

Algorithms in Molecular Biology

Lecture Notes from CSE 590BI

Richard M. Karp

Walter L. Ruzzo

Martin Tompa

Department of Computer Science and Engineering

University of Washington

Box 352350

Seattle, Washington, U.S.A. 98195

Winter 1996

Contents

Course Information	vi
Schedule of Lectures	viii
1 Introduction	1
1.1 Introduction	1
1.2 The Genome	2
1.3 Genetics	2
1.4 Cells	3
1.5 Inheritance of Simple Traits	4
1.6 Proteins	4
1.7 DNA	4
1.8 Transcription	5
1.9 Translation	7
2 Introduction to Sequence Similarity	8
2.1 Sequence Similarity	8
2.2 Biological Motivation for Studying Sequence Similarity	9
2.2.1 Hypothesizing the function of a new sequence	9
2.2.2 Researching the effects of multiple sclerosis	9
2.3 The String Alignment Problem	10

3	Algorithms for Optimal Alignment	12
3.1	An Obvious Algorithm for Optimal Alignment	12
3.2	Asymptotic Analysis of Algorithms	13
3.3	Computing an Optimal Alignment by Dynamic Programming	14
3.3.1	Example	15
3.3.2	Recovering the Alignments	15
3.3.3	Time Analysis	16
3.4	Searching for Local Similarity	16
3.4.1	An Obvious Local Alignment Algorithm	17
4	Optimal Local Alignment, and Gaps	18
4.1	Computing an Optimal Local Alignment by Dynamic Programming	18
4.1.1	Example	20
4.1.2	Time Analysis	21
4.1.3	Space Analysis	21
4.2	Optimal Alignment with Gaps	21
4.2.1	Motivations	21
4.2.2	Affine Gap Model	22
4.2.3	Dynamic Programming Algorithm	23
4.2.4	Time Analysis	23
4.3	Bibliographic Notes on Alignments	24
5	Biotechnology and Computation	25
5.1	Information Content in Biological Systems	25
5.2	Advances in Biotechnology	26
5.2.1	DNA Sequencing	26
5.2.2	Genetic Mapping	27
5.2.3	DNA Arrays	28
5.3	Analysis of the Human T-cell Receptor	29

6	Multiple Sequence Similarity	31
6.1	Biological Motivation for Multiple String Alignment	31
6.2	Multiple String Alignment	32
6.3	Computing an Optimal Multiple Alignment by Dynamic Programming	33
6.4	<i>NP</i> -completeness	33
6.5	An Approximation Algorithm for Multiple String Alignment	35
6.5.1	Algorithm	35
6.5.2	Time Analysis	36
6.5.3	Error Analysis	36
6.6	The Consensus String	38
7	Statistical Inference	39
7.1	Linkage analysis	39
7.2	Maximum Likelihood Estimation	42
7.3	Radiation Hybrid Mapping	44
8	Statistical Inference II; Sequence Analysis I	47
8.1	Radiation Hybrid Mapping (continued)	47
8.1.1	The Traveling Salesman Problem	48
8.1.2	The Mapping to Symmetric TSP	48
8.2	Sequence Analysis	51
8.2.1	Finding Genes	51
9	Sequence Analysis II	55
9.1	Review	55
9.2	Identifying Promoters via Weight Matrices	56
9.3	Identifying Other Sequence Signals	58
9.4	Identifying Genes by Codon Frequencies	59
10	Linkage Analysis	62
10.1	Conditional Independence in Genetics	63

10.2	Calculating the <i>a posteriori</i> probability of the data	63
10.3	Complex Pedigrees	64
10.4	Monte-Carlo Methods	65
11	Sequence Analysis III; Linkage Analysis II	68
11.1	Gene Prediction (continued)	68
11.1.1	Splicing and Dynamic Programming	68
11.1.2	Combining Multiple Indicators	69
11.2	Multilocus Genetic Mapping	69
12	Experimental Methods in DNA Analysis	73
12.1	Structure of DNA (the digital view)	73
12.2	Structure of DNA (the chemical view)	73
12.3	Physical Characteristics	74
12.4	Melting the Double Helix	74
12.5	Reannealing	75
12.6	DNA Chemical Synthesis	76
12.7	Detection	76
12.8	Separation	77
13	Physical Mapping Cont.	79
13.1	Overview	79
13.2	Hidden Markov Models	79
13.2.1	General Case for Hidden Markov Models	80
13.2.2	E-M Algorithm (A High Level Description)	81
13.3	Physical Mapping	81
13.3.1	(Shallow) Overview of Technology	82
14	Physical Mapping	84
14.1	Maps	84
14.2	The Whitehead-Gènèthon map	85

CONTENTS

v

15 TITLE OF LECTURE	91
16 TITLE OF LECTURE	92
17 Determining Phylogenies by Parsimony	93
17.1 Phylogeny Trees	93
17.2 Parsimony	95
17.2.1 The Small Parsimony Problem	95
17.2.2 The Big Parsimony Problem	97