

Contents

• • • • •

List of Figures	xix
Acknowledgments	xxi
0 Introduction: Building a Computing Toolbox	1
0.1 The Philosophy	2
0.2 The Structure of the Book	4
0.2.1 How to Read the Book	6
0.2.2 Exercises and Further Reading	6
0.3 Use in the Classroom	8
0.4 Formatting of the Book	10
0.5 Setup	10
1 Unix	12
1.1 What Is Unix?	12
1.2 Why Use Unix and the Shell?	13
1.3 Getting Started with Unix	14
1.3.1 Installation	14
1.3.2 Directory Structure	15
1.4 Getting Started with the Shell	17
1.4.1 Invoking and Controlling Basic Unix Commands	18
1.4.2 How to Get Help in Unix	19
1.4.3 Navigating the Directory System	20
1.5 Basic Unix Commands	22
1.5.1 Handling Directories and Files	22
1.5.2 Viewing and Processing Text Files	24
1.6 Advanced Unix Commands	27
1.6.1 Redirection and Pipes	27
1.6.2 Selecting Columns Using cut	29
1.6.3 Substituting Characters Using tr	32

1.6.4	Wildcards	35
1.6.5	Selecting Lines Using grep	36
1.6.6	Finding Files with find	39
1.6.7	Permissions	41
1.7	Basic Scripting	43
1.8	Simple for Loops	47
1.9	Tips, Tricks, and Going beyond the Basics	49
1.9.1	Setting a PATH in .bash_profile	49
1.9.2	Line Terminators	50
1.9.3	Miscellaneous Commands	50
1.10	Exercises	51
1.10.1	Next Generation Sequencing Data	51
1.10.2	Hormone Levels in Baboons	51
1.10.3	Plant–Pollinator Networks	52
1.10.4	Data Explorer	53
1.11	References and Reading	53
2	Version Control	55
2.1	What Is Version Control?	55
2.2	Why Use Version Control?	55
2.3	Getting Started with Git	56
2.3.1	Installing Git	57
2.3.2	Configuring Git after Installation	57
2.3.3	How to Get Help in Git	58
2.4	Everyday Git	58
2.4.1	Workflow	58
2.4.2	Showing Changes	64
2.4.3	Ignoring Files and Directories	65
2.4.4	Moving and Removing Files	66
2.4.5	Troubleshooting Git	66
2.5	Remote Repositories	68
2.6	Branching and Merging	70
2.7	Contributing to Public Repositories	78
2.8	References and Reading	79
3	Basic Programming	81
3.1	Why Programming?	81
3.2	Choosing a Programming Language	81
3.3	Getting Started with Python	83

3.3.1	Installing Python and Jupyter	83
3.3.2	How to Get Help in Python	84
3.3.3	Simple Calculations with Basic Data Types	85
3.3.4	Variable Assignment	87
3.3.5	Built-In Functions	89
3.3.6	Strings	90
3.4	Data Structures	93
3.4.1	Lists	93
3.4.2	Dictionaries	96
3.4.3	Tuples	100
3.4.4	Sets	101
3.5	Common, General Functions	103
3.6	The Flow of a Program	105
3.6.1	Conditional Branching	105
3.6.2	Looping	107
3.7	Working with Files	112
3.7.1	Text Files	112
3.7.2	Character-Delimited Files	115
3.8	Exercises	117
3.8.1	Measles Time Series	117
3.8.2	Red Queen in Fruit Flies	118
3.9	References and Reading	118
4	Writing Good Code	120
4.1	Writing Code for Science	120
4.2	Modules and Program Structure	121
4.2.1	Writing Functions	121
4.2.2	Importing Packages and Modules	126
4.2.3	Program Structure	127
4.3	Writing Style	133
4.4	Python from the Command Line	135
4.5	Errors and Exceptions	137
4.5.1	Handling Exceptions	138
4.6	Debugging	139
4.7	Unit Testing	146
4.7.1	Writing the Tests	147
4.7.2	Executing the Tests	149
4.7.3	Handling More Complex Tests	150

