Computer-aided organic synthesis

Matthew H. Todd

Received 10th September 2004
First published as an Advance Article on the web 8th February 2005
DOI: 10.1039/b104620a

It is tempting for those in the field of organic synthesis to liken the process of retrosynthesis to a game of chess. That the world chess champion was recently defeated by a computer leads us to think that perhaps new and powerful computing methods could be applied to synthetic problems. Here the analogy between synthesis and chess is outlined. Achievements in the 35-year history of computer-aided synthetic design are described, followed by some more recent developments.

"The view that machines cannot give rise to surprises is due, I believe, to a fallacy to which philosophers and mathematicians are particularly subject. This is the assumption that as soon as a fact is presented to a mind all consequences of that fact spring into the mind simultaneously with it. It is a very useful assumption under many circumstances, but one too easily forgets that it is false. A natural consequence of doing so is that one then assumes that there is no virtue in the mere working out of consequences from data and general principles." Alan Turing.[1]

Introduction

In 1996, Garry Kasparov, the world chess champion, and widely regarded as the strongest chess player in the game's history, went up against an IBM supercomputer, Deep Blue, for a six-game match. The match was tied after four games, but Kasparov won the final two to take the match with a decisive 4–2 final score. A rematch was arranged, which took place in May 1997 in New York City.[2] The IBM team considerably modified and improved Deep Blue.[3] The capabilities of this machine were truly formidable to any would-be human opponent (*vide infra*).

Again, the match was level after four games. Then on May 10th came game 5, with Kasparov playing white and Deep Blue black. The game quickly reached the position shown in Fig. 1. A number of moves suggest themselves for black (there are 48 legal moves in this position), with castling looking the most useful. Deep Blue decided to move the pawn on h7 forward to h5.

This was a major surprise. As Yasser Seirawan, three-times US chess champion and one of the match commentators said, "Who's been programming this machine?"[4] The move is a surprise because it is very much not the sort of move computers have tended to play in chess over the years, and feels more 'human' in character. The game continued for about another 40 moves. Deep Blue finished the game in impressive fashion, employing the great depth of its searches to salvage a draw from apparently hopeless prospects in the endgame. After the match, Kasparov repeated his earlier demand to see the printouts of the computer log, since he suspected there had been human intervention.

The tournament was decided on the final game, which was won by Deep Blue, and which therefore takes its place in history as the moment a computer finally won a tournament against a reigning world chess champion. Kasparov asked for a rematch under the usual tournament conditions (a non-IBM-organised event consisting of the usual 10 games), but IBM has so far refused. Indeed after the match the Deep Blue project

Matthew Todd

Matthew Todd completed his PhD at Cambridge University in 1998, and then spent two years as a Wellcome Trust postdoctoral fellow at the University of California, Berkeley with Professor Paul Bartlett. He returned to New Hall College, Cambridge as Fellow in Chemistry, before being appointed Lecturer in Organic Chemistry at Queen Mary, London, in September 2001. His interests range from asymmetric catalysis to chemical biology.

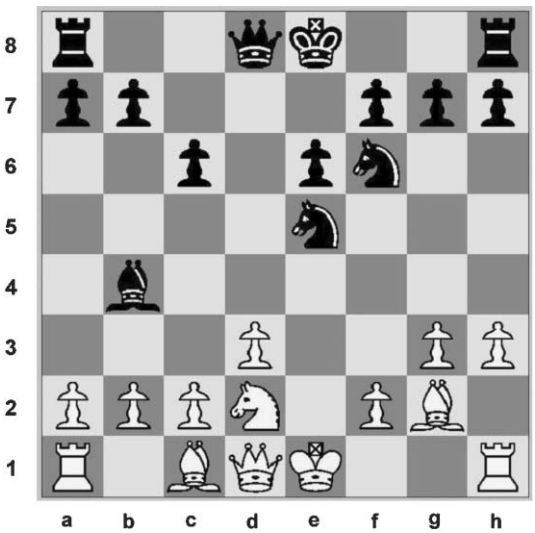


Fig. 1 Game 5, Deep Blue (black) to move.

was wound up, since they had achieved what they had set out to do.

The fascinating point about this confrontation is the contrast in the styles of play. Both were drawing on huge strategic resources; Deep Blue's originating from its programmed knowledge base, Kasparov's from years of practice and analysis. Both were assessing critical lines to great depth. But at the table, while Kasparov was focusing on a relatively small number of lines, the computer was analysing millions upon millions of moves, and with a thoroughness that no human player could ever manage. In vernacular, Kasparov was using a good deal of human intuition, whilst Deep Blue was simply cycling through endless possibilities, coldly scoring each.

The analogy between chess and organic synthesis

Is planning a retrosynthesis really like a chess game?[5] There are clear similarities:

a) Both operate on a set of fairly simple rules.

b) A complex scoring function is required for evaluation of the best move in chess by a computer, whereas humans may 'feel' which is best with a minimal analysis, using pattern recognition and intuition to great effect. A similar contrast exists with regards synthetic design.

c) Analysis of the problem generates a 'tree' of possibilities in both, with some paths being successful and others not. There is usually a multitude of good paths, except in situations where an early decision enforces some subsequent choices, like a check in chess. There is a large 'combinatorial explosion' of the move tree in chess just as there is for the synthesis tree in chemistry. The branching that arises requires pruning *via* a set of heuristics, or 'rules of thumb,' if any serious depth is to be analysed, or if those lines that appear promising are to be actively extended. Those in the respective fields may disagree on the minutiae, but most would agree which strategies are to be labelled 'good' and which 'bad'. More abstractly, exceptional synthetic routes may be described as 'beautiful' in the same way as there are 'immortal' chess games.

d) Good lines may only become apparent upon reaching a certain depth of analysis. Good lines may emerge from apparently low-scoring pathways, and similarly what appears to be a good strategy may be compromised by a hidden and irresolvable weakness that is revealed only at a certain depth. An example in chess is a piece sacrifice, where material is ceded in return for long-term gain. Strategic transforms (*vide infra*) represent a similar concept in organic synthesis.

e) It is likely to be beneficial to investigate all possible lines to a small depth initially ('breadth-first,' or in the terminology of chess programs 'iterative deepening') rather than some arbitrary branches exhaustively ('depth-first'). Subsequently, it is likely to be beneficial to investigate apparently promising branches to great depth, and to ignore completely those branches that look poor, despite the caveat in d).

Where the analogy with chess breaks down is as follows:

1) Moves in chess are binary, whereas synthetic transformations have an associated and variable yield.

2) The rules of synthesis planning are continually changing to accommodate the discovery of new chemical methodology. The rules of chess are invariant.

3) Chess is adversarial, where players are attempting to maximise their own score. This changes the nature of the searching procedure. For example, in chess the 'alpha–beta' pruning algorithm may be used for reducing the size of the search tree.[6] Similarly, the 'null-move' pruning technique involves one skipping a turn if one believes one has a good position. If the opponent, with the advantage of two sequential moves, is still unable to cause inconvenience, one's current position is judged 'good,' and does not require further search. These powerful algorithms have no direct analogue in organic synthesis.

That computers are becoming ever faster at a predictable rate is a well-known phenomenon. Is it true to say that the inexorable rise in computing power will transform what are complex problems of today into trivial problems in the future?

No. There are many problems in computer science deemed to be uncomputable, and the issue of chess falls into a class of minor severity where the problem is the combinatorial explosion. If we say that in a standard chess game, that there are on average 35 possible moves at each point, and that a game can last for 50 moves (100 'ply') then the number of possibilities that have to be examined is 35^{100}. "Even if we ignore the bookkeeping and memory space involved in a brute-force trip through all possible moves, and assume that each move can be tested in, say, a nanosecond, there is simply no way that computers can explicitly contemplate each and every possibility in any reasonable amount of time. So there is no hope for a perfect chess program. A world champion yes, but a perfect program no."[7]

Deep Blue's power relied on several crucial components.[3] At the heart of the system was a specialised chess chip, *i.e.* hardware (rather than software) that automatically generated allowed moves in any position and carried out a preliminary ranking of their worth, which greatly increasesd the rate of positional evaluation. In the 1997 rematch, Deep Blue analysed on average over 100 million positions per second, but frequently clocked twice or even three times that. The system was massively parallel, with about 500 processors.

However, we must avoid the temptation to think the success of the Deep Blue project was due to the speed of the machine. Speed is secondary to the ability to deal sensibly with the incredible growth of the tree of possible moves, and it was the corresponding advances in artificial intelligence that gave Deep Blue its real power. Crucial to the system were a complex hardware evaluation function, a heavy emphasis on search extensions on promising lines, and the ability to call upon an extensive grandmaster game database of over 700 000 games.

The combinatorial explosion is the central problem of both computer chess design and computer-aided organic synthesis. Chess is perhaps the classic example of the need for heuristics in computer science where exhaustive analysis is not possible. Computer-aided organic synthesis has had to employ similar techniques.

The advances in artificial intelligence employed by Deep Blue are desirable in any automated retrosynthetic analysis, where we would want a sound evaluation function of

strategies, a deep analysis of promising-looking lines, and knowledge of previously successful retrosynthetic strategies from the literature. A feature that was missing from Deep Blue was the ability to use its game database to reason about similar positions it may encounter, whereas humans are particularly adept at this. This absence is testament to the difficulty of defining similarity when there are so many other variables (pieces and positions) on the board. Deep Blue did, however, have a knowledge of endgame positions where for example strategies are known for any chess position with five or fewer pieces on the board. Such a method would be attractive for synthetic design, for example, where a semi-synthetic database is stored detailing routes from commercially-available starting materials to common small molecules that may arise towards the end, but not formally at the end, of a retrosynthetic analysis.

Expecting computers to become creative simply by making them faster is unreasonable if we are not first creative in the way we program them.

Fundamentals of computer-aided organic synthesis (CAOS)

There has been a fair amount of hostility to the notion of computer-aided synthesis planning over the years, owing perhaps to a certain sense of pride in the human ability to perceive and exploit the art in the process. Whether this art is solely the domain of human ability, and whether computers can produce beautiful synthetic routes to organic molecules, is a question of much interest.

We shall examine some of the major contributions to the field since its origin in the late 1960's. Retrosynthetic approaches will be examined first, followed by forward searching methods and finally the combination of both, which has long been the overall goal. Finally, the recent application of new computer science methodology to automated synthesis design will be described. A variety of techniques used to prune the synthetic tree will be seen.

The following discussion cannot be a comprehensive survey of the field, which encompasses organic chemistry, computer science and information technology. The reader is referred to several reviews that include some comprehensive listings of computer-aided synthesis programs.[8,9] The use of computers in ligand design (discovery of structures that dock biological receptors) in the pharmaceutical field,[10] and the various databases of chemical information for reaction retrieval[11,12] will not be covered in any detail.

A summary of the programs discussed here, and their interrelationships, is shown in Fig. 2.

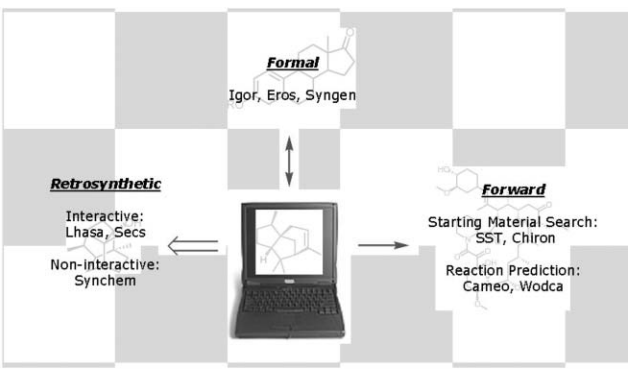


Fig. 2 Outline of several CAOS programs covered in this review.

The logic of chemical synthesis

The first major study in the field of computer-aided organic synthesis was E. J. Corey's program at Harvard. Corey realised that for a computer to become proficient in synthesis planning would require the formalisation of the rules of synthesis, which would place demands on our understanding of organic chemistry at its most basic level. As he put it in 1967: "…any technique for the automatic generation of synthetic schemes by a computer will require a complete and detailed definition of the elements of Synthesis and their mutual interaction, in a most general sense."[13]

This seminal paper defines words such as 'retrosynthesis,' 'disconnection' (hypothetical reverse of a synthetic step) and 'synthon' (hypothetical fragment of a molecule associated with a synthetic operation) that are now familiar to all those in the field. Indeed Corey describes what has become the canon of every undergraduate synthetic chemistry course on organic synthesis. Thus certain axioms are initially stated, for example that the endpoint of any synthetic analysis be a readily-available substance, and that we may judge, or score, the various possibilities according to the likely success of the individual reactions in the forward direction. Corey then describes what is required for the simplification of a molecule, for example the recognition of molecular symmetry or the perception of certain functional group or stereochemical relationships. Corey also notes that there are different types of retrosynthetic steps, in that small functional group modifications may be required to reveal significant strategic disconnections that were not initially obvious.

