

21st Century Biology

Biology, like any science, changes when technology introduces new tools that extend the scope and type of inquiry. Some changes, such as the use of the microscope, are embraced quickly and easily, because they are consonant with existing values and practices. Others, such as the introduction of multivariate statistics as performed by computers in the 1960s, are resisted, because they go against traditions of intuition, visualization, and conceptions of biology that separate it clearly from mathematics.

This chapter attempts to frame the challenges and opportunities created by the introduction of computation to the biological sciences. It does so by first briefly describing the existing threads of biological culture and practice, and then by showing how different aspects of computational science and technology can support, extend, or challenge the existing framework of biology.

Computing is only one of a large number of fields playing a role in the transformation of biology, from advanced chemistry to new fields of mathematics. And yet, in many ways, computers have proven the most challenging and the most transformative, rooted as they are in a tradition of design and abstraction so different from biology. Just as computers continue to radically change society at large, however, there is no doubt that they will change biology as well. As it has done so many times before, biology will change with this new technology, adopting new techniques, redefining what makes good science and good training, and changing which inquiries are important, valued, or even possible.

2.1 WHAT KIND OF SCIENCE?

2.1.1 The Roots of Biological Culture

Biology is a science with a deep history that can be linked to the invention of agriculture at the very dawn of civilization and, even earlier, to the first glimmerings of oral culture: “Is that safe to eat?” As such, it is a broad field, rich with culture and tradition, that encompasses many threads of observational, empirical, and theoretical research and spans scales from single molecules to continents. Such a broad field is impossible to describe simply; nevertheless, this section attempts to identify a number of the main threads of the activity and philosophy of biology.

First, biology is an *empirical* and a *descriptive* science. It is rooted in a tradition of qualitative observation and description dating back at least to Aristotle. Biological researchers have long sought to

catalog the characteristics, behaviors, and variations of individual biological organisms or populations through the direct observation of organisms in their environments, rather than trying to identify general principles through mathematical or abstract modeling. For this reason, the culture of biology is both strongly visual and specific. Identifying a new species, and adequately describing its physical appearance, environment, and life cycle, remains a highly considered contribution to biological knowledge.

It is revealing to contrast this philosophy with that of modern physics, where the menagerie of new subatomic particles discovered in the 1960s and 1970s was a source of faint embarrassment and discomfort for physicists. Only with the introduction of quarks, and the subsequent reduction in the number of fundamental particles, did physicists again feel comfortable with the state of their field. Biology, in strong contrast, not only prizes and embraces the enormous diversity of life, but also considers such diversity a prime focus of study.

Second, biology is an *ontological* science, concerned with taxonomy and classification. From the time of Linnaeus, biologists have attempted to place their observations into a larger framework of knowledge, relating individual species to the identified span of life. The methodology and basis for this catalog is itself a matter of study and controversy, and so research activity of this type occurs at two levels: specific species are placed into the tree of life (or larger taxa are relocated), still a publishable event, and the science of taxonomy itself is refined.

Biology is a *historical* science. Life on Earth apparently arose just once, and all life today is derived from that single instance. A complete history of life on Earth—which lineage arose from which, and when—is one of the great, albeit possibly unachievable, goals of biology. Coupled to this inquiry, but separate, are the questions, *How?* and *Why?* What are the forces that cause species to evolve in certain ways? Are there secular trends in evolution, for example, as is often claimed, toward increasing complexity? Does evolution proceed smoothly or in bursts? If we were to “replay the tape” of evolution, would similar forms arise? Just as with taxonomy (and closely related to it), there are two levels here: what precisely happened and what the forces are that cause things to happen.

These three strands—empirical observations of a multitude of life forms, the historical facts of evolution, and the ordering of biological knowledge into an overarching taxonomy of life—served to define the central practices of biology until the 1950s and still in many ways affect the attitudes, training, philosophy, and values of the biological sciences. Although biology has expanded considerably with the advent of molecular biology, these three strands continue as vital areas of biological research and interest.

These three intellectual strands have been reflected in biological research that has been qualitative and descriptive throughout much of its early history. For example, empirical and ontological researchers have sought to catalog the characteristics, behaviors, and variations of individual biological organisms or populations through the direct observation of organisms in their environments.

Yet as important and valuable as these approaches have been for biology, they have not provided—and cannot provide—very much detail about underlying mechanisms. However, in the last half-century, an intellectual perspective provided by molecular biology and biochemistry has served as the basis for enormous leaps forward.

2.1.2 Molecular Biology and the Biochemical Basis of Life

In the past 50 years, biochemical approaches to analyzing biological questions and the overall approaches now known as molecular biology have led to the increased awareness, identification, and knowledge of the central role of certain mechanisms, such as the digital code of DNA as the mechanism underlying heredity, the use of adenosine triphosphate (ATP) for energy storage, common protein signaling protocols, and many conserved genetic sequences, some shared by species as distinct as humans, sponges, and even single-cell organisms such as yeast.

