9

Illustrative Problem Domains at the Interface of Computing and Biology

9.1 WHY PROBLEM-FOCUSED RESEARCH?

Problems offered by nature do not respect disciplinary boundaries. That is, nature does not package a problem as a "biology" problem, a "computing" problem, or a "physics" problem. Many disciplines may have helpful insights to offer or useful techniques to apply to a given problem, and to the extent that problem-focused research can bring together practitioners of different disciplines to work on shared problems, this can only be a good thing.

This chapter describes problem domains in which the expenditure of serious intellectual effort can reasonably be expected to generate (or to require!) significant new knowledge in biology and/or computing. Biological insight could take different forms—the ability to make new predictions, the understanding of some biological mechanism, the construction of a new biological organism. The same is true for computing—insight might take the form of a new biologically inspired approach to some computing problem, different hardware, or novel architecture. It is important to note that these domains contain very difficult problems—and it is unrealistic to expect major progress in a short time.

Challenge problems can often be found in interesting problem domains. A "challenge problem" is a scientific challenge focused on a particular intellectual goal or application (Box 9.1). Such problems have a long history of stimulating important research efforts, and a list of "grand challenges" in computational biology originating with David Searls, senior vice president of Worldwide Bioinformatics for GlaxoSmithKline, includes protein structure prediction, homology search, multiple alignment and phylogeny construction, genomic sequence analysis, and gene finding.[1] Appendix B provides a sampling of grand challenge problems found in other reports and from other life scientists.

The remainder of this chapter illustrates problem domains that display the intertwined themes of understanding biological complexity and enabling a novel generation of computing and information science. It incorporates many of the dimensions of the basic knowledge sought by each field and discusses some of the technical and biological hurdles that must be overcome to make progress. However, no claim whatsoever is made that these problems exhaust the possible interesting or legitimate domains at the BioComp interface.

[1]D.B. Searls, "Grand Challenges in Computational Biology," *Computational Methods in Molecular Biology*, S.L. Salzberg, D. Searls, and S. Kasif, eds., Elsevier Science, 1999.

Box 9.1
On Challenge Problems

Challenge problems have a history of stimulating scientific progress. For example:

- The U.S. High Performance Computing and Communications Program focused on problems in applied fluid dynamics, meso- to macroscale environmental modeling, ecosystem simulations, biomedical imaging and biomechanics, molecular biology, molecular design and process optimization, and cognition.[1] These problem domains were selected because they drove applications needs for very high-performance computing.
- A second example is the Text REtrieval Conference (TREC), sponsored by the National Institute of Standards and Technology, in cooperation with the National Security Agency and the Defense Advanced Research Projects Agency. The purpose of this conference is to "support research within the information retrieval community by providing the infrastructure necessary for large-scale evaluation of text retrieval methodologies. . . . The TREC workshop series has the following goals: to encourage research in information retrieval based on large test collections; to increase communication among industry, academia, and government by creating an open forum for the exchange of research ideas; to speed the transfer of technology from research labs into commercial products by demonstrating substantial improvements in retrieval methodologies on real-world problems; and to increase the availability of appropriate evaluation techniques for use by industry and academia, including development of new evaluation techniques more applicable to current systems."[2] TREC operates by presenting a problem in text retrieval clearly and opening it up to all takers. It makes available to the community at large all basic tools, and its structure and organization have attracted a large number of research sites.
- Still another approach to challenge problems is to offer prizes for the accomplishment of certain well-specified tasks. For example, in aeronautics, the Kremer Prize was established in 1959 for the first human-powered flight over a specific course; this prize was awarded to Paul MacReady for the flight of the *Gossamer Condor* in 1977. The Kremer Prize is widely regarded as having stimulated a good deal of innovative research in human-powered flight. A similar approach was taken in cryptanalysis, in which nominal prizes were offered for the first parties to successfully decrypt certain coded messages. These prizes served to motivate the cryptanalytic community by providing considerable notoriety for the winners. On the other hand, pressures to be the first to achieve a certain result often strongly inhibit cooperation, because sharing one's own work may eliminate the competitive advantage that one has over others.

[1]See http://www.ccic.gov/pubs/blue96/index.html.
[2]See http://trec.nist.gov/overview.html.

9.2 CELLULAR AND ORGANISMAL MODELING[2]

A living cell is a remarkable package of biological molecules engaged in an elaborate and robust choreography of biological functions. Currently, however, we have very incomplete knowledge about all of the components that make up cells and how these components interact to perform those functions. Understanding how cells work is one of biology's grand challenges. If it were possible to understand more completely how at least some of the machinery of cells works, it might be possible to anticipate the onset and effects of disease and create therapies to ameliorate those effects. If it were possible to influence precisely the metabolic operations of cells, they might be usable as highly controllable factories for the production of a variety of useful organic compounds.

However, cell biology is awash in data on cellular components and their interactions. Although such data are necessary starting points for an understanding of cellular behavior that is sufficient for prediction, control, and redesign, making sense out of the data is difficult. For example, diagrams tracing all of the interactions, activities, locations, and expression times of the proteins, metabolites, and nucleic acids involved have become so dense with lines and annotations that reasoning about their functions has become almost impossible.

[2]Section 9.2 is based largely on A.P. Arkin, "Synthetic Cell Biology," *Current Opinion in Biotechnology* 12(6):638-644, 2001.

As noted in Section 5.4.2, cellular simulation efforts have for the most part addressed selected aspects of cellular functionality. The grand challenge of cellular modeling and simulation is a high-fidelity model of a cell that captures the interactions between the many different aspects of functionality, where "high fidelity" means the ability to make reasonably accurate and detailed predictions about all interesting cellular behavior under the various environmental circumstances encountered in its life cycle. Of course, a model perforce is an abstraction that omits certain aspects of the phenomenon it is representing. But the key term in this description is "interesting" behavior—behavior that is interesting to researchers. In this context, the model is intended to integrate—as a real cell would—different aspects of its functionality. Although the grand challenge may well be unachievable, almost by definition, the goal of increasing degrees of integration of what is known and understood about various aspects of cellular function remains something for which researchers strive.

The development of a high-fidelity simulation of a cell—even the simplest cell—is an enormous intellectual challenge. Indeed, even computational models that are very well developed, such as models of neural and cardiac electrophysiology, often fail miserably when they are exercised beyond the data that have been used to construct them. Yet if a truly high-fidelity simulation could be developed, the ability to predict cellular response across a wide range of environmental conditions *using a single model* would imply an understanding of cellular function far beyond what is available today, or even in the immediate future, and would be a tangible and crowning achievement in science. And, of course, the scientific journey to such an achievement would have many intermediate payoffs, in terms of tools and insights relevant to various aspects of cellular function. From a practical standpoint, such a simulation would be an invaluable aid to medicine and would provide a testbed for biological scientists and engineers to explore techniques of cellular control that might be exploited for human purposes.

An intermediate step toward the high-fidelity simulation of a real cell would be a model of a simple hypothetical cell endowed with specific properties of real cells. This model would necessarily include representations of several key elements (Box 9.2). The hundreds of molecules and hundreds of thousands of interactions required do not appear computationally daunting, until it is realized that the time scale of molecular interaction is on the order of femtoseconds, and interesting time scales of cellular response may well be hours or days.

The challenges fall into three general categories:

- *Mechanistic understanding*. High-fidelity simulations will require a much more profound physical understanding of basic biological entities at multiple levels of detail than is available today. (For example, it is not known how RNA polymerase actually moves along a DNA strand or what rates of binding or unbinding occur in vivo.) An understanding of how these entities interact inside the cell is equally critical. Mechanistic understanding would be greatly facilitated by the development of new mathematical formalisms that would enable the logical parsing of large networks into small modules

Box 9.2
Elements of a Hypothetical Cell

- An outside and inside separated by some coat or membrane (e.g., lipid)
- One or more internal compartments inside the cell
- Genes and an internal code for regulation of function
- An energy supply to keep the cell "alive" or "working"
- Reproductive capability
- At least hundreds of biologically significant molecules, with potentially hundreds of thousands of interactions between them
- Responsiveness to environmental conditions that affect the internal operation and behavior of the cell (e.g., changes in temperature, acidity, salinity)

whose behavior can be analyzed. Such modules would be building blocks that researchers could use to build functionality, understand controllable aspects, and identify points of failure.

- *Data acquisition*. Simulation models are data-intensive, and today there are relatively few systems with enough quality data to create highly detailed models of cellular function. It will be important to develop ways of measuring many more aspects of internal cellular state, and in particular, new techniques for measuring rates of processes and biochemical reactions in situ in living cells will be necessary. Besides additional reporter molecules, selective fluorescent dyes, and so on, a particular need is to develop good ways of tracking cellular state at different points in time, so that cellular dynamics can be better understood. Large volumes of data on reaction rates will also be necessary to model many cellular processes.
- *Integrative tools*. Because cellular function is so complex, researchers have used a variety of data collection techniques. Data from multiple sources—microarrays, protein mass spectroscopy, capillary and high-pressure chromatographies, high-end fluorescence microscopy, and so on—will have to be integrated—and are indeed required—if validated, high-fidelity cellular models are to be built. Moreover, because existing models and simulations relevant to a given cell span multiple levels of organizational hierarchy (temporal, spatial, etc.), tools are necessary to facilitate their integration. With such tools at the researcher's disposal, it will be possible to develop complex models rapidly, assembling molecular components into modules, linking modules, computing dynamic interactions, and comparing predictions to data.

Finally, despite the power of cellular modeling and simulation to advance understanding, models should not be regarded as an end product in and of themselves. Because all models are unfaithful to the phenomena they represent in some way, models should be regarded as tools to gain insight and to be used in continual refinement of our understanding, rather than as accurate representations of real systems, and model predictions should be taken as promising hypotheses that will require experimental validation if they are to be accepted as reliable.

The discussion above suggests that many researchers will have to collaborate in the search for an integrated understanding. Such coordinated marshaling of researchers and resources toward a shared goal is a common model for industry, but this multi-investigator approach is new for the academic environment. Large government-funded projects such as the Alliance for Cellular Signaling (discussed in Chapter 4) or private organizations like the Institute for Systems Biology[3] are the new great experiments in bringing a cooperative approach to academic biology.

Still more ambitious—probably by an order of magnitude or more—is the notion of simulating the behavior of a multicelled organism. For example, Harel proposes to develop a model of the *Caenorhabditis elegans* nematode, an organism that is well characterized with respect to its anatomy and genetics.[4] Harel describes the challenge as one of constructing "a full, true-to-all-known-facts, 4-dimensional, fully animated model of the development and behavior of this worm. . . , which is easily extendable as new biological facts are discovered."

In Harel's view, the feasibility of such a model is based on the notion that the complexity of biological systems stems from their high reactivity (i.e., they are highly concurrent and time-intensive, exhibit hybrid behavior that is predominantly discrete in nature but with continuous aspects as well, and consist of many interacting, often distributed, components). The structure of a reactive system may itself be dynamic, with its components being repeatedly created and destroyed during the system's life span. Harel notes:

---

[3]See http://www.systemsbiology.org/home.html.

[4]D. Harel, "A Grand Challenge for Computing: Towards Full Reactive Modeling of a Multi-Cellular Animal," *European Association for Theoretical Computer Science (EATCS) Bulletin*, 2003, available at http://www.wisdom.weizmann.ac.il/~dharel/papers/GrandChallenge.doc.

> [B]iological systems exhibit the characteristics of reactive systems remarkably, and on many levels; from the molecular, via the cellular, and all the way up to organs, full organisms, and even entire populations. It doesn't take much to observe within such systems the heavy concurrency, the event-driven discrete nature of the behavior, the chain-reactions and cause-effect phenomena, the time-dependent patterns, etc.

Harel concludes that biological systems can be modeled as reactive systems, using languages and tools developed by computer science for the construction of man-made reactive systems (briefly discussed in Section 5.3.4 and at greater length in the reference in footnote 4 of this chapter).

If the Harel effort is successful, a model of *C. elegans* would result that is fully executable, flexible, interactive, comprehensive, and comprehensible. By realistically simulating the worm's development and behavior, it would help researchers to uncover gaps, correct errors, suggest new experiments, predict unobserved phenomena, and answer questions that cannot be addressed by standard laboratory techniques alone. In addition, it would enable users to switch rapidly between levels of detail (from the entire macroscopic behavior of the worm to the cellular and perhaps molecular levels). Most importantly, the model would be extensible, allowing biologists to enter new data themselves as they are discovered and to test various hypotheses about aspects of behavior that are not yet known.

9.3 A SYNTHETIC CELL WITH PHYSICAL FORM

The most ambitious goal of synthetic biology (Section 8.4.2) is the biochemical instantiation of a real—if synthetic—cell with the capability to grow and reproduce. Such an achievement would necessarily be accompanied by new insights into the molecular dynamics of cells, the origins of life on Earth, and the limits of biological life. In practical terms, such cells could be engineered to perform specific functions, and thus could serve as a platform for innovative industrial and biomedical applications.

Cellular modification has a long history ranging from the development of plasmids carrying biosynthetic genes, or serving as "engineering blanks" for production of new materials, to the creation of small genetic circuits for the control of gene expression. However, the synthetic cells being imagined today would differ from the original cell much more substantially than those that have resulted from modifications to date. In principle, these cells need have no chemical or structural similarity to natural cells. Since they will be designed, not evolved, they may contain functions or structures unachievable through natural selection.

