

Some Recent Human-Computer Discoveries in Science and What Accounts for Them

Raul E. Valdes-Perez

■ My collaborators and I have recently reported in domain science journals several human-computer discoveries in biology, chemistry, and physics. One might ask what accounts for these findings, for example, whether they share a common pattern. My conclusion is that each finding involves a new representation of the scientific task: The problem spaces searched were unlike previous task problem spaces. Such new representations need not be wholly new to the history of science; rather, they can draw on useful representational pieces from elsewhere in natural or computer science. This account contrasts with earlier explanations of machine discovery based on the expert system view. My analysis also suggests a broader potential role for (AI) computer scientists in the practice of natural science.

My collaborators and I have recently reported several novel findings in the domains of biology, chemistry, and physics that represent various mixes of human-machine effort. These results, published in domain science journals, can fruitfully be analyzed as a whole by asking what accounts for them, that is, by seeking patterns within these results. The generality of any patterns that are found can be tested against other instances of human-machine discoveries. These patterns can also be tested against the conventional wisdom of the contributions to natural science that can be expected from computers and (AI) computer scientists.

This article carries out such a pattern-seeking study by making use of the concepts of

problem space and representation. My conclusion is that each of the three novel findings involves a new representation of the scientific task, in the sense that the problem spaces searched were unlike previous task problem spaces. I contrast my explanations of these human-computer discoveries with an earlier explanation of machine discovery based on the PROSPECTOR expert system. I end with some conclusions about the potential role of (AI) computer scientists in the practice of natural science.

Patterns in AI

A frequent task within an empirical science is to seek patterns within the phenomena studied by the science. In AI as empirical science (Simon 1995), the phenomena involve AI tasks and systems, their detailed designs, and the processes followed by the AI scientists who build the systems. One can seek patterns within these phenomena, a goal that was explicitly advocated by Simon (1979), arguing that the data-driven character of AI discovery systems suggests that AI research should itself be more data driven. As an early example of pattern finding, the system-building projects of the early AI scientists of the 1950s led, on detailed analysis of these systems, to the fundamental concept of heuristic search in problem spaces (Newell and Simon 1976).

Such data-driven theoretical work of inducing patterns from systems is not widespread in AI, but neither is it rare. I can cite two

examples from recent work: Valdes-Perez, Zytkow, and Simon (1993) analyzed six separately developed discovery systems and observed that all the discovery tasks shared an implicit representation named *search in matrix spaces*. Elsewhere, I proposed the concept of *generic task of scientific discovery* as a generalization of a broad array of work on computing in science (Valdes-Perez 1995a).

Representations

In a recent psychological article, Kaplan and Simon (1990) define a *representation* simply as a problem space and a *change of representation* as a change in problem space (also see Newell and Simon [1972]). The invention of a new representation for a scientific task is, then, the invention of a new problem space in which to accomplish the task.

Let us recall that a *problem space* is defined as a start state, a collection of operators that generate new states by recursive application, and a test of whether a goal state has been reached. The decision about when to apply each operator is guided by heuristics. In real-life tasks, problem spaces are often too large to be searched unselectively; the selectivity is provided by the heuristics.

Simon (1992) argued that a usual task of the scientist is not to invent completely unprecedented representations because these are historically scarce; a possible example is the invention of infinitesimals and their calculus, although even this representation might have some roots in antiquity in the paradoxes of Zeno and others (Moore 1995). Rather, the usual challenge for the creative scientist who finds new representations for scientific problems is to adapt and mix elements of known representations that are our common scientific heritage and to develop the operators and heuristics that are applicable in the new situation. In what follows, I adopt this concept of representation and ask to what extent each of the discoveries analyzed here was enabled by a new task representation. That is, I ask whether there are new operators or new types of state.

One can illustrate this notion of representation in science with a familiar example from theoretical computer science: the design of *randomized algorithms*, that is, algorithms that make random choices using a stream of random bits. Karp (1990, p. 2) states that “isolated examples of randomized algorithms can be traced back to the very early days of computer science, but the central importance of the concept of a randomized algorithm

became generally recognized only about fifteen years ago” and cites a 1976 article that drew attention to the general concept. Karp adds that by now, “randomization is an extremely important tool for the construction of algorithms” (p. 3). Hence, we see that a new representation of the algorithm design task has emerged, in the sense that algorithm designers now have an additional mental operator that incorporates coin flipping in the algorithm being designed; previously, designers would only make use of deterministic approaches. The problem space for algorithm design has thus been broadened.

