

CSCI1950-L Final Exam

Due: Monday, May 12, 2008 at 10:00 am

To be given to Prof. Istrail's assistant Erin Klopfenstein (CIT 546, ehk@cs.brown.edu)

Late submissions will not be accepted.

IMPORTANT: The Exam has 6 Sections. You **MUST** do the required problems in Sections 1-5. Extra work will be considered for extra credit.

Contents

| | |
|--|----------|
| 1 ASSEMBLY (10) | 2 |
| 1.1 Genome assembly using next generation high-throughput sequencing | 2 |
| 2 BLAST (15) | 3 |
| 2.1 Karlin-Altschul Equation and the BLAST speed-up | 3 |
| 3 HIDDEN MARKOV MODELS (30) | 4 |
| 3.1 Gene Prediction | 4 |
| 3.2 Solution to the Learning Problem | 4 |
| 3.3 Dishonest Casino: Coin Problem | 4 |
| 3.4 Three-sided die | 4 |
| 4 REGULATORY GENOMICS (30) | 6 |
| 4.1 The First Animals of Regulatory Genomics | 6 |
| 4.2 Futility Theorem | 6 |
| 4.3 Position Weight Matrices | 6 |
| 4.4 Adopt a Gene Project | 6 |
| 5 COMPARATIVE GENOMICS (15) | 7 |
| 5.1 Of Mice and Dogs and Chimps and Men | 7 |
| 6 EXTRA CREDIT | 8 |
| 6.1 Assembling a Bacteria with Mathematica | 8 |
| 6.2 HMM simulations with Mathematica | 8 |

1 ASSEMBLY (10)

1.1 Genome assembly using next generation high-throughput sequencing

The company 454 provides one of the very popular high throughput sequencing technologies. In 2005, it produced 20 millionbases in 5 hours with reads of length 100. In 2007 it produced 100 million bases in 8 hours with reads of length 300. In 2008 it produced 1 billion bases in 24 hours with reads of length 400. Do you think it would be possible to design a sequencing and genome assembly operation for a mammalian genome with this technology?

2 BLAST (15)

2.1 Karlin-Altschul Equation and the BLAST speed-up

1. The statistical theory of Maximal Segment Pairs (MSPs) developed by Sam Karlin and Steve Altschul gives the rigorous foundation for deriving statistical significance of database search matches. The central formula

$$E = kmne^{-\lambda S}$$

states that the number of alignments expected by chance E during a sequence database search is a function of the size of the search space mn , the normalized score λS and a constant k . Present informally how is this formula is used to compute the statistical significance of BLAST scores?

2. In essence, BLAST is a linear time algorithm approximating the quadratic time Smith Waterman algorithm and its dynamic programming computation. Describe how BLAST achieves such a reduction in time complexity while retaining close to optimal accuracy.

3 HIDDEN MARKOV MODELS (30)

3.1 Gene Prediction

Give an outline of the construction of a Gene Prediction algorithm using HMMs.

3.2 Solution to the Learning Problem

Present the Expectation Maximization algorithm for the Learning Problem (PB.3).

3.3 Dishonest Casino: Coin Problem

1. A dishonest casino uses either a fair or a biased coin. The fair coin comes out heads or tails with equal probability. The biased coin comes out heads with probability $\frac{3}{4}$ and tails with probability $\frac{1}{4}$. The probability of transitioning from a biased to a fair coin and from a fair to a biased coin is $\frac{1}{10}$. Using the Hidden Markov Model (HMM) just defined representing a dishonest casino compute the most probable sequence of states that generate the following sequence of coin tosses: HHHHHTTTTT. You should fill out a 2 by 10 dynamic programming table.
2. With the same HMM, what is the probability that the 'T' at the seventh position is generated by a biased coin?

3.4 Three-sided die

Consider a different game where the dealer is not flipping a coin, but instead rolling a three-sided die with labels 1, 2, and 3. (Try not to think about what a three-sided die might look like.) The dealer has two loaded dice D_1 and D_2 . For each die D_i , the probability of rolling the number i is $1/2$, and the probability of each of the other two outcomes is $1/4$. At each turn, the dealer must decide whether to (1) keep the same die, (2) switch to the other die, or (3) end the game. He chooses (1) with probability $1/2$ and each of the others with probability $1/4$. At the beginning the dealer chooses one of the two dice with equal probability.

1. Give an HMM for this situation. Specify the alphabet, the states, the transition probabilities, and the emission probabilities. Include a start state *start*, and assume that the HMM begins in state *start* with probability 1. Also include an end state *end*.
2. Suppose that you observe the following sequence of die rolls: 1 1 2 1 2 2. Find a sequence of states which best explains the sequence of rolls. What is the probability of this sequence? Find the answer by completing the Viterbi table. Include backtrack arrows in the cells so you can trace back the sequence of states.

3. There are actually two optimal sequences of states for this sequence of die rolls. What is the other sequence of states?

4 REGULATORY GENOMICS (30)

4.1 The First Animals of Regulatory Genomics

Sea Urchin and Drosophila are the leading two experimental organisms regarding research on developmental regulatory genomics. Present of comparison of the regulatory genomics sections of the two organisms as described in their Science papers describing their genome sequences.

4.2 Futility Theorem

The folklore of regulatory genomics captures the difficulty of transcription factor binding sites prediction in the following "theorem:"

Essentially all predicted transcription factor (TF) binding sites that are generated with models for the binding of individual TFs will have no functional role.

Explain why the state of the art of prediction algorithms is in such a bottleneck. What do you think we need in order to improve the ability of algorithms to predict more accurately such sites?

4.3 Position Weight Matrices

Present the Position Weight Matrices in full generality and the algorithm that takes a genomic DNA sequence and a set of PWMs and finds the highest score subsequences based on the PWMs scores. What is a consensus sequence for TF binding sites and how it compares with the PWM representation and with the Sequence Logo representation of such sites using information theory.

4.4 Adopt a Gene Project

In class we discussed the Cyrene Browser Project devoted to the construction of the cis-Lexicon database of regulatory modules and regulatory information. Give a short description of the project and how such a cis-Lexicon would lead the way to breaking the cis-Regulatory Code – i.e., the discovery of an algorithm that will read regulatory DNA sequences and create the most likely cis-module structure for it.

5 COMPARATIVE GENOMICS (15)

5.1 Of Mice and Dogs and Chimps and Men

Present a comparison of the regulatory genomics of Mice, Dogs, Chimps, and Men – focusing on two categories: *common* to all four organisms regulatory genomics structure, and *specific* to each organism regulatory genomics structure.

6 EXTRA CREDIT

These problems will be made available on the Class webpage Friday afternoon.

6.1 Assembling a Bacteria with Mathematica

6.2 HMM simulations with Mathematica