A Powerful Heuristic for the Discovery of Complex Patterned Behavior

Raúl E. Valdés-Pérez
Computer Science Department and
Center for Light Microscope Imaging and Biotechnology
Carnegie Mellon University
Pittsburgh, PA 15213 - USA
valdes@cs.cmu.edu

Aurora Pérez
Facultad de Informática
Universidad Politécnica de Madrid
Campus de Montegancedo s/n
Madrid, Spain
aurora@fi.upm.es

Abstract

A major activity of many sciences is to search for patterned *behavior* within complex phenomena. The fields of Biology and Psychology are just two examples, in which the discovery of patterns is an impetus for building explanatory models that could account for the patterns. This paper reports the invention of a powerful machine-oriented heuristic for finding complex patterned behavior in empirical data. The heuristic was developed by retrospecting on our own human reasoning during "field work" in experimental developmental biology, in which we detected a novel dynamic pattern in the mitoses of the early embryo. The new heuristic is broadly applicable: we also apply it to psychological data on memory in chess, with interesting results.

1 INTRODUCTION

The discovery of patterns from observation or experiment is a major impetus for the advance of science. Although the ultimate goal of basic natural science is, say, to *explain* phenomena, the natural scientist must choose what phenomena to explain, and the wise choice of phenomena is an important determinant of success. The research goals of explaining life, or explaining the Earth, or explaining the mind, are not problems that recommend themselves without first finding the regularities and patterns that underlie these very broad phenomena. It is the "smaller" patterns that focus the activity of most productive science, e.g., the striped magnetic pattern of the seafloor that led to the inference that the magnetic polarity of the Earth had fluctuated over the eons.

Some sciences concern themselves largely with the detection and explanation of patterns. For example, in a Science article [Melton, 1991] on "Pattern Formation During Animal Development," the author states that:

> At the beginning of this century, embryologists defined the central problems of developmental biology that remain today. These questions include how differentiated cells arise and form tissues and organs and how pattern is generated.

Oliver, in his book *The Incomplete Guide to the Art of Discovery* [Oliver, 1991, p. 78], instructs "Go for the spatial pattern" as a "specific tip on how to make important discoveries about the earth," but with analogues outside of geophysics.

In general, the discovery and elucidation of patterns are major activities in sciences that deal with complex phenomena in which causal interactions are not well understood, i.e., most sciences. Roughly, one may view patterns as manifestations of underlying order, as distinct from randomness. The word 'pattern' often refers to static *structural* patterns, such as the striped magnetic seafloor pattern cited above, or a hexagonal structural pattern that was observed in the North Pole of Saturn [Godfrey, 1988]. However, the word 'pattern' can also refer to *behavioral* patterns, in which time and dynamics are inherent. Of course, the two uses are related, since most structural patterns do not come into being out of whole cloth, but are the consequence of behavioral regularities or patterns that weave the structural patterns.

Now, the field of machine discovery is concerned with reconstructing (or constructing *de novo*) the logic, psychology, or history of scientific discovery by means of computation. Examples of work in this area include more-or-less psychological reconstructions of historical discoveries [Karp, 1993, Kulkarni and Simon, 1988, Langley et al., 1987, Ledesma et al., 1994, Ledesma et al., 1993], analysis of the computational and heuristic logic of some discovery task [Fischer and Zytkow, 1990, Lindsay et al., 1980, Valdes, 1992, Valdes, in press], conceptual generalizations of distinct discovery systems [Langley and Zytkow, 1989, Valdes et al., 1993], and theoretical analysis of the proposition that scientific discovery is heuristic search [Simon, 1966, Zytkow and Simon, 1988].

This paper examines the logic of another, important class of scientific discovery: detection of patterned behavior. We report the invention of a single, quite powerful heuristic that is capable of noticing complex patterned behavior in a wide variety of data from various sources. The heuristic has been implemented within a program PENCHANT that has seen application to data from developmental biology and from cognitive psychology.

We know of no other work in AI that addresses the task of detecting patterned behavior (which is distinct from finding static or structural patterns), much less work that focuses on science. Perhaps the closest is the discovery of sequence-generation rules from sample sequences (e.g., [Dietterich and Michalski, 1985, Laird, 1992]), in which a "pattern" is a rule that generates observed sequences, e.g., the Fibonacci rule. In our case, as will be seen, a pattern is "nonrandom" behavior of an unspecified nature. Our method does not carry out a search over pattern-generating rules, hence its applicability is broader than the methods described by Dietterich and Michalski and their successors. In contrast, our method makes weaker conclusions, e.g., proposing a specific explanation for the patterns found is outside the scope of this work. Also, unlike the case of sequence prediction, our method by itself will not result in individual predictions, only statistical ones. However, such weak conclusions about patterned behavior are crucial starting points for many discoveries in science.

