
Contents

List of Figures	xix
List of Tables	xxiii
Preface to the First Edition	xxv
Preface to the Second Edition	xxvii
Acknowledgments	xxxi
SECTION I Programming	
CHAPTER 1 ■ Introduction	3
1.1 WHO SHOULD READ THIS BOOK	3
1.1.1 What the Reader Should Already Know	4
1.2 USING THIS BOOK	4
1.2.1 Typographical Conventions	4
1.2.2 Python Versions	5
1.2.3 Code Style	5
1.2.4 Get the Most from This Book without Reading It All	6
1.2.5 Online Resources Related to This Book	7
1.3 WHY LEARN TO PROGRAM?	7
1.4 BASIC PROGRAMMING CONCEPTS	8
1.4.1 What Is a Program?	8
1.5 WHY PYTHON?	10
1.5.1 Main Features of Python	10
1.5.2 Comparing Python with Other Languages	11
1.5.3 How Is It Used?	14
1.5.4 Who Uses Python?	15
1.5.5 Flavors of Python	15
1.5.6 Special Python Distributions	16

1.6	ADDITIONAL RESOURCES	17
CHAPTER 2 ■ First Steps with Python		19
2.1	INSTALLING PYTHON	20
2.1.1	Learn Python by Using It	20
2.1.2	Install Python Locally	20
2.1.3	Using Python Online	21
2.1.4	Testing Python	22
2.1.5	First Use	22
2.2	INTERACTIVE MODE	23
2.2.1	Baby Steps	23
2.2.2	Basic Input and Output	23
2.2.3	More on the Interactive Mode	24
2.2.4	Mathematical Operations	26
2.2.5	Exit from the Python Shell	27
2.3	BATCH MODE	27
2.3.1	Comments	29
2.3.2	Indentation	30
2.4	CHOOSING AN EDITOR	32
2.4.1	Sublime Text	32
2.4.2	Atom	33
2.4.3	PyCharm	34
2.4.4	Spyder IDE	35
2.4.5	Final Words about Editors	36
2.5	OTHER TOOLS	36
2.6	ADDITIONAL RESOURCES	37
2.7	SELF-EVALUATION	37
CHAPTER 3 ■ Basic Programming: Data Types		39
3.1	STRINGS	40
3.1.1	Strings Are Sequences of Unicode Characters	41
3.1.2	String Manipulation	42
3.1.3	Methods Associated with Strings	42
3.2	LISTS	44
3.2.1	Accessing List Elements	45

3.2.2	List with Multiple Repeated Items	45
3.2.3	List Comprehension	46
3.2.4	Modifying Lists	47
3.2.5	Copying a List	49
3.3	TUPLES	49
3.3.1	Tuples Are Immutable Lists	49
3.4	COMMON PROPERTIES OF THE SEQUENCES	51
3.5	DICTIONARIES	54
3.5.1	Mapping: Calling Each Value by a Name	54
3.5.2	Operating with Dictionaries	56
3.6	SETS	58
3.6.1	Unordered Collection of Objects	59
3.6.2	Set Operations	60
3.6.3	Shared Operations with Other Data Types	62
3.6.4	Immutable Set: Frozenset	63
3.7	NAMING OBJECTS	63
3.8	ASSIGNING A VALUE TO A VARIABLE VERSUS BINDING A NAME TO AN OBJECT	64
3.9	ADDITIONAL RESOURCES	67
3.10	SELF-EVALUATION	68
CHAPTER 4 ■ Programming: Flow Control		69
4.1	IF-ELSE	69
4.1.1	Pass Statement	74
4.2	FOR LOOP	75
4.3	WHILE LOOP	77
4.4	BREAK: BREAKING THE LOOP	78
4.5	WRAPPING IT UP	80
4.5.1	Estimate the Net Charge of a Protein	80
4.5.2	Search for a Low-Degeneration Zone	81
4.6	ADDITIONAL RESOURCES	83
4.7	SELF-EVALUATION	83
CHAPTER 5 ■ Handling Files		85
5.1	READING FILES	86

5.1.1	Example of File Handling	87
5.2	WRITING FILES	89
5.2.1	File Reading and Writing Examples	90
5.3	CSV FILES	90
5.4	PICKLE: STORING AND RETRIEVING THE CONTENTS OF VARIABLES	94
5.5	JSON FILES	96
5.6	FILE HANDLING: OS, OS.PATH, SHUTIL, AND PATH.PY MODULE	98
5.6.1	path.py Module	100
5.6.2	Consolidate Multiple DNA Sequences into One FASTA File	102
5.7	ADDITIONAL RESOURCES	102
5.8	SELF-EVALUATION	103
CHAPTER 6 ■ Code Modularizing		105
6.1	INTRODUCTION TO CODE MODULARIZING	105
6.2	FUNCTIONS	106
6.2.1	Standard Way to Make Python Code Modular	106
6.2.2	Function Parameter Options	110
6.2.3	Generators	113
6.3	MODULES AND PACKAGES	114
6.3.1	Using Modules	115
6.3.2	Packages	116
6.3.3	Installing Third-Party Modules	117
6.3.4	Virtualenv: Isolated Python Environments	119
6.3.5	Conda: Anaconda Virtual Environment	121
6.3.6	Creating Modules	124
6.3.7	Testing Modules	125
6.4	ADDITIONAL RESOURCES	127
6.5	SELF-EVALUATION	128
CHAPTER 7 ■ Error Handling		129
7.1	INTRODUCTION TO ERROR HANDLING	129
7.1.1	Try and Except	131
7.1.2	Exception Types	134
7.1.3	Triggering Exceptions	135