This new knowledge both shaped and was shaped by changes in the practice of biology. Two important threads of biological inquiry, both existing long before the advent of molecular biology, came

to the forefront in the second half of the 20th century. These threads were biological experimentation and the search for the underlying mechanics of life.

Biological experimentation and the collection of data are not new, but they acquired a new importance and centrality in the late 20th century. The identification of genes and mutations exemplified by experiments on *Drosophila* became an icon of modern biological science, and with this a new focus emerged on collecting larger amounts of quantitative data.

Biologists have always been interested in how organisms live, a question that ultimately comes down to the very definition of life. A great deal of knowledge regarding anatomy, circulation, respiration, and metabolism was gathered in the 18th and 19th centuries, but without access to the instruments and knowledge of biochemistry and molecular biology, there was a limit to what could be discovered. With molecular biology, some of the underlying mechanisms of life have been identified and analyzed quantitatively.

The effort to uncover the basic chemical features of biological processes and to ascertain all aspects of the components by way of experimental design will continue to be a major aspect of basic biological research, and much of modern biology has sought to reduce biological phenomena to the behavior of molecules.

However, biological researchers are also increasingly interested in a systems-level view in which completely novel relationships among system components and processes can be ascertained. That is, a detailed understanding of the components of a biological organism or phenomenon inevitably leads to the question of how these components interact with each other and with the environment in which the organism or phenomenon is embedded.

2.1.3 Biological Components and Processes in Context, and Biological Complexity

There is a long tradition of studying certain biological systems in context. For example, ecology has always focused on ecosystems. Physiology is another example of a life science that has generally considered biological systems as whole entities. Animal behavior and systematics science also considers biological phenomena in context. However, data acquisition technologies, computational tools, and even new intellectual paradigms are available today that enable a significantly greater degree of in-context understanding of many more biological components and processes than was previously possible, and the goal today is to span the space of biological entities from genes and proteins to networks and pathways, from organelles to cells, and from individual organisms to populations and ecosystems.

Following Kitano,¹ a systems understanding of a biological entity is based on insights regarding four dimensions: (1) system structures (e.g., networks of gene interactions and biochemical pathways and their relationship to the physical properties of intracellular and multicellular structures), (2) system dynamics (e.g., how a system behaves over time under various conditions and the mechanisms underlying specific behaviors), (3) control mechanisms (e.g., mechanisms that systematically control the state of the cell), and (4) design principles (e.g., principles underlying the construction and evolution of biological systems that have certain desirable properties).²

As an example, consider advances in genomic sequencing. Sequence genomics has created a path for establishing the “parts list” for living cells, but to move from isolated molecular details to a comprehensive understanding of phenomena from cell growth up to the level of homeostasis is widely recog-

¹H. Kitano, “Systems Biology: A Brief Overview,” *Science* 295(5560):1662-1664, 2002.

²For example, such principles might occur as the result of convergent evolution, that is, the evolution of species with different origins toward similar forms or characteristics, and an understanding of the likely ways that evolution can take to solve certain problems. Alternatively, principles might be identified that can explain the functional behavior of some specific biological system under a wide set of circumstances without necessarily being an accurate reflection of what is going on inside the system. Such principles may prove useful from the standpoint of being able to manipulate the behavior of a larger system in which the smaller system is embedded, though they may not be useful in providing a genuine understanding of the system with which they are associated.

nized as requiring a very different approach. In the highly interactive systems of living organisms, the macromolecular, cellular, and physiological processes, themselves at different levels of organizational complexity, have both temporal and spatial components. Interactions occur between sets of similar objects, such as two genes, and between dissimilar objects, such as genes and their environment.

A key aspect of biological complexity is the role of chance. One of the most salient instances of chance in biology is evolution, in which chance events affect the fidelity of genetic transmission from one generation to the next. The hand of chance is also seen in the development of an organism—chance events affect many of the details of development, though generally not the broad picture or trends. But perhaps the most striking manifestation is that individual biological organisms—even as closely related as sibling cells—are unlikely to be identical because of stochastic events from environmental input to thermal noise that affect molecular-level processes. If so, no two cells will have identical macromolecular content, and the dynamic structure and function of the macromolecules in one cell will never be the same as even a sibling cell. This fact is one of the largest distinctions between living systems and most silicon devices or almost any other manufactured or human-engineered artifact.

Put differently, the digital “code of life” embedded in DNA is far from simple. For example, the biological “parts list” that the genomic sequence makes available in principle may be unavailable in practice if all of the parts cannot be identified from the sequence. Segments of the genome once assumed to be evolutionary “junk” are increasingly recognized as the source of novel types of RNA molecules that are turning out to be major actors in cellular behavior. Furthermore, even a complete parts list provides a lot less insight into a biological system than into an engineered artifact, because human conventions for assembly are generally well understood, whereas nature’s conventions for assembly are not.

