion would be wrong (e.g., speed and angle of the blades cannot be neglected). Happily these points are so obvious that a human expert would be unlikely to make the mistake. In many areas of science, such as the prediction of biological activity, there is a more serious lack of understanding of existing knowledge. Some research effort has therefore been directed to rule induction for expert systems—the creation of new knowledge by computer analysis of data.

KNOWLEDGE RELATING TO BIOLOGICAL ACTIVITY

The commercial insecticide shown in Figure 1, Bendiocarb, will be used to illustrate the main types of knowledge that may be relevant to biological activity. Bendiocarb is an insecticidal acetylcholinesterase inhibitor. An expert system may need to take account of knowledge involving a detailed understanding of the mode of action of a compound, and the implications of this requirement will be discussed, but one aspect of the mode of action is sufficient for the purposes of illustration in this paper. A dotted box around part of the structure of Bendiocarb in Figure 1 shows a substructural feature which can mimic acetylcholine, allowing competitive binding to the enzyme site. The carbamate group maps to the

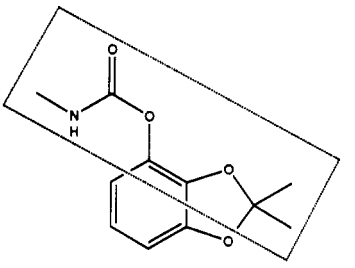

Figure 1. Insecticide, Bendiocarb.

carboxylate group of acetylcholine (its different chemical properties are important, but will not be discussed here). It appears likely that the propyleneoxy fragment, being at an appropriate distance from the carbamate group, mimics the alkyl-substituted ammonium group of acetylcholine.

Since the insecticide must enter the site of action of the esterase, there must be overall restrictions on size and shape which are satisfied by Bendiocarb.

A substance showing activity on an enzyme *in vitro* will not be active *in vivo* unless it has the right properties to move through fat/water barriers to the site of action. An expert system must therefore be able to take account of these properties. The most usual empirical measure used as a guide by human experts is log P, the octanol–water partition coefficient. To exhibit insecticidal activity through acetylcholinesterase inhibition *in vivo*, a compound should normally have a log P value between –0.5 and +3. For example, log P for Bendiocarb is approximately 2. Other properties such as water solubility, volatility, particle size, and even palatability may be important.

So some of the knowledge relevant to insecticidal activity through acetylcholinesterase inhibition might be expressed most generally as follows:

a substance is likely to be insecticidal

IF	it contains a carboxyl-like function of appropriate reactivity at a suitable distance from a center of potential positive charge bearing lipophilic substituents
AND	it is not too large to fit into the enzyme cavity
AND	it has the right physical properties to reach sites of action in an insect.

More specifically, for carbamates like Bendiocarb, a simple rule might be

* Abstract published in *Advance ACS Abstracts*, January 15, 1994.

0095-2338/94/1634-0148$04.50/0

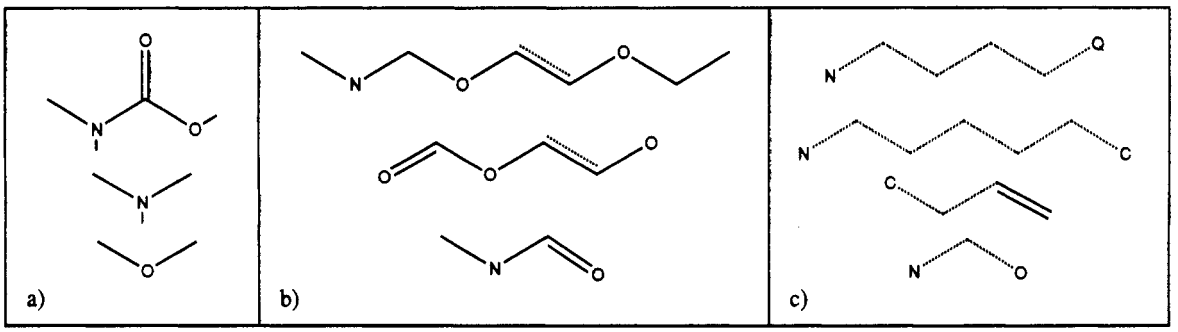

Figure 2. Types of fragments that might be used by (a) TopKat, (b) CaseTox, and (c) Rex.

a substance is likely to be insecticidal

IF it contains a carbamate group bearing small *N*-alkyl substituents approximately 5.2 Å from a nitrogen, oxygen, or sulfur atom, bearing a small- to medium-sized lipophilic substituent or substituents and ideally carrying a positive charge at biological pH

AND it is not too large to fit into the enzyme cavity

AND it has a log P = –0.5 to +3.0

A KBS needs to be able to handle all of the above types of information, and a self-contained rule induction system would have to derive the information from appropriate raw data.

