

Table of Contents

Introduction 1

What This Book Does for You.....	1
Foolish Assumptions	2
How This Book Is Organized.....	2
Part I: Getting Started in Bioinformatics	3
Part II: A Survival Guide to Bioinformatics	3
Part III: Becoming a Pro in Sequence Analysis	3
Part IV: Becoming a Specialist: Advanced Bioinformatics Techniques.....	3
Part V: The Part of Tens.....	4
Icons Used in This Book.....	4
Where to Go from Here.....	4

Part I: Getting Started in Bioinformatics 7

Chapter 1: Finding Out What Bioinformatics Can Do for You 9

What Is Bioinformatics?	9
Analyzing Protein Sequences	10
A brief history of sequence analysis.....	12
Reading protein sequences from N to C.....	13
Working with protein 3-D structures.....	14
Protein bioinformatics covered in this book.....	16
Analyzing DNA Sequences	17
Reading DNA sequences the right way.....	17
The two sides of a DNA sequence.....	18
Palindromes in DNA sequences.....	20
Analyzing RNA Sequences.....	21
RNA structures: Playing with sticky strands	22
More on nucleic acid nomenclature	23
DNA Coding Regions: Pretending to Work with Protein Sequences	23
Turning DNA into proteins: The genetic code	24
More with coding DNA sequences	25
DNA/RNA bioinformatics covered in this book.....	26
Working with Entire Genomes	26
Genomics: Getting all the genes at once	27
Genome bioinformatics covered in this book	28

