

6.095/6.895 Fall 2005 - Project Proposal and Problem Set 4

Due: November 14, 2005 at 8pm

0. **Project Proposal.** You should provide a 1 to 2 page project proposal. Your project should be either
- (a) The implementation of a computational biology algorithm found in the literature along with its analysis on some data set.
 - (b) The description and analysis of a novel computational biology algorithm.
 - (c) The careful analysis including criticism, corrections or improvements of a conference or journal article.

Your proposal should include

- (a) Your partner, if you are working on one. At most two people can work on each project. If you are working with a partner a division of labor should be indicated.
- (b) Citations to one or more papers related to your proposal.
- (c) What you specifically hope to accomplish.
- (d) Milestones for the completion of subtasks.

Projects will be graded on the basis of the level of challenge, creativity and feasibility.

1. In this problem we will explore the score of the best RNA folds we can obtain for random sequences. Thus, we can use this to infer how big a score needs to be before we suspect it is biologically significant.

You should write code (in your programming language of choice) to generate random RNA sequences and then score them using the Nussinov algorithm described in class. Use a score of 1 for $A - U$, $G - U$, $C - G$ and a score of 0 for all other pairs.

- (a) What is the average Nussinov score for 1000 RNA sequences of length 100 when the random RNA string has equal probabilities for each base?
 - (b) How does this score change as you increase or decrease the length of the RNA sequence? What sort of relationship (linear, quadratic, etc) do you see?
 - (c) How does this score change as you change the $G - C$ content of the DNA sequences? Is the score approximately symmetric about $G - C$ content 0.5? Explain why it is or is not symmetric.
 - (d) What does this tell you about how you should compute baseline scores for RNA sequences? What are other biases (i.e. background models) you may want to take into account?
 - (e) Include a printout of your code.
2. (a) Give a context free grammar for a valid set of parenthesis and brackets, i.e. a grammar that could derive $[(())((()()))]$.
- (b) Consider the stochastic context free grammar,

$$A \rightarrow Aa | \epsilon$$

where $P(A \rightarrow Aa) = p$ (thus, $P(A \rightarrow \epsilon) = 1 - p$). What is the expected length of a derived sequence?

- (c) Consider the stochastic context free grammar,

$$A \rightarrow AA|a$$

where $P(A \rightarrow AA) = p$. What is the expected length of a derived sequence? What happens if p is "large"?

- (d) Write a stochastic context free grammar that generates the same sequences as the HMM for the dishonest casino problem. Recall, that the dishonest casino plays with either a fair die where each of the 6 values has equal probability or a loaded die where the probability of rolling a 6 is $1/2$ and all other probabilities are $1/10$. The probability of switching from a fair die to a loaded die is 0.05 and the probability of switching from a loaded die to a fair die is 0.1.

3. Consider $\pi = 3 \ 4 \ 6 \ 5 \ 8 \ 1 \ 7 \ 2$

- (a) Show the sequence of reversals as well as the intermediate permutations when performing the breakpoint reversal sort algorithm on π (break ties arbitrarily).
- (b) Find a shorter sequence of reversals than the one found in part (a).
- (c) Do you know if the sequence of reversals you found in part (b) is optimal?