SOME EXISTING SYSTEMS

DEREK[2] is a KBS originally developed to deal with knowledge about toxicophores, but equally suited to dealing with pharmacophores. It uses a knowledge base created by human experts and does not carry out automatic rule induction. DEREK could use 3D information, but in the current implementation distance is normally expressed as topological distance in chemical graphs. In practice, using topological distance is closely analogous to using distance through space: a topological distance (e.g., expressed as a number of bond lengths) sets upper and lower boundaries for 3D distance values, and it is normal to use such distance ranges in 3D-based systems.[8] The structural information for the above carbamate rule is currently expressed in the linear notation language of the DEREK knowledge base, PATRAN[3] (similar to SMILES[4]) as follows:

N(-C)-C(=O)-O&X&X&N,O,S-C

where X represents any atom and & represents any bond.

PATRAN allows qualifying details about atoms and bonds and stereochemistry to be included, although the facilities are not needed in this particular case. For example, the requirement for substitution by particular atoms or groups, or their absence, and for bonds to have particular characteristics can be expressed in patterns such as the following:

N[HS=2]-C[WBOND=YES]–C[FGSNOT=
KETONE,ESTER]-C-[FUSION=NO]C

Further information relating to a rule is recorded in a second DEREK knowledge-base language, CHMTRN.[5] An example shown in Figure 6 is discussed later in this paper.

TopKat[6] is not a KBS as defined above, but it is capable of rule induction. A large, predefined set of structural fragments is used for a TopKat analysis. Thus, insecticidal activity for compounds like Bendiocarb might be expressed as a function of, for example, occurrence of carbamate, amine, ether, and/or thioether groups. A disadvantage of this approach is that while it is mathematically valid, it may not be chemically or biologically valid. For example, the fact that a suitably placed ether group can enhance activity of these insecticides leads to the erroneous conclusion that the presence of an ether group alone gives a nonzero level of activity, which a knowledge of the mode of action of the compounds shows to be invalid.

CaseTox[7] uses a similar approach to TopKat in some ways, but it uses different structural fragments. All linear fragments present in the training set of molecules are used. Recent implementations of CaseTox take account of branching in fragments, and a distinction is made between fragments that are essential for activity and those which only enhance activity.

REX[1] also uses linear fragments, but they are more loosely defined than those used in CaseTox. A *link* in REX is defined by a pair of features and their separation expressed as a number of bond lengths. The features may be specific atom types (including hydrogen atoms attached to heteroatoms), or features such as lone pairs. Thus, for example, one link in Bendiocarb is a nitrogen atom separated from an oxygen atom by five bonds.

The kinds of fragments used by TopKat, CaseTox, and REX are illustrated in Figure 2. They are not discussed further here, that being the subject of previous papers.[1,8]

PROVIDING RULE INDUCTION FOR KNOWLEDGE-BASED SYSTEMS

In a recent study,[9] human experts were apparently better than computer systems at pharmacophore-based prediction of biological activity. This strengthens the view that KBS are best used as repositories for human ideas and knowledge, with separate rule induction tools being provided for use by human rule-writers, not to replace them. Examples from the DEREK knowledge base in Figure 3 illustrate how pharmacophores perceived by human experts are usually more completely defined than the sets of simple fragments found by rule induction in other systems. REX therefore provides for manual intervention following preliminary analysis of sets of structures of active molecules, to allow the human expert to build complex pharmacophores suitable for a knowledge base.

The first stage in a complete analysis using REX to build a rule for DEREK is to choose the members of a set of compounds sharing similar biological activity and of a reference set of compounds for comparison.

