
Contents

1	Core Bioinformatics Problems	1
1.1	DNA Mapping and Sequencing	1
1.1.1	Size of Human DNA	1
1.1.2	Copying DNA: Polymerase Chain Reaction (PCR)	2
1.1.3	Hybridization and Microarrays	2
1.1.4	Cutting DNA into Fragments	2
1.1.5	Prefix Cutting	3
1.1.6	Unique Markers	3
1.1.7	Measuring the Length of DNA Molecules	4
1.1.8	Sequencing Short DNA Molecules	4
1.1.9	Mapping Long DNA Molecules	4
1.1.10	Mapping by Single Digestion	5
1.1.11	Mapping by Double Digestion	6
1.1.12	Mid-sized DNA Molecules: Shotgun Sequencing	7
1.2	String Storage and Pattern Matching	8
1.2.1	Complexity of Pattern Matching in Bioinformatics	8
1.2.2	First Look at Suffix Trees	9
1.2.3	A Couple of Suffix Tree Applications	10
1.3	String Alignment	11
1.4	Multiple Alignment	13
1.5	Gene Detection	15
1.6	Gene Comparison	16
1.7	RNA Structure Prediction	18
1.8	Protein Structure Prediction	19
1.8.1	Holy Grail Protein Structure Prediction	19
1.8.2	Secondary Structure	20
1.8.3	Tertiary Structure Prediction with Contact Maps	20
1.8.4	HP-Model Approach	21
1.8.5	Protein Threading	22
1.9	Molecular Networks	23
1.10	Bibliographic Remarks	24

2	Turning to Algorithmic Problems	25
2.1	Presentation of Algorithmic Problems	25
2.2	DNA Mapping	26
2.2.1	Mapping by Hybridization	26
2.2.2	Mapping by Single Digestion	27
2.2.3	Mapping by Double Digestion	30
2.3	Shotgun Sequencing and Shortest Common Superstrings	31
2.4	Exact Pattern Matching	35
2.4.1	Naive Pattern Matching	35
2.4.2	Knuth-Morris-Pratt Algorithm	35
2.4.3	Multi-Text Search	37
2.4.4	Multi-Pattern Search	37
2.4.5	Formal Definition of Suffix Trees	38
2.5	Soft Pattern Matching = Alignment	41
2.5.1	Alignments Restated	41
2.5.2	Evaluating Alignments by Scoring Functions	41
2.6	Multiple Alignment	43
2.6.1	Sum-of-Pairs Maximization	43
2.6.2	Multiple Alignment Along a Guide Tree	43
2.6.3	Consensus Line Optimization	45
2.6.4	Steiner String	45
2.6.5	Equivalence of Consensus Lines and Steiner Strings	45
2.6.6	Generalization to Steiner Trees/Phylogenetic Alignment	47
2.6.7	Profile Alignment	47
2.6.8	Hidden Markov Multiple Alignment	48
2.6.9	Example Heuristic for Obtaining Multiple Alignments	51
2.7	Gene Detection	52
2.8	Genome Rearrangement	53
2.9	RNA Structure Prediction	55
2.10	Protein Structure Prediction	56
2.10.1	Comparative Approaches Using Neural Networks	56
2.10.2	Protein Threading	57
2.11	Bi-Clustering	59
2.12	Bibliographic Remarks	60
3	Dynamic Programming	61
3.1	General Principle of Dynamic Programming	61
3.2	Alignment	63
3.2.1	Problem Restated	63
3.2.2	Parameterization	64
3.2.3	Bellman Principle	64
3.2.4	Recursive Solution	64
3.2.5	Number of Different Subcalls and Overall Complexity	65
3.2.6	Tabular Organization of the Bottom-Up Computation	65
3.3	Multiple Alignment	67

