

Testing Computer Models of Scientific Discovery

Deepak Kulkarni and Herbert A. Simon
NASA and Carnegie Mellon University

Laboratory notebooks of scientists, used in conjunction with published papers, retrospective accounts and other sources, are rich sources of information for theories of scientific discovery. In a few cases, such records have provided the basis for computer models of the processes of experimental science. As one example, the records of Hans Krebs' discovery of the ornithine cycle of *in vivo* synthesis of urea have been used to test two distinct simulations, KEKADA and the Causal Discovery Program (CDP), respectively. What do we learn from these experiments in modeling? In particular, what, if anything, do such models add to the lessons we can learn directly from our study of the historical sources themselves.

The kind of modeling with which we will be concerned is descriptive and explanatory, not normative. We are interested in simulating actual instances of discovery, not in testing norms of how discovery should be done. Although the latter is a legitimate objective of modeling; it is not our topic here.

Computer Simulation

Computer simulation of human thought processes aims at creating a system that, with appropriate inputs, will trace the approximate path taken by someone seeking to solve a problem. To see what this implies for the simulation of scientific discovery, we must describe what we mean by "appropriate inputs," "approximate," and "path."

Inputs. The inputs required for a simulation consist of the scientist's relevant knowledge, including both general processes, and knowledge and procedures that are specific to the domain of the problem. The simulation must incorporate processes for observing and theorizing in the ways that scientists do, knowledge of relevant subject matter, specialized observational and theorizing procedures usable in that science, and a goal, or a way of creating a goal from the scientist's knowledge.

The information given to the program may come from sources that are known prior to and independently of the data to be simulated, or may be derived from the data themselves. In the latter case, we can view this information as a set of parameters that have been estimated from the data, and which therefore reduce the number of degrees of freedom available for testing the goodness of fit of the model. Information provided to the program that is known independently of the data to be simulated does not reduce the degrees of freedom if incorporated in the program in an unselective and unbiased manner.

If the data to be simulated are known at the time the model is constructed, it may be difficult or even impossible to separate clearly the independent information from the data-derived information. One (partial) solution to this problem is to use the data from one case to construct and debug the simulation model and then to test it with data from other cases. This was done with both of the simulation models discussed in this paper.

Approximation. Any particular simulation model will only reproduce the phenomena of interest down to some level of detail, which can often be characterized by the (simulated) times required to execute the primitive processes. We may call this the temporal resolution of the simulation. Similarly, tests of goodness of fit can only be carried down to the level of the temporal resolution of the empirical data. In cases where thinking-aloud protocols are obtained from subjects, the temporal resolution of the data may be of the order of a few seconds; when scientists' laboratory notebooks provide the basic data, the temporal resolution is usually not less than a day.

If the data and the trace of the simulation have been segmented into items at a common level of resolution, then the corresponding items of the two data streams can be compared. In the simplest case they can be coded as "same" or "different," and the goodness of fit measured by the percentage that are the same. More elaborate schemes are also possible. Items can be weighted for their "importance" to the process; and the degree of similarity of each pair of items can be scored. Gross numerical evidence of goodness of fit is not of great interest, beyond assuring us that the simulation performs better than chance. Determining the nature of the discrepancies and misfits and their probable causes is more useful for improving theory.

Path. As people generally have reasons for the steps they take along a problem-solving path, the time when a particular step occurs depends on what has already transpired. Hence, the order of events is of the essence

in modeling, and we must consider the possibility that the process is dynamically unstable or chaotic: that no matter how closely the model approximates the actual human process, small differences between them will cumulate, and any slight divergence in their paths will grow steadily larger, so that all but the first few pairs of events will be coded as "different." The study of unstable and chaotic systems must proceed quite differently from the study of dynamically stable systems. Simple comparison of time series simply doesn't work for the former.

There are several ways, not necessarily exclusive, of dealing with divergence. First, when the two paths diverge, the simulation model may be "reset" (the state of the system adjusted) to put it back on the path of the human behavior stream. Then, goodness of fit may be measured by the average length of path between resettings: by how long, on average, the simulation tracks the behavior before the one diverges from the other. The degree of divergence that triggers resetting is a parameter. Meteorology, for example, evaluates the quality of its predictions by the length of time over which they do better than chance.

Second, the systems we are considering are goal-seeking systems, having the property called equifinality: In general, they continue to search until they attain the goal, or until search is terminated by the user. Hence, whenever both program and scientist are successful, the program trace and the scientist's protocol will finally converge to a common state. In this case, instead of resetting the program periodically to test how well it matches the data, we may simply let it run to conclusion. If it reaches the goal, we can evaluate the fit by comparing the regions it has visited during its search with the regions visited by the human scientist, and can pay secondary attention to the order in which they were visited. Subgoals achieved by both scientist and program can also be points of convergence.

Roughly speaking, the second method tests whether human scientist and simulation attend to the same things in seeking to solve the problem; the first method tests whether they assign the same priorities, hence examine different possibilities in the same order. Of course, matters are more complicated than this, for information revealed during the search itself commonly turns the path in new directions, and different searches will reveal different information at different times.

Level of Generality. There are several different senses in which we can test a model. First, we may compare it with the behavior of the particular scientist whose data stream is being matched. Alternatively, we may test it as a more general theory of scientific discovery. There are also

all sorts of intermediate possibilities: we may test it, for example, as a theory of scientific discovery in a particular domain: biochemistry, say. In this way, some components of the program may describe idiosyncratic elements of a scientist's style or knowledge, others may describe "methods of doing research in biochemistry," still others, aspects of "the scientific method."

If our aims are general, we fit the model to historical data as a means of advancing the sociology and psychology of science. If we wish to understand the individual case, the links between the model and history draw closer; the empirical methods of the historian provide the basic data, while the model builder tests the consistency and completeness of the historical account against a broader collection of ideas about the scientific process. We would conjecture that, at least for some time to come, modeling will have its chief value in allowing us to draw, in an unusually disciplined and rigorous way, on historical case studies as important sources of evidence for generating and testing theories of scientific method — both general and domain-specialized theories. They may, however, also help the historian to examine the completeness and consistency of his evidence for the causal links in his story of discovery. In particular — and we will elaborate on this point later — they may help the historian to examine "the paths not taken."

Implications for Historical Explanation. These methodological remarks have implications that go beyond formal modeling and computer simulation of time series. As they address the logic of explanation and of the verification of explanations in general terms, they apply with equal force to all efforts at accounting for sequences of historical events, whatever methods are used in the analysis. They imply, for example, that to account for human actions, we must have information, not just about actors' goals, but also about what the actors know and believe about the world, and the way in which they represent it.

What distinguishes computer modeling and simulation from other methods is simply the particular techniques used to assure rigor in the specification of goals, knowledge and beliefs, representations and thought processes, and the technical means available for deducing the implications of these specifications for the course of events. As with other methods, their effectiveness depends critically upon the richness and reliability of the source data upon which the analysis draws.

The Discovery of the Ornithine Cycle.

A notable example of archives recording the steps in an important scientific discovery are the notebooks of Hans Krebs and his research assistant Kurt Henseleit describing the experiments that led them to the discovery of the ornithine cycle of urea synthesis. This was a discovery of first importance that revealed a cyclical catalytic process of a kind that was relatively unfamiliar at that time. The notebooks, supplemented by the published papers and by interviews of Krebs by the science historian Frederic L. Holmes (1980, 1991), provide a day-by-day account of the course of the discovery.

Two computer models of experimental strategy have been tested against the record of the discovery of the ornithine cycle. One is the KEKADA system (Kulkarni, 1988, Kulkarni & Simon, 1988); the other is the CDP system (Grasshoff & May, 1995; Grasshoff, 1995). Each system was designed as a model of experimental science, and not just as a specific model of the ornithine cycle discovery. Because the models and their explanations of the ornithine discovery differ in important respects, they provide a useful starting point for an inquiry into the modeling process. As a basis for the inquiry we must first provide a resumé of the discovery, as recorded in the laboratory notebooks and analysed and recounted by Holmes.

