



Pavel A. Pevzner

Computational Molecular Biology

An Algorithmic Approach

Contents

Preface

xiii

1	Computational Gene Hunting	1
1.1	Introduction	1
1.2	Genetic Mapping	1
1.3	Physical Mapping	5
1.4	Sequencing	8
1.5	Similarity Search	10
1.6	Gene Prediction	12
1.7	Mutation Analysis	14
1.8	Comparative Genomics	14
1.9	Proteomics	17
2	Restriction Mapping	19
2.1	Introduction	19
2.2	Double Digest Problem	21
2.3	Multiple Solutions of the Double Digest Problem	23
2.4	Alternating Cycles in Colored Graphs	26
2.5	Transformations of Alternating Eulerian Cycles	27
2.6	Physical Maps and Alternating Eulerian Cycles	32
2.7	Partial Digest Problem	34
2.8	Homometric Sets	35
2.9	Some Other Problems and Approaches	38
2.9.1	Optical mapping	38
2.9.2	Probed Partial Digest mapping	38

3	Map Assembly	41
3.1	Introduction	41
3.2	Mapping with Non-Unique Probes	44
3.3	Mapping with Unique Probes	48
3.4	Interval Graphs	50
3.5	Mapping with Restriction Fragment Fingerprints	53
3.6	Some Other Problems and Approaches	54
3.6.1	Lander-Waterman statistics	54
3.6.2	Screening clone libraries	55
3.6.3	Radiation hybrid mapping	55
4	Sequencing	59
4.1	Introduction	59
4.2	Overlap, Layout, and Consensus	61
4.3	Double-Barreled Shotgun Sequencing	62
4.4	Some Other Problems and Approaches	63
4.4.1	Shortest Superstring Problem	63
4.4.2	Finishing phase of DNA sequencing	63
5	DNA Arrays	65
5.1	Introduction	65
5.2	Sequencing by Hybridization	67
5.3	SBH and the Shortest Superstring Problem	68
5.4	SBH and the Eulerian Path Problem	70
5.5	Probability of Unique Sequence Reconstruction	74
5.6	String Rearrangements	75
5.7	2-optimal Eulerian Cycles	78
5.8	Positional Sequencing by Hybridization	81
5.9	Design of DNA Arrays	82
5.10	Resolving Power of DNA Arrays	84
5.11	Multiprobe Arrays versus Uniform Arrays	85
5.12	Manufacture of DNA Arrays	87
5.13	Some Other Problems and Approaches	91
5.13.1	SBH with universal bases	91
5.13.2	Adaptive SBH	91
5.13.3	SBH-style shotgun sequencing	92
5.13.4	Fidelity probes for DNA arrays	92

6	Sequence Comparison	93
6.1	Introduction	93
6.2	Longest Common Subsequence Problem	96
6.3	Sequence Alignment	98
6.4	Local Sequence Alignment	98
6.5	Alignment with Gap Penalties	100
6.6	Space-Efficient Sequence Alignment	101
6.7	Young Tableaux	102
6.8	Average Length of Longest Common Subsequences	106
6.9	Generalized Sequence Alignment and Duality	109
6.10	Primal-Dual Approach to Sequence Comparison	111
6.11	Sequence Alignment and Integer Programming	113
6.12	Approximate String Matching	114
6.13	Comparing a Sequence Against a Database	115
6.14	Multiple Filtration	116
6.15	Some Other Problems and Approaches	118
6.15.1	Parametric sequence alignment	118
6.15.2	Alignment statistics and phase transition	119
6.15.3	Suboptimal sequence alignment	119
6.15.4	Alignment with tandem duplications	120
6.15.5	Winnowing database search results	120
6.15.6	Statistical distance between texts	120
6.15.7	RNA folding	121
7	Multiple Alignment	123
7.1	Introduction	123
7.2	Scoring a Multiple Alignment	125
7.3	Assembling Pairwise Alignments	126
7.4	Approximation Algorithm for Multiple Alignments	127
7.5	Assembling 1-way Alignments	128
7.6	Dot-Matrices and Image Reconstruction	130
7.7	Multiple Alignment via Dot-Matrix Multiplication	131
7.8	Some Other Problems and Approaches	132
7.8.1	Multiple alignment via evolutionary trees	132
7.8.2	Cutting corners in edit graphs	132

8	Finding Signals in DNA	133
8.1	Introduction	133
8.2	Edgar Allan Poe and DNA Linguistics	134
8.3	The Best Bet for Simpletons	136
8.4	The Conway Equation	137
8.5	Frequent Words in DNA	140
8.6	Consensus Word Analysis	143
8.7	CG-islands and the "Fair Bet Casino"	144
8.8	Hidden Markov Models	145
8.9	The Elkhorn Casino and HMM Parameter Estimation	147
8.10	Profile HMM Alignment	148
8.11	Gibbs Sampling	149
8.12	Some Other Problems and Approaches	150
8.12.1	Finding gapped signals	150
8.12.2	Finding signals in samples with biased frequencies	150
8.12.3	Choice of alphabet in signal finding	151
9	Gene Prediction	153
9.1	Introduction	153
9.2	Statistical Approach to Gene Prediction	155
9.3	Similarity-Based Approach to Gene Prediction	156
9.4	Spliced Alignment	157
9.5	Reverse Gene Finding and Locating Exons in cDNA	167
9.6	The Twenty Questions Game with Genes	169
9.7	Alternative Splicing and Cancer	169
9.8	Some Other Problems and Approaches	171
9.8.1	Hidden Markov Models for gene prediction	171
9.8.2	Bacterial gene prediction	173
10	Genome Rearrangements	175
10.1	Introduction	175
10.2	The Breakpoint Graph	187
10.3	"Hard-to-Sort" Permutations	188
10.4	Expected Reversal Distance	189
10.5	Signed Permutations	192
10.6	Interleaving Graphs and Hurdles	193
10.7	Equivalent Transformations of Permutations	196

10.8	Searching for Safe Reversals	200
10.9	Clearing the Hurdles	204
10.10	Duality Theorem for Reversal Distance	209
10.11	Algorithm for Sorting by Reversals	213
10.12	Transforming Men into Mice	214
10.13	Capping Chromosomes	219
10.14	Caps and Tails	221
10.15	Duality Theorem for Genomic Distance	223
10.16	Genome Duplications	226
10.17	Some Other Problems and Approaches	227
10.17.1	Genome rearrangements and phylogenetic studies	227
10.17.2	Fast algorithm for sorting by reversals	228
11	Computational Proteomics	229
11.1	Introduction	229
11.2	The Peptide Sequencing Problem	231
11.3	Spectrum Graphs	232
11.4	Learning Ion-Types	236
11.5	Scoring Paths in Spectrum Graphs	237
11.6	Peptide Sequencing and Anti-Symmetric Paths	239
11.7	The Peptide Identification Problem	240
11.8	Spectral Convolution	241
11.9	Spectral Alignment	243
11.10	Aligning Peptides Against Spectra	245
11.11	Some Other Problems and Approaches	248
11.11.1	From proteomics to genomics	248
11.11.2	Large-scale protein analysis	249
12	Problems	251
12.1	Introduction	251
12.2	Restriction Mapping	251
12.3	Map Assembly	254
12.4	Sequencing	256
12.5	DNA Arrays	257
12.6	Sequence Comparison	259
12.7	Multiple Alignment	264
12.8	Finding Signals in DNA	264

12.9	Gene Prediction	265
12.10	Genome Rearrangements	266
12.11	Computational Proteomics	269
13	All You Need to Know about Molecular Biology	271
	Bibliography	275
	Index	309