A second example of the complexity is that a single gene can sometimes produce *many* proteins. In eukaryotes, for example, mRNA cannot be used as a blueprint until special enzymes first cut out the introns, or noncoding regions, and splice together the exons, the fragments that contain useful code.³ In some cases, however, the cell can splice the exons in different ways, producing a series of proteins with various pieces added or subtracted but with the same linear ordering (these are known as splice variants). A process known as RNA editing can alter the sequence of nucleotides in the RNA after transcription from DNA but before translation into a protein, resulting in different proteins. An individual nucleotide can be changed into a different one (“substitution editing”), or nucleotides can be inserted or deleted from the RNA (“insertion-deletion editing”). In some cases (however rare), the cell’s translation machinery might introduce an even more radical change by shifting its “reading frame,” meaning that it starts to read the three-base-pair genetic code at a point displaced by one or two base pairs from the original. The result will be a very different sequence of amino acids and, thus, a very different protein.

Furthermore, even after the proteins are manufactured at the ribosome, they undergo quite a lot of postprocessing as they enter the various regulatory networks. Some might have their shapes and activity levels altered by the attachment, for example, of a phosphate group, a sugar molecule, or any of a variety of other appendages, while others might come together to form a multiprotein structure. In short, knowing the complete sequence of base pairs in a genome is like knowing the complete sequence of *1s* and *0s* that make up a computer program: by itself, that information does not necessarily yield insight into what the program does or how it may be organized into functional units such as subroutines.⁴

A third illustration of biological complexity is that few, if any, biological functions can be assigned to a single gene or a single protein. Indeed, the unique association between the hemoglobin molecule and the function of oxygen transport in the bloodstream is by far the exception rather than the rule.

³Virtually all introns are discarded by the cell, but in a few cases, an intron has been found to code—by itself—for another protein.

⁴A meaningful analogy can be drawn to the difference between object code and source code in a computer. Object code, consisting of binary digits, is what runs on the computer. Source code, usually written in a high-level programming language, is compiled into object code so that a program will run, but source code—and therefore program structure and logic—is much more comprehensible to human beings. Source code is also much more readily changed.

Much more common is the situation in which biological function depends on interactions among many biological components. A cell's metabolism, its response to chemical and biological signals from the outside, its cycle of growth and cell division—all of these functions and more are generally carried out and controlled by elaborate webs of interacting molecules.

François Jacob and Jacques Monod won the 1965 Nobel Prize in medicine for the discovery that DNA contained regulatory regions that governed the expression of individual genes.⁵ (They further emphasized the importance of regulatory feedback and discussed these regulatory processes using the language of circuits, a point of relevance in Section 5.4.3.3.) Since then, it has become understood that proteins and other products of the genome interact with the DNA itself (and with each other) in a regulatory web.

For example, RNA molecules have a wide range of capabilities beyond their roles as messengers from DNA to protein. Some RNA molecules can selectively silence or repress gene transcription; others operate as a combination chemoreceptor-gene transcript ("riboswitch") that gives rise to a protein at one end of the molecule when the opposite end comes in contact with the appropriate chemical target. Indeed, it may even be that a significant increase in the number of regulatory RNAs on an evolutionary time scale is largely responsible for the increase in eukaryotic complexity without a large increase in the number of protein-coding genes. Understanding the role of RNA and other epigenetic phenomena that result in alternative states of gene expression, molecular function, or organization—"systems [that] are far more complex than any problem that molecular biology, genetics or genomics has yet approached,"⁶ is critical to realizing genomics' promise.

A fourth example of biological complexity is illustrated by the fact that levels of biological complexity extend beyond the intricacies of the genome and protein structures through supramolecular complexes and organelles to cellular subsystems and assemblies of these to form often functionally polarized cells that together contribute to tissue form and function and, thereby to an organism's properties. Although the revolution of the last half of the last century in biochemistry and molecular biology has contributed significantly to our knowledge of the building blocks of life, we have only begun to scratch the surface of a data-dense and Gordian knot-like puzzle of complex and dynamic molecular interactions that give rise to the complex behaviors of organisms. In short, little is known about how the complexities of physiological processes are governed by molecular, cellular, and transcellular signaling systems and networks. Available information is deep only in limited spatial or temporal domains, and scarce in other key domains, such the middle spatial scales (e.g., 10 Å-10 μm), and there are no tools that make intelligent links between relatable pieces of scientific knowledge across these scales.

Complexity, then, appears to be an essential aspect of biological phenomena. Accordingly, the development of a coherent intellectual approach to biological complexity is required to understand systems-level interactions—of molecules, genes, cells, organisms, populations, and even ecosystems. In this intellectual universe, both "genome syntax" (the letters, words, and grammar associated with the DNA code) and "genome semantics" (what the DNA code can express and do) are central foci for investigation. Box 2.1 describes some of the questions that will arise in cell biology.

