

# Modeling and Simulating Biological Processes as Logical Enterprises<sup>1</sup>

Toni Kazic

Institute for Biomedical Computing, Washington University

Shalom Tsur

University of Texas System Center for High Performance Computing

IRI-9117005 10/1/91 - 3/30/94

## Abstract

Understanding and improving the biological machines responsible for very complex phenomena would be greatly speeded by a computational tool which simulates the outcome of laboratory manipulations. These processes are extremely difficult to think about because they involve literally hundreds of interacting parts. For modeling to be realistic enough to be useful, the model must include detailed information on both the structure of the molecular parts and their functions in biochemical reactions. Representing and using this information in databases to simulate even simple processes is the first step to achieving larger, more comprehensive models. We are studying how the molecules, reactions and constraints on their interactions can be represented in the computer. A key requirement is that facts, ideas and questions all be represented in the same language and using the same fundamental approach, while still permitting scientists to describe and manipulate ideas in the language they find most natural. We have achieved this goal by using declarative database technology and by expressing molecules, reactions and constraints in a formal grammar. The grammar allows one to pick the most congenial way to describe and model the biochemistry in a detailed and comprehensive manner.

## 1 Introduction

Imagine being able to engineer bacteria to degrade plastics, screen potential drugs for side-effects before clinical testing, or design vegetables with higher starch content – as easily as we design airplanes. Biologists now find these kinds of goals very difficult to achieve because the biological machinery which produces them is so complex, with hundreds of different interacting parts. The parts in biology are not just static objects, but also the processes that the interactions of the objects produce. Goals like these, which most likely involve more than one process, can be achieved only by understanding how existing biological machinery works and then rearranging it to suit the need. Since changing one part can affect so many others, choosing an effective combination of alterations is essentially a trial-and-error process, fraught with its inherent difficulties and inefficiencies. Understanding and engineering biological machines – that is, a group of processes which produce a desired result – would clearly benefit from a way to simulate the effects of changes before making them in the laboratory.

An increasingly important part of efficient engineering is the use of computer-aided design tools (CAD). These tools allow engineers to visually design the machine on the computer. Information about the parts and their properties, stored in the CAD tool's database, is used to help restrict their placement: for example nuts fit onto bolts, not screws. Once a preliminary design is completed, the computer is then used to simulate its performance under a wide variety of conditions. Analyzing the simulation results produces possible refinements, which are incorporated and tested again. Inventing

---

<sup>1</sup>*Proceedings of the NSF Scientific Database Projects, 1991-1993.* W. W. Chu, A. F. Cardenas and R. K. Taira, eds., pp. 16-22. National Science Foundation, Washington, D.C. (1993)

new machines is faster because prototypes can be designed and thoroughly tested before a physical device is built. New discoveries, both of general scientific interest and of specific applications, often result, since the old truism applies: one doesn't really understand something until one can take it apart, put it back together *and* make it better. However the differences between biological and conventional machines are so great that one cannot simply use existing CAD tools for biology. Unlike conventional ones, understanding and modifying biological machines is a matter of balancing the relative efficiencies of various alternative processes in the context of the cell's overall systems. This means that the database of information on the parts and their properties that the biological CAD tool (bioCAD) will use to help scientists and engineers try out new ideas must store the kinds of information which are important in biology, a very difficult task with existing techniques.

