

# CONTENTS

Preface	xi	2.5 OMIM	29
<b>Chapter 1</b>		<b>2.6 RETRIEVING NUCLEOTIDE SEQUENCES</b>	<b>31</b>
<b>Introduction to Bioinformatics and Sequence Analysis</b>	<b>1</b>	The NR and RefSeq Databases	32
		A RefSeq mRNA Record	33
<b>1.1 INTRODUCTION</b>	<b>1</b>	<b>2.7 THE GENE DATABASE</b>	<b>36</b>
<b>1.2 DATA, DATA, EVERYWHERE</b>	<b>2</b>	The NCBI Taxonomy Database	39
Origins of the Sequence Data	2	<b>2.8 SUMMARY</b>	<b>41</b>
Metagenomics	3	<b>EXERCISES</b>	<b>41</b>
Species Surveillance	3	<b>FURTHER READING</b>	<b>42</b>
Invasive Species Detection	3		
The Barcode of Life	4	<b>Chapter 3</b>	
Medically Directed Efforts	4	<b>Introduction to the BLAST Suite and BLASTN</b>	<b>43</b>
Paleogenomics	5		
<b>1.3 THE SIZE OF A GENOME</b>	<b>5</b>	<b>3.1 INTRODUCTION</b>	<b>43</b>
<b>1.4 SEQUENCE ANNOTATION</b>	<b>6</b>	Why Search a Database?	43
<b>1.5 WITNESSING EVOLUTION THROUGH BIOINFORMATICS</b>	<b>7</b>	<b>3.2 WHAT IS BLAST?</b>	<b>44</b>
Recent Evolutionary Changes to Plants and Animals	7	How Does BLAST Work?	44
<b>1.6 LARGE SOURCES OF HUMAN SEQUENCE VARIATION</b>	<b>8</b>	<b>3.3 YOUR FIRST BLAST SEARCH</b>	<b>45</b>
<b>1.7 RECENT EVOLUTIONARY CHANGES TO HUMAN POPULATIONS</b>	<b>8</b>	Find the Query Sequence in GenBank	45
<b>1.8 DNA SEQUENCE IN DATABASES</b>	<b>10</b>	Convert the File to Another Format	47
Genomic DNA Assembly	10	Performing BLASTN Searches	49
cDNA in Databases—Where Does It Come From?	13	<b>3.4 BLAST RESULTS</b>	<b>50</b>
<b>1.9 SEQUENCE ANALYSIS AND DATA DISPLAY</b>	<b>15</b>	Graphic Summary	50
<b>1.10 SUMMARY</b>	<b>22</b>	Interpretation of the Graphic	51
<b>FURTHER READING</b>	<b>22</b>	Descriptions	51
		Interpretation of the Table	53
		The Alignments	53
		Other BLASTN Hits from This Query	56
		Simultaneous Review of the Graphic, Table, and Alignments	58
		<b>3.5 BLASTN ACROSS SPECIES</b>	<b>59</b>
		BLASTN of the Reference Sequence for Human Beta Globin against Nonhuman Transcripts	59
		Paralogs, Orthologs, and Homologs	63
		<b>3.6 MORE DISTANTLY RELATED HITS</b>	<b>64</b>
		<b>3.7 BLAST OUTPUT FORMAT</b>	<b>67</b>
		<b>3.8 SUMMARY</b>	<b>68</b>
<b>Chapter 2</b>			
<b>Introduction to Internet Resources</b>	<b>23</b>		
<b>2.1 INTRODUCTION</b>	<b>23</b>		
<b>2.2 THE NCBI WEBSITE AND ENTREZ</b>	<b>23</b>		
<b>2.3 PUBMED</b>	<b>26</b>		
<b>2.4 GENE NAME EVOLUTION</b>	<b>29</b>		

<b>EXERCISES</b>	<b>69</b>	<b>5.3 cDNA</b>	<b>99</b>
Exercise 1: BLASTN vs. megaBLAST	69	Synthesis	99
Exercise 2: An Unknown	69	cDNA in Databases	100
Exercise 3: Reference Sequences vs. the Nucleotide Database	70	Normalized cDNA Libraries	103
<b>FURTHER READING</b>	<b>71</b>	<b>5.4 BLASTX</b>	<b>104</b>
<b>INTERNET RESOURCES</b>	<b>71</b>	Reading Frames in Nucleic Acids	104
		A Simple BLASTX Search	105
		A BLASTX with an Unknown	106
		Using the Annotation of Sequence Records	112