Conditions for the Reactions. Modes of biochemical experimentation that preserve cells intact permit intracellular structures and molecules contained in the cell to play their customary roles in the chemical reactions that take place there. Many experiments conducted in the presence of intact cells produce results quite different from otherwise identical experiments that are conducted after the cells have been destroyed. Several methods were devised by biologists to observe the reactions that occur in the living cell without requiring a whole organism as the site: for example, (a) the use of isolated organs continually perfused with Ringer's solution or some equivalent liquid to maintain their functioning, (b) the use of similarly treated tissue slices, and (c) the use of tissue ground coarsely enough to maintain numerous cells intact.

For a long time the requirement that intact cells be present in order to reproduce the reactions observed in living organisms served as an argument for vitalism. It was equally possible, however, to account for the dependence of reactions upon the viability of the cells in a non-vitalistic way: simply as implying that the reactions required substances and sites that were closely attached to the cells and were largely destroyed or dissipated when the cells were destroyed. The shady boundary between these two interpretations is nicely illustrated by a quotation,

which Holmes reproduces in his biography of Krebs, from the paper reporting the synthesis path for urea (pp. 35 and 53, respectively, of Krebs and Henseleit, (1932c)):

"since all essential metabolic phenomena are bound to the cell structure, the tissue *Brei* [finely mashed broth of tissue], used so often in the past -- in which the structure is destroyed -- is unsuitable for metabolic investigations."

Krebs himself, in a 1978 conversation with Holmes, described the statement as "an empirical assertion and not a philosophical viewpoint," but it is surely ambiguous on the issue of vitalism. With today's hindsight, we know that the dependence of reactions on the intact cell is due to the involvement in them of substances in the cell, especially co-enzymes and energy-rich radicals like ATP and ADP, that were, at the time of the urea research, largely unknown.

Viewing the matter from a modern, and non-vitalistic, standpoint, we can say that carrying out experiments in the presence of intact cells, by any of the methods listed above, allows the experimenter to remain ignorant of or to ignore substances that are essential components of the reactions under study but that are normally present in the cells: for example, essential enzymes and co-enzymes. Thus, as urea is normally synthesized in the mammalian liver, an experimenter can focus on supplying the substances that he hypothesizes will provide the urea nitrogen without the need to supply all the substances required to catalyze the reactions.

Resumé of the Discovery. Toward the end of July, 1931, Hans Krebs set out to determine the sources of the nitrogen used to synthesize urea *in vivo*, and the chemical reaction path for the synthesis. The basic paradigm was to measure the rate of production of urea (per mg of tissue per hour) as a function of the substances present and other experimental conditions (e.g., temperature, pH). The chemical constitutions of all of the substances examined during the experiments were known, as was the site of the synthesis of urea: the liver. It was generally believed, though not fully demonstrated, that the amino acids (mainly derived from decomposition of proteins) were the principal ultimate source of the nitrogen in urea, and that ammonia or salts of ammonia might be intermediate products on the reaction path. The various amino acids were all thought to play essentially the same role of contributing nitrogen, directly or indirectly. Previous research also suggested that the state of nutrition of the laboratory animal could influence the rate of urea

production. (More urea would be produced by a well-fed animal than a starved animal.)

The experiments that Krebs and Henseleit performed can be divided into four groups. (1) From the end of July to the 13th of November, the experimental system was established, and about 88 experiments were run (each averaging about a half dozen conditions) without any very evident specific long-term plan. These experiments compared the rate of production of urea from ammonia with the rates from a substantial number of other substances, including, but not limited to, a half dozen amino acids and some metabolites (e.g., glucose). Holmes (1991, pp. 254-283) discusses in detail the motivations, known and conjectured, for these particular experiments. Only in the last three weeks of this period, were experiments run comparing ammonia alone with ammonia *plus* some other substance.

At the outset, Krebs spent several weeks setting up his apparatus and tuning his experimental procedure, which used the method of tissue slices that he had learned in Otto Warburg's laboratory. With this method, he replicated earlier experiments on urea synthesis, which had used perfused organs instead of tissue slices, perfected his measurement of the urea output, tested the urea outputs from various substances (ammonia and several amino acids), verified that urea was not produced at significant rates in organs other than the liver, and tested the increase or decrease of the synthesis in the presence of common metabolites like glucose. On August 4, a graduate student, Kurt Henseleit, had joined him as an assistant, and Krebs set Henseleit to learning how to operate in this experimental setting. Most of the subsequent experiments were performed by Henseleit under Krebs' direction.

In these initial experiments, Krebs discovered that when ammonia was supplied to tissue slices of liver, urea that accounted for nearly 2/3 of the nitrogen in the ammonia was synthesized quite rapidly. Why the rest of the ammonia was not transformed was not clear, but enough urea was produced to demonstrate that the enzymes and other conditions essential to the reaction were present in liver cells, but not in the cells of other organs, in substantial amounts.

However, when Krebs supplied various amino acids to the tissue, but without ammonia, very little urea was produced. (Alanine was a partial exception, although generally less urea was synthesized with it than with ammonia). Nearly the same results were also obtained when both ammonia and an amino acid were supplied as when only ammonia was

supplied: the rate of urea production was approximately that obtained from the ammonia alone.

During this initial period there was no *systematic* comparison of the differential effects of different amino acids, perhaps because they were thought of as alternative sources of the nitrogen in urea. Alanine, which was found to produce urea at rates nearly comparable to ammonia was the amino acid most often tested in these experiments. As the yields of urea from particular inputs or sets of inputs were generally compared, in any given experiment, with the yield from ammonia alone, the latter treatment may be thought of as the experimental "control" condition — an "effect" being a significantly larger yield in some condition than the yield from ammonia. The presumed rationale was that, if amino acids were first converted to ammonia, and the ammonia to urea, urea should be produced at a more rapid rate from the intermediate product, ammonia, than from the amino acid.

However, formal statistical tests were given little attention, and comparisons were often made between experimental conditions (the same or different treatments) in different experiments. The important unit of analysis was the experimental condition, not the "experiment." Experiments at first averaged about four conditions each; later on, they often contained as many as nine or ten conditions each.

The first 49 experiments (about 200 treatments), summarized on September 4, showed little more than that urea could be synthesized from ammonia in tissue slices and that over a span of time most, but not all, of the ammonia that had been added was converted to urea; that the addition of the three or four amino acids they tested yielded little urea, except for alanine, which sometimes yielded nearly as much as ammonia; and that none of these results were very sensitive to the other conditions they had varied. The results suggested strongly that the source of the nitrogen in urea produced in liver tissue was ammonia, not the amino acids, without casting any light on the reaction paths that might lead from proteins and amino acids to ammonia.

About 25 more experiments were run by the 20th of October, without notable findings. During the latter part of this period, as Krebs was heavily involved in several other research problems on which he had previously been active, the urea experiments were performed by Henseleit.

(2) On October 21, the yield of urea from ammonia was compared with that from arginine, and from several other amino acids. What motivated the experiment with arginine has not been established. It was well known to Krebs, however, that arginine is present in the liver in large quantities (and not in such quantities elsewhere), and that it is split in the presence of arginase into urea and ornithine, the latter being another, and relatively rare, amino acid of then unknown function found mainly in the liver. We could hypothesize, but without any direct evidence, that at this time Krebs was motivated to test any amine-containing molecule that was found specifically in the liver. (Arginine, called to his attention by the literature, and ornithine, called to his attention by the output of the arginine reaction, are the two that fit this description.)

In this experiment, the arginine produced urea *eight* times more rapidly than did ammonia. There is no indication, from the log book, Krebs' publications, or his latter recollections, that he inferred from this result that arginine was the source of the urea normally produced by the liver. If it were, the research task would then be to explain the sources of the arginine and the disposition of the ornithine, neither then being known.

On the following day, October 22, an experiment in which ammonia was tested both alone and in combination with potassium cyanate ruled out cyanic acid as an intermediate source of urea nitrogen. This negative result refuted an earlier hypothesis in the literature that the nitrogen might be contributed jointly by cyanate and ammonia.