The concept of representation that I have borrowed is technical; thus, detecting changes in representation amounts to the reasonably well-structured task of deciding whether a different problem space is involved. I have borrowed this concept directly from the theory of heuristic search, implying no conflict with other accounts of the representation concept. For example, an earlier article in this magazine gave an account of (knowledge) representation in terms of the roles that it plays in reasoning, such as serving as a surrogate for the natural world (Davis, Shrobe, and Szolovits 1993). There is no substantial conflict between these roles and the current definition.

Recent Discoveries

In each of the three cases of novel findings I present here, I proceed by giving some background on the task and the computational approach, documenting the finding, asking whether a new task representation is involved, and contrasting some of the previous or alternative task representations. I give only enough detail to appreciate the task and the role of the new representation; further details are available through the citations.

The three findings share the circumstance of involving more or less human inference, although they do not involve following the lead of domain scientists. Hence, my conclusions address the issue of what a computer plus a (AI) computer scientist can accomplish in science rather than what a computer by itself can accomplish. To address the latter issue, one needs to analyze in detail what fractional credit should accrue to human or machine, a task that would take us far afield.

Catalytic Chemistry

For the last six years (Valdes-Perez 1994a, 1995b), I have worked to automate a task that has occupied the energies of experimental

chemists since the beginning of this century, after it was first realized that many chemical reactions take place as a series of steps rather than as a single elementary act. A recurrent chemical research problem has been to explain (elucidate) the multistep character, or *pathway*, of a particular chemical reaction. Typically, this problem is one for the experimentalist because chemistry still lacks a practical theory to predict what steps will occur based on first principles. To obtain insight that complements background theory on what steps tend to occur, one carries out experiments with the reaction: measures the overall stoichiometry, detects reaction intermediates, infers precursor relations among the products, plots concentrations over time, and so on. Much reasoning particular to chemistry is involved;¹ hence, this task is much more theory driven than many of the recent tasks addressed by data-driven-discovery research (Langley and Zytkow 1989).

The approach to automating this discovery task began with a common AI question: What is the space of hypotheses, and how can one generate the elements (reaction pathways) of this space sensibly and systematically? There have been obstacles to overcome, involving the usual problems of combinatorial explosion and also some difficult algorithm designs; the solutions are described in the articles cited in this subsection. For the present, I focus on one of these obstacles, as follows:

A key problem in this and other scientific model-building tasks is that unseen entities have to be conjectured to formulate acceptable models. For example, in a chemical reaction, typically some intermediates remain undetected because of the practical limitations of experimental technique. If a large base of chemical knowledge were needed to make such conjectures competently, then automating this task would be problematic, except perhaps on a narrow class of chemical problems.

The obstacle of conjecturing unseen entities is overcome in MECHEM with a simple and seemingly naive method: Conjecture wild cards such as *X*, *Y*, and *Z*; use these wild cards, together with the seen entities, to formulate hypotheses; and then use the domain laws of a science to constrain these variables sufficiently (within the context of a specific hypothesis) to entail a small set of possible identities for the variables. For example, using the conservation constraint of reaction balance, the unknown *X* in the single-step hypothesis



is inferred to consist of one *M*, one carbon, two hydrogens, and one oxygen atom; that

1. $\text{H}_2 + \text{M}_2 \rightarrow 2\text{U}$.
2. $\text{M}_2 + \text{ethane} \rightarrow \text{U} + \text{V}$.
3. $\text{M}_2 + \text{V} \rightarrow \text{CH}_2\text{M}-\text{CH}_2\text{M} + \text{U}$.
4. $\text{M} + \text{CH}_2\text{M}-\text{CH}_2\text{M} \rightarrow \text{U} + \text{W}$.
5. $2\text{W} \rightarrow \text{CH}_2\text{M}-\text{CH}_2\text{M} + \text{CHM}-\text{CHM}$.
6. $\text{CHM}-\text{CHM} \rightarrow 2\text{X}$.
7. $\text{U} + \text{X} \rightarrow \text{M} + \text{Y}$.
8. $2\text{Y} \rightarrow \text{X} + \text{Z}$.
9. $\text{U} + \text{Z} \rightarrow \text{M}_2 + \text{methane}$.