2.2 TOWARD A BIOLOGY OF THE 21st CENTURY

A biology of the 21st century will integrate a number of diverse intellectual themes.⁷ One integration is that of the reductionist and systems approaches. Where the component-centered reductionist

⁵F. Jacob and J. Monod, "Genetic Regulatory Mechanisms in the Synthesis of Proteins," *Journal of Molecular Biology* 3:318-356, 1961.

⁶F.S. Collins et al., "A Vision for the Future of Genomic Research," *Nature* 422:835-847, 2003.

⁷What this report calls 21st century biology has also been called "bringing the genome to life," an intentional biology, an integrative biology, synthetic biology, the new biology or even the next new biology, *Biology 21*, beyond the genome, postgenomic biology, genome-enabled science, and industrialized biology.

Box 2.1 Some Questions for Cell Biology in the 21st Century

In the Human Genome Institute's recently published agenda for research in the postgenome era, Francis Collins and his coauthors repeatedly emphasized how little biologists understand about the data already in hand. Collins et al. argue that biologists are a very long way from knowing everything there is to know about how genes are structured and regulated, for example, and they are virtually without a clue as to what's going on in the other 95 percent of the genome that does not code for genes. This is why the agenda's very first grand challenge was to systematically endow those data with meaning—that is, to “comprehensively identify the structural and functional components encoded in the human genome.”¹

The challenge, in a nutshell, is to understand the cellular information processing system—all of it—from the genome on up. Weng et al. suggest that the essential defining feature of a cell, which makes the system as a whole extremely difficult to analyze, is the following:²

[The cell] is not a machine (however complex) drawn to a well-defined design, but a machine that can and does constantly rebuild itself within a range of variable parameters. For a systematic approach, what is needed is a relatively clear definition of the boundary of this variability. In principle, these boundaries are determined by an as-yet-unknown combination of intrinsic capability and external inputs. The balance between intrinsic capability and the response to external signals is likely to be a central issue in understanding gene expression. . . . A large body of emerging data indicates that early development occurs through signaling interactions that are genetically programmed, whereas at the later stages, the development of complex traits is dependent on external inputs as well. A quantitative description of this entire process would be a culmination and synthesis of much of biology.

Some of the questions raised by this perspective include the following:

- What is the proteome of any given cell? How do these individual protein molecules organize themselves into functional subnetworks—and how do these subnetworks then organize themselves into higher- and higher-level networks?³ What are the functional design principles of these systems? And how, precisely, do the products of the genome react *back* on the genome to control their own creation?
- To what extent are active elements (such as RNA) present in the noncoding portions of the genome? What is the inventory of epigenetic mechanisms (e.g., RNA silencing, DNA methylation, histone hypoacetylation, chromatin modifications, imprinting) that cells use to control gene expression? These mechanisms play important roles in controlling an organism's development and, in some lower organisms, are defense responses against viruses and transposable elements. However, epigenetic phenomena have also been implicated in several human diseases, particularly cancer development due to the repression of tumor suppressor genes. What activates these mechanisms?
- How do these dynamically self-organizing networks vary over the course of the cell cycle (even though most cells in an organism are not proliferating and have exited from the cell cycle)? How do they change as the cell responds to its surroundings? How do they encode and process information? Also, what accounts for life's *robustness*—the ability of these networks to adapt, maintain themselves, and recover from a wide variety of environmental insults?

¹F.S. Collins, E.D. Green, A.E. Guttacher, and M.S. Guyer, “A Vision for the Future of Genomic Research,” *Nature* 422(6934):835-847, 2003. To help achieve this grand challenge, the institute has launched the ENCODE project, a public research consortium dedicated to building an annotated encyclopedia of all known functional DNA elements. See <http://www.genome.gov/10005107>.

²G. Weng, U.S. Bhalla, and R. Iyengar, “Complexity in Biological Signaling Systems,” *Science* 284(5411):92-96, 1999.

³The hierarchy of levels obviously doesn't stop at the cell membrane. Although deciphering the various cellular regulatory networks is a huge challenge in itself, systems biology ultimately has to deal as well with how cells organize themselves into tissues, organs, and the whole organism. One group that is trying to lay the groundwork for such an effort is the Physiome Project at the University of Auckland in New Zealand. See http://www.webopedia.com/TERM/W/Web_services.html.

