

Preface

In its report, *A New Biology for the 21st Century*,¹ the National Research Council defines the essence of the New Biology as “...re-integration of the many sub-disciplines of biology, and the integration into biology of physicists, chemists, computer scientists, engineers, and mathematicians to create a research community with the capacity to tackle a broad range of scientific and societal problems.” The report stipulates that “...the emergence of the New Biology signals the need for changes in how scientists are educated and trained” and calls for substantive changes in interdisciplinary education at the junction of mathematics and biology at both the undergraduate and graduate levels. This report echoes many of the recommendations of an earlier influential report, *Bio 2010*,² that “...each institution of higher education reexamine its current curricula...” and concludes that “...College and university administrators, as well as funding agencies, should support mathematics and science faculty in the development or adaptation of techniques that improve interdisciplinary education for biologists.”

Due to the high profiles of these reports, it is now widely accepted that a main push in biology during the coming decades will be toward an increasingly quantitative understanding of biological functions, and that the new generation of biologists will routinely use mathematical models and computational approaches to frame hypotheses, design experiments, and analyze results. A 2010 Society for Industrial and Applied Mathematics (SIAM) white paper, *Mathematics: An Enabling Technology for the New Biology*, further underscores the critical role that mathematicians and statisticians are asked to play toward accomplishing the New Biology’s aims. This white paper also recommends increased federal support to ensure a pipeline of such adequately trained professionals, starting at the undergraduate level.

It is thus critically important that the training of the new biologists and their collaborators, whether coming through biology or through other areas of the natural and mathematical sciences, facilitates access to a rich toolbox of diverse mathematical approaches. New educational guidelines and recommendations linked with the reports above and with others,³ have catalyzed various educational discussions and curricular changes. In particular, in the past few years the number of undergraduate and graduate programs in mathematical and computational biology has increased, institutions have added new courses in mathematical biology linked with ongoing research in biology, and the American Mathematical Society, the Mathematical Association of America (MAA), The National Science Foundation (NSF), the National Institutes of Health, and the NSF Mathematical Sciences Institutes are funding faculty

¹A New Biology for the 21st Century (2009). The National Academies Press, Washington, DC.

²BIO2010: Transforming Undergraduate Education for Future Research Biologists (2003). The National Academies Press, Washington, DC.

³See, e.g., Vision and Change in Undergraduate Biology Education: A Call to Action (2009). AAAS, Washington, DC.

development workshops, research-related experiences, and specialized research conferences in mathematical biology for undergraduates.

However, while traditional mathematical biology topics using difference equations, differential equations, and continuous dynamical systems have to a large extent worked their way into the classroom and have become standard curriculum, mathematical techniques from modern discrete mathematics (encompassing traditional discrete mathematics with combinatorics and graph theory, as well as linear algebra, algebraic geometry, and modern abstract algebra) have remained relatively invisible in these curricular changes. The 2010 SIAM white paper cited above calls for increased support in a number of mathematical subfields with strong ties to modern discrete mathematics, as there is mounting evidence that novel algebraic methods are being used with great success in current mathematical biology research. These include Boolean networks, finite/polynomial dynamical systems (including many agent-based models), elements of graph theory, Petri nets, and Gröbner (Groebner) bases and other elements from algebraic geometry and modern algebra. In spite of their accessibility to undergraduates, these topics are almost entirely absent from the undergraduate mathematical biology training landscape. Thus, while novel applications of theories from modern discrete mathematics are finding increasing use in the rapidly evolving field of mathematical biology, the already existing gap between research and education is growing wider, particularly in the area of undergraduate education. While students interested in mathematical biology have relatively easy access to courses that utilize analytic methods, and generally have an adequate exposure to such methods before deciding upon a graduate program, students interested in learning about modern discrete mathematical approaches to mathematical biology topics have fewer doors visibly open to them, and indeed may not even know such approaches exist. Faculty who want to teach courses utilizing differential equations models now have ready access to a fair number of texts and textbook resources (including the textbook *An Invitation to Biomathematics* by Robeva *et al.* published by Elsevier in 2008) focusing primarily on the use of analytic mathematical methods in biology. In contrast, materials applying modern discrete mathematical methods in biology are generally widely scattered, and, outside of a select set of topics,⁴ there are practically no educational resources reflecting the importance of algebraic methods in many of the fast-growing areas of mathematical biology. In the cases when sources for the latter are available (the 2005 text *Algebraic Statistics for Computational Biology* by Pachter and Sturmfels, published by Cambridge University Press, is an important example), the level of presentation is not necessarily aimed at the true beginner and may be more appropriate for graduate level training.