It is possible, however, that this experiment suggested to Krebs that the nitrogen might come from a *combination* of ammonia with some other source, for, in the fourteen experiments (75-88) performed during the remainder of October and up to November 13, the urea yield with ammonia alone was, for the first time, compared with ammonia *plus* a large number of other substances, including a half dozen amino acids. (Holmes infers (p. 281) that Krebs was examining the hypothesis, like that which had been proposed for cyanate and ammonia, that one of the nitrogen atoms in urea might be donated by the ammonia, the other by an amino acid.) Only one of the experiments showed production of urea at a rate much in excess of that produced by ammonia alone. However, some connection was exhibited between the general level of metabolic activity and urea production, and this lead was followed up by experiments on metabolites that continued through November 13.

(3) On November 15, new tests were made on the effects of metabolites, but one other condition was introduced — ammonia in combination with the amino acid, ornithine. Ornithine was also tested alone, but no other amino acid was tested in this experiment. Essentially no urea was obtained from ornithine alone, but it was produced at about three times the rate from the combination of ammonia with ornithine as from ammonia alone. This effect elicited surprise, and turned out to be a key finding, leading to the determination of the source of the urea produced in the liver and the reaction path of the final synthesis. There is no evidence, from the log books, Krebs' papers, or his later recollections, as to what motivated this experiment or why it was performed at this time.

The surprising urea yield from ammonia plus ornithine (the "ornithine effect") led immediately to experiments, with negative results, on substances chemically related to ornithine. The criteria of similarity were the possession by these substances of (a) the same carbon skeleton as ornithine, or (b) the same amino groups. The inclusion of the first class of substances indicates that Krebs was considering the possibility that while ornithine or an equivalent was essential to the urea synthesis, it was not contributing the nitrogen: a possible antecedent to the hypothesis that ornithine was a catalyst. The second class of substances would fit the hypothesis that the amino groups of ornithine contributed nitrogen.

It is notable that Krebs' response to the surprising ornithine effect was quite analogous to Alexander Fleming's response to his surprise upon finding the mold *Penicillium* lysing bacteria in a Petri dish that had been left unwashed. Fleming's first response was to initiate experiments to determine the scope of the phenomena (what species of bacteria would be lysed; what molds would lyse them?); his second response was to seek a mechanism, a task later completed by Chain and Florey. A similar analogy can be seen in the Curies' response to the unexpectedly high density of radiation they detected in the pitchblende they were refining, and in a number of other important discoveries that began with a surprise. We will see later that this strategy of response to surprise is incorporated in the KEKADA program.

No experiments were run from November 17 to December 8, as the laboratory was being relocated in a new building, nor from December 18 to January 6, 1932, when the new laboratory was closed through the holidays. After resumption of work in January, increased attention was paid to the ratios of inputs to outputs of the ammonia and ornithine in the November 15 experiment under wide variations of the input quantities. Beginning January 14 new equipment (Parnas-Heller apparatus), which

had been ordered earlier but had just arrived, was used to measure, for the first time, the ammonia consumed; and the initial concentrations of ornithine were varied widely. It was found (with substantial variance in results) that, when ornithine was present, the urea production varied more or less proportionately with the quantity of ammonia consumed but its rate of production was considerably less sensitive to the quantity of ornithine present. Ammonia was consumed roughly in the ratio of 2 molecules for every molecule of urea produced, as stoichiometry required if ammonia were the sole source of the nitrogen in the urea.

These findings deflected the experimental program, from the search for the steps in urea synthesis that could lead from proteins to ammonia to a search for a reaction path that could convert the ammonia into urea. This redirection was supported (or perhaps even stimulated) by parallel experiments Krebs initiated on January 23, 1932, after Henseleit, following up on the ornithine effect and testing both liver and kidney tissues, discovered rapid deamination of amino acids in the kidney (quite independently of the presence of the ornithine). The researchers then focused upon the kidney as the site for production of most of the ammonia that was subsequently converted to urea in the liver, and separated research on the deamination process from research on the conversion of ammonia into urea.

During this period, Krebs gradually arrived at the definite conclusion that ornithine was not consumed in the production of urea, and that the ammonia was the source of the urea nitrogen. He also began to connect the ornithine effect with the arginine reaction, perhaps because both actions were concentrated in the liver (Holmes, 1991, p. 304). From stoichiometric considerations, he postulated a reaction converting ornithine, ammonia, and CO_2 into arginine, and a reaction converting arginine to urea and ornithine. Perhaps the hardest step here was to envision the cyclical role of ornithine. Instead of conceptualizing it as a template for the reaction of other substances, as catalysts were generally viewed, he initially viewed it as both an input to the first step and an output from the second step of a two-step reaction.

By April 13, Krebs was satisfied that he had identified the catalytic reaction: Ornithine combines with two molecules of ammonia to form arginine, and arginine splits to form urea and ornithine, which is thereby regenerated for reuse. This reaction, providing a plausible reaction path for *in vivo* urea synthesis from ammonia, also identified a source of the arginine that was found in the liver and explained the disposition of the ornithine — two previously unanswered questions. A paper reporting the

findings and their interpretation was then prepared and published on April 30, 1932 (Krebs & Henseleit, 1932).

(4) After April 15, a literature search and further experiments showed that another substance, citrulline, could serve as an intermediate in the reaction chain, and a second paper was published describing this more detailed reaction path. The research over the four periods had involved about 188 experiments, totaling about 1,000 conditions.

Interpretation of the Chronicle. A fully successful simulation of Krebs' experimental strategy would have to account for at least the following empirical phenomena: (a) the kinds of substances that were tested for their contribution to or influence upon the yield of urea, and the times at which and order in which they were tested, (b) the choice between experiments in which substances were tested alone and those in which they were tested in combination with ammonia, (c) the decision on October 21 to study the (known) arginine reaction (d) the decision on November 15 to test ornithine in combination with ammonia, (e) the decision on January 14, 1932 to vary the amounts of ammonia and of ornithine and to measure the consumption of ammonia, (f) the inference that ornithine was a catalyst, (g) the connection of the latter conclusion with the arginine reaction to form the urea cycle, (h) the identification of citrulline as an intermediate in the cycle.

Krebs made many decisions besides these that we would like to understand, but the decisions and inferences listed above were the most critical for discovering the process of urea synthesis. Notice that the first four reflect conjectures about the substances that might be the sources of the nitrogen in urea or that might accelerate or decelerate the synthesis process, whereas the last four were directed at specifying the reaction path from ammonia to urea.

Initial Conditions for Simulation

We have seen that a model that is to simulate a process of scientific discovery must be provided with knowledge about general methods of discovery as well as knowledge specific to the domain of research. The general knowledge must include knowledge of how phenomena observed in experiments or in natural situations can be used to formulate scientific problems and hypotheses (both new hypotheses and reformulations of existing hypotheses). Conversely, it must include knowledge of how hypotheses can be used to plan new experiments and observations.

Observation of phenomena generates hypotheses; hypotheses generate experiments that produce new phenomena.

In the case of the urea research, the domain-specific knowledge includes a vast range of biochemical knowledge, although any given researcher would be acquainted with only a modest part of it, and with techniques for extending this knowledge by literature search. Even knowledge held in memory is ineffectual until evoked as potentially relevant by cues provided by hypotheses or by the observed phenomena. The domain-specific knowledge also includes knowledge of experimental procedures, and of techniques for measuring the substances produced and consumed. It includes both knowledge of phenomena that had been observed or reported, and of existing hypotheses and theories about relevant phenomena and the processes that change them. In particular, in biochemistry, it includes an understanding of reaction paths, and of methods for inferring them from data, modifying and expanding them, and balancing reactions stoichiometrically.

More specifically, in this case much of the knowledge would relate to the chemistry of metabolism, and still more specifically, the chemistry of urea, amino acids, and ammonia and the complex nitrogenous materials in the body (especially proteins) that were thought to be the ultimate sources of the nitrogen in urea.

The Problem and Methods

Krebs sought to find the reaction path for the synthesis of urea from the nitrogen in amino acids (and ultimately in protein) in living cells, using as his experimental tool the tissue slice method. Previous research had failed to produce urea in a biologically plausible way outside living tissue, and the prevailing method of experimenting with tissue was to use whole organs perfused with appropriate liquids that allowed them to function for some time outside the whole organism. This method had had only slight success. Its main accomplishment had been to show that the liver was the principal organ in which urea was synthesized and that it could be synthesized from ammonia there. The tissue slice method allowed much more rapid experimentation and more accurate measurement of inputs and outputs under better controlled conditions than the method of perfused organs.

