CAMBRIDGE
Cambridge University Press & Assessment
978-1-107-15954-9 — Topological Data Analysis for Genomics and Evolution
Raul Rabadan , Andrew J. Blumberg
Frontmatter
More Information

TOPOLOGICAL DATA ANALYSIS FOR GENOMICS AND EVOLUTION

Topology in Biology

Biology has entered the age of Big Data. A technical revolution has transformed the field, and extracting meaningful information from large biological data sets is now a central methodological challenge. Algebraic topology is a well-established branch of pure mathematics that studies qualitative descriptors of the shape of geometric objects. It aims to reduce comparisons of shape to a comparison of algebraic invariants, such as numbers, which are typically easier to work with. Topological data analysis is a rapidly developing subfield that leverages the tools of algebraic topology to provide robust multiscale analysis of data sets. This book introduces the central ideas and techniques of topological data analysis and its specific applications to biology, including the evolution of viruses, bacteria and humans, genomics of cancer, and single cell characterization of developmental processes. Bridging two disciplines, the book is for researchers and graduate students in genomics and evolutionary biology as well as mathematicians interested in applied topology.

RAÚL RABADÁN is a Professor at Columbia University, New York. He is Director of the Program for Mathematical Genomics at Columbia University, New York, and the NCI Physics and Oncology Center for Topology of Cancer Evolution and Heterogeneity. Dr Rabadán received his Ph.D. in Theoretical Physics in 2001 and went on to conduct research at the European Laboratory for Particle Physics (CERN) in Switzerland, and at the Institute for Advanced Study (IAS) in Princeton, New Jersey. At Columbia University, he leads a highly interdisciplinary laboratory with researchers from the fields of mathematics, physics, computer science, engineering, and medicine, with the common goal of solving biomedical problems through quantitative computational models.

ANDREW J. BLUMBERG is a Professor in the Department of Mathematics at the University of Texas, Austin. He completed his Ph.D. at the University of Chicago under the supervision of Peter May and Michael Mandell, and was later a National Science Foundation postdoctoral fellow at Stanford. He also spent a year as a member at the Institute for Advanced Study (IAS) in Princeton, New Jersey. His pure mathematics research focuses primarily on homotopy theory and algebraic topology and his applied research focuses on the development of topological and geometric techniques for studying genomic data.

TOPOLOGICAL DATA ANALYSIS FOR GENOMICS AND EVOLUTION

Topology in Biology

RAÚL RABADÁN
Columbia University, New York

ANDREW J. BLUMBERG
University of Texas, Austin

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment,
a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of
education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781107159549

DOI: 10.1017/9781316671665

First published 2020

A catalogue record for this publication is available from the British Library

Library of Congress Cataloging-in-Publication data
Names: Rabadán, Raúl, author. | Blumberg, Andrew J., author.
Title: Topological data analysis for genomics and evolution : topology in
biology / Raúl Rabadán, Columbia University, New York, Andrew J.
Blumberg, University of Texas, Austin.
Description: Cambridge, United Kingdom ; New York, NY : Cambridge University
Press, 2019. | Includes bibliographical references and index.
Identifiers: LCCN 2019002342 | ISBN 9781107159549 (hardback : alk. paper)
Subjects: LCSH: Bioinformatics – Mathematical models. | Computational biology.
| Mathematical analysis.
Classification: LCC QH324.2 .R33 2019 | DDC 570.285–dc23
LC record available at https://lccn.loc.gov/2019002342

ISBN 978-1-107-15954-9 Hardback

This book is dedicated to our families, for their persistent support.
To Jean-Michel, Emma, and Alex.
To Olena, Miriam, and Becky.

Contents

Cambridge University Press & Assessment
978-1-107-15954-9 — Topological Data Analysis for Genomics and Evolution
Raul Rabadan , Andrew J. Blumberg
Frontmatter
More Information

Contributors

Andrew J. Blumberg
University of Texas, Austin

Pablo G. Cámara
University of Pennsylvania, Philadelphia

Joseph Chan
Memorial Sloan-Kettering Cancer Center, New York

Kevin Emmett
Columbia University, New York

M. Riley Meth
University of Texas, Austin

Raúl Rabadán
Columbia University, New York

Daniel Rosenbloom
Columbia University, New York

Preface

Modern biology is awash in data. This situation, the result of a technical revolution in high-throughput genomics, promises rapid scientific advances. However, analyzing the data poses unique challenges. Unlike in physics, there is usually no quantitative biological model that can guide investigation and generate precise predictions; often, we do not even know what the relevant quantities are that could capture the essential behavior of the biological system.

In response to the flood of data, the use of clustering algorithms and dimensionality reduction procedures is now ubiquitous. These families of techniques can be regarded as efforts to describe the *shape* of the data set. Although there have been noted successes, such methods provide only crude descriptions of this shape. The power of these tools, as well as their evident limitations, makes it clear that there would be substantial scientific benefit from richer and more robust methods for understanding geometric structure in data.