Figure 1. A Reaction Pathway Containing Six Conjectured Wild Cards.

1. $\text{H}_2 + \text{M}_2 \rightarrow 2(\text{HM})$.
2. $\text{M}_2 + \text{ethane} \rightarrow \text{HM} + \text{CH}_3-\text{CH}_2\text{M}$.
3. $\text{M}_2 + \text{CH}_3-\text{CH}_2\text{M} \rightarrow \text{CH}_2\text{M}-\text{CH}_2\text{M} + \text{HM}$.
4. $\text{M} + \text{CH}_2\text{M}-\text{CH}_2\text{M} \rightarrow \text{HM} + \text{CH}_2\text{M}-\text{CHM}$.
5. $2(\text{CH}_2\text{M}-\text{CHM}) \rightarrow \text{CH}_2\text{M}-\text{CH}_2\text{M} + \text{CHM}-\text{CHM}$.
6. $\text{CHM}-\text{CHM} \rightarrow 2(\text{CHM})$.
7. $\text{HM} + \text{CHM} \rightarrow \text{M} + \text{CH}_2\text{M}$.
8. $2(\text{CH}_2\text{M}) \rightarrow \text{CHM} + \text{CH}_3\text{M}$.
9. $\text{HM} + \text{CH}_3\text{M} \rightarrow \text{M}_2 + \text{methane}$.

Figure 2. MECHEM's Reaction Pathway for Ethane + $\text{H}_2 \rightarrow 2$ (methane).

is, its chemical formula is MCH_2O . In more complex cases involving multiple steps and unknowns and possibly more than one unknown to a step, a generalized linear equation solver is used to infer the chemical formulas of the wild cards. The introduction of wild cards is guided by simplicity: *N* wild cards are introduced when *N* − 1 wild cards prove insufficient to account for the constraints, that is, when no adequate reaction pathway can be built. The program begins with *N* equal to zero.

However, modern chemists reason at the level of molecular structures (that is, graph-like connectivity between atoms), not just chemical formulas. However, it is also possible to infer the molecular structure of any wild cards given their already-inferred formulas and the overall multistep context in which they appear. For example, figure 1 shows a nine-step pathway in which the six wild cards *U*, *V*, *W*, *X*, *Y*, and *Z* appear (*M* and M_2 are not wild cards but catalyst reaction sites). Figure 2 shows the same pathway but with the wild cards replaced by the molecular structures that MECHEM inferred for them. MECHEM found this pathway, which I (Valdes-Perez 1994b) then proposed as an alternative explanation for the catalytic reaction *ethane* +

$H_2 \rightarrow 2$ (methane), which has been studied for over 20 years and whose pathway was considered largely solved. This discussion documents the first example of a novel finding: a simple, plausible alternative reaction pathway for a long-studied reaction.

I now consider the nature of the problem space that MECHEM searches.² Figure 1 shows a state in the problem space that is reached by applying operators that propose underdetermined chemical steps involving wild cards, or variables; each additional step that uses a variable places more constraint on the variable, resulting eventually in a completely determined step. This type of operator has not been used to date in computer applications to chemistry,³ which instead generally employ operators that result immediately in specific chemical steps (that is, not including variables); further elaborations of the reaction pathway do not constrain the steps further. Wipke, Ouchi, and Krishnan (1978) described these typical operators for an AI audience, although in the context of chemical synthesis, not elucidation, but the principle is the same. Ugi and Wochner (1988) described an alternative, more mathematical representation of chemical inference, which again differs from MECHEM's representation.

MECHEM's problem space involves new types of state and new operators and, therefore, implies that MECHEM uses a new representation of the chemical task of elucidating reaction pathways. Of course, the notion of using variables and solving for their identities by inference rules is hardly unprecedented, even though the application to molecular structure here involves complex inference. However, as Simon (1992) argued, the ordinary task of the scientist who proposes new task representations (even the scientist who carries out Kuhnian revolutionary science rather than "normal" science) is to stitch together the representation from familiar elements; the key issue is whether the resulting cloth is new to its particular dress.