Logically, compounds for the *active set* should be chosen on the best available evidence that they have the same mode of action, since the aim is to discover the pharmacophore responsible. But if the pharmacophore is not already known,

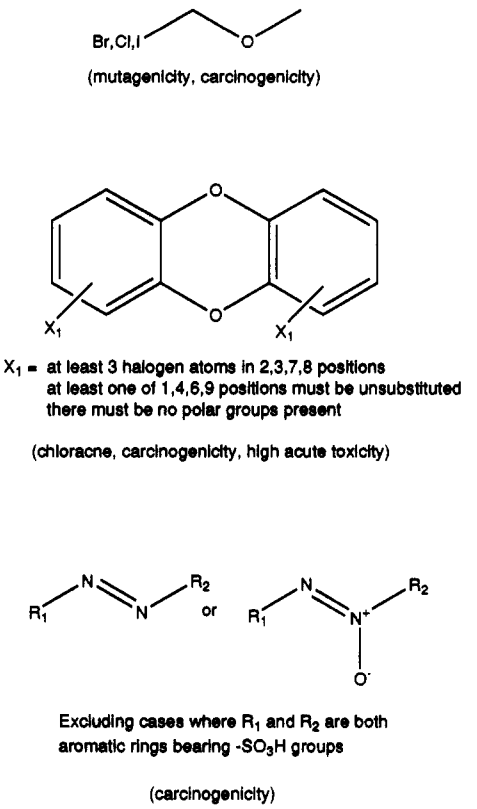

Figure 3. Examples of toxicophores used in the DEREK knowledge base.

it is unlikely that the mode of action is fully understood. Usually, only indirect evidence is available (for example, on the grounds that symptoms are similar) and at least one or two compounds included in the set will turn out to act quite differently. In information science terms they are "false drops", and REX is designed to tolerate their presence.

The ideal set of compounds for comparison, the *reference set*, representing a wide range of structure types, would include no examples of structures omitted from the active set but containing the (as yet undiscovered) pharmacophore. In practice, such structures are invariably present, because their activity has not been discovered or because some other property prevents the expression of their activity. Some can be eliminated prior to computer analysis on the basis of obvious reasons for suspicion (for example, compounds with log P or water solubility values far out of range for compounds normally active in the relevant biological test). The presence of these "missed hits" does not affect the performance of REX provided that they are a minority in the reference set.

Given active and reference sets, REX analyzes all the structures in each and counts all the links it finds. It then ranks the links found in the active set in order of how much more common they are in that set than in the reference set. This very simple approach has been found to be adequate.[8] A user-definable cutoff value determines how many structures in the active set must contain a link for it to be considered significant. At 100%, this cutoff allows a link to be included only if it appears in every active molecule. This is normally too restrictive as it prohibits the false drops mentioned above—structures included mistakenly in the active set do not contain all of the features of the pharmacophore shared by the other members of the set. Set too low, the cutoff allows a link to be included even if it appears in the structures of only one or two active molecules. This is a problem if the active set happens to include some structures with very unusual features that are irrelevant to the activity. Since they are

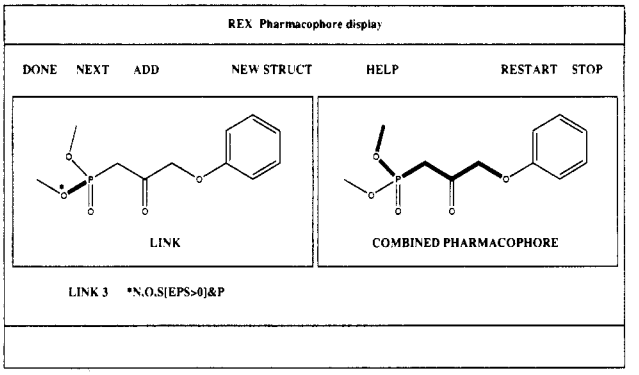

Figure 4. REX pharmacophore display screen.

rare features, perhaps not appearing in any other structures in the reference set, they are recognized as special to the active set and wrongly identified as potential features of a pharmacophore. A cutoff value of about 80% has usually been found satisfactory.

At the end of this process, the most significant links are listed to the user. Currently, links are ignored unless they are at least twice as common in the active set as in the inactive set, and the length of the list is limited to a maximum of 16. Both of these limits have been chosen empirically.