⁴E.g., some aspects of phylogenetics as they appear in portions of *Mathematical Biology* by Allman and Rhodes or more specialized texts like Semple and Steel's book *Phylogenetics* (Oxford University Press, 2003), Felsenstein's *Inferring Phylogenies* (Sinauer Associates, 2003), as well as combinatorial mathematics in Waterman's *Introduction to Computational Biology* (Chapman and Hall/CRC, 1995; second edition coming out in 2012).

We hope that our volume *Mathematical Concepts and Methods in Modern Biology: Using Modern Discrete Methods* will bring undergraduate students (and faculty interested in teaching them) face-to-face with more applications of modern discrete mathematics to biology. In its choice of topics and style of approach, this volume is not intended to be a comprehensive treatment of all current uses of modern discrete methods in biology, but to provide passageways to a diverse and expansive landscape. Consequently, the collection of chapters comprising the volume are designed to be largely independent from one another and can be viewed as “modules” for classroom use, as independent studies, as starting points for undergraduate research projects, or even as gentle entryways for more mathematically oriented readers. Each chapter begins with a question from modern biology, followed by the description of certain mathematical methods and theory appropriate in the search for answers. As such, the chapters can be viewed as fast-track pathways through the problem that begin by laying out the biological foundation, proceed by covering the relevant mathematical theory, and end by highlighting connections with ongoing research and current publication.

Multiple exercises and projects are embedded within the chapters, giving instructors the flexibility to cover material only up to a certain point and ignore later sections that may require higher mathematical sophistication. Embedding the exercises ensures that only material which has already been covered is needed for their execution. Many of the projects and exercises utilize specialized software, exemplifying the notion that familiarity and experience with computing applications which implement the mathematical theory are critical elements of the “modern biology” skills set. We have been particularly mindful of designing the exercises in a way that requires only the use of freely-available applications or mainstream proprietary software that is commonly available on college and university campuses (e.g., MATLAB).

Even though the chapters are to a large extent independent and self-contained, they are grouped, wherever appropriate, by common biological or mathematical threads. They are not organized by level of mathematical difficulty. A chapter appearing later in the volume should not be assumed to require a higher level of mathematical prerequisites. However, when the chapters consider similar biological questions or make use of the same mathematical theory, earlier chapters will usually contain more introductory details. In this sense, it would be beneficial to cover Chapters 1–3 in this order, as Chapters 2 and 3 expand upon the mathematical foundation presented in Chapter 1. We recommend the same for the following clusters: Chapters 4 and 5; Chapters 7 and 8, and (perhaps to a lesser degree) Chapters 9 and 10. Chapter 6 is self-contained. The highest level of mathematical proficiency reached in each chapter may vary significantly from topic to topic. The list below presents a brief summary of the chapters’ topics, highlights the assumed mathematical background for each chapter, and provides information regarding possible course adoptions and use of specialized software.

Chapter 1. Mechanisms of Gene Regulation: Boolean Network Models of the Lactose Operon in *Escherichia coli*, by Raina Robeva, Bessie Kirkwood, and Robin Davies.

