6

A Computational and Engineering View of Biology

Because 21st century biology is very concerned with function, it is helpful to have abstractions available that characterize the functionality of interest. By doing so, insights derived from study of those abstractions in other contexts become available for biological use. In addition, because biological systems are the products of eons of evolutionary history and decision making, viewing them through the lens of engineering yields insights that are not otherwise available from an analysis that might be based on first principles.

6.1 BIOLOGICAL INFORMATION PROCESSING[1]

As noted in Chapter 2, biological systems are extraordinarily complex—and partly as a consequence, poorly understood. Yet it is clear that biological systems demonstrate and exemplify functionality at different levels.

Artifacts such as computer hardware and software also exhibit functionality and multiple levels. To facilitate the understanding and construction of such artifacts, computer science has developed information abstractions that seek to capture and encapsulate certain kinds of functional behavior in manipulating and managing information; such abstractions are a primary focus of study of the computer scientist (Box 6.1).

One key connection to 21st century biology is that many biological problems now require the simultaneous consideration of phenomena at different scales. For example, biologists can think of genetics at the level of individual nucleotides, at the level of chromosomes, at the level of genomes, and at the level of populations. From nucleotide to population is a span of many orders of magnitude, and it is difficult to conceptualize such a range without moving seamlessly between different levels of abstraction.

Section 6.1 describes several such abstractions and their specific biological applications already in use, but the description is not intended to be exhaustive, and there are likely many more such abstractions capable of providing biological insight, including new or as yet undiscovered techniques or concepts. As such, this area represents opportunities for both biologists and computer scientists.

[1] Much of the discussion in Section 6.1 about cells as information-processing devices is adapted from R. Aviv and E. Shapiro, "Cellular Abstractions: Cells as Computation," *Nature* 419:343, 2002.

Box 6.1
On the Abstractions of the Computer Scientist and Engineer

Abstraction is a generic technique that allows the scientist or engineer to focus only on certain features of a system while hiding others. Scientists in all disciplines typically use abstractions as a way to simplify calculations for purposes of analysis, but computer scientists also use abstractions for purposes of design: to build working computer systems. Because building systems is the central focus of much work in computer science, the use of abstractions to cope with complexity over a wide range of scale, size, and levels of detail is central to a computer scientist's way of thinking.

The focus of the computer scientist in creating an abstraction is to hide the complexity of operation "underneath the abstraction" while offering a simple and useful set of services "on top of it." Using such abstractions is the principal technique for organizing and constructing very sophisticated computer systems, and they enable computer scientists to deal with large differences of scale. For example, one particularly useful abstraction uses hardware, system software, and application software as successive layers on which useful computer systems can be built. This illustrates one very important use of abstraction in computer systems: each layer provides the capability to specify that a certain task be carried out without specifying *how* it should be carried out. In general, computing artifacts embody many different abstractions that capture many different levels of detail.

A good abstraction is one that captures the important features of an artifact and allows the user to ignore the irrelevant ones. (The features decided to be important collectively constitute the interface of the artifact to the outside world.) By hiding details, an abstraction can make working with an artifact easier and less subject to error. But hiding details is not cost-free—in a particular programming problem, access to a hidden detail might in fact be quite helpful to the person who will use that abstraction. Thus, deciding how to construct an abstraction (i.e., deciding what is important or irrelevant) is one of the most challenging intellectual issues in computer science. A second challenging issue is how to manage all of the details that are hidden. The fact that they are hidden beneath the interface does not mean that they are irrelevant, only that the computer scientist must design and implement approaches to handle these details "automatically" (i.e., without external specification).

SOURCE: Adapted from Computer Science and Telecommunications Board, National Research Council, *Computing the Future: A Broader Agenda for Computer Science and Engineering,* National Academy Press, Washington, D.C., 1991.

Consider that biological processes, such as catalysis, protein synthesis, and other metabolic systems, are consumers, processors, or creators of information. As Loewenstein puts it, in biological systems, "in addition to flows of matter and energy, there is also flow of information. Biological systems are information-processing systems and this must be an essential part of any theory we may construct."[2] Sydney Brenner goes farther, arguing that ". . . this information flow, not energy per se, is the prime mover of life—that molecular information flowing in circles brings forth the organization we call 'organism' and maintains it against the ever-present disorganizing pressures in the physics universe. So viewed, the information circle becomes the unit of life."[3]

The current state of intellectual affairs with respect to biological information and complexity may have some historical analogy with the concept of energy at the beginning of the 19th century. Although the concept was intuitively obvious, it was not formally defined or measured at that time. Carnot's analysis of the performance of steam engines formalized the meaning of energy, creating the basis for

[2] W. Loewenstein, *The Touchstone of Life: Molecular Information, Cell Communication, and the Foundations of Life*, Oxford University Press, New York, 1998, p. xiv.

[3] S. Brenner, "Theoretical Biology in the Third Millennium," *Philosophical Transactions of the Royal Society B* 354(1392):1963-1965, 1999.

the science of thermodynamics. Only after energy had been identified and studied in the artificial realm of steam engines was it recognized as a prime aspect of natural systems as well.

Similarly, the existing state of the theory of biological information (or, indeed, information of any sort) is based on the work of Claude Shannon, who studied the processing of information in human technological channels of communication, and the field of computational complexity, which was created to analyze the performance characteristics of algorithms running on human-built computers. However, just as thermodynamics successfully widened its scope to the natural world from steam engines, information and computation theory may become a powerful lens for describing, measuring, and understanding processes in the natural world.

Biological information is likely to have a close relationship to information in the Shannon sense of the term, if only because biological entities depend on information to coordinate their internal activity. Cells coordinate their internal activity because they have harnessed intracellular Shannon information channels. Multicellular organisms coordinate their internal activity because they have harnessed intercellular Shannon information channels. These channels are the conduits through which genes transfer their information content to proteins, proteins serve as signaling agents, and nervous systems work. Also, Shannon's insight about the nature of information transmission allows us to understand how signals can reliably be sent through a noisy unpredictable environment (whether cell telephone signals, Internet packets, or hormone signaling proteins) and received accurately at the other end.

On the other hand, Shannon information applies in the strict sense only when it is possible to identify a sender and receiver connected by a channel. There are some places in which this applies, such as the projection of the retina to the brain. Yet in the context of information feedback and loops rather than channels, it is not clear that Shannon information continues to have a well-defined meaning.

There have been a number of attempts to generalize Shannon information to problems at the cellular and subcellular levels, of which the conceptualization by Manfred Eigen of hypercycles, quasispecies, and sequence space is one of the most notable.[4] But whether these concepts are the right ones is not as important as the recognition that new concepts are needed.

A more specific connection between biology and computation can be seen in the biological use of information to enhance the survival and reproductive functions of an organism. That is, biological organisms use information about the environment to stimulate or drive responses that boost the likelihood of survival and successful reproduction. This process is effectively a computation that transforms the inputs (which describe environmental conditions) into the appropriate outputs (the organism's behavior).[5] For example, Hartwell et al. note that signals from the environment entrain circadian biological clocks to produce responses to predicted fluctuations in light intensity and temperature.[6]

Embedded within cells are complex signaling mechanisms that transfer information from one part of a cell to another and intercellular mechanisms that transfer information from one part of a multicellular organism to another. Indeed, signal transduction pathways—and the proteins associated with them—appear to serve the functions of information processing and transfer,[7] rather than those of more "traditional" biology (e.g., chemical transformation of metabolic intermediates or the building of cellular structures).

[4] M. Eigen, "The Origin of Biological Information," presented at the Seventh International Conference on Intelligent Systems for Molecular Biology, August 6-10, 1999; Heidelberg, Germany, available at http://bioinf.mpi-sb.mpg.de/conferences/ismb99/WWW/abstracts/abs-eigen.html.

[5] Indeed, it has been asserted that the history of life can be described as the evolution of systems that manipulate one set of symbols representing inputs into another set of symbols that represent outputs. J.J. Hopfield, "Physics, Computation, and Why Biology Looks So Different," *Journal of Theoretical Biology* 171:53-60, 1994.

[6] L.H. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Murray, "From Molecular to Modular Cell Biology," *Nature* 402(6761 Suppl):C47-C52, 1999.

[7] D. Bray, "Protein Molecules as Computational Elements in Living Cells," *Nature* 376(6538):307-312, 1995. The examples in the next paragraph are also Bray's.

For example, a simple enzyme protein could be viewed as a computational element that takes an input—the concentration of its "substrate," the molecule with which it interacts—and produces an output: a concentration of the catalyzed reaction product. An enzyme that becomes active only when it binds to two separate regulator molecules will function something like a Boolean AND gate, and so on. Circuits formed from these elements can be as simple as a switch or an oscillator, or as complex as to drive a bacterium's chemotaxis response. Indeed, the cell even possesses a kind of short-term, "random-access" memory, in the sense that events in its environment have profoundly shaped the concentration and activity of many thousands of molecules in the cell. In short, these protein-based circuits constitute a kind of nervous system for the cell, providing it with much of what it needs to control its behavior. Box 6.2 provides some additional perspective on this subject.

Additional insights can be gained from the notion that both computational processes and biological pathways can be viewed as processes that affect the state of a system according to well-defined (though possibly probabilistic) rules. Thus, it is possible to describe regulatory, metabolic, and signaling pathways, as well as multicellular processes such as immune responses, as systems of interacting computations operating in parallel. In particular, languages such as Petrinets, Statecharts (discussed in Section 4.3.1), and the Pi-calculus, originally developed for the specification and study of systems of interacting computations, can be used to represent such systems.[8] Such representations enable researchers to simulate their behavior, and to support qualitative and quantitative reasoning on the properties of these systems.

