

Machine Discovery of Chemical Reaction Pathways

Raúl E. Valdés-Pérez

A fundamental question in AI is what mechanisms suffice for computer programs to make scientific discoveries. My Ph.D. thesis (Valdés-Pérez 1990e) addresses this question by automating the following scientific task to a significant extent: Given observed data about a particular chemical reaction, discover the underlying set of reaction steps from starting materials to products, that is, elucidate the reaction pathway.

The automated system of pathway elucidation, called MECHEM, consists of three parts: (1) a hypothesis-formation program called STOICH that generates pathway hypotheses based on data commonly available to the chemist, (2) heuristics for designing experiments and seeking evidence

STOICH searches for pathway hypotheses guided by simplicity...

non-experimentally, and (3) a set of programs for applying diverse types of evidence to (dis)confirm hypotheses. My scientific contribution is to describe and interpret the design of a system that forms plausible explanatory hypotheses about dynamic processes in science and that proposes unseen entities in a manner justified by simplicity.

The MECHEM system has been applied to systematically derive Hans Krebs's historic discovery of the urea pathway using his same data; this derivation complements a previous cognitive model of this discovery by Kulkarni and Simon (1988), which focused on heuristics to guide experimentation. Also, several neglected (or new)

research questions in machine discovery are proposed, motivated by the practical problems confronted during system design. Some by-products of the thesis are several novel contributions to chemistry knowledge in addition to scientific tools of immediate use.

Introduction

Chapter 1 surveys previous work in machine discovery, focusing on work that involved assembling an extensive amount of knowledge particular to a domain. The outstanding research questions in the field are then listed, some as early as Amarel (1962) and others as recent as Langley et al. (1987). Because this thesis is a case study, the issue of how case studies lead to scientific progress is examined.

Chapter 2 presents the chemistry knowledge pertinent to the dissertation; explains how the elucidation of reaction pathways is usually carried out; and evaluates relevant automation work in AI, chemistry, and chemical engineering. The work of Soo et al. (1987) and Kulkarni and Simon (1988) are the only precursors within AI that deal with elucidating reaction pathways.

In chapter 3, I attempt to distinguish between two architectures for theory formation. One architecture generates many whole hypotheses in a constrained manner and then tests them, and the other (hypothetical) architecture mimics the human scientific reasoner's tendency to establish partial results before advancing whole hypotheses. A model for pathway discovery is proposed that is more elaborate than DENDRAL's constrained generate-and-test approach (Lindsay et al. 1980): STOICH generates pathways in order of simplicity, and experimentation has a stronger role in the model. The tactical stages

of this model have been implemented, but the more strategic stages are only discussed.

Design

Chapter 4 describes the design of STOICH, the hypothesis-formation program within MECHEM. STOICH searches for pathway hypotheses guided by simplicity; pathways with fewer reaction steps and fewer chemical species are simpler. For fixed numbers of steps R and species S , the program tries to explain the data and satisfy diverse constraints based in chemical theory. If no consistent hypotheses are found in a simplicity class $\langle S, R \rangle$, then R is incremented, and constraint satisfaction resumes. When further increments in R cannot lead to any consistent hypotheses, S is then incremented, and the procedure recurs after resetting R . One view of STOICH is that it performs search in two spaces (steps and species), with the satisfaction of constraints at its core.

By incrementing S , STOICH is conjecturing an unseen reaction species that it initially denotes using a variable, say, X . STOICH then uses X in its search for a consistent pathway, eventually inferring a value for X ; for example, the pathway step $2X \rightarrow C_4H_8$ entails a value of C_2H_4 for X . These conjectured species merit attention because they enable simple explanations of the problem data.

The abstract algorithm underlying STOICH is proved to be (partially) nonredundant; that is, it generates pathways according to a canonical representation, thus avoiding the generation of permuted versions of the same pathway. It is proved that the algorithm terminates on any well-posed problem by finding a consistent pathway hypothesis.

When given evidence used by Hans Krebs to discover the metabolic pathway underlying urea synthesis (compare Holmes [1980]), STOICH systematically derives the Krebs pathway as 1 among 10 of equal simplicity and conjectures an unseen species; no simpler pathways account for the evidence. This result complements that of Kulkarni and Simon (1988), who presented a cognitive model of the discovery based on the historical record but did not automate the general pathway discovery problem, which is the focus of this thesis. Their cognitive model did not attempt to discover alternatives to the Krebs

pathway that also explain the evidence.

Chapter 5 discusses several programs for applying diverse evidence to test hypotheses. One program that formalizes the notion of catalysis was used on the Krebs example. Another program discriminates among hypotheses using data on species concentrations measured over time. Heuristics for seeking evidence, of which some pertain to searching the chemical literature, are useful for discrimination.

Integrated Analysis

In chapter 6, I apply MECHEM to three reaction examples. The urea reaction (extended) and a second example are biochemical reactions, and the third reaction is one of industrial significance. Data on the latter reaction were drawn from the chemical literature, which also contained a complex, proposed pathway to explain the data. The chapter then discusses how MECHEM could be improved by incorporating more chemistry knowledge.

Chapter 7 revisits the important questions in machine discovery; I state what answers are implicit in MECHEM's design without claiming that these answers are necessarily the best. Some answers are proposed for several new (or neglected) questions that arose during the design. Finally, I compare the methods of STOICH to those of DENDRAL and META-DENDRAL.

