RSC Theoretical and Computational Chemistry Series

Philip Judson

Knowledge-Based Expert Systems in Chemistry

Not Counting on Computers

RSC Publishing

Knowledge-based Expert Systems in Chemistry

Not Counting on Computers

Philip Judson
Consultant, Harrogate, UK

RSC Publishing

RSC Theoretical and Computational Chemistry Series No. 1

ISBN: 978-0-85404-160-2
ISSN: 2041-3181

A catalogue record for this book is available from the British Library

Published by The Royal Society of Chemistry,
Thomas Graham House, Science Park, Milton Road,
Cambridge CB4 0WF, UK

Registered Charity Number 207890

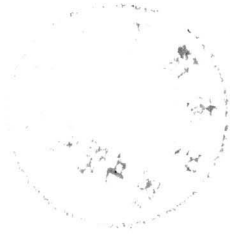
For further information see our web site at www.rsc.org

Knowledge-based Expert Systems in Chemistry
Not Counting on Computers

RSC Theoretical and Computational Chemistry Series

Series Editor:
Jonathan Hirst, *University of Nottingham, Nottingham, UK*

Titles in the Series:
1: Knowledge-based Expert Systems in Chemistry: Not Counting on Computers



How to obtain future titles on publication:
A standing order plan is available for this series. A standing order will bring delivery of each new volume immediately on publication.

For further information please contact:
Sales and Customer Care, Royal Society of Chemistry, Thomas Graham House, Science Park, Milton Road, Cambridge, CB4 0WF, UK
Telephone: +44 (0)1223 432360, Fax: +44 (0)1223 420247, Email: sales@rsc.org
Visit our website at http://www.rsc.org/Shop/Books/

Preface

Computers began to think, in their simple way, in the 1960s; they ceased to be mere adding machines. People interested in using computers to help with chemical synthesis design were among the earliest researchers in the field of artificial intelligence, and the results of their work have had a major impact on chemical software development ever since. I had the good fortune to become involved just at the time when the scientific community began to take chemical information and knowledge systems seriously – in the 1980s, twenty years after the pioneers had taken the lead. I have watched some of the systems grow from research ideas into mature products and in this book I write about them. The book is biased because I have written about what I know. However, I have been involved in some key areas. Knowledge-based and reasoning-based approaches are in routine use to predict or plan chemical reactions, to predict toxicity, and to predict metabolism and biodegradation, and spin-off from research into them has produced the best-known chemical structure and reaction database systems. There has not been a book specifically about them until now and I think it is time to fill the gap.

I came into this field almost by chance and, while chemistry has remained a central theme, it has taken me into biology, aspects of mathematics and theories of logic, and even psychology and law. I have crossed the boundaries between industry and academia and collaborated with scientists on every continent, and what I write is about the results of that collective effort. As everyone writes in the preface to a book, because it is true, it would be impractical to list the names of everyone to whom I owe gratitude. However, I do want to thank David A. Evans, A. Peter Johnson and Alan K. Long for inspiring my interest in knowledge-based systems in chemistry and their patient support while I developed an understanding of the science. I thank Alan K. Long, Anthony Long and Martin Ott for their comments and advice on some of the sections in this book.

RSC Theoretical and Computational Chemistry Series No. 1
Knowledge-based Expert Systems in Chemistry: Not Counting on Computers
By Philip Judson

Published by the Royal Society of Chemistry, www.rsc.org

Contents

RSC Theoretical and Computational Chemistry Series No. 1
Knowledge-based Expert Systems in Chemistry: Not Counting on Computers
By Philip Judson

Published by the Royal Society of Chemistry, www.rsc.org

CHAPTER 1

Artificial Intelligence – Making Use of Reasoning

Launched by half a dozen young men at a run, a three-metre long paper dart can fly successfully, dare we even say "gracefully", the length of a research station canteen before making an unfortunate landing in the director of research's Christmas lunch. It is just a question of getting the aerodynamics right. My school mathematics teacher reminded us on most days (several times on some) that all science is mathematics. But was it only the power of numbers he had in mind? Does science come down to a sweatshop full of equations mindlessly crunching numbers, real and imaginary?

Contrary to the perceptions of many people outside science, as well as too many inside it, science is not about proving facts: it is about testing hypotheses and theories; ultimately, it is about people and their opinions. Simple, rigid application of rules of aerodynamics may get you a paper dart that flies but in many fields human decision making is best supported by reasoned argument or the use of analogy and not much helped by numerical answers. The minimum braking distance for a car travelling at forty miles per hour is twenty-four metres, according to the Driving Manual from the Driving Standards Agency.[1] Assuming you can countenance the required mixing of miles and metres, does this information help you to drive more safely? Have you any more idea than I have how far ahead an imaginary twenty-four metre boundary-marker precedes you along the road?

