# BMI/CS 776 Spring 2008
# Homework #5

Prof. Colin Dewey

Due Tuesday, April 15th, 2008 by 11:59pm

The goal of this assignment is to become familiar with two pattern matching techniques: *suffix trees* and *locality-sensitive hashing*. You have three options for turning in this homework:

- Copy all relevant files to the directory:
  `/u/medinfo/handin/bmi776/hw5/USERNAME`
  where `USERNAME` is your account name for the BMI network.
- Send it to me by email
- Turn it in on paper during class on Tuesday, April 15 or put it in my mailbox by 5pm on that day.

1. Draw the suffix tree for the DNA sequence `CACTACGTACG`. Include the suffix links from internal nodes as used in Ukkonen's suffix tree construction algorithm.

2. The algorithm described in class for finding all occurences of a query string $Q$ in a suffix tree $\mathcal{T}$ takes time $O(n+k)$, where $n$ is the length of $Q$ and $k$ is the number of occurrences of $Q$ in the string encoded by $\mathcal{T}$. Write an algorithm for preprocessing the suffix tree $\mathcal{T}$ in $O(m)$ time (where $m$ is the length of the database string) such that finding a *single* occurence of $Q$ in $\mathcal{T}$ takes $O(n)$ time (hint: label each internal node of the tree with a convenient value).

3. Give updated bounds for the false negative ($\rho_{fn}$) and false positive ($\rho_{fp}$) rates for LSH-ALL-PAIRS if we form hash functions by picking $k$ of $d$ positions at random *without* replacement.

4. Let $S$ be the set of all bit strings of length $\ell$. For $x \in S$, let $ones(x) = \{i : x_i = 1\}$, i.e., the set of all positions in $x$ that are equal to one. We define the *set resemblance*, $s(x,y)$, for $x, y \in S$ as $s(x,y) = \frac{|ones(x) \cap ones(y)|}{|ones(x) \cup ones(y)|}$, i.e., the fraction of positions that are equal to one in both strings out of all positions that are equal to one in either of the

strings. For the case of $x = 0^\ell$ (the all-zero string), define $s(x, x) = 1$. From this we can define a distance function $d(x, y) = 1 - s(x, y)$. A great locality-sensitive hash function for this distance measure is called *minhash*. Given a random permutation, $\pi$, of positions $\{1, 2, \ldots, \ell\}$, $minhash_\pi(x) = argmin_{i \in ones(x)}(\pi(i))$, i.e., the index of the first non-zero position in $x$ in the position ordering given by $\pi$. Show that that $\mathbb{P}_\pi[minhash_\pi(x) = minhash_\pi(y)] = 1 - d(x, y) = s(x, y), \forall x, y \in S$.