

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

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

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

Part II Biological Applications

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

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

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
<i>Appendix A</i>	Algorithms in Topological Data Analysis	444
<i>Appendix B</i>	Introduction to Population Genetics	447
<i>Appendix C</i>	Molecular Phylogenetics	454
	<i>References</i>	468
	<i>Index</i>	495