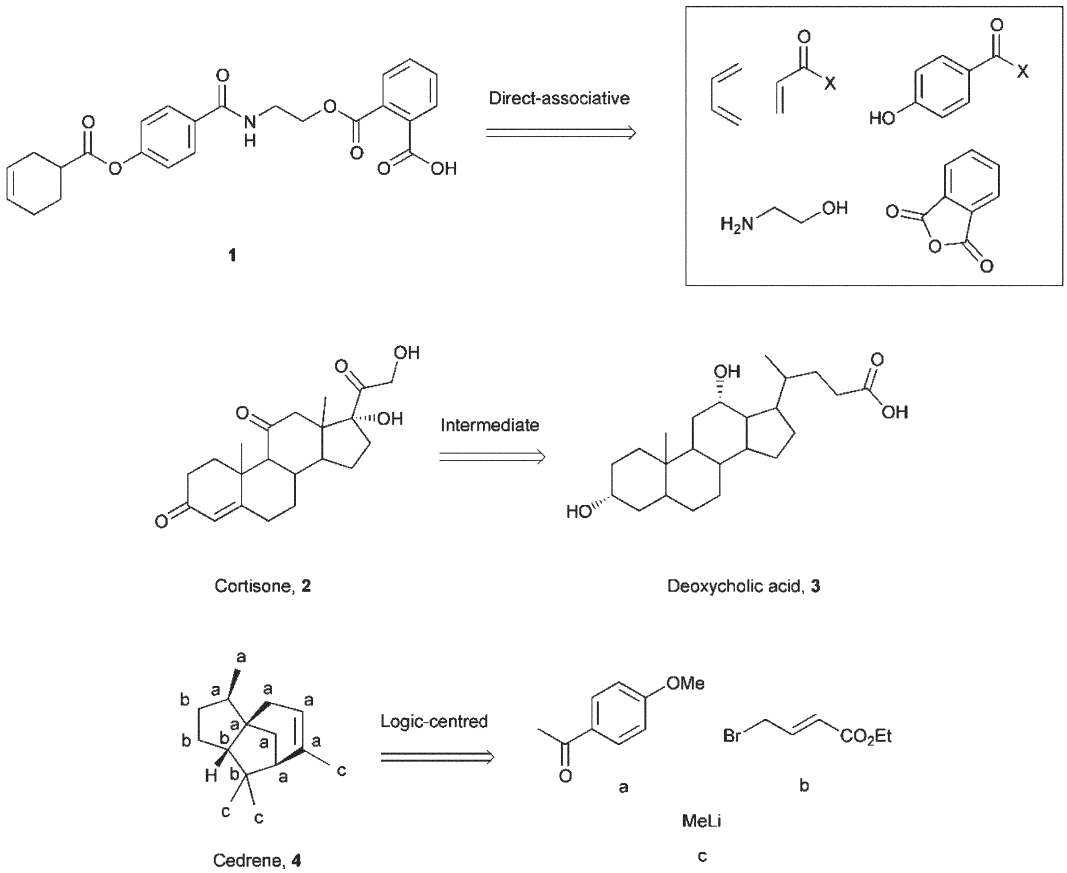
Synthetic analysis is classified into three approaches:[14]

a) *Direct associative*, where the synthetic target (*e.g.* **1**, Scheme 1) is a simple collection of 'undisguised' subunits, and where a minimal and uncontroversial analysis reveals the required starting materials.

b) *Intermediate*, where a complex synthetic target bears a close resemblance to another, but synthetically accessible, molecule, and the problem becomes finding the appropriate sequence of reactions for their interconversion. The example Corey gave was a synthesis of cortisone (**2**), which was constructed from deoxycholic acid (**3**).†

c) *Logic-centred*, where a logical analysis generates a synthetic tree without any assumptions as to the starting materials required. For example, cedrene (**4**) was synthesised by Corey from the three starting materials shown (letters refer to the carbon atoms these reagents contribute to the target.) These starting materials, and indeed the overall strategy, are not obviously suggested by the target. This approach is intellectually the most interesting.

Overall it was envisaged that a computer could perform the 'logic-centred' portion of an analysis, to which a chemist contributes creatively *via* an inherently human 'information-centred' approach. Corey's first realisation of these principles



Scheme 1 Three classes of synthetic analysis.

was in the program that arose from them, Logic and Heuristics Applied to Synthetic Analysis (LHASA).

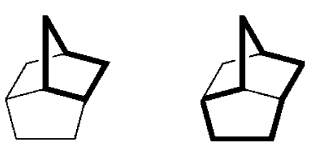
LHASA

One of the first challenges encountered at the outset was something that we now take for granted—the graphical representation of organic chemistry by a computer. While any chemist may now easily communicate chemical structures and reactions to a computer with the aid of a mouse and a commercial drawing package, no such system was in place in the late 1960's. Input to Corey's system was achieved with a 'Rand tablet' and a pen, which produced chemical structures on the 'scopes' from which a 'plotter' could produce hard copies.[14,15] The program was capable of manipulating charges, radicals and stereogenic centres. The chemical drawing system that was developed as part of this work evolved into what is now known as ChemDraw.[16]

The computer had to be instructed in how to interpret and manipulate chemical structures efficiently. The inappropriate human representation was therefore translated into a form intelligible to the software, in which the atoms (other than hydrogen) and bonds (explicitly) were arranged in a 'connection table' that initially had a maximum capacity of 36 atoms. The table also included stereochemical information.

This table was used for a 'perception' of the nature of the molecule, in which higher-level concepts such as functional groups, rings, symmetry, aromaticity and simple electronic properties were noted.[17] Functional groups were arranged into classes to reflect similarities in reactivity (*e.g.* nitro, cyano and carbonyl groups would be classed together to reflect their shared acidity of the synthetically useful α-proton). The reactivity of various functional groups was understood, including effects of the immediate molecular environment. Relationships *between* functional groups, rings and so on were then perceived. Whilst this is simple for a human chemist, programming such perception for a computer is a non-trivial problem. In the case of the perception of rings in bridged cyclohexanes for example, it is important to distinguish 'real' rings such as that shown in bold in **5** (Fig. 3), from 'pseudo' rings such as that indicated in **6**.

Once the computer had an understanding of the molecule under consideration, then the results of the perception analysis were passed to the 'strategy and control module,' which contained a set of fundamental heuristics. This is the crucial part of the program, since these heuristics are essentially our



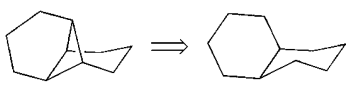
5 **6**

Fig. 3 Real *versus* pseudo rings.

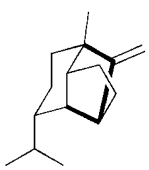
rules of thumb as to the productive or most general ways in which we might analyse a molecule retrosynthetically. For example, the program understood that it is a sound idea to remove reactive functional groups first, or to cleave carbon–heteroatom bonds.

Higher-level strategies were then considered. For example, the evaluation of which are the strategic rings[18] or strategic bonds[19] to disconnect in a molecule. Again, heuristics as to what constitutes a strategic bond needed to be clearly formulated. For example, strategic bonds tend to be *endo* to 5-, 6-, or 7- membered rings, and *endo* to those rings exhibiting the most bridging. In the polycyclic structure **7** (Scheme 2), it is clear that one of the bonds in the central four-membered ring should be disconnected back to the decalin structure **8**, and in sativene (**9**) the strategic bonds are indicated in bold. Special consideration for heterocyclic structures had to be programmed. While LHASA's choice of strategic bonds could be overridden by the user, the program performed well when tested against a range of natural products, where it correctly identified strategic bond disconnections actually used. Interestingly, the rules for identifying strategic bonds (originally based on heuristics) have been found to correspond well with more recent graph theoretical descriptions of molecular complexity.[20]

The manipulation module then carried out operations on the connection table that correspond to retrosynthetic transformations, thereby generating the next layer of the tree. This structure generation is chemically 'sensible.' For example the retrosynthetic cleavage of a C–X bond will initially generate a carbocation in the secondary target, but the program is able to generate the neutral species implied by this *via* elimination or association, a trivial process for a human analyst. For LHASA to be able to perform a range of different transforms clearly required the conversion of a great deal of chemical information into a machine-readable knowledge base.[21] A great many basic synthetic concepts were written for the program in a new 'chemical English' language CHMTRN, which was intended to be simple for a chemist (rather than a computer scientist) to add to, if the chemist wanted new methodology to be known to the computer. LHASA was capable of carrying out simple functional group interconversions (FGI's), functional group additions (FGA's, strategically important as enablers of initially invalid retrosynthetic routes, *vide infra*) and 'two-group transforms' such as the aldol reaction. Each transform is represented by an individual 'transform table' and these are called as a subroutine in response to the program's understanding of the molecule in question.



7 **8**



9

Scheme 2 Identification of strategic bonds.

Extensive use was made of qualifiers, giving a quantitative assessment of the ease of a transform in its molecular context. Individual steps that were judged as good in most situations could be judged as unsatisfactory (with a numerical penalty) if there were features of the molecule that prevented the step being effective, for example the danger of elimination during a desired nucleophilic substitution reaction. A cumulative effect of all these qualifiers had the effect of ranking the routes analysed, and those that ranked highest were displayed first. The success of such a ranking relies heavily on the heuristics contained in the transform tables, and the quality of the information generated during the perception phase.

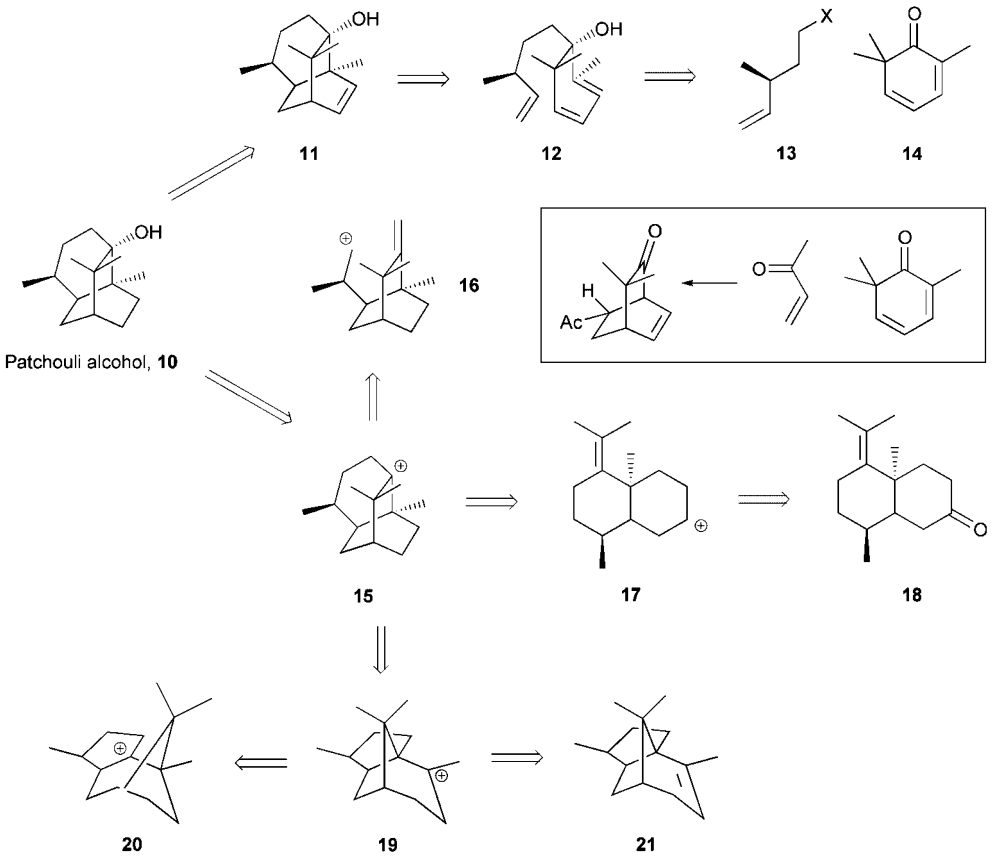
Once LHASA had generated a possible retrosynthesis, the overall path was evaluated. The idea was for a chemist to do this, but LHASA was programmed to assist. For example, the evaluation module checks for the uniqueness of structures in the synthesis tree. Common intermediates are promoted in importance, since these imply methods of getting round a practical problem with one route, in effect giving the chemist more 'outs' of a synthetic tangle.

The early version of LHASA completed a retrosynthetic analysis of patchouli alcohol (**10**, Scheme 3). Many more routes than those shown were generated by this analysis (though it is not clear quite how many were unsuitable), and some of these are reproduced here. It should be noted that extremely plausible routes arise. For example, the sequence **10** to **14** is both direct and novel, but related to a published synthesis of patchouli alcohol by Danishefsky (a key step of which is boxed). Interesting carbocationic rearrangements *via* **15** were also suggested.

The sophistication of the analysis depends on the number and quality of the heuristics known to the program. Herein lies a problem. For a sophisticated analysis, one would wish to include a large number of heuristics, in the same way as a chemist with a good knowledge of synthesis will be more adept at analysing a molecule than a novice. Indeed, Corey identified this very much as an ongoing issue, in that future development of the program required the cumulative addition of new synthetic strategies. However, this increasing sophistication comes at the cost of an increasing branching of the synthetic tree, and the corresponding combinatorial explosion in the size of the tree that may be generated.

