Can computers help to explain biology?

The road leading from computer formalisms to explaining biological function will be difficult, but **Roger Brent** and **Jehoshua Bruck** suggest three hopeful paths that could take us closer to this goal.

It doesn't require vast prophetic vision to identify developments in computers and information technology that will greatly affect the practice of biology. By 2020 we expect that biologists will use computers, numerous 'omic' data types (ref. 1) and a greatly expanded biological literature to design experiments, generate and analyse new data, and think about their own work. But we will leave forecasting about PubMed and Google, metadata and the semantic web to others. Instead, we wish to consider some of the formalisms offered by computer science[2] that developed alongside computing machines. The search for biologically relevant formalisms has a chance to greatly affect the understanding of biological function, in ways we are just starting to imagine.

Today, by contrast with descriptions of the physical world, the understanding of biological systems is most often represented by natural-language stories codified in natural-language papers and textbooks. This level of understanding is adequate for many purposes (including medicine and agriculture) and is being extended by contemporary biologists with great panache. But insofar as biologists wish to attain deeper understanding (for example, to predict the quantitative behaviour of biological systems), they will need to produce biological knowledge and operate on it in ways that natural language does not allow.

Living computers

We begin with what we know in 2006: the trajectory of living systems through developmental time and space is highly determined by the actions and interactions of functional molecules encoded by their genomes. These encoded molecules are further influenced by external perturbations. Because the dynamic behaviour of biological systems is highly determined by a central stored program, living systems differ profoundly from all other naturally occurring, time-evolving systems. The weather has no genome.

Some aspects of biological systems, such as the sequence of encoded proteins (which determines their structure), arise directly from the genome. But others, including most biological functions, arise from the genome by considerably more complex routes, with the consequence that function typically occurs simultaneously at multiple levels[3,4]. These levels include the biochemical activity of an individual protein, the function of that protein in cellular processes involving other proteins, and the developmental trajectory of those processes within a multicellular organism[4,5]. None of these levels is more true or fundamental than the other.

If biology and information science continue with business as usual, then, by 2020, most of the natural-language stories of 2006 about biological function will be subsumed into more sophisticated narratives, which will be better organized and accessed by computers. But the outlines of most of these stories will probably remain unchanged. Here, however, we imagine ways that formalisms from computer science might contribute to a deeper understanding of biological function. These approaches will not bear fruit without deliberate and difficult work.

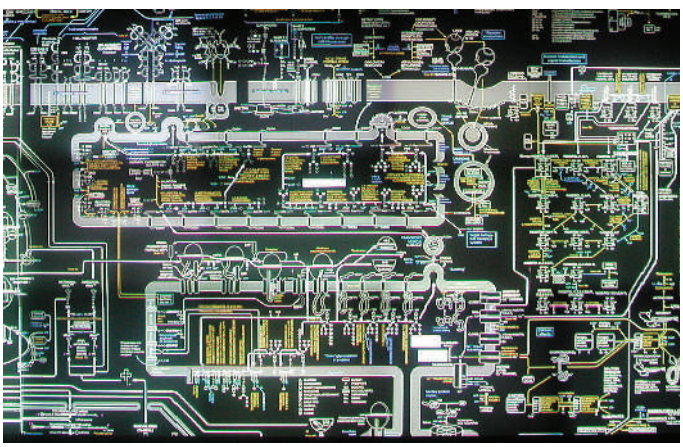
The development of computer science required both new formalisms to capture reasoning in natural languages and ways to implement those formalisms in physical devices[6]. In 2006, it seems reasonable to compare living systems to 'von Neumann', or stored-program, computers, with processing systems (here encoded by the stored program), various external and internal inputs, and outputs in the form of execution. In this view, the biological system is not primarily a factory or a chemical plant but an assemblage that takes information, processes it, decides and executes.