A reliable method was available for measuring urea production. Under the action of urease, CO_2 was extracted from the urea and its amount measured (The hourly rate of production was reported). Less

satisfactory methods were available for measuring the substances consumed during the urea synthesis reaction, although ammonia consumption was measured accurately by equipment that Krebs obtained about January 14, half-way through the period of the experiments.

Hypotheses

There were several known hypotheses about the source of urea nitrogen:

1. amino acids are the source, with ammonia an intermediate reaction product.
2. amino acids are the source, without ammonia as an intermediate product.
3. particular amino acids are sources of the nitrogen (with or without ammonia intermediate)

The first two hypotheses were regarded as more plausible than the third, for there was relatively good evidence that most of the nitrogen produced by decomposition of protein and other nitrogen-containing molecules was excreted in the form of urea.

Two specific known reactions turned out to be important. (1) One reaction that was known, but whose biological function was unknown, used the catalyst arginase to split the amino acid, arginine, a common constituent of protein, into urea and the amino acid ornithine, the latter not being found in protein. Arginase was found abundantly in the liver. Some unsuccessful efforts had been made to discover the ultimate disposition of the ornithine thus produced. Krebs became familiar with these facts from the literature during the first phase of his experimentation. (2) Just months before Krebs began work on the urea synthesis problem, it had been discovered that an obscure amino acid, citrulline, readily combined with ammonia to form arginine. Krebs obtained this latter knowledge only through a literature search after his initial discovery of the ornithine effect and its production of urea via arginine. Neither of these two reactions solved the problem of the ultimate sources of the nitrogen in urea. If the first reaction was on the main path of urea synthesis, it remained to explain how the nitrogen of the various amino acids became embedded in arginine. If the second reaction gave the answer to the first question, the analogous question had to be answered for citrulline.

Strategies

All of the knowledge described above was currently available, although which part of it was relevant to the urea synthesis problem remained to be discovered. Not all of it was in the minds of all biochemists who might attack this problem (e.g., the knowledge of citrulline or of the abundance of arginase in the liver, or of the relation of ornithine to arginine). Even if known and relevant, it might only be evoked by phenomena observed during the research or questions raised by these phenomena.

Success in solving a scientific problem of this kind requires obtaining the essential knowledge, whether from memory, from the literature, or from experimental results and observations. In order to conserve time and thought, a scientist would want to separate the relevant from the irrelevant as quickly as possible, so that attention could be focused on the former without distraction by the latter. Of course, relevance is only known with certainty after the problem has been solved. Knowledge that appears to be relevant (but, by hindsight, isn't) mainly enlarges the scientist's search space, leading him or her off into fruitless explorations. In this case, knowledge that the presence of metabolites might increase or decrease the yield of urea motivated a substantial number of experiments that, in the event, proved irrelevant.

Important stages in Krebs' progress toward the essential knowledge were marked by his first experiments with arginine, his first experiment with ornithine, his finding, by calculation, a reaction that produced arginine from ornithine and ammonia, and, at the very final stage, his discovery of the relation of citrulline to arginine. Also important was the accumulating evidence, after measurements of ammonia consumption were begun, that ammonia was the sole source of the urea nitrogen, and the inference of the consequent catalytic role of the ornithine. Negative results from experiments pursued in other directions, led Krebs gradually to abandon alternative hypotheses, and also facilitated his focusing attention on the relevant variables.

In the later stages of his search (beginning shortly after the ornithine effect was first observed), Krebs was also aided by another, and independent, line of experiment he was pursuing that began to show that the deamination of amino acids to form ammonia did not occur in the liver at all, but mainly in the kidney. With this finding, he could focus, as far as the liver was concerned, entirely on the portion of the reaction path

that led from ammonia to urea, and abandon the question of how the ammonia was produced.

These steps of progress were enabled by the heuristics of the basic hypothesis-experiment cycle. Given a hypothesis, experiments were designed that might test its correctness. From the findings of an experiment, if they challenged the hypothesis that motivated it, new or improved hypotheses were generated to explain the phenomena.

But experiments do not merely test the hypotheses that motivated them. Apart from a specific hypothesis, various pieces of knowledge held prior to an experiment provide expectations about its outcome. Experimental findings that violate expectations produce surprise that is exploited to generate new hypotheses to explain the surprising effects, and to design experiments to determine the effects' scope and generality. Surprise may be evoked not only by discrepancies between experimental findings and specific hypotheses, but also by sizeable experimental effects that have no obvious cause in terms of known mechanisms. The former discrepancies appear in the form of statistically significant differences between experimental and control conditions, but the latter arise as differences between experimental results and expectations derived from previous knowledge.

The main surprising outcomes in this research were the production of much more, or much less, urea in a given experimental condition than had been expected. If ornithine had been the first amino acid tested, the large yield of urea would not have been surprising, for it would have been consistent with the initial hypothesis that urea was produced from amino acids, with ammonia as a possible intermediate product. Because small yields had been obtained from other amino acids and because the large yield was forthcoming only when both ornithine and ammonia were present the outcome was surprising. A principal effect of surprise on Krebs was to focus his attention on the unexpected phenomenon, first to verify that it was not a mistake, then to explore its implications. The heuristics that were prominent in the search took the following forms:

1. defining an experimental condition that is hypothesized to synthesize urea; if there is a significant yield, trying to magnify the effect by modifying the conditions;
2. finding a stoichiometrically correct reaction that will explain the yield;

3. if more than two reagents are involved in a reaction, designing experiments that could yield intermediate steps in the reaction;
4. formulating a hypothesis about the source of the nitrogen;
5. formulating a hypothesis about intermediate products;
6. formulating a hypothesis about facilitating conditions;
7. formulating a hypothesis about reaction paths;
8. generalizing a hypothesis from a substance to a class of substances;
9. specializing a hypothesis to particular class members;
10. elaborating a hypothesis to incorporate (4), (5), (6) or (7).

KEKADA and CDP Compared

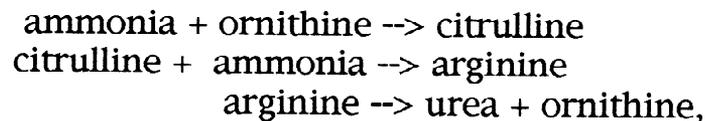
Both KEKADA and CDP fit reasonably well the description of goals, methods, hypotheses (models, in the case of CDP), and strategies that we have just outlined. This is perhaps not surprising in light of the fact that, on the urea problem, they both obtained their information from the same detailed historical record: the log books and published papers of Krebs and Henseleit, and Holmes' painstaking analysis of these and other sources. Although, on the surface, the two programs appear to be rather different, the processes they followed in simulating the discovery were, in many respects, closely similar. In order to discover why, we now compare the two programs in more detail.

Hypothesis space and instance space. Both KEKADA and CDP can be described in terms of the two-space model of learning and discovery proposed by Simon and Lea (1974): law discovery, in this scheme, derives from alternating searches in the space of hypotheses (models) and the space of instances, or phenomena. Hypotheses guide the design of experiments that produce phenomena; the phenomena reject hypotheses and guide the search for new or revised hypotheses.

In KEKADA, once a problem has been chosen for study, experiment-proposers derive possible experiments from existing hypotheses. At the same time, expectations are set, on the basis of previous knowledge, for the outcomes of the experiments. New information obtained from the experiments, especially data that violate expectations, modify the knowledge base, removing and modifying hypotheses, modifying confidence in hypotheses, and generating new hypotheses about mechanisms and phenomena; these processes lead, in turn, to the generation of new experiments. This cycle is summarized in Figure 1.

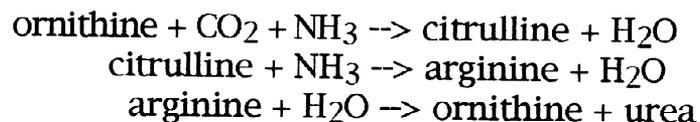
In CDP, a very similar cycle is described in causal terms. An effect (a phenomenon -- in the case at hand, the production of urea in the liver) is selected, with the goal of discovering sufficient conditions (causes) for its production. A possible explanation is chosen from a hypothesis space (space of models), and an experiment designed to test it. The experiment is evaluated to modify the model by eliminating irrelevant factors, incorporating new factors, generalizing or particularizing. Then a new experiment is designed. This cycle, as summarized by Grasshoff and May, is shown in Figure 2.