To cite two prominent researchers in this area:

> Processes, the basic interacting computational entities of these languages, have an internal state and interaction capabilities. Process behavior is governed by reaction rules specifying the response to an input message based on its content and the state of the process. The response can include state change, a change in interaction capabilities, and/or sending messages. Complex entities are described hierarchically—for example, if a and b are abstractions of two molecular domains of a single molecule, then (a parallel b) is an abstraction of the corresponding two-domain molecule. Similarly, if a and b are abstractions of the two possible behaviors of a molecule in one of two conformational states, depending on the ligand it binds, then (a choice b) is an abstraction of the molecule, with the choice between a and b determined by its interaction with a ligand process.[9]

Abstractions of the cell as a computing or information-processing device allow one to distinguish between two conceptual levels: a "low-level" view that focuses on implementation (i.e., how the system is built—where the wires go or the detailed molecular processes involved) and a "high-level" view that focuses on functionality (what the system does—analogous to a logic gate or a computational device).[10] For example, one might distinguish between the pathways involved in regulating the circadian rhythm of an organism and its functional behavior as an oscillator.

The difference between these levels of abstraction enables biologically significant comparisons to be made. For example, it would be instructive if two different organisms implemented the same function in different ways. In other words, functional equivalence between related implementations in different organisms could be regarded as a measure of the behavioral similarity of entire systems. (In the literature of evolutionary biology, the implementation of the same function in different ways is called "analogous" implementation.) Perhaps more importantly, a functional perspective is an enabler for the integration of knowledge about the function, activity, and interaction of cellular molecular systems.

[8] R. Aviv and E. Shapiro, "Cellular Abstractions: Cells as Computation," *Nature* 419:343, 2002.

[9] R. Aviv and E. Shapiro, "Cellular Abstractions," 2002.

[10] In many circumstances, different parts of a biological system may play different roles at different times or even different roles at different time scales at the same time. This is especially true in splicing variants, where the expression of a gene may produce proteins with quite different functions according to the behavior of the splicing mechanism. Indeed, in some cases, different splicings have opposite functions. Nevertheless, in understanding a given role at a given time and time scale, the high-level abstraction focused on functionality is meaningful and scientifically significant.

Box 6.2
Role of Computation in Complex Regulatory Networks

Computation . . . [is] a crucial ingredient when dealing with the description of biocomplexity and its evolution, because it turns out to be much more relevant than the underlying physics. Its dynamics is governed mainly by the transmission, storage and manipulation of information, a process which is highly nonlinear. This nonlinearity is well illustrated by the nature of signaling in cells: local events involving a few molecules can produce a propagating cascade of signals through the whole system to yield a global response. . . . If we try to make predictions about the outcomes of these signaling events in general, we are faced with the inherent unpredictability of computational systems. It is at this level where computation becomes central and where idealized models of regulatory networks seem appropriate enough to capture the essential features at the global scale.

Cells are probably the most complete example of this traffic of signals at all levels. . . . The cellular network can be divided into three major self-regulated sub-webs:

- The *genome*, in which genes can affect each other's level of expression;
- The *proteome*, defined by the set of proteins and their interactions by physical contact; and
- The metabolic network (or the *metabolome*), integrated by all metabolites and the pathways that link each other.

All these subnetworks are very much intertwined since, for instance, genes can only affect other genes through special proteins, and some metabolic pathways, regulated by proteins themselves, may be the very ones to catalyze the formation of nucleotides, in turn affecting the process of translation. . . . It is not difficult to appreciate the enormous complexity that these networks can achieve in multicellular organisms, where large genomes have structural genes associated with at least one regulatory element and each regulatory element integrates the activity of at least two other genes. . . .

Luckily, all this extraordinary complexity can be abstracted, at least at some levels, to simplified models which can help in the study of the inner-workings of cellular networks. Overall, irrespective of the particular details, biological systems show a common pattern: some low-level units produce complex, high-level dynamics coordinating their activity through local interactions. Thus, despite the many forms of interaction found at the cellular level, all come down to a single fact: the state of the elements in the system is a function of the state of the other elements it interacts with. What models of network functioning try, therefore, is to understand the basic properties of general systems composed of units whose interactions are governed by nonlinear functions. These models, being simplifications, do not allow one to make predictions at the level of the precise state of particular units. Their average overall behavior, however, can shed light into the way real cells behave as a system. . . .

. . . [M]any entities in cellular networks can be identified as the basic units of regulation, mainly distinguished by their unique roles with respect to interaction with other units. These basic units are genes, each of the proteins that the genes can produce, each of the forms of a protein, protein complexes, and all related metabolites. These units have associated values that either represent concentrations or levels of activation. Their values depend on the values of the units that affect them due to the mechanisms discussed, plus some parameters that govern each special form of interaction. . . . Computer modeling of [the] network [the segment polarity network of *Drosophila melanogaster*] has provided insight into various questions. A very important result is the fact that this network seems to be a conserved module. Evidence for this has been obtained by simulations demonstrating its robustness against the change of parameters. . . .

SOURCE: Reprinted from P. Fernandez and R.V. Sole, "The Role of Computation in Complex Regulatory Networks," Santa Fe Institute Working Paper, 2003, available at http://www.santafe.edu/sfi/publications/Working-Papers/03-10-055.pdf; to appear in a chapter in *Power Laws, Scale-Free Networks and Genome Biology,* Landes Bioscience. Reprinted with permission.

This perspective on cells as computational devices should not be taken as an argument that cells process information the way a digital computer does. The organizations are radically different. To name just a few differences, in a cell there is no clean separation between the data store and the central processing unit: the cell's memory is the same protein reaction network that does its processing. Real proteins rarely respond or act in a completely binary fashion—the levels of concentration matter. Apart from DNA, few portions of a cell's internal machinery are explicitly digital in nature—with the result that signaling in a cell must take place in a highly noisy environment.

It is also interesting that biological function often relies on what might be called exploration with selection—the production of many intermediate products resulting from stochastic subprocesses that are then refined to unique and appropriate solutions.[11] Taken across the entire population, exploration with selection exploits the difference between creating a solution and testing a solution for correctness—the first being in general a much more difficult computational task than the second.[12] Random processes are used to explore the space of possible solutions,[13] and other machinery culls these possible solutions. As Hartwell et al. argue, "Similar messy and probabilistic intermediates appear in engineering systems based on artificial neural networks—mathematical characterizations of information processing that are directly inspired by biology. A neural network can usefully describe complicated deterministic input-output relationships, even though the intermediate calculations through which it proceeds lack any obvious meaning and their choice depends on random noise in a training process."[14]

6.2 AN ENGINEERING PERSPECTIVE ON BIOLOGICAL ORGANISMS

6.2.1 Biological Organisms as Engineered Entities

Engineering insights can be useful in understanding biological organisms as engineered entities, and the rationale for seeking insights from engineering is based on three notions. First, although the physical scales may differ in some cases, human technology and natural systems operate in the same world and must obey the same physical rules. Knowledge that engineering fields have accumulated about what techniques work and the limits of those techniques can serve as a potentially valuable guide in investigating the physical basis of the operations of natural systems. This is especially true for biomechanical feats, such as structural support, locomotion, circulation, and so on.

The second rationale is that because evolution and a long history of environmental accidents have driven processes of natural selection, biological systems are more properly regarded as engineered artifacts than as objects whose existence might be predicted on the basis of the first principles of physics, although the evolutionary context means that an artifact is never "finished" and is rather evaluated on a continuous basis.[15] Both engineered artifacts and biological organisms demonstrate function, embody

[11] For example, the immune system relies on the random generation of pathogen detectors, which are then eliminated when they match some definition of "self." In single molecules, kinetic funnels direct different molecules of the same protein through multiple, different paths from the denatured state to a unique folded structure (K.A. Dill and H.S. Chan, "From Levinthal to Pathways to Funnels," *Nature Structural Biology* 4:10-19, 1997). Within cells, the shape of the mitotic spindle is due partly to selective stabilization of randomly generated microtubules whose ends happen to be close to a chromosome (R. Heald, R. Tournebize, T. Blank, R. Sandaltzopoulos, P. Becker, A. Hyman, and E. Karsenti, "Self-organization of Microtubules into Bipolar Spindles Around Artificial Chromosomes in *Xenopus* Egg Extracts," *Nature* 382(6590):420-425, 1996). Within the brain, the patterning of the nervous system is refined by the death of nerve cells and the decay of synapses that fail to connect to an appropriate target.

[12] This point can be formalized in the language of theoretical computer science. See J. Hartmanis, "Computational Complexity and Mathematical Proofs," pp. 251-256 in *Informatics: 10 Years Back, 10 Years Ahead, 2000,* Lecture Notes in Computer Science, Springer-Verlag, Berlin, Heidelberg, 2001.

[13] For example, random processes are at the heart of stochastic optimization methods that can be used for protein structure prediction and receptor ligand docking, including simulated annealing, basin hopping, and parallel tempering. (An interesting introduction to stochastic optimization methods can be found at W. Wenzel, "Stochastic Optimization Methods," available at http://iwrwww1.fzk.de/biostruct/Opti/opti.htm.) Also, the systematic exploration of ecological models discussed in Section 5.4.8 is also based on the use of random processes.

[14] The quote is taken from L.H. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Murray, "From Molecular to Modular Cell Biology," *Nature* 402(6761 Suppl.):C47-C52, 1999. Hartwell et al. credit Sejnowski and Rosenberg with the neural network example (T.J. Sejnowski and C.R. Rosenberg, "Parallel Networks That Learn to Pronounce English Text," *Complex Systems* 1:145-168, 1987).

[15] A classic paper on this subject is F. Jacob, "Evolution and Tinkering," *Science* 196(4295):1161-1166, 1977.

behavior, and manifest an evolutionary history.[16] Engineered artifacts serve the purposes of their human designers, and biological organisms serve the purposes of nature—that is, to survive and reproduce.[17] Thus, the concepts needed to understand biological function may have some resemblance to some of the concepts already developed for "synthetic" disciplines, of which engineering and computer science are prime examples.

A third rationale is that the engineering disciplines have already had a long history of systems-level thinking and, indeed, have produced artifacts that are arguably approaching biological levels of complexity. For example, a Boeing 777 jetliner contains about 150,000 subsystem modules, including 1,000 computers, a number of the same order of magnitude as the estimated 300,000 different proteins in a typical human cell. Just as in the cell, moreover, these aeronautical subsystems are linked into an immensely complex "network of networks"—a control system that just happens to fly.[18]

A related point, and a key lesson from engineering, is that large systems are built out of smaller systems that are stable. Decomposition of a complex structure into an assembly of simpler structures whose operation is coordinated tends to be a much more successful strategy that building the complex structure from scratch, and this approach can be seen in the structure of the cell. Consider that a human cell has many physical structures within it—nucleus, mitochondria, and so on; each of these can be regarded as a device, many of which compose the cell. Further, many and perhaps even most cellular functions (e.g., genetic regulatory networks, metabolic pathways, signaling cascades) are implemented in a manner that is highly robust against single-point failure (i.e., the function will continue to operate properly even when one element is missing). Section 6.2.3 addresses this point in more detail.

