ALGEBRAIC AND DISCRETE MATHEMATICAL METHODS FOR MODERN BIOLOGY

RAINA ROBEVA, EDITOR

Algebraic and Discrete Mathematical Methods for Modern Biology

Edited by

Raina S. Robeva
Department of Mathematical Sciences, Sweet Briar College, Sweet Briar, VA, USA

ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON
NEW YORK • OXFORD • PARIS • SAN DIEGO
SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO
Academic Press is an imprint of Elsevier

Academic Press is an imprint of Elsevier
32 Jamestown Road, London NW1 7BY, UK
525 B Street, Suite 1800, San Diego, CA 92101-4495, USA
225 Wyman Street, Waltham, MA 02451, USA
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK

First edition 2015

Notices
Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

Library of Congress Cataloging-in-Publication Data
A catalog record for this book is available from the Library of Congress

British Library Cataloguing in Publication Data
A catalogue record for this book is available from the British Library

For information on all Academic Press publications
visit our website at http://store.elsevier.com/

Printed and bound in the USA

ISBN: 978-0-12-801213-0

Algebraic and Discrete Mathematical Methods for Modern Biology

Contributors

Numbers in parentheses indicate the pages on which the authors' contributions begin.

Réka Albert (65), Pennsylvania State University, University Park, PA, USA

Todd J. Barkman (261), Department of Biological Sciences, Western Michigan University, Kalamazoo, MI, USA

Mary Ann Blätke (141), Otto-von-Guericke University, Magdeburg, Germany

Hannah Callender (193,217), University of Portland, Portland, OR, USA

Margaret (Midge) Cozzens (29), Rutgers University, Piscataway, NJ, USA

Kristina Crona (51), Department of Mathematics and Statistics, American University, 4400 Massachusetts Ave NW, Washington, DC 20016

Robin Davies (93,321), Department of Biology, Sweet Briar College, Sweet Briar, VA, USA

Monika Heiner (141), Brandenburg University of Technology, Cottbus, Germany

Terrell L. Hodge (261), Department of Mathematics, Western Michigan University, Kalamazoo, MI, USA

Qijun He (93,321), Department of Mathematical Sciences, Clemson University, Clemson, SC, USA

John R. Jungck (1), Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA

Winfried Just (193,217), Department of Mathematics, Ohio University, Athens, OH, USA

Bessie Kirkwood (237), Sweet Briar College, Sweet Briar, VA, USA

M. Drew LaMar (193,217), The College of William and Mary, Williamsburg, VA, USA

Matthew Macauley (93,321), Department of Mathematical Sciences, Clemson University, Clemson, SC, USA

Wolfgang Marwan (141), Otto-von-Guericke University, Magdeburg, Germany

David Murrugarra (121), Department of Mathematics, University of Kentucky, Lexington, KY, USA

Christian M. Reidys (347), Department of Mathematics and Computer Science, University of Southern Denmark, Odense M, Denmark

Raina Robeva (65), Sweet Briar College, Sweet Briar, VA, USA

Janet Steven (237), Christopher Newport University, Newport News, VA, USA

Blair R. Szymczyna (261), Department of Chemistry, Western Michigan University, Kalamazoo, MI, USA

Natalia Toporikova (193), Washington and Lee University, Lexington, VA, USA

Alan Veliz-Cuba (121), Department of Mathematics, University of Houston, and Department of BioSciences, Rice University, Houston, TX, USA

Rama Viswanathan (1), Beloit College, Beloit, WI, USA

Grady Weyenberg (293), Department of Statistics, University of Kentucky, Lexington, KY

Emilie Wiesner (51), Department of Mathematics, Ithaca College, 953 Danby Rd, Ithaca, NY 14850

Ruriko Yoshida (293), Department of Statistics, University of Kentucky, Lexington, KY

Preface

In the last 15 years, the field of modern biology has been transformed by the use of new mathematical methods, complementing and driving biological discoveries. Problems from gene regulatory networks and genomics, RNA folding, infectious disease and drug resistance modeling, phylogenetics, and ecological networks and food webs have increasingly benefited from the application of discrete mathematics and computational algebra. Modern algebra approaches have proved to be a natural fit for many problems where the use of traditional dynamical models built with differential equations is not appropriate or optimal.