The set of significant links could be used as an "ensemble pharmacophore", in the same way that sets are used in TopKat and CaseTox but, taken together with the structures used for the analysis, the set carries information about the interrelationships between the links that should not be ignored. As a trivial example, an analysis of insecticidal carbamates might show two significant links to be a lone pair on an sp^2 oxygen atom two bonds from an sp^3 oxygen atom and a lone pair on an sp^2 oxygen atom two bonds from a nitrogen atom. 3-Methoxycarbonylpropionamide has these features even though it does not contain a carbamate group—an essential component of the true pharmacophore. A second stage in the analyses creates the more rigorous pharmacophore information that is needed.

REX allows the user to influence the construction of a complete pharmacophore by looking at the mappings of links onto individual examples from the active set (or other structures), using the screen shown in Figure 4. The left-hand picture of the chosen molecule shows a mapping of the link currently under consideration. The user can decide whether this mapping of the link is appropriate and should be added to a growing, complex pharmacophore, displayed in the right-hand picture. Further links and/or molecules can be selected and studied in the same way. Thus a pharmacophore is gradually built, as illustrated for the mapping of several links onto one molecule in Figure 5. During this operation, the user may notice features missed by REX, for one reason or another, and so tools are provided for editing the pharmacophore.

Subsequently, the user can map the completed pharmacophore onto any molecule or set of molecules. In the first instance, mapping onto all the molecules in the active set provides confirmation that the pharmacophore is appropriate and may also lead to some interesting discoveries about odd members of the set. For example, if mapping fails on one or two members of the set, this may reveal that they are false drops or may draw attention to a feature which, though unusually common in the active set, is not essential for biological activity.

Once the refinement process is completed, knowledge about a potentially new pharmacophore is available to the drug

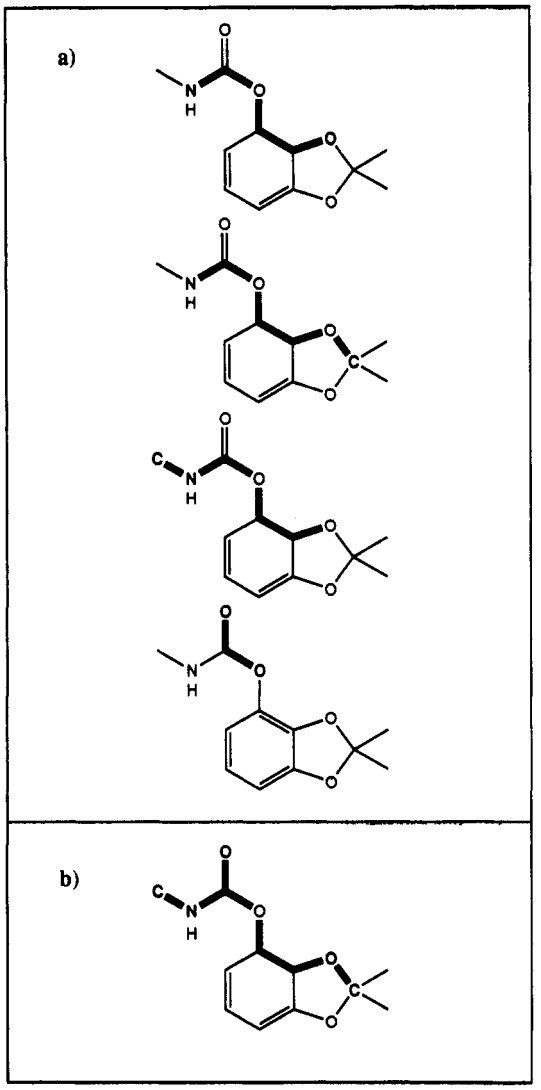

Figure 5. Mapping fragments onto an active molecule (a) to build a pharmacophore (b).

designer. The knowledge remains available in electronic form after the individual ceases work on the project and can later alert a user to the potential activity of a novel compound, even if the user has forgotten, or never knew, about the pharmacophore. REX itself may be used for this purpose, or the pharmacophore may form the basis of a rule for DEREK.

ADDING TO THE RULE FOR DEREK

Finding the putative pharmacophore in a set of active compounds is only part of the process of developing a good rule for a knowledge base. As mentioned above, it is necessary to assess structural factors which may prevent the expression of activity even though a pharmacophore is present (e.g., because of steric interference), physicochemical factors such as fat/water partition and water solubility, and physical factors such as particle size. The first may be investigated by examining molecules from the REX reference set (i.e., inactive molecules), to which the pharmacophore maps, and by using molecular modeling methods. The second and third depend on physicochemical measurements and/or calculations, such as the statistical and algebraic techniques used in QSAR work.