Enhancements to LHASA

1) Enhancements to knowledge base. LHASA's knowledge base was considerably expanded several years after the original reports.[22] For example, the program was instructed in the use of the reconnective transform strategy, where acyclic or medium-ring structures would give small-ring structures in the retrosynthetic direction. LHASA could then successfully reconnect a 1,6-dicarbonyl compound to a cycloalkene. In addition, LHASA was instructed in strategically important topological strategies,[19] and recent developments in stereoselective synthesis.[23] By 1994 the number of reactions known to LHASA stood at 2100.[24] A teaching version of LHASA was



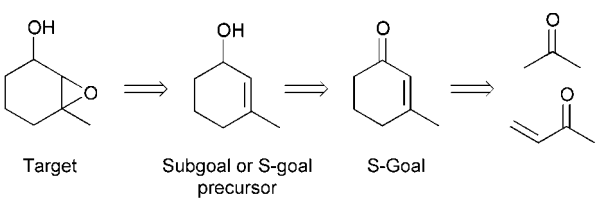
Scheme 3 LHASA's retrosynthetic analysis of patchouli alcohol.

developed, where the original set of 60 reagents known to the program was expanded to 138.[25]

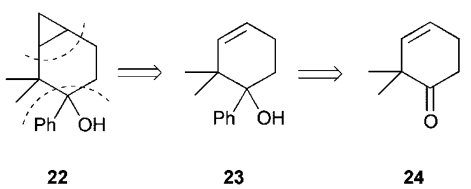
2) Deep searching and the combinatorial explosion. A retrosynthetic pathway may contain one or two particularly simplifying transforms—masterstroke disconnections that are the backbone of the strategy. It may be inappropriate to apply such transforms to the actual target molecule, and small, non-simplifying manipulations may be required to prepare the target for the key transform. For example, if we wish to employ the Robinson annulation as a simplifying transform (Scheme 4), we need to generate, retrosynthetically, an enone in the target molecule.

To map sections of the target molecule to a desired transform precursor, in other words to determine what sequence of FGI/FGA's is necessary to convert one to the other, required the introduction of the 'Localised Matching Unit' (LMU).[26] Small units of the target were considered in turn, and the sequence of reactions needed to generate the required structural unit from this could then be evaluated. Were the target of interest to be the bicyclic alcohol **22** (Scheme 5), and were the Robinson annulation to be employed, then the structure may be reasonably analysed by a two-carbon LMU whereby the cyclopropyl group is disconnected back to a double bond to give **23**, and a one-carbon LMU where the alcoholic centre is disconnected back to a ketone moiety, to generate **24**. Each LMU, or rather the chemical transformation that such a route implies, could be scored in terms of the likely success of the transformation in the laboratory.

LHASA was furnished with a 'Prior Procedure Evaluation' function whose job it was to rule out certain lines of analysis that were clearly going to be unprofitable without further analysis, essentially an initial screen for synthetic problems.[26,27] The LMU's are given a numerical score corresponding to the number of steps required to effect the transformation, along with a general assessment of the utility of isolated transforms. Thus instead of predicting yields for



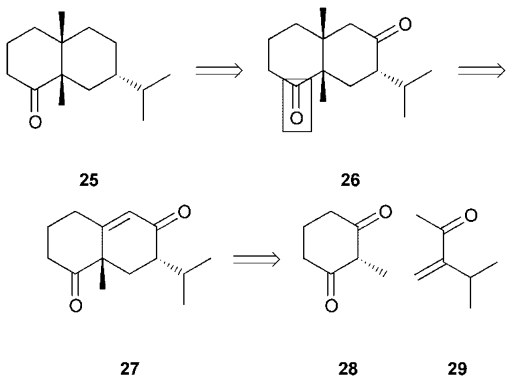
Scheme 4 Use of Robinson annulation as a transform goal.



Scheme 5 LHASA's use of the Localised Matching Unit (LMU).

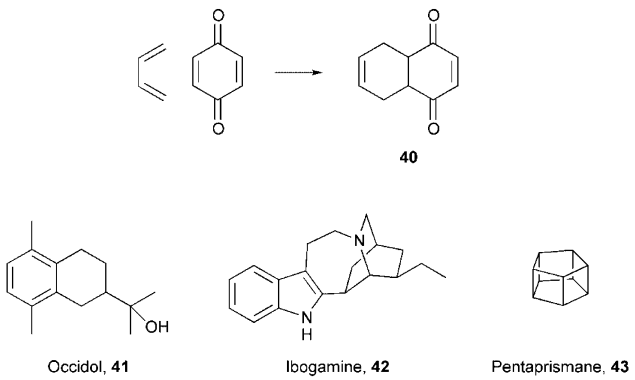
specific steps (which becomes complex in densely functionalised targets), routes were quickly assessed on their length and the general likelihood of the generic transformations succeeding. The route under consideration can then be given a rating by summing the various LMU contributions, and once this value crosses a certain difficulty Rubicon, that route is no longer analysed. This is a good example of how heavily LHASA employs heuristics to limit the combinatorial explosion. In the case of the Robinson annulation analysis for **25** (Scheme 6),[26] for example, thirty Robinson annulation routes were identified, and one of the best-ranked, which is clearly synthetically reasonable, is shown. A trivial functional group interference was noted by the program (one of the carbonyls appears boxed).

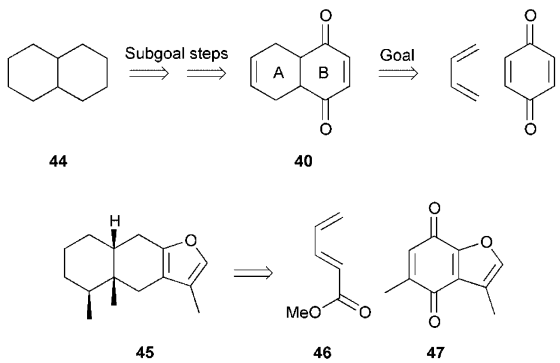
To search for a simplifying transform, the user first defined the transform or 'global strategy' of interest, for example the Diels–Alder reaction (Scheme 7).[28] The impressive length of this sequence should be noted: LHASA could search to a depth of 15 steps. Much of this depth was caused by the requirement of converting the amino group present in **30** to a methoxy group that is more relevant to the application of the Diels–Alder strategy. Of note is the step from **35** to **36** where an electron-withdrawing ketone is introduced to activate what will become the dienophile portion of the structure.



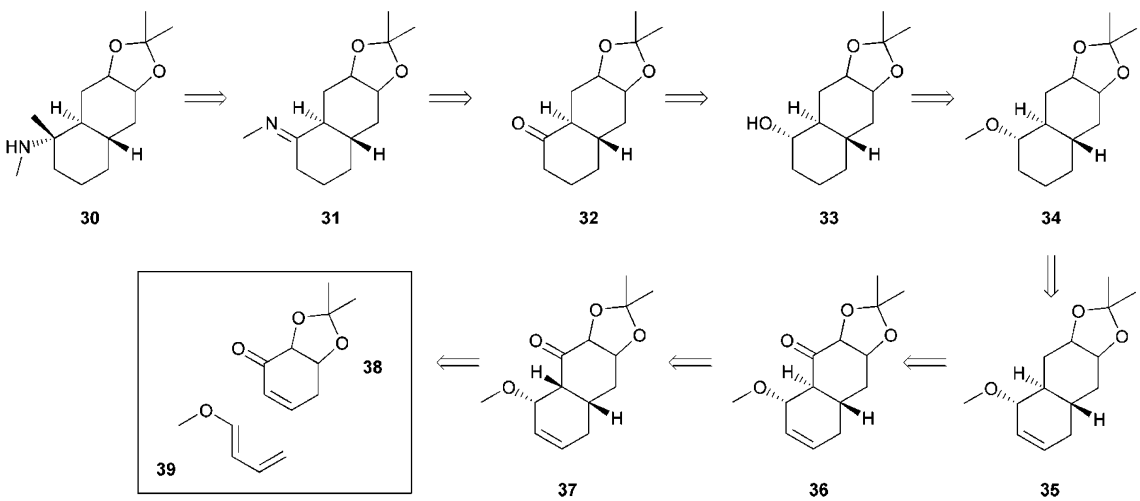
Scheme 6 LHASA's strategic use of the Robinson annulation.

More recently, the quinone Diels–Alder transform was also incorporated as a long-range strategy in LHASA.[29] The implementation of this is interesting. The quinone Diels–Alder reaction is simply as shown in Scheme 8 (butadiene and benzoquinone generate **40**). However, the *cis*-decalin that is produced may be subject to a wide range of further transformations in the course of a total synthesis, such that the original decalin ring system is unrecognisable in the target. Some of the natural products synthesised with the aid of the quinone Diels–Alder reaction are shown (**41–43**). LHASA is unsurprisingly not capable of spotting the possibility of the





Scheme 8 Quinone Diels–Alder transform as a long-range strategy.



Scheme 7 LHASA's long-range use of the Diels–Alder transform.

quinone Diels–Alder reaction as part of the retrosynthesis of molecules as far removed from the *cis*-decalin ring system as this, which is an excellent illustration of the number of retrosynthetic possibilities available to us, and the corresponding difficulty of designing a program to spot such deeply embedded strategies.

The use of this strategic disconnection therefore had to begin with the limitation that LHASA would be instructed to look only for the decalin ring system **44** in a target, and perform a long-range search on the transformation of such a target into the appropriate substructure goal (the formal quinone Diels–Alder retron **40**). LHASA used this strategy to devise a retrosynthetic analysis for furanoeremophilane (**45**).

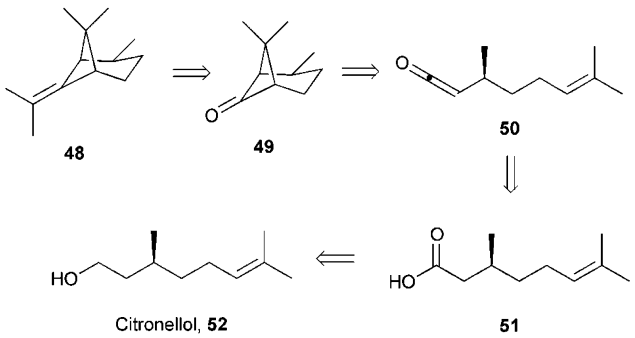
As of 1985,[23] LHASA was still not in a position to select the strategy of the retrosynthetic analysis, which was left to the user. Of course not only would the program (ideally) identify a good synthetic strategy, but would combine various strategies as part of an overall approach, making the combinatorial possibilities still greater. The latest version of the program is able to make suggestions as to the global strategies appropriate, based on perception of basic features in the target, and to employ tactical combinations of retrosynthetic transforms to widen the available strategies.[30]

3) Protecting groups. LHASA was given information on the reactivity of various functional groups to a range of representative reagents in order for it to be able to identify potential interfering functionality in any synthetic route it suggested and to suggest a suitable protecting group.[31] Higher level strategies such as the timing of the protection in the synthetic sequence, or the desirability of using a single functional group for the protection of more than one functional group, were still not possible. The enormous amount of work that went into the organisation of knowledge for protecting group strategies developed into a well-known textbook on the subject.[32]

4) Starting material-oriented searches. Also developed was a starting-material oriented search where a starting material is specified, and the task is then the long-range generation of a retrosynthetic route that includes it.[33] The starting material influences the retrosynthesis, and this approach is therefore different from other starting-material strategies (*vide infra*). LHASA was able to evaluate where in the target the starting material matched best. The program understood the quantitative idea of synthetic proximity in terms of the chemical distance between two structures (*e.g.* the number of atoms or bonds needing to be changed) and factors such as synthetic ease of the routes. The program was able to suggest citronellol as a non-obvious starting material for the bridged compound **48** (Scheme 9), and then formulate a retrosynthesis.[34]

Other retrosynthetic analysis approaches

LHASA has been discussed in detail above for two reasons. First, the development of the program is intimately bound to the development of so much fundamental understanding of the process of retrosynthetic analysis by one of the masters of the discipline. Secondly, LHASA is the CAOS program with



Scheme 9 Examples of LHASA retrosynthesis using the starting-material oriented strategies.

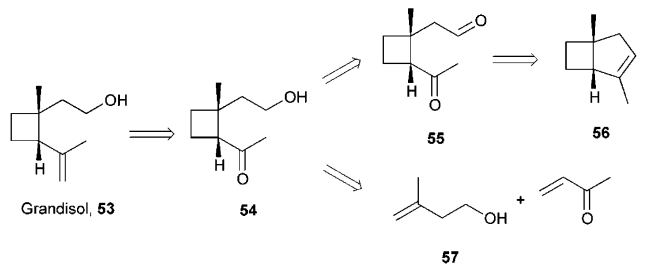
the most extensive published history, and with many seminal contributions in areas such as the representation of molecules by a computer and heuristics for the limitation of the size of the synthetic tree. LHASA acts as a representative example of many common features between synthesis programs. The examples presented above of LHASA's capabilities show the power of the program in retrosynthetic analysis, particularly when it is remembered that much of this work was carried out so early in the history of computer science. Despite these successes, there have been no published papers on LHASA since 1997, and the program has been taken in new directions, such as the development of tools for the prediction of chemical toxicity and metabolic fate.[35]

It is important to stress, however, the number of other important contributions to computer-aided organic synthesis. We shall now turn to these, dealing firstly with other retrosynthetic programs including those of the non-interactive type, before turning to forward synthesis approaches.