Particle Physics

Previously, Langley et al. (1987) pointed out resemblances between the task addressed by their DALTON program and the following task from particle physics: Given observational data on reactions among elementary particles, one notices the absence of certain reactions that do not violate any known conservation laws, such as conservation of electric charge or baryon number. Because it is expected that any reaction that violates no conservation law will occur with some proba-

bility, these persistent absences are surprising and require an explanation. One form of explanation that has occurred is to invent new quantum numbers and assign them to the known particles such that the absent reactions fail to conserve the quantum numbers, but all the observed reactions successfully conserve the quantity. Devising such explanations for observational data (or phenomena) is what goes on in particle physics phenomenology; for example, the baryon and strangeness quantum numbers historically were postulated in such a data-driven way (Ne'eman and Kirsh 1986).

To illustrate the discovery task more concretely, let us consider the following simple problem that involves a single observed and a single never-observed reaction:

observed: $\bar{\pi} + p \rightarrow \pi^0 + n$
 unobserved: $p \nrightarrow \pi + \pi^0$.

To exclude the unobserved reaction while the observed one is admitted, it suffices to postulate a new conserved quantum number having the value of unity for p and n and zero for the other particles. This quantum number is conserved by the observed reaction and not conserved by the unobserved one.

Kocabas (1991) followed up on the observation by Langley et al. and wrote a Prolog program, BR-3, capable of postulating quantum numbers by a generate-and-test method that proposed specific numbers for each particle and backtracked when necessary. BR-3 was used on historical data to rediscover some accepted quantum numbers. In some cases, the program would find two new conserved properties to account for the observations, as has been the historical case.

In turn, I followed up on BR-3 after realizing that the logic of this discovery task could be formulated systematically, with guarantees about the simplicity of the resulting conservation laws. This realization was based partly on a representation that had been used in MECHEM and that is familiar to mathematical chemists: A reaction subject to conservation conditions can be represented as an algebraic equation. For example, the reaction $A + B \rightarrow C$, in which A , B , and C are substances, particles, or whatever, implies the equation $a + b = c$ over a numeric domain, such as the number of hydrogens present in the respective substances; the equality = expresses the conservation condition.

For the particle physics task, I formulated the relevant constraints explicitly and algebraically and then implemented the PAULI program, which is a search-linear-programming hybrid based on an algebraic formula-

tion of constraints (Valdes-Perez 1994c). PAULI is guaranteed to find the simplest assignment of quantum numbers, first in terms of the number of new conserved quantum properties (fewer is simpler) and second in terms of the magnitudes of the quantum numbers (smaller is better).

PAULI was applied to historical data that led to the independent discoveries of strangeness by Gell-Mann and Nishijima (Ne'eman and Kirsh 1986), and the program verified their solution, which consists of the strangeness quantum numbers. However, further experiments with PAULI (consisting of trying various input) led to an invariant pattern: PAULI always found that one new quantum property or conservation law was enough even when BR-3 (and physicists) found several. This pattern led to my conjecture and a proof (by M. Erdmann) of the theorem that on any input data consisting of observed and nonobserved reactions among elementary particles, at most, one conservation law suffices to explain the observations, if they can be explained at all in terms of conserved quantum numbers. Erdmann and I reported this theorem, together with a description of PAULI, to the physics community (Valdes-Perez and Erdmann 1994). This theorem constitutes the second example of a novel finding; this case involved more human inference than the first case because (simple) human reasoning was involved in the conjecture of the theorem but was prompted by the invariant behavior of the discovery program.

As in my first example involving MECHEM, this result was enabled by a new task representation, which involves an algebraic representation of the relevant constraints and which leads to guarantees about the simplicity of the solutions that the program will find. These constraints were reformulated, through various linear-programming representational tricks, into a working program. The operators in this new problem space add algebraic constraints and remove them, and the states consist of sets of algebraic constraints. To illustrate the problem space, the following two reactions lead to the possible state in PAULI's problem space that is shown in figure 3:

observed: $\bar{\pi} + p \rightarrow \pi^0 + n$
 unobserved: $p \not\rightarrow \pi + \pi^0$.

The use of the subscript 1 emphasizes that these are not particles but variables over a numeric domain.

