BMI/CS 776 Spring 2015
Homework #4

Prof. Colin Dewey

Due Thursday April 30th, 2015 by 11:59pm

The goal of this assignment is to become familiar with the Nussinov algorithm and *stochastic context-free grammars (SCFGs)*. To submit your homework, please copy all relevant files to the directory:
`/u/medinfo/handin/bmi776/hw4/USERNAME`
where `USERNAME` is your account name for the BMI network. You must submit a file named `README` to this directory, which gives directions on how to compile (if necessary) and run your programs. For each question below, the `README` file should list the files relevant to that question (e.g., code, other files with written answers).

1. Write a program `Nussinov` that takes as input an RNA sequence, and outputs the base-paired positions of a structure that maximizes the number of base pairings. The program should be run with the command line

   ```
   Nussinov RNASEQUENCE
   ```

   where `RNASEQUENCE` is a string of `A`, `C`, `G`, and `U` characters. The program should print to standard output the positions of all pairs of base-paired positions (using one-based indexing), in lexicographical order. For example, the command

   ```
   Nussinov UCCAGG
   ```

   should output

   ```
   2 6
   3 5
   ```

2. (Textbook exercise 9.6) Consider the complete language generated by the CFG:

$$\begin{array}{rcl} s & \rightarrow & Aw_1U \mid Cw_1G \mid Gw_1C \mid Uw_1A \\ w_1 & \rightarrow & Aw_2U \mid Cw_2G \mid Gw_2C \mid Uw_2A \\ w_2 & \rightarrow & Aw_3U \mid Cw_3G \mid Gw_3C \mid Uw_3A \\ w_3 & \rightarrow & GAAA \mid GCAA \end{array}$$

   (a) Specify a regular grammar that generates *exactly* the same language.

   (b) Why is describing this sequence family with a regular grammar not a good idea?

3. Consider the following RNA (only consisting of $\texttt{A}$ and $\texttt{U}$) SCFG:

$$\begin{array}{rcl} s & \rightarrow & \overset{0.8}{t} \mid \overset{0.2}{l} \\ t & \rightarrow & \overset{0.3}{AtU} \mid \overset{0.3}{UtA} \mid \overset{0.3}{At} \mid \overset{0.1}{l} \\ l & \rightarrow & \overset{0.4}{Al} \mid \overset{0.4}{Ul} \mid \overset{0.1}{A} \mid \overset{0.1}{U} \end{array}$$

   (a) Give one derivation, π, for the string $\texttt{AAU}$ using this grammar.

   (b) What is the (prior) probability, $P(\pi)$, of the derivation you gave in part (a)?

   (c) What is the probability of the string $\texttt{AAU}$, $P(\texttt{AAU})$, given the grammar?

   (d) What is the posterior probability, $P(\pi|\texttt{AAU})$, of the derivation you gave in part (a)?

4. Consider the SCFG described below:

$$\begin{array}{rcl} s & \rightarrow & \overset{0.002}{ss} \mid \overset{0.7}{p} \mid \overset{0.28}{u} \mid \overset{0.018}{\epsilon} \\ u & \rightarrow & \overset{0.5}{bs} \mid \overset{0.5}{sb} \\ p & \rightarrow & \overset{0.25}{AsU} \mid \overset{0.25}{UsA} \mid \overset{0.25}{CsG} \mid \overset{0.25}{GsC} \\ b & \rightarrow & \overset{0.25}{A} \mid \overset{0.25}{U} \mid \overset{0.25}{C} \mid \overset{0.25}{G} \end{array}$$

where ϵ is the empty string. With this grammar, the log probability of a sequence (x) and its maximum likelihood parse ($\hat{\pi}$) can be written in the form

$$\log P(x, \hat{\pi}|\theta) = f_1(\hat{\pi}) \log(g_1(\theta)) + f_2(\hat{\pi}) \log(g_2(\theta)) + \log(g_3(\theta, |x|)) \qquad (1)$$

where $|x|$ is the length of sequence x.

(a) Determine the functions $f_1, f_2, g_1, g_2,$ and g_3. Provide both the symbolic form of these functions (i.e., with specific parameters as variables) and the values of these functions given the specific values of the parameters provided above. *Hints: (1) f_1 is the number of base-pairings implied by the parse and f_2 is a count of some other feature of the structure implied by the parse. (2) Take advantage of the identity $|x| = 2 \times$ (number of base-pairings) + (number of unpaired bases).*

(b) With this interpretation of the objective function for the CYK algorithm given this grammar, explain how the Nussinov algorithm differs from the CYK algorithm (with this grammar) in terms of the optimal structures that are returned.