Separating tools for the acquisition of knowledge from those for applying the knowledge and providing a suitable language for the interface between them allow the scientist to select the best tools for each part of the acquisition process and to record

```
RULE   100
RISK   Skin sensitisation &
       if physical properties allow easy skin penetration
...
STARTP
0=C[HS=1]-C=C
ENDP
...
IF GOT*LOGP
 BEGIN CHECK*LOGP
  IF LOGP .GE. LOGP*MIN AND:IF LOGP .LE. LOGP*MAX &
   THEN NEW*RISK   Skin sensitisation &
                   high risk because of easy skin penetration
  IF LOGP .LT. LOGP*MIN OR:IF LOGP .GT. LOGP*MAX &
   THEN NEW*RISK   Skin sensitisation &
                   low risk because of low skin penetration
 BLKEND CHECK*LOGP
...
```

Figure 6. Extract from a DEREK rule for skin sensitization caused by an enal (this rule has been simplified for the purpose of illustration).

and maintain the resultant knowledge in a form that is accessible to human users—qualifications and details are added to the prototype rule created by REX in the DEREK knowledge-base language. For example, the enal substructure is commonly associated with skin sensitization, but the effect is only observed if a substance containing the substructure has the right physical properties to penetrate the skin. These properties include the fat–water partition coefficient, conveniently modeled by log *P*. An extract from a corresponding rule in the DEREK knowledge base language in Figure 6 shows how the additional information is used to change the advice given to a user.

DEALING WITH KNOWLEDGE ABOUT METABOLISM AND CONTAMINANTS

The biological properties associated with many—perhaps most—substances are really the properties of their metabolites. An innocuous substance may be converted to a toxic one, or an inherently toxic substance may be converted to a harmless one.

For example, *n*-hexane, alone among the simple hydrocarbons, causes neurotoxic effects because it is enzymatically converted to a 1,4-diketone—the actual toxicant which acts via the formation of pyrroles at key amine sites *in vivo* (see Figure 7). A 1,4-diketone containing a fully substituted carbon atom between the two ketone sites does not show this toxicity, because aromatization of the amine to a pyrrole is prevented. *n*-Heptane undergoes similar metabolic oxidation at the methylene groups adjacent to the end of the carbon chain, but the resultant 1,5-diketone does not form a stable adduct with amines so readily. On the other hand, 3-hydroxyheptane, having suitably-located functionality for oxidation to a ketone, can form the toxic 2,5-heptadiene.

DEREK rules check for four-carbon units with appropriate substitution to facilitate these metabolic conversions and to allow aromatization to a pyrrole. The end user is warned of the potential hazard associated with the query substance, and an explanation is given that activity will be potentiated through metabolism.

Some other rules in DEREK warn of a different way in which toxicological hazard may be associated with a substance which is in itself harmless. A knowledge of the method of synthesis, or chemical properties of, a substance may lead a human expert to suspect the presence of dangerous contaminants. Thus, for example, one rule in DEREK recognizes the presence of a polyhalophenol (or a simple derivative) and warns of the possible toxic hazards arising from contamination with dioxins.

A more elegant approach to dealing with metabolism would be to separate the systems dealing with prediction of toxicity and prediction of metabolism, but to allow the first to call the