3.4	Affine Gap Alignment	67
3.4.1	Problem Restated	67
3.4.2	Parameterization and Conditioning	68
3.4.3	Bellman Principle	68
3.4.4	Recursive Solution	68
3.4.5	Number of Different Subcalls and Overall Complexity . .	69
3.5	Exon Assembly	69
3.5.1	Problem Restated	69
3.5.2	Parameterization and Conditioning	70
3.5.3	Bellman Principle	70
3.5.4	Recursive Solution	71
3.5.5	Number of Different Subcalls and Overall Complexity . .	72
3.5.6	Preprocessing of Simple Terms	72
3.5.7	Parallel Computation of Recursive Terms	72
3.6	RNA Structure Prediction	73
3.6.1	Problem Restated	73
3.6.2	Parameterization and Conditioning	74
3.6.3	Recursive Solution and Bellman Principle	74
3.6.4	Number of Different Subcalls and Overall Complexity . .	75
3.6.5	Variation: Free Energy Minimization	76
3.7	Viterbi Algorithm	83
3.7.1	Problem Restated	83
3.7.2	Parameterization and Conditioning	83
3.7.3	Bellman Principle	83
3.7.4	Recursive Computation	83
3.7.5	Counting and Overall Complexity	84
3.8	Baum-Welch Algorithm	84
3.8.1	Problem Motivation	84
3.8.2	A Couple of Important ‘Variables’	84
3.8.3	Computing Forward Variables	85
3.8.4	Computing Backward Variables	85
3.8.5	Computing Transition Variables	85
3.8.6	Computing State Variables	85
3.8.7	Model Improvement	86
3.9	Expressiveness of Viterbi Equations	86
3.10	Lifted Phylogenetic Alignment	91
3.10.1	Problem Restated	91
3.10.2	Parameterization and Conditioning	92
3.10.3	Bellman Principle	92
3.10.4	Recursive Solution	92
3.10.5	Counting and Overall Complexity	93
3.10.6	Preprocessing of Simple Terms	93
3.11	Protein Threading	93
3.11.1	Problem Restated	93
3.11.2	Parameterization and Conditioning	94

3.11.3	Bellman Principle	94
3.11.4	Recursive Solution	94
3.11.5	Counting and Overall Complexity	95
3.12	Bibliographic Remarks	95
4	Intelligent Data Structures	97
4.1	Representation Matters	97
4.1.1	Intelligent Data Structures Used in Computer Science..	97
4.1.2	Covering a Truncated Board	97
4.1.3	Choose Three Bits and Win	98
4.1.4	Check a Proof	99
4.2	PQ-Trees	101
4.2.1	Basic Notions	101
4.2.2	Transformation rules for marked PQ-Trees	103
4.3	Suffix Trees	111
4.3.1	Outline of Ukkonen's Algorithm and Basic Procedures ..	111
4.3.2	Insertion Procedure	112
4.3.3	Saving Single Character Insertions	115
4.3.4	Saving Navigation Steps	116
4.3.5	Estimation of the Overall Number of Visited Nodes	117
4.3.6	Ukkonen's Theorem	120
4.3.7	Common Suffix Tree for Several Strings	123
4.3.8	Applications of Suffix Trees	124
4.4	Least Common Ancestor	131
4.4.1	Motivation	131
4.4.2	Least Common Ancestor in Full Binary Trees	131
4.4.3	Least Common Ancestor in Arbitrary Trees	135
4.5	Signed Genome Rearrangement	144
4.5.1	Reality-Desire Diagrams of Signed Permutations	144
4.5.2	"Ping-Pong" of Lower and Upper Bounds	148
4.5.3	Padding RD-Diagrams	159
4.5.4	Sorting Padded Good Components	161
4.5.5	Summary of the Sorting Procedure	164
4.6	Bibliographic Remarks	166
5	NP-Hardness of Core Bioinformatics Problems	167
5.1	Getting Familiar with Basic Notions of Complexity Theory ...	167
5.2	Cook's Theorem: Primer NP-Complete Problem	171
5.3	"Zoo" of NP-Complete Problems	175
5.4	Bridge to Bioinformatics: Shortest Common Super-Sequence ..	188
5.5	NP-Completeness of Core Bioinformatics Problems	198
5.5.1	Multiple Alignment	198
5.5.2	Shortest Common Superstring	201
5.5.3	Double Digest	204
5.5.4	Protein Threading	204