A second view of biological organisms as engineered entities—as novel entities to be constructed by human beings rather than as existing organisms to be understood by human beings—is discussed in Section 8.4.2 on synthetic biology.

6.2.2 Biology as Reverse Engineering

Biological organisms are generally presented to scientists as completed entities, so the challenge of achieving an engineering understanding of them is in fact a challenge of *reverse engineering*. One definition of reverse engineering is "the process of analyzing a subject system with two goals in mind: (1) to

[16] While it is generally recognized that biology and evolution are intimately linked, the analogous connection between engineering and evolution is less well understood. Nevertheless, most human-engineered objects have a lot of historicity in them as well. Most human objects are designs based as improvements on previous designs, not de novo, and this can complicate the understanding of the relationship between functionality and design of a human artifact. One reason is a desire for backward compatibility—consider the fact that two-prong electric plugs and sockets are much more hazardous than some alternative designs and yet they are ubiquitous in appliances today. The same is true for operating systems—later versions of an operating system often incorporate large amounts of code from previous versions to facilitate backward compatibility. A second reason is that previous designs may have solved a design problem in a particularly effective way, and these solutions from the past are ignored today at the designer's peril. For example, consider the evolution of the rotary phone into today's push-button phones. Donald Norman observes that the cradle of the phone handset and the button-switch in it had two distinct functions: the cradle provided a place for the user to put the phone and the button-switch turned the phone on and off. Norman notes that whether deliberately or by accident, the particular design of the rotary phone that placed the on-off switch in a protected spot in the cradle also protected the on-off switch from the user accidentally hanging up the phone. However, the designers of newer push-button phones did not pick up on that feature; many push-button phones are designed so that the on-off switch and the hang-up cradle are separate—thus making the on-off switch much easier to bump and thereby to accidentally disconnect a phone call. See D. Norman, *The Design of Everyday Things*, Basic Books, New York, 1998.

[17] See for example L.H. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Muray, "From Molecular to Modular Cell Biology," *Nature* 402(6761 Suppl):C47-52, 1999, available at http://cgr.harvard.edu/publications/modular.pdf. Hartwell et al. further argue that it is notions of function and purpose that differentiate biology from other natural sciences such as chemistry or physics, and hence that reductionist biology—inquiry that seeks to explain biological phenomena only in chemical or physical terms—is inherently incomplete.

[18] M.E. Csete and J.C. Doyle, "Reverse Engineering of Biological Complexity," *Science* 295(5560):1664-1669, 2002, available at http://www.sciencemag.org/cgi/content/abstract/295/5560/1664.

identify the system's components and their interrelationships and (2) to create representations of the system in another form or at a higher level of abstraction."[19]

A better description could not be developed for the goal of systems biology, even without having to change any words in this definition. And yet reverse engineering, despite being a fairly standard engineering topic, is not taught to biologists.[20] One drawback is that the metaphor itself is foreign to biologists; if they wanted to do engineering of any kind, they would have been engineers. Second, reverse engineering is generally a more difficult task than forward engineering (i.e., the fabrication of a device to implement some specific functionality), and reverse engineering of a biological organism is a particularly difficult endeavor.

One important reason is that reverse engineering is often underdetermined, in the sense that multiple solutions can be developed to account for a given behavior. In such cases, choosing among them thus requires either more data or a priori assumptions about the true nature of the system being reverse-engineered. For example, in dealing with the reverse-engineering task of building detailed kinetic models of intracellular processes from time-series data, Rice and Stolovitzky note that assumptions such as linearity or sparseness or the use of predetermined model structures (e.g., reactions limited in the number of possible reactants and substrates) can help to reduce the non-uniqueness.[21]

A second and even more important reason for the difficulty of reverse engineering is that because of their evolutionary history, the organisms of interest are constructed in a highly nonoptimal manner. When engineers seek to understand how an artifact has been constructed, the basic question they ask is, Why? Why is this structure here? Why was that material used? By asking such questions of a human-engineered artifact, the engineer can often divine a reason that answers them. The reason is that engineers can be expected to design artifacts using principles such as modularity and separation of function (i.e., to minimize unnecessary links between subsystems with different purposes). These principles guard human designs against unforeseen side effects that would arise if components were not deliberately assembled in such a way as to minimize undesired or unanticipated interactions.

However, the same is not true of biological organisms. In many cases, the only answer for biological systems is, "That's the way it was built." Nature builds from accidents that happen to work and creates new mechanisms on top of old ones. While some evolved systems are quite elegant (e.g., the sensory and the motor components of the *Escherichia coli* chemotaxis mechanism), many if not most such systems at least appear to a human as inelegant, redundant, "kludgy," and inefficient—some of them extremely so. Systems engineered by humans, even very poorly engineered ones and even though they too often show their historical origins, are seldom if ever as arcane and kludgy as evolved biological organisms.

Finally, it is helpful to distinguish between two different approaches to reverse engineering. One approach to reverse engineering of biological systems—a "top-down" approach—begins with its observable behavior and characteristics, and seeks to decompose the system into components or subsystems that collectively exhibit the macroscopic behavior in question. That is, the top-down approach is based on a successive decomposition down to the system's most elemental components.

A second approach is based on a "bottom-up" approach, which begins with an understanding of the constituent parts at the lowest level, e.g., the macromolecules and the genetic regulatory networks of the

[19] E.J. Chikofsky and J.H. Cross, "Reverse Engineering and Design Recovery: A Taxonomy," *IEEE Software* 13-17, 1990.

[20] Indeed, the BIO2010 report on undergraduate education in biology (National Research Council, *Bio 2010: Undergraduate Education to Prepare Biomedical Research Scientists*, National Academies Press, Washington, DC, 2003) noted that "one approach to the study of biology is as a problem in reverse engineering. Manufactured systems are easier to understand than biological systems, because they have no unknown components, and their design principles can be explicitly stated. It is easiest to learn how to analyze systems through investigating how manufactured systems achieve their designed purpose, how their function depends on properties of their components, and how function can be reliable even with imperfect components." Also, underscoring the point that engineering is not a part of biology education today, the report emphasized the importance of exposing biology students to engineering principles and analysis in the course of their undergraduate educations. Chapter 10 has more discussion of this point.

[21] J.J. Rice and G. Stolovitzky, "Making the Most of It: Pathway Reconstruction and Integrative Simulation Using the Data at Hand," *Biosilico* 2(2):70-77, 2004.

Box 6.3
Functional Modules in Biology

An important theme in systems biology has been to look for functional modules that have been conserved and reused. The idea of breaking biological systems into small functional blocks has obvious appeal; the parts can be divided and conquered so that the most complex of machines become readily understood in terms of block diagrams or sets of subroutines. Clearly, some conserved modules exist such as the ribosome and the tricarboxylic acid cycle. One method to search for modules involves looking for higher-order structures or recurring sub-networks (often termed "motifs") in metabolic or gene regulatory networks. Another approach mentioned earlier is clustering expression profiles to produce groups of genes that appear to be co-regulated that should ideally reveal the functional modules. However, this assumption does not appear to generalize to all functional groups under all conditions, as some functional groups show well-correlated expression profiles whereas others do not. The low correlation of genes observed within some functional groups has been attributed to the fact that some of these genes belong to multiple functional classes. In another analysis in *E. coli*, 99 cases were found where one reaction existed in multiple pathways in EcoCyc. These observations suggest potential pitfalls with anticipating too much functional modularity in terms of biology being neatly partitioned into non-overlapping modules. Moreover, the tissue- or species-specific differences mentioned earlier may prevent simplistic transfer of modules from one biological system to another. It remains to be seen if biology is as modular as the system biologist might like it to be.

Biological modules may turn out be more interconnected and overlapping than independent in many systems. In addition, the experiences with pathway reconstruction suggest that the combinations of data source produce a more accurate if not more complete characterization of the system under study. These observations point to an eventual need to develop large-scale, predictive models based on a multitude of data sources. For example, metabolic models may combine data from many sources into a quantitative set of equations that can make predictions amenable to experimental verification. In another system, cardiac models can bridge data at multiple levels (i.e. molecular, cellular, organ, etc.) and their corresponding characteristic timescales. In this system, modeling efforts at the single-cell level in the heart suggested a mechanism of increased contraction force that was later confirmed in experimental studies of whole heart.

SOURCE: Reprinted by permission from J.J. Rice and G. Stolovitzky, "Making the Most of It: Pathway Reconstruction and Integrative Simulation Using the Data at Hand," *Biosilico* 2(2):70-77. Copyright 2004 Elsevier.

cells that make up the system. The philosophical notion embedded in the bottom-up approach is that a component is likely to be easier to understand than the system in which it is embedded. By successive assembly of component parts, one is able to create ever-larger assemblies whose operation is understood.

Both approaches seek as their underlying ultimate goal an understanding of how a biological system works in all of its complexity. But they require different strategies for acquiring data at different levels of scale (top-down entails data acquisition at ever-smaller scales, while bottom-up entails data acquisition at ever-larger scales). And also, it should be expected that they will generate different intermediate outputs and products along the way to this ultimate goal.

6.2.3 Modularity in Biological Entities[22]

A functional perspective on biology is centrally based on the notion that biological function is separable, into what might be called modules. The essence of a module—well known in engineering disciplines as well as computer science—is that of an entity whose function is separable from other modules. In the computer science context, a module might be a subroutine upon which various programs can build. These various programs would interact with the subroutine only through the programming interface—the set of arguments to the subroutine that parameterize its behavior. Box 6.3 describes how the search for functional modules plays into systems biology.