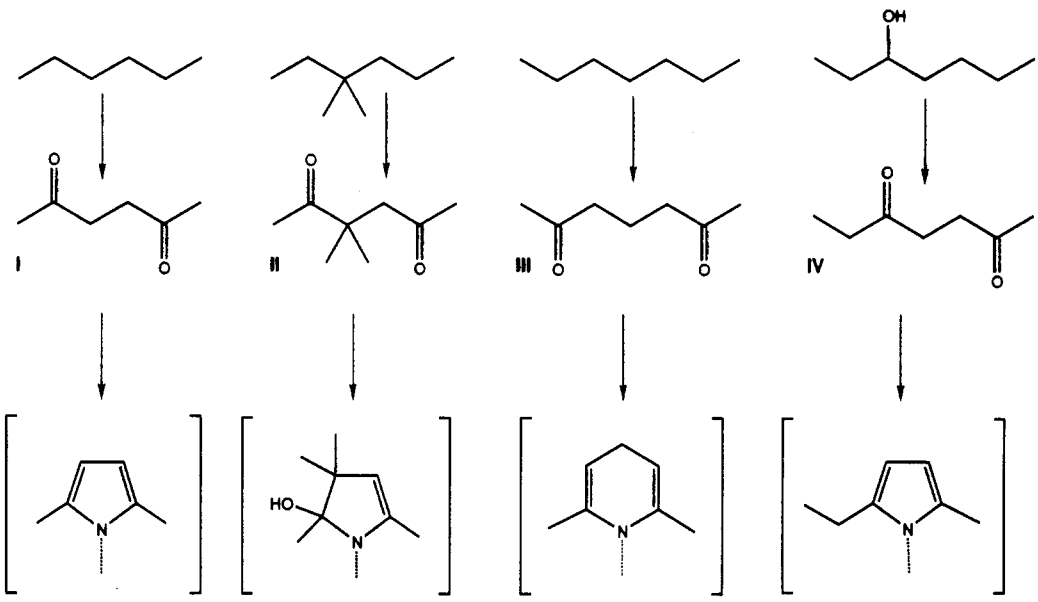

Figure 7. Metabolism of hexane and related compounds to give toxic (I, IV) and nontoxic (II, III) products.

second automatically. Existing systems for predicting metabolism (e.g., MetabolExpert[10]) have been disappointing because they have generated too many potential metabolites and have been unable to discern the more important routes. Asking the question "can this specific metabolite be generated" reduces the size of the search tree in a way analogous to the use of goals and subgoals in the synthesis design program, LHASA,[11] and research is being carried out into this approach.

An example is illustrated in Figure 8. Suppose that a user asks for suggestions for the synthesis of the compound shown in Figure 8a. LHASA recognizes that it is strategically desirable to break the bond marked by a curly line, retrosynthetically (breaking a bond adjacent to a ring normally achieves significant simplification). A knowledge-base search finds that the Grignard reaction, among others, can be used to break carbon–carbon bonds in the retrosynthetic direction (Figure 8b), so it is chosen as a goal. But, in the retrosynthetic direction, the target structure does not have the functionality required for the Grignard reaction—in the example in Figure 8b a hydroxy group β to the atom bearing the new bond is lacking. A further search in the knowledge base for ways to get hydroxy groups retrosynthetically turns up the esterification reaction shown in Figure 8c. The target structure contains an ester group in the right place for the retroreaction, and a synthesis is proposed (Figure 8d). Note how much more efficient this is than simply looking for all possible precursors to the target, and then all possible precursors to each of them.

In the toxicology case, the goal is a particular metabolite, and metabolic reactions need to be evaluated only if they lead to the goal or directly compete with reactions on that path.

KNOWLEDGE-BASE LANGUAGE

The knowledge-base languages used by DEREK are versatile, and it is rarely impossible to express an idea relating to chemical structures (when such a problem has arisen, it has been resolved by writing suitable program code to add new words to the languages). Nevertheless, recent developments in computer technology offer opportunities for a new approach.[12-14] Research is in progress to develop a new language, and details will be published.