7.2	CREATING CUSTOMIZED EXCEPTIONS	137
7.3	ADDITIONAL RESOURCES	138
7.4	SELF-EVALUATION	138
CHAPTER 8 ■ Introduction to Object Orienting Programming (OOP)		139
8.1	OBJECT PARADIGM AND PYTHON	139
8.2	EXPLORING THE JARGON	140
8.3	CREATING CLASSES	142
8.4	INHERITANCE	145
8.5	SPECIAL METHODS	149
8.5.1	Create a New Data Type Using a Built-in Data Type	154
8.6	MAKING OUR CODE PRIVATE	154
8.7	ADDITIONAL RESOURCES	155
8.8	SELF-EVALUATION	156
CHAPTER 9 ■ Introduction to Biopython		157
9.1	WHAT IS BIOPYTHON?	158
9.1.1	Project Organization	158
9.2	INSTALLING BIOPYTHON	159
9.3	BIOPYTHON COMPONENTS	162
9.3.1	Alphabet	162
9.3.2	Seq	163
9.3.3	MutableSeq	165
9.3.4	SeqRecord	166
9.3.5	Align	167
9.3.6	AlignIO	169
9.3.7	ClustalW	171
9.3.8	SeqIO	173
9.3.9	AlignIO	176
9.3.10	BLAST	177
9.3.11	Biological Related Data	187
9.3.12	Entrez	190
9.3.13	PDB	194
9.3.14	PROSITE	196
9.3.15	Restriction	197

9.3.16	SeqUtils	200
9.3.17	Sequencing	202
9.3.18	SwissProt	205
9.4	CONCLUSION	207
9.5	ADDITIONAL RESOURCES	207
9.6	SELF-EVALUATION	209

SECTION II Advanced Topics

CHAPTER 10 ■ Web Applications 213

10.1	INTRODUCTION TO PYTHON ON THE WEB	213
10.2	CGI IN PYTHON	214
10.2.1	Configuring a Web Server for CGI	215
10.2.2	Testing the Server with Our Script	215
10.2.3	Web Program to Calculate the Net Charge of a Protein (CGI version)	219
10.3	WSGI	221
10.3.1	Bottle: A Python Web Framework for WSGI	222
10.3.2	Installing Bottle	223
10.3.3	Minimal Bottle Application	223
10.3.4	Bottle Components	224
10.3.5	Web Program to Calculate the Net Charge of a Protein (Bottle Version)	229
10.3.6	Installing a WSGI Program in Apache	232
10.4	ALTERNATIVE OPTIONS FOR MAKING PYTHON-BASED DYNAMIC WEB SITES	232
10.5	SOME WORDS ABOUT SCRIPT SECURITY	232
10.6	WHERE TO HOST PYTHON PROGRAMS	234
10.7	ADDITIONAL RESOURCES	235
10.8	SELF-EVALUATION	236

CHAPTER 11 ■ XML 237

11.1	INTRODUCTION TO XML	237
11.2	STRUCTURE OF AN XML DOCUMENT	241
11.3	METHODS TO ACCESS DATA INSIDE AN XML DOCUMENT	246
11.3.1	SAX: cElementTree Iterparse	246

11.4	SUMMARY	251
11.5	ADDITIONAL RESOURCES	252
11.6	SELF-EVALUATION	252
CHAPTER 12 ■ Python and Databases		255
12.1	INTRODUCTION TO DATABASES	256
12.1.1	Database Management: RDBMS	257
12.1.2	Components of a Relational Database	258
12.1.3	Database Data Types	260
12.2	CONNECTING TO A DATABASE	261
12.3	CREATING A MYSQL DATABASE	262
12.3.1	Creating Tables	263
12.3.2	Loading a Table	264
12.4	PLANNING AHEAD	266
12.4.1	PythonU: Sample Database	266
12.5	SELECT: QUERYING A DATABASE	269
12.5.1	Building a Query	271
12.5.2	Updating a Database	273
12.5.3	Deleting a Record from a Database	273
12.6	ACCESSING A DATABASE FROM PYTHON	274
12.6.1	PyMySQL Module	274
12.6.2	Establishing the Connection	274
12.6.3	Executing the Query from Python	275
12.7	SQLITE	276
12.8	NOSQL DATABASES: MONGODB	278
12.8.1	Using MongoDB with PyMongo	278
12.9	ADDITIONAL RESOURCES	282
12.10	SELF-EVALUATION	284
CHAPTER 13 ■ Regular Expressions		285
13.1	INTRODUCTION TO REGULAR EXPRESSIONS (REGEX)	285
13.1.1	REGEX Syntax	286
13.2	THE RE MODULE	287
13.2.1	Compiling a Pattern	290
13.2.2	REGEX Examples	292