Synthetic cells are a potentially powerful therapeutic tool that may be able to deliver drugs to damaged tissue to seek and destroy foreign cells (in infections), destroy malignant cells (in cancer), remove obstructions (in cardiovascular disease), rebuild or correct defects (e.g., reattach severed nerves), or replace parts of tissue that was injured—and doing so without affecting nonproblematic tissues, while reducing the side effects of current conventional treatments.

The applications of synthetic cells undertaking cell-level process control computing are not limited to those of medicine and chemical sensing. There are also potential applications to the nanofabrication of new and useful materials and structures. Indeed, natural biology exhibits propulsive rotors and limbs at the microscale, and synthetic cells may be an enabling technology for nanofabrication—the building of structures at the microscopic level. There may be other techniques to accomplish this, but synthetic cells offer a promise of high efficiency through massively parallel reproduction. The gene regulatory networks incorporated into synthetic cells allow for the simultaneous creation of multiple oligonucleotide sequences in a programmable fashion. Conversely, self-assembled DNA nanostructures can potentially be used as control structures that interact with intracellular components and molecules. Such control could enable the engineering construction of complex extracellular structures and precise control of fabrication at the subnanometer level, which might in turn lead to the construction of complex molecular-scale electronic structures (Section 8.4.3.2) and the creation of new biological materials, much as natural biological materials result from natural biological processes.

Constructing these structures will require the ability to fabricate individual devices and the ability to assemble these devices into a working system, since it is likely to be very difficult to assemble a system directly from scratch. One approach to an assembly facility is to use a mostly passive scaffold, consisting of selectively self-assembling molecules that can be used to support the fabrication of molecular devices that are appropriately interconnected. Indeed, DNA molecules and their attendant enzymes are capable of self-assembly. By exploiting that capability, it has been possible to create a number of designed nanostructures, such as tiles and latticed sheets. Although the characteristics of these biomaterials need further exploration, postulated uses of them include as scaffolds (for example, for the crystallization of macromolecules); as photonic materials with novel properties; as designable zeolite-like materials for use as catalysts or as molecular sieves; and as platforms for the assembly of molecular electronic components or biochips.[5] Uses of DNA as a molecular "Lego" kit with which to design nanomachines, such as molecular tweezers and motors on runways, are also under investigation.

The relevance of synthetic cell engineering to nanofabrication is driven by the convergence of developments in several areas, including the miniaturization of biosensors and biochips into the nanometer-scale regime, the fabrication of nanoscale objects that can be placed in intracellular locations for monitoring and modifying cell function, the replacement of silicon devices with nanoscale, molecular-based computational systems, and the application of biopolymers in the formation of novel nanostructured materials with unique optical and selective transport properties. The highly predictable hybridization chemistry of DNA, the ability to completely control the length and content of oligonucleotides, and the wealth of enzymes available for modification of DNA make nucleic acids an attractive candidate for all of these applications.

Furthermore, by designing and implementing synthetic cells, a much better understanding will be gained of how real cells work, how they are regulated, and what limitations are inherent in their machinery. Here, the discovery process is iterative, in that our understanding and observations of living cells serve as "truthing" mechanisms to inform and validate or refute the experimental constructs of synthetic cells. In turn, the mechanisms underlying synthetic cells are likely to be more easily understood than comparable ones in natural cells. Using this combined information, the behavior of biological processes in living cells can slowly be unraveled. For such reasons, the process of creating synthetic cells will spin off benefits to biology and science, just as the Human Genome Project led to dramatic improvements in the technology of molecular biology.

To proceed with the creation of synthetic cells, three separate but interrelated problems must be addressed:

- The theoretical and quantitative problem of formulating, understanding, and perhaps even optimizing the design of a synthetic cell;
- The biological problem of applying lessons learned from real cells to such designs and using synthetic cells to inform our understanding of more complicated natural cells; and
- The engineering and chemistry problem of assembling the parts into a physical system (or to design self-assembling pieces).

One approach to building such a cell de novo is to start with a set of parts and assemble them into a functional biomolecular machine. Conceiving a cell de novo means that cellular components and their assembly are predetermined, and that the cell engineer has a quantifiable understanding of events and outcomes that can be used to predict the behavior of the components and their interactions at least probabilistically. A key aspect of de novo construction is that a de novo cellular design is not constrained by evolutionary history and hence is much more transparent than cells found in nature. Be-

[5]E. Winfree, F. Liu, L.A. Wenzler, and N.C. Seeman, "Design and Self-Assembly of Two-Dimensional DNA Crystals," *Nature* 394(6693):539-544, 1998.

cause an engineered cell would be designed by human beings, the functions of its various elements would be much better known. This fact implies that it would be easier to identify critical control points in the system and to understand the rules by which the system operates.

A second approach is to modify an existing living cell to give it new behaviors or to remove unwanted behaviors; classical metabolic engineering and natural product synthesis would be relevant to this approach. One starting point would be to use the membrane of an existing cell, but modification of these lipid bilayers to incorporate chemically inducible channels, integrated inorganic structures for sensing and catalysis, and other biopolymer structures for the identification and modification of biological substrates will provide a greater degree of freedom in the manipulation of the chemical state of the synthetic cell.

A third approach is to abandon DNA-based cells. Szostak et al.[6] argue that the "stripping-down" of a present-day bacterium to its minimum essential components still leaves hundreds of genes and thousands of different proteins and other molecules. They suggest that synthetic cells could use RNA as the repository of "genetic" information and as enzymes that catalyze metabolism. In their view, the most important requirements of a synthetic cell from a scientific standpoint are that it replicates autonomously and that it is subject to evolutionary forces. In this context, autonomous replication means continued growth and division that depends only on the input of small molecules and energy, not on the products of preexisting living systems such as protein enzymes. Evolution in this context means that the structure is capable of producing different phenotypes that are subject to forces of natural selection, although being subject to evolutionary forces has definite disadvantages from an engineering perspective seeking practical application of synthetic cells.

The elements of a synthetic cell are likely to mirror those of simulations (see Box 9.2), except of course that they will take physical representation. Inputs to the synthetic cell would take the form of environmental sensitivities of various kinds that direct cellular function. (Another perspective on "artificial cells" similar to this report's notion of synthetic cells is offered by Pohorille.[7] In general, synthetic cells share much with artificial cells, and the dividing line between them is both blurry and somewhat arbitrary. The modal use of the term "artificial cell" appears to refer to an entity with a liposome membrane, whose physical dimensions are comparable to those of natural cells, that serves a function such as enzyme delivery, drug delivery for cell therapy, and red blood cell substitutes.[8]) However, if synthetic cells are to be useful or controllable, it will be necessary to insert control points that can supply external instructions or "reprogram" the cell for specialized tasks (e.g., a virus that injects DNA into the cell to insert new pieces of code or instructions).

Researchers are interested in expanding the size and complexity of pathways for synthetic cells that will do more interesting things. But there is little low-hanging fruit in this area, and today's computational and mathematical ability to predict cellular behavior quantitatively is inadequate to do so, let alone to select for the desired behavior. To bring about the development of synthetic cells from concept to practical reality, numerous difficulties and obstacles must be overcome. Following is a list of major challenges that have to be addressed:

- *A framework for cellular simulation that can specify and model cellular function at different levels of abstraction* (as described in Section 9.2). Simulations will enable researchers to test their proposed designs, minimizing (though not eliminating) the need for in vivo construction and experimentation. Note that the availability of such a framework implies that the data used to support it are also available to assist in the engineering development of synthetic cells.

---

[6]J.W. Szostak, D.P. Bartel, and P.L. Luisi, "Synthesizing Life," *Nature* 409(6818):387-390, 2001.

[7]A. Pohorille, "Artificial Cells: Prospects for Biotechnology," *Trends in Biotechnology* 20(3):123-128, 2002.

[8]See, for example, T.M.S. Chang, "Artificial Cell Biotechnology for Medical Applications," *Blood Purification* 18:91-96, 2000, available at http://www.medicine.mcgill.ca/artcell/isbp.pdf.

Box 9.3
Tool Suites

One tool suite is a simulator and verifier for genetic digital circuits, called BioSPICE. The input to BioSPICE is the specification of a network of gene expression systems (including the relevant protein products) and a small layout of cells on some medium. The simulator computes the time-domain behavior of concentration of intracellular proteins and intercellular message-passing chemicals. (For more information, see http://www.biospice.org.)

A second tool would be a "plasmid compiler" that takes a logic diagram and constructs plasmids to implement the required logic in a way compatible with the metabolism of the target organism. Both the simulator and the compiler must incorporate a database of biochemical mechanisms, their reaction kinetics, their diffusion rates, and their interactions with other biological mechanisms.

- *Stability and robustness in the face of varying environmental conditions and noise.* For example, it is well known that nature provides a variety of redundant pathways for biological function, so that (for example) the incapacitation of one gene is often not unduly disruptive to the cell.
- *Improvement in the libraries of DNA-binding proteins and their matching repressor patterns.* These are at present inadequate, and good data about their kinetic constants are unavailable (hence signal transfer characteristics cannot be predicted). Any specific combination of proteins might well interact outside the genetic regulatory mechanisms involved, thus creating potentially undesirable side effects.
- *Control point design and insertion.*
- *Data measurement and acquisition.* To facilitate the monitoring of a synthetic cell's behavior, it is desirable to incorporate into the structure of the cell itself methods for measuring internal state parameters. Such measurements would be used to parameterize the functionality of cellular elements and compare performance to specifications.
- *Deeper understanding of biomolecular design rules.* Engineering of proteins for the modification of biointeractions will be required in all aspects of cell design, because it is relevant to membrane-based receptors, protein effectors, and transcriptional cofactors. Today, metabolic engineers are frequently frustrated in attempts to reengineer metabolic pathways for new functions because, at this point, the "design principles" of natural cells are largely unknown. To design, fabricate, and prototype cellular modules, it must be possible to engineer proteins that will bind to DNA and regulate gene expression. Current examples of DNA binding proteins are zinc fingers, response regulators, and homeodomains. The goal is to create flexible protein systems that can be modified to vary binding location and strength and, ultimately, to insert these modules into living cells to change their function.
- *A "device-packing" design framework that allows the rapid design and synthesis of new networks inside cells.* This framework would facilitate designs that allow the reuse of parts and the rapid modification of said parts for creating various "modules" (switches, ramps, filters, oscillators, etc.). The understanding available today regarding how cells reproduce and metabolize is not sufficient to enable the insertion of new mechanisms that interact with these functions in predictable and reliable ways.
- *Tool suites to support the design, analysis, and construction of biologic circuits.* Such suites are as yet unavailable (but see Box 9.3).

9.4 NEURAL INFORMATION PROCESSING AND NEURAL PROSTHETICS

Brain research is a grand challenge area for the coming decades. In essence, the goal of neuroscience research is to understand how the interplay of structural dynamics, biochemical processes, and electri-

cal signals in nervous tissue gives rise to higher-order functions such as normal or abnormal thoughts, actions, memories, and behaviors. Experimental advances of the past decades have given the brain researcher an increasingly powerful arsenal of tools to obtain data—from the level of molecules to nervous systems—and to compare differences between individuals.

Today, neuroscientists have begun the arduous process of adapting and assembling neuroscience data at all scales of resolution and across disciplines into electronically accessible, distributed databases. These information repositories will complement the vast structural and sequence databases created to catalog, organize, and analyze gene sequences and protein products. Such databases have proven enormously useful in bioinformatics research; whether equal rewards will accrue from similar efforts for tissue-level data, whole-brain imaging, physiological data, and so forth remains to be seen, but based on the successes of the molecular informatics activities and the challenge questions of the neuroscientist, big payoffs can be anticipated.

At the very least, multiscale informatics efforts for brain research will provide organizing frameworks and computational tools to manage neuroscience data, from the lab notebook to published data. An ideal and expected outcome will be the provisioning for new opportunities to integrate large amounts of biological data into unified theories of function and aid in the discovery process.

To provide some perspective on the problem, consider that animal brains are the information-processing systems of nature. A honeybee's brain contains roughly 100 million synapses; a contemporary computer contains roughly 100 million transistors. Given a history of inputs, both systems choose from among a set of possible outputs. Yet although it is understood how a digital computer adds and subtracts numbers and stores error-free data, it is not understood how a honeybee learns to find nectar-rich flowers or to communicate with other honeybees.

We do not expect a honeybee to perform numerical computations; likewise, we do not expect a digital computer to learn autonomously, at least not today. However, an interesting question is the extent to which the structure of an information-processing system and the information representations that it uses predispose the system to certain types of computation. Put another way, in what ways and under what circumstances, if any, are neuronal circuits and neural information-processing systems inherently superior to von Neumann architectures and Shannon information representations for adaptation and learning? Given the desirability of computers that can learn and adapt, an ability to answer this question might provide some guidance in the engineering of such systems.

Some things are known about neural information processing:

- Animal brains find good solutions to real-time problems in image and speech processing, motor control, and learning. To perform these tasks, nervous systems must represent, store, and process information. However, it is highly unlikely that neural information is represented in digital form.
- It is likely that neurons are the nervous system's primary computing elements. A typical neuron is markedly unlike a typical logic gate; it possesses on average 10,000 synaptic inputs and a similar number of outputs.
- The stored memory of a neural information-processing system is contained in the pattern and strength of the analog synapses that connect it to other neurons. Nervous systems use vast numbers of synapses to effect their computations: in neocortical tissue, the synapse density is roughly 3×10^8 synapses per cubic millimeter.[9] Specific memories are also known not to be localized to particular neurons or sets of neurons in the brain.[10]

[9]R. Douglas, "Rules of Thumb for Neuronal Circuits in the Neocortex," *Notes for the Neuromorphic VLSI Workshop*, Telluride, CO, 1994.