Our research is dedicated to developing the scientific and engineering infrastructure necessary for bioCAD, beginning with database technology. Since the subcellular machinery of even the simplest organisms is so complex, we've chosen to focus on just one part which controls how carbohydrates – simple and complex sugars and starches – are digested and used by the cell. This part of metabolism, the biochemical reactions which build and degrade all of the chemical molecules required by living things, is very important in its own right. Many diseases which affect humans, from birth defects to diabetes, are due to altered carbohydrate metabolism. Many drugs are changed by these reactions, and it's the storage of carbohydrates in certain plants that helps make them important food crops. Since we want to store information which will allow scientists to simulate the effects of changing portions of carbohydrate metabolism, we need to decide both what kinds of information to store and how to store it. We do this by considering a set of questions which scientists pose experimentally when they're trying to understand how metabolism works. The most fundamental is to trace the atoms through the successive molecules in a metabolic pathway. To understand this experiment, consider the U.S. transportation system, a network of roads, rail lines, air routes and navigable waterways on which many different conveyances transport people and goods. Imagine an electrical map of the country showing all these different routes, their connections, the present traffic and the contents of the conveyances. Tracing the driving of a new car home from the dealer's lights up not only the people in the car, the car itself, the act of driving, and the route, but also the assembly of parts in manufacturing plants, the manufacture of parts from steel, glass and plastic, the making of those substances from the raw materials, the refinement of petroleum, and the transport of all these objects. This set of people, parts, acts and routes is a subnetwork of the transportation system. A metabolic pathway is a set of biochemical molecules and reactions which transform the different molecules into each other. The molecules are objects, like the car and its parts, and the reactions are the chemical transformations which degrade and build the molecules from each other, like the manufacturing and driving of the car. All the molecules and their reactions form a network of metabolic processes, like the transportation network, and a pathway is a subnet of that whole, like the subnet from parts in manufacturers' plants to the car in the buyer's driveway. Tracing the atoms in a metabolic pathway is rather like tracing the parts of the car back to their origins, except that the parts *change each other and help build the car*.

Scientists solve this problem "manually" for one pathway by knowing the molecules' structures and the reactions' chemistry. But in the cell several pathways compete for each molecule, and a molecule can contribute atoms to other molecules, which can recycle these atoms further. Thus answering this question realistically by manual methods is extremely tedious and error-prone. It is also the foundation for other questions directly relevant to engineering biological machines, such as the effects of altering available pathways by drugs. Our task is to develop database representations – ways of encoding information in the computer – for molecular structures, reaction chemistry, and

reasoning steps which capture the important information in a way that make sense to biologists.

## 2 Methods

Because computational efficiency is important, it is essential that the representations be in the same language. This view is very different from that of traditional databases, which separate data (molecules), process (reactions) and queries (questions or reasoning) into rigid categories, each expressed in a different computational language. An analogy would be an entirely English poem with a rich but consistent set of metaphors versus one in English, Finnish and Chinese, each language using a different, largely unrelated set of metaphors. By requiring our representations to be symmetric in this way, we make it easier to translate information contained in the biological literature into the database, and to use terms already familiar to biologists.

We achieve symmetry using two techniques. First, the database is implemented using deductive technology. Its virtues are the abilities to express data, processes and queries in the same language, Prolog, and to easily specify the deductive reasoning required to solve a problem. So the database can store not just hard facts but the concepts biologists use to understand the natural world. For example, the solution of the trace-the-atoms problem requires knowing the ways molecules are rearranged in reactions; the equations which describe how fast the reactions occur; and how to put these ideas together, reaction after reaction. Deductive technology allows us to express ideas and chains of reasoning symmetrically to facts, and to use them to carry out the deductions which simulate the behavior of the biological machine. Thus the database is both the simulation system and the information repository – the bioCAD tool.

The second technique for achieving symmetry is to mimic the human language biologists use to describe biochemistry. Molecules are composed of groups of atoms, each having different chemical properties, and their structure is most often described in terms of these groups. Reactions occur by altering the various groups in specific ways. The human language uses the name of the group as the root of the verb describing the chemistry which takes place at or with the group: so *carboxylate* means *transfer a carboxyl group*. This and many similar ways is how the human language integrates the ideas of data, processes and queries: the language expresses the laws of chemistry both explicitly, in terms of rules for forming chemical bonds among atoms, and implicitly, by describing the reactions which occur. To represent this, we have extended the techniques for representing human languages like English to biochemistry. The reactions are analogous to sentences, the molecules to nouns, the mechanisms of reactions to verbs, and the constraints on which groups react to modifiers such as adjectives and adverbs. Each of these elements include facts and rules, or a grammar, which specify how groups are combined to make molecules according to the laws of chemistry. The elements and computational language used to describe molecular structure in terms of groups are also used in describing the mechanisms of the reactions, the choice of groups, and to pose queries to the database, thus achieving in practice the goal of a symmetric representation.

