

Contents

Preface

1 Introduction

2 Algorithms and Complexity

- 2.1 What Is an Algorithm?
- 2.2 Biological Algorithms versus Computer Algorithms
- 2.3 The Change Problem
- 2.4 Correct versus Incorrect Algorithms
- 2.5 Recursive Algorithms
- 2.6 Iterative versus Recursive Algorithms
- 2.7 Fast versus Slow Algorithms
- 2.8 Big-O Notation
- 2.9 Algorithm Design Techniques
 - 2.9.1 Exhaustive Search
 - 2.9.2 Branch-and-Bound Algorithms
 - 2.9.3 Greedy Algorithms
 - 2.9.4 Dynamic Programming
 - 2.9.5 Divide-and-Conquer Algorithms
 - 2.9.6 Machine Learning
 - 2.9.7 Randomized Algorithms
- 2.10 Tractable versus Intractable Problems
- 2.11 Notes
 - Biobox: Richard Karp
- 2.12 Problems

3	Molecular Biology Primer	57
3.1	What Is Life Made Of?	57
3.2	What Is the Genetic Material?	59
3.3	What Do Genes Do?	60
3.4	What Molecule Codes for Genes?	61
3.5	What Is the Structure of DNA?	61
3.6	What Carries Information between DNA and Proteins?	63
3.7	How Are Proteins Made?	65
3.8	How Can We Analyze DNA?	67
3.8.1	Copying DNA	67
3.8.2	Cutting and Pasting DNA	71
3.8.3	Measuring DNA Length	72
3.8.4	Probing DNA	72
3.9	How Do Individuals of a Species Differ?	73
3.10	How Do Different Species Differ?	74
3.11	Why Bioinformatics?	75
	Biobox: Russell Doolittle	79
4	Exhaustive Search	83
4.1	Restriction Mapping	83
4.2	Impractical Restriction Mapping Algorithms	87
4.3	A Practical Restriction Mapping Algorithm	89
4.4	Regulatory Motifs in DNA Sequences	91
4.5	Profiles	93
4.6	The Motif Finding Problem	97
4.7	Search Trees	100
4.8	Finding Motifs	108
4.9	Finding a Median String	111
4.10	Notes	114
	Biobox: Gary Stormo	116
4.11	Problems	119
5	Greedy Algorithms	125
5.1	Genome Rearrangements	125
5.2	Sorting by Reversals	127
5.3	Approximation Algorithms	131
5.4	Breakpoints: A Different Face of Greed	132
5.5	A Greedy Approach to Motif Finding	136
5.6	Notes	137

Biobox: David Sankoff

5.7 Problems

6 Dynamic Programming Algorithms

6.1 The Power of DNA Sequence Comparison

6.2 The Change Problem Revisited

6.3 The Manhattan Tourist Problem

6.4 Edit Distance and Alignments

6.5 Longest Common Subsequences

6.6 Global Sequence Alignment

6.7 Scoring Alignments

6.8 Local Sequence Alignment

6.9 Alignment with Gap Penalties

6.10 Multiple Alignment

6.11 Gene Prediction

6.12 Statistical Approaches to Gene Prediction

6.13 Similarity-Based Approaches to Gene Prediction

6.14 Spliced Alignment

6.15 Notes

Biobox: Michael Waterman

6.16 Problems

7 Divide-and-Conquer Algorithms

7.1 Divide-and-Conquer Approach to Sorting

7.2 Space-Efficient Sequence Alignment

7.3 Block Alignment and the Four-Russians Speedup

7.4 Constructing Alignments in Subquadratic Time

7.5 Notes

Biobox: Webb Miller

7.6 Problems

8 Graph Algorithms

8.1 Graphs

8.2 Graphs and Genetics

8.3 DNA Sequencing

8.4 Shortest Superstring Problem

8.5 DNA Arrays as an Alternative Sequencing Technique

8.6 Sequencing by Hybridization

8.7 SBH as a Hamiltonian Path Problem

8.8	SBH as an Eulerian Path Problem	272
8.9	Fragment Assembly in DNA Sequencing	275
8.10	Protein Sequencing and Identification	280
8.11	The Peptide Sequencing Problem	284
8.12	Spectrum Graphs	287
8.13	Protein Identification via Database Search	290
8.14	Spectral Convolution	292
8.15	Spectral Alignment	293
8.16	Notes	299
8.17	Problems	302
9	Combinatorial Pattern Matching	311
9.1	Repeat Finding	311
9.2	Hash Tables	313
9.3	Exact Pattern Matching	316
9.4	Keyword Trees	318
9.5	Suffix Trees	320
9.6	Heuristic Similarity Search Algorithms	324
9.7	Approximate Pattern Matching	326
9.8	BLAST: Comparing a Sequence against a Database	330
9.9	Notes	331
	Biobox: Gene Myers	333
9.10	Problems	337
10	Clustering and Trees	339
10.1	Gene Expression Analysis	339
10.2	Hierarchical Clustering	343
10.3	k -Means Clustering	346
10.4	Clustering and Corrupted Cliques	348
10.5	Evolutionary Trees	354
10.6	Distance-Based Tree Reconstruction	358
10.7	Reconstructing Trees from Additive Matrices	361
10.8	Evolutionary Trees and Hierarchical Clustering	366
10.9	Character-Based Tree Reconstruction	368
10.10	Small Parsimony Problem	370
10.11	Large Parsimony Problem	374
10.12	Notes	379
	Biobox: Ron Shamir	380
10.13	Problems	384

11 Hidden Markov Models	387
11.1 <i>CG</i> -Islands and the “Fair Bet Casino”	387
11.2 The Fair Bet Casino and Hidden Markov Models	390
11.3 Decoding Algorithm	393
11.4 HMM Parameter Estimation	397
11.5 Profile HMM Alignment	398
11.6 Notes	400
Biobox: David Haussler	403
11.7 Problems	407
12 Randomized Algorithms	409
12.1 The Sorting Problem Revisited	409
12.2 Gibbs Sampling	412
12.3 Random Projections	414
12.4 Notes	416
12.5 Problems	417
Using Bioinformatics Tools	419
Bibliography	421
Index	428