[10]The essential reason is that specific memories are generally richly and densely connected to other memories, and hence can be reconstructed through that web of connections.

- The disparity between the information processing that can be done by digital computers and that done by nervous systems is likely to be a consequence of the different way in which nerve tissue represents and processes information, although this representation is not understood.
- At the device level, nervous tissue operates on physical principles that are similar to those that underlie semiconductor electronics.[11] Thus, differences between neural and silicon computation must be the result of differences in computational architecture and representation. It is thus the higher-level organization underlying neural computation that is of interest and relevance. Note also that for the purposes of understanding neural signaling or computation, a neuron-by-neuron simulation of nervous tissue per se cannot be expected to reveal very much about the principles of organization, though it may be necessary for the development of useful artifacts (e.g., neural prostheses).

Some of the principles underlying neural computation are understood. For example, neurobiology uses continuous adaptation rather than absolute precision in responding to analog inputs. The dynamic range of the human visual system is roughly 10 decades in input light intensity—about 32 bits. But biology doesn't process visual signals with 32-bit precision; rather, it uses a 7- or 8-bit instantaneous dynamic range and adapts the visual pathway's operating point based on the background light intensity. Although this approach is similar to the automatic gain control used in electronic amplifiers, biology takes the paradigm much farther: adaptation pervades every level of the visual system, rather than being concentrated just at the front end.[12]

There are essentially two complementary approaches toward gaining a greater understanding of neural information processing. One approach is to reproduce physiological phenomena to increase our understanding of the nervous system.[13] A second approach is based on using a manageable subset of neural properties to investigate emergent behavior in networks of neuron-like elements.[14] Those favoring the first approach believe that these details are crucial to understanding the collective behavior of the network and are developing probes that are increasingly able to include the relevant physiology. Those favoring the second approach make the implicit assumption that reproducing many neurophysiological details is secondary to understanding the collective behavior of nervous tissue, even while acknowledging that only a detailed physiological investigation can reveal definitively whether the details are in fact relevant.

What can be accomplished by building silicon circuits modeled after biology? First, once the neuronal primitives are known, it will be possible to map them onto silicon. Once it is understood how biological systems compute with these primitives, biologically based silicon computing will be possible. Second, we can investigate how physical and technological limits, such as wire density and signal delays and noise, constrain neuronal computation. Third, we can learn about alternative models of computation. Biology demonstrates nondigital computing machines that are incredibly space- and energy-efficient and that find adequate solutions to ill-posed problems naturally.

---

[11]In both integrated circuits and nervous tissue, information is manipulated principally on the basis of charge conservation. In the former, electrons are in thermal equilibrium with their surroundings and their energies are Boltzmann distributed. In the latter, ions are in thermal equilibrium with their surroundings and their energies also are Boltzmann distributed. In semiconductor electronics, energy barriers are used to contain the electronic charge, by using the work function difference between silicon and silicon dioxide or the energy barrier in a *pn* junction. In nervous tissue, energy barriers are also erected to contain the ionic charge, by using lipid membranes in an aqueous solution. In both systems, when the height of the energy barrier is modulated, the resulting current flow is an exponential function of the applied voltage, thus allowing devices that exhibit signal gain. Transistors use populations of electrons to change their channel conductance, in much the same way that neurons use populations of ionic channels to change their membrane conductance.

[12]Adaptation helps to explain why some biological neural systems never settle down—they can be built so that when faced with unchanging inputs, the inputs are adapted away. This phenomenon helps to explain many visual aftereffects. A stabilized image on the retina disappears after a minute or so, and the whole visual field appears gray.

[13]M.A. Mahowald and R.J. Douglas, "A Silicon Neuron," *Nature* 354(6354):515-518, 1991.

[14]J. Hertz, A. Krogh, and R.G. Palmer, *Introduction to the Theory of Neural Computation*, Addison-Wesley, Reading, MA, 1991.

The challenges of neural information processing fall into two primary categories: the semantics of neural signaling and the development of neural prostheses. Signaling is the first challenge. It is known that the spike trains of neurons carry information in some way—neurons that cannot "fire" are essentially dead.[15] Also, the physical phenomena that constitute "firing" are known—electrical spikes of varying amplitude and timing. However, the connections among these patterns of signaling in multiple neurons to memories of specific events, motor control of muscles, sensory perception, or mental computation are entirely unknown. How do neurons integrate data from large numbers of multimodal sensors? How do they deal with data overload? How do they decide a behavioral response from multiple alternatives under severe and ill-posed constraints?

Today's neural instrumentation (e.g., positron emission tomography [PET] scans, functional magnetic resonance imaging [fMRI]) can identify areas of the brain that are active under various circumstances, but since the spatial resolution of these probes is wholly inadequate to resolve individual neuronal activity,[16] such instrumentation can provide only the roughest guidance about where researchers need to look for more information about neuronal signaling, rather than anything specific about that information itself. The primary challenge in this domain is the development of a formalism for neuronal signaling (most likely a time-dependent one that takes kinetics into account), much like the Boolean algebra that provides a computational formalism based on binary logic levels in the digital domain.

A step toward a complete molecular model of neurotransmission for an entire cell is provided by MCell, briefly mentioned in Chapter 5. MCell is a simulation program that can model single synapses and groups of synapses. To date, it been used to understand one aspect of biological signal transduction, namely the microphysiology of synaptic transmission. MCell simulations provide insights into the behavior and variability of real systems comprising finite numbers of molecules interacting in spatially complex environments. MCell incorporates high-resolution physical structure into models of ligand diffusion and signaling, and thus can take into account the large complexity and diversity of neural tissue at the subcellular level. It models the diffusion of individual ligand molecules used in neural signaling using a Brownian dynamics random walk algorithm, and bulk solution rate constants are converted into Monte Carlo probabilities so that the diffusing ligands can undergo stochastic chemical interactions with individual binding sites, such as receptor proteins, enzymes, and transporters.[17]

The second challenge is that of neural prosthetics. A neural prosthesis is a device that interfaces directly with neurons, receiving and transmitting signals that affect the function and activity of those neurons, and that behaves in predictable and useful ways. Perhaps the "simplest" neural prosthesis is an artificial implant that can seamlessly replace nonfunctioning nerve tissue.

Today, some measure of cognitive control of artificial limbs can be achieved through bionic brain-machine or peripheral-machine interfaces. William Craelius et al.[18] have designed a prosthetic hand that offers amputees control of finger flexion using natural motor pathways, enabling them to undertake slow typing and piano playing. The prosthetic hand is based on the use of natural tendon movements in the forearm to actuate virtual finger movement. A volitional tendon movement within the residual limb causes a slight displacement of air in foam sensors attached to the skin in that location, and the resulting pressure differential is used to control a multifinger hand.

---

[15]It is also known that not all neural signaling is carried by spikes. A phenomenon known as graded synaptic transmission also carries neural information and is based on a release of neurotransmitter at synaptic junctions whose volume is voltage dependent and continuous. Graded synaptic transmission appears to be much more common in invertebrates and sometimes exists alongside spike-mediated signaling (as in the case of lobsters). The bandwidth of this analog channel is as much as five times the highest rates measured in spiking neurons (see, for example, R.R. de Ruyter van Steveninck and S.B. Laughlin, "The Rate of Information Transfer at Graded-Potential Synapses," *Nature* 379:642-645, 1996), but the analog channel is likely to suffer a much higher susceptibility to noise than do spike-mediated communications.

[16]The spatial resolution of neural instrumentation is on the order of 1 to 10 mm. See D. Purves et al., *Neuroscience*, Sinauer Associates Inc., Sunderland, MA, 1997. Given about 3×10^8 synapses per cubic millimeter, not much localization is possible.

[17]See http://www.mcell.cnl.salk.edu/.

[18]W. Craelius, R.L. Abboudi, and N.A. Newby, "Control of a Multi-finger Prosthetic Hand," *ICORR '99: International Conference on Rehabilitation Robotics*, Stanford, CA, 1999.

A second example of a neural prosthesis is a retinal prosthesis intended to provide functionality when the retina of the eye is nonfunctional. In one variant, a light-sensitive microchip is implanted into the back of the eye. Light striking the microchip (which has thousands of individual sensors) generates electrical signals that travel through the optic nerve to the brain and are interpreted as an image.[19] In another variant, the retina is bypassed entirely through the use of a camera mounted on a pair of eyeglasses to capture and transmit a light image via a radio signal to a chip implanted near the ganglion cells, which send nerve impulses to the brain.[20] In a third variant, an implanted microfluidic chip that controls the flow of neurotransmitters translates digital images into neurochemical signals that provide meaningful visual information to the brain. The microfluidic chip has a two-dimensional array of small controllable pores, corresponding to pixels in an image. An image is created by the selective drip of neurotransmitters onto specific bipolar cells, which are the cells that carry retinal information to the brain.[21]

A third example of work in this area is that of Musallam et al., who have demonstrated the feasibility of a neural interface that enables a monkey to control the movement of a cursor on a computer screen by thinking about a goal the monkey would like to achieve and assigning a value to that goal.[22] The interesting twist to this work is the reliance of signals from parts of the brain related to higher-order ("cognitive") brain functions for movement planning for the control of a prosthetic device. (Previous studies have relied on lower-level signals from the motor cortex.[23])

The advantage of using higher-level cognitive signals is that they capture information about the monkey's goal (moving the cursor) and preferences (the destination on the screen the monkey wants). Musallam et al. point out that once the signals associated with the subject's goals are decoded, a smart external device can perform the lower-level computations necessary to achieve the goals. For example, a smart robotic arm would be able to understand what the intended goal of an arm movement is and then compute—on its own—the trajectory needed to move the arm to that position. Furthermore, the abstract nature of a cognitive command would allow it to be used for the control and operation of a number of different devices. If higher-level signals associated with speech or emotion could be decoded, it would become possible to record thoughts from speech areas (reducing the need for the use of cumbersome letter boards and time-consuming spelling programs) or to provide online indications of a patient's emotional state.

A fourth example is provided by Theodore Berger of the University of Southern California, who is attempting to develop an artificial hippocampus—a silicon implant that will behave neuronally in a manner identical to the brain tissue that it replaces.[24] The hippocampus is the part of the brain responsible for encoding experiences so that they can be stored as long-term memories elsewhere in the brain; without the hippocampus, a person is unable to store new memories but can recall ones stored prior to its loss. Because the manner in which the hippocampus stores information is unknown, Berger's approach is based on designing a chip that can provide the identical input-output response. The input-

---

[19]N.S. Peachey and A.Y. Chow, "Subretinal Implantation of Semiconductor-based Photodiodes: Progress and Challenges," *Journal of Rehabilitation Research and Development* 36(4):371-376, 1999.

[20]W. Liu, E. McGucken, M. Clements, S.C. DeMarco, K. Vichienchom, C. Hughes, et al., "Multiple-Unit Artificial Retina Chipset System to Benefit the Visually Impaired," to be published in *IEEE Transactions on Rehabilitation Engineering*. Available at http://www.icat.ncsu.edu/projects/retina/files/MARC_system_paper.pdf.

[21]B. Vastag, "Future Eye Implants Focus on Neurotransmitters," *Journal of the American Medical Association* 288(15):1833-1834, 2002.

[22]S. Musallam, B.D. Corneil, B. Greger, H. Scherberger, and R.A. Andersen, "Cognitive Control Signals for Neural Prosthetics," *Science* 305(5681):258-262, 2004. A Caltech press release of July 8, 2004, available at http://pr.caltech.edu/media/Press_Releases/PR12553.html, describes this work in more popular terms.

[23]J. Wessberg, C.R. Stambaugh, J.D. Kralik, P.D. Beck, M. Laubach, J.K. Chapin, J. Kim, S.J. Biggs, M.A. Srinivasan, and M.A.L. Nicolelis, "Real-Time Prediction of Hand Trajectory by Ensembles of Cortical Neurons in Primates," *Nature* 408(6810):361-365, 2000. Similar work on rats is described in J.K. Chapin, K.A. Moxon, R.S. Markowitz, and M.A.L. Nicolelis, "Real-Time Control of a Robot Arm Using Simultaneously Recorded Neurons in the Motor Cortex," *Nature Neuroscience* 2(7):664-670, 1999.

[24]R. Merritt, "Nerves of Silicon: Neural Chips Eyed for Brain Repair," *EE Times*, March 17, 2003 (10:37 a.m. EST), available at http://www.eetimes.com/story/OEG20030317S0013.

output response of a hippocampal slice was determined by stimulating it with a random-signal generator, and a mathematical model was developed to account for its response to these different stimuli. This model is then the basis for the chip circuitry.

By December 2003, Berger and his colleagues had completed the first test of using a microchip model to replace a portion of the hippocampal circuitry contained in a specific hippocampal brain slice. In that slice is the major intrinsic circuitry of the hippocampus that consists of three major cell fields, designated A, B, and C. Field A projects to and excites field B, which projects to and excites field C. Berger et al. developed a predictive mathematical model of the signal transformations that field B performs on the input signals that come from field A, and that field B then projects onto field C, and implemented the model in a field-programmable gate array (FPGA) for field B. When field B was surgically removed and the FPGA model of B was substituted, the result was that the output from area C of the hippocampal slice remained unchanged in all meaningful respects. Next steps beyond this work (e.g., developing circuitry that is less sensitive to the details of slice preparation, understanding the hardware in terms of meaningful abstractions) remain to be realized.

