

Brief Contents

Preface	v
Chapter 1	Introduction 1
Chapter 2	NumPy and SciPy 13
Chapter 3	Image Manipulation 29
Chapter 4	The Akando and Dancer Modules 37
Chapter 5	Statistics 53
Chapter 6	Parsing DNA Data Files 71
Chapter 7	Sequence Alignment 89
Chapter 8	Dynamic Programming 101
Chapter 9	Tandem Repeats 119
Chapter 10	Hidden Markov Models 125
Chapter 11	Genetic Algorithms 139
Chapter 12	Multiple Sequence Alignment 157
Chapter 13	Gapped Alignments 179
Chapter 14	Trees 197
Chapter 15	Text Mining 227
Chapter 16	Measuring Complexity 241
Chapter 17	Clustering 255
Chapter 18	Self-Organizing Maps 275
Chapter 19	Principal Component Analysis 289
Chapter 20	Species Identification 309
Chapter 21	Fourier Transforms 319
Chapter 22	Correlations 333
Chapter 23	Numerical Sequence Alignment 345
Chapter 24	Gene Expression Array Files 365

Chapter 25	Spot Finding and Measurement	375
Chapter 26	Spreadsheet Arrays and Data Displays	389
Chapter 27	Applications with Expression Arrays	401
Index		413

Contents

Preface v

Chapter 1 Introduction 1

- 1.1 The Purpose of This Book 1
- 1.2 Use of Third-Party Software 2
- 1.3 Required Background of Readers 2
- 1.4 Object-Oriented Programming 3
- 1.5 Presentation Convention 3
- 1.6 Conversion from C/C++ to Python 3
 - 1.6.1 Similarities 4
 - 1.6.2 Fundamental Python Commands that Differ from C/C++ 7
- 1.7 The Environment 11
- 1.8 Biopython 12
 - Bibliography 12

Chapter 2 NumPy and SciPy 13

- 2.1 Introduction to NumPy and SciPy 13
- 2.2 Basic Array Manipulations 13
- 2.3 Basic Math 14
- 2.4 More on Multiplication 16
- 2.5 More Math 17
 - 2.5.1 Equals or Copy 17
 - 2.5.2 Comparisons 18
 - 2.5.3 More on Slicing 20
 - 2.5.4 Sorting and Shaping 21
 - 2.5.5 Random Numbers 23
 - 2.5.6 Statistical Methods 24
- 2.6 Thinking About Problems 24
- 2.7 Array Conversions 25
- 2.8 SciPy 27
- 2.9 Summary 27
 - Bibliography 27
 - Problems 27

Chapter 3	Image Manipulation	29
3.1	The Image Module	29
3.2	Colors and Conversions	30
3.3	Digital Image Formats	31
3.4	Simple Image Manipulations	33
3.5	Conversions to and from Arrays	34
3.6	Summary	36
	Bibliography	36
	Problems	36
Chapter 4	The Akando and Dancer Modules	37
4.1	The Akando Module	37
4.1.1	Plotting Routines	37
4.1.2	Algebraic and Geometric Functions	41
4.1.3	Correlation	47
4.1.4	Image Conversions	48
4.2	The Dancer Module	48
4.3	Summary	52
	Problems	52
Chapter 5	Statistics	53
5.1	Simple Statistics	53
5.2	Distributions	55
5.3	Normalization	56
5.4	Multivariate Statistics	63
5.5	Probabilities	66
5.6	Odds	68
5.7	Decisions from Distributions	68
5.8	Summary	69
	Problems	69
Chapter 6	Parsing DNA Data Files	71
6.1	FASTA Files	71
6.2	Genbank Files	72
6.2.1	File Overview	73
6.2.2	Parsing the DNA	73
6.2.3	Gene and Protein Information	75
6.2.4	Gene Locations	76
6.2.5	Normal and Complement	77
6.2.6	Splices	78
6.2.7	Extracting All Gene Locations	79
6.2.8	Coding DNA	80
6.2.9	Proteins	82
6.2.10	Extracting Translations	83
6.3	ASN.1 File Format	84

- 6.4 Summary 87
- Bibliography 87
- Problems 87

Chapter 7 Sequence Alignment 89

- 7.1 Alphabets 89
- 7.2 Matching Sequences 90
 - 7.2.1 Perfect Matches 90
 - 7.2.2 Insertions and Deletions 90
 - 7.2.3 Rearrangements 90
 - 7.2.4 Global Versus Local Alignments 91
 - 7.2.5 Sequence Length 91
- 7.3 Simple Alignments 91
 - 7.3.1 Direct Alignment 91
 - 7.3.2 Statistical Alignment 92
 - 7.3.3 Brute Force Alignment 96
- 7.4 Summary 98
- Bibliography 98
- Problems 99

