University of Wisconsin-Madison	Spring 2016
BMI/CS 776: Advanced Bioinformatics	Homework #4
Prof. Anthony Gitter	Due: Tue, Apr 26, 2016, 11:59 PM

Assignment Goals

- Gain experience with interpolated Markov models and suffix trees
- Learn how to design a generalized hidden Markov model for a new task

Part 1: Interpolated Markov Models

We will use the interpolated Markov model approach from GLIMMER to estimate the probability $P_{IMM,3}(\mathbf{A}|\mathbf{TTA})$. For the sub-parts below, suppose we have the following counts in our training data. Show your work for partial credit.

TTAA	15
TTAC	20
TTAG	10
TTAT	5
Total	50

TAA	80
TAC	70
TAG	40
TAT	10
Total	200

AA	450
AC	260
AG	150
AT	40
Total	900

1A: χ^2 *test*

In order to calculate the λ values, we must first perform the χ^2 statistical test to determine whether the distributions of the current character depend on the order of the history. First, compute the χ^2 test statistic, rounded to the tenths place, comparing the 3rd order and 2nd order counts in the training data. Then use the *p*-value table for a χ^2 test with 3 degrees of freedom at https://www.biostat.wisc.edu/bmi776/hw/hw4_chisquare_df3_pvalues.txt to lookup the *p*-value for this test statistic and round to the thousandths place. Finally, compute $d = 1 - p$ to obtain the GLIMMER confidence score.

Repeat the χ^2, *p*-value, and *d* calculations for the 2nd order and 1st order comparison.

Recall that the χ^2 test statistic for an *n* by *m* contingency table is defined as

$$\chi^2 = \sum_{i=1}^{n}\sum_{j=1}^{m}\frac{\left(O_{i,j} - E_{i,j}\right)^2}{E_{i,j}}$$

where $O_{i,j}$ is the observed count in the contingency table and $E_{i,j}$ is the expected count

$$E_{i,j} = \frac{R_i C_j}{N}$$

R_i is the sum of the entries in row *i*, C_j is the sum of the entries in column *j*, and N is the sum of all entries in the contingency table. In this test there are $n = 4$ rows for the nucleotides and $m = 2$ columns for the n^{th} and $(n-1)^{th}$ order histories so there are 3 degrees of freedom.

1B: *Calculating* λ

Use the values of *d* calculated above, the training data counts, and the λ definition from GLIMMER to calculate λ_3(**TTA**), λ_2(**TA**), and λ_1(**A**).

1C: *Interpolated Markov model probability*

Use the λ values and the probabilities estimated from the training data counts to compute $P_{IMM,3}$(**A**|**TTA**).

Part 2: Generalized Hidden Markov Models

Consider a generative probabilistic model of the *cis*-regulatory elements that are upstream of a gene's translation start site that can be used to predict the binding sites of regulatory proteins. This is analogous to the gene finding task but for *cis*-regulatory elements. Specifically, consider a simplified model in which the DNA sequence upstream of the gene contains these elements in the specified order:

- Noncoding DNA, defined here as intergenic DNA and all other DNA that is not assigned to different category
- An optional enhancer region, which contains a single protein binding site if present
- Noncoding DNA
- A promoter region, which contains protein binding sites for one or more transcription factors that are separated by noncoding DNA if there are multiple binding sites
- The 5' untranslated region
- The three base translation start site **ATG**

We will design a generalized hidden Markov model to represent the sequence elements above. There is more than one reasonable way to design the model. You may find it helpful to review the GENSCAN model to help make these design decisions. When considering the protein binding sites in the enhancer and promoter regions, you may assume that you are working in a species like human where transcription factor binding preferences have been previously characterized. Specifically, you may use a library of transcription factor position weight matrices or other prior knowledge if you state your assumptions.

2A: *Defining variables*

Your model will need to include hidden variables to represent the sequence elements above. Define these hidden variables so that you can refer to them in the transition diagram below. What other variables, both hidden and observed, must be included in the generalized hidden Markov model?

2B: *Hidden states*

Draw the valid transitions among the hidden states in the generalized hidden Markov model. These also define the transition probabilities that must be modeled.

2C: *Defining probability distributions*

In addition to the transition probabilities, each hidden state requires two additional types of probability distributions to be associated with it. What are these, and how will you represent them for each of the hidden states above?

Part 3: Generalized Suffix Trees

We will use the generalized suffix tree data structure from MUMmer to identify unique matches and maximal unique matches in the following genomes:

Genome A: **`ACTACGTA$`**
Genome B: **`CGTTACTT#`**

The special terminator characters **`$`** and **`#`** have already been added to the genome sequences.

3A: *Suffix tree construction*

Construct and draw the generalized suffix tree as shown in slide 15 of the Alignment of Long Sequences lecture. Include sequence labels on the edges, and label the leaf nodes with the genome and position.

3B: *Repeated sequences*

In terms of the nodes in the suffix tree you constructed, how can we determine the number of times the sequence **`T`** is repeated in genome A and in genome B?

3C: *Unique matches*

List the sequences of the unique matches and indicate in the suffix tree the internal nodes that correspond to these unique matches. You may annotate or highlight these internal nodes in the drawing from Part 3A.

3D: *Maximal unique match(es)*

List the maximal unique match(es).

Submission Instructions

a) Login to the biostat server **`mi1.biostat.wisc.edu`** or **`mi2.biostat.wisc.edu`** using your BMI (biostat) username and password.
b) Compile your answers in a single file **`<USERNAME>.pdf`** and copy it to the directory **`/u/medinfo/handin/bmi776/hw4/<USERNAME>`** where **`<USERNAME>`** is your BMI (biostat) username.
c) Make sure to write the number of late days used at the first line of your pdf file.