Of added methodological interest here is the circumstance that the heuristic was invented by reflecting on our own recent discovery of patterned behavior within developing *Drosophila* embryos [Valdes and Minden]. We posed the question of how a program could, in a general manner, make the same discovery that we had just made. This experience suggests that "field work" in some science can be a fruitful technique for researchers in machine discovery.

We proceed by recounting our own experience of pattern discovery in developmental biology. Then, we show how a machine-oriented heuristic is extracted and purified from the human-oriented inference that we ourselves had employed. The PENCHANT program, which uses this single heuristic, is applied to the same data we used and successfully notices the same pattern. This reconstruction is not trivial, since the machine-oriented heuristic is not altogether similar to the human-oriented inference. We then illustrate further the heuristic's generality by adducing the results of applying PENCHANT to data on the psychology of memory in chess. Finally, we place the program's role within the broader context of discovery, and reconsider the widespread belief that "deep knowledge is critical for creativity" [Kim et al., 1993] in the light of these results, in which a single, moderately shallow heuristic is capable of surprising discoveries.

2 BACKGROUND

This paper could proceed by describing the heuristic and the computer program, but we believe it is more instructive in this case to consider the context of their development, especially since the field of machine discovery concerns itself with the process and context of scientific discovery, not only with its fruits.

As "field work" in machine discovery, we collaborated with a developmental biologist who studies the early embryo using fluorescent labelling. That is, certain constituents of the developing embryo, such as rapidly dividing cells within the embryo, are highlighted with "markers" that fluoresce under ordinary lighting, hence emit radiation at wavelengths that distinguish them from the other parts of the embryo. These fluorescent wavelengths are separated from the other light by filters, and the result is a picture of quite specific components of the embryo, for example, of dividing cells. The pictures are then filmed, digitized, and stored on a computer for subsequent analysis.

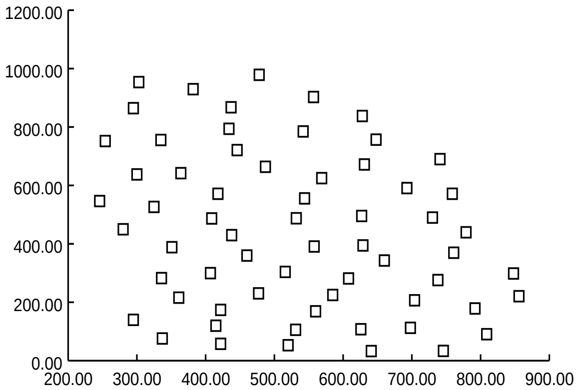

Figure 1: Undivided Nuclei in Early Developing Embryo

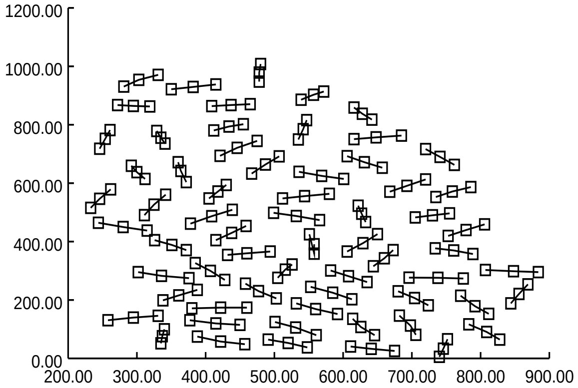

Figure 2: The Daughter Nuclei after a Round of Divisions

Our biological collaborator had been interested in a phase of embryonic development in which the nuclei (cells that have not yet acquired a membrane) on the surface of the embryo undergo several divisions (or mitoses) seemingly to fill up the space with nuclei as much as possible, in preparation for acquiring their membranes and later assuming their fated roles in the organism. Figure 1 shows (abstracted) image data just before a round of divisions, and Figure 2 superimposes the original nucleus and the two daughter nuclei that result from each division. In their search for any patterns that might govern this process, developmental biologists

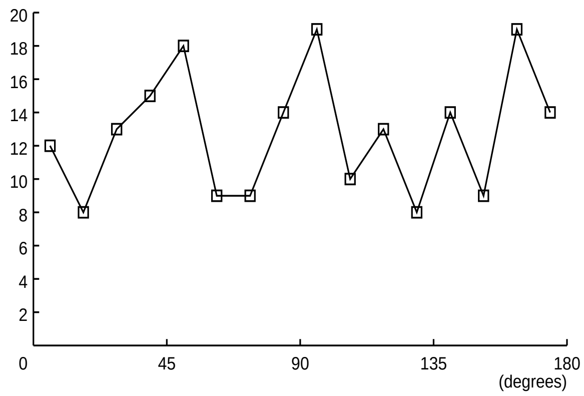

Figure 3: Histogram of Angles of Nuclear Division

had plotted histograms of the angles at which nuclei divide. For example, the angles observed in one set of experimental data are plotted as a histogram in Figure 3. No simple pattern is evident in such histograms, which led experimenters to infer that the process of nuclear division is disordered or patternless.