Form of hypotheses. In application to the urea synthesis problem, hypotheses for both systems consist of possible synthesis reaction paths, the required input substances and conditions being the final causes, the substances produced being intermediate or final effects in a short or long chain of processes. CDP employs a more formal language than KEKADA for stating hypotheses, expressing them in terms of causal chains and networks, and Boolean functions for the conjunction and disjunction of inputs and conditions at each node in the network. The final causal chain in the urea synthesis problem can be described in this language by:



with the ornithine cycling back as input to the first reaction.

The inputs and outputs can be stated in qualitative form, as above, or, in the final form of the hypothesis, can include the quantities of substances that balance the equations stoichiometrically. Thus, CDP represents Krebs' solution quantitatively by:



In either KEKADA or CDP, other conditions for the experiment (e.g., whether the experimental animals are fed or starved, composition of the Ringer's solution, presence of glucose, etc.) can be included among the causes.

Initial hypotheses. Initial hypotheses about the problem solution are included in the initial knowledge base. In the case of KEKADA, two main classes of hypotheses were provided:

Amino acids donate their amino groups to form urea, with ammonia as an intermediate product in the process.

Amino acids and ammonia react to form urea, each contributing one of the two nitrogen atoms.

Both CDP and KEKADA begin with the first of these hypotheses, answering, in part, the question raised earlier: what kinds of substances were tested for their contribution to or influence upon the yield of urea. As "amino acids" refers to a class of twenty or so substances, individual experiments are designed by selecting elements from this class and comparing their yields of urea with the yield from ammonia alone. As the heuristics that the systems possess for choosing among the various amino acids are relatively weak, this leads to an essentially random search, which produces at best (with alanin) a modest yield of urea, only a fraction of that available from the nitrogen of the amino acid.

KEKADA upon several failures to verify the first hypothesis, and after testing only a few amino acids, turns to the second hypothesis, now testing amino acids in combination with ammonia until it chooses ornithine and discovers the "ornithine effect."

Similarly, CDP tests the hypothesis:

amino acid --> ammonia --> urea;

then, at some point, switches to combining the amino acid with ammonia, which is represented as:

amino acid + ammonia --> urea.

Notice that the first hypothesis implies that the amino acid is the source of the nitrogen in the urea, while the second hypothesis is noncommittal on the source of the nitrogen.

Neither KEKADA nor CDP motivates very specifically the experiments that were actually run by Krebs up to the time when the ornithine effect was discovered, and in particular, neither provides a convincing answer to the remaining key questions raised earlier: the times at which and the

order in which different substances were tested; the choice between testing substances alone or in combination with ammonia; the decision to study the known arginine reaction; and the decision to test ornithine in combination with ammonia.

First, on the evidence, Krebs was not testing all members of the class of amino acids, either systematically or randomly. Indeed, his initial hypothesis that the amino acids were the source of the nitrogen in urea would have predicted a similar rate of production of urea from all of them. He tested three or four acids, usually several times each, with unencouraging results, but did not extend his search systematically to the others. In fact, it has been impossible to establish from the historical record (including interviews with Krebs) what motivated the test of ornithine at the time it took place.

Both KEKADA and CDP succeed about equally well, on the basis of very weak initial hypotheses, to stage a series of experiments (not exactly those of Krebs) that do, presently, lead to the ornithine test. This illustrates that chemical experiments designed around weak hypotheses covering the potentially relevant substances may discover crucial phenomena that are not explicitly incorporated in the hypotheses and that require their radical reformulation. Thus, the hypothesis that any or all of the amino acids, in conjunction with ammonia, might be sources of the urea could motivate the test of ornithine and ammonia. This does not imply that the computer systems and Krebs had the same motives for the behavior. As Holmes has shown convincingly, the historical evidence is not sufficient to pin down Krebs' specific motives for the experimental sequence he followed prior to his discovery of the ornithine effect.

Exploiting the ornithine effect. Once the large yield of urea from a combination of ornithine and ammonia is discovered (which KEKADA interprets as "surprising," and CDP as "statistically significant"), both systems undertake a new line of experimentation motivated by the discovery. KEKADA first tests, with negative results, whether substances closely resembling ornithine (provided to KEKADA by its "knowledge") will produce the same effect. It tests both the hypothesis that ornithine is a condition for the synthesis and that it is a source of (some of) the nitrogen.

To do this, KEKADA begins to measure the ratio of urea produced to ammonia consumed, finding that all of the nitrogen in the urea can be accounted for by the ammonia consumption. Next, KEKADA, begins to vary the quantity of ornithine used, finding that the yield of urea varies much less than proportionately with the amount of ornithine, and that the

reaction is sustained with quite small quantities of ornithine, consuming all the ammonia if continued long enough.

During the corresponding period, CDP also carries out experiments varying the amount of ornithine used and experiments measuring the amount of ammonia consumed. Comparison of the programs' motivations with those of Krebs at this time are complicated by the fact that Krebs just then received and put into operation a new instrument that, for the first time, allowed him to measure the amount of ammonia consumed. Hence, he did not need any reasons for computing the ammonia/urea ratio other than that he was now in a position to do so, and that this would be a reasonable measure to take whenever it was experimentally feasible.

With these measurements at hand, a chemist might well begin considering possible reaction paths and their stoichiometry. Both KEKADA and CDP have procedures for balancing reactions when the input and output quantities are known, but unless intermediate products are also known, this leaves open a large field of possible reactions among which to search. In the present case, if one computes how two molecules of ammonia could be converted into one of urea, one finds that, on balance, one molecule of carbon dioxide is required and one molecule of water is produced. This makes the transformation chemically plausible, but says nothing about the reaction path or the intermediate products. In particular, it suggests nothing about the role of ornithine.

One initial step in the reaction path that might suggest itself is to consider the synthesis of the two ammonia molecules and the molecule of carbon dioxide, and to observe what remains when a molecule of water is removed. What remains is a molecule of the amino acid arginine. It was already known to Krebs (and there is testimony that he recalled at this time) that a molecule of arginine combines with a molecule of water (in the presence of arginase) to produce a molecule of ornithine and one of urea. Krebs, as we have seen, was also aware that both ornithine and arginine were found chiefly in the liver. Here is a path that produces urea from ammonia, and recovers the ornithine that it uses initially.

Valdes-Perez (1994) has shown, with the MECHEM program, that an artificial intelligence system (an "expert system") designed to find chemical reaction paths, starting with the simplest possibilities and working toward more complex ones, will find the two-step reaction path just described with a moderate amount of computation, even without previous knowledge that arginine might be an intermediate product.

With the heuristics provided to CDP and KEKADA, which included information about the relevance of arginine (i.e., that arginine separates into urea and ornithine), both of them discover this path. Subsequently, they both elaborate it by showing that the ornithine and one molecule of ammonia convert into an intermediate substance, citrulline; and citrulline with a molecule of ammonia converts into arginine. Again, they are explicitly informed of the relevance of citrulline. (Krebs discovered the reaction path from arginine to citrulline, or rather its inverse, in the literature, where it had been published quite recently.)

Some Generalizations. We can summarize our comparison of the two systems, and the match of both to Krebs' behavior in a couple of generalizations, one concerning the discovery of the ornithine effect; the other the discovery of the detailed reaction path and the role of ornithine as a catalyst.

In the first phase of experimentation, KEKADA and CDP behave quite similarly, pursuing a general hypothesis that both ammonia and amino acids are involved in the synthesis of urea. Having no detailed model of a possible reaction path, they both compare the yields of urea from individual amino acids with the yield from ammonia. When these strategies fail to produce large yields, they shift (mainly on the basis of this failure) to testing individual amino acids *in combination with ammonia*, hitting by chance on the ornithine effect. In CDP, this shift is justified by regarding the reaction, amino acid + ammonia \rightarrow urea, as a generalization of the reaction, amino acid \rightarrow ammonia \rightarrow urea, because the latter, but not the former, places ammonia in the special role of an intermediate product and not a source of the urea nitrogen, while the latter is noncommittal on the ultimate source of the nitrogen. In KEKADA, both hypotheses are equally specific, the former being that the amino acid and the ammonia each contribute an atom of nitrogen to the urea.