Search within this representation, as carried out by PAULI, created the phenomena (that is, the program invariably found single conservation-law solutions) that led to the

$$\begin{array}{rcl}
 (\pi_1^0 + n_1) - (\bar{\pi}_1 + p_1) & = & 0 \\
 \pi_1 + \pi_1^0 - p_1 & \geq & 1/100 \\
 \pi_1 + \pi_1^0 - p_1 & \leq & 99 \\
 \pi_1 + \pi_1^0 - p_1 & \geq & -99 \\
 \bar{\pi}_1 + \pi_1 & = & 0 \\
 \bar{\pi}_1 & \leq & 0 \\
 \pi_1^0 & \leq & E_{\pi_1^0} \\
 -E_{\pi_1^0} & \leq & \pi_1^0 \\
 n_1 & \leq & E_{n_1} \\
 -E_{n_1} & \leq & n_1 \\
 \bar{\pi}_1 & \leq & E_{\bar{\pi}_1} \\
 -E_{\bar{\pi}_1} & \leq & \pi_1 \\
 p_1 & \leq & E_{p_1} \\
 -E_{p_1} & \leq & p_1 \\
 \pi_1 & \leq & E_{\pi_1} \\
 -E_{\pi_1} & \leq & \pi_1
 \end{array}$$

Figure 3. A State in PAULI's Problem Space.

data-driven conjecture and Erdmann's proof of the theorem. As with MECHEM, this problem space is new to computer applications in particle physics, including AI applications. The new representation involved piecing together known elements of other representations with which the author was familiar, such as representing reactions as algebraic equations. Some technical problems had to be overcome, but the representational pieces were, nevertheless, not without precedent.

Cell Biology

A developmental biologist with whom I have collaborated has studied a developmental stage of early *Drosophila* embryos in which nuclei, or cells not yet having acquired a membrane, undergo several successive rounds of division. These processes are illustrated in figure 4, which superimposes the original nucleus and the two daughter nuclei that result from each division. Earlier analysis of such data concluded that nuclear divisions were disordered, or random, based on a histogram of division angles that appeared roughly uniform between 0 and 180 degrees (Minden et al. 1989).

Our analysis led to the conclusion that the divisions displayed patterned behavior (Valdes-Perez and Minden 1995) rather than

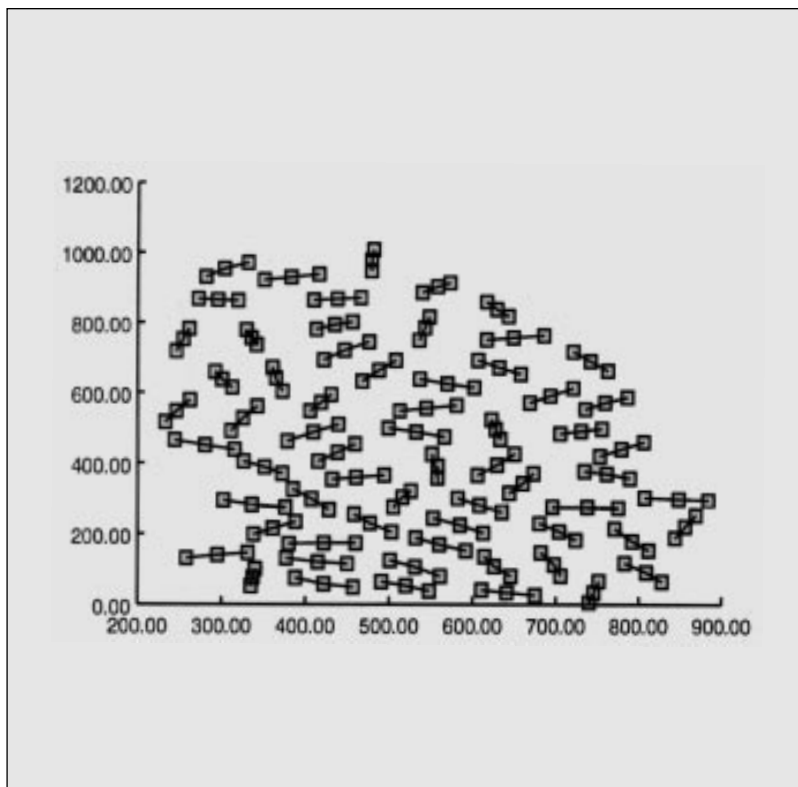


Figure 4. Nuclear Divisions within Early *Drosophila* Embryos.

the randomness cited earlier. This conclusion of patterning is my third and final example of a new finding in science. Detecting the patterning was followed up with hypothesizing a model that would give rise to the observations, although this follow-up is not relevant to the analysis of this article.