Our observations of these imaging data led to the following inferences. If the division angles were truly random, then by chance some localized collisions or crowding would occur. However, observations of these and other data indicate that excessive crowding *does not occur*. That is, in the face of a growing population, the nuclei seem to maintain respectable distances from their neighboring nuclei. Hence, one concludes that *there is complex patterned behavior*: division angles are not random, but are causally influenced by some factor not yet identified, which acts to alleviate the potential crowding. The schematic form of these inferences is *modus tollens*: A $\Rightarrow$ B, but $\neg$ B, hence $\neg$ A.

After detecting that an underlying pattern or regularity was present, the subsequent stages of this research involved hypothesizing an abstract causal mechanism (expressed computationally) that could give rise to the pattern, testing and confirming the hypothesis against observations, and finally conjecturing a biological mechanism that could "implement" the abstract, computational mechanism [Valdes and Minden]. These subsequent stages are not the focus of this paper, but it is valuable to point them out as typical follow-ups to the discovery of patterned behavior. Simon has examined these stages more generally in an early paper [Simon, 1977] on discovering and explaining empirical regularities, and in fact his paper served as a conscious normative guide for the latter stages of this work.

After completing the above work, we posed the following question: how could a computer program accomplish what we did, namely the discovery of a pattern, and yet do it in a general manner so that patterns from other scientific datasets could be successfully handled? The next section describes how we succeeded in purifying and generalizing our reasoning.

3 A HEURISTIC FOR PATTERN DISCOVERY

To summarize our reasoning on the embryological data, our own human inferences were:

1. if random, then crowding
2. observe a lack of crowding
3. conclude not random

Knowledge of the implication in step 1 is commonly acquired by tertiary education in science or engineering: random movements give rise to disorder and collisions. The second inference was made simply by observing that the divided nuclei maintain roughly equal spacing between neighbors. The third inference is an instance of *modus tollens*. A program capable of making these inferences would need to define algorithmically what "random" means and how to detect "crowding." Humanly, the latter was done by visual inspection, but a practical program would need some other means more suitable to machine inference. However, the best course would be to identify a general heuristic of which the application to nuclear divisions was only an instance, so that a broader class of data and patterns could be handled. We believe to have succeeded: the following sections develop this heuristic in the general case. Before describing the heuristic, we will need to introduce several concepts.

3.1 PROCESSES AND PARAMETERS

Let us consider that the source of the data is a transition from a time τ to a next time τ'; multiple transitions could be handled by considering each transition separately. Between τ and the subsequent time τ' there are a known set of processes that some entities undergo. For example, in the case of nuclear divisions, the processes of interest are of the schematic form $A \rightarrow B \,\&\, C$, in which one entity disappears by begetting two new entities. The converse process would be fusion, and is of the schematic form $A \,\&\, B \rightarrow C$. Table 1 lists a number of processes that occur in scientific studies. These processes were generated by drawing on our own (limited) knowledge; there is certainly room for others (and for multiple combinations) but these serve to illustrate the scope of the heuristic.

There are parameters associated with each process that specify exactly the effect of a process instance. For example, the division of an object into two offspring is described by two parameters: the angle of the segment joining the two offspring, and the distance between the two offspring. If the division is not exactly symmetric, then more parameters may be introduced to describe the relation of each offspring to the parent; in this paper we assume that the process of division is symmetric. An additional possible parameter is *time*, in cases where temporal distinctions within a transition are warranted; this enables representing processes that are strictly consecutive rather than concurrent.

Table 1: Some Common Processes in Science

Process	Schema	Parameters
division	$A \rightarrow B\,\&\,C$	angle, distance(s)
fusion	$A\,\&\,B \rightarrow C$	angle, distance(s)
death	$A \rightarrow \emptyset$	location
birth	$\emptyset \rightarrow A$	location
translation	$A(x, y) \rightarrow A(x', y')$	angle, distance
growth	$A(\theta) \rightarrow A(\theta')$	percentage change (or magnitude)

3.2 QUANTITIES

Before describing the heuristic, we need to introduce the notion of a quantity, which will, in the case of the mitosis data, make the connection to "crowding." A quantity is any statistic or calculation that is carried out on the data during or after a transition, and which yields one or more numbers. For example, a quantity relevant to crowding in the mitosis data is the distance to nearest non-sibling neighbor, since a preponderance of short distances implies crowding. Thus, a transition could give rise to a number of distance quantities, one for each nucleus offspring. Or, a quantity could perhaps yield a single number for each of many transitions, e.g., by calculating the area of the convex hull of all planar entities after each transition, or some other global quantity, one per transition.