SECS

SECS (Simulation and Evaluation of Chemical Synthesis), developed by W. T. Wipke, arose as an offshoot of LHASA.[36] SECS worked in a similar fashion, with graphical input, connection table and perception modules all operating essentially as described above. SECS was instructed in the perception of stereogenic centres and double bond geometry, as well as using this information in retrosynthetic analysis from an early stage in the program's development, and this was the main focus of Wipke's extensive effort.[37] Crucially the program was able to use stereochemical features of a target molecule to screen for the suitability of candidate transforms, for example in the requirement for inversion in an S_N2 reaction. Further, the program understood the basic shape of the molecule under consideration in three dimensions, and could perform a basic energy minimisation. This allowed it to calculate whether, for example, any reactive site would be subject to unusual steric hindrance.

SECS was used in the retrosynthetic analysis of the insect pheromone grandisol (**53**, Scheme 10) the synthesis of which had been extensively studied in the literature.[36] After a human-guided search, the synthesis tree contained 300 structures, and eight of the twelve published synthesis routes were discovered, of which two are shown. Those literature routes not found involved intermediates that were considerably more complex



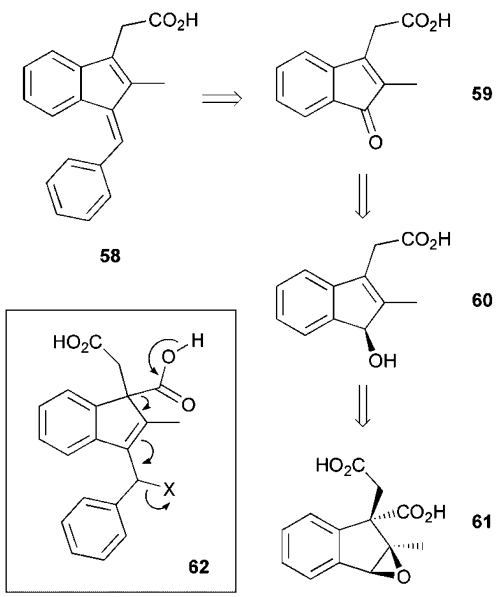
Scheme 10 SECS retrosynthetic analysis of grandisol.

than the target, and SECS did not have a strategy to allow for this.

SECS was extensively tested in an industrial setting, with the aim of generating a routinely-useful synthesis aid.[38] Compound **58** (Scheme 11) was analysed since there was a wealth of in-house experience of this target, and of the routes SECS generated, one (shown) proceeded *via* an unusual fragmentation pathway. This route had not been considered by the chemists working on the synthesis, and possible uses of such a method (such as that from **62** directly to the target molecule) originated from the subsequent discussion of this process. While these proposals did not lead to the strategy being adopted, this illustrates the way a computer can *aid* the chemist in generating novel ideas. The main conclusion of the lengthy evaluation of SECS was that the knowledge base needed considerable improvement, both in the sophistication of the qualifiers attached to existing transforms, and in a basic expansion of its reaction database.

Non-interactive programs

Interactive programs such as LHASA use human perception to reduce the need for severe tree-pruning heuristics in the program, and there is a clear advantage in this approach. However, might not a non-interactive program allow us to generate synthetic routes that are independent of the bias of the operator, and even routes that predict new chemistry? This is, after all, what often happens in the course of a total synthesis, where the invention of new methodology allows us to achieve what was thought to be a problematic route.



Scheme 11 Novel fragmentation pathway suggestion by SECS.

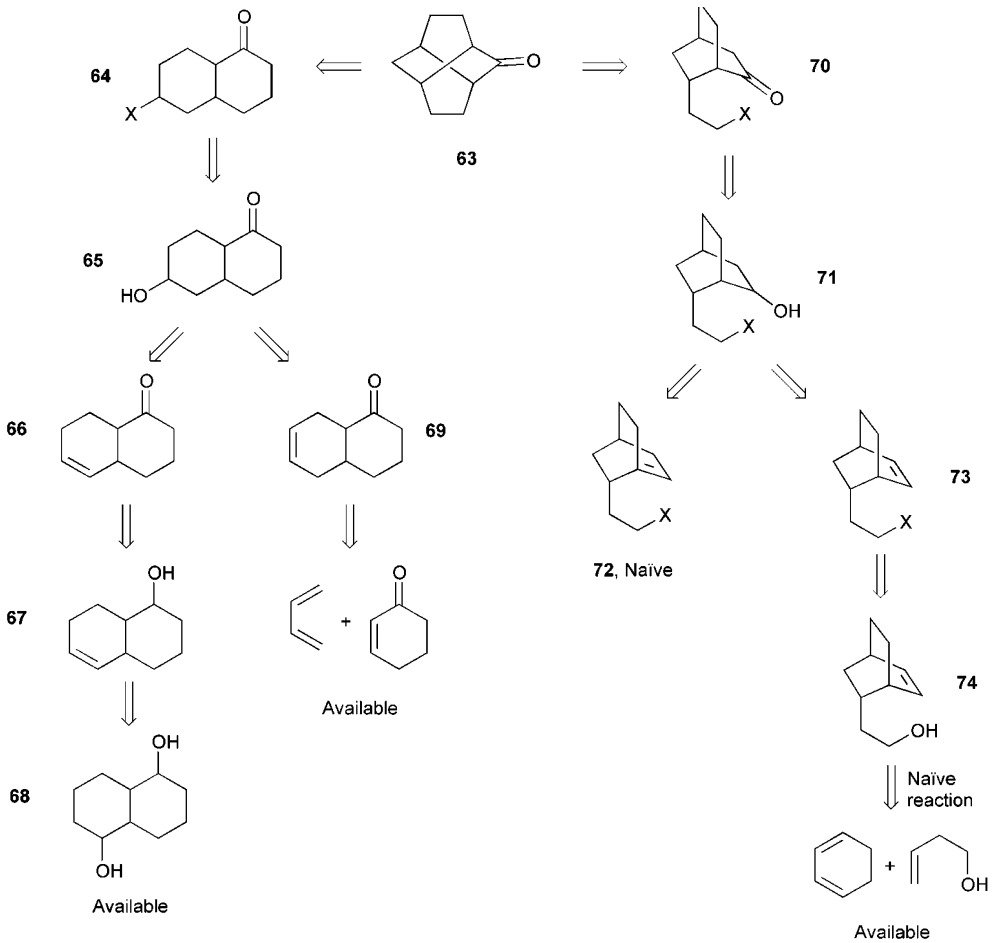
SYNCHEM

Gelernter developed the non-interactive, retrosynthetic program SYNCHEM in an effort to apply some new concepts in artificial intelligence and heuristic computer programming to organic synthesis route discovery.[39] The general outline of function resembled LHASA. SYNCHEM accepted the target molecule in a variety of representational formats, and then perceived the various functional groups and other relevant features of the molecular environment. Strategies, heuristics and qualifiers for the synthesis of each feature were stored in the program's knowledge base, which also contained a starting material library of 3000 compounds (originally available from Aldrich as 'punched cards'). The program generated the retrosynthetic tree, scoring each intermediate as it went with a merit function, followed by composite scores for overall routes based on estimates of reaction merit and yields. Such scores influenced the direction the tree was grown, but tree growth stopped when a link was made with an available starting material or a fatal flaw in the pathway emerged.

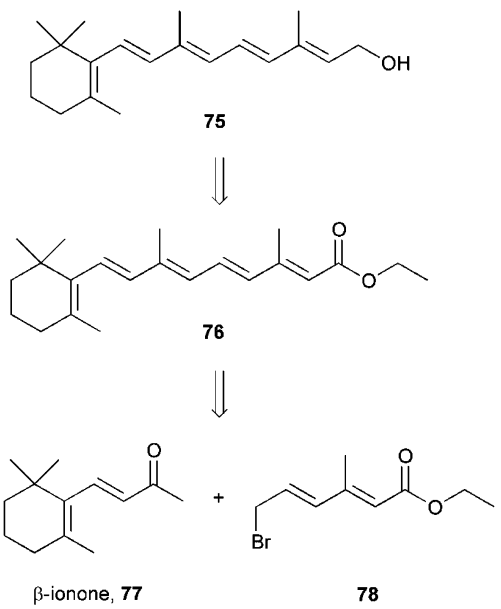
An early version of the program was able to suggest reasonable and precedented routes to twistanone (**63**, Scheme 12). There are clearly technical faults in some of the suggestions, such as the Bredt's rule violation in **72** and a suggested Diels–Alder reaction between cyclohexadiene and an unactivated dienophile in the generation of **74**, but these mistakes were a consequence of the small size of the program's knowledge base. There was also some human intervention in the generation of these routes, in that the program explored several very barren-looking paths and had to be stopped.

An example of SYNCHEM's shortcomings is the undue emphasis the program placed on the discovery of β-ionone (**77**) in its retrosynthesis of Vitamin A (**75**, Scheme 13). SYNCHEM perceived this starting material early in the retrosynthesis, and then rated the route highly, necessitating a poor coupling partner **78**. In other words, SYNCHEM removed this starting material first, and then constructed **78** in a convergent fashion. A chemist analysing this target would probably disconnect *to* β-ionone gradually, in the knowledge that the target could be built up in a repetitive stepwise fashion. SYNCHEM's error was a result of the way it generated the retrosynthetic intermediates. At each level of the synthetic tree, the intermediate molecule under consideration is treated as the target, and no global history of the route is included, making strategies somewhat short-termist.

This and other limitations led to SYNCHEM's burial, and the creation of a successor containing many modifications and improvements, SYNCHEM2.[40] This program described chemical transforms by patterns, rather than the comparison of individual features and allowed for a consideration of stereochemistry. SYNCHEM was also upgraded to operate in a multiprocessor format, *vide infra*.[41]



Scheme 12 Retrosynthetic analysis of twistanone by SYNCHEM.



Scheme 13 Retrosynthetic analysis of Vitamin A by SYNCHEM revealing a weakness in the approach.

Formal approaches

The CAOS approaches described above use empirically-derived heuristics to limit the combinatorial explosion of the synthesis tree. Conversely, formal approaches to synthetic design are based upon abstract definitions of possible chemical reactions. Such programs do not have a knowledge base of known, or good, reactions because they can draw upon all theoretically possible transformations, including those that have no known experimental counterpart. Abandoning a synthetic knowledge base means a return to the problem of the combinatorial explosion of possibilities. How is it possible to limit this explosion if an organic chemist's sense of restraint is abandoned? On the other hand we enter a realm where the prediction of new reactions is possible, where the generation of the search tree is unguided by the bias of the user, and where a possible route will never be missed simply because with a formal approach we are able to generate them all.

For a computer to predict all possible reactions in a non-empirical manner that a given molecule may undergo requires a formal description of transformations. In addition, for synthetic progress to be assessed, a quantitative assessment of molecular complexity is useful. These are large and on-going fields of activity. With regard to the former, Hendrickson has compared several of the systems developed for the

formalisation of organic transformations.[42] With regards the latter, Bertz was the first to describe a 'general index of molecular complexity.'[43] Naturally this also means that the *change* in molecular complexity as a result of a chemical reaction may also be calculated.[44] More recently, Bertz has exploited graph theory to devise a conceptually simple measure of strategic bond disconnections.[45] The graphs used are skeletal simplifications of a molecule, where, for example, methane is a point, ethane is two points connected by a line, *etc*. One then performs a disconnection to generate candidate retrosynthetic intermediates, and for each one calculates the number of different subgraphs as well as the number of different types of subgraphs, rather like the children's puzzle of working out the number of squares or rectangles present in a grid. It seems as though these two simple considerations (essentially topological complexity) give a good indication of the level of simplification achieved by a disconnection.

The question of how one measures a synthetic route's efficiency, rather than judging it subjectively beautiful or inefficient, is still receiving detailed attention.[46,47]

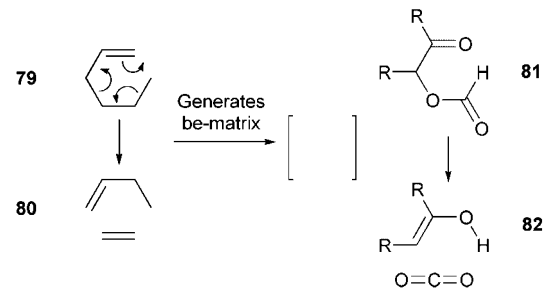
IGOR

Ugi developed the computer-aided synthesis program IGOR (Interactive Generation of Organic Reactions) through a formal treatment of reactions.[48] One of the key tenets of the so-called Dugundji–Ugi (DU) model[49] is that reactions of ensembles of molecules (EM's) may be treated as isomerism. In this model, a molecule or EM of n atoms is represented by a symmetric n by n 'BE-' matrix ('bond–electron' matrix). Diagonal entries give the number of free valence electrons, off-diagonals give the bond orders between atoms. Chemical reactions are represented by '***R***' matrices, which when added to the matrix representing the starting material EM matrix, generates a product matrix ***E*** giving the structure of the product. An example is shown in Scheme 14, for the addition of a cyanide ion to formaldehyde to generate the cyanohydrin anion. Since the matrices are commutative, synthetic reactions and retroreactions may be analysed with similar ease.