The evidence does not permit us to decide which, if either, of these is the actual reason for the shift. We have very little information about what guided Krebs' search during the period that ended with the successful experiment with ornithine. We do know that Krebs did *not* systematically test successive amino acids. Only six such acids had been used at all before he turned for a time to a miscellany of other experiments, mainly with metabolites, then ran an experiment on arginine and, more than three weeks later, the first ornithine experiment. We simply cannot determine, with even moderate certainty, how he was led to the experiment with ornithine, although the earlier experiment with arginine might have had something to do with it, and also the knowledge that both arginine and

ornithine were present in quantity in the liver. Plausibility, however, is not evidence.

The most reasonable conclusion from the evidence we do have, and from the simulations, is that experiments, whatever motivated them, seeking to obtain urea from preparations that include amino acids and ammonia can, along a variety of routes, lead after some time to the discovery of the ornithine effect. Even in a situation like this one, where we have an enormous body of evidence from the lab notebooks of the experimenter's path, the critical evidence that would permit us to test the detailed hypotheses proposed by the simulation programs is simply absent. Had Krebs undertaken systematically to test all of the amino acids, alone or in combination with ammonia, at most about 40 experimental conditions, and on average about 20, would have led him to the ornithine effect. In fact, he had performed about 100 experiments, totaling perhaps 500 conditions, before he found it. Of course a substantial number of these conditions were concerned with establishing and refining the experimental procedure and testing for possible effects of substances other than amino acids, so that his total search space was considerably larger than was indicated above.

The second segment of the simulation, leading to the discovery of the reaction path, is much less conjectural. Once we have some knowledge of the input and output substances involved in the reaction, and their quantities, standard methods of stoichiometric analysis, perhaps aided by the hypotheses that arginine and (later) citrulline are intermediate products, lead rather directly to the correct reaction path and then to the conclusion that ornithine has the role of a catalyst. The precise computational schemes used by KEKADA and CDP are rather different, and again, we have little specific knowledge about the exact form of Krebs' computations, but these differences are not of great psychological interest in this context.

Divergence and Equifinality

In discussing the methodology of comparing simulation programs with human processing, we mentioned two important factors, one of which makes matching two dynamic processes difficult, the other of which facilitates matching. On the side of difficulty, dynamic processes with different initial conditions are often unstable or chaotic, so that their divergences may magnify rapidly and without limit. At best, they can only match over shorter or longer intervals, and the simulations must then be "reset" to test a next interval of matching. On the side of facilitation,

purposeful, goal oriented, actions, to the degree that they are successful, exhibit equifinality. However much two sequences of actions may diverge, they cannot succeed unless they find their way to the same goal.

Both Krebs and the two simulations (KEKADA and CDP) did, in fact, discover the reaction path for urea synthesis. But an attempt to match in detail the routes along the way shows a rapid parting of the three paths terminated by sudden convergence at the end of each of two episodes: at the common discovery of the ornithine effect and at the discovery of the structure of the reaction path. Both programs resemble the actual history in having this two-episode structure; there is much less resemblance among the paths within each episode. What the simulations mainly showed was that discovery of the ornithine effect was a necessary, and nearly sufficient, condition to finding an *in vivo* reaction path for the synthesis of urea from ammonia, and that systematic application of stoichiometric methods like those employed by the MECHEM program could complete the task without much excess search.

For both programs, the essential conditions for discovering the ornithine effect were (a) to include amino acids and ammonia among the input substances, and (b) to test specific amino acids *together with* ammonia. Both conditions were readily suggested by prior knowledge: the former by knowledge of the principal molecular repositories of nitrogen in the organism, the second by the more specific idea that amino acids and/or ammonia could be the sources of the nitrogen in urea. As long as there was sufficient motivation to test the whole range of amino acids, the crucial experiment would sooner or later be performed. Notice that the most probable motive for testing ornithine -- that it might provide all or part of the nitrogen in the urea -- led to discovering the ornithine effect but turned out to be irrelevant to ornithine's actual role in urea production. In this sense, the inclusion of ornithine among the substances tested was fortuitous.

The evidence from the data of the actual discovery, however, creates more than a little doubt that this was the whole story. Krebs did not, in fact, test a wide range of amino acids, and additional motivation is needed to explain the initial experiments with arginine in October, and with ornithine in November. Ornithine is not a common constituent of ordinary proteins. However, both ornithine and arginine were known to be present in considerable quantities in the liver, and not in other organs; arginine was known to decompose into urea and ornithine; and the ultimate disposition of the ornithine in liver was unknown. Any or all of these facts, when and if recalled by Krebs, could have suggested a special role for

arginine and ornithine and motivated the crucial experiments, even without postulating that ornithine was the source of the nitrogen in urea. The difficulty with this easy explanation, as Holmes shows, is that there is no empirical evidence, either from the logbooks or from Krebs' publications or later recollections, that this knowledge played a role in the conduct of the experiments.

The question must now be raised whether Krebs solved the problem he set out to solve: to show how the nitrogen in amino acids is converted into urea. What he actually showed was how ammonia was converted into urea without settling the issue of how amino acids are converted into ammonia. The same is true of the two models, KEKADA and CDP. With respect to Krebs, it was remarked earlier that, just at the time he was solving the problem of the ornithine cycle, he had discovered that the conversion of amino acids into ammonia appeared to be occurring mainly in the kidneys, not in the liver. He was therefore able to separate the two problems, and did, in fact, continue successfully the research on deamination of amino acids in the kidney. Without being provided with additional information, neither of the two models could take this step. Nor, as the original problem was put to them -- Show how some combination of amino acids and/or ammonia could produce urea. -- would they be motivated to try this step, or even be aware that a problem remained.

Sensitivity Analysis

From the foregoing discussion we conclude that the existing empirical evidence is insufficient to determine a unique discovery path, and that the two simulations, although different in many details, tell basically the same story, which is consistent with the evidence. Perhaps the most valuable products of the simulations are their explorations of alternative possibilities that tell us something about the potential size of the exploration space, and help account for the large number of experimental conditions that Krebs tested. Such explorations provide a disciplined method for reasoning about the consequences of lengthening or shortening Cleopatra's nose, and thus considering "what if?" histories different from the actual sequence of events.

To this end, a number of different variants of KEKADA were tested in order to see what effect the changes would have on the discovery path and outcome. In these variants, certain of KEKADA's choices of path were not made automatically but by our intervention. For example, the hypothesis was present in the knowledge base that any one of the substances present

in an experiment might be a catalyst, and not contribute urea to the output. As there is no evidence that this hypothesis was evoked in Krebs' mind until quite late in the discovery process, it becomes important to see how crucial it is, and when it has to be evoked in order to enable the discovery. Simulation shows that the hypothesis does not play any essential role at all. It is not involved in discovery of the ornithine effect, and we have seen that elucidation of the reaction path follows from quite other considerations.

First, the quantitative measurements carried out after January 14 pretty well settled, without any prior hypothesis, that ammonia was the source of the nitrogen in the urea (two molecules of ammonia for every molecule of urea). But the ornithine, while a condition of the reaction, was not necessarily acting as a catalyst in it.

Second, familiarity with arginine and stoichiometric calculations readily lead to the discovery that combining ornithine with two molecules of ammonia (and one of carbon dioxide) produces arginine. Ornithine was an input, not a catalyst, in this reaction. In the examples of catalysis known at the time of these experiments, the idea that the catalyst engages in actual chemical reactions with the other substances, then is again released, was relatively new, and still had little empirical support. For example, metal surfaces acting as catalysts were sometimes regarded as "facilitating" a reaction without engaging in it.

Third, prior knowledge of the decomposition of arginine into ornithine and urea (plus water), enabled the path to be completed with correct stoichiometry, and independently of the catalytic interpretation of ornithine's role.

