Chapter 1

Introduction

1.1 A NEW TALE OR SAME STORY, DIFFERENT DAY?

In this day and age of informatic and '*omic* efforts within biology, experimentalists are challenged to exploit information systems and computational models within their research. Computational tools that are already commonly used include image analysis software, structural modeling programs, and sequence alignment tools. Mellman and Misteli (2003) suggest that it is time for computation to become recognized as a tool on a par with molecular tools in cell biology research. At the molecular level, one rarely questions if Web-based sequence alignment and database tools should be used but rather asks if they are being used efficiently and accurately to produce trustworthy results. At the cellular level, informatics tools are helpful in identifying putative molecular components and functions. However, the dynamic behaviors of cellular systems require the development of computational models (mathematical models). Mathematical models are becoming increasingly visible in the cell biology literature, and yet the methods of creating such models are less obvious to many of us trained as experimental cell biologists.

A number of opinion pieces have been published in journals commonly read by cell biologists—*Nature*, *Science*, *Journal of Cell Biology*, and *Cell*—that engage the question of what cell biologists are to do with the now-existent parts lists generated from the Human Genome Project and ongoing genomic and proteomic work (Bray, 1997; Hartwell *et al.*, 1999). Does having such an extensive catalogue of molecular data change our understanding of the nature of biology? Hartwell argued that biology may be better understood as modular. Others have described biology as an information science, relating to the implementation of instructions encoded in DNA and RNA and operationalized by protein, carbohydrate, and lipid machinery. These perspectives draw on lessons and advantages gained from engineering principles and information sciences. Whether

A Cell Biologist's Guide to Modeling and Bioinformatics. By Raquell M. Holmes

these perspectives are helpful frameworks for further understanding biology has yet to be determined. A common theme within these perspectives is an increased focus on the complex and formalized relationships between biological factors. These include evolutionary relationships that provide evidence for gene function and protein and network interactions that describe molecular circuitry underlying cellular processes and formalized kinetic relationships in the form of rate equations used to simulate dynamic behaviors.

Despite the recent fervor and attention to the benefits of computing in biology, the use of computational approaches in biology has existed for years although not always in association with computers. What we now commonly refer to as computational models appear historically as mathematical and theoretical models. As such, computational models in biology can be found as early as 1952 when Turing hypothesized short-range action, long-range inhibitor reaction diffusion as an explanation of pattern formation. This theory was later applied to shells, cheetahs, and drosophila (Nagorcka and Mooney, 1992). Membrane physiologists have used mathematical models to characterize membrane proteins such as pumps and channels; and the use of mathematical models and computation has led to better understanding of actin polymer dynamics, muscle contraction, and drosophila development (Julian, 1969; Pollard, 1986; Wachsstock and Pollard, 1994). Although computations have been part of our work, these aspects of our research have not been explicitly discussed as often by us as experimentalists. We are most familiar with mathematics in terms of probability associated with graphs and tables to demonstrate that our experimental results are not likely due to chance.

Recently, standard journals read by cell biologists—*Science*, *Nature*, *Journal of Cell Biology*—are publishing more papers that include computational analysis. Reviews discuss the importance of modeling and simulation in understanding the dynamics of cellular systems and how we approach biological research (Hartwell *et al.*, 1999). Research papers describe modeling results in an effort to better understand specific biological systems. Vesicular transport, membrane ruffling, and cell adhesion are just a few examples (Hirschberg *et al.*, 1998; Waterman-Storer and Danuser, 2002; Lee *et al.*, 2003). Within these papers, there is either an implicit or explicit reference to the importance of modeling.

1.2 COMPUTATIONAL BIOLOGY

Computational biology is a broad discipline, as broad as the numerous fields of biology and methods of computation. In its simplest description, computational biology is the use of computers and mathematics to solve problems within biology. Computation involves applying known and hypothesized relationships in mathematical form to the description of phenomena. The use of computational methods in biological research has been referred to in many ways, typically dependent on the biological focus and computational method of the speaker. Overall, the terms refer to the development and use of mathematical descriptions of a working hypothesis.

Bioinformatics Algorithms and database designs focused on molecular data or information management, sometimes including protein folding efforts.

Quantitative Biology Used in reference to various biological scales (e.g., molecular, cellular, tissue, etc.) and refers primarily to measuring quantities of biological factors. The quantitative data is then subjected to informatics efforts for cataloguing and is available for numerical modeling.