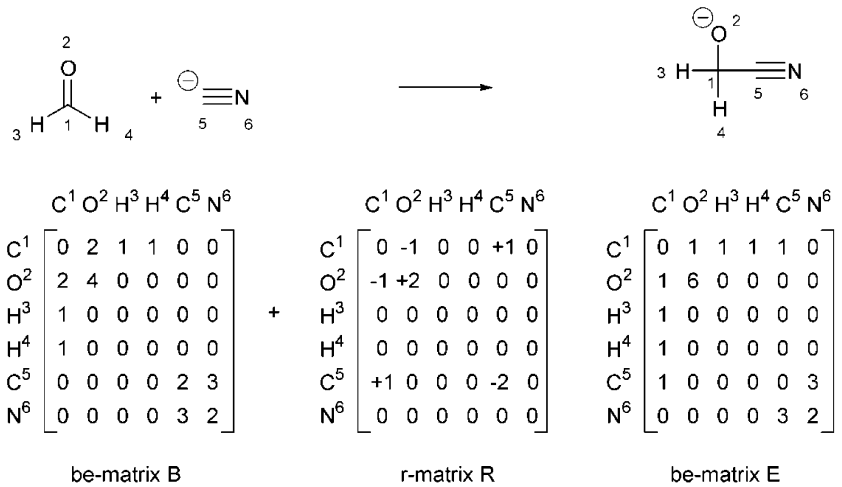
The conversion of a reaction into such a mathematical formulism allows for the quantitative description of synthetic progress as 'chemical distance.' If it is imagined that the n^2 entries of the BE-matrices are coordinates of points in an n^2-dimensional space, then the EMs are nodes connected by reaction 'vectors', and the chemical distance between any two EMs may be easily given a precise value. The 'Principle of Minimum Chemical Distance' of the DU model states that reactions will preferentially proceed by redistribution of the minimum number of valence electrons. Synthetic routes may therefore easily be generated by navigation of this space between target and starting materials.

The approach not only provides a convenient system for reaction classification, but allows for reaction prediction. The IGOR program essentially uses the electron-pushing patterns represented by ***R***-matrices to 'invent' chemical reactions, and its sister program RAIN explores reaction networks implied by the BE-matrices for the purposes of looking at plausible reaction mechanisms, and thereby acts as a 'reaction generator.' Reactions suggested can be screened for basic chemical reasonableness, for example on the basis that atoms must be in allowed valence states. The pyrolysis of α-formyloxy ketones (**81** to **82**, Scheme 15) was discovered by application of the known BE- matrix for the general class of reactions **79** to **80**. This (at the time unknown) reaction was later confirmed empirically in the laboratory.

A problem associated with the matrix approach is that all of the atoms involved in a transformation need to be accounted for (including such byproducts as water, or sodium chloride). In the retrosynthetic direction this is particularly problematic,



Scheme 15 Discovery of unprecedented reactions by IGOR.



Scheme 14 Reaction description according to the DU model.

since the reaction byproducts will depend on the disconnection chosen. A separate program, STOECH, was created to generate automatically all the species implied by a certain transform.[50]

The advantage of this approach therefore is that all possible synthesis routes can be investigated exhaustively. The synthesis tree may be pruned by exploitation of the principle of minimum chemical distance, in the sense that the most efficient routes (shortest vector paths) are rated highest. However, the enormous number of possible combinations of such matrices with any given EM still generates a serious combinatorial explosion were such a system applied to the design of the synthesis of a molecule of even moderate complexity. For a chemist to select between these would not only be impractical, but would also introduce a danger that novel reactions suggested by the analysis would be discarded by the chemist, thereby negating one of the attractive features of the approach. This formal approach should be regarded as a method of exploring the synthetic space, rather than isolating a reliable route. Pruning of the comprehensive synthetic tree generated by a formal approach requires the introduction of heuristics, in a semi-formal method, and this was realised in the EROS program.

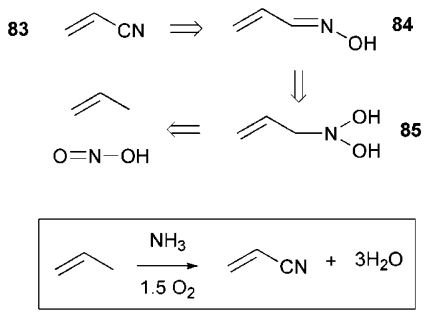
EROS

Gasteiger designed the program EROS (Elaboration of Reactions for Organic Synthesis) to apply a set of heuristics to limit the combinatorial explosion of the DU-model synthetic tree.[51] Reaction sites in molecules were defined by breakable bonds, which were taken to be multiple bonds, bonds to heteroatoms, or those bonds proximal to these. The size of the reaction generator is thus reduced in comparison to IGOR. The most significant input of heuristics came from physical chemistry, however, whereby reaction enthalpies were calculated from the relevant bond enthalpies for the transformation. This allowed the rejection of reaction products in the synthetic tree that are either unlikely or would be so unreactive as to be useless for subsequent reaction.

EROS developed a synthesis tree retrosynthetically in a non-interactive fashion, using these heuristics as guidelines. Bonds for disconnection (and consequently the ***R***-matrices to be employed) are rated quantitatively and examined best-first, a choice the user may override. Prior to an analysis, the user may further prune the search tree by specifying its gross dimensions, such as the number of levels to be generated. The search of a given pathway proceeded until a suitable starting material was encountered. Simple molecules may easily be analysed with this program. One of the routes devised for the synthesis of acrylonitrile (**83**, Scheme 16), where propene and nitrous acid are suggested as starting materials, looks unlikely at first sight. The ammoxidation of propene (boxed) is, however, an industrial route to this molecule.

SYNGEN

Hendrickson developed a novel description of organic reactions based not on matrices but on a symbolic representation of bond formation and cleavage (which will not be described in detail here).[52] Again, Hendrickson's approach is a formal one,

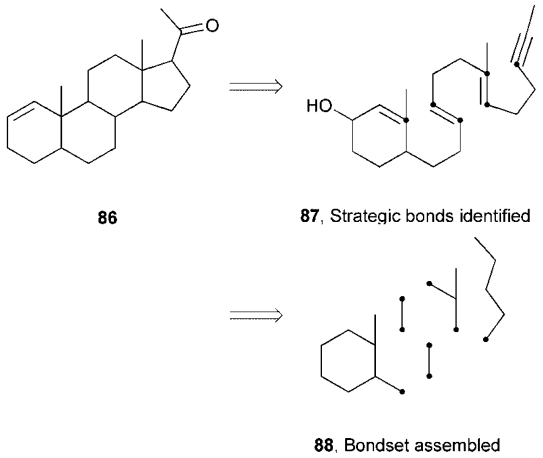


Scheme 16 EROS discovery of industrial route to acrylonitrile.

so we leave aside mechanistic considerations and return to thinking of reactions in terms of net structural change. As with Ugi's system, such a formal description of reactions is exhaustive, and allows for both reaction classification, and by extension, reaction prediction.

Hendrickson developed the term 'ideal synthesis' for the shortest and most atom economical route from starting materials to target, one where the construction proceeds with no intermediate functionalisation. If all the required functionality is present at the outset, such that the molecule that results from one step is ready for the subsequent transformation (a 'self-consistent sequence') then the synthesis tree is drastically pruned. An example of overlap is where an aldol reaction generates an α,β-unsaturated ketone, which may then be used in a conjugate addition reaction. The imposition of this kind of constraint on the generation of self-consistent sequences greatly reduces the number of routes considered, sometimes with the rejection of over 99% of possible two- or three-step sequences. The problem then of course becomes identifying the starting materials that make this possible, and navigating the synthesis space to find the shortest route between them.

The emphasis on the synthesis is one of skeletal construction, with the introduction of functionality and stereochemistry being treated as secondary to the obligatory reactions involved in forming the backbone of a molecule. One may subdivide the synthesis tree by 'bondsets,' the bonds that are formed in a given route. This implies at the outset the starting material skeleta required (without defining exactly how they are put together), and thereby reduces the number of synthetic options. An example of this is shown for the Johnson approach to the synthesis of the steroid skeleton (**86**, Scheme 17). The designation of strategic bonds according to Johnson's polyene cascade is followed by the selection of a bondset likely to give an efficient construction of the implied precursor. The detailed design of the synthetic route may then begin on this basis, with a pruned synthetic tree. The selection of the bondset requires a set of heuristics. Synthons are said to be capable of undergoing half-reactions. The oxidation state change at any atom gives an assignment of polarity as a result of that half-reaction, *i.e.* negative for oxidative or nucleophilic transforms and positive for reductive or electrophilic half-reactions. For reasonable transforms involving pairs of synthons, these polarities have to overlap, and the synthesis tree is constructed on this basis. Undesirable sequences may be eliminated on grounds such as undesired product functionality (violating the 'ideal synthesis concept'), competing



Scheme 17 The formulation of a bondset for steroid synthesis.

functionality or unavailable starting materials. Hendrickson showed that a large number of reactions (such as those in online databases) could be analysed in the same way, but by characterising the transforms as redistribution of bonds in cyclic 'transitions state' of various ring sizes, essentially with the same result of condensing reactions into an efficient classification.[42]

The sequences thus generated are exhaustive, since all possible reaction combinations may be considered, and they are not pre-judged on likely yield. From a statistical analysis of literature syntheses, Hendrickson concluded it is desirable to synthesise roughly one bond in four of a given target, which correspondingly implies the size of the bondsets required.

Hendrickson employed this reaction classification and logic in the program SYNGEN.[53] An example of the program's capabilities is shown in Scheme 18. The route shown is a literature synthesis of estrone (**93**), where the starting materials **89** to **91** were assembled into an intermediate compound possessing the estrone skeleton (**92**), followed by a couple of functional group interconversions to generate the desired product. SYNGEN found this synthesis (rated with high priority) when given the intermediate as the target, but not when estrone itself was entered. If functional groups are effectively removed in the final stages of a synthesis, they do not illuminate a reaction history, and SYNGEN has nothing to go on. Hendrickson developed the FORWARD program to allow for functional group additions to targets, which basically allows for a greater diversity of starting materials to be selected for any given bond set, making the selection more 'fuzzy.' The combination of SYNGEN and FORWARD was able to generate the synthesis of estrone itself.

We saw above that IGOR was capable of discovering new reaction types that were subsequently validated in the laboratory. An exceptional example of such discovery was demonstrated more recently by Herges in the search for new methodology towards butadienes.[54] A computer-generated set of all conceivable 7-centre/8-electron pericyclic reactions contained 72 general schemes. Three were found to be suitable candidates for the synthesis of butadienes, only one of which was known. For the two others (**94** to **95** and **100** to **101**, Scheme 19), various 'real-world' variables such as activation enthalpy and possible side reactions were introduced again *via* computation. The result was the design of two unprecedented reactions that were demonstrated in the laboratory (**96** to **98** and **102** to **103**). This example of reaction discovery is notable for its lack of serendipity!

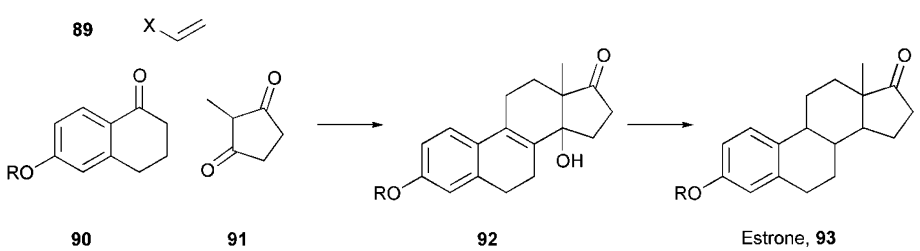
Concluding remarks on the retrosynthetic approach

We have seen a wide spectrum of approaches to CAOS in the retrosynthetic direction. On the one hand, LHASA represents the application of a large expert system coupled with a great many empirically-derived heuristics that achieve a human-like pruning of the synthetic tree. In contrast the formal approaches are able to consider all possible reactions for a given molecule, and predict the shortest possible synthetic paths unfettered by chemical knowledge. The control of what is an otherwise fully-fledged combinatorial explosion in the latter may be effectively controlled either by the insistence on the shortest, or most convergent, possible routes being considered only, or by a semi-formal approach where simple rules of physical chemistry are included.