		BLASTX Alignments with the Reverse Strand	115
<b>Chapter 4</b>		<b>5.5 TBLASTN</b>	<b>115</b>
<b>Protein BLAST: BLASTP</b>	<b>73</b>	A TBLASTN Search	116
		Metagenomics and TBLASTN	118
<b>4.1 INTRODUCTION</b>	<b>73</b>	<b>5.6 SUMMARY</b>	<b>121</b>
<b>4.2 CODONS AND THE GENETIC CODE</b>	<b>73</b>	<b>EXERCISES</b>	<b>121</b>
Memorizing the Genetic Code	76	Exercise 1: Analyzing an Unknown Sequence	121
<b>4.3 AMINO ACIDS</b>	<b>76</b>	Exercise 2: Identify the Origin of the Alignments	121
Amino Acid Properties	77	Exercise 3: Snake Venom Proteins: EST Identification Using BLASTX	121
<b>4.4 BLASTP AND THE SCORING MATRIX</b>	<b>78</b>	<b>FURTHER READING</b>	<b>123</b>
Building a Matrix	78		
<b>4.5 A SAMPLE BLASTP SEARCH</b>	<b>81</b>		
Retrieving Protein Records	81		
Running BLASTP	82		
The Filter Results Window	82		
The Alignments	85		
Distant Homologies	86		
<b>4.6 ALIGNING TWO OR MORE SEQUENCES WITH BLAST</b>	<b>87</b>		
<b>4.7 YOUR QUERY IS AN UNKNOWN: WHAT IS IT?</b>	<b>88</b>	<b>Chapter 6</b>	
<b>4.8 SUMMARY</b>	<b>92</b>	<b>Advanced Topics in BLAST</b>	<b>125</b>
<b>EXERCISES</b>	<b>92</b>		
Exercise 1: Typing Contest	92	<b>6.1 INTRODUCTION</b>	<b>125</b>
Exercise 2: How Mammoths Adapted to Cold	93	<b>6.2 RECIPROCAL BLAST: CONFIRMING IDENTITIES</b>	<b>125</b>
Exercise 3: Identifying Orthologs	94	Demonstration of a Reciprocal BLASTP	126
<b>FURTHER READING</b>	<b>95</b>	<b>6.3 ADJUSTING BLAST PARAMETERS</b>	<b>128</b>
		Gap Cost	129
		<b>6.4 EXON DETECTION</b>	<b>135</b>
		Exon Detection with BLASTN	135
		Looking at the Coordinates	139
		Exon Detection with TBLASTN	140
		Orthologous Exon Searching with TBLASTN	143
		<b>6.5 REPETITIVE DNA</b>	<b>145</b>
		Simple Sequences	147
		Satellite DNA	147
		Mini-satellites	147
		LINES and SINES	147
		Tandemly Arrayed Genes	149
<b>Chapter 5</b>			
<b>Cross-Molecular Searches: BLASTX and TBLASTN</b>	<b>97</b>		
<b>5.1 INTRODUCTION</b>	<b>97</b>		
<b>5.2 MESSENGER RNA STRUCTURE</b>	<b>97</b>		

<b>6.6 INTERPRETING DISTANT RELATIONSHIPS</b>	<b>149</b>	<b>7.8 SUMMARY</b>	<b>181</b>
Name of the Protein	150	<b>EXERCISE</b>	<b>181</b>
Percentage Identity	150	Spider Silk: A Workflow of Analysis	181
Alignment Length and Length Similarity between Query and Subject	150	<b>FURTHER READING</b>	<b>183</b>
E Value	151		
Gaps	153	<b>Chapter 8</b>	
Conserved Amino Acids	154	<b>Protein Analysis</b>	<b>185</b>
<b>6.7 SUMMARY</b>	<b>154</b>		
<b>EXERCISES</b>	<b>155</b>	<b>8.1 INTRODUCTION</b>	<b>185</b>
Exercise 1: Simple Sequences	155	<b>8.2 FINDING FUNCTIONAL PATTERNS</b>	<b>185</b>
Exercise 2: Reciprocal BLAST	155	A Repeating Pattern within a Zinc Finger	186
Exercise 3: Exon Identification with TBLASTN	156	<b>8.3 ASSISTANCE IN SEQUENCE ANNOTATION</b>	<b>190</b>
<b>FURTHER READING</b>	<b>156</b>	A Zinc Protease Pattern	191
		A Domain Profile	192
		Domain Architecture View	194
