Theory-Driven Discovery of Reaction Pathways in The MECHEM System

Raúl E. Valdés-Pérez

School of Computer Science and
Center for Light Microscope Imaging and Biotechnology
Carnegie Mellon University

Abstract

One goal of machine discovery is to automate creative tasks from human scientific practice. This paper describes a project to automate in a general manner the theory-driven discovery of reaction pathways in chemistry and biology. We have designed a system - called MECHEM - that proposes credible pathway hypotheses from data ordinarily available to the chemist. MECHEM has been applied to reactions drawn from the history of biochemistry, from recent industrial chemistry as reported in journals, and from organic chemistry textbooks.

The paper first explains the chemical problem and discusses previous AI treatments. Then are presented the architecture of the system, the key algorithmic ideas, and the heuristics used to explore the very large space of chemical pathways. The system's efficacy is demonstrated on a biochemical reaction studied earlier by Kulkarni and Simon in the KEKADA system, and on another reaction from industrial chemistry.

Our project has also resulted in separate novel contributions to chemical knowledge, demonstrating that we have not simplified the task for our convenience, but have addressed its full complexity.

Introduction

The goal of machine discovery is to interpret scientific reasoning in computational terms by automating specific tasks or formulating general conceptions of scientific activities (e.g., experiment design, theory choice). The usual focus has been on scientific *discovery* because of its status in the scientific pantheon, although it has been argued by [Simon, 1966] that the processes of discovery are not distinct from less exalted aspects of scientific reasoning or from more mundane reasoning.

This paper describes the current state of a project that attempts to automate the discovery of reaction pathways in chemistry and biology and which began as the author's Ph.D. thesis [Valdes, 1990b, Valdes, 1990c]. This theory-driven task is of a long tradition in science, and is important today both for its scientific aspect and economic significance , since understanding the pathway underlying a chemical synthesis is a first step toward improving its industrial yield.

We have designed a system - called MECHEM - that can propose credible pathway hypotheses from data ordinarily available to the chemist. MECHEM has been applied to reaction data drawn from the history of biochemistry, from recent industrial chemistry as reported in journals, and from organic chemistry textbooks.

The following section first explains the chemical problem and discuss previous AI treatments of pathway discovery. Then are presented the architecture of the system, the key algorithmic ideas, and the heuristics used to explore the very large space of chemical pathways. Then the system's efficacy is demonstrated on two specific reactions, one of which was studied earlier by [Kulkarni and Simon, 1988] in the KEKADA system. Finally, the issue of generality is taken up.

Discovery of reaction pathways

The chemical problem of discovering reaction pathways is as follows. Given empirical data on starting materials and observed products (not necessarily final products) of a specific chemical reaction,

1. infer simple, plausible pathways that explain the empirical data and are consistent with existing theory, and
2. seek and apply further evidence to discriminate among the plausible pathways.

For example, a reaction that consumes the starting materials ornithine, ammonia, and carbon dioxide and that is observed to form the products urea, arginine, and water can be explained by the following pathway:

$$\begin{aligned} ornithine + NH_3 + CO_2 &\rightarrow H_2O + C_6H_{13}N_3O_3 \\ NH_3 + C_6H_{13}N_3O_3 &\rightarrow arginine + H_2O \\ arginine + H_2O &\rightarrow urea + ornithine \end{aligned}$$

in which the chemical species $C_6H_{13}N_3O_3$ has been conjectured, i.e., was not observed and was not part of

the input data. (Ornithine is $C_5H_{12}N_2O_2$, ammonia is NH_3, carbon dioxide is CO_2, arginine is $C_6H_{14}N_4O_2$, and urea is CH_4N_2O. The reader can verify that all steps are balanced.) The celebrated discovery of this pathway by the biochemist Hans Krebs was studied historically by [Holmes, 1980] and was the subject of a cognitive model by [Kulkarni and Simon, 1988].

The *idealized* problem of pathway discovery is to find the exact set of steps that occur in a given chemical reaction. In current practice it is infeasible to determine the pathway with certainty, for several reasons. The most recalcitrant reason is that, given usual experimental data and a pathway to explain those data, one can always create a more elaborate pathway that explains the data equally well. Hence, the *practical* task is to find a set of steps that is consistent with all or most of the experimental data, is consistent with current knowledge of what types of reaction tend to occur, and is simple, i.e., no more complex than needed.