And there is a further problem. "Numbers out" implies "numbers in", so what do you do if you have no numbers to put in? A regrettably popular solution is to invent them – or at least to come up with dubious estimates to feed into a model that demands them, which is close to invention. It is the only option if you want to apply numerical methods and to give numbers to the people asking for solutions. That numbers make people feel comfortable is a bigger problem than it may at first appear to be, too. Uncritical recipients of

RSC Theoretical and Computational Chemistry Series No. 1
Knowledge-based Expert Systems in Chemistry: Not Counting on Computers
By Philip Judson

Published by the Royal Society of Chemistry, www.rsc.org

numerical answers tend to believe them, and to act on them, without probing very deeply. More sceptical recipients want to judge for themselves how meaningful the answers are but often find that the kind of supporting evidence associated with a numerical method is not much help. Many are the controversies over whether this or that numerical method is more precise but they are missing the point if the data are far less precise than the method. Perhaps numbers are unnecessary – even unsuitable – for expressing some kinds of scientific knowledge.

There are circumstances in which numerical methods are highly reliable. Aeroplanes stay up in the sky and make it safely to earth where they are supposed to do. Chemical plants run twenty-four hours a day, year in year out. Numerical methods work routinely in physical chemistry laboratories, and toxicology and pharmacology departments. But it is unlikely that the designers of the three-metre paper dart that took flight at the start of this chapter did any calculations at all. My guess is that they just went with a gut feeling based on years of experience making little ones.

This book is about uses of artificial intelligence (AI) and databases in computational chemistry and related science where qualitative output may be of more practical use than quantitative output. It touches on quantitative structure–activity relationships (QSAR) and how they can inform qualitative predictions, but it is not about QSAR. Neither is it a book about molecular modelling. Both subjects are well-covered in too many books to list comprehensively. A few examples are given in the references at the end of this chapter.[2–6] This book focuses on less widely described and yet, probably, more widely-used applications of AI in chemistry.

The term "artificial intelligence" carries with it notions of thinking computers but, as a radio personality in former times would have had it,[7] it all depends on what you mean by intelligence. If you type "Liebig Consender" into the Google™ search box, Google™ responds with "Did you mean *Liebig Condenser*" and provides a list of corresponding links without waiting for an answer. That is worryingly *like* intelligent behaviour whether it *is* intelligent behaviour or not. Arguments continue about whether tests for artificial intelligence such as the Turing test[8] are valid and whether a categorical test or set of tests can be devised. Perhaps it is sufficient to require that to be intelligent a system must be able to learn, be able to reason, be creative, and be able to explain itself persuasively. Currently, no artificial intelligence system can claim to have all of these characteristics. Individual systems typically have two or three.

To count as intelligent, solving problems needs to involve a degree of novel thinking, *i.e.* creativity. Restating the known, specific answer to a question requires only memory. Compare the following questions and answers. The first answer merely reproduces a single fact. Generating the second answer, simple though it is, requires reasoning and a degree of creativity.

"Where's the sugar?"

"In the sugar bowl".

"Where will the sugar be in this supermarket?"

"A lot of supermarkets put it near the tea and coffee, so it could be along the aisle labelled 'tea and coffee'. Alternatively, it might be in the aisle labelled 'baking'. Let's try 'baking' first – it is nearer".

One of the first computer systems to behave like an expert using a logical sequence of questions and answers to solve a problem was MYCIN,[9] a system to support medical diagnosis.

"Doctor, I keep getting these terrible headaches".

"Sorry to hear that. Is there any pattern to when the headaches occur?"

"Now you ask, they do seem to come mostly on Sunday mornings".

"And what do you do on Saturday evenings?"

The doctor's questions are not arbitrary. You can see how they are directed by the patient's responses. You can probably see where they are leading, too, but the doctor would still want to ask further questions to rule out all the possibilities before jumping to the obvious conclusion about the patient's Saturday nights out on the town. The aim of the MYCIN experiment was to design a computer system capable of choosing appropriate sequences of questions similarly, in order to reach a diagnosis efficiently.

This kind of reasoning is common throughout science although it often does not involve a dialogue; the questions may be implicit in a process of thought rather than consciously asked. Suppose you know that:

many α,β-unsaturated aldehydes cause skin sensitisation;

for activity to be expressed a compound must penetrate the skin;

compounds with low fat/water partition coefficients do not penetrate the skin easily:

many imines can be hydrolysed easily in living systems to generate aldehydes.

