

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



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](http://www.cambridge.org)  
 Information on this title: [www.cambridge.org/9781107159549](http://www.cambridge.org/9781107159549)

DOI: 10.1017/9781316671665

© Raúl Rabadán and Andrew J. Blumberg 2020

This publication is in copyright. Subject to statutory exception and to the provisions  
 of relevant collective licensing agreements, no reproduction of any part may take  
 place without the written permission of Cambridge University Press & Assessment.

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://lcn.loc.gov/2019002342>

ISBN 978-1-107-15954-9 Hardback

Cambridge University Press & Assessment has no responsibility for the persistence  
 or accuracy of URLs for external or third-party internet websites referred to in this  
 publication and does not guarantee that any content on such websites is, or will  
 remain, accurate or appropriate.

This book is dedicated to our families, for their persistent support.

To Jean-Michel, Emma, and Alex.

To Olena, Miriam, and Becky.

Contents

<i>List of Contributors</i>	<i>page</i> xiii
<i>Preface</i>	xv
Introduction	1
0.1 Why Algebraic Topology?	5
0.2 Combinatorial Algebraic Topology	7
0.3 Topological Data Analysis (TDA)	10
0.4 Genetics and Genomics	13
0.5 Why Is Topological Data Analysis Useful in Genomics?	15
0.6 What Is in This Book?	19

**Part I Topological Data Analysis**

1 Basic Notions of Algebraic Topology	23
1.1 Sets	25
1.2 Metric Spaces	29
1.3 Topological Spaces	38
1.3.1 Maps between Topological Spaces	43
1.3.2 Homeomorphisms	46
1.4 Continuous Deformations and Homotopy Invariants	49
1.4.1 Homotopy Groups	53
1.5 Gluing and CW Complexes	56
1.6 Algebra	62
1.6.1 Groups	62
1.6.2 Homomorphisms	66
1.6.3 New Groups from Old	67
	vii

viii	<i>Contents</i>	
	1.6.4 The Group Structure on $\pi_n(X, x)$	71
	1.6.5 Rings and Fields	75
	1.6.6 Vector Spaces and Linear Algebra	76
1.7	Category Theory	80
	1.7.1 Functors	90
1.8	Simplicial Complexes	92
1.9	The Euler Characteristic	101
1.10	Simplicial Homology	102
	1.10.1 Chains and Boundaries	103
	1.10.2 Homology Groups	105
	1.10.3 Homology of Chain Complexes	108
	1.10.4 Simplicial Homology with Coefficients in an Abelian Group	112
1.11	Manifolds	114
1.12	Morse Functions and Reeb Spaces	118
1.13	Summary	120
1.14	Suggestions for Further Reading	121
2	Topological Data Analysis	122
	2.1 Simplicial Complexes Associated to Data	123
	2.2 The Niyogi-Smale-Weinberger Theorem	128
	2.3 Persistent Homology	132
	2.4 Stability of Persistent Homology under Perturbation	141
	2.5 Zigzag Persistence	149
	2.6 Multidimensional Persistence	153
	2.6.1 Multidimensional Persistence	154
	2.6.2 The Persistent Homology Transform	155
	2.7 Efficient Computation of Persistent Homology	158
	2.8 Multiscale Clustering: Mapper	161
	2.9 Towards Persistent Algebraic Topology	167
	2.10 Summary	168
	2.11 Suggestions for Further Reading	169
3	Statistics and Topological Inference	170
	3.1 What Can Topological Data Analysis Tell Us?	171
	3.1.1 Persistent Homology and Sampling	173
	3.1.2 Topological Inference	181
	3.2 Background: Geometric Sampling and Metric Measure Spaces	183
	3.2.1 Metric Measure Spaces	183

<i>Contents</i>		ix
3.2.2	The Fréchet Mean and Variance of a Metric Measure Space	189
3.2.3	Distances on Measures and Metric Measure Spaces	191
3.3	Probability Theory in Barcode Space	195
3.3.1	Polish Spaces of Barcodes	195
3.3.2	Sampling and Hypothesis Testing in Barcode Space	197
3.4	Stability Theorems for Persistent Homology of Metric Measure Spaces	199
3.5	Estimating Persistent Homology from Samples	205
3.5.1	Estimating Persistent Homology by Density Estimation	210
3.5.2	Estimating Persistent Homology by Resampling	213
3.6	Summarizing Persistence Diagrams	216
3.6.1	Tractable Features from Persistence Diagrams	218
3.6.2	Kernel Methods for Barcodes	221
3.6.3	Persistence Landscapes	221
3.6.4	Coordinates on Persistent Homology	225
3.7	Stochastic Topology and the Expected Persistent Homology of Random Complexes	226
3.8	Euler Characteristics in Topological Data Analysis	228
3.9	Exploratory Data Analysis with Mapper	231
3.10	Summary	233
3.11	Suggestions for Further Reading	234
4	Dimensionality Reduction, Manifold Learning, and Metric Geometry	235
4.1	A Quick Refresher on Eigenvectors and Eigenvalues	238
4.2	Background on PCA and MDS	239
4.3	Manifold Learning	242
4.3.1	Isomap	242
4.3.2	Local Linear Embedding (LLE)	244
4.3.3	Laplacian Eigenmaps	246
4.3.4	Manifold Learning and Kernel Methods	248
4.3.5	Discrete Harmonic Analysis	249
4.3.6	Other Manifold Learning Techniques	251
4.3.7	Manifolds of Differing Dimension	252
4.4	Neighbor Embedding Algorithms	252
4.4.1	Stochastic neighbor Embedding (SNE)	253