The transcription of genes (mRNA synthesis) and translation of mRNA (protein synthesis) are energetically expensive processes and cells have the ability to make certain proteins only when the environmental conditions warrant. Otherwise, if a cell had to make all of its proteins all of the time, it would be expending a lot of cellular energy in the making of proteins for which it has no use. Understanding the relevant mechanisms of gene expression, controlled via so-called gene regulatory networks, is thus critically important to understanding the regulation of cellular behavior. The lactose (*lac*) operon is a relatively simple but important example of a gene regulatory network for the metabolism of lactose in the bacterium *E. coli*. Since its discovery in the late 1950s, the *lac* operon has served as a model system for understanding many aspects of gene regulation.

The chapter is an introduction to mathematical modeling with Boolean networks in the context of gene regulatory networks, using the *lac* operon as a main example. Students who are prepared mathematically to enroll in a discrete mathematics course can read this chapter and work through all exercises. No specific mathematical background is required, as the chapter includes a primer on Boolean arithmetic. All substantive computations beyond the initial introductory exercises are done using the web-based suite *DVD*, which is freely available. Even though no prior knowledge of modern algebra is required, students enrolled in an undergraduate modern algebra course that covers algebraic rings and ideals can use elements of the chapter to introduce and motivate the question of solving polynomial systems of equations and the connections with Groebner bases of polynomial ideals. The chapter provides an on-line appendix on using Groebner bases for solving systems of polynomial equations.

Chapter 2. Bistability in the Lactose Operon of *Escherichia coli*: A Comparison of Differential Equation and Boolean Network Models, by Raina Robeva and Necmettin Yildirim.

Bistability is the ability of a system to achieve two different steady states under the same external conditions. The *lac* operon of *E. coli* is a bistable system: under certain external conditions, the *lac* operon may be turned on or turned off depending on the history of the cell (determined by the environmental conditions under which it has been grown). The chapter introduces several ordinary differential equation (ODE) models of the *lac* operon and their Boolean network analogs and compares these two types of models with regard to their ability to capture the bistable behavior of the *lac* operon system. The ODE and Boolean parts of the chapter could be considered independent if the reader would be willing to accept the ODE models without justification. Some of the exercises related to the ODE models require MATLAB, while the Boolean networks are analyzed using *DVD*, as in Chapter 1. The first part of the chapter is appropriate as an introduction to the modeling of biochemical reactions in differential equations courses, while the second part is appropriate for courses in discrete mathematics. The entire chapter can be used in a mathematical biology course, or as a student research project to highlight connections between abstract algebra and differential equations in the context of gene regulation.

Chapter 3. Inferring the Topology of Gene Regulatory Networks: An Algebraic Approach to Reverse Engineering, by Brandilyn Stigler and Elena Dimitrova.

Key features of gene regulatory networks can be represented diagrammatically through graphs whose nodes are genes or gene-related products, and whose interactions are, at least partially, captured through certain types of edges. The topology of a gene regulatory network is the essential shape of this graph. It is a very important and difficult biological task to try, from knowing only partial information (generally observed only during snapshots in time) about the expressions of genes or gene products, to infer this topology, and hence discover the relationships (edges) among the nodes.

This chapter uses aspects of the algebra of polynomials to recreate such networks from time series data when the levels of gene or gene product expression can be captured by a finite number of states. Such systems generalize the Boolean models treated previously in Chapters 1 and 2. At their most elementary level, they can be approached through elementary multivariable polynomials by a reader familiar with modular arithmetic (an approach also taken initially in Chapter 5). The presentation in Chapter 3 generally assumes familiarity with elementary modern algebra at the level of rings and ideals and is appropriate for an undergraduate modern algebra course. Some advanced topics such as the Chinese remainder theorem for rings, the ideal-variety correspondence of algebraic geometry, primary decomposition of ideals, and Jacobson radical of an ideal make an appearance, but one need not be familiar with these more advanced concepts in order to work through the entire chapter. Some exercises do require the reader to compute the intersection of ideals in a polynomial ring, the primary decomposition of an ideal, and the Jacobson radical of an ideal, but readers with only an elementary background in modern algebra (and, just as well, those with more experience!) can perform these computations quite easily using the freely available computational algebra system *Macaulay 2*.