Actually, the story for skin sensitisers is better understood and can be more fully and more usefully described than this, but what we have will do for the purposes of illustration. Suppose you are shown the structure of a novel α,β-unsaturated imine and asked for an assessment of its potential to cause skin sensitisation. You will be aware that the imine might be converted into a potentially skin-sensitising aldehyde. If you have access to suitable methods you will get an estimate of the fat/water partition coefficient for the imine in order to make a judgement about whether it will penetrate the skin (most likely you will use a calculated log*P* value as a measure of fat/water partition coefficient, but there is more about that later in this book). You will presumably have the gumption to consider the partition coefficient for the aldehyde as well, in case the imine is unstable enough to hydrolyse on the surface of the skin.

Depending on the information, you will come up with conclusions and explanations such as:

"the query substance is likely to be a skin sensitiser because it has the right partition coefficient to penetrate the skin and the potential to be converted

into an α,β-unsaturated aldehyde – a class of compounds including many skin sensitisers";

"the query is not likely to be a skin sensitiser because although it is an imine which could be converted into an α,β-unsaturated aldehyde – a class of compounds including many skin sensitisers – both compounds have such low fat/water partition coefficients that they are unlikely to penetrate the skin";

"the situation is equivocal because the imine has too high a fat/water partition coefficient to penetrate the skin easily but the related aldehyde has a lower fat/water partition coefficient and I do not know how readily the imine will hydrolyse to the aldehyde on the skin surface."

Systems in which a reasoning engine solves problems by applying rules from a knowledge base compiled by human experts were originally called "expert systems", on the grounds that they behave like experts. In this book they are distinguished by being called "knowledge-based systems". They use reasoning to varying degrees and they are creative in the sense that they solve novel problems and make predictions. The particular strength of the best of them is their ability to explain themselves. For example, there is fairly good understanding of why α,β-unsaturated aldehydes are skin sensitisers. The human compilers of a knowledge base can include that information so that the expert system can present it to a user when it makes a prediction and can explain how it reached its conclusion.

Given access to structures and biological data for lots of compounds, you might discover the rule that α,β-unsaturated aldehydes are often skin sensitisers, assuming you were not overwhelmed by the quantity of data. Knowledge-based systems as defined here make no attempt to discover rules from patterns in data – they simply apply the rules put into them by human experts. In terms of the criteria for intelligence, they are unable to learn for themselves. The more general term, "expert system", was later extended to include systems that generate their own models by statistical methods and apply them. While these systems are perhaps nearer to all-rounders in the stakes for showing intelligence than knowledge-based systems, they fall down on explaining themselves. They cannot go beyond presenting the statistical evidence for their rules.

A speaker remarked at a meeting I attended that "An expert system is one that gives the answers an expert would give . . . including the wrong ones". It might be fairer to compare consulting a knowledge-based system (which is what he was talking about at the time) with consulting a group of human experts rather than one, since knowledge bases are normally compiled from collective knowledge, not just individual knowledge, but his warning stands. Other people have, only half-jokingly, suggested that an expert system is one suitable only for use by an expert. That may be over-cautious but users of expert systems should at least be thinking and well-informed: it is what you would expect of someone taking advice from a team of experts.

References

1. N. Lynch and A. Wood, *The Driving Manual*, Her Majesty's Stationery Office, Norwich, England, 1992.
2. L. Eriksson, E. Johansson, N. Kettaneh-Wold and S. Wold, *Multi- and Megavariate Data Analysis*, Umetrics AB, Umeå, Sweden, 2001.
3. C. Hansch and A. Leo, *Exploring QSAR: Fundamentals and Applications in Chemistry and Biology*, Am. Chem. Soc., Washington DC, USA, 1995.
4. D. Livingstone, *Data Analysis for Chemists: Applications to QSAR and Chemical Product Design*, OUP, England, 1995.
5. T. Schlick, *Molecular Modelling and Simulations*, Springer, New York, 2002.
6. A. R. Leach, *Molecular Modelling: Principles and Applications*, Pearson Education EMA, Essex, England, 2nd edn, 2001.
7. C. E. M. Joad quoted by G. E. Penketh, *Anal. Proc.*, 1980, **17**, 163–4.
8. A. Turing, Computing Machinery and Intelligence, *Mind*, 1950, **50**, 433–60.
9. E. H. Shortcliffe, *Computer-Based Medical Consultation: MYCIN*, Elsevier Science Publications, New York, 1976.