Systems Biology Complex systems exist within and across multiple biological scales. Systems biology has been defined as the determination of the components of these biological systems as well as the simulation of system behaviors with kinetic models in a given scale (Henry, 2003).

Computational Cell Biology Mathematical models of cellular systems including molecular motors, vesicle transport, cell signaling, and actin dynamics.

Bioinformatics is an aspect of computational biology that uses statistical approaches to model biological relationships (e.g., evolutionary traits, structural and gene regulatory networks). Broadly, bioinformaticians work with and develop computational methods and information management systems to discover biological principles. In practice, bioinformatics has developed the tools by which researchers can aggregate background information on already characterized genes and proteins as well as tools to predict genetic and biochemical networks across species.

A key characteristic of the ‘omic efforts is the need for high-throughput methods of generating data. The development of bioinformatics from genomics was largely due to the ability to mass-produce nucleic acid data for DNA mapping, gene identification, sequencing, and expression mapping. The great success of the genomics effort has led to the search for methods of generating large reproducible, reliable, and biologically relevant data sets for proteomics, metabolomics, cellomics, and physiomics. Researchers in these areas focus on engineering high-throughput data methods and computational tools to mine and analyze the data. ‘Omic research enables discovery science, where statistical models are used to identify patterns of biological significance from the data.

Molecular sequence data and computational tools are used to develop arguments of molecular identity, homology, and function that are based on evolutionary relationships and the presence of domain motifs/signatures, and 3D structures. DNA sequences of known genomes are used to search for and predict homologous or orthologous biological functions in species where they have yet to be identified through experimentation. Appropriate annotations are sought for genes in the databases, and researchers attempt to infer function based on sequence and expression data. This includes determining participation in genetic or metabolic networks. Computational tools are employed to (1) draw comparisons between DNA sequences, gene structures, and determine possible evolutionary relationships, (2) predict biochemical properties of proteins (i.e., protein folds, protein binding sites, and posttranslational modifications), and (3) predict the functional role of genes in cellular or physiological processes.

Molecular sequence databases are useful for determining what is known experimentally or predicted computationally about the molecular components of biological systems and their biological function. Sequence similarity searches are used to infer identity and biochemical properties of a novel cDNA clone based on already characterized genes and proteins. Protein family and domain databases use sequence and structure data to construct family assignments or characteristic patterns for functional domains. Databases of protein-protein interactions provide evidence for *in vivo* interactions and participation in biochemical pathways. Protein interaction data, microarrays, and protein profiles are subjected to cluster analyses to infer coregulated genes. Genomic comparisons are used to identify conserved pathways between species.

The challenge that we face as experimentalists is how to isolate relevant information from the large pool of data.

Which data resource has the type of data one is looking for?

Is the data within the database trustworthy?

How does one know if one's search results are statistically and, importantly, biologically significant?

In the following chapters on sequence alignment and protein family and domain databases, we provide initial insights to answer these questions. More in-depth answers and direction can be found in books and resources dedicated to informatics discussion and training. The treatments in this book provide background information and directions that can be useful to researchers as they engage with more commonly used Web-based molecular research tools.

1.3 MODELING AND SIMULATION

As the use of information systems increases in biological research, allowing us both to catalogue and search large amounts of data, biologists are faced with a new opportunity to study complex relationships that were previously not feasible. We are challenged to create models as a means of (1) making explicit the understood functional relationship between biological entities (i.e., protein, nucleic acid, or organelle), (2) finding fault with a hypothesis, and (3) providing colleagues an additional means of testing our results.

Experimental biologists typically use conceptual models to illustrate current hypotheses or understandings of cellular systems. These explanations are usually provided in the form of illustrations, information flow charts, and diagrams. These diagrams replace the need for developing paragraphs of text to define the relationships among biological factors within any system. However, it is impossible to use these diagrams to test the hypothesis. In a computational model of dynamic systems, the relationships among factors in a diagram are made explicit by defining the relationships in terms of rates, quantities, or state changes. This transforms conceptual models into *working hypotheses*.

The phrase *mathematical description* can invoke images of a series of equations consisting of unfamiliar symbols or incomprehensible numbers. It is helpful to state here that generating a mathematical description is not necessarily about numbers or exact quantities but rather about formalizing the relationships between biological objects such that the relationship itself and the product of the relationship can be tested. Developing models therefore is the mathematical assertion of a hypothesis about our experimental system in a testable form. Transforming our hypothesis into explicit relationships and assumptions is a rigorous reflective process that helps to expose missing components and inconsistencies.

