BMI/CS 776
Spring 2009
HW #2

(1) Suppose sequence A is m bases long and sequence B is n bases long, and we have an anchor match that starts at coordinates (a1, b1) and ends at coordinates (a2, b2). Show pseudocode for a global alignment algorithm that constrains the alignment to be "close" to this anchor. See the figure for "Step 3" in the LAGAN notes for an illustration of what "close" means in this context.

(2) Consider the two toy genomes:

Genome X: **taccggt**

Genome Y: **ccgctac**

(a) Show how MUMmer would use a suffix tree to find the MUMs in these sequences.

(b) Show the threaded trie that LAGAN would construct to index the 2-mers in Genome X.

(c) Show the path that would be followed through the threaded trie as Genome Y is processed

(3) For the following SCFG, show all of the possible parse trees for the sequence **cggcg**. Which parse is the most probable one? Note that the probability of each production is shown in parentheses following it.

S → B S	(0.1)
S → B	(0.1)
S → **g** S **c**	(0.4)
S → **c** S **g**	(0.4)
B → **c**	(0.5)
B → **g**	(0.5)

(4) Show how the Inside algorithm would determine the probability of the string **ccgg** under the following grammar.

S → D D	(0.3)
S → B T	(0.7)
T → B D	(0.6)
T → D B	(0.4)
D → B B	(1.0)
B → **c**	(0.5)
B → **g**	(0.5)