Hypothesis Formation and Qualitative Reasoning in Molecular Biology

Peter D. Karp

My Ph.D. dissertation describes an investigation of scientific reasoning from a computational perspective (Karp 1989).[1] The investigation focuses on a program of research in molecular biology that culminated in the discovery of a new mechanism of bacterial gene regulation called *attenuation.* In the first phase of my work, I performed a historical study of this program of biological research in which I reconstructed the different theories that the biologists possessed at different points in time and then analyzed the differences between these successive theories. In the second phase, I developed a qualitative biochemistry for representing theories of molecular biology. In the third phase of the research, I constructed a machine learning program that solves some of the hypothesis-formation problems that the biologists solved during their research.

Hypothesis-formation problems occur when the observed outcome of an experiment conflicts with the outcome predicted by an existing scientific theory. The problem is to modify either the theory or the conception of the experiment such that the prediction agrees with the observation. This statement of the problem assumes that we are able to represent a scientific theory in a manner that allows a simulation program to use the theory to predict experimental outcomes and that allows a hypothesis-formation program to reason about the theory to understand why its predictions are faulty.

Qualitative Modeling

Chapter 3 of the dissertation presents three related methods for representing scientific theories. Only the third representation method was used for both prediction and hypothesis formation. This method breaks theories in molecular biology into three parts. A class knowledge base defines a taxonomic hierarchy of the classes of biological objects that exist in the trp operon gene-regulation system. A process knowledge base describes the chemical reactions that can occur between the biological objects in this system. An experiment is described in a third knowledge base by creating the particular objects (instantiated from the known classes of objects) that are present in the experiment.

A qualitative reasoning program called Gensim (genetics simulator) predicts experimental outcomes by determining what reactions occur between the objects in an experiment; these reactions create new objects, which can cause additional reactions. Gensim processes are similar to both Forbus's (1984) processes and production rules. The Gensim framework defines a qualitative biochemistry—an ontology for reasoning about chemical reactions. The dissertation identifies a number of constraints that a biochemical modeling program must satisfy to correctly simulate biochemical reactions. Some of the constraints have a negative impact on efficiency; thus, I also developed several optimizations to Gensim. Further, chapter 3 presents new qualitative representations for capturing the partial knowledge that biologists (and other scientists) have of the mathematical relationships that characterize the systems they study (see also Karp and Friedland [1989]).

A Historical Study of a Biological Discovery

Chapter 4 is a detailed historical study of the process by which Charles Yanofsky and his colleagues discovered attenuation. The study is based on information that my colleagues Peter Friedland and Rene Bach and I obtained from the scientific publications that the biologists produced and interviews that we conducted with the biologists. This biological research program consumed over 50 person-years of effort and is the most complex instance of scientific reasoning ever studied in AI.

In the first phase of the analysis, I produced a conceptual reconstruction of what knowledge the biologists possessed about the trp operon at different points in time. In the next phase, I searched for patterns in the differences between successive states of the biologists' knowledge. These differences were the result of changes the biologists had made to their theories of the trp operon. Patterns in the differences indicate reasoning methods that were used to derive new theories from old ones. My analysis identified theory-modification operators that the biologists used to modify their theories; these operators form the core of the hypothesis-formation program that I developed. These patterns also support the conjecture that scientists use four different modes of scientific exploration to determine what types of experiments to perform next from a given state of knowledge. A mode of exploration selects experiments based on the number of theories entertained at a given moment and the relative credibilities of the theories.

Hypothesis Formation

Chapter 5 describes methods for solving hypothesis-formation problems (see also Karp [1990]). A hypothesis is a proposed modification to either Gensim's theory or the initial conditions of the experiment (which are often not known with certainty), such that the predicted outcome of the experiment matches its observed outcome. The thesis treats the problem of hypothesis formation as a design problem. The Hypgene (hypothesis generator) program is a designer of hypotheses whose goal is to eliminate the difference between the observed and the predicted outcomes of the experiment—the *prediction error.* Hypgene uses a sophisticated planning system to reason backward from the prediction error to determine what modifications to the theory or initial experimental conditions will eliminate the error.

Four classes of design operators form the core of the Hypgene planner. Process-design operators modify

the chemical reaction theory, class–knowledge base design operators modify the class knowledge base, initial-condition design operators modify the presumed initial conditions of the experiment, and quantity-hypothesis design operators generate quantitative hypotheses. Only the last two classes of operators were implemented in Hypgene. Each operator performs a specific syntactic hypothesis-generation operation to rectify a specific type of prediction error. For example, one quantity-hypothesis design operator attempts to rectify a prediction error in which too little of some object is predicted to be present by postulating that more of the object is being produced by the occurrence of an additional chemical reaction.

This design problem is a search problem because often more than one design operator can be used to eliminate a given prediction error and because a single operator can sometimes be applied in several ways. The synthetic, goal-directed search used here should prove more efficient than past approaches to hypothesis generation, which often used heuristic search to guide a purely syntactic generator of hypotheses. Hypgene uses heuristic search to guide a generator that is already focused on errors in the prediction. Its search can be further guided by the results of a second experiment (which we term a reference experiment) whose initial conditions are similar to those of the first experiment. The dissertation shows that the difference between the initial conditions of the two experiments is likely to have caused the prediction error, and thus, we can use this difference to evaluate hypotheses that Hypgene has generated. In addition, the dissertation describes a second reasoning strategy for hypothesis formation that generates hypotheses by reasoning forward from the difference in initial conditions. This forward-reasoning strategy should be more efficient than the backward strategy for some problems.