Chapter 2: How Most People Use Bioinformatics29

 Becoming an Instant Expert with PubMed/Medline29

 Finding out about a protein by its name30

 Searching PubMed using author’s names32

 Searching PubMed using fields.....35

 Searching PubMed using limits38

 A few more tips about PubMed41

 Retrieving Protein Sequences.....42

 ExPASy: A prime Internet site for protein information42

 More advanced ways to retrieve protein sequences45

 Retrieving a list of related protein sequences48

 Retrieving DNA Sequences.....51

 Not all DNA is coding for protein51

 Going from protein sequences to DNA sequences.....52

 Retrieving the DNA sequence relevant to my protein53

 Using BLAST to Compare My Protein Sequence

 to Other Protein Sequences57

 Making a Multiple Protein Sequence Alignment with ClustalW62

Part II: A Survival Guide to Bioinformatics.....67

Chapter 3: Using Nucleotide Sequence Databases69

 Reading into Genes and Genomes70

 Prokaryotes: Small bugs, simple genes70

 Eukaryotes: Bigger bugs, complex genes72

 Making Use (and Sense) of GenBank73

 Making sense of the GenBank entry of a prokaryotic gene73

 Making sense of the GenBank entry of an eukaryotic mRNA78

 Making sense of a GenBank eukaryotic genomic entry.....79

 Working with related GenBank entries84

 Retrieving GenBank entries without accession numbers85

 Using a Gene-Centric Database86

 Working with Whole-Genome Databases88

 Working with complete viral genomes89

 Working with complete bacterial genomes.....92

 More bacterial genomics at TIGR.....94

 Microbes from the environment at DoE96

 Exploring the Human Genome.....97

 Finding out about the Ensembl project98

Chapter 4: Using Protein and Specialized Sequence Databases . . .105

 From Translated ORFs to Mature Proteins107

 ORFs: What you see is NOT what you get107

 A personal final destination for each protein109

 A combinatorial diversity of folds and functions.....109

Reading a Swiss-Prot Entry	110
Deciphering the EGFR Swiss-Prot entry	110
General information about the entry	111
Name and origin of the protein	112
The References	114
The Comments	114
The Cross-References	116
The Keywords	118
The Features	119
Finally, the sequence itself	123
Finding Out More about Your Protein	123
Finding out more about “modified amino acids”	124
Some advanced biochemistry sites	125
Finding out more about biochemical pathways	125
Finding out more about protein structures	126
Finding out more about major protein families	127
Chapter 5: Working with a Single DNA Sequence	129
Catching Errors Before It’s Too Late	130
Removing vector sequences	130
Cases when you shouldn’t discard your sequence	133
Computing/Verifying a Restriction Map	134
Designing PCR Primers	135
Analyzing DNA Composition	138
Establishing the G+C content of your sequence	138
Counting words in DNA sequences	139
Counting long words in DNA sequences	140
Experimenting with other DNA composition analyses	142
Finding internal repeats in your sequence	142
Identifying genome-specific repeats in your sequence	145
Finding Protein-Coding Regions	145
ORFing your DNA sequence	146
Analyzing your DNA sequence with GeneMark	148
Finding internal exons in vertebrate genomic sequences	149
Complete gene parsing for eukaryotic genomes	151
Analyzing your sequence with GenomeScan	151
Assembling Sequence Fragments	153
Managing large sequencing projects with public software	154
Assembling your sequences with CAP3	155
Beyond This Chapter	157
Chapter 6: Working with a Single Protein Sequence	159
Doing Biochemistry on a Computer	160
Predicting the main physico-chemical properties of a protein	161
Interpreting ProtParam results	164
Digesting a protein in a computer	166

Doing Primary Structure Analysis.....	166
Looking for transmembrane segments.....	168
Looking for coiled-coil regions	174
Predicting Post-Translational Modifications in Your Protein.....	174
Looking for PROSITE patterns	175
Interpreting ScanProsite results	177
Finding Known Domains in Your Protein	180
Choosing the right collection of domains	182
Finding domains with InterProScan	183
Interpreting InterProScan results.....	185
Finding domains with the CD server	187
Interpreting and understanding CD server results	189
Finding domains with Motif Scan	190
Discovering New Domains in Your Proteins	194
More Protein Analysis for Free over the Internet	194

Part III: Becoming a Pro in Sequence Analysis 197

Chapter 7: Similarity Searches on Sequence Databases 199

Understanding the Importance of Similarity	200
The Most Popular Data-Mining Tool Ever: BLAST	201
BLASTing protein sequences	201
Understanding your BLAST output.....	209
BLASTing DNA sequences	216
The BLAST way of doing things.....	218
Controlling BLAST: Choosing the Right Parameters	219
Controlling the sequence masking.....	220
Changing the BLAST alignment parameters	223
Controlling the BLAST output.....	224
Making BLAST Iterative with PSI-BLAST	226
PSI-BLASTing protein sequences.....	226
Avoiding mistakes when running PSI-BLAST	228
Discovering and using protein domains with BLAST and PSI-BLAST	230
Similarity Searches for Free over the Internet	231

Chapter 8: Comparing Two Sequences 235

Making Sure You Have the Right Sequences and the Right Methods.....	236
Choosing the right sequences	236
Choosing the right method	237
Making a Dot Plot	239
Choosing the right dot-plot flavor.....	240
Using Dotlet over the Internet	241
Doing biological analysis with a dot plot	249

Making Local Alignments over the Internet.....	254
Choosing the right local-alignment flavor	255
Using Lalign to find the ten best local alignments	256
Interpreting the Lalign output	258
Making Global Alignments over the Internet.....	261
Using Lalign to Make a Global Alignment.....	262
Aligning Proteins and DNA.....	262
Free Pairwise Sequence Comparisons over the Internet.....	262
Chapter 9: Building a Multiple Sequence Alignment	265
Finding Out if a Multiple Sequence Alignment Can Help You	266
Identifying situations where multiple alignments do not help	267
Helping your research with multiple sequence alignments	267
Choosing the Right Sequences	270
The kinds of sequences you're looking for	271
Gathering your sequences with online BLAST servers	275
Choosing the Right Method of Multiple Sequence Alignment.....	281
Using ClustalW	282
Aligning sequences and structures with Tcoffee	287
Crunching large datasets with MUSCLE	291
Interpreting Your Multiple Sequence Alignment.....	291
Recognizing the good parts in a protein alignment	292
Taking your multiple alignment further	294
Comparing Sequences That You Can't Align	297
Making multiple local alignments with the Gibbs sampler	298
Searching conserved patterns.....	299
Internet Resources for Doing Multiple Sequence Comparisons	299
Making multiple alignments with ClustalW around the clock	300
Finding your favorite alignment method.....	300
Searching for motifs or patterns	301
Chapter 10: Editing and Publishing Alignments	303
Getting Your Multiple Alignment in the Right Format	305
Recognizing the main formats	307
Working with the right format	307
Converting formats	309
Watching out for lost data.....	312
Using Jalview to Edit Your Multiple Alignment Online.....	313
Starting Jalview	314
Editing a group of sequences.....	316
Useful features of Jalview	318
Saving your alignment in Jalview	318
Preparing Your Multiple Alignment for Publication	319
Using Boxshade	319
Logos.....	322

