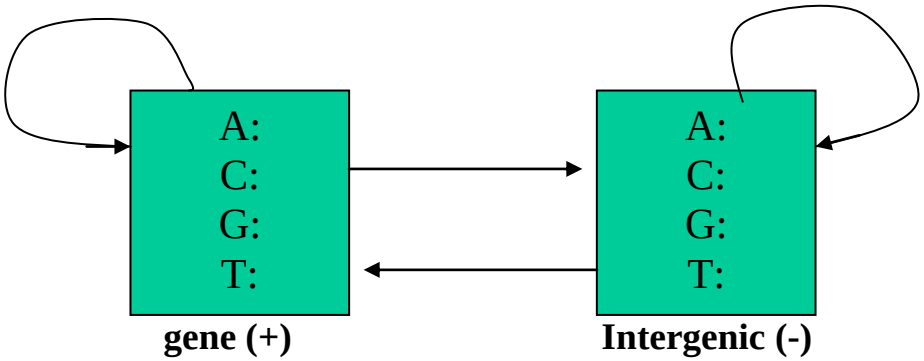
CISC 436/636 Computational Biology & Bioinformatics
Fall 2016
Homework 2

Due date: October 25, 2016

Hidden Markov Models

Assume the genomic sequences are comprised of genes and intergenic regions, and can be modeled by a hidden Markov model with an architecture depicted as follows.



a. **Model Training**.

(636 Section) Estimate the model parameters using Maximum Likelihood (ML) with **Viterbi** training. Note: you should download the unlabelled training data.

i. Randomly initialize model parameters
ii. Run Viterbi algorithm (see part b) on the training sequence and update the model parameters
iii. Check the predicted labels. Repeat step ii until the difference for the predicted labels between two iterations is less than 1%.

Report the final values for model parameters and plot the difference (Y-axis) as a function of the number of iterations (X-axis).

(436 Section) Estimate the model parameters using Maximum Likelihood (ML) by counting. Note: you should treat the labeled training data. Report the model parameters.

b. **Predicting.** Implement the Viterbi algorithm as discussed in the class. Note: you need to implement the logarithm transformation in order to avoid the underflow.

c. **Testing**.

- Run the Viterbi algorithm from part b on the test sequence.
- Evaluate the performance by compare the predicted labels with the true labels. If a gene (+) position is predicted correctly, we say that prediction is a true positive, whereas if an intergenic (-) position is predicted as gene, we say that prediction is false positive. Similarly, true negative and false negative can be defined by switching the roles of gene and intergenic.

- Calculate the sensitivity Sn, specificity Sp, and correlation coefficient CC, as defined in the following.

  $$Sn = TP/(TP+FN),$$
  $$Sp = TP/(TP+FP)),$$
  $$CC= (TP \times TN - FP \times FN)/ \sqrt{(PP \times PN \times AP \times AN)},$$

  where TP, FP, AP, and PP are true positive, false positive, actual positive, and predicted positive respectively; and TN, FN, AN, and PN are true negative, false negative, actual negative, and predicted negative respectively.

- Report the Sn, Sp, and CC.

Name your program as "xxxx_align", where xxxx is your last name appended with first initial.

Synopsis: **`xxxx_hmm <train_data> <test_data>`**

Download the training and testing sequences from

`http://www.cis.udel.edu/~lliao/cis636f16/hw2_data/`