3.3 RANDOMIZATION

Above we saw that a process has several associated parameters that characterize exactly any instance of the process. For example, the division $A \rightarrow B\,\&\,C$ may have the two parameters *angle-of-division* and *distance-between-siblings*. Let us consider a process that takes place from a time τ to the next time τ'. Each process instance involves a specific value of a process parameter, hence there is a pairwise correspondence between instances and parameter values, which is directly determined from the experimental or observational data.

We then define a randomization of the observed pairwise correspondence to be a random permutation or shuffling of the pairs, such that given N distinct parameter values, the probability that a process is assigned a given value is 1/N. The intent is that after a randomization, the quantities of Section 3.2 are re-computed using the artificially generated process/parameter combinations. For example, in nuclear division, the distance to nearest neighbor can be computed by keeping the observed inter-sibling distances, but using the shuffled angles of division to generate by computer simulation new artificial "observations" at time τ'. The artificial data are created by simulating the processes under a shuffled assignment of parameter values. Notably, these artificial data will lack the patterning (if any) of the original data, which has been destroyed by the shuffling.

3.4 THE HEURISTIC

We are now ready to state the heuristic. Given a process of known associated parameters that operates on entities during a transition, calculate some quantity over all the entities and then form (or plot) the empirical distribution of the quantity values. Then, randomize the process parameter values, simulate its effects, re-calculate the specified quantities in the new situation, and again form (or plot) the "randomized" distribution of the quantities. If the empirical distribution is sufficiently different from the "randomized" distribution, then conclude that *there is patterned behavior* underlying the process. Furthermore, a focused research problem is immediately suggested: look for a causal mechanism that can give rise to the pattern.

For example, if the mean of the empirical distribution is significantly larger than the mean of the randomized distribution, then suggest that the researcher/user look for a causal mechanism that would tend to *increase* the specified quantity. Or, if the means are similar, but the variance of the randomized distribution is larger, then look for a causal mechanism that would decrease the variability in the specified quantity.

There are four things that the heuristic needs to define for any specific application: the process(es), the parameters, the parameter(s) to be randomized, and the quantity. Typically, the first two are known from the scientific background. For example, if one is studying cell division, then the fact that one is studying a division process is clear and requires no new scientific insight. The parameters that characterize each type of process can be stored within a program as generic process descriptors. On the other hand, such a program should have knowledge of several commonly interesting quantities to calculate, such as distance, angle, and so on, and should perform a search among these quantities while looking for a pattern.

3.5 REDISCOVERING THE PATTERNS OF NUCLEAR DIVISION

We now illustrate the machine-oriented heuristic by applying it to the same data from developmental biology that we had studied earlier (the data correspond to the divisions that occur during a transition similar to the transition from Figure 1 to Figure 2, except that more nuclei are involved). The histogrammed frequency distribution of the quantity *distance-to-nearest-nonsibling-neighbor* is shown as the curve through rectangular points in Figure 4. Randomizing the correspondence of observed parameter values (angles) to process instances and then simulating the results, one obtains in the same Figure 4 the superimposed frequency distribution (through the crossed points) of the same *distance-to-nearest-nonsibling-neighbor*.[1] In this pa-

[1] The distances to nearest neighbor are not computed for points on the periphery, since full information on their neighbors is not available. The peripheral points are determined by stripping off one or more convex hull layers [Preparata and Shamos, 1988].

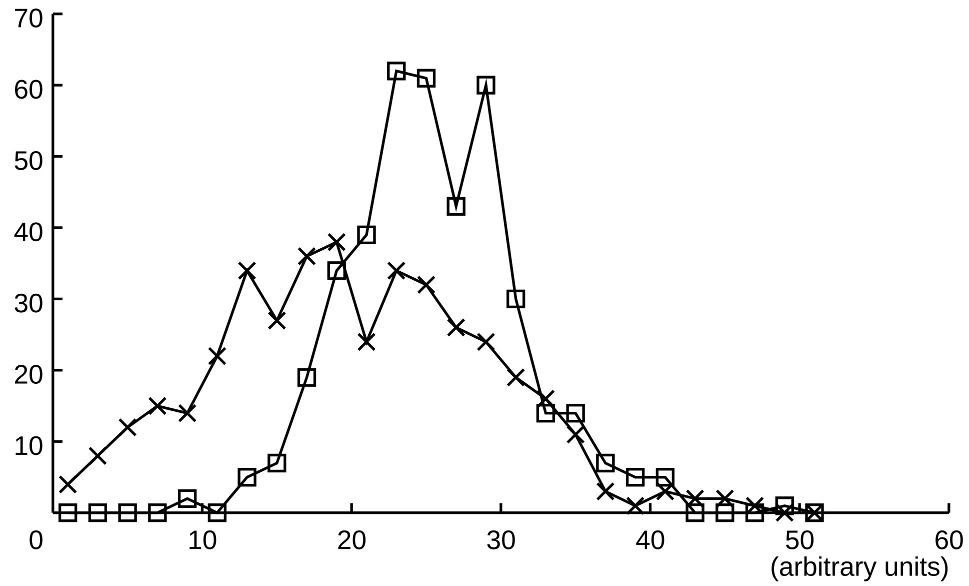

Figure 4: Distributions of Inter-Nuclear Distances in Observed (□) and Randomized (X) Cases

per, rectangular points in plots will always signify the empirical data, and crossed points the randomized data.