Our finding of patterned behavior exclusively involved human reasoning, which after the event I was able, using retrospection, to reconstruct, as shown in figure 5. This discovery can be credited to possessing or actively applying the heuristic rule that is the first premise of the deduction. As mentioned previously, earlier studies considered a different problem space in the search for patterns, for example, constructing a histogram of division angles, which revealed no simple pattern because the distribution was roughly uniform. Our finding involved, instead, reasoning in a different problem space that used the operators-inferences of Figure 5.

In contrast to the first two cases from chemistry and physics, the current biological example involved only human reasoning, not machine discovery. Nevertheless, afterward, we were able to devise a general, systematic method (PENCHANT) for finding subtle patterned behavior that is broadly applicable

(Valdes-Perez and Perez 1994). This method, which is a complex instance of a permutation test (Good 1994), generalizes the three steps in figure 5 as follows: Nuclear division becomes the process under study, division orientation becomes a process parameter, and crowding becomes an arbitrary feature or quantity.

PENCHANT reaches the same conclusion about patterned divisions that Minden and I had, which was not a trivial exercise because the human and computer methods are similar only on close analysis. Perez and I have since begun to apply the method to other problems both within biology and elsewhere. No significant machine discoveries have yet been made, but PENCHANT has uncovered unsuspected patterns of moderate interest to at least one experimentalist (Valdes-Perez and Perez 1994).

When PENCHANT makes an independent discovery of significance, rather than retraces human steps, then a stronger case for the role of the computer could be made than is warranted in this final example of a human-computer finding in science. Again, the task representation or problem space involves familiar elements drawn from elsewhere; nothing in figure 5 is new to the history of science.

Discussion

I argued that new task representations or problem spaces have been involved in the three documented discoveries in science that this article analyzed. It is likely that the use of a new representation was a crucial cause of the new findings rather than a mere coincidence.

The representation-based explanation differs from some earlier accounts of machine discovery that involved expert systems. For example, in a report of the novel recognition of a hidden mineral deposit by the PROSPECTOR program, Campbell et al. (1982) state that "expert systems...encode valuable knowledge and judgment for the purpose of making that expertise more broadly available than in the past" (p. 927). They explained PROSPECTOR's achievement as follows: "We believe that this result was achieved because PROSPECTOR programming adequately reflects limited but appropriate selections from the knowledge and judgment of a known porphyry molybdenum expert" (p. 928). Thus, PROSPECTOR's achievement is the result of confronting a model of a human expert (embodied in PROSPECTOR) with new data unavailable to the

original expert. This expert system view focuses on expanding the availability of expertise through the use of expert systems rather than expanding the expertise itself. In contrast, my representation-change explanation of the three recent discoveries focuses on a change in task problem space, that is, a change in expertise.

The new task representations that I claim were, of course, humanly designed or even searched (in the biology case) rather than invented by machine. However, much scientific work involves problem solving within a given representation rather than the invention of a new one. Hence, the computer's role is not unlike the role of many an individual scientist who makes a publishable discovery within a given representation that served as a starting point.

Given the research interest in automating changes in representation (Benjamin 1990), one can also ask how drastic were the representation changes described here. I believe that these changes were not minor; for example, they do not involve simple changes to the problem space such as deleting an operator, making an operator more specific by instantiating one of its parameters, and so on. Rather, the new problem spaces differ from the previous spaces in more complex ways. Moreover, these representation changes were largely theory driven (rather than data driven) in the sense that they involved analytic insights into the special task structure that were not obtained in any obvious way through the use of data. This theory-driven character makes it doubtful whether any attempt to automate these representation-change processes in a general way can succeed in practice, at least for scientific applications.

Besides the expert system viewpoint discussed earlier, other common views credit the achievements and/or potential of machine discovery to the computer's greater speed and memory or to capabilities for doing exhaustive searches within known representations. More broadly, the role for computer scientists in natural science is commonly perceived as limited to inventing or adapting faster algorithms and more sophisticated data structures or devising means to deal with large databases. However, the three discoveries in this article all involved modest computations (whether computer or mental), straightforward data structures, and modest amounts of input data.