One result of such work may be the creation of building blocks that can be used to calculate universal mathematical functions and ultimately be the basis of families of devices for neural pattern matching. Such building blocks may also serve as a point of departure for understanding neural functions at a higher level of abstraction than is possible today.

An analogy might be drawn to finding a mathematical representation of a particular dataset. The approach of mapping an exhaustive input-output response is similar to a curve-fitting process that generates a function capable of reproducing the dataset perfectly. Knowledge of such a function does not necessarily entail any understanding of the casual mechanisms underlying that dataset; thus, a function resulting from a curve-fitting process is highly unlikely to be able to account for new data. Still, developing such a function may be the first step toward such understanding.

As suggested above, building a successful neural prosthetic implies some understanding of the semantics of neural information processing: how the relevant nerve tissue stores and replicates and processes information. However, it also requires a well-understood interface between a biological organism (e.g., a person) and the engineered device.

One of the primary challenges in the area of neural interface design is the physical connection of neurons to a chip—the right neurons must make connection with the right electrodes. The body's natural response to an electrode implanted in living tissue is to wall it off with glial cells that prevent neuron and electrode from making contact. One approach to solving this problem is to coat the electrode with a substance that does not trigger the glial reaction. Another is to rely on the neural tissue to reconfigure itself. Based on the knowledge that auditory nerves can reconfigure themselves to accommodate the signals emitted by cochlear implants, it may be possible to send out a signal that attracts the right nerves to the right contacts.

Prosthetic devices that restore or augment human physical abilities are increasingly sophisticated, and follow-on work will focus on enabling control of more complex actions by robotic arms and other devices. On the other hand, although some early work on prostheses that help to replace cognitive abilities has been successful, prostheses that improve cognitive abilities, by enhancing perception (superhuman sense) and decision-making (superhuman computation or knowledge) capabilities, must at present be regarded as being on the distant horizon.

9.5 EVOLUTIONARY BIOLOGY[25]

Although the basic principles of evolution (natural selection and mutation) are understood in the large, both population genetics and phylogenetics have been radically transformed by the recent avail-

[25]Section 9.5 is adapted largely from the Web page of John Huelsenbeck, University of California, San Diego, http://biology.ucsd.edu/faculty/huelsenbeck.html.

ability of large quantities of molecular data. For example, in population genetics (the study of mutations in populations), more molecular variability was found in the 1960s than had been expected, and this finding stimulated Kimura's neutral theory of molecular evolution.[26] Phylogenetics (the study of the evolutionary history of life) makes use of a variety of different kinds of data, of which DNA sequences are the most important, as well as whole-genome, metabolic, morphological, geographical, and geological data.[27]

Evolutionary biology is founded on the concept that organisms share a common origin and have diverged through time. The details and timing of these divergences—that is, the estimation or reconstruction of an evolutionary history—are important for both intellectual and practical reasons, and phylogenies are central to virtually all comparisons among species. From a practical standpoint, phylogenetics has helped to trace routes of infectious disease transmission (e.g., dental transmission of AIDS/HIV) and to identify new pathogens such as the New Mexico hantavirus. Moret (footnote 27) notes that phylogenetic analysis is useful in elucidating functional relationships within living cells, making functional predictions from sequence data banks of gene families, predicting ligands, developing vaccines, antimicrobials, and herbicides, and inferring secondary structure of RNAs. A clear picture of how life evolved from its humble origins to its present diversity would answer the age-old question, Where do we come from?

There are many interesting phylogenetic problems. For example, consider the problem of estimating large phylogenies, which is a central challenge in evolutionary biology. Given three species, there are only three possible trees that could represent their phylogenetic history: (A,(B,C)); (B,(A,C)); and (C,(A,B)). (The notation (A,(B,C)) means that B and C share a common ancestor, who itself shares a different common ancestor with A. Thus, even if one picks a tree at random, there is a one in three chance that the tree chosen will be correct. But the number of possible trees grows very rapidly with the number of species involved. For a "small" phylogenetic problem involving 10 species, there are 34,459,425 possible trees. For a problem involving 22 species, the number of trees exceeds 10^{23}. Today, most phylogenetic problems involve more than 80 species and some data sets contain more than 500 species. (For 500 species, there are approximately 1.0085×10^{1280} possible trees, only one of which can be correct.) Of course, the grandest of all challenges in this area is the construction of the entire phylogeny of all organisms on the planet—the complete "Tree of Life" involving some 10^7 to 10^8 species.

Given the existence of such large state spaces, it is clear that exhaustive search for the single correct phylogenetic tree is not a feasible strategy, regardless of how fast computers become in the foreseeable future. Researchers have developed a number of methods for coping with the size of these problems, but many of these methods have serious deficiencies. For example, the optimality criteria used by these methods often have dubious statistical justifications. In addition, many of these methods are simply stepwise addition algorithms and make no effort to explore the space of trees. Methods with the best statistical justification, such as maximum likelihood and Bayesian inference, are also the most difficult to implement for large problems.

Thus, the algorithmics of evolutionary biology are a fertile area for research. Moret (footnote 27) notes that reconstruction of the Tree of Life will require either the scaling-up of existing reconstruction methods or the development of entirely new ones. He notes that sequence-based reconstruction methodologies are available that are likely to scale effectively from 15,000 to 100,000 taxa, but that these methodologies are not likely to scale to millions of taxa. Moret also points out that the use of gene-order data (i.e., lists of genes in the order in which they occur along one or more chromosomes) can circumvent many of the difficulties associated with using sequence data. On the other hand, there are relatively

[26]M. Kimura, "Evolutionary Rate at the Molecular Level," *Nature* 217(129):624-626, 1968; Motoo Kimura, *The Neutral Theory of Molecular Evolution*, Cambridge University Press, Cambridge, MA, 1983.

[27]B.M.E. Moret, "Computational Challenges from the Tree of Life," *Proceedings of the 7th Workshop on Algorithm Engineering and Experiments*, ALENEX '05, Vancouver, SIAM Press, Philadelphia, PA, 2005. This paper presents a number of computational challenges in evolutionary biology, of which only a few are mentioned in the subsequent discussion in this section.

few whole-genome data today, few models for the evolution of gene content and gene order, and a far greater complexity of the mathematics for gene orders compared to that for DNA sequences.

A related problem is that of comparing one or more features across species. The comparative method has provided much of the evidence for natural selection and is probably the most widely used statistical method in evolutionary biology. But comparative analyses must account for phylogenetic history, since the similarity in features common to multiple species that originate in a common evolutionary history can inappropriately and seriously bias the analyses. A number of methods have been developed to accommodate phylogenies in comparative analyses, but most of these methods assume that the phylogeny is known without error. However, this is patently unrealistic, because almost all phylogenies have a large degree of uncertainty. An important question is therefore to understand how comparative analyses can be performed that accommodate phylogenetic history without depending on any single phylogeny being correct.

Still another interesting problem concerns the genetics of adaptation—the genomic changes that occur when an organism adapts to a new set of selection pressures in a new environment. Because the process of adaptive change is difficult to study directly, there are many important and unanswered questions regarding the genetics of adaptation. For example, how many mutations are involved in a given adaptive change? Does this figure change when different organisms or different environments are involved? What is the distribution of fitness effects implied by these genetic changes during a bout of adaptation? How and to what extent are adaptations constrained by phylogenetic history? To what extent are specific genetic changes inevitable given a change of selection pressures?

9.6 COMPUTATIONAL ECOLOGY[28]

The long-term scientific goal of computational ecology is the development of methods to predict the response of ecosystems to changes in their physical, biological, and chemical components. Computational ecology seeks to combine realistic models of ecological systems with the often large datasets available to aid in analyzing these systems, utilizing techniques of modern computational science to manage the data, visualize model behavior, and statistically examine the complex dynamics that arise.[29] Questions raised immediately by computational ecology have a direct bearing on issues of important policy significance today—potential losses of biodiversity, achievement of sustainable futures, and impact of global change on local communities.[30]

The scientific questions to be addressed by computational ecology have both theoretical and applied significance. These questions include the following:[31]

- How are communities organized in space and time?
- What factors maintain or reduce biodiversity?
- What are the implications for ecosystem function?
- How should biodiversity be measured?
- How is ecological robustness maintained?

Consider, for example, ecological robustness. In ecological communities, many of the salient features remain unchanged, despite the fact that the identities of the relevant actors are continually in flux.

---

[28]Much of the discussion in this section is based on J. Helly, T. Case, F. Davis, S. Levin, and W. Michener, eds., *The State of Computational Ecology*, National Center for Ecological Analysis and Synthesis, Santa Barbara, CA, 1995, available at http://www.sdsc.edu/compeco_workshop/report/report.html.

[29]J. Helly et al., eds., *The State of Computational Ecology*, National Center for Ecological Analysis and Synthesis, Santa Barbara, CA, 1995, available at http://www.sdsc.edu/compeco_workshop/report/report.html.

[30]J. Lubchenco et al., "The Sustainable Biosphere Initiative: An Ecological Research Agenda," *Ecology* 72(2):371-412, 1991.

[31]Much of this list is taken from Helly et al., *The State of Computational Ecology*, 1995.

Species richness, species abundance relations, and biogeochemical cycles exhibit remarkable regularity, despite changes at lower levels of organization. In marine systems, the Redfield ratios,[32] which characterize the mean stoichiometry of plankton and of the water column, summarize the great constancy seen in the concentration ratios of carbon, nitrogen, and phosphorus relative to each other, although absolute levels vary considerably across the oceans. Similarly, Sheldon et al.[33] observed that the size spectrum, from the smallest particles to large fish, follows a power law with a characteristic exponent, valid across a range of trophic levels.

Ecosystems and the biosphere are complex adaptive systems,[34] in which macroscopic patterns emerge from interactions at lower levels of organization and feed back to influence dynamics on those scales. Although macroscopic investigations, such as those of Carlson and Doyle,[35] can shed considerable light on designed or managed systems, or on organ systems that have been the direct products of evolution, they provide at best a benchmark for comparisons for complex adaptive systems in which selection acts well below the level of the whole system.

The robustness of complex adaptive systems is dependent upon the same suite of characteristics that govern the robustness of any system—heterogeneity and diversity, redundancy and degeneracy, modularity, and the tightness of feedback loops. Heterogeneity, for example, provides the adaptive capacity that allows a system to persist in a changing environment; indeed, the robustness of the macroscopic features of such systems may arise despite, in fact even because of, the lack of robustness of their components. Yet these systems are neither designed nor selected for their macroscopic features. How different then are such systems from those in which the level of selection is the whole system? Should robustness be expected to emerge from the bottom up, and how does this self-organized robustness differ from what would be optimal for the robustness of systems as a whole?

Given that selection is most effective at much lower levels of organization, it is unclear what sustains ecological robustness at the macroscopic level. A key problem is to understand the properties of such self-organized, complex adaptive systems—to develop theories that facilitate scaling from individuals to whole systems and relating structure to function in order to identify signals warning of collapse. What are the consequences of the erosion of biodiversity, the homogenization of systems, and the breakdown of ecological barriers? How, indeed, will such changes affect the spread of disturbances, from forest fires to novel infectious diseases? Addressing these questions will require iterative integration of computational approaches with explorations into large-scale stochastic and distributed dynamical systems, with the goal of developing more parsimonious descriptors of essential aspects.

General theory concerning the robustness of complex systems focuses on a few key features: heterogeneity and diversity, redundancy and degeneracy, modularity, and the tightness of feedback loops.[36] Robustness is a design objective for most engineering applications, and investigations such as those of Carlson and Doyle have demonstrated how one might select on complex systems as a whole to achieve tolerance to particular classes of perturbations. One general principle that emerges from such studies is that there are trade-offs between robustness on diverse scales. Systems in general may be characterized as "robust, yet fragile." That is, their robustness to one class of perturbations, or on one scale, may

---

[32]A.C. Redfield, "On the Proportions of Organic Derivatives in Sea Water and Their Relation to the Composition of Plankton," pp. 176-192 in *James Johnstone Memorial Volume*, R.J. Daniel, ed., University Press of Liverpool, Liverpool, UK, 1934.

[33]R.W. Sheldon and T.R. Parsons, "A Continuous Size Spectrum for Particulate Matter in the Sea," *Journal of the Fisheries Research Board of Canada* 24:909-915, 1967; R.W. Sheldon, A. Prakash, and W.H. Sutcliffe, Jr., "The Size Distribution of Particles in the Ocean," *Limnological Oceanography* 17:327-340, 1972.

[34]S.A. Levin, *Fragile Dominion: Complexity and the Commons*, Perseus Books, Reading, MA, 1999; S.A. Levin, "Complex Adaptive System: Exploring the Known, the Unknown and the Unknowable," *Bulletin of the American Mathematical Society* 40:3-19, 2003.

[35]J.M. Carlson and J. Doyle, "Highly Optimized Tolerance: Robustness and Design in Complex Systems," *Physical Review Letters* 84(11):2529-2532, 2000.

[36]S.A. Levin, *Fragile Dominion: Complexity and the Commons*, Perseus Books, Reading, MA, 1999; S.A. Levin, "Complex Adaptive Systems; Exploring the Known, the Unknown and the Unknowable," *Bulletin of the American Mathematical Society* 40:3-19, 2003.

necessarily lead to fragility to other classes of perturbations, or on other scales. Understanding such trade-offs is one dimension of considerable intellectual challenge and problem richness.

These general points are instantiated in many different problem areas. Two illustrative areas—each important in its own right—include the dynamics of infectious diseases and the dynamics of marine microbial systems. In the first case, increased computational resources have fostered the development of models that relate individual behaviors to the spread of novel diseases, including smallpox and new strains and subtypes of influenza. Such models have been given added stimulus by concerns about the introduction and spread of infectious agents as weapons of bioterror, but the potential for new pandemics of influenza and other infectious diseases is probably a greater motivation for their development.