Visually, the two distributions in Figure 4 are clearly dissimilar (due to discrepant means); since the number of samples is in the hundreds, a confident judgment can be made. The statistic described in Section 3.7 can quantify this dissimilarity judgment, but inspection suffices in this case to reject the null hypothesis that the two distributions reflect the same underlying behavior. Hence, one can conclude that there is patterned behavior, and moreover that some causal influence acts to increase the mean *distance-to-nearest-nonsibling-neighbor*.

A method that reports patterns always and everywhere is of doubtful utility; our heuristic can indeed fail to find patterned behavior. For example, an attempt to find patterned behavior using the same quantity as above, but randomizing the distances between sibling nuclei, which vary, does not reveal significant distributional differences, and this visual judgment is confirmed by the statistic discussed in Section 3.7. No pattern is revealed because, apparently, there is no causal influence on the intersibling distances which affects crowding.

Curiously, the heuristic does not handle some simple patterns that are readily apparent to the eye, and that are susceptible to simple statistics or plotting. For example, if the angles of nuclear division were uniformly 43°, then there would be a clear pattern. However, this heuristic could not detect it, since randomizing the parameter values (all of which are 43°) would have no effect. This is why the heuristic is said to detect *complex* patterned behavior.

3.6 JUSTIFYING THE HEURISTIC

The heuristic is complex enough that its soundness may not be immediately evident. We can justify the heuristic by a couple of thought experiments. First, let us imagine a single process that is characterized by a single parameter. Assume that the parameter values for the process instances are generated randomly by a uniform distribution over a finite interval. Then, some quantity is calculated for each situation that results from a process instance. If we now randomize the parameter values and simulate their effects, we should expect (and obtain) no significant difference between the empirical and randomized distributions, since both are the result of similar random instigations.

As a second thought experiment, we now imagine that the parameter values are causally influenced by some factor that reveals itself through the quantity to be calculated. Then, we can expect that randomizing the parameter values will undermine the basis for the causal influence and destroy the patterning, and that the resulting randomized distribution will not resemble the empirical distribution of the calculated quantity. Thus, the presence of patterned behavior is detected.

We add that if the causal influence in the second thought experiment is due to factors completely unrelated to the calculated quantity, then no evidence for patterned behavior will be detected. This is not unreasonable, since one cannot experimentally detect a pattern unless one's measurements are sensitive to manifestations of the pattern.

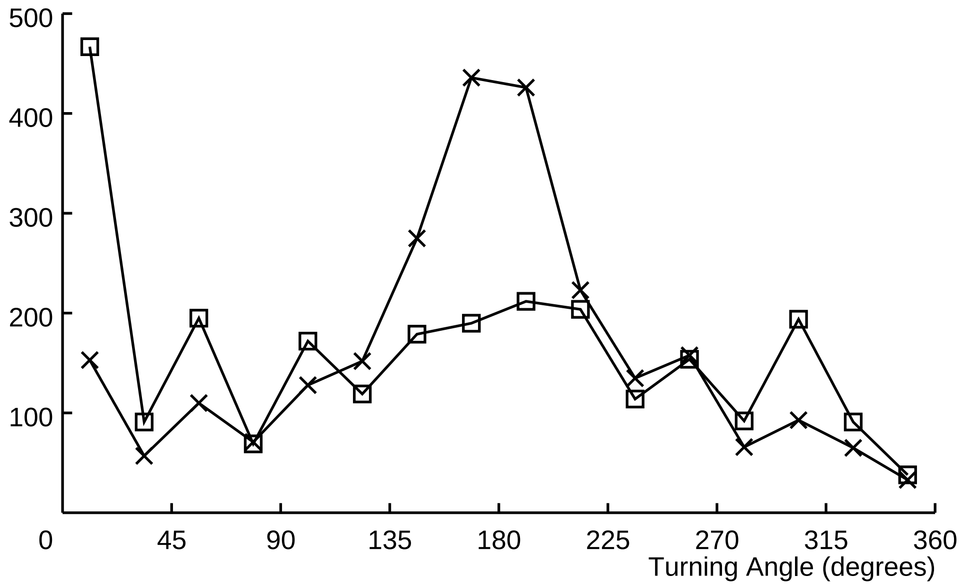

Figure 5: Turning Angles between Successive Placements (Observed □ and Randomized X)

3.7 A MEASURE OF DIFFERENCE BETWEEN TWO DISTRIBUTIONS

To complete the design of the heuristic, we need a measure of the dissimilarity between two distributions. We have referred above to an empirical distribution and to a "randomized" one. However, in statistical terms, we need a test to compare two specific distributions, i.e., obtained by measurement and simulation. Further, it is better not to assume anything about the underlying "true" nature of the distributions, since we want the heuristic to be generally applicable with as little prior knowledge as one can get away with and still have a powerful heuristic.

