Approaches to Encoding Biochemical Function

Toni Kazic
Institute for Biomedical Computing
Washington University, St. Louis, Missouri USA

In biological systems there is an intimate relationship between structure and function. The structure of molecules, large and small, determines how they function in the cellular milieu; natural selection acts on the basis of biological function, so that structures performing those functions change during evolution; and it is the functioning of many structures which cause or permit mutation to occur, supplying the raw material for evolutionary selection. Contemporary molecular biology strongly emphasizes the determination of primary structure, especially of DNA, and much of computational biology is occupied with the analysis of the resulting sequences. While neither is inappropriate, it is wise to remember that the answer to the question, "how does biology work?" is to be found at least equally by an analysis of the functional properties of molecules. The history of the last century of biochemical and physiological investigation is extraordinary, and has constructed a clear if incomplete outline of biological systems. Such work is directly responsible for nearly all advances in medicine, agriculture, and industry in which a rational approach has been attempted, and most of this has proceeded with relatively crude structural knowledge.

Yet if functional analysis is critical to understanding, the state of the experimental and computational arts are lagging compared to those for structural determination. On the experimental side, there is no methodology comparable to high-throughput DNA sequencing: the elucidation and quantitation of function are necesarily highly individualized to the system under study, and do not readily lend themselves to assembly-line tactics. The fruits of research are now only faintly captured electronically, and must be recovered from a direct examination of the scientific literature. In contrast, databases of both primary and tertiary structures are growing daily, and are often a sufficient substrate for computations. On the computational side, the lack of suitable data directly reflects the difficulties of their representation, which in turn reflects the intricacies of the structure-function relationships involved. If one is to express function in ways which support advanced computations, such as pattern recognition and classification, it is insufficient to represent molecules by name tokens or reactions by arbitrarily chosen verbal descriptors: one must instead represent these things more concretely, closer to the way they are in nature.

We are exploring the issues in representing function by attempting to represent glycolysis, a highly conserved biochemical pathway responsible for much of the anaerobic degradation of sugars containing six carbons. For our purposes, a reaction is a biochemical transformation of at least one substrate (input molecule) to at least one product (output molecule); we define a biochemical pathway as an arbitrary set of reactions which together with their substrates and products form a connected graph[1]. Our goal is to be able to trace the fate of each atom of a starting molecule to a product molecule, where each molecule is arbitrarily chosen and both are separated by an arbitrary number of reactions: we refer to this as the "trace-the-atoms" query. Since it is ultimately a question about the fates of molecules, we have named our project *Moirai* after the trio of ancient Greek goddesses who determined human and divine fates. Computing the query requires

[1] As just described, the graph is bipartite in nodes (reactions and molecules) and arcs (substrates and products), but one is obviously not limited to this presentation. Adding catalysts makes the graph tripartite in arcs (substrate, product, catalyst), and one can contract or uncolor the graph arbitrarily to simplify its presentation.

symmetric representations of three categories of information: the molecules (objects); the reactions (processes); and the specificities of the enzymatic catalysts (constraints). From the proceeding discussion it should come as no surprise that successful representation depends on our mastery of the inherent complexities of the information — for example, structure-function relationships — but many of these operate as implicit assumptions which accumulate over the course of the many years required for a thorough biological education. As a result, they tend to be deeply buried and to require information from many disciplines for their expression. We have found the philosophical approach and methodology of logic programming to be invaluable.

Biochemists describe what reactions do by naming the molecules, which parts of those molecules, and the type of biochemistry involved in the reaction. The "natural language" of biochemistry includes many verbs formed from nouns describing parts of molecules: "to acetylate" is "to transfer a acetyl". Thus any attempt to represent biochemical function must necessarily begin with representation of the structures of the molecules. We must also be to able to efficiently indicate which parts of the molecule are involved in the reaction, and describe the changes in the molecule's structure which occur during the course of the reaction. These needs militate against representations which individually enumerate all atoms and bonds of a molecule for expressing reactions, and favor a representation which summarily describes the molecule in terms of its substituent groups and which can be automatically parsed to other representations as required. We have developed a stereochemically accurate representation which meets these criteria by encoding a graph grammar which parses a high-level molecular description into a family of equivalent representations. In effect the grammar is the intensional representation of the generated ones. The grammar and the compound representations constitute *Klotho*, the first component of *Moirai* and the fate responsible for spinning the thread of life. The high-level extensional description and some of the generated representations for over three hundred compounds are available on the World-Wide Web (http://ibc.wustl.edu/klotho/). Because the high-level description is written in terms of the structural groups which compose the molecule, one can use *Klotho* to briefly and accurately point to those groups which participate in a reaction's biochemistry, then parse the high-level description into one in which individual atoms, bonds, and electrons can be recognized and manipulated.

Now that *Klotho* is sufficiently mature, we have turned our attention to describing how the structure of the molecule is manipulated during a reaction — the reaction's mechanism. We have found ourselves excavating the many layers of meaning which biochemists invest in the notion of "reaction", and the fact that the meaning of each layer depends on others. To capture these complexities, we have partitioned the overall biochemical reaction into three informational dimensions which describe the reaction's chemistry, kinetics, and enzymatic mechanism. It turns out that for many reactions, a knowledge of the reaction chemistry is sufficient to trace the atoms; in other instances, detailed expression of the enzymatic mechanism is required. To a first approximation each dimension can be represented as a set of biochemical equations equivalent to those used to represent the overall reaction. However the equations and compound structures are insufficient for describing how the reaction occurs, since by themselves they give no indication of what roles each molecule plays in the chemistry: the reactions are complex enough, and our knowledge insufficient enough, that this information cannot reliably be deduced *a priori* for all cases. There can be many possible mechanisms and new ones are being discovered. Biochemists have long circumvented this problem by supplementing equations and structural diagrams with graphical signs and textual commentary to abstract the experimental results into models of how reactions occur. So the problem of representing mechanism devolves into two parts: how to partition the information so that the dimensions can be expanded and contracted in a biologically sensible way; and how the supplementary information can be expressed. We have named this component of *Moirai Atropos*, after the goddess who spun the thread into the fabric of one's life.