[22] Section 6.2.3 is based largely on L.H. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Murray, "From Molecular to Modular Cell Biology," *Nature* 402(6761 Suppl.):C47-C52, 1999.

Important insights into biological organisms can be gained by seeking to identify general principles that govern the structure and function of modules (Box 6.4). In a biological context, a module might be an entity that performs some biochemical function apart from other modules, isolated from those other modules by spatial localization (i.e., it is physically separated from those other modules) or by chemical specificity (i.e., its biochemical processes are sensitive only to the specific chemical signals of that module and not to others that may be present). Furthermore, modules must be able to interact with each other selectively. Specific connectivity enables module A to influence the functional behavior of module B, but not to affect the operation of modules C through Z. Also, the particular pattern of connectivity can account for some emergent properties of these modules, such as an ability to integrate information from multiple sources.

As noted by Hartwell et al., "Higher-level functions can be built by connecting modules together. For example, the super-module whose function is the accurate distribution of chromosomes to daughter cells at mitosis contains modules that assemble the mitotic spindle, a module that monitors chromosome alignment on the spindle, and a cell-cycle oscillator that regulates transitions between interphase and mitosis." When a function of a protein is restricted to one module, and the connections of that module to other modules are through such proteins, it becomes much easier to alter connections to other modules without global consequences for the entire organism.

Modular structures have many advantages. For example, the imposition of modular design on an entity allows a module to be used repeatedly by different parts of the entity. Furthermore, changes internal to the module do not have global impact if those changes do not affect its functional behavior. Modules can be combined and recombined in ways that alter the functionality of the complete system—

Box 6.4
Some Mechanisms Underlying the Structure and Function of Modules

1. Positive feedback loops can drive rapid transitions between two different stable states of a system. For example, positive feedback drives cells rapidly into mitosis, and another makes the exit from mitosis a rapid and irreversible event.[1]
2. Negative feedback loops can maintain an output parameter within a narrow range, despite widely fluctuating input. For example, negative feedback in bacterial chemotaxis[2] allows the sensory system to detect subtle variations in an input signal whose absolute size can vary by several orders of magnitude.[3] (This topic—robustness against noise—is described in more detail in Section 6.2.5.)
3. Coincidence detection systems require two or more events to occur simultaneously in order to activate an output. For example, coincidence detection is central in eukaryotic gene transcription, in which several different transcription factors must be present simultaneously at a promoter site before transcription can occur. (Note the similarity to a multi-input AND gate.)
4. Parallel circuits allow devices to survive failures in one of the circuits. For example, DNA replication involves proofreading by the DNA polymerase backed up by a mismatch repair process that removes incorrect bases after the polymerase has moved on. Both of these must fail before a cell cannot produce viable progeny, and these two mechanisms, combined with a system for killing potentially cancerous cells, reduce the frequency at which individual cells give rise to cancer to about 1 in 10^{15}.
5. Quality control systems monitor the output of many biological processes to ensure that the processes have executed correctly. Such systems can be seen in cell-cycle checkpoints, DNA replication and repair, choices between cell survival and death after insults to cells, or quality control in protein folding and/or sorting events.

[1] D.O. Morgan, "Cyclin-dependent Kinases: Engines, Clocks, and Microprocessors," *Annual Review of Cell and Developmental Biology* 13:261-291, 1997.

[2] Chemotaxis is the propensity of certain bacteria, such as *E. coli*, to swim toward higher concentrations of nutrients.

[3] H.C. Berg, "A Physicist Looks at Bacterial Chemotaxis," *Cold Spring Harbor Symposium on Quantitative Biology* 53(1):1-9, 1988.

SOURCE: Items 1-4 adapted from L. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Murray, "From Molecular to Modular Cell Biology," *Nature* 402(Suppl.):C47-C52, 1999.

the building blocks remain more or less stable, while the connectivity among them determines the character of the system.

If biological modules really do exist, one might expect to find them reused in different cellular contexts, performing the same function but to different ends. Understanding the function and behavior of a cellular pathway would entail the discovery and characterization of such modular building blocks, tasks that should be simpler than trying to understand biological networks of different organisms as an irreducible whole.

Several independent pieces of evidence have emerged supporting the modularity hypothesis. For example, evidence is accruing that certain regions of DNA are "conserved" from one species to another. These regions may be associated with genes coding for proteins or with regulatory and structural functionality. Caenepeel et al. found that the human and mouse kinomes (i.e., the collection of protein kinases in an organism) are 99 percent identical, although the percentage of identity between orthologues (i.e., genes or proteins from different organisms that have the same function) ranges from 70 percent to 99 percent (with single nucleotide insertions or deletions in many cases).[23] Dermitzakis et al. found that perhaps a third of the highly conserved DNA regions between mouse and human code for proteins, while much of the rest probably codes for regulatory and structural functionality.[24]

Genetic expression networks may also display regular patterns of interconnections (motifs) recurring in many different parts of a network at frequencies much higher than those found in randomized networks.[25] Such motifs might be regarded as building blocks that can be used to assemble entities of more complex functionality.[26] For example, Shen-Orr et al. discovered a series of simple, recurring network motifs in the gene interaction map of the bacterium *E. coli*.[27] Shortly afterwards, Richard Young and colleagues found the same motifs to recur at statistically surprising frequencies in yeast.[28] Milo et al. found that these motifs were also overrepresented in a neuronal connectivity network of *Caenorhabditis elegans* as well as the connectivity networks in the ISCAS89 benchmark set of sequential logic electronic circuits, but not in ecosystem food webs.[29] Milo et al. speculate that these motifs reflect the underlying processes that generated each type of network, in this case one set of motifs for those that process information (the genetic regulation, neuronal connectivity, and electronic logic networks) and another set of motifs for those that process and carry energy.

Finally, a collaborative project led by Eric Davidson and his group at the California Institute of Technology, and involving Bolouri and Hood at the Institute for Systems Biology, also suggests simple design principles and building blocks in genetic networks. Figure 6.1 is a map of the interactions among

[23] S. Caenepeel, G. Charydezak, S. Sudarsanam, T. Hunter, and G. Manning, "The Mouse Kinome: Discovery and Comparative Genomics of All Mouse Protein Kinases," *Proceedings of the National Academy of Sciences* 101(32):11707-11712, 2004.

[24] E.T. Dermitzakis, A. Reymond, R. Lyle, N. Scamuffa, C. Ucla, S. Deutsch, B.J. Stevenson, et al., "Numerous Potentially Functional But Non-genic Conserved Sequences on Human Chromosome 21," *Nature* 420(6915):578-582, 2002.

[25] R. Milo, S. Shen-Or, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon, "Network Motifs: Simple Building Blocks of Complex Networks," *Science* 298(5594):824-827, 2002.

[26] Alon refines the notion of module as building block to suggest that modules and motifs are related but separate concepts. In Alon's view, a module in a network is a set of nodes that have strong interactions and a common function. Some nodes are internal and do not interact significantly with nodes outside the module. Other nodes accept inputs and produce outputs that control the module's interactions with the rest of the network. Alon argues that one reason modules evolve in biology is that new devices or entities can be constructed out of existing, well-tested modules; thus, adaptation to new conditions (and new forces of natural selection) is more easily accomplished. If modules are to be swapped in and out, they must possess the property that their input-output response is approximately independent of what is connected to them—that is, that the module is functionally encapsulated. By contrast, a motif is an overrepresented patterns of interconnections in a network that is likely to perform some useful behavior. However, it may not be functionally encapsulated, in which case it is not a module. For more discussion, see U. Alon, "Biological Networks: The Tinkerer as an Engineer," *Science* 301(5641):1866-1867, 2003.

[27] S.S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network Motifs in the Transcriptional Regulation Network of *Escherichia coli*," *Nature Genetics* 31(1):64-68, 2002.

[28] T.I. Lee, H.J. Yang, S.Y. Lin, M.T. Lee, H.D. Lin, L.E. Braverman, and K.T. Tang, "Transcriptional Regulatory Networks in *Saccharomyces cerevisiae*," *Science* 298(5594):799-804, 2002.

[29] R. Milo et al., "Network Motifs," 2002.

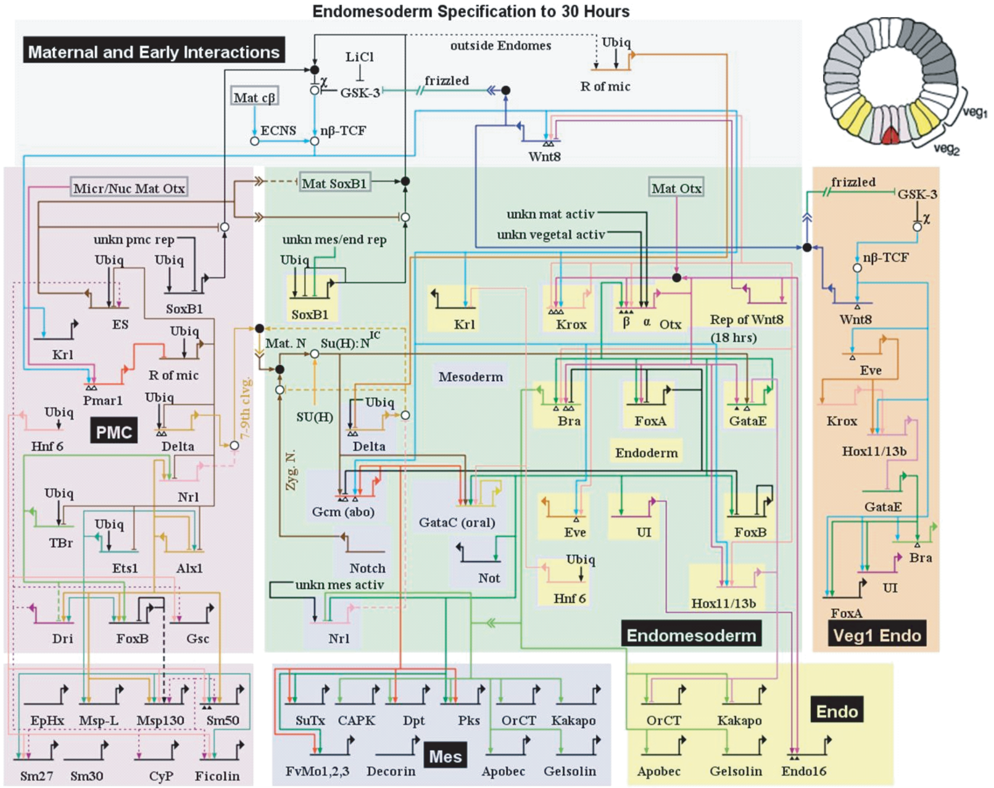

FIGURE 6.1 The endomesoderm specification network in the sea urchin species *Strongylocentrotus purpuratus*.