A test from nonparametric statistics provides the solution: the symmetric two-sample Kolmogorov-Smirnov test [Conover, 1980]. This test serves to answer the following question: Are two samples plausibly generated from a same underlying distribution? The test is sensitive both to discrepant means and shapes, and works as follows. First, convert each sample to a cumulative frequency distribution. Then, determine the maximum vertical difference between the two superimposed cumulative distributions. If this vertical difference is large and enough data are available for a sound judgment, then the null hypothesis of identical distributions is rejected at some confidence level.

4 AN APPLICATION TO PSYCHOLOGY

The Lisp program PENCHANT is a prototype, but it was able to rediscover the above pattern of mitoses in the developing embryo. The program's heuristic is of quite general applicability since, for example, its ability to uncover patterned behavior is not limited by what is contained in a stored catalogue of behavioral patterns. To give concrete evidence for the heuristic's (and program's) generality, we describe an application to experimental cognitive psychology.

4.1 MEMORY IN CHESS

A traditional experiment in the psychology of chess is as follows. A chess position is displayed to a human subject, often a chess master, for a few seconds. Then, the display is withdrawn and the subject attempts to reconstruct the exact position that was shown. Such experiments are relevant to test the theory that chess expertise consists partly of the ability to recognize and recall many chunks, which are defined as frequently seen configurations of several chess pieces. For example, a chunk recognizable by most any player is a castled king flanked by a rook and nestled behind three adjacent pawns.

We have applied the heuristic to experimental data collected over several years by Gobet and Simon [Gobet and Simon]. As with any application of the heuristic, one needs to identify four things: the processes involved, the parameters, the parameter to randomize, and the quantity to be computed.

In this case, the relevant process consists of the subject replacing a chess piece on a square of the chess board. In terms of the categories in Table 1, this process corresponds to a birth, in which a new piece is "born" on the chess board. The parameters that characterize birth instances can be the identity of the newborn (e.g., black bishop), the location (i.e., square on the chess board), and the time instant (i.e., the first placement is at time 1, the second at time 2, etc.). One good choice for the quantity to compute is the euclidean distance between one placement and its temporally nearest neighbor (i.e., subsequent placement). Another quantity can be the angle between two temporally nearest neighbors. In both cases, one discards the piece's identity. Randomization of the process parameter value *time* is, in this application, equivalent to random shuffling of the order of piece placements.

The observed and randomized distributions of distance between current and last placements reveal a pattern: subjects tend to place successive pieces adjacently. This behavioral pattern was not surprising to the expert (Gobet), since it fits well with the chunking hypothesis, and the expert had already carried out analysis on successive distances. We then calculated the angle between the last and current placements, rather than the distance. In this case, we uncovered patterned behavior that had not been noticed: subjects show a tendency to horizontal successive placements (at angles of 0° or 180°), but do not show the same tendency in the vertical, and this behavioral pattern is destroyed by randomization.

Encouraged by the novel results with the angle quantity, we then calculated the "turning angle," which is the relative angle within a triple p_1,p_2,p_3 of placements thus: $angle(p_2, p_3) - angle(p_1, p_2)$. In this case, randomization was only within placements of pieces belonging to the same chunk. Figure 5 reveals a pronounced tendency to place pieces successively in the same direction (0° turning angle), although not necessarily horizontally. Once again, this pattern had not been noticed by the experimenters.

5 DISCUSSION

One lesson from the work on BACON [Langley et al., 1987] and its successors is that a single heuristic can by itself be quite powerful in making significant discoveries in scientific data. The concept of "data-driven discovery" describes such cases, in which little theoretical knowledge is employed; most of the power comes from the data and from the typically few heuristics of broad generality. The concept of "theory-driven discovery" describes a complementary activity in science, in which significant theoretical reasoning and assumptions are involved in problem-solving; typically this activity occurs in rather mature sciences that have accumulated significant knowledge about an object of study. We may cite DENDRAL [Lindsay et al., 1993], GRAFFITI [Fajtlowicz, 1988], and MECHEM [Valdes, 1994, Valdes, 1992] as just three examples of theory-driven discovery systems that address significant problems from science and mathematics.

We suggest that the present PENCHANT program is a significant contribution to knowledge of data-driven scientific discovery. The current program is rather unlike the programs described in the book by Langley, Simon, Bradshaw, and Zytkow [Langley et al., 1987] because PENCHANT does not report laws. Rather, the program detects patterned behavior; these patterns immediately suggest the research problem of explaining the patterns, and suggest some hints about what explanatory mechanisms to look for. PENCHANT is not proposed as a plausible cognitive model of anything, but as a human/machine discovery heuristic that is based on a rational, meta-scientific analysis of a discovery task.