Chapter 4. Global Dynamics Emerging from Local Interactions: Agent-Based Modeling for the Life Sciences, by David Gammack, Elsa Schaefer and Holly Gaff.

Biological research into areas as widely varied as the population dynamics of prairie dog colonies and tick populations, bird flocking and evolutionary patterns, impact of individual behavioral choices on important societal problems, disease spreading, and blood vessel growth and leukocyte rolling, has been pursued through the use of scientific models that are agent-based. This chapter is an introduction to agent-based (also called individual-based) modeling through *Netlogo* (available for free download). It does not require any mathematical background except for some very elementary probability and provides the reader, through a rich set of hands-on exercises, with the opportunity to observe how the global behavior of a complex system of interacting “agents” arises from the local rules established for their interactions. The examples and projects presented in the chapter cover a wide range of models and topics, from basic classroom illustrations to models being used in ongoing research, including the following agent-based models that are examined and analyzed in detail: a model of axon guidance, a model for the spread of cholera, and two models describing the dynamics of tick-borne diseases. The chapter would be useful for mathematical modeling classes and in introductory programming classes.

Chapter 5. Agent-based Models and Optimal Control in Biology: A Discrete Approach, by Reinhard Laubenbacher, Franziska Hinkelmann, and Matt Oremland.

In this chapter, a wide class of agent-based models is investigated through several concrete examples and captured mathematically as polynomial dynamical systems over finite fields. This approach uses multivariable polynomials to represent the transitions between agents' states in time and polynomial functions to encode the dynamics of the entire system. It provides a broad mathematical framework for analyzing agent-based models, finding the long-term dynamic behavior of the systems, and implementing optimal control strategies.

The first seven sections of the chapter require very little mathematical background, although the first section would be best understood by a reader with some background in elementary differential equations. Section 8 can serve as a brief introduction to finite fields and to polynomial dynamical systems over finite fields (introduced in Chapter 3 as well). No modern algebra is required as a prerequisite here, though it would certainly be helpful. This section could also be used as motivation to learn more about polynomial rings and ideals over finite fields. The last section is more advanced and would be most appropriate for use in modern algebra courses or with students who have had a proof-based course in discrete mathematics and/or are engaged in student research.

Many of the chapter examples and one of the chapter exercises require the use of *Netlogo*. The web-based and freely-available application suite ADAM is used for obtaining and visualizing the characteristics of polynomial dynamical systems.

Chapter 6. Neuronal Networks: A Discrete Model, by Winfried Just, Sungwoo Ahn, and David Terman.

It is commonly believed that everything the brain does is the result of the collective electrical activity of neurons. Neurons communicate with other neurons by synaptic connections forming complex neuronal networks. Simple discrete dynamical system models of neuronal dynamics can be constructed by assuming that at any given time step each neuron can either fire or be at rest, that after it has fired each neuron needs to be at rest for a specified refractory period, and that the firing of a neuron is induced by firing of a sufficient number of other neurons with synaptic connections to it.

This chapter explores the relationship between the network connectivity and important features of the network dynamics such as the number and lengths of attractors, lengths of transients, and sizes of the basin of attraction. A variety of mathematical tools, ranging from combinatorics to probability theory, are used. The chapter also discusses some issues involved in choosing the appropriate model for a given biological system, including a result on the relation between the discrete dynamical systems models introduced in the chapter and certain more detailed ODE models. For the first four sections, students should have some experience with elementary notions of discrete mathematics such as the greatest common divisor, modular arithmetic, and the floor function at the level of writing proofs. Familiarity with graph theory would be beneficial, but is not required. Sections 5 and 6 require basic background in discrete probability. The material would be most appropriate for

courses that assume proof-based discrete mathematics as a prerequisite. Some basic knowledge of ordinary differential equations is assumed in section 7. Online supplemental material containing extensions of the mathematical theory and providing a number of additional projects and exercises is also included. Use of MATLAB is suggested for some exercises and projects, and specialized MATLAB code is made available as part of the online supplement.

