
Contents

List of Code Challenges	xviii
About the Textbook	xxi
Meet the Authors	xxi
Meet the Development Team	xxii
Acknowledgments	xxiii
1 Where in the Genome Does DNA Replication Begin?	2
A Journey of a Thousand Miles.	3
Hidden Messages in the Replication Origin	5
<i>DnaA</i> boxes	5
Hidden messages in “The Gold-Bug”	6
Counting words	7
The Frequent Words Problem	8
Frequent words in <i>Vibrio cholerae</i>	10
Some Hidden Messages are More Surprising than Others	11
An Explosion of Hidden Messages	13
Looking for hidden messages in multiple genomes	13
The Clump Finding Problem	14
The Simplest Way to Replicate DNA	16
Asymmetry of Replication	18
Peculiar Statistics of the Forward and Reverse Half-Strands	22
Deamination	22
The skew diagram	23
Some Hidden Messages are More Elusive than Others	26
A Final Attempt at Finding <i>DnaA</i> Boxes in <i>E. coli</i>	29
Epilogue: Complications in <i>oriC</i> Predictions	31

Open Problems	33
Multiple replication origins in a bacterial genome	33
Finding replication origins in archaea	35
Finding replication origins in yeast	36
Computing probabilities of patterns in a string	37
Charging Stations	39
The frequency array	39
Converting patterns to numbers and vice-versa	41
Finding frequent words by sorting	43
Solving the Clump Finding Problem	44
Solving the Frequent Words with Mismatches Problem	47
Generating the neighborhood of a string	49
Finding frequent words with mismatches by sorting	51
Detours	52
Big-O notation	52
Probabilities of patterns in a string	52
The most beautiful experiment in biology	57
Directionality of DNA strands	59
The Towers of Hanoi	60
The overlapping words paradox	62
Bibliography Notes	64
2 Which DNA Patterns Play the Role of Molecular Clocks?	66
Do We Have a “Clock” Gene?	67
Motif Finding Is More Difficult Than You Think	68
Identifying the evening element	68
Hide and seek with motifs	69
A brute force algorithm for motif finding	71
Scoring Motifs	72
From motifs to profile matrices and consensus strings	72
Towards a more adequate motif scoring function	75
Entropy and the motif logo	76
From Motif Finding to Finding a Median String	77
The Motif Finding Problem	77
Reformulating the Motif Finding Problem	77
The Median String Problem	80
Why have we reformulated the Motif Finding Problem?	82

Greedy Motif Search	83
Using the profile matrix to roll dice	83
Analyzing greedy motif finding	85
Motif Finding Meets Oliver Cromwell	86
What is the probability that the sun will not rise tomorrow?	86
Laplace’s Rule of Succession	87
An improved greedy motif search	88
Randomized Motif Search	91
Rolling dice to find motifs	91
Why randomized motif search works	93
How Can a Randomized Algorithm Perform So Well?	96
Gibbs Sampling	98
Gibbs Sampling in Action	100
Epilogue: How Does Tuberculosis Hibernate to Hide from Antibiotics?	104
Charging Stations	107
Solving the Median String Problem	107
Detours	108
Gene expression	108
DNA arrays	108
Buffon’s needle	109
Complications in motif finding	112
Relative entropy	112
Bibliography Notes	115
3 How Do We Assemble Genomes?	115
Exploding Newspapers	117
The String Reconstruction Problem	120
Genome assembly is more difficult than you think	120
Reconstructing strings from k -mers	120
Repeats complicate genome assembly	123
String Reconstruction as a Walk in the Overlap Graph	124
From a string to a graph	124
The genome vanishes	127
Two graph representations	129
Hamiltonian paths and universal strings	130
Another Graph for String Reconstruction	131
Gluing nodes and de Bruijn graphs	131