4.8	Profiling	153
4.9	Beyond the Basics	155
4.9.1	Arithmetic of Data Structures	155
4.9.2	Mutable and Immutable Types	156
4.9.3	Copying Objects	158
4.9.4	Variable Scope	160
4.10	Exercises	161
4.10.1	Assortative Mating in Animals	161
4.10.2	Human Intestinal Ecosystems	162
4.11	References and Reading	163
5	Regular Expressions	165
5.1	What Are Regular Expressions?	165
5.2	Why Use Regular Expressions?	165
5.3	Regular Expressions in Python	166
5.3.1	The <code>re</code> Module in Python	166
5.4	Building Regular Expressions	167
5.4.1	Literal Characters	168
5.4.2	Metacharacters	168
5.4.3	Sets	169
5.4.4	Quantifiers	170
5.4.5	Anchors	171
5.4.6	Alternations	172
5.4.7	Raw String Notation and Escaping Metacharacters	173
5.5	Functions of the <code>re</code> Module	175
5.6	Groups in Regular Expressions	179
5.7	Verbose Regular Expressions	181
5.8	The Quest for the Perfect Regular Expression	181
5.9	Exercises	182
5.9.1	Bee Checklist	182
5.9.2	A Map of <i>Science</i>	182
5.10	References and Reading	184
6	Scientific Computing	185
6.1	Programming for Science	185
6.1.1	Installing the Packages	185

6.2	Scientific Programming with NumPy and SciPy	185
6.2.1	NumPy Arrays	186
6.2.2	Random Numbers and Distributions	194
6.2.3	Linear Algebra	196
6.2.4	Integration and Differential Equations	197
6.2.5	Optimization	200
6.3	Working with pandas	202
6.4	Biopython	208
6.4.1	Retrieving Sequences from NCBI	208
6.4.2	Input and Output of Sequence Data Using SeqIO	210
6.4.3	Programmatic BLAST Search	212
6.4.4	Querying PubMed for Scientific Literature Information	214
6.5	Other Scientific Python Modules	216
6.6	Exercises	216
6.6.1	Lord of the Fruit Flies	216
6.6.2	Number of Reviewers and Rejection Rate	217
6.6.3	The Evolution of Cooperation	217
6.7	References and Reading	219
7	Scientific Typesetting	220
7.1	What Is L ^A T _E X?	220
7.2	Why Use L ^A T _E X?	220
7.3	Installing L ^A T _E X	223
7.4	The Structure of L ^A T _E X Documents	223
7.4.1	Document Classes	224
7.4.2	L ^A T _E X Packages	224
7.4.3	The Main Body	225
7.4.4	Document Sections	227
7.5	Typesetting Text with L ^A T _E X	228
7.5.1	Spaces, New Lines, and Special Characters	228
7.5.2	Commands and Environments	228
7.5.3	Typesetting Math	229
7.5.4	Comments	231
7.5.5	Justification and Alignment	232
7.5.6	Long Documents	232
7.5.7	Typesetting Tables	233
7.5.8	Typesetting Matrices	236

7.5.9	Figures	237
7.5.10	Labels and Cross-References	240
7.5.11	Itemized and Numbered Lists	241
7.5.12	Font Styles	241
7.5.13	Bibliography	242
7.6	\LaTeX Packages for Biologists	244
7.6.1	Sequence Alignments with \LaTeX	245
7.6.2	Creating Chemical Structures with \LaTeX	246
7.7	Exercises	246
7.7.1	Typesetting Your Curriculum Vitae	246
7.8	References and Reading	247
8	Statistical Computing	249
8.1	Why Statistical Computing?	249
8.2	What Is R?	249
8.3	Installing R and RStudio	250
8.4	Why Use R and RStudio?	250
8.5	Finding Help	251
8.6	Getting Started with R	251
8.7	Assignment and Data Types	253
8.8	Data Structures	255
8.8.1	Vectors	255
8.8.2	Matrices	257
8.8.3	Lists	261
8.8.4	Strings	262
8.8.5	Data Frames	263
8.9	Reading and Writing Data	264
8.10	Statistical Computing Using Scripts	267
8.10.1	Why Write a Script?	267
8.10.2	Writing Good Code	267
8.11	The Flow of the Program	270
8.11.1	Branching	270
8.11.2	Loops	272
8.12	Functions	275
8.13	Importing Libraries	278
8.14	Random Numbers	279
8.15	Vectorize It!	280
8.16	Debugging	283
8.17	Interfacing with the Operating System	284