- How do cells develop spatial structure? The cytoplasm is far from a uniform mixture of all of the biomolecules that exist in a cell; proteins and other macromolecules are often bound to membranes or isolated inside various cellular compartments (especially eukaryotes). A full account of the regulatory networks has to take this compartmentalization into account, along with such spatial factors as diffusion and the transport of various species through the cytoplasm and across membranes.
- How do the networks organize and reorganize themselves over the course of embryonic development, as each cell decides whether its progeny are going to become skin, muscle, brain, or whatever?⁴ Then, once the cells are through differentiating, how do the networks actually vary from one cell type to the next? What constitutes the difference, and what happens to the networks as cells age or are damaged? How do flaws in the networks manifest themselves as maladies such as cancer?
- How do the networks vary between individuals? How do those variations account for differences in morphology and behavior? Also—especially in humans—how do those variations account for individual differences in the response to drugs and other therapies?
- How do multicellular organisms operate? A full account of multicellular organisms will have to include an account of signaling (in all its varieties, including cell-cell; cell-substratum; autocrine, paracrine, and exocrine signaling), cellular differentiation, cell motility, tissue architecture, and many other “community” issues.
- How do the networks vary between species? To put it another way, how have they changed over the course of evolution? Since the “blueprint” genes for proteins and RNA seem to be quite highly conserved from one species to the next, is it possible that most of evolution is the result of rearrangements in the genetic regulatory system?⁵

⁴Physiological processes such as metabolism, signal transduction, and the cell cycle take place on a time scale that ranges from milliseconds to days and are reversible in the sense that an activity flickers on, gene expression is adjusted as needed, and then everything returns to some kind of equilibrium. But the commitments that the cell makes during development are effectively *irreversible*. Becoming a particular cell line means that the genetic regulatory networks in each successive generation of cells have to go through a cascade of decisions that end up turning genes on and off by the thousands. Unless there is some drastic intervention, as in the cloning experiments that created Dolly the Sheep, those genes are locked in place for the life span of the organism. Of course, the developmental program does not proceed in an isolated, “open-loop” fashion, as a computer scientist might say. Very early in the process, for example, the growing embryo lays out its basic body plan—front versus back, top versus bottom, and so on—by establishing embryo-wide chemical gradients, so that the concentration of the appropriate compound tells each cell what to do. Similar tricks are used at every stage thereafter: each cell is always receiving copious feedback from its neighbors, with chemical signals providing a constant stream of instructions and course corrections.

⁵After all, even very small changes in the timing of events during development, and in the rates at which various tissues grow, can have a profound impact on the final outcome.

approach is based on identifying the constituent parts of an organism and understanding the behavior of the organism in terms of the behavior of those parts (in the limit, a complete molecular-level characterization of the biological phenomena in question), systems biology aims to understand the mechanisms of a living organism across all relevant levels of hierarchy.⁸ These different foci—a focus on components of biological systems versus a focus on interactions among these components—are complementary, and both will be essential for intellectual progress in the future.

Twenty-first century biology will bring together many distinct strands of biological research: taxonomic studies of many species, the enormous progress in molecular genetics, steps towards understanding the molecular mechanisms of life, and an emerging systems biology that will consider biological entities in relationship to their larger environment. Twenty-first century biology aims to understand fully the mechanisms of a living cell and the increasingly complex hierarchy of cells in metazoans, up to

⁸As a philosophical matter, the notion of reductionist explanation has had a long history in the philosophy of science. Life is composed of matter, and matter is governed by the laws of physics. So, the ultimate in reductionist explanation would suggest that life can be explained by the properties of Schrödinger’s equation.

processes operating at the level of the organism and even populations and ecosystems. However, this kind of understanding is fundamentally dependent on synergies between a systems understanding as described above and the reductionist tradition.

Twenty-first century biology also brings together empirical work in biology with computational work. Empirical work is undertaken in laboratory experiments or field observations and has led to both hypothesis testing and hypothesis generation. Hypothesis testing relies on the data provided by empirical work to accept or reject a candidate hypothesis. However, data collected in empirical work can also suggest new hypotheses, leading to work that is exploratory in nature. In 21st century biology, computational work provides a variety of tools that support empirical work, but also enables much of systems biology through techniques such as simulation, data mining, and microarray analysis—and thus underlies the generation of plausible candidate hypotheses that will have to be tested. Note also that hypothesis testing is relevant to both reductionist and systems biology, in the sense that both types of biology are formulated around hypotheses (about components or about relationships between components) that may—or may not—be consistent with empirical or experimental results.

In this regard, a view expressed by Walter Gilbert in 1991 seems prescient. Gilbert noted that “in the current paradigm [i.e., that of 1991], the attack on the problems of biology is viewed as being solely experimental. The ‘correct’ approach is to identify a gene by some direct experimental procedure—determined by some property of its product or otherwise related to its phenotype—to clone it, to sequence it, to make its product and to continue to work experimentally so as to seek an understanding of its function.” He then argued that “the new paradigm [for biological research], now emerging [i.e., in 1991], is that all the genes will be known (in the sense of being resident in databases available electronically), and that the starting point of a biological investigation will be theoretical. An individual scientist will begin with a theoretical conjecture, only then turning to experiment to follow or test that hypothesis. The actual biology will continue to be done as ‘small science’—depending on individual insight and inspiration to produce new knowledge but the reagents that the scientist uses will include a knowledge of the primary sequence of the organism, together with a list of all previous deductions from that sequence.”⁹