		<b>8.4 LOOKING AT THREE-DIMENSIONAL PROTEIN STRUCTURES</b>	<b>194</b>
<b>Chapter 7</b>		JSmol: A Protein Structure Viewer	197
<b>Bioinformatics Tools for the Laboratory</b>	<b>157</b>	Exploring and Understanding a Structure	198
		JSmol Scripting	199
<b>7.1 INTRODUCTION</b>	<b>157</b>	UniProtKB Structures	200
<b>7.2 RESTRICTION MAPPING AND GENETIC ENGINEERING</b>	<b>158</b>	<b>8.5 THE IMPACT OF SEQUENCE ON STRUCTURE</b>	<b>202</b>
Restriction Enzymes	158	<b>8.6 BUILDING BLOCKS: A MULTIPLE-DOMAIN PROTEIN</b>	<b>204</b>
Restriction Enzyme Mapping: The Polylinker Site	160	<b>8.7 POST-TRANSLATIONAL MODIFICATION</b>	<b>205</b>
NEBcutter	160	Secretion Signals	205
Generating Reverse Strand Sequences: Reverse Complement	162	Prediction of Protein Glycosylation Sites	206
DNA Translation: The ExpASy Translate Tool	162	<b>8.8 TRANSMEMBRANE DOMAIN DETECTION</b>	<b>208</b>
<b>7.3 FINDING OPEN READING FRAMES</b>	<b>163</b>	<b>8.9 SUMMARY</b>	<b>210</b>
The NCBI ORF Finder	164	<b>EXERCISES</b>	<b>210</b>
<b>7.4 PCR AND PRIMER DESIGN TOOLS</b>	<b>166</b>	Aquaporin-5	210
Primer-BLAST	167	<b>FURTHER READING</b>	<b>212</b>
<b>7.5 MEASURING DNA AND PROTEIN COMPOSITION</b>	<b>169</b>		
DNA Stats	169	<b>Chapter 9</b>	
Composition/Molecular Weight Calculation Form	169	<b>Explorations of Short Nucleotide Sequences</b>	<b>213</b>
<b>7.6 QUERIES AT THE UNIPROTKB WEBSITE</b>	<b>171</b>		
<b>7.7 DotPlots</b>	<b>176</b>	<b>9.1 INTRODUCTION</b>	<b>213</b>
DotPlot of Alternative Transcripts	177	<b>9.2 SINGLE NUCLEOTIDE POLYMORPHISMS</b>	<b>214</b>
Using DotPlots to Understand Protein Structure	178	How Are SNPs Represented in a Database?	215
DotPlots between Orthologous Genes	178	Viewing SNP Data at the Ensembl Website	217

SNP Categories and Distribution within an Entire Gene	219	10.12 USING TarBase TO IDENTIFY CELL CYCLE miRNAs	254
How Are SNPs Distributed?	219	TargetScan Predictions for Cell Cycle Transcripts	255
<b>9.3 VIEWING FEATURES WITHIN WHOLE GENES</b>	<b>220</b>	<b>10.13 EXPANDING miRNA REGULATION OF THE CELL CYCLE USING TarBase AND TargetScan</b>	<b>259</b>
<b>9.4 TRANSLATION INITIATION: THE KOZAK SEQUENCE</b>	<b>222</b>	<b>10.14 MAKING SENSE OF miRNAs AND THEIR MANY PREDICTED TARGETS</b>	<b>261</b>
<b>9.5 EXON SPLICING</b>	<b>224</b>	<b>10.15 miRNAs ASSOCIATED WITH DISEASES</b>	<b>262</b>
Renin: A Striking Example of a Small Exon	226	PubMed	262
Another Striking Splice: Human ISG15 Ubiquitin-Like Modifier	228	miRBase	262
Alternative Splicing	229	<b>10.16 LONG NON-CODING RNA (LNCRNA)</b>	<b>263</b>
Human Plectin: Alternative Splicing at the 5P End	231	<b>10.17 SUMMARY</b>	<b>264</b>
Consensus Splice Junctions, Translated	232	<b>EXERCISES</b>	<b>264</b>
<b>9.6 POLYADENYLATION SIGNALS</b>	<b>234</b>	Exercise 1: miRBase	264
<b>9.7 SUMMARY</b>	<b>236</b>	Exercise 2: Texel Sheep Mutation	264