The period of activity represented spans embryonic growth from single cell to gastrulation (approximately 600 cells). The different background colors denote different cell types, as indicated on the cartoon of an early blastula-stage embryo on the top right. The short, thick horizontal lines represent regulatory DNA of a particular gene in the network, to which transcription factors bind to activate or repress transcription. The bent arrow emanating from each regulatory domain represents the basal transcription apparatus of the gene, and the line(s) emerging from it represent the interactions of the product of the gene with other proteins (via the white and black interaction boxes) or *cis*-regulatory DNA.

The architecture of the network is based on perturbation and expression data, on data from *cis*-regulatory analyses for several genes, and on other experiments discussed in the references below. For quantitative results of perturbation experiments and temporal details and the latest view of the network, see http://sugp.caltech.edu/endomes/.

The repression cascade motif referred to in the text is indicated by the thick black (upstream gene) and gray (downstream genes) arrows. This work is described in the following:

1. E.H. Davidson, J.P. Rast, P. Oliveri, A. Ransick, C. Calestani, C.H. Yuh, T. Minokawa, et al., "A Genomic Regulatory Network for Development," *Science* 295(5560):1669-1678, 2002.

2. H. Bolouri and E.H. Davidson, "Modeling DNA Sequence-based *cis*-Regulatory Gene Networks," *Developmental Biology* 246(1):2-13, 2002.

3. C.T. Brown, A.G. Rust, P.J.C. Clarke, Z. Pan, M.J. Schilstra, T. De Buysscher, G. Griffin, et al., "New Computational Approaches for Analysis of *cis*-Regulatory Networks," *Developmental Biology* 246(1):86-102, 2002.

4. A. Ransick, J.P. Rast, T. Minokawa, C. Calestani, and E.H. Davidson, "New Early Zygotic Regulators of Endomesoderm Specification in Sea Urchin Embryos Discovered by Differential Array Hybridization," *Developmental Biology* 246(1):132-147, 2002.

5. C.H. Yuh, C.T. Brown, C.B. Livi, L. Rowen, P.J.C. Clarke, and E.H. Davidson, "Patchy Interspecific Sequence Similarities Efficiently Identify Positive *cis*-Regulatory Elements in the Sea Urchin," *Developmental Biology* 246(1):148-161, 2002.

6. E.H. Davidson, J.P. Rast, P. Oliveri, A. Ransick, C. Calestani, C.H. Yuh, T. Minokawa, et al., "A Provisional Regulatory Gene Network for Specification of Endomesoderm in the Sea Urchin Embryo," *Developmental Biology* 246(1):162-190, 2002.
7. J.P. Rast, R.A. Cameron, A.J. Poustka, and E.H. Davidson, "Brachyury Target Genes in the Early Sea Urchin Embryo Isolated by Differential Macroarray Screening," *Developmental Biology* 246(1):191-208, 2002.
8. P. Oliveri, D.M. Carrick, and E.H. Davidson, "A Regulatory Gene Network That Directs Micromere Specification in the Sea Urchin Embryo," *Developmental Biology* 246(1):209-228, 2002.

SOURCE: Figure from M. Levine and E.H. Davidson, "Gene Regulatory Networks for Development," *Proceedings of the National Academy of Sciences* 102(14):4936-4942, 2005, available at http://www.pnas.org/cgi/content/full/102/14/4936. Copyright 2005 National Academy of Sciences.

approximately 50 genes underlying an early cell-type specification event in sea urchin embryos that includes several recurring interaction motifs. For example, there are several cases in which a gene (thick black arrow), instead of activating another gene directly, represses a repressor of the target gene (thick gray arrows). Such an arrangement can provide a number of possible advantages, including a sharper activation profile for the target gene, important in defining spatial boundaries between cell types.

Modularity and conservation suggest a potential for comparative studies across species (e.g., pufferfish, mice, humans) to contribute to an understanding of biological function. That is, understanding the role of a certain protein in mice, for example, may suggest a similar role for that same protein if it is found in humans.

These comments should not be taken to mean that functional modules in biological entities are necessarily simple or static. Biological systems are often made up of elements with multiple functions interacting in ways that are complex and difficult to separate, and nature exploits multiple linkages that a human engineer would not tolerate in the design of an artifact.[30] For example, a component of one module may (or may not) play a role in a different module at a different time. A module's functional behavior may be quantitatively regulated or switched between qualitatively different functions by chemical signals from other modules. Despite these important differences between biological modules and the modules that constitute humanly engineered artifacts, the notion of a collection of parts that can be counted on to perform a given function—that is, a module—is meaningful from an analytical perspective and our understanding of that function.

6.2.4 Robustness in Biological Entities

Robustness is one of the characteristics of biological systems that is most admired and most desired for engineered systems. Especially as compared to software and information systems, which are notoriously brittle, biological systems maintain functionality in the face of a range of perturbations. More traditional hardware engineering, however, has studied the questions of robustness (under various names including fault-tolerance and control systems). Applying the analytical techniques developed in engineering to studying the mechanics of robustness in biology, the logic goes, might reveal new insights not only about biology, but about robust system design.

In biology, the term robustness is used in many different ways in different subfields, including the preservation of species diversity, a measure of healing, comprehensibility in the face of incomplete information, continuity of evolutionary lineages, phenotypic stability in development, cell metabolic stability in the face of stochastic events, or resistance to point mutations.[31] Its most general usage,

[30] This is not to say that human-engineered artifacts are not affected by their origins. "Capture by history" characterizes many human artifacts as well, but likely not as strongly. For more discussion of these points, see D. Norman, 1998, cited in Footnote 16.

[31] D.C. Krakauer, "Robustness in Biological Systems—A Provisional Taxonomy," *Complex Systems Science in Biomedicine*, T.S. Deisboeck, J.Y. Kresh, and T.B. Kepler, eds., Kluwer, New York, 2003.

however, refers to the ability of a structure or process to persist in the face of perturbations of internal components or the environment. Those perturbations might include outright component failure, unexpected behavior from components or other cooperating systems, stochastic changes in chemical concentrations or reaction rates, mutations, or the motion of external biochemical parameters. These sorts of perturbations, such as stochastic changes of molecular concentrations, are intrinsic to the nature of biology, from the molecular scale to the ecological.

A robust response to these perturbations generally consists of one of three types: (1) parameter insensitivity, meaning that a robust process does not depend on a single ideal value of an input; (2) graceful degradation, in which the level of functionality of the system is indeed lessened by component failures, but it continues to function; and (3) adaptation, in which internal components reconfigure to react to a change to maintain the same level of functionality.[32]

Kitano notes that robustness is attained in biological systems by using mechanisms well known to human engineers. He describes four mechanisms or approaches to biological robustness:[33]

1. System control mechanisms such as negative-feedback and feed-forward control;
2. Redundancy, whereby multiple components with equivalent functions are introduced for backup;
3. Structural stability, where intrinsic mechanisms are built to promote stability; and
4. Modularity, where subsystems are physically or functionally insulated so that failure in one module does not spread to other parts and lead to system-wide catastrophe.

Kitano then notes that these approaches used in engineering systems are also found in biological systems, pointing out that "redundancy is seen at the gene level, where it functions in control of the cell cycle and circadian rhythms, and at the circuit level, where it operates in alternative metabolic pathways in *E. coli*." Furthermore, engineering approaches have proven to be a useful lens when investigating biological robustness.

For example, Barkai and Leibler[34] established a model (later confirmed experimentally) to explain perfect robust adaptation in bacterial chemotaxis, or the ability of bacteria to move toward increased concentrations of certain ligands. It had long been known that the mechanism responsible for this ability had several key attributes, among them a high sensitivity to changes in chemical concentration, together with an ability to adapt to the absolute level of that concentration. Working with the known molecular makeup of these cells (e.g., the receptors, kinases, and diffusible messenger proteins), Barkai and Leibler showed that when varied separately, many of the rate constants (such as molecular concentrations of elements of the signaling network or reaction rates) could be varied by orders of magnitude without affecting the magnitude of the response.[35]

Later work by Yi et al. used the mathematics of control systems to show how the Barkai-Leibler model was a special case of integral feedback control, a well-studied approach of control theory.[36] In addition to control theory (including feedback and feed-forward control), many other engineering approaches are found in biological systems, including redundancy, modularity, purging (quickly eliminating failing components), and spatial compartmentalization.[37]

[32] H. Kitano, "Systems Biology: A Brief Overview," *Science* 295(5560):1662-1664, 2002. Available at http://www.sciencemag.org/cgi/content/abstract/295/5560/1662.

[33] H. Kitano, "Systems Biology," 2002.

[34] N. Barkai and S. Leibler, "Robustness in Simple Biochemical Networks," *Nature* 387(6636):913-917, 1997.

[35] However, the mechanism does not account for the full dynamic range of the sensor patches at a molecular level. (It may be that some sort of emergent property of the sensor patch as a whole, as opposed to some property of the individual sensor complexes, is necessary to obtain the full dynamic range. See, for example, T.S. Shimizu, S.V. Aksenov, and D. Bray, "A Spatially Extended Stochastic Model of the Bacterial Chemotaxis Signaling Pathway," *Journal of Molecular Biology* 329(2):291-309, 2003.)

[36] T.M. Yi, Y. Huang, M.I. Simon, and J. Doyle, "Robust Perfect Adaptation in Bacterial Chemotaxis Through Integral Feedback Control," *Proceedings of the National Academy of Sciences* 97(9):4649-4653, 2000.

[37] D.C. Krakauer, "Robustness in Biological Systems," 2003.

