Contents

Chapter 1 Introduction to Bioinformatics and Sequence Analysis 1

1.1 Introduction 1

1.2 The Growth of GenBank 2

1.3 Data, Data, Everywhere 2

Further examples of human genome sequencing 4

Personal genome sequencing 4

Paleogenetics 4

Focused medical genomic studies 5

1.4 The Size of a Genome 5

1.5 Annotation 6

1.6 Witnessing Evolution Through Bioinformatics 7

Recent evolutionary changes to plants and animals 7

1.7 Large Sources of Human Sequence Variation 7

1.8 Recent Evolutionary Changes to Human Populations 8

1.9 DNA Sequence in Databases 9

Genomic DNA assembly 10

cDNA in databases—where does it come from? 12

1.10 Sequence Analysis and Data Display 14

1.11 Summary 20

Further Reading 20

Internet resources 21

Chapter 2 Introduction to Internet Resources 23

2.1 Introduction 23

2.2 The NCBI Website and ENTREZ 23

2.3 PubMed 25

2.4 Gene Name Evolution 27

2.5 OMIM 29

2.6 Retrieving Nucleotide Sequences 30

2.7 Searching Patents 31

2.8 Public Grants Database: NIH RePORTER 33

2.9 Gene Ontology 34

2.10 The Gene Database 36

2.11 UniGene 38

2.12 The UniGene Library Browser 43

2.13 Summary 44

Exercises 44

Williams syndrome and oxytocin: research with Internet tools 44

Further Reading 45

Chapter 3 Introduction to the BLAST Suite and BLASTN 47

3.1 Introduction 47

Why search a database? 47

3.2 What is BLAST? 48

How does BLAST work? 48

3.3 Your First BLAST Search 49

Find the query sequence in GenBank 49

Convert the file to another format 51

Performing BLASTN searches 52

3.4 BLAST Results 54

Graphic 54

Interpretation of the graphic 55

Results table 55

Interpretation of the table 57

The alignments 57

Other BLASTN hits from this query 60

Simultaneous review of the graphic, table, and alignments 63

3.5 BLASTN Across Species 64

BLASTN of the reference sequence for human beta hemoglobin against nonhuman transcripts 64

Paralogs, orthologs, and homologs 66

3.6 BLAST Output Format 68

3.7 Summary 68

Exercises 68

Exercise 1: Biofilm analysis 68

Exercise 2: RuBisCO 70

Further Reading 71

Internet resources 71

Chapter 4 Protein BLAST: BLASTP 73

4.1 Introduction 73

4.2 Codons and the Genetic Code 73

Memorizing the genetic code 76

4.3 Amino Acids 76

Amino acid properties 77

4.4 **BLASTP and the Scoring Matrix** 78
Building a matrix 78
4.5 **An Example BLASTP Search** 80
Retrieving protein records 81
Running BLASTP 81
The results 82
The alignments 84
Distant homologies 84
4.6 **Pairwise BLAST** 85
4.7 **Running BLASTP at the ExPASy Website** 86
Searching for pro-opiomelanocortin using a protein sequence fragment 87
Searching for repeated domains in alpha-1 collagen 91
4.8 **Summary** 94
Exercises 94
Exercise 1: Typing contest 94
Exercise 2: How mammoths adapted to cold 95
Exercise 3: Longevity genes? 96
Further Reading 97

Chapter 5 Cross-Molecular Searches: BLASTX and TBLASTN 99
5.1 **Introduction** 99
5.2 **Messenger RNA Structure** 100
5.3 **cDNA** 101
Synthesis 101
cDNA in databases 102
ESTs 103
Normalized cDNA libraries 104
An EST record 106
5.4 **BLASTX** 107
Reading frames in nucleic acids 107
A simple BLASTX search 108
A more complex BLASTX 109
Using the annotation of sequence records 115
BLASTX alignments with the reverse strand 117
5.5 **TBLASTN** 117
A TBLASTN search 118
Metagenomics and TBLASTN 120
5.6 **Summary** 122
Exercises 122
Exercise 1: Analyzing an unknown sequence 122
Exercise 2: Snake venom proteins 123
Exercise 3: Metagenomics 124
Further Reading 125

Chapter 6 Advanced Topics in BLAST 127
6.1 **Introduction** 127
6.2 **Reciprocal BLAST: Confirming Identities** 127
Demonstration of a reciprocal BLASTP 128
6.3 **Adjusting BLAST Parameters** 131
Gap cost 131
Compositional adjustments 133
6.4 **Exon Detection** 134
Exon detection with BLASTN 135
Look at the coordinates 138
Exon detection with TBLASTN 138
Orthologous exon searching with TBLASTN 141
6.5 **Repetitive DNA** 144
Simple sequences 145
Satellite DNA 145
Mini-satellites 145
LINEs and SINEs 145
Tandemly arrayed genes 146
6.6 **Interpreting Distant Relationships** 147
Name of the protein 147
Percentage identity 148
Alignment length and length similarity between query and hit 148
E value 149
Gaps 149
Conserved amino acids 150
6.7 **Summary** 152
Exercises 152
Exercise 1: Simple sequences 152
Exercise 2: Reciprocal BLAST 153
Exercise 3: Exon identification with TBLASTN 153
Exercise 4: Identification of orthologous exons with TBLAṢTN 154
Further Reading 155

Chapter 7 Bioinformatics Tools for the Laboratory 157
7.1 **Introduction** 157
7.2 **Restriction Mapping and Genetic Engineering** 158
Restriction enzymes 158
Restriction enzyme mapping: the polylinker site 160
NEBcutter 160
Generating reverse strand sequences: Reverse Complement 162
DNA translation: the ExPASy Translate tool 162