Chapter 8 discusses the implications of the thesis. There are several lessons about the design of machine discovery programs. Other scientific problems are identified that could benefit from the ideas of the thesis. Finally, as a by-product of the design, there are several novel contributions to chemistry knowledge, as detailed in the appendixes.

Lessons

I suggest three provisional lessons for AI from this thesis. First, a systematic combinatorial search constrained by the evidence, domain constraints, and simplicity is sufficient to discover credible hypotheses in a creative scientific task of current importance. This lesson is succinctly expressed as the following reaction:

evidence + domain constraints + simplicity + combinatorial search → credible hypotheses .

Second, the specific means of conjecturing hidden entities has been shown effectual in a real task. During

construction of a hypothesis, unknowns are introduced whose values are later inferred by the use of domain constraints, for example, balance in the current case or conservation laws in the more general case. These conjectured entities have some initial credibility when they permit simple explanations of available evidence. These two lessons are the subject of a paper that has been submitted for publication.

Third, it has proved remarkably easy in this case to discover novel theoretical results in the domain science by designing programs to carry out hypothesis formation and experimental reasoning; some of these results are discussed in the references cited at the end. Moreover, one pathway discovered by MECHEM on the urea reaction suggests an unusual, general mechanism for catalysis that seemingly has never been proposed as a theoretical possibility, according to Michael Domach of Carnegie Mellon University. This mechanism will be submitted for peer review to check its novelty before making any claims that MECHEM has made a true discovery.

Acknowledgments

I am indebted to the members of the thesis committee for their encouragement and suggestions: Herbert Simon (chairman), Bruce Buchanan, Tom Mitchell, and Gary Powers. François Lecouat was my collaborator during the first year of this work. I was supported by a United States Air Force laboratory graduate fellowship administered by the Southeastern Center for Electrical Engineering Education.

Bibliography

- Amarel, S. 1962. An Approach to Automatic Theory Formation. In *Principles of Self Organization*, eds. H. Von Foerster and G. W. Zopf, Jr., 443–483. New York: Pergamon.
- Holmes, F. L. 1980. Hans Krebs and the Discovery of the Ornithine Cycle. In *Proceedings of the Sixty-Third Annual Meeting of the Federation of American Societies for Experimental Biology, Symposium on Aspects of the History of Biochemistry*.
- Jourdan, J., and Valdés-Pérez, R. E. 1990. Constraint Logic Programming Applied to Hypothetical Reasoning in Chemistry. In *Logic Programming: Proceedings of the 1990 North American Conference*, 154–172. Cambridge, Mass.: MIT Press.
- Kulkarni, D., and Simon, H. 1988. The Processes of Scientific Discovery: The Strategy of Experimentation. *Cognitive Science* 12:139–175.

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

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

Soo, V.; Kulikowski, C.; Garfinkel, D.; and Garfinkel, L. 1987. Theory Formation in Postulating Kinetic Mechanisms: Reasoning with Constraints, Technical Report, CBM-TR-150, Dept. of Computer Science, Rutgers Univ.

Valdés-Pérez, R. E. 1991a. A Canonical Representation of Multistep Reactions. *Journal of Chemical Information and Computer Sciences*. Forthcoming.

Valdés-Pérez, R. E. 1991b. On the Concept of Stoichiometry of Reaction Mechanisms. *Journal of Physical Chemistry* 95:4918–4921.

Valdés-Pérez, R. E. 1991c. Symbolic Computing on Reaction Pathways. *Tetrahedron Computer Methodology* 3(5). Forthcoming.

Valdés-Pérez, R. E. 1990a. A Correspondence between Reaction-Network Equilibria and Boolean Functions. *Chemical Engineering Science* 45:3384–3386.

Valdés-Pérez, R. E. 1990b. A Linear-Programming Reformulation of Chemical Stoichiometry and Catalysis, Technical Report, CMU-CS-90-172, Dept. of Computer Sciences, Carnegie Mellon Univ.

Valdés-Pérez, R. E. 1990c. Coarse Judgment of Reaction Model Plausibility Using Linear Estimation of Reaction Extent. Presented at the Annual Meeting of the American Institute of Chemical Engineers, 11–16 November, Chicago, Illinois.

Valdés-Pérez, R. E. 1990d. Deductive Assistance for Elucidation of Reaction Pathways. Presented at the Annual Meeting of the American Institute of Chemical Engineers, 11–16 November, Chicago, Illinois.

Valdés-Pérez, R. E. 1990e. Machine Discovery of Chemical Reaction Pathways. Ph.D. diss., CMU-CS-90-191, School of Computer Science, Carnegie Mellon Univ. To obtain this thesis, contact Computer Science Documentation, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA 15213-3890, (412) 268-2596, reports@cs.cmu.edu.

Valdés-Pérez, R. E. 1989. Learning Retrospective Knowledge from Scientific Laws: The Case of Chemical Kinetics, Technical Report, CMU-CS-89-179, Dept. of Computer Science, Carnegie Mellon Univ.

Raúl Valdés-Pérez is on the research faculty of the School of Computer Science and the Department of Biological Sciences and a member of the Center for Light Microscope Imaging and Biotechnology at Carnegie Mellon University. He obtained a Ph.D. in computer science from Carnegie Mellon University in 1991. His current interests center on machine discovery in the natural sciences.