Walking in the de Bruijn Graph	134
Eulerian paths	134
Another way to construct de Bruijn graphs	135
Constructing de Bruijn graphs from k -mer composition	137
De Bruijn graphs versus overlap graphs	138
The Seven Bridges of Königsberg	139
Euler’s Theorem	142
From Euler’s Theorem to an Algorithm for Finding Eulerian Cycles	146
Constructing Eulerian cycles	146
From Eulerian cycles to Eulerian paths	146
Constructing universal strings	147
Assembling Genomes from Read-Pairs	150
From reads to read-pairs	150
Transforming read-pairs into long reads	151
From composition to paired composition	153
Paired de Bruijn graphs	154
Complications of paired de Bruijn graphs	155
Epilogue: Genome Assembly Faces Real Sequencing Data	158
Breaking reads into k -mers	158
Splitting the genome into contigs	159
Assembling error-prone reads	161
Inferring multiplicities of edges in de Bruijn graphs	162
Charging Stations	164
The effect of gluing on the adjacency matrix	164
Generating all Eulerian cycles	165
Reconstructing a string spelled by a path in the paired de Bruijn graph	166
Maximal non-branching paths in a graph	169
Detours	170
A short history of DNA sequencing technologies	170
Repeats in the human genome	172
Graphs	173
The icosian game	175
Tractable and intractable problems	176
From Euler to Hamilton to de Bruijn	177
The seven bridges of Kaliningrad	178
The BEST Theorem	179
Bibliography Notes	180

4	How Do We Sequence Antibiotics?	182
	The Discovery of Antibiotics	183
	How Do Bacteria Make Antibiotics?	184
	How peptides are encoded by the genome	184
	Where is Tyrocidine encoded in the <i>Bacillus brevis</i> genome?	186
	From linear to cyclic peptides	188
	Dodging the Central Dogma of Molecular Biology	188
	Sequencing Antibiotics by Shattering Them into Pieces	190
	Introduction to mass spectrometry	190
	The Cyclopeptide Sequencing Problem	191
	A Brute Force Algorithm for Cyclopeptide Sequencing	193
	A Branch-and-Bound Algorithm for Cyclopeptide Sequencing	194
	Mass Spectrometry Meets Golf	197
	From theoretical to real spectra	197
	Adapting cyclopeptide sequencing for spectra with errors	198
	From 20 to More than 100 Amino Acids	201
	The Spectral Convolution Saves the Day	203
	Epilogue: From Simulated to Real Spectra	205
	Open Problems	208
	The Beltway and Turnpike Problems	208
	Sequencing cyclic peptides in primates	209
	Charging Stations	211
	Generating the theoretical spectrum of a peptide	211
	How fast is CYCLOPEPTIDSEQUENCING ?	212
	Trimming the peptide leaderboard	214
	Detours	215
	Gause and Lysenkoism	215
	Discovery of codons	216
	Quorum sensing	217
	Molecular mass	217
	Selenocysteine and pyrrolysine	218
	Pseudo-polynomial algorithm for the Turnpike Problem	219
	Split genes	220
	Bibliography Notes	221
5	How Do We Compare Biological Sequences?	222
	Cracking the Non-Ribosomal Code	223

The RNA Tie Club	223
From protein comparison to the non-ribosomal code	224
What do oncogenes and growth factors have in common?	225
Introduction to Sequence Alignment	226
Sequence alignment as a game	226
Sequence alignment and the longest common subsequence	227
The Manhattan Tourist Problem	229
What is the best sightseeing strategy?	229
Sightseeing in an arbitrary directed graph	232
Sequence Alignment is the Manhattan Tourist Problem in Disguise	233
An Introduction to Dynamic Programming: The Change Problem	236
Changing money greedily	236
Changing money recursively	237
Changing money using dynamic programming	239
The Manhattan Tourist Problem Revisited	241
From Manhattan to an Arbitrary Directed Acyclic Graph	245
Sequence alignment as building a Manhattan-like graph	245
Dynamic programming in an arbitrary DAG	246
Topological orderings	247
Backtracking in the Alignment Graph	251
Scoring Alignments	253
What is wrong with the LCS scoring model?	253
Scoring matrices	254
From Global to Local Alignment	255
Global alignment	255
Limitations of global alignment	257
Free taxi rides in the alignment graph	259
The Changing Faces of Sequence Alignment	261
Edit distance	261
Fitting alignment	263
Overlap alignment	263
Penalizing Insertions and Deletions in Sequence Alignment	264
Affine gap penalties	264
Building Manhattan on three levels	266
Space-Efficient Sequence Alignment	269
Computing alignment score using linear memory	269
The Middle Node Problem	270