As yet there is no theory that can specify the meaning or purpose of a string of computer code, but Sussman has suggested how elements of such theory might arise[7]. He points out that mathematics had its roots in a workaday human activity, that of Egyptian surveyors redefining the boundaries of fields after the Nile floods receded. Rigorous thinking about this activity led eventually to mathematics: geometry, trigonometry, algebra and beyond. In the same way, a workaday activity — the design and use of procedural imperative languages to write code ('do this, now, do that, if such a thing happens, then do this!') to program computers — may lead to new formalisms describing information processing and eventually to new mathematics.

Blurred boundaries

In biological systems it seems reasonable to view the DNA script in the genome as executable code, code that could have been specified by a set of commands in a procedural imperative language. And in the same spirit, we can view any signal-transduction pathway as a collection of protein machines that takes inputs from inside and outside the cell, performs processing operations on those inputs to arrive at decisions, and communicates those decisions to an apparatus that executes it.

However, to make the analogy between biological systems and von Neumann computers is to reveal important differences between them (Fig. 1). At the level of cells and organisms, biological systems differ from computers in many ways, including (but not limited to): lack of modularity and boundaries in code; lack of fixed order of execution in code; self-



ARGONNE NAT'L UNIV.

Biology needs to move beyond natural-language descriptions of biomolecules and pathways. Adopting new formalisms from computing may lead to greater insight than even graphical displays (such as this wall chart) can offer.

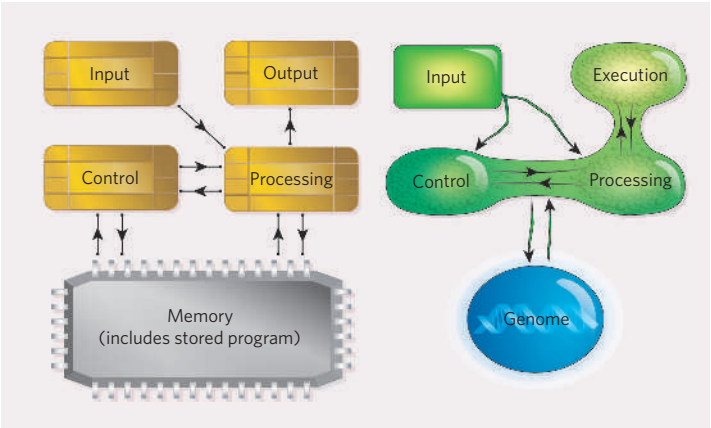


Figure 1 | Biological systems (right) have similarities and differences to von Neumann computers (left). In biological systems there is no distinction between processor and output, as function, phenotype and selection act at many levels.

assembly of encoded components; lack of intelligible sentient design; and lack of crisp boundaries between memory, processor, input and output components.

Most importantly, biological systems usually lack a clear boundary between processing apparatus and output. This distinction arises because function in biology is a consequence of selection, and selection usually acts at many different levels. Thus, the simplest human question 'what does the system do?' (which translates into 'what was the system selected for?') usually has simultaneous multiple correct answers. This fact will continue to frustrate analyses of biological systems in terms of the 'objective functions' they are 'optimized' to 'execute'. Because of this difference, it is unlikely that even a mature theory of stored-program machines will be adequate to explain biological systems. However, even in the absence of grand theory, one can work on intermediate steps. Here we describe three avenues worth exploring.

One fruitful approach formalizes cause-and-effect relationships between named proteins and regulatory sites by translating these into defined chemical reactions undergone by defined molecular species. These reactions can be modelled as differential equations constrained by the rules of chemical kinetics, more formally codified as the 'chemical master equation'[8]. In biology, differential-equation models have a mixed history; they were vital for understanding transmission of the nerve impulse[9] and for helping to identify reaction types before the channel molecules were discovered, but were less successful in circadian-rhythm research until biologists identified molecular entities and relevant reactions.

Measure of meaning

For most biological narratives, the resulting sets of differential equations are too complex to be analytically tractable. But their dynamic behaviour can be approached down a second path — by simulating approximate numerical and stochastic methods[10]. These simulations already constitute 'theory', in the narrow sense that they can generate hypotheses that can then be tested by direct experiment. Equally important, they have inspired mathematicians and computer scientists to apply existing means to reduce complexity and seek new ones. For example, biological reaction networks do not have an order of execution, but probabilistic methods can be used to explore the most likely chains of reactions executed by a given network (M. Riedel and J. Bruck, personal communication).