Discovery of reaction pathways is a *theory-driven* task because most of the heuristics are derived from strong theoretical presuppositions drawn from the domain science.

Pertinent chemistry knowledge

This section describes the minimal chemistry knowledge needed to understand the task domain of MECHEM. [King, 1963] is a good elementary introduction to the basic concepts.

The notation $A + B \rightarrow X + Y$ expresses a process, or reaction, by which the chemical substances A and B are transformed into new substances X and Y. The starting materials are the chemicals that are placed in a medium to start a reaction or sequence of reactions. A molecular formula (or just formula) is a list of the type and number of atoms making up a chemical substance, e.g., CH_2O for formaldehyde. In this paper, a chemical species corresponds to a unique formula, and vice versa. A pathway is defined as a set of individual reactions, each of which is a step.

Ordinary chemical reactions conserve mass; molecules are transformed into product molecules only in ways that conserve total numbers and types of atoms. The concentration of a reaction species refers to the number of species molecules per unit volume; concentrations can vary over time as species react or are formed.

Role of pathway discovery in chemistry

Chemists often undertake to determine a reaction pathway for purely scientific reasons. Other reasons arise in the context of enhancing the yield of a synthesis, because even partial knowledge of the pathway can suggest changes to the reaction conditions that may improve the yield. For example, another reagent might be introduced in order to increase the rate of a desirable step. Studies of the pathways underlying chemical reactions appear frequently in chemical and chemical engineering journals. Even a paper whose main purpose is to describe a new chemical synthesis typically includes at least some discussion of the pathway underlying the synthesis.

The discovery of reaction pathways has not been successfully automated ([Valdes, 1990b] has details). In particular, the automatic generation of plausible pathways has received little attention; for example, [Carpenter, 1984] in the preface to a textbook on the subject states: "Regrettably (or not, depending on one's point of view) there are no algorithms for generating scientific hypotheses."

Finally, we note that the *space* of pathway solutions is well understood in chemistry: the space consists of a set of reaction steps, each step consisting of a small number of reactants giving rise to a small number of products. What was not understood prior to our work was how to formulate a *problem space*, i.e., to find operators or an algorithm that explore the space systematically.

Prior work

The only prior work in AI on discovering reaction pathways is that of [Soo *et al.*, 1987] and [Kulkarni and Simon, 1988]. Soo et al. describe a program that uses known rules to interpret experimental concentrations data represented in a special form (as Lineweaver-Burk plots). Given a catalogue of reaction pathways, the program tested each catalogued entry against constraints inferred by the rules. Hypothesis generation was not otherwise addressed, nor experiment design. Also, knowledge of all reaction species and their roles in the reaction (i.e., substrate or product) was assumed.

Kulkarni and Simon developed a cognitive model of the historic discovery of a biochemical reaction pathway by Hans Krebs, but they focussed on the heuristic strategies that Krebs used to design and interpret experiments. They did not attempt to automate in a systematic manner the general problem of discovering reaction pathways.

The MECHEM System

Architecture

Figure 1 is the model of pathway discovery underlying MECHEM; the only aspect of the model not currently addressed by our work is deciding whether the current hypotheses and evidence constitute a scientific contribution (deciding when to suspend work on a problem is discussed briefly in [Valdes, 1990b]).

As depicted, the theory-formation cycle starts with the initial data from a problem instance. These data suggest initial values of the hypothesis complexity parameters, which are the number of *steps* R and *species* S to be contained in a pathway hypothesis. Then the hypothesis generator (called STOICH) cycles until it finds a non-empty set of pathways, at