These results suggest that new task representations, developed perhaps by a systematic analysis of the heuristic relation between data and discovery, will constitute a significant

1. If the nuclear divisions were truly random, that is, randomly oriented between 0 and 180 degrees, then one would see occasional crowding as a result of chance.
2. Observations of time-lapse embryo images reveal little crowding; on the contrary, the nuclei maintain good spacing.
3. Therefore, the nuclear divisions are not randomly oriented but must somehow be patterned.

Figure 5. Reasoning That Led to the Finding of Patterned Behavior.

fraction of the machine-aided discoveries that will occur in the short and medium term. As illustrated previously, these new representations need not involve elements wholly new to the history of science; rather, they can draw on a wide array of representational pieces that have proved useful elsewhere in natural or computer science. I surmise that a detailed analysis of other productive machine discovery programs (such as META-DENDRAL [Lindsay et al. 1993] and TETRAD [Spirtes, Glymour, and Scheines 1993] can confirm the pattern reported here.

Acknowledgments

This work was supported in part by a grant from the National Science Foundation (NSF) (#IRI-9421656), by the Center for Light Microscope Imaging and Biotechnology at CMU, and by the W. M. Keck Center for Advanced Training in Computational Biology. I thank Paul Rosenbloom for asking during a seminar at the University of Southern California Information Sciences Institute what accounted for these recent discoveries; this question eventually led to my writing this article.

Notes

1. Much of the reasoning pertains to generic pathway-elucidation tasks because pathways occur in branches of science other than chemistry.
2. Actually, MECHEM searches in more than one space because, for example, the space of molecular structures is searched to provide structural information on the conjectured intermediates. However, the analysis focuses on the top-level problem space of possible pathways, that is, steps of the form *reactants* → *products*.
3. This type of operator is also absent from human practice; for example, a recent book on writing reaction pathways (Miller 1992) contains nothing like this operator.

References

- Benjamin, D. P., ed. 1990. *Change of Representation and Inductive Bias*. Boston, Mass.: Kluwer Academic.