8.18	Running R from the Command Line	285
8.19	Statistics in R	287
8.20	Basic Plotting	290
8.20.1	Scatter Plots	290
8.20.2	Histograms	291
8.20.3	Bar Plots	292
8.20.4	Box Plots	292
8.20.5	3D Plotting (in 2D)	293
8.21	Finding Packages for Biological Research	293
8.22	Documenting Code	294
8.23	Exercises	295
8.23.1	Self-Incompatibility in Plants	295
8.23.2	Body Mass of Mammals	296
8.23.3	Leaf Area Using Image Processing	296
8.23.4	Titles and Citations	297
8.24	References and Reading	297
9	Data Wrangling and Visualization	300
9.1	Efficient Data Analysis and Visualization	300
9.2	Welcome to the tidyverse	300
9.2.1	Reading Data	301
9.2.2	Tibbles	302
9.3	Selecting and Manipulating Data	304
9.3.1	Subsetting Data	305
9.3.2	Pipelines	307
9.3.3	Renaming Columns	308
9.3.4	Adding Variables	309
9.4	Counting and Computing Statistics	310
9.4.1	Summarize Data	310
9.4.2	Grouping Data	310
9.5	Data Wrangling	313
9.5.1	Gathering	313
9.5.2	Spreading	315
9.5.3	Joining Tibbles	316
9.6	Data Visualization	318
9.6.1	Philosophy of ggplot2	319
9.6.2	The Structure of a Plot	320
9.6.3	Plotting Frequency Distribution of One Continuous Variable	321

9.6.4	Box Plots and Violin Plots	322
9.6.5	Bar Plots	323
9.6.6	Scatter Plots	324
9.6.7	Plotting Experimental Errors	325
9.6.8	Scales	326
9.6.9	Faceting	328
9.6.10	Labels	329
9.6.11	Legends	330
9.6.12	Themes	331
9.6.13	Setting a Feature	332
9.6.14	Saving	332
9.7	Tips & Tricks	333
9.8	Exercises	335
9.8.1	Life History in Songbirds	335
9.8.2	Drosophilidae Wings	335
9.8.3	Extinction Risk Meta-Analysis	335
9.9	References and Reading	336
10	Relational Databases	337
10.1	What Is a Relational Database?	337
10.2	Why Use a Relational Database?	338
10.3	Structure of Relational Databases	340
10.4	Relational Database Management Systems	341
10.4.1	Installing SQLite	341
10.4.2	Running the SQLite RDBMS	341
10.5	Getting Started with SQLite	342
10.5.1	Comments	342
10.5.2	Data Types	342
10.5.3	Creating and Importing Tables	343
10.5.4	Basic Queries	344
10.6	Designing Databases	352
10.7	Working with Databases	355
10.7.1	Joining Tables	355
10.7.2	Views	358
10.7.3	Backing Up and Restoring a Database	359
10.7.4	Inserting, Updating, and Deleting Records	360
10.7.5	Exporting Tables and Views	361
10.8	Scripting	362
10.9	Graphical User Interfaces (GUIs)	362

10.10	Accessing Databases Programmatically	362
10.10.1	In Python	363
10.10.2	In R	363
10.11	Exercises	364
10.11.1	Species Richness of Birds in Wetlands	364
10.11.2	Gut Microbiome of Termites	364
10.12	References and Reading	365
11	Wrapping Up	366
11.1	How to Be a More Efficient Computational Biologist	367
11.2	What Next?	368
11.3	Conclusion	371
	Intermezzo Solutions	373
	Bibliography	389
	Indexes	393
	Index of Symbols	393
	Index of Unix Commands	395
	Index of Git Commands	397
	Index of Python Functions, Methods, Properties, and Libraries	399
	Index of \LaTeX Commands and Libraries	401
	Index of R Functions and Libraries	403
	Index of SQLite Commands	405
	General Index	407