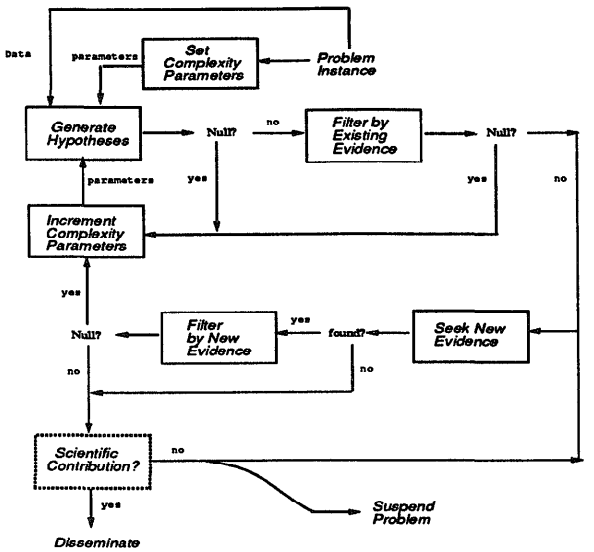

Figure 1: A Model for Pathway Discovery

which point these are tested against the available evidence not already incorporated in the generator. If all pathways are discredited to the point of rejection, then hypothesis generation resumes with incremented complexity parameters.

If the current pathways explain the existing evidence, then an experimentation cycle is entered. Initially, evidence is sought to discriminate among them. Any evidence found is applied to filter the pathways; if all of the latter are rejected, then hypothesis formation resumes after incrementing the complexity. If some pathways survive, then one might decide that a scientific result has been achieved, e.g., a small set of highly credible pathways.

If, on the other hand, no evidence can be found with an effort deemed appropriate, then still one might decide that a disseminable result has been reached, or perhaps decide to suspend the problem, or decide to incur a greater cost in order to find more evidence. Strategic decisions such as whether to suspend work on a problem, publish one's results, or incur greater cost to obtain evidence are not yet addressed by our work. Their inclusion in the diagram is intended to convey a picture of the theory-formation cycle that complements MECHEM, which to date primarily addresses detailed tactical reasoning about hypotheses and evidence.

Key algorithmic ideas

The space of possible reaction pathways has been well understood by chemists for nearly a century. Each element of the space is a set of reaction steps, and a step is a small set of reactants giving rise to a small set of products, syntactically separated by an arrow, i.e., reactants $\rightarrow$ products. *Our key discovery has not been one of perceiving this space, but of identifying an algorithm to generate the elements of the space in an order determined by an unproblematic simplicity metric.* That is, we have formulated a *problem space* where only an unstructured space was known beforehand. The two main operators, which are implicit within the pathway generator, add an extra step to a pathway fragment, and introduce a new lexical item (a new species variable). The generation algorithm is more fully described in forthcoming papers.

Another key algorithmic idea concerns how unseen species are introduced into pathways. An unseen species is initially designated as a variable (e.g., 'X') and later is instantiated by balancing the steps in which it appears. For example, on the one-step pathway fragment

$$C_3H_8 + O_2 \rightarrow X + C_2H_4O_2$$

$X = CH_4$ is deduced by balancing the step. If instead the one-step fragment were

$$C_3H_8 + O_2 \rightarrow X + Y$$

then X, Y would not be constrained to unique values. However, as further extensions to the fragment use X, Y in new steps, enough constraint can be added to result in unique values. In general, solving for the pathway variables is done by solving a matrix equation $Ax = b$, where A is the matrix of stoichiometric coefficients of the pathway variables, x is the vector of pathway variables, and vector b sums for each step the molecular formulas that are already known. The matrix equation can be solved by adapting a standard Gaussian Elimination algorithm to work on molecular formulas (i.e., vectors) rather than scalars.

MECHEM's search for pathways can well be viewed as a sequence of breadth-first searches that start each time with a different theoretical vocabulary; introducing a new unseen species (incrementing the number of species S) in effect augments the vocabulary of entities by a new variable. The breadth-first search corresponds to systematically searching for all pathways, given a fixed vocabulary, up to a certain depth R of number of steps. When MECHEM fails to find a satisfying hypothesis in complexity class S,R, it increments R and adds another level in the breadth-first search. How the breadth-first search is pruned by heuristics, and at what point S is incremented while R is reset, is described in [Valdes, 1990b] and in forthcoming papers.

It is worth emphasizing that the search is comprehensive due to its breadth-first nature, that is, all pathway hypotheses of specified complexity that are consistent with the data are found. Further details on the algorithm are found in [Valdes, 1990b] and [Valdes, submitted for publication].