Ornithine operates here catalytically if we view the whole path as a single step: then ammonia combines with carbon dioxide, under the catalytic influence of ornithine, to produce urea and water. When we write the story in this form, arginine and citrulline do not appear in it at all. It is quite possible that Krebs interpreted the role of ornithine as catalytic only late in the discovery process, after he had determined the reaction path by stoichiometric methods; and none of the evidence contradicts this possibility. The catalytic action of ornithine would then be one of the results of the investigation, in itself an interesting contribution to the theory of catalysis, not one of the hypotheses guiding the discovery.

We have mentioned only in passing the numerous experiments that Krebs conducted, not with amino acids, but with metabolites. Indeed, in

the experiment where the ornithine effect was discovered, the other experimental conditions involved such substances. The simulation would perform these experiments only if the possible relevance of these substances to urea synthesis were included among Krebs' hypotheses. The historical evidence showed that such hypotheses had been explored extensively in previous research on urea synthesis, thus motivating their inclusion in the knowledge base.

We have also said little about the numerous experiments, especially in the early phases of the work, aimed at improving the instrumentation and experimental procedures and creating conditions that would be conducive to urea production. To motivate these experiments would again require that specific knowledge about such matters be included in KEKADA's and CDP's knowledge bases.

To motivate the two kinds of experiments mentioned in the previous paragraph, appropriate hypotheses were inserted in the knowledge bases of the two programs, but independent historical evidence motivated their introduction. When independent evidence is not available, explicit user intervention to cause the experiments to be performed calls attention to the fact that these particular events along the discovery path are unexplained. Of course, the choice between motivating events by altering or augmenting the knowledge base and activating them by user intervention is not a black-or-white matter. One will be influenced by the strength or weakness of the independent evidence for the knowledge and by one's conservatism in testing the simulation. Our own attitude in testing KEKADA has been that, when in doubt, we should adopt the conservative strategy of marking such interventions as acts of the user.

Experimental Controls and Critical Experiments

In the history and philosophy of science the concepts of experimental control and critical experiment have been prominent. To determine whether X has a causal effect upon Y, we test for the presence (and often the magnitude) of Y in the presence or absence of X; the former condition being referred to as the experimental condition, the latter as the control. The values of Y for the two conditions are compared, the hypothesis that they are equal is referred to as the *null hypothesis*, and a causal effect is regarded as statistically significant if the probability is small that an effect of that size or larger would be observed if the null hypothesis were true. Controlled experiments are implementations of John Stuart Mill's celebrated *Method of Difference*.

A critical experiment, on the other hand, is one in which the predicted outcome is different depending upon whether one or another of two hypotheses is true. Thus, the Michelson-Morley experiment was a critical experiment, which had a negative outcome, for the theory of ether drift (i.e., that the observed velocity of light depends on the movement of the Earth relative to a hypothetical substance, ether, in space).

In this experiment, the velocity of light is measured in two conditions having different directions of movement of the Earth in space, and the difference in velocity is interpreted as the drift relative to the ether. The absence of ether drift can be taken as the null hypothesis, and the statistical significance of the observed drift evaluated; but there is no asymmetry between the two experimental conditions, and only by a strained interpretation can either be regarded as a control for the other. On the other hand, an experiment with a control can be interpreted as a critical experiment to choose between the null hypothesis and the hypothesis of a causal connection between X and Y.

KEKADA's approach to comparison between experimental outcomes and their implications for hypotheses is Bayesian in conception, and somewhat different from the notions of control or criticality. By "Bayesian," we mean that experience leads cumulatively to the formation of expectations about the outcomes of proposed experiments — especially, expectations about yields of substances added to the tissue. Experimental findings that are consistent with expectations provide no new knowledge, and hence do not cause changes in the current strategy of experimentation; whereas findings that are surprising (violate expectations) introduce new hypotheses and motivate new experiments. In CDP, on the other hand, statistically significant differences between control and experimental conditions are the principal source of new knowledge and changes in strategy and models.

The search for the causal conditions for urea synthesis can be viewed as a series of controlled experiments, in each of which X is a substance or set of substances and conditions that might produce urea, and Y is the amount of urea produced. Absence of X can be interpreted as the control condition, but with some ambiguity. As has often been pointed out, the failure of X to produce Y does not demonstrate the causal irrelevance of X. The correct hypothesis may be that Z must be present in order for X to produce Y ($X + Z \rightarrow Y$). If so, then the correct control is a condition in which Z is present, and not X, and the correct experimental condition one in which both Z and X are present. But, of course, to perform this experiment, one must be aware of the possible significance of Z.

In the case before us, the nature and identity of the "possible Z's," (a term it would not be easy to define) were only partly known. In all of the experiments, the tissues were infused with a serum (initially, Ringer's Solution), because it was generally understood that tissue would not otherwise be biologically viable. However, if X in the presence of serum failed to produce urea, it could be argued that the failure was due to the composition of the serum. In fact, a series of experiments was undertaken by Henseleit during September 1931 to compare Ringer's Solution, the serum actually used, with human blood serum. In January 1932, Krebs returned to this problem and substituted a physiological solution of his own design for the Ringer's Solution used previously. Similarly, the numerous experiments with metabolites and other substances that aren't potential contributors of nitrogen may be interpreted as searches for necessary conditions for urea synthesis, hence as possible components of the control condition for X.

Thus, the concept of "control condition" is not absolute, but is relative to what is known about the whole experimental environment and the factors that could be relevant to the value of the dependent variable. In point of fact, Krebs generally ran a "control" condition — Ringer's solution or its equivalent with no other addends — in each experiment, although, as time went on, this condition was more and more often omitted. When it was omitted, the condition most nearly resembling a control was that in which only ammonia was added to the basic preparation. In this case, the yield of urea in the other conditions was compared with that employing ammonia alone.

From almost the beginning of the experiments, and increasingly with time, particular conditions in an experiment were compared not only with the "control" in the same experiment, but also with various relevant conditions that had been run in previous experiments. Thus, the yield with ammonia alone was often compared from one experiment to another to see if it was constant or whether other conditions (e.g., whether the rat from which the liver tissue was obtained had been fed or not) may have affected it. Similarly, the surprise and attention produced by the ornithine effect did not derive from comparing the yield with ornithine and ammonia to the yield without addends, but from comparing it to the yield with ammonia alone. A large number of substances had been tested, and had generally been found not to increase the urea yield much, if at all, above the level produced by ammonia. These outcomes were treated as negative.

As different experiments had different auxiliary conditions, comparing two conditions from different experiments using the same addends provided a valuable check on the effects of the auxiliary conditions. The expectations mechanisms incorporated in KEKADA enable it to simulate these kinds of strategies, which are observed frequently in Krebs' experiments.

After the observation of the ornithine effect, when emphasis shifts to determining the quantitative contributions of particular inputs to the urea production, and to finding a stoichiometrically balanced reaction path, a principal goal of experiments is to estimate parameters — e.g., the ratio of ammonia consumed to urea produced, or the ratio of ornithine supplied to urea produced. In the former case, comparison among conditions is only relevant in determining how well the predicted 2-1 ratio of ammonia molecules to urea molecules is maintained, and comparisons are again made between, as well as within, experiments. In the latter case, conditions are compared mainly to check how strongly the production of urea is influenced by the amount of ornithine supplied, and to see if the experimental conditions (e.g., duration of the experiment) can explain deviations from constancy.

Conclusion: Answers to the Initial Questions

By way of a general summary, we first return to the eight questions we posed at the beginning of our discussion of the simulations and see what answers they have received. Then we comment on the implications of our findings for the use of computer simulation as an aid to the study of scientific discovery and to the analysis of historical evidence about particular cases of discovery.

The Central Questions

(a) What kinds of substances were tested, either as sources of the urea nitrogen or as conditions for urea synthesis; when were they tested, and in what order?

The main guide to the selection of substances, prior to the discovery of the ornithine effect, was existing knowledge about the possible sources of nitrogen for urea and about the effects of the presence or absence of metabolites for the quantity of urea produced. Perhaps knowledge about the presence of arginine and ornithine in the liver was also influential in focusing special attention, after a time, on those two substances, but there is no concrete evidence to support this supposition. Neither computer

simulation succeeded, to any significant degree, in providing additional heuristics to narrow the search or sharpen priorities. However, as long as the search was limited largely to amino acids and ammonia, the search space was not large, so that order was only of secondary importance.