A surprisingly fast and memory-efficient alignment algorithm	273
The Middle Edge Problem	275
Epilogue: Multiple Sequence Alignment	277
Building a three-dimensional Manhattan	277
A greedy multiple alignment algorithm	280
Detours	282
Fireflies and the non-ribosomal code	282
Finding an LCS without constructing a city	283
Constructing a topological ordering	284
PAM scoring matrices	285
Divide-and-conquer algorithms	287
Scoring multiple alignments	289
Bibliography Notes	291
6 Are There Fragile Regions in the Human Genome?	292
Of Mice and Men	293
How different are the human and mouse genomes?	293
Synteny blocks	294
Reversals	294
Rearrangement hotspots	295
The Random Breakage Model of Chromosome Evolution	297
Sorting by Reversals	299
A Greedy Heuristic for Sorting by Reversals	304
Breakpoints	306
What are breakpoints?	306
Counting breakpoints	307
Sorting by reversals as breakpoint elimination	308
Rearrangements in Tumor Genomes	310
From Unichromosomal to Multichromosomal Genomes	311
Translocations, fusions, and fissions	311
From a genome to a graph	313
2-breaks	314
Breakpoint Graphs	316
Computing the 2-Break Distance	320
Rearrangement Hotspots in the Human Genome	323
The Random Breakage Model meets the 2-Break Distance Theorem	323
The Fragile Breakage Model	324

Epilogue: Synteny Block Construction	325
Genomic dot-plots	325
Finding shared k -mers	326
Constructing synteny blocks from shared k -mers	329
Synteny blocks as connected components in graphs	331
Open Problem: Can Rearrangements Shed Light on Bacterial Evolution? . . .	333
Charging Stations	335
From genomes to the breakpoint graph	335
Solving the 2-Break Sorting Problem	338
Detours	340
Why is the gene content of mammalian X chromosomes so conserved? .	340
Discovery of genome rearrangements	340
The exponential distribution	341
Bill Gates and David X. Cohen flip pancakes	342
Sorting linear permutations by reversals	343
Bibliography Notes	346
Bibliography	349
Image Courtesies	355

List of Code Challenges

Chapter 1	2
(1A) Compute the Number of Times a Pattern Appears in a Text	8
(1B) Find the Most Frequent Words in a String	8
(1C) Find the Reverse Complement of a DNA String	12
(1D) Find All Occurrences of a Pattern in a String	13
(1E) Find Patterns Forming Clumps in a String	15
(1F) Find a Position in a Genome Minimizing the Skew	25
(1G) Compute the Hamming Distance Between Two Strings	27
(1H) Find All Approximate Occurrences of a Pattern in a String	27
(1I) Find the Most Frequent Words with Mismatches in a String	28
(1J) Find Frequent Words with Mismatches and Reverse Complements	29
(1K) Generate the Frequency Array of a String	40
(1L) Implement <code>PATTERNTONUMBER</code>	42
(1M) Implement <code>NUMBERTOPATTERN</code>	43
(1N) Generate the d -Neighborhood of a String	50
Chapter 2	66
(2A) Implement <code>MOTIFENUMERATION</code>	71
(2B) Find a Median String	81
(2C) Find a <i>Profile</i> -most Probable k -mer in a String	85
(2D) Implement <code>GREEDYMOTIFSEARCH</code>	85
(2E) Implement <code>GREEDYMOTIFSEARCH</code> with Pseudocounts	91
(2F) Implement <code>RANDOMIZEDMOTIFSEARCH</code>	93
(2G) Implement <code>GIBBSAMPLER</code>	100
(2H) Implement <code>DISTANCEBETWEENPATTERNANDSTRINGS</code>	107