Heuristics

Heuristics are the strategies used to guide search over a problem space. MECHEM's heuristics are not ex-

pressed declaratively within the program, but we will state them in such a form here, roughly in perceived order of importance.

1. MECHEM organizes its search in terms of **simplicity**. The space of pathways consistent with usual empirical data is infinite, so a severe selection criterion is needed. Simplicity is determined by the number of pathway steps and of species, with the latter the more important of the two.

2. Any reaction step that is not **balanced** is ruled out at once. This criterion follows the conservation law according to which the total number and identity of all reactant atoms remain unchanged after a chemical reaction. This conservation law is not an inviolable law of nature, since certainly mass can be converted into energy, but it does describe virtually all reactions studied by chemists.

3. MECHEM only considers reaction steps involving at most **two reactants** and at most **two products**. The representation of steps itself enforces this heuristic, but there is a way to rescind the restriction when needed, as in the case of the urea reaction below. The restriction is justified by chemical theory.

4. A **canonical representation** of pathways ensures that identical pathways will not be generated twice in the search [Valdes, 1991a]. The importance of MECHEM's canon for search efficiency is reminiscent of the DENDRAL case [Lindsay *et al.*, 1980], although the two canons are very different, since the DENDRAL canon was about molecular structures, not pathways.

5. Chemical species are **represented** as **molecular formulas**. If instead the species were represented immediately as molecular structures with all atoms and bonds explicit, then severe combinatoric problems would ensue due to the multiple ways that bonds could break and rearrange during a reaction step. This heuristic can be viewed as planning in an abstract space, since a molecular formula typically corresponds to many structural isomers.

The above heuristics are critical to the efficacy of MECHEM as a discoverer of reaction pathways. The next group of heuristics below are not critical, although each adds power in the sense of discriminating among otherwise equally plausible pathways.

There are a group of heuristics that MECHEM shares with DENDRAL at a conceptual level, although their justification and detail differ from DENDRAL's case. First is the concept of **superatoms**: a group of atoms deemed to be stable through the reactions can be defined as a superatom which is conserved intact just as single atoms are. Second is the degree of **unsaturation** of a molecular formula, which serves to rule out as implausible certain formulas proposed by the system. Third is the use of a **badlist**, which is a list of implausible components of pathways, e.g., a pair of species that are implausible co-reactants, or a specific step that is implausible. Badlists are currently input by the author in consultation with the reaction expert.

Another class of heuristics derives from **experiment**. These heuristic methods test whether a given pathway can account for experimental evidence. The methods used so far on actual reactions involve the following types of experimental evidence:

- precursor relations between two species, i.e., X is on the path to formation of Y.
- starting material A catalyzes the reaction [Valdes, 1992].
- concentrations of species measured over time [Valdes, 1990a].
- overall stoichiometry (net consumption/formation of observed species).

An important lesson from the DENDRAL project was the efficiency gain from incorporating heuristics as early as possible in a generator of candidate solutions [Lindsay *et al.*, 1980]. Our experience with MECHEM has confirmed the value of this lesson, and we have exerted much conceptual and programming effort in applying it. For example, species unknowns are instantiated as soon as unique values can be inferred from their use in a pathway (this involves solving a matrix equation in the general case of several unknowns and several steps).

Finally, [Valdes, 1990b] discusses some heuristics for experiment planning. So far MECHEM has been applied only to reactions derived from the chemical literature and books, so we have not had occasion to recommend or carry out experiments. Hence we omit discussion of this important topic, pending future demonstration of its role in MECHEM.

Validation

The efficacy of MECHEM can be validated in several ways. Firstly, the output of the system can be correctly seen as a deductive statement of the form:

theoretical assumptions + experimental evidence $\Rightarrow$ set of simplest pathways.

This formulation allows carrying out a systematic, comprehensive search, while providing confidence that no simpler satisfying pathways have been overlooked.

Our algorithms for generating and testing pathways, as described in the chemical literature, are guaranteed to embody correctly the relevant chemical concepts (although in at least one case we have argued for an algorithmic redefinition of a standard chemical concept [Valdes, 1991b]). MECHEM is guaranteed, short of a mistake in our programming or in our published theory, to find the simplest pathways consistent with the assumptions and evidence.