7.3 Finding Open Reading Frames 163
The NCBI ORF Finder 163
7.4 PCR and Primer Design Tools 165
Primer3 166
Primer-BLAST 169
7.5 Measuring DNA and Protein Composition 170
DNA Stats 170
Composition/Molecular Weight Calculation Form 171
7.6 Asking Very Specific Questions: The Sequence Retrieval System (SRS) 172
7.7 DotPlot 174
DotPlot of alternative transcripts 175
DotPlots of orthologous genes 176
7.8 Summary 179
Exercises 179
Spider silk: a workflow of analysis 179
Further Reading 181

Chapter 8 Protein Analysis 183
8.1 Introduction 183
8.2 Finding Functional Patterns 183
A repeating pattern within a zinc finger 184
8.3 Annotating an Unknown Sequence 187
A zinc protease pattern 188
The ADAM_MEPRO profile 188
8.4 Looking at Three-dimensional Protein Structures 190
Jmol: a protein structure viewer 192
Exploring and understanding a structure 193
Jmol scripting 194
8.5 ProPhylER 195
The Interface view 196
The CrystalPainter view 198
8.6 The Impact of Sequence on Structure 201
8.7 Building Blocks: A Multiple Domain Protein 204
8.8 Post-translational Modification 204
Secretion signals 206
Prediction of protein glycosylation sites 208
8.9 Transmembrane Domain Detection 208
8.10 Summary 211
Exercises 211
Aquaporin-5 211
Further Reading 213
Internet resources 214

Chapter 9 Explorations of Short Nucleotide Sequences 215
9.1 Introduction 215
9.2 Transcription Factor Binding Sites 216
Transfac 216
Identifying other binding sites for the estrogen receptor 219
Predicting transcription factor binding sites 220
An experiment with MATCH 221
An experiment with PATCH 224
9.3 Translation Initiation: The Kozak Sequence 226
9.4 Viewing Whole Genes 228
9.5 Exon Splicing 231
Renin: a striking example of a small exon 234
Another striking splice: human *ISG15* ubiquitin-like modifier 235
Alternative splicing 236
Human plectin: alternative splicing at the 5P end 237
Consensus splice junctions, translated 238
9.6 Polyadenylation Signals 239
9.7 Summary 240
Exercises 242
Inhibitor of Kappa light polypeptide gene enhancer in B-cells (*IKBKAP*) 242
Further Reading 243

Chapter 10 MicroRNAs and Pathway Analysis 245
10.1 Introduction 245
10.2 miRNA Function 245
10.3 miRNA Nomenclature 247
10.4 miRNA Families and Conservation 247
10.5 Structure and Processing of miRNAs 248
10.6 miRBase: The Repository for miRNAs 250
10.7 Numbers and Locations 251
10.8 Linking miRNA Analysis to a Biochemical Pathway: Gastric Cancer 251
10.9 KEGG: Biological Networks at Your Fingertips 253
miRNAs in the cell cycle pathway 255
10.10 TarBase: Experimentally Verified miRNA Inhibition 256
Verified miRNA-driven translation repression 256

10.11 TargetScan: miRNA Target Site Prediction **258**
TargetScan predictions for cell cycle transcripts 260
10.12 Expanding miRNA Regulation of the Cell Cycle Using TarBase and TargetScan **263**
10.13 Making Sense of miRNAs and Their Many Predicted Targets **265**
10.14 miRNAs Associated With Diseases **266**
10.15 Summary **267**
Exercises **267**
GDF8 267
Further Reading **269**

Chapter 11 Multiple Sequence Alignments **271**
11.1 Introduction **271**
11.2 Multiple Sequence Alignments Through NCBI BLAST **271**
11.3 ClustalW from the ExPASy Website **274**
11.4 ClustalW at the EMBL-EBI Server **276**
MARK1 kinase 277
MAPK15 kinase 280
DNA versus protein identities 282
11.5 Modifying ClustalW Parameters **282**
Gap-opening penalty 282
The clustering method 283
11.6 Comparing ClustalW, MUSCLE, and COBALT **286**
11.7 Isoform Alignment Problem: Internal Splicing **288**
11.8 Aligning Paralog Domains **292**
11.9 Manually Editing a Multiple Sequence Alignment **294**
Jalview 294
Editing with a word processor 296
11.10 Summary **296**
Exercises **296**
FOXP2 296
Further Reading **297**

Chapter 12 Browsing the Genome **299**
12.1 Introduction **299**
12.2 Chromosomes **299**
Human chromosome statistics 300
Chromosome details and comparisons 302
12.3 Synteny **303**
Synteny of the sex chromosomes 304
12.4 The UCSC Genome Browser **304**
OPN5: a sample gene to browse 305
Simple view changes in the UCSC Genome Browser 308
Configuring the UCSC Genome Browser window 310
Searching genomes and adding tracks through BLAT 312
Viewing the Multiz alignments 314
Zooming out: seeing the big picture 316
Very large genes: dystrophin and titin 318
Gene density 320
Interspecies comparison of genomes 323
The beta globin locus 324
12.5 Summary **325**
Exercises **325**
Olfactory genes 325
Further Reading **327**

Appendix 1 Formatting Your Report **329**
A1.1 Introduction **329**
A1.2 Font Choice and Pasting Issues **329**
A1.3 Find and Replace **331**
Changing file format 332
A1.4 Hypertext **333**
Creating hypertext 334
Selecting a column of text 334
A1.5 Summary **334**

Appendix 2 Running NCBI BLAST in "batch" Mode **337**

Abbreviations **340**
Glossary **341**
Web Resources **344**
Index **347**