Chapter 7. Predicting Population Growth: Modeling with Projection Matrices, by Janet Steven and James Kirkwood.

In many models of population growth, life stages are defined based on morphological changes during growth, or changes in size. In some organisms, development leads to natural categories; seeds, seedlings, and reproductive plants, for example, or egg, larva, pupa, and adult in butterflies. In other organisms, sometimes it makes more sense to categorize individuals on the basis of age. Matrix algebra is often used to build models that incorporate the different stages an organism goes through during its life. The model can then be used to predict both the overall growth of the population and the distribution of individuals across these life stages.

The first several sections of the chapter provide an introduction to the modeling of population dynamics with projection matrices, through segmentation into various life stages. For these sections, only the very basics of matrix algebra are required (e.g., matrix notation, matrix multiplication, vectors), and concrete applications to a ginseng population are explored. Section 8 and beyond use linear algebra (eigenvalues and eigenvectors) to determine the steady-state stage distribution of a population. Familiarity with elementary linear algebra is a necessary prerequisite for these later sections. The chapter provides MATLAB and R commands for performing the necessary matrix operations, but *GNU Octave* can be used as a free alternative. Graphing calculators (e.g., the *TI-89*) may also be used to perform the calculations. Early material would be appropriate for any course introducing basic matrix theory, while the later material would be appropriate for linear algebra courses and could be used to demonstrate an important application of eigenvectors.

Chapter 8. Metabolic Pathways Analysis: A Linear Algebraic Approach, by Terrell L. Hodge.

At the cellular level, metabolic processes are biochemical reaction systems that enable a cell to extract energy and other necessities for life from nutrients, and to build new structures it needs to live and reproduce. The chains of biochemical reactions involved are called metabolic pathways, and the manipulation of them, and the complex networks into which they fit, is the domain of metabolic engineering. In this chapter, the underlying pathways and networks of metabolism are modeled mathematically through the use of matrix analysis and linear algebra associated to these systems of biochemical reaction equations. The initial material can be used to motivate the basics of matrix representations of linear equations, and the remainder fits well into a course covering the fundamentals of linear algebra, including analyzing null spaces, interpreting linear independence, bases, and more. Graphing calculators or standard mathematics software may be used to carry out calculations. A tutorial for a freely downloadable package *ExPA* appears in the supplementary materials.

Chapter 9. Identifying CpG Islands: Sliding Window and Hidden Markov Model Approaches, by Raina Robeva, Aaron Garrett, James Kirkwood, and Robin Davies.

In the strings of adenine (A), cytosine (C), guanine (G), and thymine (T) out of which DNA is formed, the dinucleotide CG appears with a probability that differs notably from what naïve randomness would predict. Regions with relatively low frequencies of the CG nucleotide contain clusters, known as “CpG islands,” within which the CG content is much higher. CpG islands are often associated with the promoter regions of genes. Methylation of these promoter islands is associated with the transcriptional silence of the gene while promoter-associated CpG islands in constitutively-expressed housekeeping genes are unmethylated. Inappropriate methylation of the CpG islands in tumor suppressor promoters has been associated with the development of numerous human cancers. Thus, identifying the locations of CpG islands in DNA sequences is an important task.