Secondly, MECHEM's pathway generator STOICH has also been implemented very differently in Prolog III [Jourdan and Valdes-Perez, 1990], and comparative tests between the two versions have resulted in exact agreement, thus corroborating MECHEM's reliability.

Thirdly, MECHEM has rediscovered in a systematic manner Krebs's pathway for the urea reaction. Currently we are applying MECHEM to some novel industrial chemistry [Smith and Savage, 1991] in consultation with Phillip Savage of the University of Michigan, with promising preliminary results. The rest of the current section describes in detail the results of the urea and industrial-chemistry reactions.

Urea synthesis

KEKADA, already mentioned above, implemented a historical reconstruction by Holmes of Hans Krebs's discovery of the cyclical biochemical pathway underlying the synthesis of urea *in vivo*. We have applied MECHEM to the problem of inferring pathways for the urea reaction, using a portion of the evidence discovered by Krebs and KEKADA. Krebs at one point had concluded that three starting materials were involved in forming three reaction products thus:

$$ornithine,\ NH_3,\ CO_2 \rightsquigarrow urea,\ H_2O,\ arginine$$

Also, Krebs had inferred from experiment that ornithine catalyzed the reaction, i.e., ornithine was completely re-generated after initially undergoing reaction. Lastly, Krebs knew from the chemical literature that arginine produces ornithine and urea in other contexts (see page 149 of [Kulkarni, 1988] or page 221 of [Holmes, 1980]), and considered this fact relevant to the urea pathway under study. We will require that arginine be a precursor of urea, although not necessarily in a single step.

We will use the above data together with a heuristic that merges two pathway steps into one three-reactant step if the following condition is met: a conjectured species appearing in the original two steps is eliminated from the new, merged step. This heuristic is formalized in a small program MERGE-STEPS. The sole purpose of MERGE-STEPS is to allow three-reactant steps like Krebs did. General reasoning about pathway catalysis in MECHEM is formalized in another program YIELD; this program and its supporting theoretical results have been reported [Valdes, 1991b, Valdes, 1992].

Thus STOICH, YIELD, and MERGE-STEPS combine to discover 10 simplest pathways consistent with Krebs's evidence. We show here only three of these, with the pathway proposed by Krebs shown first:

$$\begin{aligned}
ornithine + NH_3 + CO_2 &\rightarrow H_2O + C_6H_{13}N_3O_3\\
NH_3 + C_6H_{13}N_3O_3 &\rightarrow arginine + H_2O\\
arginine + H_2O &\rightarrow ornithine + urea\\
\\
NH_3 + CO_2 &\rightarrow H_2O + CHNO\\
ornithine + 2\,(CHNO) &\rightarrow CO_2 + arginine\\
arginine + H_2O &\rightarrow ornithine + urea\\
\\
ornithine + NH_3 &\leftrightarrow H_2O + C_5H_{13}N_3O\\
NH_3 + CO_2 + C_5H_{13}N_3O &\rightarrow arginine + H_2O\\
NH_3 + arginine &\rightarrow C_5H_{13}N_3O + urea
\end{aligned}$$

Each pathway involves seven species, meaning that STOICH needed to conjecture two unseen species to find a satisfying pathway; in each case one of the conjectured species was eliminated by MERGE-STEPS. The conjectured species $C_6H_{13}N_3O_3$ is the formula of citrulline, the intermediate proposed by Krebs (ornithine is $C_5H_{12}N_2O_2$, arginine is $C_6H_{14}N_4O_2$, urea is CH_4N_2O, ammonia is NH_3, and carbon dioxide is CO_2).

evidence	#pathways	class
starting materials & urea	0	4S2R
	0	5S2R
	6	6S2R
arginine is formed	0	7S2R
	71	8S3R
water is formed	16	8S3R
ornithine effect	98	8S4R
arginine precursor of urea	45	8S4R
merge step (allow three-reactant steps)	10	7S3R