A third path to better understanding of function begins with deeper analysis of the natural language now used to describe it. The cause-and-effect stories of function of proteins and regulatory sites use an impoverished vocabulary: many proper nouns, few verbs and some prepositional phrases denoting location. Like information in Geographical Information Systems, which also have a limited vocabulary, biological narratives of cause and effect are readily systematizable by computers. There are at least four commercial companies working to provide such systematizations, which are already providing some insight[11].

"We imagine ways that formalisms from computer science might contribute to a deeper understanding of biological function."

But for biological function, just beyond cause-and-effect narratives and before the ultimate truths of fitness and selection, there lies a muddy patch of ground known as 'teleology'. Teleology is hard to avoid: it is difficult to explain why the lens of the eye is transparent without at some point mentioning that the eye is 'for' seeing. But in that mud there may be hope.

The twentieth-century architects of information theory[6] deliberately restricted their concept of information because they were limited by their ability to define and measure it. Information theorists wanted to build a theory that involved the meaning (the semantic content) of messages, but could not measure meaning in sender, recipient or at any point in between. Instead, they chose a meaning for information that was restricted to the carrying capacity of communications channels such as telephone lines — the information technology of their era.

Similarly, biologists would like to cast their descriptions in terms of meaning and purpose, but are limited in their ability to measure those things. As we have said, biology does offer a clear definition of meaning ('it was selected'), but the multiple levels at which selection acts means that meaning is always difficult to determine.

Deeper understanding

Happily, there is considerable interest in wanting to build one element of biological semantics — the passage of time — into information theory. Formalizations of information processing that embodied this and other semantic concepts relevant to biology might help biologists to go beyond quantifying reaction rates and molecular species of biological systems to understand their dynamic behaviour. They might also help to suggest new experiments — perhaps on synthetic biological systems engineered to have a crisper division between process and output, which could then be evolved by artificial selection. This approach might bring a deeper understanding of function at its most fundamental level of fitness and selection.

However marvellous developments in computation are by 2020, if their impact is limited to information generation, handling, visualization and integration, it will mean that their potential contribution to a more predictive understanding of biological function will have failed. By laying out three paths from current computer science that might lead to deeper insights, we at least hope to stir things up. But we also observe growing frustration with business as usual. If we knew better how biological systems worked, we could better perturb existing ones (such as ours, for human medicine) and we could design and build better ones. The fact that both possibilities and frustrations are now starkly evident should make the next 16 years interesting indeed. ■

Roger Brent is at The Molecular Sciences Institute, Berkeley, California 94704, USA; Jehoshua Bruck is at the California Institute of Technology, Pasadena, California 91125, USA.

1. Weinstein, J. N. *Science* **282,** 687 (1998).
2. Simon, H. R. *The Sciences of the Artificial* 1st edn (MIT Press, Cambridge, Massachusetts, 1969).
3. Fields, S. *Nature Genet.* **15,** 325–326 (1997).
4. Kirschner, M. W. *Cell* **121,** 503–504 (2005).
5. Brent, R. *Cell* **100,** 169–183 (2000).
6. Shannon, C. E. & Weaver, W. *The Mathematical Theory of Communication* (Univ. Illinois Press, 1948).
7. Sussman, G. J. *The Legacy of Computer Science* NCR/CSTB Symposium on the Fundamentals of Computer Science (2001).
8. Gillespie, D. T. *Physica A* **188,** 404–425 (1992).
9. Hodgkin, A. L. & Huxley, A. F. *J. Physiol. Lond.* **117,** 500–544 (1952).
10. Gillespie, D. T. *J. Phys. Chem.* **81,** 2340–2361 (1977).
11. Calvano, S. E. *et al. Nature* **437,** 1032–1037 (2005).

Acknowledgements We are grateful to L. Lok, K. Tahashashi, D. Endy, O. Resnekov, P. Rabinow, D. Gillespie, M. Cook and M. Riedel for useful discussions and comments on the manuscript.