Biological Sequence Analysis

Lecture Notes from CSE 527

Martin Tompa
Department of Computer Science and Engineering
University of Washington
Box 352350
Seattle, Washington, U.S.A. 98195-2350

Winter 2000

Contents

1 Basics of Molecular Biology 1

1.1 Proteins 1

1.1.1 Classification of the Amino Acids 2

1.2 DNA 3

1.2.1 Structure of a Nucleotide 3

1.2.2 Base Pair Complementarity 4

1.2.3 Size of DNA molecules 4

1.3 RNA 4

1.4 Residues 5

1.5 DNA Replication 5

1.6 Synthesis of RNA and Proteins 5

1.6.1 Transcription in Prokaryotes 6

1.6.2 Translation 6

2 Basics of Molecular Biology (continued) 8

2.1 Course Projects 8

2.2 Translation (continued) 8

2.3 Prokaryotic Gene Structure 9

2.4 Prokaryotic Genome Organization 9

2.5 Eukaryotic Gene Structure 9

2.6 Eukaryotic Genome Organization 10

2.7 Goals and Status of Genome Projects 10

2.8 Sequence Analysis 11

3 Introduction to Sequence Similarity 12

3.1 Sequence Similarity 12

3.2 Biological Motivation for Studying Sequence Similarity 12

3.2.1 Hypothesizing the Function of a New Sequence 12
3.2.2 Researching the Effects of Multiple Sclerosis 13
3.3 The String Alignment Problem 13
3.4 An Obvious Algorithm for Optimal Alignment 14
3.5 Asymptotic Analysis of Algorithms 15

4 Alignment by Dynamic Programming 17
4.1 Computing an Optimal Alignment by Dynamic Programming 17
4.1.1 Example 18
4.1.2 Recovering the Alignments 18
4.1.3 Time Analysis 19
4.2 Searching for Local Similarity 19
4.2.1 An Obvious Local Alignment Algorithm 19
4.2.2 Set-Up for Local Alignment by Dynamic Programming 20

5 Local Alignment, and Gap Penalties 21
5.1 Computing an Optimal Local Alignment by Dynamic Programming 21
5.1.1 Example 22
5.1.2 Time Analysis 22
5.2 Space Analysis 23
5.3 Optimal Alignment with Gaps 23
5.3.1 Motivations 23
5.3.2 Affine Gap Model 24
5.3.3 Dynamic Programming Algorithm 24
5.3.4 Time Analysis 25
5.4 Bibliographic Notes on Alignments 25

6 Multiple Sequence Alignment 27
6.1 Biological Motivation for Multiple Sequence Alignment 27
6.1.1 Representing Protein Families 27
6.1.2 Repetitive Sequences in DNA 27
6.2 Formulation of the Multiple String Alignment Problem 28
6.3 Computing an Optimal Multiple Alignment by Dynamic Programming 29
6.4 *NP*-completeness 29
6.5 An Approximation Algorithm for Multiple String Alignment 31
6.5.1 Algorithm 31

6.5.2 Time Analysis 31
6.5.3 Error Analysis 32
6.5.4 Other Approaches 33
6.6 The Consensus String 33
6.7 Summary 34

7 Finding Instances of Known Sites 35
7.1 How to Summarize Known Sites 35
7.2 Using Probabilities to Test for Sites 37

8 Relative Entropy 39
8.1 Weight Matrices 39
8.2 A Simple Site Example 39
8.3 How Informative is the Log Likelihood Ratio Test? 40
8.4 Nonnegativity of Relative Entropy 42

9 Relative Entropy and Binding Energy 44
9.1 Experimental Determination of Binding Energy 44
9.2 Computational Estimation of Binding Energy 45
9.3 Finding Instances of an Unknown Site 45

10 Finding Instances of Unknown Sites 47
10.1 Greedy Algorithm 48
10.2 Gibbs Sampler 48
10.3 Other Methods 50

11 Correlation of Positions in Sequences 51
11.1 Nonuniform Versus Uniform Distributions 51
11.2 Dinucleotide Frequencies 52
11.3 Disymbol Frequencies 53
11.4 Coding Sequence Biases 53
11.4.1 Codon Biases 54
11.4.2 Recognizing Genes 54

12 Maximum Subsequence Problem 55
12.1 Scoring Regions of Sequences 55
12.2 Maximum Subsequence Problem 55

12.3 Finding *All* High Scoring Subsequences 56

13 Markov Chains 59

13.1 Introduction to Markov Chains 59

13.2 Biological Application of Markov Chains 60

13.3 Using Markov Chains to Find Genes 61

14 Using Interpolated Context Models to Find Genes 62

14.1 Problems with Markov Chains for Finding Genes 62

14.2 Glimmer 62

14.2.1 Training Phase 62

14.2.2 Identification Phase 63

14.2.3 Resolving Overlap 64

15 Start Codon Prediction 66

15.1 Experimental Results of Glimmer 66

15.2 Start Codon Prediction 66

15.3 Finding SD Sites 68

16 RNA Secondary Structure Prediction 70

16.1 RNA Secondary Structure 70

16.2 Notation and Definitions 71

16.3 Anatomy of Secondary Structure 71

16.4 Free Energy Functions 72

16.5 Dynamic Programming Arrays 73

17 RNA Secondary Structure Prediction (continued) 74

17.1 Recurrence Relations 74

17.1.1 $\boldsymbol{W(j)}$ 74

17.1.2 $\boldsymbol{V(i,j)}$ 74

17.1.3 $\boldsymbol{VBI(i,j)}$ 75

17.1.4 $\boldsymbol{VM(i,j)}$ 75

17.2 Order of Computation 75

17.3 Speeding Up the Multibranched Computation 76

17.4 Running Time 76

18 Speeding Up Internal Loop Computations 78

18.1 Assumptions About Internal Loop Free Energy 78

18.2 Asymmetry Penalty 78

18.3 Comparing Interior Pairs 79