Table 1: Pruning Hypotheses on Urea Pathway

Table 1 summarizes the evolution of hypotheses on the urea reaction. Starting only with knowledge of starting materials and of urea, STOICH finds six 6S2R pathways. Data on the formation of arginine and then of water changes the current pathways found by STOICH as shown. Then the current 16 pathways, together with the fact that ornithine is a catalyst, are input to YIELD, which finds that none of the 16 explains the "ornithine effect." STOICH then searches for incrementally more complex pathways, and the new batch in 8S4R is again tested by YIELD, leaving 98 pathways able to explain the ornithine effect. Of these, only 45 pathways show arginine to be a precursor of urea. On each remaining pathway the program MERGE-STEPS is called to merge two steps into one three-reactant step (as was allowed by Krebs) if the merge eliminates a conjectured species; this results in 10 simplest merged pathways, of which 3 were shown above.

We have limited MECHEM's input to the experimental evidence discussed by Holmes; other evidence could be sought for further discrimination. For example, consultation of the Merck Index of chemicals reveals the formula *CHNO* - which is a conjectured intermediate in two of the ten pathways - to be either cyanic or isocyanic acid, both of which are toxic and hence rather implausible biological intermediates.

Industrial chemistry

[Smith and Savage, 1991] have recently studied the novel chemistry of pyrolysis of 1-dodecylpyrene (DDP).[1] The chemistry of DDP is relevant to the processing of heavy crude oils in the petroleum industry. The authors identified experimentally seven major products of DDP, and measured the product concentrations at sampled times under various reaction con-

[1] Pyrolysis refers to the splitting (-lysis) of bonds due to heat (pyro-).

ditions. We have used their data from experiments at 400°C (personal communication).

Twelve pathways able to account for observed products and concentrations are found by MECHEM, of which three are shown here below. (The superatom Z refers to a polycyclic pyrene missing a hydrogen atom at the reactive site; its formula is $C_{16}H_9$. Also, for space reasons, the string 'pyrene' is replaced by 'Π'). Each pathway contains two conjectured species that were not input to the system, shown as formulas.

$$\begin{aligned} DDP &\rightarrow Decane + VinylΠ \\ DDP &\rightarrow Undecene + MethylΠ \\ MethylΠ + VinylΠ &\rightarrow Π + ZC_3H_5 \\ Undecene + ZC_3H_5 &\rightarrow EthylΠ + C_{12}H_{22} \\ Decane + EthylΠ &\rightarrow Π + Dodecane \end{aligned}$$

$$\begin{aligned} DDP &\rightarrow Decane + VinylΠ \\ DDP &\rightarrow Undecene + MethylΠ \\ DDP + Undecene &\rightarrow EthylΠ + C_{21}H_{42} \\ Decane + EthylΠ &\rightarrow Π + Dodecane \\ Π + Decane &\rightarrow MethylΠ + C_9H_{20} \end{aligned}$$

$$\begin{aligned} DDP &\rightarrow Decane + VinylΠ \\ DDP &\rightarrow Undecene + MethylΠ \\ DDP + Undecene &\rightarrow EthylΠ + C_{21}H_{42} \\ MethylΠ + VinylΠ &\rightarrow Π + ZC_3H_5 \\ Decane + EthylΠ &\rightarrow Π + Dodecane \end{aligned}$$

Phillip Savage has evaluated these results in a personal communication thus: "I find your identification of ZC_3H_5 and C_9H_{20} as key intermediates to be fascinating, because we have seen these products in modest yields."

Discussion

Currently, MECHEM is unable, with practicable amounts of computation, to handle cases where the number of conjectured species is greater than four. The execution time of the system increases greatly with this number, and the current practical boundary on a 14 MIPS workstation is four.

With respect to theoretical knowledge, MECHEM's main deficiency is its very limited ability to reason about molecular structure, which provides a strong constraint on plausible reactions. This deficiency is not intrinsic: structural reasoning was not our highest priority in bringing the system to its current state. Adding such reasoning is the next step in our agenda.

Related systems

Most work on machine discovery has concerned largely data-driven tasks such as law discovery, partly because such tasks offer the most scope for generality. Some recent work on theory-driven discovery includes that of [Karp, 1989], [Hayes-Roth *et al.*, 1986], [Kocabas, 1991], and [Sleeman *et al.*, 1990].