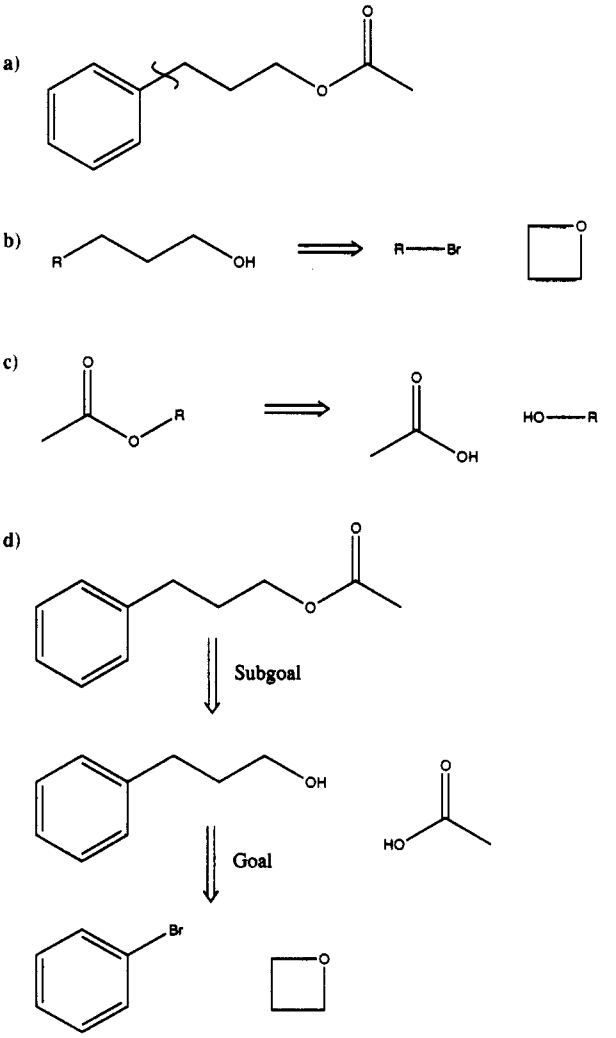

Figure 8. Goals and Subgoals in a synthesis design system: (a) the target, showing a strategic bond perceived by the system; (b) a goal reaction, breaking a carbon–carbon bond retrosynthetically; (c) a subgoal reaction, creating an alcohol gorup retrosynthetically; (d) a possible retrosynthesis.

CONCLUSION

Knowledge-based expert systems are now in practical use for the storage of information about the biological activities

of substances. Increasingly better tools are being provided to help scientists to discover what makes certain substances active. These tools must be linked to KBS in ways that allow scientists to scrutinize and modify the intermediate knowledge bases, and languages used in end-user interfaces and knowledge bases must be readily comprehensible to human users.

REFERENCES AND NOTES

(1) Judson, P. N. QSAR and Expert Systems in the Prediction of Biological Activity. *Pestic. Sci.* **1992**, *36*, 155–60.

(2) Sanderson, D. M.; Earnshaw, C. G. Computer Prediction of Possible Toxic Action from Chemical Structure; The DEREK System. *Hum. Exp. Toxicol.* **1991**, *10*, 261–73.

(3) (a) Marshall, C. Computer Assisted Design of Organic Synthesis. Ph.D. Thesis. University of Leeds, U.K., 1984. (b) Hopkinson, G. A. Computer-Assisted Organic Synthesis Design. Ph.D. Thesis. University of Leeds, U.K., 1985.

(4) Weininger, D. SMILES, A Chemical Language and Information System. 1. Introduction to Methodology and Encoding Rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36.

(5) Orf, H. W. Ph.D. Thesis. Harvard University, Cambrige, MA, 1976; Appendix C.

(6) Enslein, K.; Borgstedt, H. H.; Blake, B. W.; Hart, H. B. Estimation of Maximum Tolerated Dose for Long-Term Bioassays from Acute Lethal Dose and Structure by QSAR. *Risk Anal.* **1991**, *11*, 509–17.

(7) Klopman, G. Predicting Toxicity through a Computer Automated Structure Evaluation Program. *Environ. Health Perspect.* **1985**, *61*, 269–74.

(8) Judson, P. N. Structural Similarity Searching Using Descriptors Developed for Structure–Activity Relationship Studies. *J. Chem. Inf. Comput. Sci.* **1992**, *32*, 657–663.

(9) Predicting Chemical Carcinogens in Rodents. A meeting held at the National Institute of Environmental Health Science, Research Triangle Park, NC, May 24–25, 1993.

(10) HazardExpert and MetabolExpert are supplied by CompuDrug Chemistry Ltd., H-1136 Budapest, Fürst Sándor utca 5, Hungary.

(11) The LHASA and DEREK systems are supplied by LHASA UK, School of Chemistry, University of Leeds, Leeds LS2 9JT, U.K., and by A. K. Long, Harvard Chemistry Department, 12 Oxford St., Cambridge, MA 02138.

(12) Krause, P. J.; Clark, D. A. *Representing uncertain knowledge-an artificial intelligence approach*; Intellect: Oxford, England, 1993.

(13) Fox, J.; Clarke, M. Towards a formalisation of arguments in decision making. In *Proceedings of AAAI Spring Symposium on Argumentation and Belief*; AAAI: Stanford, CA, 1991.

(14) Krause, P. J.; Ambler, S. J.; Fox, J. The development of a logic of argumentation. In *Advanced methods in artificial intelligence*; Bouchon-Meunier, B., Valverde, L., Yager, R., Eds.; Springer-Verlag: Heidelberg, Germany, 1993.