CHAPTER 2

Synthesis Planning by Computer

Organic synthesis chemists are used to working with ideas and rules of thumb. They are not inclined to plan reaction sequences to novel compounds on the basis of kinetic or thermodynamic calculations – indeed, they are rarely in the position to do so because data of sufficient reliability are not available for the calculations – but they have a reasonable success rate. How do they do it? Could a computer emulate the thinking of a chemist who works out a practical synthetic route to a complicated organic compound?

The tale is told of a conversation over a few beers one evening between three eminent chemists famed for their work in organic synthesis – Elias J. Corey, Alexander R. Todd and Robert B. Woodward. Corey, it is said, expressed the view that computers would eventually be capable of matching or even outclassing human reasoning; soon there would be machines capable of designing chemical syntheses just as well as chemists do. Todd and Woodward were sceptical, it is said, arguing that chemical synthesis was an art more than a science, calling for imagination and creativity well beyond the capacity of a computer. Corey saw how a computer might reason like a chemist and he proposed to set up a project to demonstrate the feasibility of his ideas. The story may be apocryphal but it does not matter if it is. The exciting thing is that Corey recognised a new challenge well beyond the everyday goals of most researchers and took it on. He was not alone in seeing and taking up the challenge – there were others who will feature in this chapter and the next – but his project proliferated like the mustard tree in the parable so that by now every chemist is familiar with at least one spin-off computer application that roosts in its branches.

Corey's project to develop a synthesis-planning program, OCSS ("Organic Chemical Simulation of Synthesis"), started in the 1960s and was described in a paper in *Science* in 1969.[1] By 1971, when a paper was submitted to the Journal of the American Chemical Society,[2] the program had been re-implemented as

RSC Theoretical and Computational Chemistry Series No. 1
Knowledge-based Expert Systems in Chemistry: Not Counting on Computers
By Philip Judson

Published by the Royal Society of Chemistry, www.rsc.org

LHASA (Logic and Heuristics Applied to Synthetic Analysis) and the project was expanding.

Right from the start the plan was to develop a computer system that did not just think like a chemist, but communicated like one, too. Computer graphics was in its infancy. The computer mouse was yet to come to public notice – Douglas Engelbart filed his application for a patent in 1967[3] – but there were systems that linked a graphics tablet, or "bit pad", to a vector graphics screen (a line is displayed on a vector graphics screen by scanning the electron beam between the coordinates of the ends of the line, whereas in a television or a modern personal computer system the screen is scanned systematically from side to side and top to bottom and the beam is activated at the right moments to illuminate the pixels on the screen that lie on the line). Other researchers interested in using computers for chemistry were developing representations of chemical structures to suit computers, but in this project the computer would be expected to use the representations favoured by organic chemists – structural diagrams. In their paper in 1969,[1] Corey and Wipke wrote, "The following general requirements for the computer system were envisaged at the outset: (i) that it be an 'interactive system' allowing facile graphical communication of both input and output in a form most convenient and natural for the chemist . . . ".

A structural diagram is full of implicit information for a chemist that would not be perceived by someone not trained in chemistry. It is not a picture of a molecule, in as much as there can be a picture of one; it tells you what is connected to what, and how, but it does not tell you the three dimensional locations of atoms: like the map of the London Underground it is a graph. To make useful inferences, the computer needs to be able to "see" the graph like a chemist sees it, and so a chemical perception module in LHASA fills checklists for the atoms and bonds in a molecule for use in subsequent processing. For example, if a carbon atom is found to be bonded through a double bond to one oxygen atom and through a single bond to another oxygen atom which itself bears a hydrogen atom, the carbon atom can be flagged as the centre of a carboxylic acid group; if an atom is at a fusion point between two rings (which would have implications for its reactivity) it can be flagged as a "fusion atom".

Computer perception of a molecule may put the computer in the position to think about it the way a chemist would, but how does a chemist think of ways to synthesise even a simple molecule? The question embodies a host of others each of which probably has more than one answer. Corey would have been well-placed to look for answers suited to computer-implementation, having formulated his ideas for the retrosynthetic approach to chemical synthesis design for which he was later to receive a Nobel Prize in Chemistry,[4,5,6,7] and his thinking on the subject and his work on a computer system must surely have fed each other.

The essence of the retrosynthetic approach is that the target molecule contains the clues to the ways in which it might be constructed. That might be obvious but stating something explicitly and letting it lead your thinking can