A brief comparison with the STAHLp system [Rose and Langley, 1987] and its successor REVOLVER [Rose and Langley, 1988] is of some interest. STAHLp finds componential models of substances that are implied by a set of given reactions, which is a step in MECHEM's search: given a pathway fragment with species variables, infer its component formula. STAHLp's task also involves finding plausible revisions to the input reactions when an illegal model is inferred. MECHEM, however, simply removes that node from its search. Finally, STAHLp obtains its componential models by a search involving substitution of one variable for another, while MECHEM carries out a Gaussian elimination algorithm to infer the component formulas of species variables.

Generality

MECHEM is a general system in the sense that it is prepared to handle *any* data from chemical reactions by proposing pathways accounting for the data. All the theory and heuristics incorporated into MECHEM thus far are completely general, i.e., they are not restricted to any type of chemistry, whether organic, biological, or industrial. Although the system can be improved by incorporating more chemical theory, it constitutes currently a theory of pathway discovery of considerable power, as was demonstrated in the above reaction examples.

The abstract mechanisms of MECHEM bear a strong similarity to parallel work carried out by a group headed by Jan Zytkow. Specifically, the Gell-Mann program in particle physics [Fischer and Zytkow, 1990] and the Mendel program in classical genetics (personal communication) are similar in the sense of conjecturing unseen quark and gene entities to account for observational data in a manner controlled by simplicity. This similarity lends credence to claims of MECHEM's generality at a conceptual level. We add that additional interest is generated in MECHEM's case by its application to problems of current importance, and by the link with experimentation.

[Langley *et al.*, 1987] state that "... generality is more likely to reside in data-driven approaches than in theory-driven ones." A reasonable implication is not that theory-driven science should not be an object of AI study, but that the role of generality in theory-driven science is unclear. One way to elucidate its role is to examine the generality (or lack thereof) of the heuristics used in successful theory-driven approaches to real-life problems, as we have done here. Our separate contributions to chemical knowledge show that we have not simplified the scientific task for our own convenience; rather, we have addressed the full complexity of the pathway-discovery task.

We intend to investigate further the generic scientific task of pathway elucidation by trying to apply ideas from MECHEM to the elucidation of the endocytic pathways of cell biology [Dautry-Varsat and Lodish, 1984].

Conclusion

This paper reports on the current status of the MECHEM project, which attempts to automate in a general manner the theory-driven discovery of reaction pathways in chemistry and biology. Significant progress has been made as demonstrated by application of the system to the rediscovery of a celebrated biochemical pathway and to industrial chemistry of current importance. The key algorithmic ideas underlying MECHEM are novel, even if appreciable intellectual debt is owed to the DENDRAL project. Our project has also resulted in novel contributions to chemical knowledge, and we expect MECHEM by itself to make a publishable discovery of a pathway in the medium term.

Acknowledgments

This paper is based on the author's Ph.D. thesis in Computer Science at Carnegie Mellon, supervised by Herbert Simon (chairman), Bruce Buchanan, Tom Mitchell, and Gary Powers. This research was supported by an Air Force Laboratory Graduate Fellowship, administered by the Southeastern Center for Electrical Engineering Education. Support for writing and publication derives from NSF grant DIR-8920118 to the Center for Light Microscope Imaging and Biotechnology.

References

Carpenter, Barry K. 1984. *Determination of Organic Reaction Mechanisms.* John Wiley & Sons.

Dautry-Varsat, Alice and Lodish, Harvey F. 1984. How receptors bring proteins and particles to cells. *Scientific American* 250:52–58.

Fischer, P. and Zytkow, Jan M. 1990. Discovering quarks and hidden structure. In *International Symposium on Methodologies for Intelligent Systems.*

Hayes-Roth, B.; Buchanan, B.; Lichtarge, O.; Hewett, M.; Altman, R.; Brinkley, J.; Cornelius, C.; Duncan, B.; and Jardetzky, O. 1986. Protean: Deriving protein structure from constraints. In *Proc. of AAAI.* 904–909.