Algebraic topology is a well-established branch of pure mathematics that studies qualitative descriptors of the shape of geometric objects. Roughly speaking, the goal of algebraic topology is to reduce questions about comparing shapes to questions about comparing algebraic invariants (e.g., numbers), which are typically easier to solve. Moreover, algebraic topology has had a long tradition of employing combinatorial models of geometric objects, *simplicial complexes*, that are well suited to algorithmic computation.

Topological data analysis is a rapidly developing subfield that leverages the tools and outlook of algebraic topology to provide a methodology for analyzing the shape of data sets. The basic strategy is to assign a family of simplicial complexes to a data set; invariants of the complexes integrate information about the shape of the data across different feature scales.

Our aim in this book is to provide a concise introduction to the central ideas and techniques of topological data analysis and to explain in detail a number of specific applications to biology. We imagine as our idealized readers a modern quantitative biologist or a graduate student in mathematics with a background in topology or geometry and an interest in applied problems. We have three central goals:

1. to equip the modern quantitative biologist with techniques from topological data analysis,
2. to direct mathematicians with training in geometry and topology towards problems of interest to biologists, and
3. to make it easier for mathematicians and biologists to communicate and collaborate.

These goals pose an expositional challenge, as we expect two quite different audiences with different backgrounds. To address this, we have attempted as much as possible to provide a self-contained introduction to the relevant topics along with abundant and detailed references. We assume that the reader has some familiarity with calculus, linear algebra, elementary probability, and basic statistics.

The first part of this book presents the mathematical background necessary to understand topological data analysis and then provides an overview of techniques in the area. These chapters are intended to be read in order, as each one builds on the previous chapters. The second part of this book consists of a collection of distinct biological applications; each chapter can be read independently.

Acknowledgements

This work grew out of the efforts of many people. We would like to thank Arnold Levine, for his vision, his scientific insights, and his enthusiasm. He created an exceptional creative interdisciplinary environment at the Institute for Advanced Study in Princeton, providing the seeds of many of the ideas discussed in this book. Pablo G. Cámara, Joseph Chan, Kevin Emmett, and Daniel Rosenbloom contributed to several sections in the initial draft of the book. M. Riley Meth made many invaluable corrections and contributions to the second draft of the book. Juan Patino Galindo provided feedback on using genomic data for studying evolutionary processes, and, in particular, helped to write an introduction on different methods to study recombination. We are particularly thankful to Timothy Chu, Oliver Elliott, and M. Riley Meth for using their artistic talents to design the illustrations that enliven the book. William Blumberg and Michael Walfish provided careful readings and helpful comments on previous drafts. Jacqueline Aw, Kyle Bolo, Andrew Chen, Ioan Filip, Chioma Madubata, Patrick Van Nieuwenhuizen, Samuel J. Resnick, Richard T. Wolff, and Sakellarios Zairis proofread different sections of the book. Michael Lesnick and Jun-Hou Fung gave the entire book a very careful reading and made numerous helpful comments correcting errors and improving the exposition. The authors gratefully acknowledge many interesting discussions with Nils Baas, Gunnar Carlsson, Ben Greenbaum, Gillian Grindstaff, Hossein Khiabanian, Michael Lesnick, Arnold Levine, Michael Mandell, M. Riley Meth, Bud Mishra, Anthea Monod, Sayan Mukherjee, Vladimir Trifonov, Stephen Walker, and Jiguang Wang. In addition, Raúl Rabadán would like to acknowledge many of his collaborators in biology for the time shared, their patience, and the fun solving many problems together: Uttiya Basu, Riccardo Dalla Favera, Adolfo Ferrando, Antonio Iavarone, Anna Lasorella, Tom Maniatis, Do-Hyun Nam, Gustavo Palacios, Teresa Palomero, Laura Pasqualucci, Abbas Rizvi, and Sagi Shapira among many others.

This book was possible in part due to the funding from the Center for Topology and Evolution of Cancer at Columbia University through the National Cancer Institute (U54 CA193313). The Center brings together mathematicians and cancer biologists to solve some interesting problems in cancer. This book was born from many interesting interactions between mathematicians, computational biologists and cancer researchers, where with more or less success, but always with enthusiasm, we have tried to cross the interdisciplinary borders that separate our disciplines. In addition, Raúl Rabadán would like to acknowledge the National Institute of Health grants, R01 CA179044, R01 GM109018, R01 CA185486 and U54 CA209997, and the Convergence program of Stand Up to Cancer together with National Science Foundation. Both authors acknowledge the National Institute of Health grant R01 GM117591. Andrew Blumberg would also like to acknowledge AFOSR research grant FA9550-15-1-0302.