<b>EXERCISES</b>	<b>236</b>	<b>FURTHER READING</b>	<b>265</b>
Elongator Acetyltransferase Complex Subunit 1 ( <i>ELP1</i> )	236	<b>Chapter 11</b>	
<b>FURTHER READING</b>	<b>237</b>	<b>Multiple Sequence Alignments</b>	<b>267</b>
<b>Chapter 10</b>		<b>11.1 INTRODUCTION</b>	<b>267</b>
<b>MicroRNAs</b>	<b>239</b>	<b>11.2 MULTIPLE SEQUENCE ALIGNMENTS THROUGH NCBI BLAST</b>	<b>267</b>
<b>10.1 INTRODUCTION</b>	<b>239</b>	<b>11.3 CLUSTAL OMEGA FROM THE UNIPROT WEBSITE</b>	<b>270</b>
<b>10.2 miRNA FUNCTION</b>	<b>239</b>	<b>11.4 CLUSTAL OMEGA AT THE EMBL-EBI SERVER</b>	<b>274</b>
<b>10.3 miRNA NOMENCLATURE</b>	<b>241</b>	MARK1 Kinase	274
<b>10.4 miRNA FAMILIES AND CONSERVATION</b>	<b>241</b>	MAPK15 Kinase	276
<b>10.5 STRUCTURE AND PROCESSING OF miRNAs</b>	<b>243</b>	DNA versus Protein Identities	279
<b>10.6 miRBase: THE REPOSITORY FOR miRNAs</b>	<b>244</b>	<b>11.5 COBALT, THE NCBI MULTIPLE SEQUENCE ALIGNMENT TOOL</b>	<b>279</b>
<b>10.7 NUMBERS AND LOCATIONS</b>	<b>245</b>	Gap-Opening Penalty	281
<b>10.8 TarBase: EXPERIMENTALLY OBSERVED miRNA TARGETING OF mRNAs</b>	<b>247</b>	<b>11.6 ISOFORM ALIGNMENT PROBLEM: INTERNAL SPLICING</b>	<b>284</b>
<b>10.9 TargetScan: miRNA TARGET SITE PREDICTION</b>	<b>247</b>	<b>11.7 ALIGNING PARALOG DOMAINS</b>	<b>284</b>
<b>10.10 LINKING miRNA ANALYSIS TO A BIOCHEMICAL PATHWAY: GASTRIC CANCER</b>	<b>249</b>	<b>11.8 MANUALLY EDITING A MULTIPLE SEQUENCE ALIGNMENT</b>	<b>288</b>
<b>10.11 KEGG: BIOLOGICAL NETWORKS AT YOUR FINGERTIPS</b>	<b>250</b>		

<b>11.9 SUMMARY</b>	<b>289</b>	Searching Genomes and Adding Tracks through BLAT	314
<b>EXERCISES</b>	<b>289</b>		
FOXP2	289		
<b>FURTHER READING</b>	<b>290</b>		
<b>Chapter 12</b>			
<b>Browsing the Genome</b>	<b>291</b>		
<b>12.1 INTRODUCTION</b>	<b>291</b>		
<b>12.2 CHROMOSOMES</b>	<b>291</b>		
Human Chromosome Statistics	292		
Chromosome Details and Comparisons	293		
<b>12.3 SYNTENY</b>	<b>296</b>		
Synteny of the Sex Chromosomes	296		
<b>12.4 THE UCSC GENOME BROWSER</b>	<b>299</b>		
<i>WWCI</i> : A Sample Gene to Browse	300		
Simple View Changes in the UCSC Genome Browser	302		
Data Exploration	302		
Viewing Details of the Alignments	305		
Very Large Genes: Dystrophin and Titin	309		
Gene Density	311		
Interspecies Comparison of Genomes	312		
The Beta Globin Locus	312		
		<b>12.5 SUMMARY</b>	<b>316</b>
		<b>EXERCISES</b>	<b>316</b>
		Olfactory Genes	316
		<b>FURTHER READING</b>	<b>318</b>
		<b>Appendix: Formatting Your Report</b>	<b>319</b>
		<b>A.1 INTRODUCTION</b>	<b>319</b>
		<b>A.2 FONT CHOICE AND PASTING ISSUES</b>	<b>319</b>
		<b>A.3 FIND AND REPLACE</b>	<b>320</b>
		<b>A.4 CHANGING FILE FORMAT</b>	<b>323</b>
		<b>A.5 HYPERTEXT</b>	<b>324</b>
		<b>A.6 SELECTING A COLUMN OF TEXT</b>	<b>325</b>
		<b>A.7 SUMMARY</b>	<b>325</b>
		<b>Abbreviations</b>	<b>327</b>
		<b>Glossary</b>	<b>329</b>
		<b>Web Resources</b>	<b>333</b>
		<b>Index</b>	<b>335</b>