Finally, 21st century biology encompasses what is often called discovery science. Discovery science has been described as “enumerat[ing] the elements of a system irrespective of any hypotheses on how the system functions” and is exemplified by genome sequencing projects for various organisms.¹⁰ A second example of discovery science is the effort to determine the transcriptomes and proteomes of individual cell types (e.g., quantitative measurements of all of the mRNAs and protein species).¹¹ Such efforts could be characterized as providing the building blocks or raw materials out of which hypotheses can be formulated—metaphorically, words of a biological “language” for expressing hypotheses. Yet even here, the Human Genome Project, while unprecedented in its scope, is comfortably part of a long tradition of increasingly fine description and cataloging of biological data.

All told, 21st century biology will entail a broad spectrum of research, from laboratory work directed by individual principal investigators, to projects on the scale of the human genome that generate large amounts of primary data, to the “mesoscience” in between that involves analytical or synthetic work conducted by multiple collaborating laboratories. For the most part, these newer research strategies involving discovery science and analytical work will complement rather than replace the traditional, relatively small laboratory focusing on complementary empirical and experimental methods.

⁹W. Gilbert, “Towards a Paradigm Shift in Biology,” *Nature* 349(6305):99, 1991.

¹⁰R. Aebersold, L.E. Hood, and J.D. Watts, “Equipping Scientists for the New Biology,” *Nature Biotechnology* 18:359, 2000.

¹¹These examples are taken from T. Ideker, T. Galitski, and L. Hood, “A New Approach to Decoding Life: Systems Biology,” *Annual Review of Genomics and Human Genetics* 2:343-372, 2001. The transcriptome is the complete collection of transcribed elements of the genome, including all of the genetic elements that code for proteins, all of the mRNAs, and all noncoding RNAs that are used for structural and regulatory purposes. The proteome is the complete collection of all proteins involved in a particular pathway, organelle, cell, tissue, organ, or organism that can be studied in concert to provide accurate and comprehensive data about that system.

Grand questions, such as those concerning origins of life, the story of evolution, the architecture of the brain, and the interactions of living things with each other in populations and ecosystems, are up for grabs in 21st century biology, and the applications to health, agriculture, and industry are no less ambitious. For example, 21st century biology may enable the identification of individuals who are likely to develop cancer, Alzheimer's, or other diseases, or who will respond to or have a side effect from a particular disease treatment. Pharmaceutical companies are making major investments in transcriptomics to screen for plausible drug targets. Forward-thinking companies want to develop more nutritious plants and animals, commandeer the machinery of cells to produce materials and drugs, and build interfaces to the brain to correct impaired capabilities or produce enhanced abilities. Agencies interested in fighting bioterrorism want to be able to rapidly identify the origins and ancestry of pathogen outbreaks, and stewards of natural systems would like to make better predictions about the impacts of introduced species or global change.

2.3 ROLES FOR COMPUTING AND INFORMATION TECHNOLOGY IN BIOLOGY

To manage biological data, 21st century biology will integrate discovery science, systems biology, and the empirical tradition of biological science and provide a quantitative framework within which the results of efforts in each of these areas may be placed. The availability of large amounts of biological data is expected to enable biological questions to be addressed globally, for example, examining the behavior of all of the genes in a genome, all of the proteins produced in a cell type, or all of the metabolites created under particular environmental conditions. However, enabling the answering of biological questions by uncovering the raw data is not the same as answering those questions—the data must be analyzed and used in intellectually meaningful and significant ways.

2.3.1 Biology as an Information Science

The data-intensive nature of 21st century biology underlies the dependence of biology on information technology (IT). For example, even in 1990 it was recognized that IT would play a central role in the International Human Genome Consortium for the storage and retrieval of biological gene sequence data—recording the signals, storing the sequence data, processing images of fluorescent traces specific to each base, and so on. Also, as biology unfolds in the 21st century, it is clear that the rate of production of biological data will not abate. Data acquisition opportunities will emerge in most or all life science subdisciplines and fields, and life scientists will have to cope with the coming deluge of highly multivariate, largely nonreducible data, including high-resolution imaging and time series data of complex dynamic processes.

Yet beyond data management issues, important and challenging though they are, it has also become clear that computing and information technology will play crucial roles in identifying meaningful structures and patterns in the genome (e.g., genes, genetic regulatory elements), in understanding the interconnections between various genomic elements, and in uncovering functional biological information about genes, proteins, and their interactions. This focus on information—on acquiring, processing, structuring, and representing information—places genomic studies squarely in the domain of computing and information science.