Chapter 8 Dynamic Programming 101

- 8.1 The Problem with the Brute Force Approach 101
- 8.2 The Dynamic Programming Algorithm 101
 - 8.2.1 The Scoring Matrix 102
 - 8.2.2 The Arrow Matrix 103
 - 8.2.3 Extracting the Aligned Sequences 105
- 8.3 Efficient Programming 107
 - 8.3.1 Flowing along the Diagonals 107
 - 8.3.2 Slicing Matrices 108
 - 8.3.3 Extracting Diagonal Element Locations 108
 - 8.3.4 Extracting Values from the Substitution Matrix 109
 - 8.3.5 Computing the Scoring Matrix Values for a Single Diagonal 110
 - 8.3.6 An Efficient Computation of the Scoring Matrix 110
- 8.4 Global Versus Local Alignments 112
- 8.5 Gap Penalties 114
- 8.6 Does Dynamic Programming Find the Best Alignments? 114
- 8.7 Summary 116
- Problems 117

Chapter 9 Tandem Repeats 119

- 9.1 Tandem Repeats 119
- 9.2 Hauth's Solution 119
 - 9.2.1 Foundation 119
 - 9.2.2 Multiple Words 122

	9.2.3 Tandem Repeats	123
9.3	Summary	123
	Bibliography	123
	Problems	124
Chapter 10	Hidden Markov Models	125
10.1	The Emission HMM	125
10.2	The Transition HMM	128
10.3	The Recurrent HMM	130
10.4	Constructing a Transition HMM	132
10.5	Considerations	136
	10.5.1 Assuming Data	136
	10.5.2 Spurious Strings	136
	10.5.3 Recurrent Probabilities	137
10.6	Summary	137
	Problems	137
Chapter 11	Genetic Algorithms	139
11.1	Simulated Annealing	139
11.2	The Genetic Algorithm	143
	11.2.1 Energy Surfaces	143
	11.2.2 The Genetic Algorithm Approach	144
	11.2.3 Checking the Solution	149
11.3	Nonnumerical Genetic Algorithms	149
	11.3.1 Notes on Copying	149
	11.3.2 Creating Random Arrangements	151
	11.3.3 The Genetic Algorithm	152
11.4	Summary	155
	Problems	155
Chapter 12	Multiple Sequence Alignment	157
12.1	The Greedy Approach	157
	12.1.1 Sequence Comparison	158
	12.1.2 Assembly	160
12.2	Nongreedy Approach	169
	12.2.1 Creating Genes	170
	12.2.2 Steps in the Genetic Algorithm	174
	12.2.3 The Test Run	176
	12.2.4 Improvements	177
12.3	Summary	178
	Problems	178
Chapter 13	Gapped Alignments	179
13.1	Theory of Gapped Alignments	179
13.2	Chopping the Data	180

- 13.3 Pairwise Alignments 182
- 13.4 Building the Assembly 185
 - 13.4.1 Creating New Contigs 186
 - 13.4.2 Adding to a Contig 187
 - 13.4.3 Joining Contigs 191
 - 13.4.4 Performing the Assembly 193
- 13.5 Summary 194
- Bibliography 194
- Problems 194

Chapter 14 Trees 197

- 14.1 Basic Tree Theory 197
- 14.2 Python and Trees 198
- 14.3 An Example Using UPGMA 199
- 14.4 Examples of Trees 203
 - 14.4.1 Sorting Trees 203
 - 14.4.2 Dictionary Trees 207
 - 14.4.3 Percolation Trees 209
 - 14.4.4 Suffix Trees 217
- 14.5 Decision Trees and Random Forests 220
- 14.6 Summary 224
- Problems 225

Chapter 15 Text Mining 227

- 15.1 An Introduction to Text Mining 227
- 15.2 Collecting Bioinformatic Textual Data 227
- 15.3 Creating Dictionaries 228
- 15.4 Methods of Finding Root Words 229
 - 15.4.1 Porter Stemming 230
 - 15.4.2 Suffix Trees 230
 - 15.4.3 Combining Simplified Porter Stemming with Slicing 231
- 15.5 Document Analysis 232
 - 15.5.1 Text Mining Ten Documents 232
 - 15.5.2 Word Frequency 232
 - 15.5.3 Indicative Words 236
 - 15.5.4 Document Classification 237
- 15.6 Summary 238
- Bibliography 238
- Problems 239

Chapter 16 Measuring Complexity 241

- 16.1 Linguistic Complexity 241
- 16.2 Suffix Trees 244
- 16.3 Superstrings 246

- 16.4 Summary 254
- Bibliography 254
- Problems 254

Chapter 17 Clustering 255

- 17.1 The Purpose of Clustering 255
- 17.2 k -Means Clustering 259
- 17.3 Solving More Difficult Problems 262
 - 17.3.1 Preprocessing Data 264
 - 17.3.2 Modifications of k -Means 266
- 17.4 Dynamic k -Means 268
- 17.5 Comments on k -Means 272
- 17.6 Summary 273
 - Bibliography 273
 - Problems 273