While the use of modern algebra methods is now in the mainstream of mathematical biology research, this trend has been slow to influence the undergraduate mathematics and biology curricula, where difference and differential equation models still dominate. Several high-profile reports have been released in the past 5 years, including Refs. [1–3], calling urgently for broadening the undergraduate exposure at the interface of mathematics and biology, and including methods from modern discrete mathematics and their biological applications. However, those reports have been slow to elicit the transformative change in the undergraduate curriculum that many of us had hoped for. The anemic response may be attributed to a relative lack of educational undergraduate resources that highlight the critical impact of algebraic and discrete mathematical methods on contemporary biology. It is this niche that our book seeks to fill.

The format of this volume follows that of our earlier book, *Mathematical Concepts and Methods in Modern Biology: Using Modern Discrete Methods*, Robeva and Hodge (Editors), published in 2013 by Academic Press. At the time of its planning, we considered the modular format of that text (with chapters largely independent from one another) experimental, but we felt reassured when the book was selected as 1 of 12 contenders for the 2013 Society of Biology Awards in its category. We have adopted the same format here, as we believe that it provides readers and instructors with the independence to choose biological topics and mathematical methods that are of greatest interest to them.

Due to the modular format, the order of the chapters in the volume does not necessarily imply an increased level of difficulty or the need for more prerequisites for the later chapters. When chapters are connected by a common biological thread, they are grouped together, but they can still be used independently. Each chapter begins with a question or a number of related questions from modern biology, followed by the description of certain mathematical methods and theory appropriate in the search of answers. As in our earlier book, chapters can be viewed as fast-track pathways through the problem, which start by presenting the biological foundation, proceed by covering the relevant mathematical theory and presenting numerous examples, and end by highlighting connections with ongoing research and current publications. The level of presentation varies among chapters—some may be appropriate for introductory courses, while others may require more mathematical or biological background. Exercises are embedded within the text of each chapter, and their execution requires only material discussed up to that point. In addition, many chapters feature challenging open-ended questions (designated as projects) that provide starting points for explorations appropriate for undergraduate research, and supply references to relevant publications from the recent literature. In their most general form, some of the projects feature truly open questions in mathematical biology.

The book's companion website (http://booksite.elsevier.com/9780128012130) contains solutions to the exercises, as well all figures and relevant data files for the examples and exercises in the chapters. In addition, the site hosts software code, project guidelines, online supplements, appendices, and tutorials for selected chapters. The specialized software utilized throughout the book highlights the critical importance of computing applications for visualization, simulation, and analysis in modern biology. We have been careful to feature software that is in the mainstream of current mathematical biology research, while also being mindful of giving preference to freely available software.

We hope that the book will be a valuable resource to mathematics and biology programs, as it describes methods from discrete mathematics and modern algebra that can be presented, for the most part, at a level completely accessible to undergraduates. Yet the book provides extensions and connections with research that would also be helpful to graduate students and researchers in the field. Some of the material would be appropriate for mathematics courses such as finite mathematics, discrete structures, linear algebra, abstract/modern algebra, graph theory, probability, bioinformatics, statistics, biostatistics, and modeling, as well as for biology courses such as genetics, cell and molecular biology, biochemistry, ecology, and evolution.

The selection of topics for the volume and the choice of contributors grew out of the workshop "Teaching Discrete and Algebraic Mathematical Biology to Undergraduates" organized by Raina Robeva, Matthew Macauley, and Terrell Hodge and funded and hosted by the Mathematical Biosciences Institute (MBI) on July 29-August 2, 2013 at The Ohio State University. The editor and contributors of this volume greatly appreciate the encouragement and assistance received from the MBI's leadership and staff. Without their support, this volume would not have been possible. We also acknowledge with gratitude the support of the National Institute for Mathematical and Biological Synthesis (NIMBioS) in providing an opportunity to further test selected materials as part of the tutorial "Algebraic and Discrete Biological Models for Undergraduate Courses" offered on June 18-20, 2014 at NIMBioS.