Just as an English grammar is used to determine the role of words in sentences, to generate proper sentences and to test specimen sentences for validity, we can use the biochemical grammar to describe the roles groups and molecules play in reactions, to generate members of a family of reactions, and to determine if reactions meet particular chemical criteria. This latter includes both checking for accuracy of data in the database and the development of new ways to understand reactions by searching for similarities among them. In addition to fulfilling our goal of symmetry, the representation is compact (relatively few rules specify the myriad combinations of molecules and reactions), modular (increases in scientific knowledge are easily incorporated into the database),

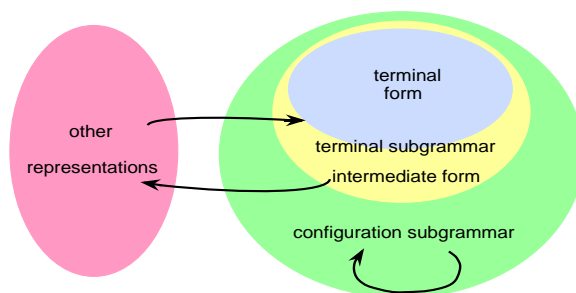


Figure 1: A layered grammar.

and flexible (providing data checking and exploration in addition to simulating biological machines). Searching for common features among the thousands of reactions known is particularly important, since these can suggest appropriate ways to modify particular subcellular biological machines.

### 3 Results

Since molecules and their associated ideas are most fundamental, we have focussed on these first [1]. The overall configuration of the molecule in terms of substituent groups is specified by a configuration fact, which allows simple, high-level specification of many molecules. This configuration fact can be parsed by the transformational grammar to produce the complete, stereochemically correct specification of the molecule as a series of terminals detailing the atom-by-atom, bond-by-bond connections in three-dimensional space (the terminal form). The configuration subgrammar recognizes alternative descriptions of the same molecule, so that a range of user preferences in description is accommodated. We can write rules to describe classes of molecules, and use these to generate all molecules in the class. This is particularly important because it can be used to represent common and variable features of molecules which help determine the specificity of enzymes and which can be used to suggest candidate enzymes for modification. The configuration facts can also refer to each other to describe new molecules in much the same way as biochemists do, by naming a parental molecule which provides the pattern and enumerating the deviations from the parent. This metaconfiguration facility will permit rapid expansion of the number of molecules contained in the database. We have successfully tested the terms configuration subgrammar on a wide variety of biochemically important molecules.

The configuration and terminal forms are interconverted by the necessary generation and recognition predicates and serve different functions (Figure 1). They are connected by an intermediate form which renders group and connectivity information as a set of edges, and points back to the terminal form for isotopes and forward to the configuration form for spatial relationships. The use of multiple representations allows us to choose which is most suitable for a given query, and has helped keep computational intensity low. Humans are trained to view molecules by groups, so a group-based form is essential for rapid molecule specification and the description of molecule classes. We also use it to describe reaction biochemistries since it is so compact and expressive. The terminal form stores information on relative spatial position, atomic isotope, bond type and atom number in a form which can be easily manipulated by Prolog. The intermediate edge form is quite useful for direct topological comparison of molecules; we have used it to identify substructures

common to molecules in reactions. This is one view of the trace-the-atoms problem, and relies on topological identities between two molecules rather than descriptions of reaction biochemistry.

In the course of developing the molecule representations, we have used the grammar to check the accuracy of our molecule facts in the database. This is extremely important because no ready mechanism for the automated validation of large databases of biochemical reactions exists. It also permits the identification of particular molecules as members of a set of related molecules, thus more accurately expressing the notion of broad enzyme specificity.

On the process dimension, we have been defining the types of information conveyed and sorting out the boundaries between the formal chemistry and the detailed mechanisms of the reactions. This is important because the formal chemistry can be defined by inspection of the substrates and products, while the mechanism requires extensive experimentation and cannot be predicted. Both views are important because they reflect different degrees of knowledge, types of experiments, and scientific needs. Viewing reactions by their net chemical effect allows one to specify the final result the enzyme must achieve, thus constraining the possible enzymatic mechanisms. In the absence of other information, knowledge of the net effect can be used to classify reactions [2] and to identify possible similarities among enzymes based on their function. Such similarities can suggest suitable targets for protein engineering, explanations for physiological differences among species, and which enzymes might react with a novel compound such as a drug. The definitive catalytic mechanism is important because it reveals further similarities [3] and provides specific criteria for protein engineering and drug design. For example in an bioengineering problem which requires modification of one or more pathways to produce a novel compound, knowing the chemical effect would allow one to suggest which enzymatic steps might be modified; similarly knowing the enzymes' mechanisms would suggest specific modifications. The boundary between the two types of information is not sharp, and common classification schemes tend to blur it [2, 3]. Our efforts to distinguish formal chemistry and reaction mechanism have prompted some ideas for improving the classification of enzymatic reactions, both in accuracy and in expressing multiple scientific views [4].

