Notes

You must choose to answer one of the following questions. Your answer must be around 5 typewritten pages, with additional material in appendices if needed. The questions are difficult and vary in skills required, ranging from algorithms to biology, so please leave some time to answer them. Let me know if you cannot find references.

Questions

1. Describe an efficient algorithm for computing the sensitivity of a gapped seed for database search.
   (This problem needs algorithmic skills. You might want to read the discussion in lecture 7 and the comments on the final slide. Relevant papers are the reading for L7).

2. Define a \((d, l, m)\)-repeat as a sequence of length \(l \pm d\) that occurs \(m\) times in the genome of length \(n\), with each pair of sequences being \(d\) edit-distance apart. To detect repeats without a pre-defined library of repeats, we often need to do genome to genome comparisons in \(O(n^2)\) time. For a large genome, this is computationally expensive. Describe an efficient \(o(n^2)\) algorithm for detecting if a given subsequence in a genome is repetitive.
   This problem requires some theoretical ideas. The goal is to compute a hash function for hashing every length \(l\) string, so that strings which are close in edit distance get hashed to the same location. A relevant starting point might be the following paper:

3. Summarize the arguments in the following 4 papers: the first should describe known mitochondrial phylogenies, and the next two address whether mtDNA could have been paternally inherited. Redo the phylogenetic scenarios the first 2 papers using perfect phylogeny. You should be able to do it by hand.
   (Skills needed: some knowledge of biology, genetics).

4. A “structural polymorphism” is a large scale genetic event (insertion, deletion, or inversion) that is polymorphic in a population. Review the following papers on structural polymorphisms. Your review must summarize computational techniques used to detect such polymorphisms. Review additional literature and/or do some computational tests to answer the following: Is there a genomic signal (low copy repeats/motifs) that can cause such polymorphisms to occur?
   Common deletion polymorphisms in the human genome
   Steven A McCarroll1, 2, 3, Tracy N Hadnott1, George H Perry4, Pardis C Sabeti3, Michael C Zody3, Jeffrey C Barrett3, Stephanie Dallaire4, Stacey B Gabriel3, Charles Lee4, 5, Mark J Daly2, 3, 5, 6, David M Altshuler
A high-resolution survey of deletion polymorphism in the human genome Nature Genetics 38, pp75 - 81 (2006)

-Donald F Conrad, T Daniel Andrews, Nigel P Carter, Matthew E Hurles and Jonathan K Pritchard