Of course, genomic studies are not the whole of modern biology. For life sciences ranging from ecology, botany, zoology, and developmental biology to cellular and molecular biology—all of which can be characterized as science with diverse data types and high degrees of data heterogeneity and hierarchy—IT is essential to collect key information and organize biological data in methodical ways in order to draw meaningful observations. Massive computing power, novel modeling approaches, new algorithms and mathematical or statistical techniques, and systematic engineering approaches will provide biologists with vital and essential tools for managing the heterogeneity and volume of the data and for extracting meaning from those data.

Ultimately, what calculus is to the language of the physical sciences, computing and information will be to the language of 21st century biology, or at least to its systems biology thread.¹² The processes of biology, the activities of living organisms, involve the usage, maintenance, dissemination, transformation or transduction, replication, and transmittal of information across generations. Biological systems are characterized by individuality, contingency, historicity, and high digital information content—every living thing is unique. Furthermore, the uniqueness and historical contingency of life means that for population-scale problems, the potential state space that the population actually inhabits is huge.¹³ As an information science, the life sciences use computing and information technology as a language and a medium in which to manage the discrete, asymmetric, largely irreducible, unique nature of biological systems and observations.

In the words above, those even marginally familiar with the history of biology will recognize hints of what was once called theoretical biology or mathematical biology, which in earlier days meant models and computer simulations based on such then-fashionable ideas as cybernetics and general systems theory.¹⁴ The initial burst of enthusiasm waned fairly quickly, as it became clear that the available experimental data were not sufficient to keep the mathematical abstractions tethered to reality. Indeed, reliable models are impossible when many or most of the quantitative values are missing. Moreover, experience since then has indicated that biological systems are much more complex and internally interlinked than had been imagined—a fact that goes a long way towards explaining why the models of that era were not very successful in driving productive hypothesis generation and research.

The story is radically different today. High-throughput data acquisition technologies (themselves enabled and made practical by today's information technologies), change a paucity of data into a deluge of it, as illustrated by the use of these technologies for sequencing of many eukaryotic organisms. This is not to say that more data are not needed, merely that the acquisition of necessary data now seems to be possible in reasonable amounts of time.

The same is true for the information technologies underpinning 21st century biology. In the past, even if data had been available, the IT then available would have been inadequate to make sense out of those data. But today's information technologies are vastly more powerful and hold considerable promise for enabling the kinds of data management and analytical capabilities that are necessary for a systems-level approach. Moreover, information technology as an underlying medium has the advantage of growing ever more capable over time at exponential rates. As information technology becomes more capable, biological applications will have an increasingly powerful technology substrate on which to draw.

¹²Biological Sciences Advisory Committee on Cyberinfrastructure for the Biological Sciences, *Building a Cyberinfrastructure for the Biological Sciences (CIBIO): 2005 and Beyond: A Roadmap for Consolidation and Exponentiation*, July 2003. Available from http://research.calit2.net/cibio/archived/CIBIO_FINAL.pdf. This is not to deny that calculus also has application in systems biology (mostly through its relevance to biochemistry and thermodynamics), but calculus is not nearly as central to systems biology as it is to the physical sciences nor as central as computing and information technology are to systems biology.

¹³The number of possible different 3-billion-base-pair genomes, assuming only simple base substitution mutations, is 4 to the 3-billionth power. That's a big number. In fact, it is so big that the ratio of that number (big) to the number of particles in the known universe (small) is much greater than the ratio of the diameter of the universe to the diameter of a carbon atom. Thus, exhaustive computer modeling of that state space is effectively precluded. Even more tractable state spaces, such as the number of different possible human haploid genotypes, still produce gigantic numbers. For example, if we assume that the entire human population is heterozygous at just 500 locations throughout the genome (a profound underestimate of existing diversity), with each site having only two states, then the number of possible human haplotypes is 2 to the 500th power, which also exceeds the number of electrons in the known universe. These back-of-the-envelope calculations also show that it is impossible for the state space of existing human genotypes to exist in anything approaching linkage equilibrium.

¹⁴N. Wiener, *Cybernetics, or Control and Communication in the Animal and the Machine*, 2nd ed., MIT Press, Cambridge, MA, 1961; L. von Bertalanffy, *General Systems Theory: Foundations, Development, Applications*, George Braziller, New York, 1968. This history was recently summarized in O. Wolkenhauer, "Systems Biology: The Reincarnation of Systems Theory Applied in Biology?" *Briefings in Bioinformatics* 2(3):258-270, 2001.

In short, the introduction of computing into biology has transformed, and continues to transform, the practice of biology. The most straightforward, although often intellectually challenging, way involves computing tools with which to acquire, store, process, and interpret enormous amounts of biological data. But computing (when used wisely and in combination with the tools of mathematics and physics) will also provide biologists with an alternative and possibly more appropriate language and set of abstractions for creating models and data representations of higher-order interactions, describing biological phenomena, and conceptualizing some characteristics of biological systems.