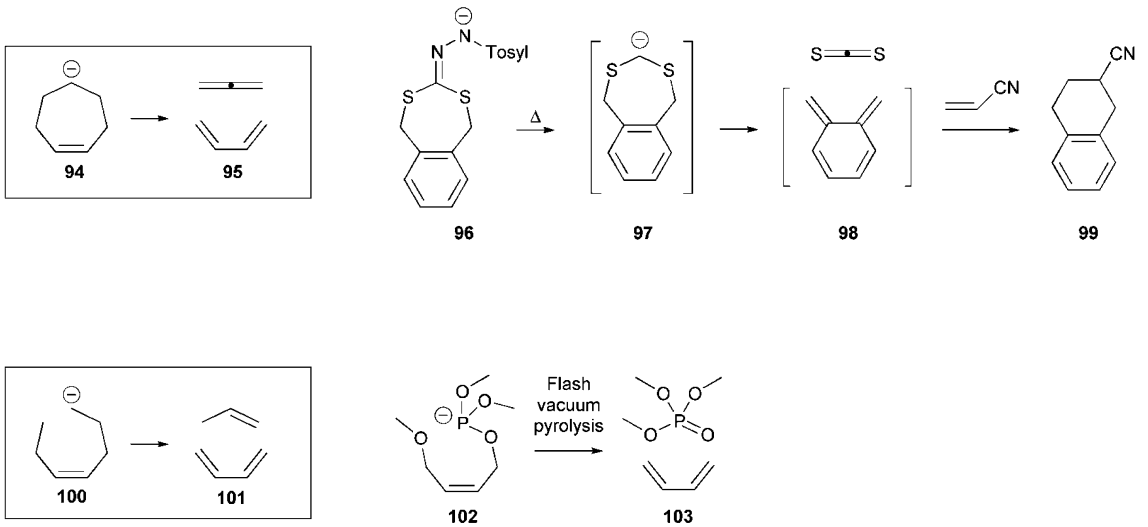
It may be desirable that retrosynthetic analyses of intermediate compounds to starting materials are not repeated if those intermediates keep recurring during various analyses. It should be possible to store successful analyses of these materials, and retrieve them, rather than working through them again. This is analogous to a chess program recognising certain near-endgame positions and knowing solutions for mate. An industrial group has used such a storage of retrosynthesis of intermediates as part of an in-house synthesis program.[55]

The forward approach

The ultimate aim of the CAOS field has always been the union of forward and backward search processes. A starting-material oriented search strategy was developed for LHASA, a knowledge of commercially-available starting materials was included



Scheme 18 SYNGEN's synthesis of estrone.



Scheme 19 Herges' discovery of novel butadiene syntheses *via* a formal approach.

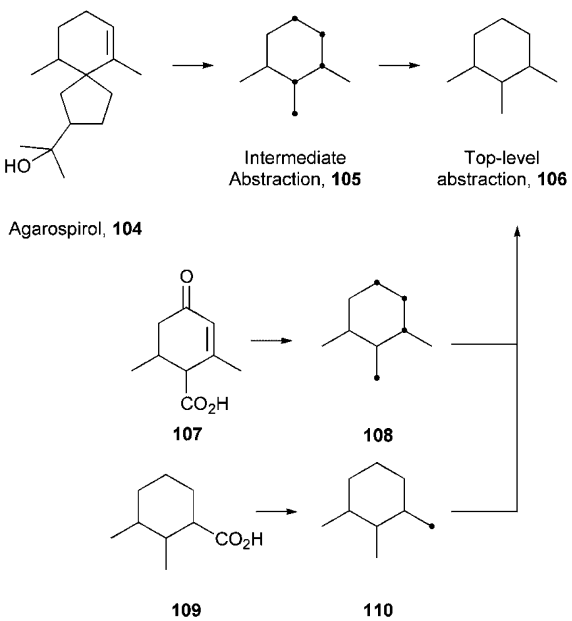
in SYNCHEM from an early stage. Besides knowledge of starting materials, a forward synthesis approach requires methods of comparison between target and candidate materials, as well as the capability of reaction prediction.

Search for starting materials

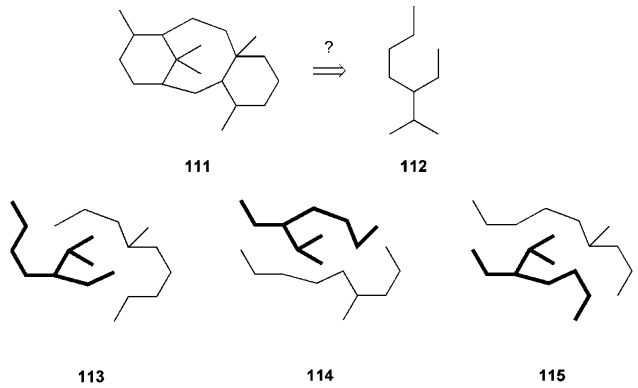
SST. Wipke developed the program SST for the search for starting materials, with the intention of mimicking the chemist's *Gestalt* approach of using pattern recognition and intuitive leaps.[56] The method used was one of abstraction of the target and starting materials—lowering the level of detail and attempting to spot superficial synthetic relationships between them. Two levels of abstraction were carried out, one high-level where all functionality is removed, and one intermediate where markers indicating the location (but not type) of functionality in the molecule are included.

From a survey of literature syntheses, Wipke concluded we tend to synthesise 20% of all carbon–carbon bonds (Hendrickson had concluded roughly 25% in a similar analysis) and 40% of all carbon–heteroatom bonds in most syntheses, with the remainder originating from the starting materials used. This implied that carbon–heteroatom bonds are a less important feature for recognising potential starting materials. This and related conclusions allowed the generation of abstraction rules for converting both starting materials and targets to abstracted graphs (an abstracted library of starting materials may of course be stored), which were then compared. High-level abstracted graphs were compared first, and to distinguish between similar starting materials retrieved, the intermediate level abstracted graphs were then compared, to test for similarity in functionality. The program successfully found the starting materials used in a large number of literature syntheses. An example of the approach is shown in Scheme 20, where a search was made for starting materials for agarospirol (**104**). An intermediate level of abstraction generated structure **105** for this molecule (dots indicate where functionality was present), whereas the high-level abstraction generated **106**. Two candidate starting materials in the program's library gave graphs that are substructures of **104** when they were fully abstracted, and these would both be selected as appropriate. Based on the intermediate level of abstraction, however, **107** was revealed as the preferred starting material, and indeed this was the molecule used in the literature synthesis of this compound. This kind of hierarchical search successfully limits the problem space of graphical matching, and is independent of reaction knowledge.

The carbon skeleton of complex natural products may contain, hidden, the carbon skeleta of simple commercially-available starting materials. A more recent program, SESAM, was written to search for these.[57] The user could enter the framework of a target molecule and that of a simple starting material, and the program would indicate whether a match was found. Stereochemical and functional group aspects of



Scheme 20 Levels of abstraction in the search for starting materials by SST.



Scheme 21 Abstraction of the carbon skeleton in the search for hidden approaches to the Taxol™ skeleton.

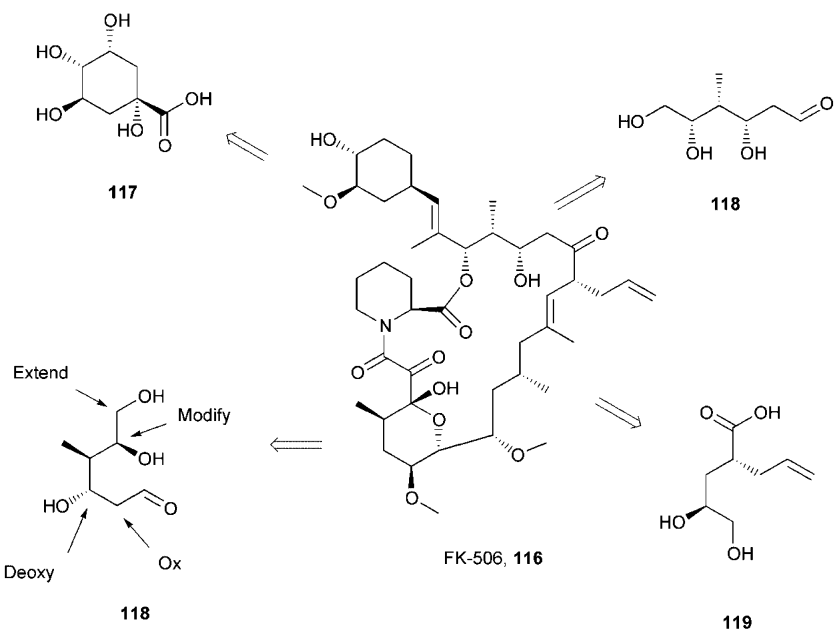
both were ignored, but the program was able to reveal what were perhaps hidden similarities between the two, and thereby perhaps suggest new synthetic strategies. For example, the Taxol™ framework (**111**) was examined for the occurrence of skeleton **112** (Scheme 21) which is available from a monoterpene. Eighty-eight 'solutions' were found where this arrangement of carbon atoms was found in the framework of the natural product (three examples **113–115** shown here).

The chiron approach

An important contribution to starting material-oriented synthetic planning was Hanessian's 'chiron' approach.[58] Here the emphasis of a retrosynthetic analysis is on leaving much of the stereochemical (and functional group) information in the target intact. The synthetic path conceptually proceeds forward *via* 'chirons' as starting materials rather than with Corey's 'synthons' to reflect this change in emphasis. Clearly in some cases, the choice of starting material may be apparent from simple visual inspection of the target, but in cases where an appropriate chiron is hidden, the application of computer-aided methods has the potential to reveal appropriate starting materials by exhaustive comparison of the target with databases of available enantiopure starting materials.

Hanessian developed the CHIRON program for this purpose. The program was designed to search for maximal similarity (carbon skeleton, functionality and stereochemistry) between the target and databases of available chemicals, and scored the candidates accordingly. CHIRON searched for those structures that are either commercially available or easily accessible *via* literature methodology, and which contain at least three carbons. The program's understanding of functional group interconversions permitted the identification of starting materials that are not an exact match. For example, unsaturation present in cyclic molecules may be hypothetically cleaved, and functionality may be introduced adjacent to carbonyl groups (with a higher score being given to such functionalisation in the α position than the β position). Currently the database the program uses (which is being continually updated) consists of the online Available Chemicals Directory, as well as over 3000 commercially unavailable chiral, non-racemic precursors selected from the literature, with total knowledge base of over 150 000 compounds.[59]

The CHIRON program has been shown capable of selecting starting materials for highly complex synthetic targets. For example, an analysis of the immunosuppressant FK 506 (**116**, Scheme 22) was performed. Quinic acid (**117**) was found as a suitable precursor for the northwest portion of the structure, a starting material that was also used by the Merck group that carried out the total synthesis of FK-506. The common precursor **118** was suggested for the northeast and southwest portions; this common precursor is in fact assigned a lower score for the match with the latter owing to a requirement for a hydroxylation–deoxygenation sequence, and for a new stereogenic centre at the aldehyde carbon. The program suggests the



Scheme 22 Starting materials suggested by CHIRON for the synthesis of FK-506.

necessary changes for each fragment, and these are shown for the southwest fragment in the scheme. Clearly the chiron approach is of particular interest in the pharmaceutical industry where large-scale syntheses, or syntheses of commercially valuable compounds far simpler than the likes of FK 506, are required, and the CHIRON program has consequently found wide use in this setting.

Reaction prediction

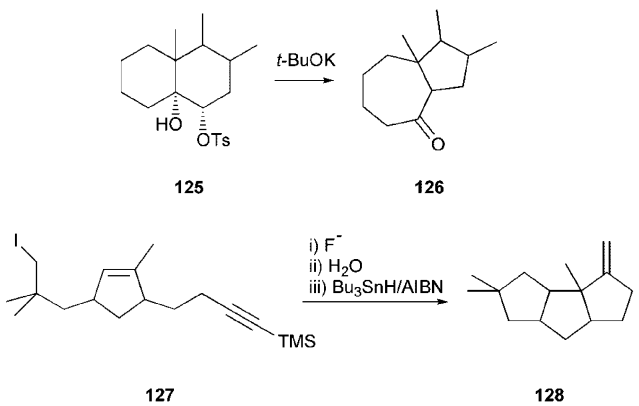
CAMEO. A program to predict the outcome of organic reactions based on a mechanistic approach rather than an empirical one was developed by Jorgensen, called CAMEO (Computer Assisted Mechanistic Evaluation of Organic Reactions).[60] This is of fundamental importance to computer-aided organic synthesis in the forward sense, but, like Corey's logical analysis of retrosynthesis, also questions our understanding of organic chemistry at its base.

The program was designed to indicate the feasibility of individual steps (including novel reactions) in a synthetic plan, including the identity of any likely side products. The powerful inclusions in CAMEO's perception module were concerned with pK_a and the identification of electrophiles and nucleophiles. Fifteen pK_a 'levels' were defined from the extremes of the hydrohalic acids to alkyl groups (*i.e.* from −7 to about 40), and CAMEO was thereby given an understanding of whether a base used was sufficiently strong to remove any given proton to a synthetically significant extent. CAMEO combined pK_a perception with recognition of nucleophiles and electrophiles. For example, if acetophenone is treated with sodium ethoxide, CAMEO perceived three nucleophilic sites—the reagent alkoxide oxygen and the enolate at both carbon and oxygen. On the other hand, if the base used is LDA, then only the enolate sites were identified owing to the program's understanding of pK_a. Further, pK_a was employed as a measure of leaving group ability, which feeds into the identification of electrophiles. From a pedagogical viewpoint, one can only approve of CAMEO's focus on pK_a to help understand a wide variety of reaction mechanisms.

