

Table of Contents

Preface	v
Chapter 1: Python and the Surrounding Software Ecology	1
Introduction	1
Installing the required software with Anaconda	2
Installing the required software with Docker	7
Interfacing with R via rpy2	9
Performing R magic with IPython	16
Chapter 2: Next-generation Sequencing	19
Introduction	19
Accessing GenBank and moving around NCBI databases	20
Performing basic sequence analysis	25
Working with modern sequence formats	28
Working with alignment data	37
Analyzing data in the variant call format	44
Studying genome accessibility and filtering SNP data	47
Chapter 3: Working with Genomes	61
Introduction	61
Working with high-quality reference genomes	62
Dealing with low-quality genome references	68
Traversing genome annotations	73
Extracting genes from a reference using annotations	76
Finding orthologues with the Ensembl REST API	80
Retrieving gene ontology information from Ensembl	83
Chapter 4: Population Genetics	89
Introduction	89
Managing datasets with PLINK	91
Introducing the Genepop format	97

Exploring a dataset with Bio.PopGen	101
Computing F-statistics	107
Performing Principal Components Analysis	113
Investigating population structure with Admixture	118
Chapter 5: Population Genetics Simulation	125
Introduction	125
Introducing forward-time simulations	126
Simulating selection	132
Simulating population structure using island and stepping-stone models	138
Modeling complex demographic scenarios	143
Simulating the coalescent with Biopython and fastsimcoal	149
Chapter 6: Phylogenetics	155
Introduction	155
Preparing the Ebola dataset	156
Aligning genetic and genomic data	162
Comparing sequences	164
Reconstructing phylogenetic trees	170
Playing recursively with trees	174
Visualizing phylogenetic data	179
Chapter 7: Using the Protein Data Bank	187
Introduction	187
Finding a protein in multiple databases	188
Introducing Bio.PDB	192
Extracting more information from a PDB file	197
Computing molecular distances on a PDB file	201
Performing geometric operations	205
Implementing a basic PDB parser	208
Animating with PyMol	212
Parsing mmCIF files using Biopython	220
Chapter 8: Other Topics in Bioinformatics	223
Introduction	223
Accessing the Global Biodiversity Information Facility	224
Geo-referencing GBIF datasets	230
Accessing molecular-interaction databases with PSIQUIC	236
Plotting protein interactions with Cytoscape the hard way	239
Chapter 9: Python for Big Genomics Datasets	247
Introduction	247
Setting the stage for high-performance computing	248
Designing a poor human concurrent executor	254

Performing parallel computing with IPython
Computing the median in a large dataset
Optimizing code with Cython and Numba
Programming with laziness
Thinking with generators

Index