Kitano makes the point that robustness is a property of an entire system;[38] it may be that no individual component or process within a system would be robust, but the system-wide architecture still provides robust behavior. This presents a challenge for analysis, since elucidating such behaviors can be counterintuitive and computationally demanding.[39] In one such example, von Dassow and colleagues investigated the development of striped patterns in *Drosophila*.[40] They computationally modeled a network of interactions between genes and regulatory proteins active during embryogenesis and explored the parameter space to see which sets of parameters produced stable striping. In their first attempt, they were unable to reproduce such behavior computationally. However, once they added two more molecular events and their interactions to the network, a surprisingly high proportion of the randomly chosen parameters produced the desired results. This strongly implies that such a network, taken as a whole, is a robust developmental module, able to produce a particular effect despite wide variation in reaction parameters.

In a refinement to that work, Ingolia investigated the architecture of that network to attempt to determine the structural sources of such robust behavior.[41] He determined that the source of the robustness at the network level was a pair of positive feedback loops of gene expression, which led to cells being forced to one of two stable states (bistability). That is, small perturbations or changes in certain parameters would necessarily result in individual cells reaching one of two states. Ingolia showed that such bistability, at both an individual cell level and a network level, is an important architectural property leading to robust behavior and that the latter is in fact a consequence of the former. Moreover, it is this bistability that is responsible for the ability of the network to maintain a fixed pattern of gene expression even in the face of cell division and growth.[42]

Robustness comes at a cost of increased complexity. The simplest bacteria can survive only within narrow ranges of environmental parameters, while more complex bacteria, such as *E. coli* (with a genome an order of magnitude larger than mycoplasma), can withstand more severe environmental fluctuations.[43] This increased complexity can in turn be the root of cascading failures, if the elements of the network responsible for the adaptive response fail. This implies that increased robustness of a certain aspect or element of a system with respect to a certain perturbation may come at the cost of increased vulnerability in a different aspect or element or to a different attack.

Robustness can also serve as a signpost for discovering the details of biological function. Although there may be a prohibitively large number of ways that a genetic network could produce a given result, for example, only a few of those ways are likely to do so robustly. Knowledge of the robust qualities of a biological system, coupled with theoretical or simulated analysis of networks, could aid in reverse engineering the system to determine its actual configuration.[44]

An open and intriguing question is the relationship between robustness and evolution. Because robustness is the quality of maintaining stability, in some sense it stands as a potential inhibitor to evolution, for example, by masking the effects of point mutations. And yet robust modules or organisms are more likely to survive, and thus pass on into succeeding generations. How does robustness evolve? How do robust systems evolve? One engineering approach to this problem is to consider biological systems as sets of components interacting through protocols,[45] with one critical measure of a

[38] H. Kitano, "Systems Biology," 2002. Available at http://www.sciencemag.org/cgi/content/abstract/295/5560/1662.

[39] A.D. Lander, "A Calculus of Purpose," *PLoS Biology* 2(6):e164, 2004.

[40] G. von Dassow, E. Meir, E.M. Munro, and G.M. Odell, "The Segment Polarity Network Is a Robust Developmental Module," *Nature* 406(6792):188-192, 2000.

[41] N.T. Ingolia, "Topology and Robustness in the *Drosophila* Segment Polarity Network," PLoS Biology 2(6):e123, 2004.

[42] A.D. Lander, "A Calculus of Purpose," 2004.

[43] J.M. Carlson and J. Doyle, "Complexity and Robustness," *Proceedings of the National Academy of Sciences* 99(Suppl. 1):2538-2545, 2002.

[44] U. Alon, "Biological Networks: The Tinkerer as an Engineer," *Science* 301:1866-1867, 2003.

[45] M.E. Csete and J.C. Doyle, "Reverse Engineering of Biological Complexity," *Science* 295:1664-1669, 2002.

good protocol being its ability to support both robustness and evolvability, a key consideration in technical protocols of human engineering such as TCP/IP.

6.2.5 Noise in Biological Phenomena[46]

As one illustration of how engineering disciplines might shed light on biological mechanism, consider the opposition of robustness and noise in biological phenomena. Biological organisms exhibit high degrees of robustness in the face of changing environments. Engineered artifacts designed by human beings have used mechanisms such as negative feedback to provide stability, redundancy to provide backup, and modularity for the isolation of failures to enhance robustness. As the discussion below indicates, these mechanisms are used for these purposes in biological organisms, as well.[47]

In a biological context, noise can take the form of fluctuations in quantities such as reaction rates, concentrations, spatial distributions, and fluxes. In addition, fluctuations may also occur at the molecular level. However, despite the noise inherent in the internal environment of a cell, cells operate—often robustly and quite stably—within strict parameters, and robustness has been hypothesized as an intrinsic property of intracellular networks. (For instance, the chemotaxis pathway in *E. coli* functions over a wide range of enzymatic activities and protein concentrations.[48] Robustness is also illustrated in some developmental processes[49] and phage lambda regulation.[50]) This robustness suggests that cells use and reject noise in a systematic manner.

For the analysis of biological noise, much of the analysis originally derived from signal processing and control theory is applicable.[51] Indeed, pathways can be regarded as analog filters and classified in terms of frequency response, where the differences between filtering electronic noise and filtering biological noise are reflected only in the details of the underlying mechanisms rather than in high-level abstractions of filtering theory.

Cascades and relays such as two-component systems and the mitogen-activated protein kinase pathway function as low-pass filters (i.e., filters that attenuate high-frequency noise).[52] As a general rule, longer cascades are more effective at reducing noise. However, because noise arises in the pathway itself, the amount of internally generated noise increases with cascade length—suggesting that there is an optimal cascade length for attenuating noise.[53]

It is not surprising that low-pass filters are components of biological systems. As noted above, biological systems operate homeostatically,[54] and the essential principle underlying homeostasis is that of negative feedback. From the standpoint of signal processing, a negative feedback loop functions as a low-pass filter.

[46] Section 6.2.5 is based on and incorporates several excerpts from C.V. Rao, D.M. Wolf, and A.P. Arkin, "Control, Exploitation and Tolerance of Intracellular Noise," *Nature* 420(6912):231-237, 2002.

[47] H. Kitano, "Systems Biology: A Brief Overview," Science 295(5560):1662-1664, 2002. Available at http://www.sciencemag.org/cgi/content/abstract/295/5560/1662.

[48] N. Barkai and S. Leibler, "Robustness in Simple Biochemical Networks," *Nature* 387:913-917, 1997; U. Alon, M.G. Surette, N. Barkai and S. Leibler, "Robustness in Bacterial Chemotaxis," *Nature* 397:168-171, 1999. (Cited in Rao et al., 2002.)

[49] G. von Dassow, E. Meir, E.M. Munro, and G.M. Odell, "The Segment Polarity Network Is a Robust Developmental Module," *Nature* 406:188-192, 2000; E. Meir, G. von Dassow, E. Munro, and G.M. Odell, "Robustness, Flexibility, and the Role of Lateral Inhibition in the Neurogenic Network," *Current Biology* 12:778-786, 2002. (Cited in Rao et al., 2002.)

[50] J.W. Little, D.P. Shepley, and D.W. Wert, "Robustness of a Gene Regulatory Circuit," *EMBO Journal* 18:4299-4307, 1999.

[51] A.P. Arkin, "Signal Processing by Biochemical Reaction Networks," pp. 112-144, *Self-organized Biological Dynamics and Nonlinear Control*, J. Walleczek, ed., Cambridge University Press, London, 2000; M. Samoilov, A. Arkin, and J. Ross, "Signal Processing by Simple Chemical Systems," *Journal of Physical Chemistry* 106:10205-10221, 2002. (Cited in Rao et al., 2002.)

[52] P.B. Detwiler, S.A. Ramanathan, A. Sengupta, and B.I. Shraiman, "Engineering Aspects of Enzymatic Signal Transduction: Photoreceptors in the Retina," *Biophysical Journal* 79(6):2801-2817, 2000. (Cited in Rao et al., 2002.)

[53] M. Thattai and A.Van Oudenaarden, "Attenuation of Noise in Ultrasensitive Signaling Cascades," *Biophysical Journal* 82(6):2943-2950, 2002. (Cited in Rao et al., 2002.)

[54] Homeostasis is the property of a system that enables it to respond to changes in its environment in such a way that it tends to maintain its original state.

A second useful construct from signal processing is the bandpass filter, which is based on the control theory notion of integral feedback. Integral feedback is a kind of negative feedback that amplifies intermediate frequencies and attenuates low and high frequencies. A biological instantiation of integral feedback is contained in bacterial chemotaxis.[55]

In addition to the filters described above, other mechanisms attenuate noise in systems. These include the following:

- *Redundancy*. Noise in a single channel might be misinterpreted as a genuine signal. However, redundancy—in the form of multiple channels serving the same function—can help to minimize the likelihood of such an occurrence. In a biological context, redundancy has been demonstrated in mechanisms such as gene dosage and parallel cascades,[56] which attenuate the effects of noise by increasing the likelihood of gene expression or establishing a consensus from multiple signals.
- *Checkpointing*. Noise can interfere with the successful completion of various biological operations that are essential in a pathway. However, a checkpoint can ensure that each step in a pathway is completed successfully before proceeding with the next step. Such checkpoints have been characterized in the cell cycle and flagellar biosynthesis.[57]
- *Proofreading*. Noise can introduce errors into a process. But error-correcting mechanisms can reduce this effect of noise, as is the case of kinetic proofreading in protein translation.[58]

A final, and surprising, mechanism is that complexity itself in some cases can be implicated in the robustness of an organism against noise. In 1942, Waddington noted the stability of phenotypes (from the same species) against a backdrop of considerable genetic variation, a phenomenon known as canalization.[59] In principle, such stability could result from explicit genetic control of phenotype features, such as the number of fingers on a hand or the placement of wings on an insect's body. However, Siegal and Bergman modeled the developmental process responsible for the emergence of such features as a network of interacting transcriptional regulators and found that the network constrains the genetic system to produce canalization.[60] Furthermore, the extent of canalization, measured as the insensitivity of a phenotype to changes in the genotype (i.e., to mutations), depends on the complexity of the network, such that more highly connected (i.e., more complex) networks evolve to be more canalized. (Box 6.5 provides more details.)