I would like to express my personal thanks to all contributors who embraced the project early on and committed time and energy into producing the chapter modules for this unconventional textbook. Your enthusiasm for the project was remarkable, and you have my deep gratitude for the dedication and focus with which you carried it out. My special thanks also go to Daniel Hrozencik and Timothy Comar for providing feedback on a few of the chapter drafts. I am indebted to the editorial and production teams at Elsevier and particularly to the book's editors, Paula Callaghan and Katey Birtcher, our editorial project managers, Sarah Watson and Amy Clark, and our production manager, Vijayaraj Purushothaman. It has been a pleasure and a privilege to work with all of you. Finally, I would like to thank my husband, Boris Kovatchev, for his patience and support throughout.

Raina S. Robeva
October 20, 2014

REFERENCES

[1] Committee on a New Biology for the 21st Century: Ensuring the United States Leads the Coming Biology Revolution, Board on Life Sciences, Division on Earth and Life Studies, National Research Council. A new biology for the 21st century. Washington, DC: The National Academies Press; 2009.

[2] Brewer CA, Anderson CW (eds). Vision and change in undergraduate biology education: a call to action. Final report of a National Conference organized by the American Association for the Advancement of Science with support from the National Science Foundation, July 15-17, 2009, Washington, DC. The American Association for the Advancement of Science; 2011. http://visionandchange.org/files/2013/11/aaas-VISchange-web1113.pdf (accessed March 1, 2015).

[3] Committee on the Mathematical Sciences in 2025, Board on Mathematical Sciences and Their Applications, Division on Engineering and Physical Sciences, National Research Council. The mathematical sciences in 2025. Washington, DC: The National Academies Press; 2013.

Companion website: http://booksite.elsevier.com/9780128012130/

Supplementary Resources for Instructors

The website features the following additional resources available for download:

- All figures from the book
- Solutions to all exercises
- Computer code, data files, and links to software and materials carefully chosen to supplement the content of the textbook
- Appendices, tutorials, and additional projects for selected chapters

Contents

Companion website: http://booksite.elsevier.com/9780128012130/.

Chapter 1

Graph Theory for Systems Biology: Interval Graphs, Motifs, and Pattern Recognition

John R. Jungck[1] and Rama Viswanathan[2]
[1]*Center for Bioinformatics and Computational Biology, University of Delaware, Newark, DE, USA,* [2]*Beloit College, Beloit, WI, USA*

1.1 INTRODUCTION

Systems thinking is perceived as an important contemporary challenge of education [1]. However, *systems biology* is an old and inclusive term that connotes many different subareas of biology. Historically two important threads were synchronic: (a) the systems ecology of the Odum school [2–4], which was developed in the context of engineering principles applied to ecosystems [5, 6], and (b) systems physiology that used mechanical principles [7] to understand organs as mechanical devices integrated into the circulatory system, digestive system, anatomical system, immune system, nervous system, etc. For example, the heart could be thought of as a pump, the kidney as a filter, the lung as a bellows, the brain as a wiring circuit (or later as a computer), elbow joints as hinges, and so on. It should be noted that both areas extensively employed *ordinary* and *partial differential equations* (ODEs and PDEs). Indeed, some systems physiologists argued that all mathematical biology should be based on the application of PDEs. On the other hand, evolutionary biologists argued that these diachronic systems approaches too often answered only "how" questions that investigated optimal design principles and did not address "why" questions focusing on the constraints of historical contingency.

Not surprisingly, one of the leading journals in the field—*Frontiers in Systems Biology*—announces in its mission statement, [8] "Contrary to the reductionist paradigm commonly used in Molecular Biology, in Systems Biology the understanding of the behavior and evolution of complex biological systems need not necessarily be based on a detailed molecular description of the interactions between the system's constituent parts." Therefore, in this chapter we emphasize two major macroscopic and global aspects of contemporary systems biology: (i) the graph-theoretic relationships between components in networks and (ii) the relationship of these patterns to the historical contingencies of evolutionary constraints. Numerous articles and several books [9, 10] exist on graph theory and its application to systems biology, so the reader may ask what are we doing in this chapter that is different. Our main purpose is to help biologists, mathematicians, students, and researchers recognize which graph-theoretic tools are appropriate for different kinds of questions, including quantitative analyses of interactions for mining large data sets, visualizing complex relationships, modeling the consequences of perturbation of networks, and testing hypotheses.