In this chapter, a heuristic model for locating CpG islands using sliding windows is briefly introduced, followed by mathematical methods based on hidden Markov models. Familiarity with discrete probability (e.g., conditional probability, independence, geometric distribution) and finite Markov chains is assumed for the whole chapter, although a brief refresher on Markov chains is included. Many introductory and intuitive examples are included in order to illustrate the nature of hidden Markov models and their application as modeling tools for locating CpG islands in the genome. The natural place for the material would be in a discrete probability course, but the chapter can also be used in computer science courses since it covers decoding and training algorithms. The companion suite of freely-available web-based software applications *CpG Educate* is utilized for many of the chapter projects and exercises. The chapter includes an online project “Investigating Predicted Genes” appropriate for biology courses with no mathematics prerequisites.

Chapter 10. Phylogenetic Tree Reconstruction: Geometric Approaches, by David Haws, Terrell Hodge, and Rudy Yoshida.

Comparing the DNA sequences of individual species or groups of related species can often provide essential insights into evolutionary biology. This chapter’s topic is the recovery of the evolutionary history of gene families, species, or other levels of biological organisms by means of phylogenetic trees, easily pictured as the equivalent of “family trees” but created only from DNA sequence data of the “family” members alive today, with no prior knowledge of their ancestors and their relationships. Reconstructing the evolutionary history of genes or organisms, based on molecular and genetic data, has a multiplicity of modern applications. The most obvious and historically revolutionary application is the classification of species and organisms not by their outward looks (classical taxonomy via morphology), but by their genetic similarities. Tracing the evolutionary history of genetic data has also informed our understanding of human and animal population movements across the globe over generations and millennia. In addition, phylogenetic tree reconstruction makes it possible to track, prepare for, and try to attack outbreaks of disease, such as HIV or the flu. As another important outcome, knowledge of phylogenetic trees has made it possible to reconstruct, e.g., biochemically recreate, potential ancestors

of genes and to then use these ancestors to test hypotheses about their roles in the evolution of traits.

Through the study of a subset of tree reconstruction methods, the “distance-based” methods, such phylogenetic trees are represented as points in a high-dimensional real vector space, and the process of finding of a good tree that fits the real-world sequence data is treated as a geometric projection in this space. Freely accessible on-line programs are used to illustrate phylogenetic trees and implement some tree reconstruction methods. The first section can be used early in an elementary discrete mathematics or linear algebra course to introduce elementary matrix notation, basics on trees (as graphs), and high-dimensional spaces, in a biologically relevant context. Later sections explore two key distance-based tree reconstruction methods, and the relationship between them, through geometric structures in the aforementioned space, including cones and the use of linear optimization over a convex polytope whose vertices correspond to certain phylogenetic trees.

For the book as a whole, supplemental materials, including online projects, software and data files, appendices and extensions of the chapter materials, are gathered together and are available from the volume’s web site (<http://booksite.elsevier.com/9780124157804>). Complete solutions to the chapter exercises and guidelines for the projects are also available from the website.

The materials authored or co-authored by Robeva and Hodge grew out of a set of educational modules based upon work supported by the NSF under grant DUE-0737467.⁵ This material has been tested in multiple classroom settings at Sweet Briar College and at Western Michigan University. Those materials were also used in the faculty professional development PREP workshops “Mathematical Biology: Beyond Calculus” sponsored by the MAA (under NSF grant DUE-0817071) and offered in 2010 and 2011 at Sweet Briar College. We greatly appreciate these organizations’ support.

We express our sincere gratitude to all of the authors who contributed their excellent work to this volume. We thank our wonderful editorial team at Elsevier and specifically our project managers, Catherine (Cassie) Mullane and Julia Haynes, and editor, Christine Miniham. We are particularly indebted to our former Elsevier editor, Patricia Osborn, who embraced this project early on, encouraged us to pursue it, and invested many hours of her time at the planning stages to ensure the publication of this collection. Robert Kipka at Western Michigan University was indispensable during the book’s initiation and editing stages and we thank him warmly for his time and dedication. Finally, we thank our husbands, Boris Kovatchev and Robert McNutt, for their patience and support throughout this process.

*Raina S. Robeva
Terrell L. Hodge
August 20, 2012*

⁵Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.