Consider that noise can also make positive contributions to biological systems. For example, it is well known from the agricultural context that monocultures are less robust than ecosystems that involve multiple species—the first can be wiped out by a disease that targets the specific crop in question, whereas the second cannot. Thus, some degree of variation in a populating species is desirable, and noise is one mechanism for introducing variation that results in population heteroge-

[55] The size of a single bacterium is so small that the bacterium is unable to sense a spatial gradient across the length of its body. Thus, to sense a spatial gradient, the bacterium moves around and senses chemical concentrations in different locations at different times; the result is a motion bias toward attractants. See T.M. Yi, Y. Huang, M.I. Simon, and J. Doyle, "Robust Perfect Adaptation in Bacterial Chemotaxis Through Integral Feedback Control," *Proceedings of the National Academy of Sciences* 97(9):4649-4653, 2000. (Cited in Rao et al., 2002.)

[56] H.H. McAdams and A. Arkin, "It's a Noisy Business! Genetic Regulation at the Nanomolar Scale," *Trends in Genetics* 15(2):65-69, 1999; D.L. Cook, A.N. Gerber, and S.J. Tapscott, "Modeling Stochastic Gene Expression: Implications for Haploinsufficiency," *Proceedings of the National Academy of Sciences* 95(26):15641-15646, 1998. (Cited in Rao et al., 2002.)

[57] L.H. Hartwell and T.A. Weinert, "Checkpoints: Controls That Ensure the Order of Cell Cycle Events," *Science* 246(4930):629-634, 1989. (Cited in Rao et al., 2002.)

[58] M.V. Rodnina and W. Wintermeyer, "Ribosome Fidelity: tRNA Discrimination, Proofreading and Enduced Fit," *Trends in Biochemical Science* 26(2):124-130, 2001. (Cited in Rao et al., 2002.)

[59] C.H. Waddington, "Canalization of Development and the Inheritance of Acquired Characters," *Nature* 150:563-565, 1942.

[60] M.L. Siegal and A. Bergman, "Waddington's Canalization Revisited: Developmental Stability and Evolution," *Proceedings of the National Academy of Sciences* 99(16):10528-10532, 2002.

Box 6.5
Canalization and the Connectivity of Transcriptional Regulatory Networks

To explore the possibility that genetic canalization may be a by-product of other selective forces, . . . [we start with] the model of A. Wagner, who treats development as the interaction of a network of transcriptional regulatory genes, phenotype as the equilibrium state of this network, and fitness as a function of the distance between an individual's equilibrium state and the optimum state. . . . Evolution in the model [a generalized version of Wagner's] consists of three phases: mating, development, and selection. Mating and selection are modeled in accord with traditional population-genetic approaches. . . . [To handle development] one can represent a network of transcriptional regulators by a state vector containing the concentration of each gene product and a matrix, the entries of which represent the effects of each gene product on the expression of each gene. Entries may be either positive (activating) or negative (repressing) and may differ in magnitude. Zero elements in the matrix represent the absence of interaction between the given gene product and gene. The developmental process is then fully described by a set of nonlinear coupled difference equations. . . . Wagner draws an analogy between the rows of the interaction matrix and the enhancer regions of the genes in the network and further justifies the biological realism of this type of model by reference to data from actual genetic networks. An important assumption in the model, also justified by A. Wagner, is that functional genetic networks will reach a stable equilibrium gene-expression state, and that unstable networks reflect, in a sense, the failure of development. Thus, in his model and ours, development itself enforces a kind of selection, because we require that the network of regulatory interactions produce a stable equilibrium gene-expression state (its "phenotype"), whose distance to an optimum state can then be measured during the selection phase.

. . . We report here the results of numerical simulations of our model of an evolving developmental-genetic system. We demonstrate an important, perhaps primary, role for the developmental process itself in creating canalization, in that insensitivity to mutation evolves even when stabilizing selection is absent. We go on to demonstrate that the complexity of the network is a key factor in this evolutionary process, in that networks with a greater proportion of connections evolve greater insensitivity to mutation.

. . . One is led to wonder whether the evolution of canalization under no stabilizing selection on the gene-expression pattern is an artifact of the modeling framework or whether it represents a finding of real biological significance. We argue that the latter is true on a number of counts. To begin, we acknowledge that it is difficult to envision a scenario in nature in which the stability of a developmental module is required, but the phenotype produced by that module is not subject to selection. One situation in which this condition may hold is when a species colonizes a new territory with virtually unlimited resources, so selection is only for those that develop to reproduce. Furthermore, even if such a scenario does not pertain, the conceptual decomposition of stabilizing selection into selection for an optimum and selection for developmental stability is important. Thus, even in scenarios in which members of a population are subject to selection for an optimum, the evolution of canalization may proceed because of the underlying selection for stability of the developmental outcome. Our results suggest that this underlying selection can occur very fast. Because others have argued that the evolution of canalization under stabilizing selection may be slow, developmental stability may therefore be the dominant force in the evolution of canalization.

SOURCE: Reprinted by permission from M.L. Siegal and A. Bergman, "Waddington's Canalization Revisited: Developmental Stability and Evolution," *Proceedings of the National Academy of Sciences* 99(16):10528-10532, 2002. Copyright 2002 National Academy of Sciences. (References and figures are omitted above and can be found in the original article.)

neity and diversity. For example, noise (in the form of molecular fluctuations) introduced into the genetic circuit governing development in phage lambda can cause an initially homogeneous population to separate into lytic and lysogenic populations.[61] (In this case, the basic mechanism involves

[61] A. Arkin, J. Ross, and H.H. McAdams, "Stochastic Kinetic Analysis of Developmental Pathway Bifurcation in Phage Lambda-infected *Escherichia coli* Cells," *Genetics* 149(4):1633-1648, 1998. (Cited in Rao et al., 2002.)

two antagonistic feedback loops that create a switch and molecular fluctuations that partition the initial population stochastically.)

Noise can be used to enhance a signal when certain nonlinear effects are present, as demonstrated by the phenomenon of stochastic resonance.[62] Stochastic resonance is found in many biological systems, including the electroreceptors of paddlefish,[63] mechanoreceptors in the tail fins of crayfish,[64] and hair cells in crickets.[65] A similar phenomenon can potentially increase sensitivity in certain signaling cascades.[66]

Finally, noise can be useful for introducing stability. The network that controls circadian rhythms consists of multiple, complex, interlocking feedback loops. Both deterministic and stochastic mechanisms for noise resistance in circadian rhythms have been explored,[67] and it turns out that stochastic models are able to produce regular oscillations when the deterministic models do not,[68] suggesting that the regulatory networks may utilize molecular fluctuations to their advantage.

The discussion above suggests that biological robustness is in some ways a problem of controlling the effects of noise and in other ways one of exploiting those effects. Considerations of noise and robustness thus offer insight into the design and function of intracellular networks.[69] That is, the function of an intracellular network may require specific regulatory and information structures, and certain design features are necessary for a stable network phenotype.

Finally, note that mechanisms of the sorts described above do not generally function in isolation, but rather interact in complex networks involving multiple feedback loops, and the resulting networks can produce diverse phenomena, including switches, memory, and oscillators.[70] Such coupling also has an important analytical consequence—namely, that the composite behavior of multiple coupled mechanisms is much more difficult to predict than the behavior of individual components. To analyze multiple coupled systems, computational models are highly useful.

6.3 A COMPUTATIONAL METAPHOR FOR BIOLOGY

In addition to the abstractions described above, computing and computer science can also provide life scientists with a rich source of language, metaphors, and analogies with which to describe biological phenomena and insights from a computational perspective. These linguistic and cognitive aspects may well make it easier for insights originating in computing to be made relevant to biology, and thus

[62] L. Gammaitoni, P. Hanggi, P. Jung, and F. Marchesoni, "Stochastic Resonance," *Reviews of Modern Physics* 70:223-287, 1998. (Cited in Rao et al., 2002.)

[63] D.F. Russell, L.A. Wilkens, and F. Moss, "Use of Behavioural Stochastic Resonance by Paddle Fish for Feeding," *Nature* 402(6759):291-294, 1999. (Cited in Rao et al., 2002.)

[64] J.K. Douglass, L. Wilkens, E. Pantazelou, and F. Moss, "Noise Enhancement of Information Transfer in Crayfish Mechanoreceptors by Stochastic Resonance," *Nature* 365(6444):337-340, 1993. (Cited in Rao et al., 2002.)

[65] J.E. Levin and J.P. Miller, "Broadband Neural Encoding in the Cricket Cercal Sensory System Enhanced by Stochastic Rresonance," *Nature* 380(6570):165-168, 1996. (Cited in Rao et al., 2002.)

[66] J. Paulsson, O.G. Berg, and M. Ehrenberg, "Stochastic Focusing: Fluctuation-enhanced Sensitivity of Intracellular Regulation," *Proceedings of the National Academy of Sciences* 97(13):7148-7153, 2000. (Cited in Rao et al., 2002.)

[67] N. Barkai and S. Leibler, "Circadian Clocks Limited by Noise," *Nature* 403(6767):267-268, 2000; D. Gonze, J. Halloy, and A. Goldbeter, "Robustness of Circadian Rhythms with Respect to Molecular Noise," *Proceedings of the National Academy of Sciences* 99(2):673-678, 2002; P. Smolen, D.A. Baxter, and J.H. Byrne, "Modeling Circadian Oscillations with Interlocking Positive and Negative Feedback Loops," *Journal of Neuroscience* 21(17):6644-6656, 2001. (Cited in Rao et al., 2002.)

[68] J.M. Vilar, H.Y. Kueh, N. Barkai, and S. Leibler, "Mechanisms of Noise Resistance in Genetic Oscillators," *Proceedings of the National Academy of Sciences* 99(9):5988-5992, 2002. (Cited in Rao et al., 2002.)