Every network construct in systems biology is a hypothesis. For example, Rios and Vendruscolo [11] describe the network hypothesis as the assumption "according to which it is possible to describe a cell through the set of interconnections between its component molecules." They then conclude, "it becomes convenient to focus on

Algebraic and Discrete Mathematical Methods for Modern Biology. http://dx.doi.org/10.1016/B978-0-12-801213-0.00001-0

these interactions rather than on the molecules themselves to describe the functioning of the cell." In this chapter, we go a step further. We believe that a mathematical biology perspective also studies such questions as: Which molecules are involved? What do they do functionally? What is their three-dimensional structure? Where are they located in a cell? We stress that every network and pathway that we discuss is a useful construct from a biological perspective. They do not exist *per se* inside of cells. Imagine a series of biological macromolecules (proteins, nucleic acids, polysaccharides) that are crowded and colliding with one another in a suspension. The networks and pathways for the interactions between these molecules constructed by biologists may represent preferred associations defined by tighter bindings of specific macromolecules or the product of a reaction catalyzed by one macromolecule (an enzyme) as the starting material (substrate) of another enzyme. Thus, biologists have already drawn mathematical diagrams and graphs in the sense that they have abstracted, generalized, and symbolized a set or relationships.

Too often biologists produce networks as visualizations without further analysis. In this chapter, using Excel and Java-based software that we have developed, we show readers how to make mathematical measurements (average degree, diameter, clustering coefficient, etc.) and discern holistic properties (small world versus scale-free, see Hayes [12] for a complete overview) of the networks being studied and visualized, and obtain insights that are relevant and meaningful in the context of systems biology. We show how the network hypothesis can be investigated by complementary and supplementary mathematical and biological perspectives to yield key insights and help direct and inform additional research.

Palsson [10] suggests that twenty-first century biology will focus less on the reductionist study of components and more on the integration of systems analysis. He identifies four principles in his "systems biology paradigm": "First, the list of biological components that participated in the process of interest is enumerated. Second, the interactions between these components are studied and the 'wiring diagrams' of genetic circuits are constructed .... Third, reconstructed network[s] are described mathematically and their properties analyzed.... Fourth, the models are used to analyze, interpret, and predict biological experimental outcomes." Here, we assume that the first two steps exist in databases or published articles; this allows us to focus on the mathematics of the third step as a way that allows biologists to better direct their work on the fourth step. Thus, the goals for this chapter are as follows.

- Learn how graph theory can be used to help obtain meaningful insights into complex biological data sets.
- Analyze complex biological networks of diverse types (restriction maps, food webs, gene expression, disease etiology) to detect patterns of relationships.
- Visualize ordering of modules/motifs within complex biological networks by first testing the applicability of simple linear approaches (interval graphs).
- Demonstrate that even when strict mathematical assumptions do not apply fully to a given biological data set, there is still benefit in applying an analytical approach because of the power of the human mind to discern prominent patterns in data rearranged through the application of mathematical transformations.
- Show that the visualizations help biologists obtain insights into their data, examine the significance of outliers, mine databases for additional information about observed associations, and plan further experiments.

To accomplish this, we first emphasize *how* graph theory is a natural fit for biological investigations of relationships, patterns, and complexity. Second, graph theory lends itself easily to questions about *what* biologists should be looking for among representations of relationships. We introduce concepts of hubs, maximal cliques, motifs, clusters, interval graphs, complementary graphs, ordering, transitivity, Hamiltonian circuits, and consecutive ones in adjacency matrices. Finally, graph theory helps us interrogate *why* these relationships are occurring. Basically, we examine the triptych of form, function, and phylogeny to differentiate between evolutionary and engineering constraints.

The chapter is structured as follows. We begin by introducing some background concepts from graph theory that will be utilized later in the chapter. We then introduce interval graphs through two biological examples related to chromosome sequencing and food webs. The rest of the chapter is devoted to two extended examples of biological questions related to recently published studies on gene expression and disease etiology. The analyses