Marine microbial systems represent a vast and important storehouse of biodiversity, about which much too little is known. Recent efforts, stimulated by the success of genomics, have directed attention to characterizing the massive genetic diversity found in these systems. The computational challenges are substantial, even to catalog the vast array of data being collected. Yet just as sequencing efforts in genomics have highlighted the importance of knowing what the catalog of genetic detail reveals about how systems function in their ecological environments, the mass of accumulating information about marine microbial diversity spurs efforts at understanding how those marine ecosystems are organized and what maintains the robustness of features such as microbial diversity.

To address the scientific questions described above, researchers need techniques for dealing with systems across scales of space, time, and organizational complexity. Ultimately, an essential enabling tool will be a statistical mechanics of heterogeneous and nonindependent entities, in which the components of a system of interest are continually changing through processes of mutation and other forms of change.[37] Such a system differs dramatically from systems that have traditionally been analyzed through the machinery of traditional statistical mechanics (e.g., systems composed of identical, independently moving particles), and analytical methods for dealing with heterogeneous, nonindependent entities are generally very sophisticated. In general, such methods rely on the ability to capture the heterogeneity of the distribution (e.g., of traits) in terms of a small number of moments or other descriptors or rely on "equation-free" approaches[38] that finesse the need for explicit closures. In the absence of such an analytical characterization, computation is generally the only alternative to gaining insights about ensemble behavior, although computation may often provide analytical insights (and vice versa).

Today, computational ecology makes use of continuum and individual descriptions. Continuum modeling focuses on the impact on local ecological communities of large-scale (global) influences such as climate and fluxes of key elements such as carbon and nitrogen. These models are typically characterized by parameterized partial differential equations that represent appropriately averaged continuum quantities of ecological significance (e.g., density of a species). A central intellectual challenge of the top-down approach is reconciling the hundred-kilometer resolution of models that predict global climate change and elemental fluxes with the meter and centimeter scales of interest in natural and managed ecosystems.

The ab initio formulation of realistic continuum models is difficult, because the details of the underlying populations and entities matter a great deal. For example, naïve assumptions of independence, random motion, zero mixing time, or infinite propagation speed, which are often used in the ab initio formulation of continuum models, simply do not hold at the underlying individual level.[39] Accordingly, great care must be taken to derive a continuum description from knowledge of the individual elements in play.

---

[37]S. Levin, *Mathematics and Biology: The Interface,* Lawrence Berkeley Laboratory Pub-701, Berkeley, CA, 1992, available at http://www.bio.vu.nl/nvtb/Interface.html.

[38]C. Theodoropolous, Y. Quan, and I.G. Kevrekidis, "Coarse Stability and Bifurcation Analysis Using Time-Steppers: A Reaction-Diffusion Example," *Proceedings of the National Academy of Sciences* 97(18):9840-9843, 2000.

[39]S.A. Levin, "Complex Adaptive Systems: Exploring the Known, the Unknown and the Unknowable," *Bulletin (New Series) of the American Mathematical Society* 40(1):3-19, 2002.

Individual-based modeling seeks to extrapolate from the level of effects on individual plants and animals to changes in community-level patterns, which are necessarily characterized by longer time scales and broader space scales than those of individuals. Individual-based models, an ecological form of agent-based models, are rule-based approaches that can track the growth, movement, and reproduction of many thousands of individuals across the landscape[40] and, in looking at the global consequences of local interactions of individuals, are particularly well suited to address questions that relate to spatial heterogeneities (e.g., ecological sanctuaries).

In individual-based models, the inherent parallelism of ecological systems—that organisms interact concurrently across space—is manifest.[41] (By contrast, the parallelism in many computational models of other biological systems such as genomes and proteins is primarily a speedup mechanism for computation-intensive problems.) Individual-based models have been used to represent populations of predators, trees, and endangered species, and they are very useful in understanding the detailed response of the population of interest to alternative environmental circumstances.

In general, individual-based models are powerful tools for investigating systems that are analytically intractable, and they provide opportunities for the consideration of various scenarios and for exploring ecosystem management protocols that would not otherwise be possible. Nevertheless, such simulations often contain too many degrees of freedom to allow robust prediction. Thus, efforts to develop macroscopic representations that reduce dimensionality and that suppress irrelevant detail are essential—a point that reinforces the desirability of developing an appropriate statistical mechanics as described above.

Individual-based modeling is generally computation-intensive, for two reasons. The first is that a multitude of individuals must be represented, the behavior of each must be computed, and the entire ecosystem being modeled must be time-stepped at appropriately fine intervals. The second is that realism demands a certain amount of stochasticity; thus, an ensemble of simulations must be run in order to understand how changes in environmental and other parameters affect predicted outcomes. Grid implementations, taking advantage of the inherent parallelism of ecosystems, are one recent effort to advance individual-based modeling. The development of algorithms implementing parallelization for individual-based ecological models has enabled a number of simulations, including simulations for fish populations in the Everglades[42] and for more general models aimed ultimately at resource management.[43]

Data issues in computational ecology are also critical. Information technology has been a key enabler for a great deal of ecological data. For example, high-resolution multispectral images captured by satellites provide a wealth of information about ecosystems, resulting in maps that can depict how ecologically significant quantities can vary across large areas. While such images cannot yield significant information on the behavior of individuals, modern telemetry can be used to follow the movements of many individual organisms, a method applied routinely for certain endangered and threatened species.

At the same time, much remains to be done. Ground-based sensors take data only in their immediate locality. Thus, the spatial resolution provided by such sensors is a direct function of their areal density. Therefore, the advent of inexpensive networked sensors, described in Chapter 7, is potentially the harbinger of a new explosion of ecological data. For example, a survey of thirty papers chosen randomly from the journal *Ecology* illustrates that most ecological sampling is conducted with measurements being taken in small areas or at low frequency (often including one-time sampling).[44] Wireless

---

[40]See D.L. DeAngelis and L.J. Gross, eds., *Individual-Based Models and Approaches in Ecology*, Routledge, Chapman and Hall, New York, 1992.

[41]J. Haefner, "Parallel Computers and Individual-Based Models: An Overview," pp. 126-164 in D. DeAngelis and L. Gross, eds., *Individual-Based Models and Approaches in Ecology*, Chapman and Hall, New York, 1992.

[42]D. Wang, M.W. Berry, E.A. Carr, and L.J. Gross, "A Parallel Landscape Model for Fish as Part of a Multi-Scale Ecological System," available at http://www.tiem.utk.edu/gem/papers/dalipaper.pdf.

[43]D. Wang, E.A. Carr, M.R. Palmer, M.W. Berry, and L.J. Gross, "A Grid Service Module for Natural-Resource Managers," *IEEE Internet Computing* 9(1):35-41, 2005, available at http://www.tiem.utk.edu/gem/papers/gridservice.pdf.

[44]J. Porter et al., "Wireless Sensor Networks for Ecology," *Biosciences*, 2005, in press.

sensor networks can fill a gap in our current capabilities by enabling researchers to sample at finer spatial scales or faster rates not currently possible. It is this range of space-time (widely distributed spatial sensing with high temporal frequency) that will be critical to address the grand challenges of the environmental sciences (biogeochemical cycles, biological diversity and ecosystem functioning, climate variability, hydrologic forecasting, infectious disease and the environment, institutions and resource use, land-use dynamics, reinventing the use of materials) proposed by the National Research Council.[45] Similarly, an explosion of data and of information will arise from sensors carried by individual animals. The extent of information potentially provided by continuous monitoring of position and physiological data, compared to tags and radio collars, is obvious.

Note also an important synergy between modeling and the use of sensor networks. The effective use of sensor networks relies on modeling and analytical work to guide the placement of sensors. In turn, sensor data provide data to models that allow for prediction and interpretation of models, to understand the underlying processes. In this sense, models are the basis for an adaptive sampling scheme for sensor use.

Another data issue is progress in capturing specimen data in electronic form. Over the years, hundreds of millions of specimens have been recorded in museum records. While the information in extant collections could provide numerous opportunities for modeling and increased understanding, very few records are in electronic form and even fewer have been geocoded. Museum records carry a wealth of image and text data, and digitizing these records in a meaningful and useful way remains a serious challenge, in terms of both appropriate technical methods and the practical effort and resources required.

9.7 GENOME-ENABLED INDIVIDUALIZED MEDICINE

By many accounts, knowledge of the sequence of the human genome has enormous potential for changing the practice of medicine and the delivery of health care services. As more is understood about human biology, it is increasingly feasible for medicine to be predictive—to have advance knowledge of how a person's health status will respond (positively or negatively) to various exposures to different foods and environmental events, and to prevent disease and sustain lifelong health and well-being. Both these goals depend on a personalized medicine that begins with deep knowledge of the implications of the genetic makeup of any given individual, as well as his or her health and medical life history. Indeed, one of the most important implications of knowledge of the genome is the possibility that medical treatment and interventions might be more customized to the genetic profile of individuals or groups in ways that maximize the likelihood of successful outcomes.[46]

One necessary precondition for genome-based individualized medicine is technology for the inexpensive acquisition of sequence information—perhaps a few hundred dollars for an individual's complete genome, for example.[47] On the other hand, from a cost-effectiveness standpoint, it is better to stratify individuals into subcategories that are relevant to various treatment or intervention regimes by looking at a limited number of genetic markers, rather than to acquire the complete genetic sequence of all individuals involved. This vision has led major pharmaceutical companies to proclaim that genomic

[45]National Research Council, *Grand Challenges in Environmental Sciences*, National Academy Press, Washington, DC, 2001.

[46]One of the most ambitious efforts to exploit the potential of genome-enabled individualized medicine is being undertaken by Mexico, whose population is composed of more than 65 native Indian groups and Spaniards. Because the overall genetic makeup of this population is associated with a characteristic set of disease susceptibilities, Mexico has undertaken this initiative to reduce the social and financial burden of health problems, since new strategies for prevention, early diagnosis, and more effective treatment are essential to meet the mid- and long-term health care goals in Mexico. See Gerardo Jimenez-Sanchez, "Developing a Platform for Genomic Medicine in Mexico," *Science* 300:295-296, 2003.

[47]Note that this is 10^5 times less expensive than the sequencing of the first genome. Whether the least expensive approach turns out to be sequencing individual genomes from scratch, or sequencing only those portions specific to individuals and integrating those portions into the genome of the generic human, remains to be seen.

medicine and related technologies will allow physicians to provide the right drug to the right patient at the right time. Thus, the term "individualized medicine" should be regarded as one that ranges from single individuals (likely in the farther-term future) to genetically differentiated subpopulations (more likely to happen in the near term).

The fundamental challenge is to correlate genetic variation to susceptibility for specific diseases, specific drug reactions, and specific responses to environmental insult. But even with these correlations in hand, it is a very long way from examination of individual drug-gene interactions to individualized medicine—what might be called translational medicine—that affects the well-being of the citizenry at large. Traversing this distance will require considerable advances on multiple fronts: in the laboratory, on the computer, and in how scientists conceptualize the relationships between all of the individual components involved.

9.7.1 Disease Susceptibility[48]

It has been known for many years that many medical conditions have a genetic basis. Indeed, for many illnesses, the strongest predictor of risk is an individual's family history. The association of specific genomic differences with the likelihood of disease will provide physicians and patients with more specific and more certain information. Such knowledge will allow individuals to takes steps that reduce the likelihood and/or severity of such disease in the future. These steps might include greater medical surveillance or screening, environmental changes, diet, exercise, or preventive drug therapy (e.g., more frequent colonoscopies starting earlier in life for individuals with genetic profiles that imply a high degree of risk for colon cancer).

It is useful to distinguish between genetic signatures that are highly penetrant and those that are highly prevalent. A highly penetrant genetic signature associated with a disease is one whose presence implies a high likelihood that the disease will develop in an individual with that signature: examples provided by Guttmacher and Collins (Footnote 48) include mutations in the BRCA1 and BRCA2 genes that increase the risk of breast and ovarian cancer, in the HNPCC gene set that increases the risk of hereditary nonpolyposis colorectal cancer, and in the gene for synuclein that causes Parkinson's disease. A highly prevalent genetic signature is one that occurs frequently in the population, but its presence may or may not be associated with a large increase in the likelihood that a disease will develop in an individual with that signature: as examples, Guttmacher and Collins (Footnote 48) include a mutation in the factor V Leiden gene that increases the risk of thrombosis, in the APC (adenomatosis polyposis coli) gene that increases the risk of colorectal cancer, and in the apolipoprotein gene that increases the risk of Alzheimer's disease.

From the standpoint of the individual, identification of a *highly penetrant* genetic signature associated with disease will have important clinical ramifications. However, from a public health standpoint, it is the identification of *highly prevalent* genetic signatures associated with disease that is most significant.

