

Contents

Preface	xiii
0 Introduction	1
0.1 Molecular Biology	3
0.2 Mathematics, Statistics, and Computer Science	3
1 Some Molecular Biology	5
1.1 DNA and Proteins	6
1.1.1 The Double Helix	6
1.2 The Central Dogma	7
1.3 The Genetic Code	8
1.4 Transfer RNA and Protein Sequences	12
1.5 Genes Are Not Simple	16
1.5.1 Starting and Stopping	16
1.5.2 Control of Gene Expression	16
1.5.3 Split Genes	17
1.5.4 Jumping Genes	18
1.6 Biological Chemistry	18
2 Restriction Maps	29
2.1 Introduction	29
2.2 Graphs	31
2.3 Interval Graphs	33
2.4 Measuring Fragment Sizes	38
3 Multiple Maps	41
3.1 Double Digest Problem	42
3.1.1 Multiple Solutions in the Double Digest Problem	43
3.2 Classifying Multiple Solutions	48
3.2.1 Reflections	48
3.2.2 Overlap Equivalence	48
3.2.3 Overlap Size Equivalence	51
3.2.4 More Graph Theory	52

3.2.5	From One Path to Another	53
3.2.6	Restriction Maps and the Border Block Graph	56
3.2.7	Cassette Transformations of Restriction Maps	58
3.2.8	An Example	61
4	Algorithms for DDP	65
4.1	Algorithms and Complexity	65
4.2	DDP is <i>NP</i> -Complete	67
4.3	Approaches to DDP	68
4.3.1	Integer Programming	68
4.3.2	Partition Problems	69
4.3.3	TSP	70
4.4	Simulated Annealing: TSP and DDP	70
4.4.1	Simulated Annealing	70
4.4.2	Traveling Salesman Problem	75
4.4.3	DDP	76
4.4.4	Circular Maps	78
4.5	Mapping with Real Data	79
4.5.1	Fitting Data to a Map	80
4.5.2	Map Algorithms	81
5	Cloning and Clone Libraries	83
5.1	A Finite Number of Random Clones	85
5.2	Libraries by Complete Digestion	85
5.3	Libraries by Partial Digestion	87
5.3.1	The Fraction of Clonable Bases	88
5.3.2	Sampling, Approach 1	91
5.3.3	Designing Partial Digest Libraries	92
5.4	Genomes per Microgram	98
6	Physical Genome Maps: Oceans, Islands and Anchors	101
6.1	Mapping by Fingerprinting	102
6.1.1	Oceans and Islands	102
6.1.2	Divide and Conquer	110
6.1.3	Two Pioneering Experiments	111
6.1.4	Evaluating a Fingerprinting Scheme	114
6.2	Mapping by Anchoring	119
6.2.1	Oceans, Islands and Anchors	119
6.2.2	Duality Between Clones and Anchors	126
6.3	An Overview of Clone Overlap	127
6.4	Putting It Together	129

7	Sequence Assembly	135
7.1	Shotgun Sequencing	135
7.1.1	SSP is <i>NP</i> -complete	137
7.1.2	Greedy is at most Four Times Optimal	138
7.1.3	Assembly in Practice	143
7.1.4	Sequence Accuracy	145
7.1.5	Expected Progress	147
7.2	Sequencing by Hybridization	148
7.2.1	Other SBH Designs	154
7.3	Shotgun Sequencing Revisited	156
8	Databases and Rapid Sequence Analysis	161
8.1	DNA and Protein Sequence Databases	162
8.1.1	Description of the Entries in a Sequence Data File	163
8.1.2	Sample Sequence Data File	164
8.1.3	Statistical Summary	166
8.2	A Tree Representation of a Sequence	167
8.3	Hashing a Sequence	168
8.3.1	A Hash Table	169
8.3.2	Hashing in Linear Time	170
8.3.3	Hashing and Chaining	170
8.4	Repeats in a Sequence	171
8.5	Sequence Comparison by Hashing	172
8.6	Sequence Comparison with at most <i>l</i> Mismatches	176
8.7	Sequence Comparison by Statistical Content	180
9	Dynamic Programming Alignment of Two Sequences	183
9.1	The Number of Alignments	186
9.2	Shortest and Longest Paths in a Network	190
9.3	Global Distance Alignment	192
9.3.1	Indel Functions	194
9.3.2	Position-Dependent Weights	197
9.4	Global Similarity Alignment	198
9.5	Fitting One Sequence into Another	201
9.6	Local Alignment and Clumps	202
9.6.1	Self-Comparison	206
9.6.2	Tandem Repeats	207
9.7	Linear Space Algorithms	209
9.8	Tracebacks	212
9.9	Inversions	215
9.10	Map Alignment	219
9.11	Parametric Sequence Comparisons	223
9.11.1	One-Dimension Parameter Sets	225
9.11.2	Into Two-Dimensions	228