The topics discussed in this book might be considered the basis for computational cell or systems biology. Molecular databases provide information on cellular components, and mathematical models provide a means of studying dynamic properties of biochemical reactions. The computational methods differ for informatics and dynamics modeling. For instance, sequence alignment and search algorithms use statistical models to determine the significance of an alignment. This is distinct from the hidden Markov models used to calculate protein domain motifs, graphical models for network inference, differential equations used to simulate dynamical systems, or Monte Carlo simulations of stochastic processes.

1.3.1 From Databases to Dynamics: A View of Network and Pathways

The terms *cell signaling*, *pathways*, and *networks* are used to describe a series of protein interactions that are the basis of dynamic, observable cellular behaviors. Intricate sets of interactions have been created from the examination of interactions between subsets of proteins. For example, we investigate growth factor stimulation of a signaling pathway by examining changes in the state of one or more downstream molecules (e.g., cell adhesion activation of PI3 kinase; Epidermal Growth Factor (EGF)-triggered phosphorylation of Mitogen Activated Protein Kinase (MAPK)). Changes in the phosphorylation state or localization of a protein can serve as the biomarker for the activation of the pathway. This biomarker is then used to identify additional players. The reduction and subsequent reconstruction of the complex system of interactions is ideal for experimentation. However, the properties of the system are not reflected in the discussion of the details or parts of the system but rather only in the discussion of the system as a whole (*Nature* 403: 345–346). The network is a view of the global aspects of the system. A distinction between pathways and networks then is the level of abstraction.

To infer of gene and biochemical networks, researchers use information science, statistics, and graph theory to integrate data and elucidate complex biological relationships. They take advantage of data from genomic and gene and protein sequence databases to map newly identified sequences onto preexisting, already established networks or to predict new interaction networks. Network and pathway databases like molecular sequence databases catalogue information. They differ from the sequence database by focusing on the interactions between genes and proteins. Kyoto Encyclopedia of Genes and Genomes (KEGG) maps known metabolic pathways in yeast and computes similar pathways in other species (Table 1.1; Goto *et al.*, 1997; Ogata *et al.*, 1999; Kanehisa *et al.*, 2006). EcoCyc contains *Escherichia coli* metabolic and some signaling pathways and is the metabolism template for predicting metabolic pathways in other species (Karp *et al.*, 1999). Biomolecular Interactions Network Database (BIND), now encompassed by BondPlus and IntAct developed by the European Bioinformatics Institute, catalogues protein-protein interactions (Hermjakob *et al.*, 2004; Unleashed Informatics, Ltd., 2006). Collectively, these databases gather and organize data on networks and pathways, as well as provide computational tools that primarily use binary relationships to predict metabolic pathways in species where they have yet to be confirmed experimentally.

When examining networks and pathways, there are a series of questions that can be posed with which computational and bioinformatics tools can help. These include: What genes or proteins are involved? What are their functions? What other pathways or networks are they involved in? Is the associated cellular behavior dependent on the concentration or location of the factor?

TABLE 1.1 Gene and Protein Interaction—Network Databases

Database	Type of Data	Primary Species	Data Sources
BIND	Protein interactions	Yeast	
KEGG	Gene and metabolic networks	Yeast	WIT, LIGAND, Enzyme handbook, Japanese catalogue
EcoCyc	Metabolic pathways	*E. coli*, coding DNA only	Genbank, Enzyme, primary literature
CSNDB	Cell signaling	Human	Transfac, journal literature

TABLE 1.2 Computational Resources for Addressing Biological Questions

Question	Computational Resources
What are the functions of the proteins?	Molecular sequence, family, and domain databases
What proteins are involved in a given pathway?	Genetic and metabolic pathway database
What is the dynamic behavior of protein interactions?	Computational simulation
How are dynamics of cellular behaviors affected by changes in molecular concentrations and kinetics?	Computational simulation

These questions can be addressed by different bioinformatics and computational resources (Table 1.2). The database tools and graphical maps are helpful toward understanding the protein components and interaction flows in a pathway, and mathematical modeling tools enable us to examine and better understand the behavior of the pathways and networks.

Mathematical models take many forms. The models discussed in this book are systems of differential equations also known as continuous or population models. These equations are solved numerically by providing numerical values for concentrations, rates of reactions, diffusion rates, and binding constants. Quantitative data for some of these values exists in literature and databases, however many are missing for a large fraction of known proteins and enzymes. National Institutes of Health—funded efforts of the Alliance for Cellular Signaling (AFCS) and the National Technology Centers for Networks and Pathways have focused on the development and use of methods to obtain this quantitative data. The premise is that by obtaining quantitative data within biological systems, it will be possible to model the dynamics of cellular networks and pathways and thus predict the behavior of a system.