The best-understood genetic disorders leading to disease are those associated with the inheritance of a single gene. Such disease conditions have been cataloged in the Online Mendelian Inheritance in Man (OMIM) catalog.[49] Examples of single-gene conditions cited by Guttmacher and Collins include hereditary hemochromatosis, cystic fibrosis, alpha$_1$-antitrypsin deficiency, and neurofibromatosis. These

[48]The discussion in this section on monogenic and highly penetrant signatures is based on excerpts from A.E. Guttmacher and F.S. Collins, "Genomic Medicine—A Primer," *New England Journal of Medicine* 347(19):1512-1520, 2002. The discussion in this section on polygenic and highly prevalent signatures is based on excerpts from P.D. Pharoah, A. Antoniou, M. Bobrow, R.L. Zimmern, D.F. Easton, and B.A. Ponder, "Polygenic Susceptibility to Breast Cancer and Implications for Prevention," *Nature Genetics* 31(1):33-36, 2002. A highly positive and optimistic view of the impact of the genome on medicine can be found in F.S. Collins and V.A. McKusick, "Implications of the Human Genome Project for Medical Science," *Journal of the American Medical Association* 285(5):540-544, 2001. A somewhat contrary view can be found in N.A. Holtzman and T.M. Marteau, "Will Genetics Revolutionize Medicine?" *New England Journal of Medicine* 343(2):141-144, 2000.

[49]See http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?db=OMIM.

disorders are highly penetrant, but occur relatively rarely in the population (with approximate incidences of one in several hundred or less).

On the other hand, multifactor genetic causality for disease is almost certainly much more common than monogenic causality. In principle, knowledge of the range of genetic variations across many loci in the population will allow researchers to estimate risks arising from the combined effect of such variations.

Using breast cancer as a case study, Pharoah et al. (footnote 48) compared the potential for prediction of risk based on common genetic variations with the predictions that could be made using known and established risk factors. They concluded that a typical polygenic approach for analysis would suggest that the half of the population at highest risk would account for 88 percent of all affected individuals, if all of the susceptibility genes could be identified. However, using currently known factors for breast cancer to stratify the population, they estimated that the half of the population at highest risk would account for only 62 percent of all cases. Pharoah et al. thus suggest that genetic profiles may provide significant improvement in the ability to differentiate at-risk individuals from individuals not at risk.

Nevertheless, for a variety of reasons, identifying the relevant genetic signatures over multiple genes that account for disease susceptibility will pose significant intellectual challenges. Probably the most important point is that the contribution of any given gene involved is likely to be weak; hence detecting its clinical significance may be problematic. Nongenomic effects, such as posttranslational modifications, may also be relevant. Zimmern[50] notes that even monogenic conditions can result in variable expressivity and incomplete penetrance, and that similar disease phenotypes may result from genetic heterogeneity, whether in the form of allelic heterogeneity (different mutations at the same locus) or locus heterogeneity (where mutations occur at different loci). Different mutations of the same gene may also give rise to separate clinical effects. Environmental factors may be difficult to disentangle from genetic ones. As a consequence of such issues, definitive conclusions about the relationship of a given polygenic genotype to a specific disease condition may well be difficult to draw.

An extension of the genomic approach to disease susceptibility applies to understanding the impact of an individual's genomic composition on that individual's response to various environmental insults to the body, such as those caused by exposure to chemicals (e.g., from drinking water or air pollution) or electromagnetic fields (e.g., from cell phones or ambient radiation). Furthermore, in dealing with certain environmental insults, stochasticity is likely to play an important role. For example, in considering the effects of radiation on the genome, macroscopic parameters that characterize radiation such as duration and intensity are insufficient to determine its effect, simply because what part of a genome is affected is mostly a matter of chance. Thus, a given dose of a certain kind of radiation will not affect individuals in equal measure and, more to the point, could not be expected to affect even an ensemble of identical twins similarly.

Overall, there is wide variability in individual responses to environmental influences. While existing diseases, differences in gender, or differences in nutritional status affect such variability, genetic influences are also important. Genes that affect the human response to environmental exposure (called environmentally responsive genes by the Environmental Genome Project [EGP] of the National Institute of Environmental Health Sciences [NIEHS] tend to fall into several categories.[51] That is, they affect the cell cycle, DNA repair, cell division, cell signaling, cell structure, gene expression, apoptosis, and metabolism. The initial phases of the EGP are focused on identifying single nucleotide polymorphisms (SNPs) associated with 554 genes identified by the scientific community as environmentally responsive. Identification of the SNPs associated with environmentally responsive genes would make it possible to conduct epidemiological studies that classify subjects by SNPs, thus increasing the utility of these

[50]R.L. Zimmern, "The Human Genome Project: A False Dawn?" *British Medical Journal* 319(7220):1282, 1999.

[51]See http://www.niehs.nih.gov/envgenom/egp.htm.

studies in detecting genetic contributions to the likelihood of various diseases with at least partial environmental causation.

The challenges of polygenic data analysis are formidable. An example of methodological research in this area is that of Nelson et al.,[52] who developed the combinatorial partitioning method (CPM) for examining multiple genes, each containing multiple variable loci, to identify partitions of multilocus genotypes that predict interindividual variation in quantitative trait levels. The CPM offers a strategy for exploring the high-dimensional genotype state space so as to predict the quantitative trait variation in the population at large that does not require the conditioning of the analysis on a prespecified genetic model, such as a model that assumes that interacting loci can each be identified through their independent, marginal contribution to trait variability. On the other hand, a brute-force approach to this correlation problem explodes combinatorially. Therefore, it is likely that finding significant correlations will depend on the ability to prune the search space before specific combinations are tested—and the ability to prune will depend on the availability of insight into biological mechanisms.

9.7.2 Drug Response and Pharmacogenomics[53]

As with disease susceptibility, it has been known for many years that different individuals respond differently to the same drug at the same dosages and that the relevant differences in individuals are at least partly genetic in origin. However, characterization of the first human gene containing DNA sequence variations that influence drug metabolism did not take place until the late 1980s.[54] Today, pharmacogenomics—the impact of an individual's genomic composition on his or her response to various drugs—is an active area of investigation that many believe holds significant promise for changing the practice of medicine by enabling individual-based prescriptions for compound and dosage.[55] An individual's genetic profile may well suggest which of several drugs is most appropriate for a given disease condition. Because genetics influence drug metabolism, an individual's weight will no longer be the determining factor in setting the optimal dosage for that individual.

Similarly, many drugs are known to be effective in treating specific disease conditions. However, because of their side effects in certain subpopulations, they are not available for general use. Detailed "omic" knowledge about individuals may help to identify the set of people who might benefit from certain drugs without incurring undesirable side effects, although some degree of empirical testing will be needed if such individuals can be identified.[56] In addition, some individuals may be more sensitive than others to specific drugs, requiring differential dosages for optimal effect.

As in the case of disease susceptibility, the best-understood genetic polymorphisms that affect drug responses in individuals are those that involve single genes. As an example, Evans and Relling note that

---

[52]M.R. Nelson, S.L.R. Kardia, R.E. Ferrell, and C.F. Sing, "A Combinatorial Partitioning Method to Identify Multilocus Genotypic Partitions That Predict Quantitative Trait Variation," *Genome Research* 11(3):458-470, 2001.

[53]Much of the discussion in Section 9.7.2 is based on excerpts from W.E. Evans and M.V. Relling, "Moving Towards Individualized Medicine with Pharmacogenomics," *Nature* 429(6990):464-468, 2004.

[54]F.J. Gonzalez, R.C. Dkoda, S. Kimura, M. Umeno, U.M. Zanger, D.W. Nebert, H.V. Gelboin, et al., "Characterization of the Common Genetic Defect in Humans Deficient in Debrisoquine Metabolism," *Nature* 331(6155):442-446, 1988. Cited in Evans and Relling, 2004.

[55]If the promise of pharmacogenomics is realized, a number of important collateral benefits follow as well. Drug compounds that have previously been rejected by regulatory authorities because of their side effects on some part of the general population at large may become available to those individuals genomically identified as not being subject to those side effects. Thus, these individuals would have options for treatment that would not otherwise exist. Furthermore, clinical trials for drug testing could be much more targeted to appropriate subpopulations with a higher likelihood of ultimate success, thus reducing expenses associated with failed trials. Also, in the longer term, pharmacogenomics may enable the customized creation of more powerful medicines based on the specific proteins and enzymes associated with genes and diseases.

[56]Another application often discussed in this context is the notion of drugs customized to specific individuals based on "omic" data. However, the business model of pharmaceutical companies today is based on large markets for their products. Until it becomes possible to synthesize and manufacture different drug compounds economically in small quantity, custom-synthesized drugs for small groups of individuals will not be feasible.

individuals that are deficient in thiopurine *S*-methyltransferase (TPMT) can be treated with much lower doses of the thiopurine drugs mercaptopurine and azathiopurine used as immunosuppressants and to treat neoplasias. There is a clinical diagnostic test available for the genomic detection of the TPMT deficiency, but routine use of TPMT genotyping to make treatment decisions is limited. A second example also discussed by Evans and Relling is that polymorphisms in a gene known as CYP2D6 have a strong effect on individuals' responses to the antihypertensive drug debrisoquine and in the metabolism of the oxytocic drug sparteine.

A second example is found in the area of certain drugs for the treatment of cardiovascular disease. Numerous examples of differences among individuals have been seen as potential candidate pharmacodynamic loci (e.g., those for angiotensinogen, angiotensin-converting enzyme, and the angiotensin II receptor). Polymorphisms at these loci predict responses to specific treatments such as the inhibition of angiotensin-converting enzyme. Here, researchers hope to establish and utilize antihypertensive drugs that are matched to the genetic variations among individuals, and thus to optimize blood pressure control and reduce side effects.[57]

A number of monogenic polymorphisms have been found, encoding drug-metabolizing enzymes, drug transporters, and drug targets, as well as disease-modifying genes, that have been linked to drug effects in humans. However, these are the "low-hanging fruit" of pharmacogenetics, and for most drug effects and treatment outcomes, monogenic polymorphisms with clearly recognizable drug-response phenotypes do not characterize the situation. For example, as in the case of disease susceptibility, nongenomic effects (e.g., posttranslational modifications) on protein function may be relevant. Or, multiple genes may act together in networks to create a single drug-response phenotype.

As Evans and Relling note, genome-wide approaches, such as gene expression arrays, genome-wide scans, or proteomic assays, can contribute to the identification of as-yet-unrecognized candidate genes that may have an influence on a drug response phenotype. For example, it may be possible to detect genes whose expression differentiates drug responders from nonresponders (or those for whom certain drugs are toxic from those for whom they are not), genomic regions with a paucity of heterozygosity in responders compared with nonresponders, or proteins whose abundance differentiates drug responders from nonresponders.

In expression-array and proteomic approaches, the level of the signal may directly reflect functional variation—a distinct advantage from an experimental point of view. Yet there can be many other reasons for differences in signal level, such as the choice of tissue from which the samples are drawn (which may not be the tissue of interest where toxicity or response is concerned) or changes in function not reflected by levels of mRNA or protein. Thus, when such studies suggest that a given gene or gene product is relevant to drug response, Evans and Redding point out that large-scale molecular epidemiological association studies (in vivo or in vitro with human tissues), biochemical functional studies, and studies on preclinical animal models of candidate gene polymorphisms become necessary to further establish the link between genetic polymorphism and drug response.

A second challenge in pharmacogenomics relates to integrating pharmacogenomics with the everyday practice of medicine. Although there are cultural and historical sources of resistance to such integration, it is also true that definitive clinical pharmacogenomic studies have not been conducted that demonstrate unambiguously the benefits of integration on clinical outcomes. Indeed, there are many difficulties in conducting such studies, including the multigenic nature of most drug effects and the difficulty in controlling for nongenetic confounders such as diet or exercise. Until such difficulties are overcome, it is unlikely that a significant change will occur in clinical practice.

One of the most important databases for the study of pharmacogenomics is a database known as the Stanford PharmGKB, described in Box 3.4. Supported by the National Institute of General Medical

[57]P. Cadman and D. O'Connor, "Pharmacogenomics of Hypertension," *Current Opinion in Nephrology and Hypertension* 12(1):61-70, 2003.

Science (NIGMS), PharmGKB is a publicly available, Internet-accessible database for pharmacogenetics and pharmacogenomics. Its overall aim is to aid researchers in understanding how genetic variation among individuals contributes to differences in reactions to drugs.[58] The database integrates pharmacodynamics (drug actions), pharmacokinetics (drug metabolism), toxicity, sequence and other molecular data, pathway information, and patient data.

9.7.3 Nutritional Genomics

Traditional nutrition research has had among its goals the establishment of overarching dietary recommendations for everyone—in principle, for the world's entire population. Today, and more so in the future, the implications of individual genetic makeup for optimal diet have changed that perspective. To understand and exploit the interplay of diet and genetics, nutritional genomics is a relatively new specialization within the life sciences with two separate but related foci. One focus relates an individual's genetic makeup to dietary regimes that are more or less healthy for him or her. For example, it is well known that some individuals are more likely to suffer from high blood pressure if they consume salt in relatively large quantities, while others are not. Poch et al.[59] found a possible genetic basis on which to differentiate salt-sensitive individuals and salt-insensitive ones. If it is possible to develop genetic tests for salt sensitivity, salt-sensitive individuals could be advised specifically to limit their salt intake, and salt-insensitive individuals could continue to indulge at will their taste for salty snacks.

The traditional focus of nutrition research is not in any way rendered irrelevant by nutritional genomics. Still, beyond general good advice and informed common sense, in the most ambitious scenarios, recommended dietary profiles could be customized for individuals based on their specific genomic composition. Ordovas and Corella write:[60]

> Nutritional genomics has tremendous potential to change the future of dietary guidelines and personal recommendations. Nutrigenetics will provide the basis for personalized dietary recommendations based on the individual's genetic makeup. This approach has been used for decades for certain monogenic diseases; however, the challenge is to implement a similar concept for common multifactorial disorders and to develop tools to detect genetic predisposition and to prevent common disorders decades before their manifestation. . . . [P]reliminary evidence strongly suggests that the concept should work and that we will be able to harness the information contained in our genomes to achieve successful aging using behavioral changes; nutrition will be the cornerstone of this endeavor.