10 Multiple Sequence Alignment	233
10.1 The Cystic Fibrosis Gene	233
10.2 Dynamic Programming in r -Dimensions	236
10.2.1 Reducing the Volume	237
10.3 Weighted-Average Sequences	238
10.3.1 Aligning Alignments	242
10.3.2 Center of Gravity Sequences	242
10.4 Profile Analysis	242
10.4.1 Statistical Significance	244
10.5 Alignment by Hidden Markov Models	245
10.6 Consensus Word Analysis	248
10.6.1 Analysis by Words	249
10.6.2 Consensus Alignment	250
10.6.3 More Complex Scoring	251
11 Probability and Statistics for Sequence Alignment	253
11.1 Global Alignment	254
11.1.1 Alignment Given	254
11.1.2 Alignment Unknown	255
11.1.3 Linear Growth of Alignment Score	256
11.1.4 The Azuma-Hoeffding Lemma	257
11.1.5 Large Deviations from the Mean	259
11.1.6 Large Deviations for Binomials	261
11.2 Local Alignment	263
11.2.1 Laws of Large Numbers	263
11.3 Extreme Value Distributions	275
11.4 The Chein-Stein Method	278
11.5 Poisson Approximation and Long Matches	280
11.5.1 Headruns	280
11.5.2 Exact Matching Between Sequences	282
11.5.3 Approximate Matching	288
11.6 Sequence Alignment with Scores	294
11.6.1 A Phase Transition	294
11.6.2 Practical p -Values	299
12 Probability and Statistics for Sequence Patterns	305
12.1 A Central Limit Theorem	307
12.1.1 Generalized Words	313
12.1.2 Estimating Probabilities	313
12.2 Nonoverlapping Pattern Counts	314
12.2.1 Renewal Theory for One Pattern	314
12.2.2 Li's Method and Multiple Patterns	318
12.3 Poisson Approximation	321
12.4 Site Distributions	323

12.4.1 Intersite Distances	324
13 RNA Secondary Structure	327
13.1 Combinatorics	327
13.1.1 Counting More Shapes	332
13.2 Minimum Free-energy Structures	334
13.2.1 Reduction of Computation Time for Hairpins	336
13.2.2 Linear Destabilization Functions	338
13.2.3 Multibranch Loops	339
13.3 Consensus folding	340
14 Trees and Sequences	345
14.1 Trees	345
14.1.1 Splits	347
14.1.2 Metrics on Trees	351
14.2 Distance	353
14.2.1 Additive Trees	353
14.2.2 Ultrametric Trees	357
14.2.3 Nonadditive Distances	359
14.3 Parsimony	361
14.4 Maximum Likelihood Trees	367
14.4.1 Continuous Time Markov Chains	367
14.4.2 Estimating the Rate of Change	369
14.4.3 Likelihood and Trees	372
15 Sources and Perspectives	377
15.1 Molecular Biology	377
15.2 Physical Maps and Clone Libraries	377
15.3 Sequence Assembly	379
15.4 Sequence Comparisons	379
15.4.1 Databases and Rapid Sequence Analysis	379
15.4.2 Dynamic Programming for Two Sequences	380
15.4.3 Multiple Sequence Alignment	382
15.5 Probability and Statistics	382
15.5.1 Sequence Alignment	382
15.5.2 Sequence Patterns	383
15.6 RNA Secondary Structure	384
15.7 Trees and Sequences	385
References	387
I Problem Solutions and Hints	401
II Mathematical Notation	421

Algorithm Index 423

Author Index 425

Subject Index 428

13.1.1 Counting Pair Spaces 338

13.1.2 Minimum Free-energy RNA Secondary Structure 339

13.2.1 Reduction of Computation Time for 340

13.2.2 Linear Decomposition 341

13.2.3 Multichain 342

13.3 Consensus 343

14.1 Trees and Sequences 344

14.1.1 Trees 345

14.1.2 346

14.1.3 347

14.2 348

14.3 349

14.3.1 350

14.3.2 351

14.3.3 352

14.3.4 353

14.4 354

14.4.1 355

14.4.2 356

14.4.3 357

15.1 Molecular Biology 358

15.2 359

15.3 360

15.4 361

15.4.1 362

15.4.2 363

15.4.3 364

15.5 365

15.5.1 366

15.5.2 367

15.6 368

15.7 369

16.1 370

16.2 371

16.3 372

16.4 373

16.5 374

16.6 375

16.7 376

16.8 377

16.9 378

16.10 379

16.11 380

16.12 381

16.13 382

16.14 383

16.15 384

16.16 385

16.17 386

16.18 387

16.19 388

16.20 389

16.21 390

16.22 391

16.23 392

16.24 393

16.25 394

16.26 395

16.27 396

16.28 397

16.29 398

16.30 399

16.31 400

16.32 401

16.33 402

16.34 403

16.35 404

16.36 405

16.37 406

16.38 407

16.39 408

16.40 409

16.41 410

16.42 411

16.43 412

16.44 413

16.45 414

16.46 415

16.47 416

16.48 417

16.49 418

16.50 419

16.51 420

16.52 421

16.53 422