5.5.5	Bi-Clustering	206
5.5.6	Pseudoknot Prediction	209
5.5.7	The Picture of Presented Reductions	213
5.5.8	Further NP-Complete Bioinformatics Problems	213
5.6	Bibliographic Remarks	215
6	Approximation Algorithms	217
6.1	Basics of Approximation Algorithms	217
6.1.1	What is an Approximation Algorithm?	217
6.1.2	First Example: Metric Travelling Salesman	217
6.1.3	Lower and Upper Bounding	218
6.1.4	What are Approximation Algorithms good for?	219
6.2	Shortest Common Superstring Problem	221
6.2.1	Feasible Lower Bound	221
6.2.2	Cycle Covers Related to Perfect Matchings Problem ...	222
6.2.3	Computing Maximum Cost Perfect Matchings	223
6.2.4	Cycle Cover Versus Cyclic Covering by Strings	225
6.2.5	Some Combinatorial Statements on Cyclic Covering ...	226
6.2.6	4-Approximation Algorithm	228
6.2.7	Putting Things Together	229
6.3	Steiner String	230
6.3.1	Problem Restated	230
6.3.2	Feasible Lower Bound	230
6.3.3	2-Approximation Algorithm	231
6.4	Phylogenetic Alignment	232
6.5	Multiple Alignment	234
6.5.1	Feasible Lower Bound	234
6.5.2	2-Approximation Algorithm	235
6.6	Sorting Unsigned Genomes	236
6.6.1	Feasible Lower Bound	236
6.6.2	4-Approximation Algorithm	237
6.6.3	2-Approximation Algorithm	239
6.7	HP-Model Protein Structure Prediction	243
6.7.1	Feasible Upper Bound	243
6.7.2	Asymptotic 4-Approximation Algorithm	245
6.8	Bibliographic Remarks	248
7	A Selection of Metaheuristics and Various Projects	249
7.1	Multi-Layer Perceptron	249
7.1.1	Architecture of Multi-Layer Perceptrons	250
7.1.2	Forward Propagation of Input Signals	251
7.1.3	Input and Output Encoding	251
7.1.4	Backpropagation of Error Signals	252
7.1.5	Tuning Backpropagation	253
7.1.6	Generalization and Overfitting	253

7.1.7	Application to Secondary Structure Prediction	254
7.1.8	Further Applications	254
7.2	Support Vector Machines	255
7.2.1	Architecture of Support Vector Machines	255
7.2.2	Margin Maximization	256
7.2.3	Primal Problem	257
7.2.4	Dual Problem	258
7.2.5	Kernel Function Trick	259
7.2.6	Generalization and Over-Fitting	260
7.2.7	Application to Contact Map Prediction	261
7.2.8	Soft Margin Maximization	263
7.2.9	Semi-Supervised Support Vector Machines	264
7.3	Hidden Markov Models	265
7.3.1	Architecture of Hidden Markov Models	266
7.3.2	Causes and Effects	266
7.3.3	Bioinformatics Application: CG Islands	268
7.3.4	Further Applications	269
7.4	Genetic Algorithms	269
7.4.1	Basic Characteristics of Genetic Algorithms	269
7.4.2	Application to HP-Fold Optimization	271
7.4.3	Further Applications	272
7.5	Ant Colony Optimization	272
7.5.1	Basic Characteristics of Ant Colony Algorithms	273
7.5.2	Application to HP-Fold Optimization	275
7.5.3	Further Applications	276
7.6	Bibliographic Remarks	276
References		279
Index		285

Bioinformatics

Problem Solving Paradigms

Sperschneider, V.

2008, XVIII, 290 p. 208 illus., Hardcover

ISBN: 978-3-540-78505-7