A second focus of nutritional genomics is on exploiting the potential for modifying foodstuffs to be more healthy, and so dietary advice and discipline might be supplanted *in part* by such modifications. For example, it may be possible to redesign the lipid composition of oil seed crops using genetic modification techniques (through either selective breeding or genetic engineering). However, whether this is desirable depends on how consumption of a different mix of lipids affects human health. Watkins et al.[61] argue for an understanding of the overall metabolomic expression of lipid metabolism to ensure that a particular metabolite composition truly improves overall health, so that a change in lipid composition that is deemed healthy when viewed as lowering the risk of one disease does not simultaneously increase the risk of developing another.

---

[58]T.E. Klein and R.B. Altman, "PharmGKB: The Pharmacogenetics and Pharmacogenomics Knowledge Base," *Pharmacogenomics Journal* 4(1):1, February 2004.

[59]E. Poch, D. Gonzalez, V. Giner, E. Bragulat, A. Coca, and A. de La Sierra, "Molecular Basis of Salt Sensitivity in Human Hypertension: Evaluation of Renin-Angiotensin-Aldosterone System Gene Polymorphisms," *Hypertension* 38(5):1204-1209, 2001.

[60]J.M. Ordovas and D. Corella, "Nutritional Genomics," *Annual Review of Genomics and Human Genetics* 5:71-118, 2004.

[61]S.M. Watkins, B.D. Hammock, J.W. Newman, and J.B. German, "Individual Metabolism Should Guide Agriculture Toward Foods for Improved Health and Nutrition," *American Journal of Clinical Nutrition* 74(3):283-286, 2001.

More generally, Watkins et al. point out that the goal of nutritional improvement of agriculture—to produce changes in crops and foods that provide health benefits to all—is difficult to achieve because modifications of existing foodstuffs are likely to advantage some people while disadvantaging others. Watkins et al. cite the example of recent attempts to increase the carotenoid content of the food supply—a move that was thought to have protective value against certain cancers, especially lung cancer. In the midst of this effort, it was found that high intakes of ß-carotene as a supplement actually *increased* the incidence of lung cancer in smokers—and the move was abandoned.

The intellectual underpinning of this effort is thus metabolomics, the quantitative characterization of the set of metabolites—generally small, nonprotein molecules—involved in the metabolism or a cell, tissue, or organism over its lifetime. In the context of nutritional genomics, metabolomic studies attempt to characterize the levels, activities, regulation, and interactions of all metabolites in an individual and determine how this characterization changes in response to various foods that are consumed. Genomics is important because genetic makeup is an important influence on the specific nature of the metabolomic changes that result as a function of food consumption.

9.8 A DIGITAL HUMAN ON WHICH A SURGEON CAN OPERATE VIRTUALLY

A surgical act on a human being is by definition an invasive process, one that inflicts many insults on the body. Prior to the advent of medical imaging techniques, surgeons relied on their general knowledge of anatomy to know where and what to cut. Today's imaging technologies provide the surgeon with some idea of what to expect when he or she opens the patient.

At the same time, a surgeon in the operating room has no opportunity to practice the operation on this particular patient. Experience with other patients with similar conditions helps immeasurably, of course, but it is still not uncommon even in routine surgical operations to find some unexpected problem or complication that the surgeon must manage. Fortunately, most such problems are minor and handled easily. Surgeons-in-training operate first on cadavers and move to live patients only after much practice and under close supervision.

Consider then the advantages that a surgeon might have if he or she were to be able to practice a difficult operation before doing it on a live patient. That is, a surgeon (or surgeon-in-training) would practice or train on a digital model of a human patient that incorporates static and dynamic physical properties of the body in an operating room environment (e.g., under anesthesia, in real gravity) when it is subject to surgical instruments.

In this environment, the surgeon would likely wear glasses that projected an appropriate image to his or her retina and use implements that represented real instruments (e.g., a scalpel). Kinetic parameters of the instrument (e.g., speed, velocity, orientation) would be monitored and registered onto the image that the surgeon sees. When "touched" by the instrument, the image would respond appropriately with a change in shape and connectivity (e.g., when a scalpel touches a piece of skin, it might separate into two parts and a cut would appear). Blood would emerge at realistic rates, and tissue under the skin would appear.

Even in this very simple example, many challenges can be seen. To name just a few:

- *Realistic modeling of body subsystems.* From the perspective of a surgeon's scalpel, the body is simply a heterogeneous and spatially organized mass of tissue. Of course, this mass of tissue is functionally a collection of subsystems (e.g., organs, muscle tissue, bone) that have different properties. These subsystems must be separated so that the physiological responses of surgery are appropriately propagated through them when surgery occurs.
- *Integration of person-specific information with a generic model of a human being.* Because of the labor involved in constructing a digital model of a human being, it makes sense to consider an approach in which a model of a generic human being is developed and then adjusted according to person-specific information of any given patient.

- *Spatial registration and alignment of instruments, the surgeon's hands, and the digital body being operated on.* The surgeon must see an instrument move to the position in the body to which the surgeon has moved it. When a cutting motion is made, the appropriate tissue should split in the appropriate place and amount.
- *The different feel and texture of tissue depending on whether the instrument is a scalpel or a finger.* A digital human for surgical use must provide appropriate force feedback ("haptic capability") to the surgeon so that, for example, cutting into soft tissue feels different than cutting into bone.
- *Incorporation of gravity in the model.* Many organs consist of soft tissue that is deformed easily under pressure from instruments and touch. As importantly, tissues are subject to gravitational forces that will change their shape as their orientation is changed (the breast of a woman lying on her back has an entirely different shape than when she is lying on her side).

Some first steps have been taken in many of these areas. For example, a project at the Ohio Supercomputer Center (OSC) in 1996 sought to develop a virtual reality-based simulation of regional anesthesia that employed haptic techniques to simulate the resistance felt when an injection is given in a certain area (Box 9.4).

A second example is work in computational anatomy, one application of which has sought to characterize the structure of human brains in a formal manner. Structure is interesting to neuroscientists because of a presumed link between physical brain structure and neurological function. Through mathematical transformations that can deform one structure into another, it is possible to develop metrics that can characterize how structurally different two brains are. These metrics can then be correlated with understanding of the neurological functions of which each brain is capable (Box 9.5). Such metrics can also be used to identify normal versus diseased states that are reflected anatomically.

Box 9.4
A Virtual Reality Simulation of Regional Anesthesia

A collaborative effort between researchers at the Ohio State University Hospitals, Immersion Corporation, and the Ohio Supercomputer Center has led to the creation of a virtual reality simulator that enables anesthesiologists-in-training to practice in a realistic environment the injection of a local anesthetic into the epidural space of the spinal column. The system includes a workstation capable of stereo display, a real-time spatial volume renderer, a voice-activated interface, and most importantly, a one-dimensional haptic probe capable of simulating the resistive forces of penetrated tissues.

Although this procedure appears simple, it is in fact a delicate manual operation that requires the placement of a catheter into a small epidural space using only haptic cues (i.e., cues based on tactile sensations of pressure) to guide the needle. By feeling the resistive forces of the needle passing through various tissues, the anesthesiologist must maneuver the tip of the needle into the correct space without perforating or damaging the spinal cord in the process.

The system is designed to enable the trainee to practice the procedure on a variety of datasets representative of what he or she might experience with real patients. That is, the pressure profile as a function of needle penetration would vary from patient to patient. By training in this environment, the trainee can gain proficiency in the use of this technique in a non-harmful manner.

SOURCE: L. Hiemenz, J.S. McDonald, D. Stredney, and D. Sessanna, "A Physiologically Valid Simulator for Training Residents to Perform an Epidural Block," *Proceedings of the 15th Southern Biomedical Engineering Conference*, March 29-31, 1996, Dayton, OH. See also http://www.osc.edu/research/Biomed/past_projects/anesthesia/index.shtml.

Box 9.5
Computational Anatomy

Computational anatomy seeks to make more precise the commonsense notion that samples of a given organ from a particular species are both all the same and all different. They are the same in the sense that all human brains, for example, exhibit similar anatomical characteristics and can be associated with the canonical brain of *Homo sapiens*, rather than the canonical brain of a dog. They are all different in the sense that each individual has a slightly different brain, whose precise anatomical characteristics differ somewhat from those of other individuals.

Computational anatomy is based on a mathematical formalism that allows one structure (e.g., a brain) to be deformed reversibly into another. (Reversibility is important because irreversible processes destroy information about the original structure.) In particular, the starting structure is considered to be a deformable template. The template anatomy is morphed into the target structure via transformations applied to subvolumes, contours, and surfaces. These computationally intensive transformations are governed by generalizations of the Euler equations of fluid mechanics and are required only to preserve topological relationships (i.e., to transform smoothly from one to the other).

Key to computational anatomy is the ability to calculate a measure of difference between similar structures. That is, a distance parameter should represent in a formalized manner the extent to which two structures differ—and a distance of zero should indicate that they are identical. In the approach to computational anatomy pioneered by Grenander and Miller,[1] the distance parameter is the square root of the energy required to transform the first structure onto the metric of the second with the assumption that normal transformations follow the least-energy path.

One instance in which computational anatomy has been used is in understanding the growth of brains as juveniles mature into adults. Thompson et al.[2] have applied these deformation techniques to the youngest brains, with results that accord well with what was seen in older subjects. In particular, they are able to predict the most rapid growth in the isthmus, which carries fibers to areas of the cerebral cortex that support language function. A second application has sought to compare monkey brains to human brains.

[1]U. Grenander and M.I. Miller, "Computational Anatomy: An Emerging Discipline," *Quarterly Journal of Applied Mathematics* 56:617-694, 1998.

[2]P.M. Thompson, J.N. Giedd, R.P. Woods, D. Macdonald, A.C. Evans, and A.W. Toga, "Growth Patterns in the Developing Brain Detected by Using Continuum Mechanical Tensor Maps," *Nature* 404:190-193, March 9, 2000; doi:10.1038/35004593.

SOURCE: Much of this material is adapted from "Computational Anatomy: An Emerging Discipline," *EnVision* 18(3), 2002, available at http://www.npaci.edu/envision/v18.3/anatomy.html#establishing.

9.9 COMPUTATIONAL THEORIES OF SELF-ASSEMBLY AND SELF-MODIFICATION[62]

Self-assembly is any process in which a set of components joins together to form a larger, more complex structure without centralized or manual control. For example, it includes biologically significant processes ranging from the joining of amino acids to form a protein and embryonic development to nonbiological chemical processes such as crystallization. More recently, the term has become widely used as researchers attempt to create artificial self-assembling systems as a way to fabricate structures efficiently at nanometer scale.

One kind of structure—that can be described as a simple repeating pattern in which molecules form into a regular structure or lattice—is the basis for creating artifacts such as crystals or batteries that can be extended to potentially macroscopic scale; this process is known as periodic self-assembly. However, for applications such as electronic circuits, which cannot be described as a simple repeating pattern, a

[62]Section 9.9 is based largely on material from L. Adleman, Q. Cheng, A. Goel, M.-D. Huang, D. Kempe, P. Moisset de Espanés, P. Wilhelm, and K. Rothemund, "Combinatorial Optimization Problems in Self-Assembly," STOC '02, available at http://www.usc.edu/dept/molecular-science/optimize_self_assembly.pdf.

more expressive form of self-assembly is required. Ideally, a designer could select a set of components and a set of rules by which they connect, and the system would form itself into the desired final shape.

This kind of self-assembly, called nonperiodic or programmable self-assembly, would allow the creation of arbitrary arrangements of components. Nonperiodic self-assembly would be useful for the efficient execution of tasks such as electronic circuit design, material synthesis, micro- and nanomachine construction, and many other technological feats. For the purposes of artificial self-assembly technology, the pinnacle result of a theory would be to be able to select or design an appropriate set of components and assembling rules to produce an arbitrary desired result.

Self-assembly, both as a biological process and as a potential technology, is poorly understood. A range of significant (and possibly insuperable) engineering and technological challenges stands in the way of effectively programming matter to form itself into arbitrary arrangements. A less prominent but no less important challenge is the lack of a theoretical foundation for self-assembly.

A theory of self-assembly would serve to guide researchers to determine which structures are achievable, select appropriate sets of components and assembling rules to produce desired results, and estimate the likely time and environmental conditions necessary to do so. Such a theory will almost certainly be based heavily on the theory of computation and will more likely be a large collection of theoretical results and proofs about the behavior of self-assembling systems, rather than a single unified theory such as gravity.

The grandest form of such a theory would encompass and perhaps unify a number of disparate concepts from biology, computer science, mathematics, and chemistry—such as thermodynamics, catalysis and replication, computational complexity, and tiling theory[63] and would require increases in our understanding of molecular shape, the interplay between enthalpy and entropy, and the nature of noncovalent binding forces.[64] A central caveat is that self-assembly occurs with a huge variety of mechanisms, and there is no a priori reason to believe that one theory can encompass all or most of self-assembly and also have enough detail to be helpful to researchers. In more limited contexts, however, useful theories may be easier to achieve, and more limited theories could serve in guiding researchers to determine which structures are achievable or stable, to identify and classify failure modes and malformation, or to understand the time and environmental conditions in which various self-assemblies can occur. Furthermore, theories in these limited contexts may or may not have anything to do with how real biological systems are designed.