[69] M.E. Csete and J.C. Doyle, "Reverse Engineering of Biological Complexity," *Science* 295(5560):1664-1669, 2002; M. Morohashi, et al., "Robustness as a Measure of Plausibility in Models of Biochemical Networks," *Journal of Theoretical Biology* 216(1):19-30, 2002; L.H. Hartwell, J.J. Hopfield, S. Leibler, and A.W. Murray, "From Molecular to Modular Cell Biology," *Nature* 402(6761 Suppl):C47-C52, 1999. (Cited in Rao et al., 2002.)

[70] M.B. Elowitz and S. Leibler, "A Synthetic Oscillatory Network of Transcriptional Regulators," *Nature* 403(6767):335-338, 2000; T.S. Gardner, C.R. Cantor, and J.J. Collins, "Construction of a Genetic Toggle Switch in *Escherichia coli*," *Nature* 403(6767):339-342, 2000. (Cited in Rao et al., 2002.)

information abstractions can be used to communicate about or to explain biological processes and concepts. Consider, for example, the Jacob and Monod description of the genome as a "genetic program," capable of controlling its own execution.[71] (Conversely, biological metaphors and language might offer analogous benefits to computing, which is the subject of Chapter 8.) At the same time, poorly chosen metaphors can limit understanding by carrying over misleading or irrelevant details. For example, the "genetic program" metaphor described above might lead one to think of protein synthesis as being executed one instruction at a time (as most computer programs would be), obscuring the parallel and interconnected nature of the genetic protein synthesis network.[72]

The use of a metaphor (to look at a problem in field A through the lens of field B) invites one to apply insights from field B to the problem in field A. Metaphors are often (indeed, almost always) imprecise and somewhat vague, because they are not specific about which insights from field B are relevant to field A. They can nevertheless be useful, because they constitute an additional source of insight and new ways of thinking to be brought to bear on field A that might not otherwise be available in the absence of those metaphors. Moreover, field B—as a discipline—constitutes an existence proof that the insights in question can in fact be part of an intellectually coherent whole.

Consider, for example, extending the notion of the "genetic program." In some sense, the DNA sequence can be analogized to the binary code of a program. However, in many real computer programs, a program structure or architecture or individual components may be apparent from representing the program in its source code form, where things such as variable declarations and subroutines make manifestly obvious what is obscured in the binary representation. Calling sequences between program and subprogram define program interfaces and protocols for how different components of a program may communicate—data definitions, formats, and semantics, for instance. Thus, it may be meaningful to inquire about the analogous things in biology, and indeed, a gene contained in DNA might well be one analogue of a subprogram or the action potential in neuroscience one analogue of a communications protocol.

Another analogy can be drawn between the evolution of computing and the biological transition from single-cell organisms to multicell organisms. Multicellular life exploits four broad strategies: collaboration between highly specialized cells; communication by polymorphic messages; self, defined by a stigmergic structure; and self, protected by programmed cell death. These strategies are rare in single-cell organisms but nearly universal in multicellular organisms, and evolved before or coincident with the emergence of multicellular life. As described in Table 6.1, each of these strategies may be analogous to trends seen in computing today.

To illustrate how the use of a computational metaphor can provide insight and lead to deeper exploration, note that cellular processes are concurrent (i.e., changes in the surrounding environment can trigger the execution of many parallel processes); operate at many levels including the submolecular, molecular, subcellular, and cellular; and involve relationships among many subcellular and molecular objects. Computer scientists have devised a number of formalisms that are capable of representing such processes, and Kam et al.[73] modeled aspects of T-cell activation using the formalism of Statecharts,[74] as they have been adapted to the framework of object-oriented modeling.[75] Because the object-oriented Statechart approach supports

[71] F. Jacob and J. Monod, "Genetic Regulatory Mechanisms in the Synthesis of Proteins," *Journal of Molecular Biology* 3:318-356, 1961.

[72] E.F. Keller, *Making Sense of Life—Explaining Biological Developments with Models, Metaphors, and Machines*, Harvard University Press, Cambridge, MA, 2003.

[73] N. Kam, I.R. Cohen, and D. Harel, "The Immune System as a Reactive System: Modeling T Cell Activation with Statecharts," *Proceedings of a Symposium on Visual Languages and Formal Methods* (VLFM'01), part of IEEE Symposium on Human-centric Computing (HCC'01), 2001, pp. 15-22.

[74] D. Harel, "Statecharts: A Visual Formalism for Complex Systems," *Science of Computer Programming* 8:231-274, 1987. (Cited in Kam et al., " The Immune System as a Reactive System," 2001.)

[75] G. Booch, *Object-Oriented Analysis and Design, with Applications*, Addison-Wesley, Menlo Park, CA, 1994; D. Harel and E. Gery, "Executable Object Modeling with Statecharts," *Computer*, 31-42, 1997; J. Rumbaugh, M. Blaha, W. Premerlani, F. Eddy, and W. Lorensen, *Object-Oriented Modeling and Design*, Prentice Hall, Englewood Cliffs, NJ, 1991. (Cited in Kam et al., 2001.)

TABLE 6.1 Principles of Operation for Multicellular Organisms and Networked Computing

Principle	Multicellular Organisms	Networked Computing
Collaboration between highly specialized cells	Cells in biofilms specialize temporarily according to "quorum" cues from neighbors. Cells in "true" multicellular organisms permanently specialize (differentiate) during development. Loss of differentiation is an early sign of cancer.	Today most computers retain a large repertoire of unused general behavior susceptible to viral or worm attack. Biology suggests that more specialization and less monoculture would be advantageous (although market forces may oppose this).
Communication by polymorphic messages	Cells in multicelled organisms communicate with each other via messenger molecules, *never* DNA. The "meaning" of cell-to-cell messages is determined by the receiving cell, not the sender.	Executable code is the analogue of DNA. Most PCs permit easy, and hidden, download of executable code (Active-X or even exe). However, importing executable code is well known to create security risks, and secure systems minimize or eliminate this capability.
"Self" defined by a stigmergic structure	Multicelled organisms and biofilms build extracellular stigmergic structures (bone, shell, or just slime) that define the persistent self. "Selfness" resides as much in the extracellular matrix as in the cells.	Determination of self is largely ad hoc in today's systems. However, an organization's intranet is a stigmergic structure, as are its persistent databases.
"Self" protected by programmed cell death (PCD)	Every healthy cell in a multicelled organism is prepared to commit suicide. PCD evolved to deal with DNA replication errors, viral infection, and rogue undifferentiated cells. PCD reflects a multicellular perspective—sacrificing the individual cell for the good of the multicellular organism.	A familiar example in computing is the Blue Screen of Death, which is a programmed response to an unrecoverable error. An analogous computer should sense its own rogue behavior (e.g., download of uncertified code) and disconnect itself from the network or reboot itself periodically to give itself a clean initial state.

SOURCE: Steve Burbeck, IBM, personal communication, October 11, 2004.

concurrency, multilevel description, and object orientation, Kam et al. constructed a T-cell simulation that presents its results by displaying animated versions of the model's Statecharts.

A second example is provided by the work of Searls. It is a common, if not inescapable, metaphor that DNA represents the language of life. In the late 1980s and early 1990s, David B. Searls and collaborators made the metaphor much more concrete, applying formal language theory to the analysis of nucleic acid sequences.[76] Linguistics theory considers four levels of interpretation of text: lexical (the

[76] D.B. Searls, "The Linguistics of DNA," *American Scientist* 80:579-591, 1992. Formal language theory is a major subfield of computer science theory; it is based on Noam Chomsky's work on linguistics in the 1950s and 1960s, especially the Chomsky hierarchy, a categorization of languages by their inherent complexity. Formal languages are at the heart of parsers and compilers, and there exists a wide range of both theoretic analysis and practical software tools for the production, transformation, and analysis of text. The main algorithmic tool of language theory is the generative grammar, a series of rules that transforms higher-level abstract units of meaning (such as "sentence" or "noun phrase") into more concrete potential statements in a given language. Grammars can be categorized into regular, context-free, context-sensitive, and recursively enumerable, each of which requires more algorithmic complexity to recognize than the level before it.

identification of specific words), syntactic (the grouping of words into grammatically correct phrases), semantic (the assignment of meaning to words and phrases), and pragmatic (the role of a piece of text in the larger context). These match entirely well to genomic analysis: grouping bases into codons, genes, the function of the resulting protein, and the role of that protein in the larger molecular system.[77]

Linguistic analyses can reveal or explain relationships between bases that are far apart in a sequence. For example, an RNA structure called a stem-loop has a palindrome-like sequence, with Watson-Crick pairs at equal distances away from the center. Traditional probabilistic or pattern-searching approaches would have some difficulty recognizing this structure, but it is quite simple with a grammar that produces palindromes. Some sequences of nucleic acids result in ambiguous linguistic interpretations; while this is a difficulty for computer languages, it represents a strength of biological linguistic analysis, because these ambiguities correctly represent alternative secondary structures.[78]

This approach has been fruitful for analyzing genetic sequences and characterizing the complexity and structure of genes. GenLang, a software system that employs linguistic approaches, has successfully identified tRNA genes, group I introns, protein-encoding genes, and the specification of gene regulatory elements.[79] Other important findings include placing RNA in the Chomsky hierarchy as at least beyond context-free languages. Finally, the approach provides a powerful tool for understanding the evolution of nucleic acid sequences; since the first sequences were most likely random (and thus regular languages), there must be a mechanism that somehow promoted sequence language into more powerful linguistic categories. This can be seen as an algebraic problem of operational closure, and the question is, For which string operations are regular languages and context-free languages not closed?[80]

[77] D.B. Searls, "Reading the Book of Life," *Bioinformatics* 17(7):579-580, 2001.

[78] D.B. Searls, "The Language of Genes," *Nature* 420(6912):211-217, 2002.

[79] D.B. Searls, and S. Dong, "A Syntactic Pattern Recognition System for DNA Sequences" in *Proceedings of the Second International Conference on Bioinformatics, Supercomputing, and Complex Genome Analysis*, H.A. Lim, J. Fickett, C.R. Cantor, and R.J. Robbins, eds., World Scientific Publishing Co., pp. 89-101, 1993.

[80] D.B. Searls, "Formal Language Theory and Biological Macromolecules," *Series in Discrete Mathematics and Theoretical Computer Science* 47:117-140, 1999.