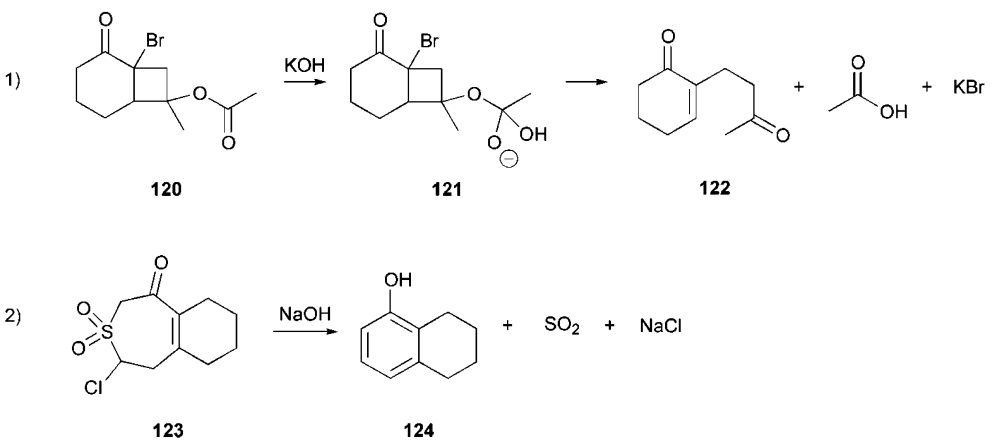
For mechanistic evaluation, a qualification value was ascribed to each nucleophile to describe whether, due to steric factors, it prefers substitution *versus* elimination pathways. The various possible mechanisms were then considered, and heuristics (derived from literature precedents) were applied to decide between them, for example in the competition between E2 and S_N2 mechanisms. Output structures were finally screened for unstable structures or more stable tautomers. The combination of all these factors led to sophisticated prediction of reaction products. For example, CAMEO successfully predicted the cyclobutane cleavage reaction shown (**120** to **122**, Scheme 23). In another example, the importance of post-mechanism screening is illustrated, where the episulfone formed after the treatment of **123** with sodium hydroxide was recognised as unstable, and the product that results was perceived to have a more stable aromatic tautomer (**124**), which was the only product displayed.

CAMEO's predictive power increased with its knowledge of reaction mechanisms and the power of the computers running it. CAMEO has been shown capable of correctly predicting the outcome of diastereoselective addition reactions in accordance with Cram's rule, for example.[61]

The program has been extensively tested against literature reports. Two examples are shown in Scheme 24. In the first, the program was able to predict successfully that treatment of the tosylate **125** with a hindered base would not lead to epoxide formation due to an absence of the appropriate stereochemical relationship between the functional groups, and that the likely (and reported) product is the rearranged



Scheme 24 Advanced reaction prediction by CAMEO.

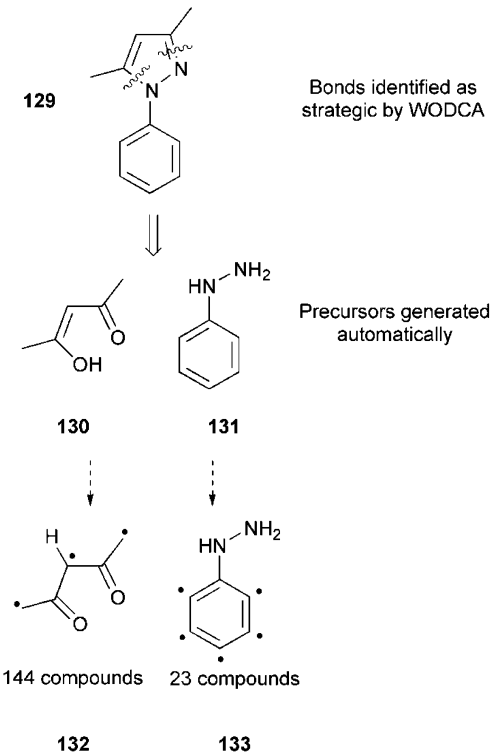


Scheme 23 Successful reaction prediction by CAMEO.

compound **126**. In the second example, CAMEO was able to predict the correct product from treatment of **127** with Bu_3SnH and AIBN (as well as minor reduction and bicyclic byproducts), which is the final step in Curran's hirsutene synthesis.

WODCA. Gasteiger incorporated the EROS reaction generation program (*vide supra*) into a larger set of programs known as WODCA (Workbench for the Organisation of Data for Chemical Application).[62] Retrosynthetic routes were first generated that suggest synthetic intermediates mid-way in the synthetic tree. The subsequent search for precursors is necessary in the large majority of cases where there is no direct link between target and starting materials. WODCA dissects the target, identifying and scoring strategic bonds. EROS predicts the likely ease of the forward transformations implied by the retrosynthesis. The program contains databases of several commercial suppliers as well as the CHIRON database, altogether containing more than 10 000 structures.

WODCA was designed to be highly modular, in the sense that additions to its knowledge base could be easily implemented. Heterocyclic transforms have tended to require treatment separate from traditional retrosynthetic methods owing to the number of effective and highly specific routes to the formation of these compounds, and a heterocyclic module was added to WODCA.[63] The combined functions of WODCA, strategic bond identification, starting material identification and reaction prediction, have been shown suitable for the synthetic design of combinatorial libraries.[64] Automated reaction prediction finds particular use in such an application, where it is desirable to know if the diverse inputs to the library may mean that the generation of certain library compounds would be problematic. WODCA's plan for a library of substituted pyrazoles is shown in Scheme 25. For each bond, the two possible heterolytic breakages were judged by physicochemical effects as described above. The bonds rated 'most strategic' by WODCA are those shown. Generalised precursor structures **132** and **133** were then generated (variable positions are indicated with an asterisk in the scheme). A search was then made for suitable starting materials from catalogues, and in this example WODCA found a large number of suitable inputs for this library. A similar tool for selection of inputs to combinatorial libraries was developed by another group, where an automatic assessment of a molecule's reactive centre was made from a set of descriptors that accounted for electronic and steric effects.[65]



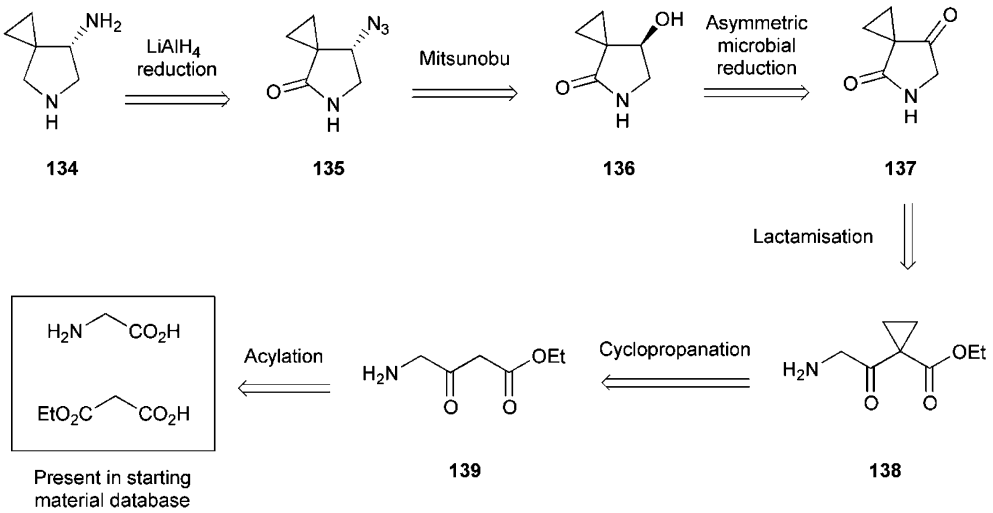
Scheme 25 Application of WODCA to the planning of combinatorial libraries.

New applications of artificial intelligence

Automated learning. A reaction database contains a great deal of raw knowledge, and is continuously updated. In contrast, knowledge-based CAOS systems have a smaller core of information, but one that is heavily processed to be both useful and relevant. Is there a way of automatically extracting information from large databases of reaction information, such that a machine may 'learn' rules and heuristics for application in synthetic design? Systems of this type appeared in the early 1990's.[66] Gelernter reported the adaptation of SYNCHEM to machine learning in both a deductive and inductive sense in order to increase the program's knowledge base automatically.[67] Funatsu's KOSP system (Knowledge base-Oriented system for Synthesis Planning) uses a knowledge base extracted from reaction databases.[68] This program, of the interactive retrosynthetic type, firstly groups similar reactions and exhaustively compares combinations of up to six bonds in the target molecule to its knowledge base, to see if it 'knows' a relevant transformation. A match corresponds to a strategic disconnection site, the disconnection is carried out, and dummy atoms are placed where the relevant bond was cut, which are then replaced by suitable 'leaving groups' in a subsequent step of the analysis. This cycle repeats until proposed fragments are those in a database of available starting materials. The program was rapidly able to find a synthetic route to the pharmaceutical intermediate **134** (Scheme 26) that involved an asymmetric microbial reduction, and this route was validated in the laboratory. No pruning of the synthetic tree takes place, and this will ultimately become problematic as the knowledge base expands.

Neural networks are a powerful computer science tool for automated learning.[69] Knowledge of reaction outcomes may be used as 'training sets' for the automated extraction of heuristics without any prior knowledge of chemistry. Such methods have been applied with success to the prediction of the major enantiomer in stereoselective reactions, for example.[70]

There is an interesting problem in the exploitation of databases of reactions in machine learning. Machine learning methodology requires failed instances as part of the training



Scheme 26 Example retrosynthetic analysis by the KOSP program.

procedure. Reaction databases are conspicuously lacking in reports of failed reactions. The recent emergence of a database dedicated to failed reactions may remedy this deficiency.[71]

Minimisation of complexity. We saw above how SYNGEN and SECS both employed an abstraction of the carbon skeleton of the target molecule to simplify the problem, or to search for hidden transforms. A similar approach has been used in smaller, more recent efforts. The HOLOWin program for example, uses abstracted graphs to search only for highly simplifying transforms where several bonds are disconnected at the same time along the lines of the minimisation of complexity described in general terms by Bertz, above.[72] Similar approaches, where the carbon skeleton is treated as the primary consideration, and functional groups are treated in a most general sense, have been implemented by Zefirov.[73]

Parallelisation. Gelernter's program SYNCHEM, described earlier, was adapted to run on multiple workstations, for example in a single building.[41] Clearly the speed with which this distributed version found literature syntheses was increased over a single-processor machine, and the search space could be covered with greater breadth. This is the only reported case where parallelisation has been exploited in the running of CAOS, and the lesson from the exercise was that the architecture of the implementation is crucial. The 'Master–Worker' model was selected, consisting of one workstation as the Master, and several others as Workers. The Master assesses the scores of individual pathways and directs where the search should proceed next, whereas the 'Workers' actually carry out the retrosynthetic manipulations, and report structures back to the Master. (This computer science terminology, and the following description of the program's function, may have some appeal to academics in the chemistry field). As mentioned above, the original SYNCHEM program was primarily a depth-first system, with promising lines being investigated to a point where either a starting material was encountered, or some flaw in the route was found. The distributed version could explore both the most promising intermediate, and at the same time a number of other near-best intermediates. This gave the distributed version much more breadth of analysis, and substantially more of the search space was analysed as a result. With this increased power came management problems of both Master and Worker. Approximately 35% of the nodes in a synthetic tree were duplicates of other structures in that tree. This entailed the Master closely monitoring the Workers to ensure they did not replicate an analysis that had already been carried out, and the Master also had to apply the knowledge gained in one analysis to the scoring of other pathways of relevance in a recursive fashion. Further, when workers had finished scouting a potential pathway, and had reported back to the Master, the Worker had to wait for the Master to decide the next route to explore. As a synthesis tree grew, the Master became overwhelmed unless Workers were given freedom to choose an appropriate local best intermediate to analyse. Conversely, towards the end of an analysis, Workers fell idle, since only one Worker could examine a path at any one time, and this effectively meant Workers had to be laid off as the analysis neared completion. The speed of the analysis here could clearly be compromised by inappropriate systems architecture.

The future

We would agree with others[74] that predictions of the future perhaps do not belong in a review article. There are, however, several advances in computer science that seem particularly appropriate for application to CAOS. DNA computing[75] has been shown able to solve particularly demanding problems by virtue of its massively parallel nature.[76] Support Vector Machines are known to be very adept at pattern recognition after only a brief training period.[77] And Genetic Programming has solved similar problems in the design of circuit boards *via* iterative improvement where only the circuit output is specified.[78] This latter is the only one of these areas of computer science that has been applied to CAOS, in an unpublished study by an industrial group.[79]

We do need to be clear, however, about the nature of our expectations for computer-aided organic synthesis. If we are hopeful of a computer program that is capable of devising, autonomously, a retrosynthetic analysis of a very complex natural product then we may be disappointed in the short term. The complexity of the analysis is severe, and grows worse as the knowledge base of organic chemistry increases (a self-inflicted problem!)

However, recent progress in new areas of computer science is beginning to point to dramatic improvements in the abilities of CAOS. As with the success of Deep Blue described at the start of this article, bringing increased brute-force computing power to the problem is only one part of this progress. Gradual improvements in heuristics and scoring functions, efficient computer architecture and modern algorithms are all needed for genuinely improved approaches to CAOS.