Holmes, Frederic L. 1980. Hans Krebs and the discovery of the ornithine cycle. In *Proc. of 63rd Annual Meeting of the Federation of American Societies for Experimental Biology,* volume 39. Symposium on Aspects of the History of Biochemistry.

Jourdan, Jean and Valdes-Perez, Raul E. 1990. Constraint logic programming applied to hypothetical reasoning in chemistry. In *Logic Programming: Proc. of the 1990 North American Conference,* Cambridge, Mass. MIT Press. 154–172.

Karp, Peter D. 1989. Hypothesis formation as design. In Shrager, J. and Langley, P., editors 1989, *Computational Models of Scientific Discovery and Theory Formation.* Morgan Kaufmann, San Mateo, CA. 275–317.

King, Edward L. 1963. *How chemical reactions occur, an introduction to chemical kinetics and reaction mechanisms.* W.A. Benjamin.

Kocabas, Sakir 1991. Conflict resolution as discovery in particle physics. *Machine Learning* 6(3):277–309.

Kulkarni, D. and Simon, H.A. 1988. The processes of scientific discovery: The strategy of experimentation. *Cognitive Science* 12:139–175.

Kulkarni, D.S. 1988. *The Processes of Scientific Research: The Strategy of Experimentation.* Ph.D. Dissertation, Carnegie Mellon University. (Computer Science Department).

Langley, P.; Simon, H.A.; Bradshaw, G.L.; and Zytkow, J.M. 1987. *Scientific Discovery: Computational Explorations of the Creative Processes.* MIT Press, Cambridge, Mass.

Lindsay, R.K.; Buchanan, B.G.; Feigenbaum, E.A.; and Lederberg, J. 1980. *Applications of Artificial Intelligence for Organic Chemistry: The Dendral Project.* McGraw Hill, New York.

Rose, D. and Langley, P. 1987. Chemical discovery as belief revision. *Machine Learning* 1:423–451.

Rose, Donald and Langley, Pat 1988. A hill-climbing approach to machine discovery. *Proceedings of the Machine Learning Conference* 367–373.

Simon, H.A. 1966. Scientific discovery and the psychology of problem solving. In Colodny, R., editor 1966, *Mind and Cosmos.* University of Pittsburgh Press. 22–40.

Sleeman, D.H.; Stacey, M.K.; Edwards, P.; and Gray, N.A.B. 1990. An architecture for theory-driven scientific discovery. In *Proceedings of the Fourth European Working Session on Learning,* San Mateo, Calif. Morgan Kaufmann. 11–23.

Smith, C. Michael and Savage, Phillip E. 1991. Reactions of polycyclic alkylaromatics. 1. Pathways, kinetics, and mechanisms for 1-dodecylpyrene pyrolysis. *Ind. Eng. Chem. Res.* 30:331–339.

Soo, V.; Kulikowski, C.A.; Garfinkel, D.; and Garfinkel, L. 1987. Theory formation in postulating kinetic mechanisms: Reasoning with constraints. Technical Report CBM-TR-150, Department of Computer Science, Rutgers.

Valdes-Perez, Raul E. An algorithm to generate reaction pathways for computer-assisted elucidation. Submitted to *Journal of Computational Chemistry.*

Valdes-Perez, Raul E. 1990a. Coarse judgment of reaction model plausibility using linear estimation of reaction extent. Presented at Annual Meeting of American Institute of Chemical Engineers. Chicago, 11-16 November.

Valdes-Perez, Raul E. 1990b. *Machine Discovery of Chemical Reaction Pathways.* Ph.D. Dissertation, Carnegie Mellon University. (School of Computer Science, CMU-CS-90-191).

Valdes-Perez, Raul E. 1990c. Symbolic computing on reaction pathways. *Tetrahedron Computer Methodology* 3(5):277–285.

Valdes-Perez, Raul E. 1991a. A canonical representation of multistep reactions. *Journal of Chemical Information and Computer Sciences* 31(4):554–556.

Valdes-Perez, Raul E. 1991b. On the concept of stoichiometry of reaction mechanisms. *Journal of Physical Chemistry* 95(12):4918–4921.

Valdes-Perez, Raul E. 1992. A necessary condition for catalysis in reaction pathways. *Journal of Physical Chemistry.* In press.