Chapter 3	115
(3A) Generate the k -mer Composition of a String	120
(3B) Reconstruct a String from its Genome Path	125
(3C) Construct the Overlap Graph of a Collection of k -mers	128
(3D) Construct the de Bruijn Graph of a String	132
(3E) Construct the de Bruijn Graph of a Collection of k -mers	137
(3F) Find an Eulerian Cycle in a Graph	146
(3G) Find an Eulerian Path in a Graph	147
(3H) Reconstruct a String from its k -mer Composition	147
(3I) Find a k -Universal Circular String	148
(3J) Reconstruct a String from its Paired Composition	157
(3K) Generate the Contigs from a Collection of Reads	160
(3L) Construct a String Spelled by a Gapped Genome Path	169
(3M) Generate All Maximal Non-Branching Paths in a Graph	169
Chapter 4	182
(4A) Translate an RNA String into an Amino Acid String	186
(4B) Find Substrings of a Genome Encoding a Given Amino Acid String	187
(4C) Generate the Theoretical Spectrum of a Cyclic Peptide	191
(4D) Compute the Number of Peptides of Given Total Mass	193
(4E) Find a Cyclic Peptide with Theoretical Spectrum Matching an Ideal Spectrum	196
(4F) Compute the Score of a Cyclic Peptide Against a Spectrum	198
(4G) Implement LEADERBOARDCYCLOPEPTIDSEQUENCING	200
(4H) Generate the Convolution of a Spectrum	203
(4I) Implement CONVOLUTIONCYCLOPEPTIDSEQUENCING	205
(4J) Generate the Theoretical Spectrum of a Linear Peptide	211
(4K) Compute the Score of a Linear Peptide	214
(4L) Implement TRIM to Trim a Peptide Leaderboard	215
(4M) Solve the Turnpike Problem	219
Chapter 5	222
(5A) Find the Minimum Number of Coins Needed to Make Change	240
(5B) Find the Length of a Longest Path in a Manhattan-like Grid	245
(5C) Find a Longest Common Subsequence of Two Strings	252
(5D) Find the Longest Path in a DAG	253
(5E) Find a Highest-Scoring Alignment of Two Strings	255

(5F) Find a Highest-Scoring Local Alignment of Two Strings	260
(5G) Compute the Edit Distance Between Two Strings	262
(5H) Find a Highest-Scoring Fitting Alignment of Two Strings	263
(5I) Find a Highest-Scoring Overlap Alignment of Two Strings	264
(5J) Align Two Strings Using Affine Gap Penalties	268
(5K) Find a Middle Edge in an Alignment Graph in Linear Space	275
(5L) Align Two Strings Using Linear Space	276
(5M) Find a Highest-Scoring Alignment of a Collection of Strings	279
(5N) Find a Topological Ordering of a DAG	285
Chapter 6	292
(6A) Implement GREEDYSORTING to Sort a Permutation by Reversals	305
(6B) Compute the Number of Breakpoints in a Permutation	308
(6C) Compute the 2-Break Distance Between a Pair of Genomes	321
(6D) Find a Shortest Transformation of One Genome into Another via 2-Breaks	322
(6E) Find All Shared k -mers of a Pair of Strings	326
(6F) Implement CHROMOSOMETOCYCLE	336
(6G) Implement CYCLETOTOCYCLE	337
(6H) Implement COLOREDEDGES	337
(6I) Implement GRAPHTOGENOME	338
(6J) Implement 2-BREAKONGENOMEGRAPH	339
(6K) Implement 2-BREAKONGENOME	339