We should also remember that we are interested in computer-*aided* organic synthesis. A cooperative effort exploits the strengths of both parties. We pride ourselves, quite rightly, on having exceptional abilities at perception, heuristics, deductive and inductive learning and judgement as to the global nature of a problem. Moreover, our abilities increase with practice. Grandmaster chess players use different parts of their brains in analysing a position from amateurs, employing high-level processing to identify the key features of a problem quickly.[80] In contrast, computers are tireless, exhaustive and unbiased in both analysis and comparative functions, and are immune to difficulties of three-dimensional perception. Further, assuming they are instructed correctly, computers will avoid blunders, which makes them unsettling chess opponents, but potentially valuable consultants for relatively simple synthesis problems. We should therefore be hopeful of an ever-more fruitful man–machine effort in synthetic design. Recent years have seen the inception of man/machine *vs*. man/machine matches in tournament chess, where each player employs their program of choice to assist in the analysis.[81] Such unions embrace the contribution of a computer in the analysis of a complex problem, and we may reasonably expect great things of such a collaboration in the future.

Acknowledgements

I am very grateful to Murray Campbell and Carl J. Spencer (IBM) for advice on the Deep Blue project and comments on the analogy with chess. My thanks go also to William Mydlowec (Pharmix Corporation) for providing the original stimulus for this article. I would also like to thank P. Judson (LHASA UK) and S. Hanessian (University of Montreal) for helpful comments.

Matthew H. Todd
Department of Chemistry, Queen Mary, University of London, Mile End Road, London, UK E1 4NS. E-mail: M.H.Todd@qmul.ac.uk

References

† Throughout this review, references to the original literature reports of syntheses of relevance to or matching those devised by a computer program may be found in the article describing that program. Here, for example, a discussion of the semisynthesis of cortisone from deoxycholic acid may be found in ref. 82, which is referenced in Corey's article.

1 A. Turing, in *Minds and Machines*, ed. A. R. Andersen, Prentice-Hall, Englewood Cliffs, NJ, 1964, pp. 4–30.
2 B. Pandolfini, *Kasparov and Deep Blue. The Historic Chess Match Between Man and Machine*, Simon & Schuster, New York, 1997.
3 M. Campbell, A. J. Hoane, Jr. and F.-H. Hsu, *Artif. Intell.*, 2002, **134**, 57–83.
4 Chess match commentary may be found at the IBM website covering the event: http://www.research.ibm.com/deepblue/games/game5/html/c.2.html.
5 M. A. Ott and J. H. Noordik, *Recl. Trav. Chim. Pays-Bas*, 1992, **111**, 239–246.
6 S. J. Russell and P. Norvig, *Artificial Intelligence: Modern Approach*, Prentice Hall, Upper Saddle River, NJ, 1995.
7 D. Harel, *Computers Ltd.: What They Really Can't Do*, Oxford University Press, Oxford, 2000.
8 N. S. Zefirov and E. V. Gordeeva, *Russ. Chem. Rev.*, 1987, **56**, 1753–1773.
9 R. Barone and M. Chanon, in *Encyclopedia of Computational Chemistry*, ed. P. Von R. Schleyer, Wiley, Chichester, 1998, pp. 2931–2948.
10 P. Bamborough and F. E. Cohen, *Curr. Opin. Struct. Biol.*, 1996, **6**, 236–241.
11 P. Bador, M. N. Surrel and J. P. Lardy, *New. J. Chem.*, 1992, **16**, 413–423.
12 R. Moll, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 131–133.
13 E. J. Corey, *Pure Appl. Chem.*, 1967, **14**, 19–37.
14 E. J. Corey and W. T. Wipke, *Science*, 1969, **166**, 178–192.
15 E. J. Corey, W. T. Wipke, R. D. Cramer, III and W. J. Howe, *J. Am. Chem. Soc.*, 1972, **94**, 421–430.
16 E. J. Corey, 2002, personal communication. ChemDraw is available from CambridgeSoft, Cambridge, MA 02140, USA, see www.cambridgesoft.com.
17 E. J. Corey, W. T. Wipke, R. D. Cramer, III and W. J. Howe, *J. Am. Chem. Soc.*, 1972, **94**, 431–439.
18 E. J. Corey and G. A. Petersson, *J. Am. Chem. Soc.*, 1972, **94**, 460–465.
19 E. J. Corey, W. J. Howe, H. W. Orf, D. A. Pensak and G. Petersson, *J. Am. Chem. Soc.*, 1975, **97**, 6116–6124.
20 C. Rücker, G. Rücker and S. H. Bertz, *J. Chem. Inf. Comput. Sci.*, 2004, **44**, 378–386.
21 E. J. Corey, R. D. Cramer, III and W. J. Howe, *J. Am. Chem. Soc.*, 1972, **94**, 440–459.
22 E. J. Corey and W. L. Jorgensen, *J. Am. Chem. Soc.*, 1976, **98**, 189–203.
23 E. J. Corey, A. K. Long and S. D. Rubenstein, *Science*, 1985, **228**, 408–418.
24 E. L. M. van Rozendaal, M. A. Ott and H. W. Scheeren, *Recl. Trav. Chim. Pays-Bas*, 1994, **113**, 297–303.
25 R. D. Stolow and L. J. Joncas, *J. Chem. Educ.*, 1980, **57**, 868–873.
26 E. J. Corey, A. P. Johnson and A. K. Long, *J. Org. Chem.*, 1980, **45**, 2051–2057.
27 E. J. Corey, A. K. Long, J. Mulzer, H. W. Orf, A. P. Johnson and A. P. W. Hewett, *J. Chem. Inf. Comput. Sci.*, 1980, **20**, 221–230.
28 E. J. Corey, W. J. Howe and D. A. Pensak, *J. Am. Chem. Soc.*, 1974, **96**, 7724–7737.
29 M. A. Ott and J. H. Noordik, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 98–108.
30 A. K. Long and J. C. Kappos, *J. Chem. Inf. Comput. Sci.*, 1994, **34**, 915–921.
31 E. J. Corey, A. K. Long, T. W. Greene and J. W. Miller, *J. Org. Chem.*, 1985, **50**, 1920–1927.
32 T. W. Greene and P. G. M. Wuts, *Protective Groups in Organic Synthesis*, John Wiley & Sons, New York, 3rd edn., 1999.
33 A. P. Johnson, C. Marshall and P. N. Judson, *J. Chem. Inf. Comput. Sci.*, 1992, **32**, 411–417 and the two succeeding papers.
34 A. P. Johnson, C. Marshall and P. N. Judson, *Recl. Trav. Chim. Pays-Bas*, 1992, **111**, 310–316.
35 N. Greene, P. N. Judson, J. J. Langowski and C. A. Marchant, *SAR QSAR Environ. Res.*, 1999, **10**, 299–313. See also the LHASA webpage at http://www.chem.leeds.ac.uk/LUK/.
36 W. T. Wipke, G. I. Ouchi and S. Krishnan, *Artif. Intell.*, 1978, **11**, 173–193.

37 W. T. Wipke and T. M. Dyott, *J. Am. Chem. Soc.*, 1974, **96**, 4825–4834 and the succeeding paper.
38 P. Gund, E. J. J. Grabowski, D. R. Hoff, G. M. Smith, J. D. Andose, J. B. Rhodes and W. T. Wipke, *J. Chem. Inf. Comput. Sci.*, 1980, **20**, 88–93.
39 H. L. Gelernter, A. F. Sanders, D. L. Larsen, K. K. Agarwal, R. H. Boivie, G. A. Spritzer and J. E. Searleman, *Science*, 1977, **197**, 1041–1049.
40 K. K. Agarwal, D. L. Larsen and H. L. Gelernter, *Comput. Chem.*, 1978, **2**, 75–84.
41 D. Krebsbach, H. Gelernter and S. M. Sieburth, *J. Chem. Inf. Comput. Sci.*, 1998, **38**, 595–604.
42 J. B. Hendrickson, *J. Chem. Inf. Comput. Sci.*, 1997, **37**, 852–860.
43 S. H. Bertz, *J. Am. Chem. Soc.*, 1981, **103**, 3599–3601.
44 S. H. Bertz, *J. Am. Chem. Soc.*, 1982, **104**, 5801–5803.
45 S. H. Bertz, *New J. Chem.*, 2003, **27**, 870–879.
46 P. L. Fuchs, *Tetrahedron*, 2001, **57**, 6855–6875.
47 R. Barone and M. Chanon, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 269–272.
48 I. Ugi, J. Bauer, K. Bley, A. Dengler, A. Dietz, E. Fontain, B. Gruber, R. Herges, M. Knauer, K. Reitsam and N. Stein, *Angew. Chem., Int. Ed. Engl.*, 1993, **32**, 201–227.
49 J. Dugundji and I. Ugi, *Top. Curr. Chem.*, 1973, **39**, 19–64.
50 A. Dengler, E. Fontain, M. Knauer, N. Stein and I. Ugi, *Recl. Trav. Chim. Pays-Bas*, 1992, **111**, 262–269.
51 J. Gasteiger and C. Jochum, *Top. Curr. Chem.*, 1978, **74**, 93–126.
52 J. B. Hendrickson, *Angew. Chem., Int. Ed. Engl.*, 1990, **29**, 1286–1295.
53 J. B. Hendrickson, D. L. Grier and A. G. Toczko, *J. Am. Chem. Soc.*, 1985, **107**, 5228–5238.
54 R. Herges and C. Hoock, *Science*, 1992, **255**, 711–713.
55 I. Dogane, T. Takabatake and M. Bersohn, *Recl. Trav. Chim. Pays-Bas*, 1992, **111**, 291–296.
56 W. T. Wipke and D. Rogers, *J. Chem. Inf. Comput. Sci.*, 1984, **24**, 71–81.
57 G. Mehta, R. Barone and M. Chanon, *Eur. J. Org. Chem.*, 1998, 1409–1412.
58 S. Hanessian, J. Franco and B. Larouche, *Pure Appl. Chem.*, 1990, **62**, 1887–1910.
59 Details of updates to the Chiron program may be found on the relevant page of Hanessian's website at the University of Montreal: http://osiris.corg.umontreal.ca/chiron.html.
60 T. D. Salatin and W. L. Jorgensen, *J. Org. Chem.*, 1980, **45**, 2043–2051.
61 J. M. Fleischer, A. J. Gushurst and W. L. Jorgensen, *J. Org. Chem.*, 1995, **60**, 490–498.
62 W.-D. Ihlenfeldt and J. Gasteiger, *Angew. Chem., Int. Ed. Engl.*, 1995, **34**, 2613–2633.
63 R. Fick, W.-D. Ihlenfeldt and J. Gasteiger, *Heterocycles*, 1995, **40**, 993–1007.
64 J. Gasteiger, M. Pfortner, M. Sitzmann, R. Hollering, O. Sacher, T. Kostka and N. Karg, *Perspect. Drug Discovery Des.*, 2000, **20**, 245–264.
65 M. Braban, I. Pop, X. Willard and D. Horvath, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 1119–1127.
66 E. S. Blurock, *J. Chem. Inf. Comput. Sci.*, 1990, **30**, 505–510.
67 H. Gelernter, J. R. Rose and C. Chen, *J. Chem. Inf. Comput. Sci.*, 1990, **30**, 492–504.
68 K. Satoh and K. Funatsu, *J. Chem. Inf. Comput. Sci.*, 1999, **39**, 316–325.
69 J. Gasteiger and J. Zupan, *Angew. Chem., Int. Ed. Engl.*, 1993, **32**, 503–527.
70 J. Aires-de-Sousa and J. Gasteiger, *J. Chem. Inf. Comput. Sci.*, 2001, **41**, 369–375.
71 The 'failed reactions' database may be found on Accelrys' website: http://www.accelrys.com/chem_db/failedreact.html.
72 F. Barberis, R. Barone and M. Chanon, *Tetrahedron*, 1996, **52**, 14625–14630.
73 N. S. Zefirov and E. V. Gordeeva, *J. Org. Chem.*, 1988, **53**, 527–532.
74 M. Bersohn and A. Esack, *Chem. Rev.*, 1976, **76**, 269–282.
75 L. M. Adleman, *Science*, 1994, **266**, 1021–1024.
76 Y. Liu, J. Xu, L. Pan and S. Wang, *J. Chem. Inf. Comput. Sci.*, 2002, **42**, 524–528.
77 Y. Takaoka, Y. Endo, S. Yamanobe, H. Kakinuma, T. Okubo, Y. Shimazaki, T. Ota, S. Sumiya and K. Yoshikawa, *J. Chem. Inf. Comput. Sci.*, 2003, **43**, 1269–1275.
78 J. R. Koza, F. H. Bennett, III, M. Keane and D. Andre, *Genetic Programming III: Automatic Programming and Automatic Circuit Synthesis*, Morgan Kaufmann Publishers, San Francisco, 1999.
79 W. J. Mydlowec, J. S. Yu and G. Lanza, in *Abstr. Pap. Am. Chem. Soc*, 2001, **221**:84-CINF.
80 O. Amidzic, H. J. Riehle, T. Fehr, C. Wienbruch and T. Elbert, *Nature*, 2001, **412**, 603.
81 Details of man/machine, or 'advanced' chess may be found on the Chessbase website: http://www.chessbase.com/events/events.asp?pid=133.
82 L. F. Fieser and M. Fieser, *Steroids*, Reinhold Publishing Corporation, New York, 1959.