For example, progress so far on a theory of self-assembly has drawn heavily from the theory of tilings and patterns,[65] a broad field of mathematics that ties together geometry, topology, combinatorics, and elements of group theory such as transitivity. A tiling is a way for a set of shapes to cover a plane, such as M.C. Escher's famous tesselation patterns. Self-assembly researchers have focused on nonperiodic tilings, those in which no regular pattern of tiles can occur. Most important among aperiodic patterns are Wang tiles, a set of tiles for which the act of tiling a plane was shown to be equivalent to the operation of a universal Turing machine.[66] (Because of the grounding in the theory of Wang tiles in particular, the components of self-assembled systems are often referred to as "tiles" and collections of tiles and rules for attaching them as "tiling systems.")

With a fundamental link between nonperiodic tilings and computation being established, it becomes possible to consider the possibility of programming matter to form desired shapes, just as Turing machines can be programmed to perform certain computations. Additionally, based on this relationship, computationally inspired descriptions might be sufficiently powerful to describe biological self-assembly processes.

Today, one of the most important approaches to a theory of self-assembly focuses on this abstract model of tiles, which are considered to behave in an idealized, stochastic way. Tiles of different types are present in the environment in various concentrations, and the probability of a tile of a given type

[63]L.M. Adleman, "Toward a Mathematical Theory of Self-Assembly," USC Tech Report, 2000, available at http://www.usc.edu/dept/molecular-science/papers/fp-000125-sa-tech-report-note.pdf.

[64]G. Whitesides, "Self-Assembly and Nanotechnology," *Fourth Foresight Conference on Molecular Nanotechnology*, 1995.

[65]B. Grunbaum and G.C. Shephard, *Tilings and Patterns*, W. H. Freeman and Co., New York, 1987.

[66]H. Wang, "Notes on a Class of Tiling Problems," *Fundamenta Mathematicae* 82:295-305, 1975.

attempting to connect to an established shape is proportional to its share of the total concentration of tiles. Then, the "glues" of touching sides of the adjacent tiles have a possibility of attaching. Simulating such self-assembly is actually relatively simple. Given a set of tiles and glues, a simulation can predict with arbitrary accuracy the end result. However, this is complicated by the fact that a given tiling system might not have a unique end result; situations could arise in which two different tiles join the assembly in the same location. While this may seem an undesirable situation, such ambiguous systems may be necessary to perform universal computation.[67]

The more challenging question is the converse of simulation: Given a desired result, how do we get there? Research into the theory of self-assembly has focused on two more specific framings of this question. First, what is the minimum number of tile types necessary to create a desired shape (the "Minimum Tile Set Problem") and, given a specific tiling system, what concentrations produce the end result the fastest (the "Tile Concentrations Problem")? The former has been shown to be NP-complete, but has polynomial solutions given certain restrictions on shape and temperature.

The current state of the art in the theory of self-assembly abstracts away much of the details of chemistry. First, the theory considers only the assembly of two-dimensional patterns. For artificial DNA tiles, designed to be flat, rigid, and square, this may be a reasonable approximation. For a more general theory that includes the self-assembly in three dimensions of proteins or other nonrigid and highly irregularly shaped macromolecules, it is less clear that such a theory is sufficient. Extending the current theory to irregular shapes in three dimensions is a key element of this challenge problem.

The history of the motivation of research into the theory of self-assembly provides a lesson for research at the BioComp interface. Originally, researchers pursued the link between self-assembly and computation because they envisioned self-assembled systems constructed from DNA as potential competitors to electronic digital computing hardware, that is, using biochemistry in the service of computation. However, as it became less obvious that this research would produce a competitive technology, interest has shifted to using the computational theory of self-assembly to increase the sophistication of the types of molecular constructs being created. In other words, today's goal is to use computational theory in the service of chemistry. This ebb and flow of both source theory and application between computation and biochemistry is a hallmark of a successful model of research at the interface.

Another area related to theories of self-assembly is what might be called adaptive programming. Today, most programs are static; although variables change their values, the structure of the code does not. Because computer hardware does not fundamentally differentiate between "code" and "data" (at the machine level, both are represented by 1's and 0's), there is no reason in principle that code cannot modify itself in the course of execution. Self-modifying code can be very useful in certain contexts, but its actual execution path can be difficult to predict and, thus, the results that might be obtained from program execution are uncertain.

However, biological organisms are known to learn and adapt to their environments—that is, they self-modify under certain circumstances. Such self-modification occurs at the genomic level, where the DNA responsible for the creation of cellular proteins contains both genetic coding and regions that regulate the extent to which, and the circumstances under which, genes are activated. It also occurs at the neural level, where cognitive changes (e.g., a memory or a physical skill) are reflected in reorganized neural patterns. Thus, a deep understanding of how biology organizes self-modification in using DNA or in a neural brain may lead to insights about how one might approach human problems that call for self-modifying computer programs.

9.10 A THEORY OF BIOLOGICAL INFORMATION AND COMPLEXITY

Much of this report is premised on the notion of biology as an information science and has argued that information technology is essential for acquiring, managing, and analyzing the many types of

[67]P.W. Rothemund, "Using Lateral Capillary Forces to Compute by Self-Assembly," *Proceedings of the National Academy of Sciences* 97(3):984-989, 2000.

biological data. This fact—that an understanding of biological systems depends on so many different kinds of biological data, operating at so many different scales, and in such volume—suggests the possibility that biological information and/or biological complexity might be notions with some formal quantitative meaning.

How much information does a given biological system have? How should biological complexity be conceptualized? Can we quantify or measure the amount of information or the degree of complexity resident in, say, a cell, or perhaps even more challengingly, in an organelle, an ecosystem, or a species? In what sense is an organism more complex than a cell or an ecosystem more complex than an individual organism? Establishing an intellectually rigorous methodology through which such information could be measured, capturing not only the raw scale of information needed to describe the constituent elements of a system but also its complexity, could be a powerful tool for answering questions about the nature of evolution, for quantifying the effects of aging and disease, and for evaluating the health of ecologies or other complex systems.

Developing such a theory of biological information and complexity will be extraordinarily challenging, however. First, complexity and information exist at a vast range of orders of magnitude in size and time, as well as in the vast range of organisms on Earth, and it is not at all clear that a single measure or approach could be appropriate for all scales or creatures. Second, progress toward such a theory has been made in fields traditionally separate from biology, including physics and computer science. Transferring knowledge and collaboration between biology and these fields is difficult at the best of times, and doubly challenging when the research is at an early stage. Finally, such a theory may prove to be the basis of a new organizing principle for biology, which may require a significant reorientation for practicing biologists and biological theory.

Some building blocks for such a theory may already be available. These include information theory, formulated by Claude Shannon in the mid-20th century for analyzing the performance of noisy communication channels; an extension of information theory, developed over the last few decades by theoretical physicists, that defines information in thermodynamic terms of energy and entropy; the body of computational complexity theory, starting from Turing's model of computation and extending it to include classes of complexity based on the relative difficulty of families of algorithms; and complexity theory (once called "chaos theory"), an interdisciplinary effort by physicists, mathematicians, and biologists to describe how apparently complex behavior can arise from the interaction of large numbers of very simple components.

Measuring or even defining the complexity of a biological system—indeed, of any complex, dynamic system—has proven to be a difficult problem. Traditional measures of complexity that have been developed to analyze and describe the products of human technological engineering are difficult to apply or inappropriate for describing biological systems. For example, although both biological systems and engineered systems often have degrees of redundancy (i.e., multiple instances of the same "component" that serve the same function for purposes of reliability), biological systems also show many other systems-level design behaviors that are rarely if ever found in engineered systems. Indeed, many such behaviors would be considered poor design. For example, "degeneracy" in biological systems refers to the property of having different systems produce the same activity. Similarly, in most biological systems, many different components contribute to global properties, a design that if included in a human-engineered system would make it very difficult to understand.

Other attempts at measuring biological complexity include enumerating various macroscopic properties of an organism, such as the number of distinct parts, number of distinct cell types, number of biological functions performed, and so forth. In practice this can be difficult (what is considered a "distinct" part?) or inconclusive (is an organism with more cell types necessarily more complex?).

More conveniently, the entire DNA sequence of an organism's genome can be analyzed. Since DNA plays a major role in determining the structure and functions of an organism, one approach is to consider the information content of the DNA string. Of course, biological knowledge is nowhere close to actually being able to infer the totality of an organism merely from a DNA sequence, but the argu-

ment is that sequence complexity will be highly correlated with organismal complexity. (Some advantages of dealing with strings of letters as an abstraction are discussed in Section 4.4.1.)

Because information theory treats all bits as alike and of equal significance, a purely information-theoretic view would suggest that a gene of a thousand base pairs that encode a crucial protein required for the development of a human characteristic has the same information content (about 2,000 bits) as a random sequence of the same length with no biological function. This view strains plausibility or, rather, would have limited applicability to biology. Thus, the example suggests that something more is needed.

Generally, the types of complexity measures applied to DNA sequences are defined by their relationship to the process of computation. For example, a string might be considered to be a program, an input to a program, or the output of a program, and the resulting complexity measure might include the size of the Turing machine that produced it, its running time, or the number of states. Each measure captures a different sense of complexity of the DNA string and will consider different strings to be relatively more or less complex.

One such approach is the notion of Kolmogorov (or more formally, Kolmogorov-Chaitin-Solomonoff) complexity. Kolmogorov complexity is a measure of the extent to which it is possible to eliminate redundancies from a bit string without loss of information. Specifically, a program is written to generate the bit string in question. For a truly random string, the program is at least as long as the string itself. But if there are information redundancies in the string, the string can be compressed, with the compressed representation being the program needed to reproduce it. A string with high Kolmogorov complexity is one in which the difference in length between the string and its program is small; a string with low Kolmogorov complexity is one that contains many redundancies and thus for which the generating program is shorter than the string.

However, for the purpose of analyzing overall complexity, a purely random string will have a maximal Kolmogorov score, which is not what seems appropriate intuitively for estimating biological complexity. In general, a desired attribute of measures of biological complexity is the so-called one-hump criterion. A measure that incorporated this criterion would indicate a very low complexity for both very ordered sequences (e.g., a purely repeating sequence) and very random sequences and the highest complexity for sequences in the middle of a notional continuum, neither periodic nor random.[68] Feldman and Crutchfield further suggest that biological complexity must also be defined in a setting that gives a clear interpretation to what structures are quantified.[69]

Other measures that have been proposed include thermodynamic depth, which relates a system's entropy to the number of possible histories that produced its current state; logical depth, which considers the minimal running time of a program that produced a given sequence; statistical complexity measures, which indicate the correlation among different elements of an entity's components and the

[68]A related phenomenon, highly investigated but poorly understood, is the ubiquity of so-called $1/f$ spectra in many interesting phenomena, including biological systems. The term "$1/f$ spectra" refers to a type of signal whose power distribution as a function of frequency obeys an inverse power law in which the exponent is a small number. A $1/f$ signal is not random noise (random noise would result in an exponent of zero; i.e., the power spectrum of a random noise source is flat). On the other hand, there is some stochastic component to $1/f$ spectra as well as some correlation between signals at different nonadjacent times (i.e., $1/f$ noise exhibits some degree of long-range correlation). Similar statistical analyses have been applied to spatial structures, such as DNA, although power and frequency are replaced by frequency of base-pair occurrence and spatial interval, respectively (see, for example, A.M. Selvam, "Quantumlike Chaos in the Frequency Distributions of the Bases A, C, G, T in Drosophila DNA," *APEIRON* 9(4):103-148, 2002; W. Li, T.G. Marr, and K. Kaneko, "Understanding Long-range Correlations in DNA Sequences," *Physica D* 75(1-3):392-416, 1994 [erratum:82, 217,1995]). $1/f$ spectra have been found in the temporal fluctuations of many biological processes, including ion channel kinetics, auditory nerve firings, lung inflation, fetal breathing, human cognition, walking, blood pressure, and heart rate. (See J.M. Hausdorff and C.K. Peng, "Multiscaled Randomness: A Possible Source of $1/f$ Noise in Biology," *Physical Review E* 54(2):2154-2157, 1996, and references therein. Hausdorff and Peng suggest that if the time scales of the inputs affecting a biological system are "structured" and there are a large number of inputs, it is very likely that the output will exhibit $1/f$ spectra, even if individual input amplitudes and time scales are loosely correlated.)

[69]D.P. Feldman and J.P. Crutchfield, "Measures of Statistical Complexity: Why?," *Physics Letters A* 238:244-252, 1997.

degree of structure or pattern in that entity; and physical complexity, which interprets the shared Kolmogorov complexity of an ensemble of sequences as information stored in the genome about the environment. This last makes the interesting point that one cannot know anything about the meaning of a DNA sequence without considering the environment in which the corresponding organism is expected to live.

All of these capture some aspect of the way in which complexity might arise over time through an undirected evolutionary process and be stored in the genome of a species. However, in their physics-inspired search for minimal descriptions, they may be missing the fact that evolution does not produce optimal or minimal descriptions. That is, because biological organisms are the result of their evolutionary histories, they contain many remnants that are likely to be irrelevant to their current environmental niches, yet contribute to their complexity. Put differently, any given biological organism is almost certainly not optimized to perform the functions of which it is capable.

Another difficulty with many of these measures' application to biology is that, regardless of their theoretical soundness, they will almost certainly be hard to determine empirically. More prosaically, they often involve a fair amount of mathematics or theoretical computational reasoning (e.g., to what level of the Chomskian hierarchy of formal languages does this sequence belong?) completely outside the experience of the majority of biologists. Regardless, this is an area of active research, and further integration with actual biological investigation is likely to produce further progress in identifying accurate and useful measures of complexity.