We have articulated formal biochemical effects and mechanisms for all the reactions in the model pathway, and are now implementing the transformational grammar to carry out the reaction chemistries using mechanistic information. This grammar describes the reaction using the terms of the compound configuration form, and implements the rearrangements as a series of rules. Thus the representations for data and process (and constraints, see below) are symmetric in the database, just as they are in the "natural language" of biochemistry. Using the description of formal biochemical effect to trace the atoms allows an approximate answer to the query by employing the metaconfiguration facility of the configuration subgrammar. Expressing the mechanism allows a more precise answer, since nearly all reactions occur by a series of steps involving other compounds: while no net global change occurs, the series of individual steps can produce either molecular rearrangements where one identical group is substituted for another or a distribution of labelled atoms to other participating molecules. These changes are invisible at the level of pure structure, but quite apparent and important when the process of the making the structures is considered. We are also developing methods to use information on enzymes and biochemical reactions contained in other databases directly in our computations.

In our current research the constraints – enzyme specificity – are relatively trivial. There is now no theory to let one predict enzyme specificity, so we represent these constraints as facts, again using the language of the configuration form. The set of such facts for a reaction are called by the predicates carrying out the chemistry, and each must be satisfied for the reaction's "proof" to

succeed. We have used the same approach previously in representing ionization reactions (Kazic, unpublished). As we come tackle more complex questions, such as those involving reaction rates and genetic modifications, the constraints will become more complex to express and reason with.

We are completing the development of the components and the grammar, but we have already used the underlying theory and our results to solve some important technical problems. The molecule component allows scientists to pick that description of molecules which most facilitates answering a particular query, just as people normally do to simplify problem-solving [5, 1]. It automatically converts among the different descriptions, an important tool in checking databases for the accuracy of their encoded information [4]. Studying how to express chemical mechanism has helped illuminate several common confusions in the classification of biochemical reactions [2, 3], and allowed us to offer suggestions for the automated classification of reactions and the exploration of similarities among them [4]. We have used these description to answer the trace-the-atoms problem by comparing molecular structures directly, a first approach to solving the problem.

## 4 Conclusion

BioCAD tools are important for accelerating the rate of both scientific discovery and biotechnological innovation. The problem we are studying and the techniques we are developing form the foundation for such tools. Deductive technology allows us to design databases which more closely match the language and needs of the biological community and which serve as both an information resource and a modeling tool. The representations of data, processes, constraints and queries we are developing express the detailed biochemical knowledge required for successful, realistic simulation of biological processes. By laying the foundation for bioCAD tools, we hasten the day when biological machines can be systematically understood and rationally engineered.

## References

- [1] Toni Kazic. Reasoning about biochemical compounds and processes. In Hwa W. Lim, James W. Fickett, Charles R. Cantor, and Robert J. Robbins, editors, *Second International Conference on Bioinformatics, Supercomputing and the Human Genome Project. Proceedings*, pages 35–49, Singapore, 1993. World Scientific.
- [2] International Union of Biochemistry and Molecular Biology. *Enzyme Nomenclature. Recommendations (1992) of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology*. Academic Press, Inc., London, 1992.
- [3] Christopher Walsh. *Enzymatic Reaction Mechanisms*. W. H. Freeman and Co., San Francisco, 1979.
- [4] Toni Kazic. Biochemical databases: challenges and opportunities. In P. S. Glaeser and M. T. L. Millward, editors, *New Data Challenges in Our Information Age. Proceedings of the Thirteenth International CODATA Conference*, pages C133–C140, Paris, 1994. CODATA Secretariat.
- [5] Toni Kazic. Representation, reasoning and the intermediary metabolism of *Escherchia coli*. In Trevor N. Mudge, Veljko Milutinovic, and Lawrence Hunter, editors, *Proceedings of the Twenty-Sixth Annual Hawaii International Conference on System Sciences*, volume 1, pages 853–862, Los Alamitos CA, 1993. IEEE Computer Society Press.

# Modeling Biology Realistically