What is ironic about the result is that the apparent motive for experimenting with ornithine (that it might be the source of the nitrogen in urea) had nothing to do with the ornithine effect (for ornithine served as a catalyst, permitting two ammonia molecules to be converted into urea without affecting the amino radicals in the ornithine molecule). The discovery was accidental, not in the superficial sense that many alternatives had to be tested to find the right one, but in the much more fundamental sense that the correct alternative was tried, as far as the evidence indicates, for an entirely irrelevant reason. If the reaction were catalytic, but the catalyst happened to be a complex protein rather than an amino acid, then no amount of testing of amino acids and nitrogen could have led to success.

After the November 15 discovery of the ornithine effect, the search was much more directed, as predicted by both simulation programs. The KEKADA strategy of exploiting surprise by testing substances related to ornithine, and by measuring input and output quantities in order to search for a mechanism (a stoichiometrically balanced reaction path) accounts for most of the remaining experiments. CDP achieves this same result by turning from qualitative to quantitative experiments after identifying a qualitative causal mechanism for the synthesis (i.e., ammonia in the presence of ornithine produces urea).

(b) How was the choice made between testing substances individually and testing them in the presence of ammonia?

The hypotheses from prior knowledge included ammonia, amino acids and combinations thereof as possible sources of the nitrogen in urea. The ammonia was also viewed as a possible intermediate product. Depending upon which of these hypotheses was given the greater priority, different experiments would be run by either KEKADA or CDP. In point of fact, experiments using ammonia together with another substance as inputs first appeared on October 22. In this case, the substance was potassium cyanate and not an amino acid, but experiments with ammonia and several amino acids began a few days later. Neither simulation throws any clear light on the initiation of this shift, although the initial experiment that used the combination could have served to test the hypothesis that

the potassium cyanate contributed one of the amino radicals to the urea, the ammonia the other.

(c) How was the decision reached to study the known decomposition of arginine into urea and ornithine?

We have already shown that the hypotheses about how this decision *might* have been reached lack supporting evidence. The simulations provide no help in answering this question.

(d) How was the decision reached to study the yield of urea from ammonia and ornithine in combination?

The answer is much the same as the answer to (c).

(e) How was the decision reached to measure the quantities of inputs and outputs of the reagents?

In the case of KEKADA this decision follows from the decision to discover the mechanism of the production of urea from ammonia in the presence of ornithine, which, in turn, is an application of the general heuristic to search for a mechanism to explain a surprising phenomenon. In the case of CDP, this decision follows from the rule that when a qualitative causal solution to a problem has been found, the solution should be sharpened by quantification.

(f) How was it inferred that ornithine is a catalyst?

Part of the inference was made as soon as quantitative measurements were made of ammonia consumption. It was seen that, while the presence of ornithine was necessary for the reaction, all of the urea nitrogen was attributable to the ammonia alone. The full understanding of the catalytic role of ornithine was not reached until the reaction cycle with arginine as an intermediary had been constructed. Hence, it was not necessary to infer as soon as the ornithine effect was observed, as KEKADA did, the possibility that ornithine had a catalytic role.

(g) How was the basic urea cycle discovered?

Both programs discover the cycle relatively directly by using stoichiometric methods to obtain a balanced reaction path that is consistent with the quantitative measurements of inputs of ornithine and ammonia, and outputs of urea.

(h) How was the intermediate role of citrulline discovered?

Both programs are motivated to search for an intermediate by the rule that three or more molecules cannot interact simultaneously in a single reaction step, as was required in the two-step cycle. Again, application of stoichiometric methods by both programs quickly found citrulline, a recently described amino acid, as the intermediate.

Computer Simulation of Scientific Discovery

We cannot claim that the simulations, singly or severally, cast any blinding light (perhaps any light at all) on the uncertainties that remained about the path of Krebs' discovery processes and its motivation after Holmes had carried out his very close and competent analysis of the available empirical evidence. Their chief contribution lies in connecting that historical account with the theories we find in the literature, as encapsulated in the two simulation programs, about the processes of scientific discovery.

Central to the whole experimental process in both programs is some version of Mill's Method of Differences. Experimental conditions and inputs are varied, generally one at a time, and differences are noted in the values of the dependent variables. This method was incorporated, if in somewhat different ways, in both computer programs. Conditions that produced no effect, as measured against expectations (KEKADA) or a control (CDP), were ignored.

The space of potential experiments was defined in part by prior knowledge and hypotheses (i.e., experiments thought capable of producing effects relevant to the goal of the scientist) and in part by the techniques and instrumentation available. The latter is illustrated by the new experiments undertaken after the acquisition, midway in the research, of equipment for measuring ammonia consumption. As the hypotheses bounding it were rather weak and general, the space of experiments was of considerable size — fortunately, large enough to include ornithine and to encompass testing the conjunction of ammonia and an amino acid.

Effects were interpreted, not only in terms of previously generated hypotheses, but also in terms of hypotheses evoked by the effects themselves. Prior to the discovery of the ornithine effect, attention focused mainly on substances that could provide the urea nitrogen, or substances that, in some general sense, might provide appropriate

conditions for the reaction's occurring. After discovery of the effect, and especially after attribution of the source of the nitrogen to ammonia, attention shifted to determining a reaction path and the role of ornithine in that path. Thus, observation of the ornithine effect produced a major shift in problem representation, and even in the research goal.

Krebs' success can be attributed in some measure to the "good fortune" that the critical ornithine experiment lay within the space of potential experiments, although probably for the "wrong" reasons. But Pasteur has already reminded us that "accidents favor the prepared mind."

References

Grasshoff, G. (1995). *De Entdeckung des Harnstoffzyklus*. Technical Report, University of Hamburg, January, 1995.

Grasshoff G. & May, M. (1995). From historical case studies to systematic methods of discovery, Proceedings of the AAI Spring Symposium on *Systematic Methods of Scientific Discovery*. Stanford University, March 27-29, 1995, pp. 46-57.

Holmes, F. L. (1980), Hans Krebs and the discovery of the ornithine cycle. *Federation Proceedings*, 39 :216-225.

Holmes, F. L. (1991), *Hans Krebs: The formation of a scientific life: 1900-1933*, New York, NY: Oxford University Press.

Krebs, H. A., & Henseleit, K. (1932a). Untersuchungen über die Harnstoffbildung in Tierkörper, *Klinische Wochenschrift*, 11 :757-759.

Krebs, H. A., & Henseleit, K. (1932b). Untersuchungen über die Harnstoffbildung in Tierkörper, II, *Klinische Wochenschrift*, 11 : 1137-1139

Krebs, H. A., & Henseleit, K. (1932c). Untersuchungen über die Harnstoffbildung in Tierkörper, *Zeitschrift für Physiologische Chemie.*, 210 :33-66.

Kulkarni, D. (1988), *The processes of scientific research: The strategy of experimentation*. Ph. D. Dissertation, Department of Computer Science, Carnegie Mellon University.

- Kulkarni, D. & Simon, H. A. (1988) The processes of scientific research: The strategy of experimentation. *Cognitive Science* 12 :139-175.
- Langley, P., Simon, H. A., Bradshaw, G. L. and Zytkow, J. M. (1987). *Scientific discovery: Computational explorations of the creative process*. Cambridge, MA: The MIT Press.
- Simon, H. A. (1995) Commentary on Grasshoff & May (1995). Proceedings of the *AAAI Spring Symposium on Systematic Methods of Scientific Discovery*. Stanford University, March 27-29, 1995, pp. 58-59.
- Simon, H. A. & Lea, G. (1974), Problem solving and rule induction: A unified view. In Lee W. Gregg (Ed.), *Knowledge and Cognition*, Potomac, MD: Lawrence Erlbaum Associates.
- Valdes-Perez, R. (1994). Conjecturing Hidden Entities via Simplicity and Conservation Laws: Machine Discovery in Chemistry. *Artificial Intelligence*, 65 :247-280.

Figures

Figure 1. Figure 3 — p. 153 of K&S.

Figure 2. Figure [5- p. 51 of G&M Stanford].

Some Further Issues

Rate of production of urea versus amount of urea produced.

Rate of production would not be independent of amount of catalyst