Chapter 18 Self-Organizing Maps 275

- 18.1 SOM Theory 275
- 18.2 An SOM Example 276
 - 18.2.1 Reading an Image 276
 - 18.2.2 Initializing the SOM 277
 - 18.2.3 The Best Matching Unit (BMU) 278
 - 18.2.4 Updating the SOM 280
 - 18.2.5 SOM Iterations 281
 - 18.2.6 Interpreting the SOM 282
- 18.3 Summary 286
 - Bibliography 287
 - Problems 287

Chapter 19 Principal Component Analysis 289

- 19.1 The Purpose of PCA 289
- 19.2 Eigenvectors 290
- 19.3 The PCA Process 291
 - 19.3.1 Case 1: More Dimensions than Vectors 292
 - 19.3.2 Case 2: Linear Combinations in the Data 294
 - 19.3.3 Case 3: Imperfect Dimensionality Reductions 295
 - 19.3.4 Coordinate Selection 296
- 19.4 Using SVD to Compute PCA 297
- 19.5 Describing Systems with Eigenvectors 298
- 19.6 Eigenimages 302
- 19.7 Summary 306
 - Bibliography 306
 - Problems 306

Chapter 20 Species Identification 309

- 20.1 Data Collection 309
- 20.2 The First Clustering 311
- 20.3 Using Principal Component Analysis 312
- 20.4 The Second Clustering 313
- 20.5 Using a Self-Organizing Map 314
- 20.6 Summary 317
- Bibliography 317
- Problems 317

Chapter 21 Fourier Transforms 319

- 21.1 Fourier Theory 319
- 21.2 Digital Fourier Transform 320
 - 21.2.1 DFT Theory 320
 - 21.2.2 Example with a Simple Sawtooth Signal 320
 - 21.2.3 Features of the DFT 321
 - 21.2.4 Power Spectrum 322
- 21.3 Fast Fourier Transform 322
 - 21.3.1 Duplicate Computations 322
 - 21.3.2 The FFT Method 323
 - 21.3.3 FFTs in SciPy 324
 - 21.3.4 The Swap Function 325
- 21.4 Frequency Analysis 325
 - 21.4.1 Simple Signals 325
 - 21.4.2 DNA Coding Regions 327
- 21.5 Summary 331
- Bibliography 331
- Problems 331

Chapter 22 Correlations 333

- 22.1 Correlation Theory 333
- 22.2 Random Signal Correlation 334
- 22.3 Structured Signal Correlation 335
- 22.4 Correlation of DNA Strings 337
- 22.5 Higher Dimensions 338
 - 22.5.1 Two-Dimensional FFTs in SciPy 338
 - 22.5.2 Image Frequencies 339
- 22.6 The Onset of Image Processing 341
- 22.7 Two-Dimensional Correlations 342
- Summary 343
- Bibliography 343
- Problems 343

Chapter 23	Numerical Sequence Alignment	345
23.1	Alternate Encodings	345
23.1.1	Hydrophobicity	345
23.1.2	GC Content	347
23.1.3	Numerical Methods	348
23.2	Numerical Alignments	350
23.3	Measuring the Hurst Exponent	351
23.4	Chaos Representation	354
23.4.1	Representing the Data	354
23.4.2	A Simpler Method	356
23.4.3	Comparing Chaos Images of Different Species	357
23.4.4	Organizing the Data	358
23.5	Summary	362
	Bibliography	362
	Problems	363
Chapter 24	Gene Expression Array Files	365
24.1	Raw Data	365
24.1.1	Reading Raw Data in Python	365
24.1.2	Dealing with 16-Bit Data	367
24.2	GEL Files	369
24.2.1	TIFF Headers	370
24.2.2	The Image File Directory	370
24.2.3	Reading the Data	372
24.3	Summary	373
	Bibliography	374
	Problems	374
Chapter 25	Spot Finding and Measurement	375
25.1	Spot Finding	375
25.1.1	Intensity Variations	376
25.1.2	Block Location	376
25.1.3	The Coarse Grid	380
25.1.4	Fine-Tuning the Spot Locations	381
25.2	Spot Measurements	383
25.3	Summary	386
	Bibliography	386
	Problems	386
Chapter 26	Spreadsheet Arrays and Data Displays	389
26.1	Reading Spreadsheets	389
26.1.1	The Platform File	389
26.1.2	The Z-Ratio File	390
26.1.3	Reading Two Channel Files	391

26.2	Displaying the Data	393
26.2.1	The Heat Map	393
26.2.2	The R Versus G Graph	395
26.2.3	The R/G Versus I Graph	396
26.2.4	M Versus A Graph	397
26.3	Summary	398
	Bibliography	399
	Problems	399

Chapter 27 Applications with Expression Arrays 401

27.1	LOESS Normalization	401
27.2	Expressed Genes	404
27.3	Multiple Slides	405
27.3.1	Normalization	405
27.3.2	Extracting Outliers	409
27.4	Summary	410
	Bibliography	411
	Problems	411

Index 413