BIBLIOGRAPHY

Alfarano C, Andrade CE, Anthony K, *et al*. (2005). The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Research* 33(Database issue):D418–D424.

Anonymous (2000). Can biological phenomena be understood by humans? *Nature* 403(6768):345.

Baldi P, Brunak S (1998). *Bioinformatics: The Machine Learning Approach*. Cambridge: MIT Press; 351p.

Bray D (1997). Reductionism for biochemists: how to survive the protein jungle. *Trends in Biochemical Sciences* 22(9):325–326.

Goto S, Bono H, Ogata H, *et al*. (1997). Organizing and computing metabolic pathway data in terms of binary relations. *Pacific Symposium on Biocomputing* 2:175–186.

Hartwell LH, Hopfield JJ, Leibler S, *et al*. (1999). From molecular to modular cell biology. *Nature* 402:C47–C52.

Henry CM (2003). Systems biology. *Chemical and Engineering News* 81(20):45–55.

Hermjakob H, Montecchi-Palazzi L, Lewington C, *et al*. (2004). IntAct: an open source molecular interaction database. *Nucleic Acids Research* 32:D452–D455.

Hirschberg K, Miller CM, Ellenberg J, *et al*. (1998). Kinetic analysis of secretory protein traffic and characterization of Golgi to plasma membrane transport intermediates in living cells. *Journal of Cell Biology* 143:1485–1503.

Julian FJ (1969). Activation in a skeletal muscle contraction model with a modification for insect fibrillar muscle. *Biophysical Journal* 9:547–570.

Kanehisa M, Goto S, Hattori M, *et al*. (2006). From genomics to chemical genomics: new developments in KEGG. *Nucleic Acids Research* 34:D354–357.

Karp PD, Riley M, Paley SM, *et al*. (1999). EcoCyc: an encyclopedia of Escherichia coli genes and metabolism. *Nucleic Acids Research* 27:55–58.

Lee K, Dinner AR, Tu C, *et al*. (2003). The immunological synapse balances T cell receptor signaling and degradation. *Science* 302:1218–1222.

Lippincott-Schwartz J, Snapp E, Kenworthy A (2001). Studying protein dynamics in living cells. *Nature Reviews* 2:444–456.

Mellman I, Misteli T (2003). Computational cell biology. *Journal of Cell Biology* 161:463–464.

Nagorcka BN, Mooney JR (1992). From stripes to spots: prepatterns which can be produced in the skin by a reaction-diffusion system. *IMA Journal of Mathematics Applied in Medicine and Biology* 9(4):249–267.

Ogata H, Goto S, Sato K, *et al*. (1999). KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Research* 27:29–34.

Pollard TD (1986). Mechanism of actin filament self-assembly and regulation of the process by actin-binding proteins. *Biophysiology Journal* 49:149–151.

Regev A, Shapiro E (2002). Cellular abstractions: cells as computation. *Nature* 419:343.

Rives AW, Galitski T (2003). Modular organization of cellular networks. *Proceedings of the National Academy of Sciences USA* 100(3):1128–1133.

Setubal JC (1997). *Introduction to Computational Molecular Biology*. Boston: PWS Publishing; 296p.

Sibley CG (1997). Proteins and DNA in systematic biology. *Trends in Biochemical Science*s 22(9):364–367.

Takai-Igarashi T, Kaminuma T (1998). A pathway finding system for the cell signaling networks database. *In Silico Biology* 1:129–146.

Takai-Igarashi T, Nadaoka Y, Kaminuma T (1998). A database for cell signaling networks. *Journal of Computational Biology* 5(4):747–754.

Tass PA (1999). *Phase Resetting in Medicine and Biology: Stochastic Modelling and Data Analysis*. New York: Springer-Verlag; 329p.

Turing AM (1952). The chemical basis of morphogenesis. *Philosophical Transactions of the Royal Society (B)* 237:37–72.

Unleashed Informatics Ltd. BOND. Available at http://bond.unleashedinformatics.com/index.jsp?pg=0.

Wachsstock DH, Pollard TD (1994). Transient state kinetics tutorial using the kinetics simulation program, KINSIM. *Biophysical Journal* 67:1260–1273.

Waterman-Storer CM, Danuser G (2002). New directions for fluorescent speckle microscopy. *Current Biology* 12:R633–R640.