4.4.2	<i>t</i> -Distributed Stochastic Neighbor Embedding ( <i>t</i> -SNE)	254
4.4.3	Reliable Use of <i>t</i> -SNE	256
4.5	Mapper and Manifold Learning	257
4.6	Dimensionality Estimation	257
4.7	Metric Trees and Spaces of Phylogenetic Trees	260
4.7.1	Inferring Trees from Metric Data	262
4.7.2	The Billera-Holmes-Vogtmann Metric Spaces of Phylogenetic Trees	264
4.7.3	Metric Geometry	266
4.8	Summary	269
4.9	Suggestions for Further Reading	269
 <b>Part II Biological Applications</b>		
5	Evolution, Trees, and Beyond	273
5.1	Introduction	273
5.2	Evolution and Topology	279
5.3	Viral Evolution: Influenza A	287
5.3.1	Influenza A	287
5.3.2	Reassortments in Influenza through TDA	293
5.3.3	Influenza Virus Evolution and the Space of Phylogenetic Trees	300
5.4	Viral Evolution: HIV	303
5.4.1	Human Immunodeficiency Virus	303
5.4.2	Viral Recombination in HIV	307
5.4.3	Viral Recombination in Late-Stage HIV Infection	308
5.5	Other Viruses	314
5.6	Bacterial Evolution	315
5.6.1	Horizontal Gene Transfer in Bacteria	316
5.6.2	Pathogenic Bacteria	318
5.6.3	Multilocus Sequence Typing Analysis	318
5.6.4	Protein Family Analysis	320
5.6.5	Antibiotic Resistance in <i>Staphylococcus aureus</i>	322
5.7	Persistent Homology Estimators in Population Genetics	324
5.7.1	Coalescent Process	324
5.7.2	Statistical Model	325
5.7.3	Coalescent Simulations	327



	<i>Contents</i>	xi
5.8	Recombination Landscape in Humans	328
5.8.1	Fine-Scale Resolution of Human Recombination	331
5.9	Gene Trees and Species Trees	333
5.10	Extensions: Median Complex and Topological Minimal Graphs	337
5.10.1	The Median Complex Construction	339
5.10.2	Topological Minimal Graphs and Barcode Ensembles	342
5.11	Summary	351
5.12	Suggestions for Further Reading, Databases, and Software	353
6	Cancer Genomics	356
6.1	A Brief History of Cancer	356
6.2	Cancer in the Era of Molecular Biology	360
6.3	The Standard Model of Tumor Evolution	363
6.4	Cancer in the Era of Genomic Data	365
6.4.1	Point Mutations	366
6.4.2	Copy Number Alterations	370
6.4.3	Gene Fusions and Translocations	371
6.4.4	Viruses	374
6.5	Differential Gene Expression Analysis in Cancer	376
6.6	The Space of Glioblastomas	377
6.7	Cross-Sectional Data in Cancer and Patient Stratification Using Expression Data	379
6.8	Cross-Sectional Data in Cancer and Identifying Driver Genes in Cancer	383
6.9	The Tissue of Origin of Melanomas	385
6.10	Association between Drug Sensitivity and Genomic Alterations	391
6.11	Summary	396
6.12	Suggestions for Further Reading and Databases	398
7	Single Cell Expression Data	399
7.1	Introduction to Single Cell Technologies	400
7.2	Identifying Distinct Cell Subpopulations in Cancer	402
7.2.1	Clonal Heterogeneity from Single Cell Tumor Genomics	404
7.3	Asynchronous Differentiation Processes	405
7.4	Differentiation in Human Preimplantation Embryos	408

xii	<i>Contents</i>	
7.5	Summary	410
7.6	Suggestions for Further Reading, Databases, and Software	411
8	Three-Dimensional Structure of DNA	412
8.1	Background	413
8.2	TDA and Chromatin Structure	414
8.3	Simulations	416
8.4	The Topology of Bacterial DNA	417
8.5	The Topology of Human DNA	419
8.6	Summary	421
8.7	Suggestions for Databases and Software	422
9	Topological Data Analysis beyond Genomics	423
9.1	Topological Study of Series Analysis	424
9.1.1	Time Series Analysis of Gene Expression Data	427
9.1.2	Time Series Analysis Using Topological Data Analysis	432
9.1.3	Topological Data Analysis of Sliding Windows	433
9.1.4	Identification of Copy Number Alterations	434
9.2	Topological Data Analysis in Networks and Neuroscience	436
9.2.1	Cellular Scales: Neuronal Activity	436
9.2.2	Mesoscopic Scales: Brain Functional Networks	437
9.3	Topological Approaches to Biomedical Imaging	438
9.4	Spreading of Infectious Diseases	440
9.5	Summary	441
9.6	Suggestions for Further Reading	442
10	Conclusions	443
Appendix A	Algorithms in Topological Data Analysis	444
Appendix B	Introduction to Population Genetics	447
Appendix C	Molecular Phylogenetics	454
	<i>References</i>	468
	<i>Index</i>	495

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

*Preface*

xvii

improving the exposition. The authors gratefully acknowledge many interesting discussions with Nils Baas, Gunnar Carlsson, Ben Greenbaum, Gillian Grindstaff, Hossein Khiabani, 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.