Editing and Analyzing Multiple Sequence Alignments for Free over the Internet	323
Finding multiple-sequence-alignment editors	323
Finding tools to interpret your multiple sequence alignment	324
Finding tools for beautifying your multiple alignments	325

***Part IV: Becoming a Specialist: Advanced Bioinformatics Techniques*327**

Chapter 11: Working with Protein 3-D Structures 329

From Primary to Secondary Structures	330
Predicting the secondary structure of a protein sequence	330
Predicting additional structural features	334
From the Primary Structure to the 3-D Structure	336
Retrieving and displaying a 3-D structure from a PDB site	337
Guessing the 3-D structure of your protein	340
Looking at sequence features in 3-D	343
Beyond This Chapter	350
Finding proteins with similar shapes	350
Finding other PDB viewers	350
Classifying your PDB structure	351
Doing homology modeling	351
Folding proteins in a computer	351
Threading sequences onto PDB structures	351
Looking at structures in movement	352
Predicting interactions	352

Chapter 12: Working with RNA 353

Predicting, Modeling and Drawing RNA Secondary Structures	354
Using Mfold	355
Interpreting mfold results	359
Forcing interaction in mfold	361
Searching Databases and Genomes for RNA Sequences	362
Finding tRNAs in a genome	363
Using PatScan to look for RNA patterns	363
Finding the "New" RNAs: miRNAs and siRNAs	367
Doing RNA Analysis for Free over the Internet	368
Studying evolution with ribosomal RNA	369
Finding the small, non-coding RNA you need	369
Generic RNA resources	370

Chapter 13: Building Phylogenetic Trees371

Finding Out What Phylogenetic Trees Can Do for You.....	372
Preparing Your Phylogenetic Data.....	373
Choosing the right sequences for the right tree	374
Preparing your multiple sequence alignment.....	380
Building the Kind of Tree You Need.....	383
Computing your tree.....	383
Knowing what's what in your tree.....	398
Displaying your phylogenetic tree	399
Doing Phylogeny for Free over the Internet	400
Finding online resources	400
Finding generic resources	401
Collections of orthologous genes.....	402

Part V: The Part of Tens403**Chapter 14: The Ten (Okay, Twelve) Commandments for Using Servers405**

Keep in Mind: Your Data Is Never Secure on the Web.....	406
Remember the Server, the Database, and the Program	
Version You Used.....	406
Write Down the Sequence-Identification Numbers.....	407
Write Down the Program Parameters.....	407
Save Your Internet Results the Right Way.....	407
Use E-Values.....	408
Make Sure You Can Trust Your Alignments	408
Use Different Programs to Check Borderline Results.....	409
Stay Away from Unpublished Methods!	409
Databases Are Not Like Good Wine	409
Just Because It Looks Free Doesn't Mean It Is Free . . .	410
Biting the Bullet at the Right Time.....	410

Chapter 15: Some Useful Bioinformatics Resources411

Ten Major Databases	411
Ten Major Bioinformatics Software Programs	412
Ten Major Resource Locators	414
Some Places to Find Out What's Really Going On.....	415

Index417