We are not able to claim yet that the heuristic has made any significant new discoveries, since we have only applied it to make a rediscovery in developmental biology and to make minor discoveries in the psychology of chess. However, the heuristic follows very closely on the heels of a (modest) scientific discovery, and works on the very same data that we ourselves used in our field work. We are actively seeking further applications of the heuristic.

5.1 KNOWLEDGE AND CREATIVITY

In their report on a AAAI Symposium on AI and Creativity, Kim, Dartnall, and Sudweeks [Kim et al., 1993] revealed that the symposiasts reached agreement on the proposition that "deep knowledge is critical for creativity." We are in sharp disagreement with this view, and believe that the results of this paper constitute further evidence that scientific creativity can be displayed by a program (or a scientist) that has only a very shallow knowledge of a subject matter. The earlier BACON program, together with its relatives and successors, constitute quite good evidence that, at least for law discovery, "naive" data-driven methods having little theoretical knowledge can be powerfully creative.

PENCHANT has minimal knowledge of the subject of developmental biology, and yet is able to reproduce a recent human discovery which itself was made only with some quite general heuristics but little domain knowledge. The power of the program depends on having sufficient data and a powerful heuristic of broad applicability.

The issue of whether deep knowledge is critical for creativity, including scientific creativity, is not only an issue of philosophical interest, it also has practical methodological implications. If it were largely true, for example, it would imply that to build successful machine discovery systems must be an arduous task, involving the codification of much scientific knowledge, or that such systems in practice will fill only very limited creative roles in joint human-machine discovery systems. Even more broadly, it would imply that crossing interdisciplinary lines to make significant scientific contributions is futile because of the deep knowledge store that would be newly required of the human researcher. Hence, it is important to test and refute this proposition with evidence such as presented here, in which a heuristic method, drawing on very limited knowledge, can make significant findings of patterned behavior in science.

5.2 RELATION TO STATISTICS

The heuristic bears some resemblance to randomization tests in statistics, which are surveyed and defined by Edgington [Edgington, 1980] as "procedures for determining statistical significance directly from experimental data without recourse to significance tables." Such procedures involve repeatedly permuting the experimental data (e.g., response of subjects to one of two treatments) and calculating a test statistic (e.g., T test). If the statistic calculated on the original data is extreme within all the cases, permuted or not, then the effect is judged significant.

The resemblance to our heuristic lies in the use of randomization to create derived data. However, there are several differences: First, rather than calculate a simple test statistic on one arrangement of the data, our "test statistic" is a comparison between two distributions, the original and the randomized, which moreover are somewhat removed (due to computer simulation) from the parameter being randomized. Second, the heuristic is grounded, by means of the four questions that it asks, in the context of science terminology and concepts, making it easier to perceive its applications within scientific practice.

6 CONCLUSION

This paper has reported the invention of a powerful, machine-oriented heuristic for the discovery of complex patterned behavior in empirical data. The heuristic was extracted from our own "field work" in developmental biology, and resulted from posing the question of how a computer could notice the same patterned behavior we had noticed, but in a general and machine-oriented way. This experience suggests that field work in some specific science can be a fruitful technique for researchers in computational scientific discovery.

The heuristic was applied to re-discover the pattern that we had noticed, which was not a trivial exercise since the heuristic is not altogether similar to the human-oriented inference we had employed. We also applied the heuristic to psychological data on memory in chess, and found patterns that had not been articulated by researchers in that area. In both applications of the heuristic, the phenomena (developing embryos and memory in chess) have been studied for some years, hence the heuristic's accomplishments are not largely due to having exclusive access to new data or phenomena. Further applications of the heuristic are in progress.

Acknowledgements

RVP was supported partly by a Science and Technology Center grant from the National Science Foundation, #BIR-8920118, by the W.M. Keck Center for Advanced Training in Computational Biology, and by a High Performance Computing and Communications grant from the National Science Foundation, #ASC-9217091. AP was supported by a grant from the Comunidad de Madrid. Fernand Gobet provided his chess data and discussed with us the significance of the patterns uncovered by the heuristic. Jonathan Minden was our collaborator in developmental biology. David Banks helped us to choose the statistical tests.

References

[Conover, 1980] Conover, W. (1980). *Practical Nonparametric Statistics*. John Wiley & Sons, New York. 2nd Edition.

[Dietterich and Michalski, 1985] Dietterich, T. G. and Michalski, R. S. (1985). Discovering patterns in sequences of events. *Artificial Intelligence*, 25:187–232.

[Edgington, 1980] Edgington, E. S. (1980). *Randomization Tests*. Marcel Dekker, New York.

