CS 224 Fall 2009 HW 5, Due thursday Nov. 19

1. Read the proof of Theorem 4.3.2 on page 67 of the notes on lower bounds. Is it essential to the proof and the method for finding $S^*(I)$, that M is first modified to be $\tilde{M}$? Explain. A more specific question is: If we don't change M, how must the statement of the theorem be changed in order to identify $S^*(I)$ in terms of M instead of in terms of $\tilde{M}$?

2. As commented on the bottom of page 51 of the notes on lower bounds, if sites p and $q > p$ in M are incompatible, then in any ARG for M, p and q must be together on some recombination cycle whose crossover point is in the range $(p, q]$. Lemma 4.1.1 on that page proves that there must be such a crossover point in any ARG for M, and an earlier result showed that p and q must be contained in some common recombination cycle. Your problem is to prove that the common recombination cycle must have crossover point in the range $(p, q]$.

3. Given a set K of k binary strings, each of length n, we want to find each triple of strings S_1, S_2, S_3 such that a single crossover recombination between S_1 and S_2 produces S_3. For any triple, S_1, S_2, S_3, we can easilly determine in $O(n)$ time whether S_1 and S_2 can recombine to create S_3. Explain one such way.

Therefore, all desired triples could be found in $O(k^3 n)$ time. However, this problem can be solved in $O(nk + k^3)$ time. Explain how (hint: think suffix tree).

Can you also see a way to solve the problem in $O(nk + w)$ time, where w is the number of desired triples? I don't know the answer to that.

4. Given a set K of k binary strings, each of length n, and a binary string S of length n, we want to create S from K by a series of single crossover recombinations, *minimizing* the total number of recombination events. A string in K can be used several times in such a scenario. Show how to do this in $O(nk)$ time.

5. Lemma 4.3.2 in the notes (the self-derivability lemma) is correct in the context of checking whether $H(M(S^*(I)))$ should be used for $b(I)$, or if $b(I)$ should be $H(M(S^*(I))) + 1$. That is, it may not be true for arbitrary S, but it is true for $S^*(I)$. So replace S with $S^*(I)$, and explain now why the proof works. The key issue before was the implicit claim in the proof that there are only $D_c(M(S))$ tree nodes, so now the key issue is why there are only

$D_c(M(S^*(I)))$ tree nodes.