13.2.3	Pattern Replace	294
13.3	REGEX IN BIOINFORMATICS	294
13.3.1	Cleaning Up a Sequence	296
13.4	ADDITIONAL RESOURCES	297
13.5	SELF-EVALUATION	298
CHAPTER 14 ■ Graphics in Python		299
14.1	INTRODUCTION TO BOKEH	299
14.2	INSTALLING BOKEH	299
14.3	USING BOKEH	301
14.3.1	A Simple X-Y Plot	303
14.3.2	Two Data Series Plot	304
14.3.3	A Scatter Plot	306
14.3.4	A Heatmap	309
14.3.5	A Chord Diagram	309
14.4	ADDITIONAL RESOURCES	311
SECTION III Python Recipes with Commented Source Code		
CHAPTER 15 ■ Sequence Manipulation in Batch		315
15.1	PROBLEM DESCRIPTION	315
15.2	PROBLEM ONE: CREATE A FASTA FILE WITH RANDOM SEQUENCES	315
15.2.1	Commented Source Code	315
15.3	PROBLEM TWO: FILTER NOT EMPTY SEQUENCES FROM A FASTA FILE	316
15.3.1	Commented Source Code	317
15.4	PROBLEM THREE: MODIFY EVERY RECORD OF A FASTA FILE	319
15.4.1	Commented Source Code	320
CHAPTER 16 ■ Web Application for Filtering Vector Contamination		321
16.1	PROBLEM DESCRIPTION	321
16.1.1	Commented Source Code	322
16.2	ADDITIONAL RESOURCES	326
CHAPTER 17 ■ Searching for PCR Primers Using Primer3		329

17.1	PROBLEM DESCRIPTION	329
17.2	PRIMER DESIGN FLANKING A VARIABLE LENGTH REGION	330
17.2.1	Commented Source Code	331
17.3	PRIMER DESIGN FLANKING A VARIABLE LENGTH REGION, WITH BIOPYTHON	332
17.4	ADDITIONAL RESOURCES	333
CHAPTER 18 ■ Calculating Melting Temperature from a Set of Primers		335
18.1	PROBLEM DESCRIPTION	335
18.1.1	Commented Source Code	336
18.2	ADDITIONAL RESOURCES	336
CHAPTER 19 ■ Filtering Out Specific Fields from a GenBank File		339
19.1	EXTRACTING SELECTED PROTEIN SEQUENCES	339
19.1.1	Commented Source Code	339
19.2	EXTRACTING UPSTREAM REGIONS OF SELECTED GENES	340
19.2.1	Commented Source Code	340
19.3	ADDITIONAL RESOURCES	341
CHAPTER 20 ■ Inferring Splicing Sites		343
20.1	PROBLEM DESCRIPTION	343
20.1.1	Infer Splicing Sites with Commented Source Code	345
20.1.2	Sample Run of Estimate Intron Program	347
CHAPTER 21 ■ Web Server for Multiple Alignment		349
21.1	PROBLEM DESCRIPTION	349
21.1.1	Web Interface: Front-End. HTML Code	349
21.1.2	Web Interface: Server-Side Script. Commented Source Code	351
21.2	ADDITIONAL RESOURCES	353
CHAPTER 22 ■ Drawing Marker Positions Using Data Stored in a Database		355
22.1	PROBLEM DESCRIPTION	355
22.1.1	Preliminary Work on the Data	355

22.1.2	MongoDB Version with Commented Source Code	358
CHAPTER 23 ■ DNA Mutations with Restrictions		363
23.1	PROBLEM DESCRIPTION	363
23.1.1	Introduce Point Mutations and Get Restriction Profile	363
23.1.2	Running Introduce Point Mutations Program	367
23.2	ADDITIONAL RESOURCES	368
 SECTION IV Appendices		
APPENDIX A ■ Collaborative Development: Version Control with GitHub		371
A.1	INTRODUCTION TO VERSION CONTROL	372
A.2	VERSION YOUR CODE	373
A.3	SHARE YOUR CODE	380
A.4	CONTRIBUTE TO OTHER PROJECTS	386
A.5	CONCLUSION	387
A.6	METHODS	387
A.7	ADDITIONAL RESOURCES	387
 APPENDIX B ■ Install a Bottle App in PythonAnywhere		393
B.1	PYTHONANYWHERE	393
B.1.1	What Is PythonAnywhere	393
B.1.2	Installing a Web App in PythonAnywhere	393
 APPENDIX C ■ Scientific Python Cheat Sheet		403
C.1	PURE PYTHON	404
C.2	VIRTUALENV	410
C.3	CONDA	412
C.4	IPYTHON	413
C.5	NUMPY	415
C.6	MATPLOTLIB	420
C.7	SCIPY	422
C.8	PANDAS	423