[Fajtlowicz, 1988] Fajtlowicz, S. (1988). On conjectures of Graffiti. *Discrete Mathematics*, 72:113–118.

[Fischer and Zytkow, 1990] Fischer, P. and Zytkow, J. M. (1990). Discovering quarks and hidden structure. In Ras, Z., Zemankova, M., and Emrich, M., editors, *Proceedings of the Fifth International Symposium on Methodologies for Intelligent Systems*, pages 362–370. North Holland.

[Gobet and Simon] Gobet, F. and Simon, H. A. Expert chess memory: Revisiting the chunking hypothesis. Submitted for publication.

[Godfrey, 1988] Godfrey, D. (1988). A hexagonal feature around Saturn's north pole. *Icarus*, 76(2):335–356.

[Karp, 1993] Karp, P. D. (1993). Design methods for scientific hypothesis formation and their application to molecular biology. *Machine Learning*.

[Kim et al., 1993] Kim, S., Dartnall, T., and Sudweeks, F. (1993). AAAI 1993 spring symposium series reports. *AI Magazine*, 14(3):32. AI and Creativity.

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

[Laird, 1992] Laird, P. (1992). Discrete sequence prediction and its applications. In *Proceedings of the 10th National Conference on Artificial Intelligence*, pages 135–146, Menlo Park, CA. American Association for Artificial Intelligence.

[Langley et al., 1987] Langley, P., Simon, H., Bradshaw, G., and Zytkow, J. (1987). *Scientific Discovery: Computational Explorations of the Creative Processes*. MIT Press, Cambridge, Mass.

[Langley and Zytkow, 1989] Langley, P. and Zytkow, J. (1989). Data-driven approaches to empirical discovery. *Artificial Intelligence*, 40(1-3):283–312.

[Ledesma et al., 1994] Ledesma, L., Perez, A., Borrajo, D., and Laita, L. (1994). La lógica de Boole como consecuencia del método de separacíon de símbolos de Gregory-Boole. Estudio histórico y su emulacíon por computador. In *Actas del Congreso Hispano-Frances sobre Historia de la Ciencia*, pages 289–301, Madrid. C.S.I.C.

[Ledesma et al., 1993] Ledesma, L., Perez, A., Laita, L., and Borrajo, D. (1993). Descubrimiento científico e inteligencia artificial. In *Real Academia de Ciencias*

Exactas Físicas y Naturales: Segundo Curso de Conferencias sobre Inteligencia Artificial, pages 115–132, Madrid, Spain.

[Lindsay et al., 1980] Lindsay, R., Buchanan, B., Feigenbaum, E., and Lederberg, J. (1980). *Applications of Artificial Intelligence for Organic Chemistry: The Dendral Project*. McGraw Hill, New York.

[Lindsay et al., 1993] Lindsay, R., Buchanan, B., Feigenbaum, E., and Lederberg, J. (1993). DENDRAL: a case study of the first expert system for scientific hypothesis formation. *Artificial Intelligence*, 61(2):209–261.

[Melton, 1991] Melton, D. (1991). Pattern formation during animal development. *Science*, 252:234–241.

[Oliver, 1991] Oliver, J. E. (1991). *The Incomplete Guide to the Art of Discovery*. Columbia University Press, New York.

[Preparata and Shamos, 1988] Preparata, F. P. and Shamos, M. I. (1988). *Computational Geometry: An Introduction*. Springer Verlag, New York.

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

[Simon, 1977] Simon, H. (1977). On judging the plausibility of theories. In *Models of Discovery*, pages 25–45. Reidel, Boston.

[Valdes, in press] Valdes-Perez, R. E. Algebraic reasoning about reactions: Discovery of conserved properties in particle physics. *Machine Learning*. in press.

[Valdes, 1992] Valdes-Perez, R. E. (1992). Theory-driven discovery of reaction pathways in the MECHEM system. In *Proceedings of 10th National Conference on Artificial Intelligence*, pages 63–69.

[Valdes, 1994] Valdes-Perez, R. E. (1994). Conjecturing hidden entities via simplicity and conservation laws: Machine discovery in chemistry. *Artificial Intelligence*, 65(2):247–280.

[Valdes and Minden] Valdes-Perez, R. E. and Minden, J. S. *Drosophila melanogaster* syncytial nuclear divisions are patterned: Time-lapse images, hypothesis, and computational evidence. *Journal of Cell Biology*. Submitted for publication.

[Valdes et al., 1993] Valdes-Perez, R. E., Zytkow, J. M., and Simon, H. A. (1993). Scientific model-building as search in matrix spaces. In *Proceedings of 11th National Conference on Artificial Intelligence*, pages 472–478.

[Zytkow and Simon, 1988] Zytkow, J. and Simon, H. (1988). Normative systems of discovery and logic of search. *Synthese*, 74:65–90.