- Campbell, A. N.; Hollister, V. F.; Duda, R. O.; and Hart, P. E. 1982. Recognition of a Hidden Mineral Deposit by an Artificial Intelligence Program. *Science* 217:927–929.
- Davis, R.; Shrobe, H.; and Szolovits, P. 1993. What Is a Knowledge Representation? *AI Magazine* 14(1): 17–33.
- Good, P. 1994. *Permutation Tests*. New York: Springer Verlag.
- Kaplan, C. A., and Simon, H. A. 1990. In Search of Insight. *Cognitive Psychology* 22:374–419.
- Karp, R. M. 1990. An Introduction to Randomized Algorithms. Technical Report ICSI 90-024, International Computer Science Institute, Berkeley, California.
- Kocabas, S. 1991. Conflict Resolution as Discovery in Particle Physics. *Machine Learning* 6(3): 277–309.
- Langley, P., and Zytkow, J. M. 1989. Data-Driven Approaches to Empirical Discovery. *Artificial Intelligence* 40(1–3): 283–312.
- Langley, P.; Simon, H. A.; Bradshaw, G. L.; and Zytkow, J. M. 1987. *Scientific Discovery: Computational Explorations of the Creative Processes*. Cambridge, Mass.: MIT Press.
- Lindsay, R. K.; Buchanan, B. G.; Feigenbaum, E. A.; and Lederberg, J. 1993. DENDRAL: A Case Study of the First Expert System for Scientific Hypothesis Formation. *Artificial Intelligence* 61(2): 209–261.
- Miller, A. 1992. *Writing Reaction Mechanisms in Organic Chemistry*. San Diego, Calif.: Academic.
- Minden, J. S.; Agard, D. A.; Sedat, J. W.; and Alberts, B. M. 1989. Direct Cell Lineage Analysis in *Drosophila Melanogaster* by Time-Lapse, Three-Dimensional Optical Microscopy of Living Embryos. *Journal of Cell Biology* 109(2): 505–516.
- Moore, A. W. 1995. A Brief History of Infinity. *Scientific American* 272(4): 112–116.
- Ne'eman, Y., and Kirsh, Y. 1986. *The Particle Hunters*. New York: Cambridge University Press.
- Newell, A., and Simon, H. A. 1976. Computer Science as Empirical Inquiry: Symbols and Search. *Communications of the ACM* 19:111–126.
- Newell, A., and Simon, H. A. 1972. *Human Problem Solving*. Englewood Cliffs, N.J.: Prentice Hall.
- Simon, H. A. 1995. Artificial Intelligence: An Empirical Science. *Artificial Intelligence*. Forthcoming.
- Simon, H. A. 1992. Scientific Discovery as Problem Solving: Reply to Critics. *International Studies in the Philosophy of Science* 6(1): 69–88.
- Simon, H. A. 1979. Artificial Intelligence Research Strategies in Light of AI Models of Scientific Discovery. In Proceedings of the Sixth International Joint Conference on Artificial Intelligence, 1086–1093. Menlo Park, Calif.: International Joint Conferences on Artificial Intelligence.
- Spirtes, P.; Glymour, C.; and Scheines, R. 1993. *Causation, Prediction, and Search*. New York: Springer Verlag.
- Ugi, I., and Wochner, M. 1988. Molecular Logic and Computer Assistance in Chemistry. *THEOCHEM* 42(3–4): 229–242.
- Valdes-Perez, R. E. 1995a. Generic Tasks of Scientific Discovery. Paper Presented at AAAI Spring Symposium on Systematic Methods of Scientific Discovery, 27–29 March, Stanford, California.
- Valdes-Perez, R. E. 1995b. Machine Discovery in Chemistry: New Results. *Artificial Intelligence* 74(1): 191–201.
- Valdes-Perez, R. E. 1994a. Conjecturing Hidden Entities via Simplicity and Conservation Laws: Machine Discovery in Chemistry. *Artificial Intelligence* 65(2): 247–280.
- Valdes-Perez, R. E. 1994b. Human/Computer Interactive Elucidation of Reaction Mechanisms: Application to Catalyzed Hydrogenolysis of Ethane. *Catalysis Letters* 28(1): 79–87.
- Valdes-Perez, R. E. 1994c. Algebraic Reasoning about Reactions: Discovery of Conserved Properties in Particle Physics. *Machine Learning* 17(1): 47–68.
- Valdes-Perez, R. E., and Erdmann, M. 1994. Systematic Induction and Parsimony of Phenomenological Conservation Laws. *Computer Physics Communications* 83(2–3): 171–180.
- Valdes-Perez, R. E., and Minden, J. S. 1995. *Drosophila Melanogaster* Syncytial Nuclear Divisions Are Patterned: Time-Lapse Images, Hypothesis, and Computational Evidence. *Journal of Theoretical Biology*. Forthcoming.
- Valdes-Perez, R. E., and Perez, A. 1994. A Powerful Heuristic for the Discovery of Complex Patterned Behavior. In *Proceedings of the Eleventh International Conference on Machine Learning*, 326–334. San Francisco, Calif.: Morgan Kaufmann.
- Valdes-Perez, R. E.; Zytkow, J. M.; and Simon, H. A. 1993. Scientific Model-Building as Search in Matrix Spaces. In Proceedings of the Eleventh National Conference on Artificial Intelligence, 472–478. Menlo Park, Calif.: American Association for Artificial Intelligence.
- Wipke, W. T.; Ouchi, G. I.; and Krishnan, S. 1978. Simulation and Evaluation of Chemical Synthesis—SECS: An Application of Artificial Intelligence Techniques. *Artificial Intelligence* 11:173–193.

Raul Valdes-Perez is a research computer scientist in the Computer Science Department at Carnegie Mellon University (CMU), where his research centers on the novel use of computers for discovery in the sciences. He received a Ph.D. in computer science in 1991 under Herbert Simon at CMU and minored in the philosophy of science under Adolf Grunbaum at the University of Pittsburgh. Before entering CMU, he worked at the Massachusetts Institute of Technology AI Lab and in the Brazilian computer industry after obtaining engineering degrees from the University of Illinois at Chicago. He has taught AI courses in Hefei, China, and Cordoba, Argentina. He chaired a 1995 AAAI Spring Symposium entitled Systematic Methods of Scientific Discovery and is coediting a special issue on the same topic for *Artificial Intelligence*.