Finally, it should be noted that although computing and information technology will become an increasingly important part of life science research, researchers in different subfields of biology are likely to understand the role of computing differently. For example, researchers in molecular biology or biophysics may focus on the ability of computing to make more accurate quantitative predictions about enzyme behavior, while researchers in ecology may be more interested in the use of computing to explore relationships between ecosystem behavior and perturbations in the ambient environment. These perspectives will become especially apparent in the chapters of this report dealing with the impact of computing and IT on biology (see Chapter 4 on tools and Chapter 5 on models).

This report distinguishes between computational tools, computational models, information abstractions and a computational perspective on biology, and cyberinfrastructure and data acquisition technologies. Each of these is discussed in Chapters 4 through 7, respectively, preceded by a short chapter on the nature of biological data (Chapter 3).

2.3.2 Computational Tools

In the lexicon of this report, computational tools are artifacts—usually implemented as software, but sometimes as hardware—that enable biologists to solve very specific and precisely defined problems. For example, an algorithm for gene finding or a database of genomic sequences is a computational tool. As a rule, these tools reinforce and strengthen biological research activities, such as recording, managing, analyzing, and presenting highly heterogeneous biological data in enormous quantity. Chapter 4 focuses on computational tools.

2.3.3 Computational Models

Computational models apply to specific biological phenomena (e.g., organisms, processes) and are used for several purposes. They are used to test insight; to provide a structural framework into which observations and experimental data can be coherently inserted; to make hypotheses more rigorous, quantifiable, and testable; to help identify key or missing elements or important relationships; to help interpret experimental data; to teach or present system behavior; and to predict dynamical behavior of complex systems. Predictive models provide some confidence that certain aspects of a given biological system or phenomenon are understood, when their predictions are validated empirically. Chapter 5 focuses on computational models and simulations.

2.3.4 A Computational Perspective on Biology

Coming to grips with the complexity of biological phenomena demands an array of intellectual tools to help manage complexity and facilitate understanding in the face of such complexity. In recent years, it has become increasingly clear that many biological phenomena can be understood as performing information processing in varying degrees; thus, a computational perspective that focuses on information abstractions and functional behavior has potentially large benefit for this endeavor. Chapter 6 focuses on viewing biological phenomena through a computational lens.

2.3.5 Cyberinfrastructure and Data Acquisition

Cyberinfrastructure for science and engineering is a term coined by the National Science Foundation to refer to distributed computer, information, and communication technologies and the associated organizational facilities to support modern scientific and engineering research conducted on a global scale. Cyberinfrastructure for the life sciences is increasingly an enabling mechanism for a large-scale, data-intensive biological research effort, inherently distributed over multiple laboratories and investigators around the world, that facilitates the integration of experimental data, enables collaboration, and promotes communication among the various actors involved.

Obtaining primary biological data is a separate question. As noted earlier, 21st century biology is increasingly a data-intensive enterprise. As such, tools that facilitate acquisition of the requisite data types in the requisite amounts will become ever more important in the future. Although they are not by any means the whole story, advances in IT and computing will play key roles in the development of new data acquisition technologies that can be used in novel ways.

Chapter 7 focuses on the roles of cyberinfrastructure and data acquisition for 21st century biology.

2.4 CHALLENGES TO BIOLOGICAL EPISTEMOLOGY

The forthcoming integration of computing into biological research raises deep epistemological questions about the nature of biology itself. For many thousands of years, a doctrine known as vitalism held that the stuff of life was qualitatively different from that of nonlife and, consequently, that living organisms were made of a separate substance than nonliving things or that some separate life force existed to animate the materials that composed life.

While this belief no longer holds sway today (except perhaps in bad science fiction movies), the question of how biological phenomena can be understood has not been fully settled. One stance is based on the notion that the behavior of a given system is explained wholly by the behaviors of the components that make up that system—a view known as reductionism in the philosophy of science. A contrasting stance, known as autonomy in the philosophy of science, holds that in addition to understanding its individual components, understanding of a biological system must also include an understanding of the specific architecture and arrangement of the system's components and the interactions among them.

If autonomy is accepted as a guiding worldview, introducing the warp of computing into the weft of biology creates additional possibilities for intellectual inquiry. Just as the invention of the microscope extended biological inquiry into new arenas and enlarged the scope of questions that were reasonable to ask in the conduct of biological research, so will the computer. Computing and information technology will enable biological researchers to consider heretofore inaccessible questions, and as the capabilities of the underlying information technologies increase, such opportunities will continue to open up.

New epistemological questions will also arise. For example, as simulation becomes more pervasive and common in biology, one may ask, Are the results from a simulation equivalent to the data output of an experiment? Can biological knowledge ever arise from a computer simulation? (A practical example is the following: As large-scale clinical trials of drugs become more and more expensive, under what circumstances and to what extent might a simulation based on detailed genomic and pharmacological knowledge substitute for a large-scale trial in the drug approval process?) As simulations become more and more sophisticated, pre-loaded with more and more biological data, these questions will become both more pressing and more difficult to answer definitively.