Chapter 6 is an empirical investigation of the methods in chapters 3 and 5, in which I ran Gensim and Hypgene on several test problems. Gensim correctly predicted the outcomes of several biological experiments. Hypgene was tested on three sets of hypothesis-formation problems from the historical study of attenuation. For the most difficult problem, the program produced three of the four solutions that the biologists had proposed plus an additional solution that the biologists did not generate. Hypgene missed the one hypothesis because it did not have knowledge of the observational techniques used by the biologists; it generated the extra hypothesis because it did not have quantitative knowledge that the biologists had. Overall, Hypgene exhibited good performance on hypothesis-formation problems of realistic complexity.

Chapter 7 describes the implementation of the Hypgene program. It also compares and contrasts my methods for hypothesis formation with the machine learning work of Dietterich (1984); Simmons (1988); Rajamoney (1988); Lenat (1982); Wilkins (1988); Langley, Simon, and Zytkow (1987); Rose and Langley (1986); Kulkarni and Simon (1988); and others.

Summary

The thesis advances a view of scientific hypothesis formation as a synthetic, goal-directed activity—as a process of design—that is focused on errors in an experimental prediction. Hypgene provides a flexible framework for hypothesis formation because its framework is syntactically complete: Its design operators can modify the initial conditions of the experiment, the process knowledge base, or the class knowledge base. Hypgene's flexibility is also enhanced because its planner can manipulate complex predicate calculus expressions, allowing the program to reason about complex domain processes. Because Hypgene's planner and operators do not contain domain concepts, the framework is largely domain independent. The framework is efficient because Hypgene's planner works backward from prediction errors using operators that associate several syntactic classes of prediction errors with different types of theory modifications. The framework allows one to integrate domain-specific knowledge (such as general knowledge of chemistry) into the hypothesis generator to prune partial solutions during the generation process. Efficiency is further increased by the use of reference experiments, which provide information for both filtering and generating hypotheses. I validated the Hypgene program by testing it on actual hypothesis-formation problems that biologists solved during their discovery of a new mechanism of gene regulation.

Acknowledgments

This work benefited greatly from the guidance of Bruce Buchanan, Edward Feigenbaum, Peter Friedland, and Charles Yanofsky.

This work was supported by funding from the National Science Foundation, grant MCS83-10236; the National Institute of Health, grant RR-00785; and the Defense Advanced Research Projects Agency, contract N00039-83-C-0136.

References

Dietterich, T. 1984. Constraint Propagation Techniques for Theory-Driven Data Interpretation. Ph.D. diss., Dept. of Computer Science, Stanford Univ.

Forbus, K. 1984. Qualitative Process Theory. Ph D diss., Massachusetts Institute of Technology.

Karp, P 1990 Hypothesis Formation as Design. In *Computational Models of Discovery and Theory Formation.* San Mateo, Calif.: Morgan Kaufmann Also 1989. KSL-89-11, Knowledge Systems Laboratory, Stanford Univ

Karp, P 1989. Hypothesis Formation and Qualitative Reasoning in Molecular Biology. Ph.D. diss., Dept of Computer Science, Stanford Univ.

Karp, P, and Friedland, P. 1989. Coordinating the Use of Qualitative and Quantitative Knowledge in Declarative Device Modeling In *Artificial Intelligence, Modeling and Simulation.* New York: Wiley. Also 1987. KSL-87-09, Knowledge Systems Laboratory, Stanford Univ

Kulkarni, D., and Simon, H. 1988. The Process of Scientific Discovery: The Strategy of Experimentation. *Cognitive Science* 12:139–175

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

Lenat, D. 1982. AM: Discovery in Mathematics as Heuristic Search. In *Knowledge-Based Systems in Artificial Intelligence.* New York: McGraw-Hill.

Rajamoney, S. 1988 Explanation-Based Theory Revision: An Approach to the Problems of Incomplete and Incorrect Theories, Ph.D diss., Dept of Computer Science, Univ. of Illinois.

Rose, D., and Langley, P. 1986. Chemical Discovery as Belief Revision. *Machine Learning* 1:423–451.

Simmons, R. 1988 Combining Associa-

Solicitation for Videos About Research Efforts in AI Academic and Industrial Laboratories in the US and Abroad

As an experiment, the AAAI would like to communicate the different research activities within AI research laboratories in the U.S. and abroad using the video media. This is an opportunity for your lab's research efforts to be conveyed to a larger audience.

We are looking for short, 10 minute tapes which we plan to run in parallel in one large room. Please do not send us videos of a particular research project or taped lectures. We're looking for broad descriptions of different programs and projects within a lab.

If you are interested in submitting such a tape, please send it by **March 1, 1991** to:

AAAI-90 Videos
445 Burgess Drive
Menlo Park, CA 94025-3496

with the following basic information:

- Title
- Full names, postal addresses, phone numbers and email addresses of all authors
- Tape Format (e.g. VHS, 3/4" U-matic, NTSC, PAL, SECAM) and its duration in minutes
- One abstract briefly describing the lab's research programs, etc.; and
- Author or institution's permission to copy tape for reviewing purposes.

All tapes will be previewed. Only those tapes judged to be inappropriate will be returned.

tional and Causal Reasoning to Solve Interpretation and Planning Problems, Ph.D diss , Massachusetts Institute of Technology.

Wilkins, D 1988 Apprenticeship Learning Techniques for Knowledge Based Systems Ph.D. diss , Dept. of Computer Science, Stanford Univ.

Peter Karp is a research scientist with the Artificial Intelligence Center at SRI International. He recently completed a postdoctoral fellowship at the National Library of Medicine His research interests include knowledge representation, qualitative reasoning, and machine learning.

Note

1. For information on obtaining this dissertation, contact Publications Coordinator, Computer Science Department, Margaret Jacks Hall, Stanford University, Stanford, CA 94305 Phone: (415) 723-4776