Algorithms for Bioinformatics (autumn 2010)

Study group (Wed 13.10, 10-12, B222)

1. **Sequencing by hybridization.**

   A measurement from a hybridization experiment estimates that the 3-mer spectrum of s would be $Spectrum(s, 3) = \{\texttt{GAG,GAT,TAG,ATA,ATA,AGA,TAC}\}$. Construct s by the Eulerian path approach described at the lecture, taking into account that there might be one k-mer missing from the measured spectrum.

2. **Computation around ℓ-mers.**

   Show that computing the frequency/count of each ℓ-mer in a string $s \in \Sigma^*$ can be done in $O(|\Sigma|^\ell + |s|)$ time by filling a table of size $|\Sigma|^\ell$ and scanning s once from left to right. Here $|\Sigma|$ is the alphabet size (e.g. 4 for DNA).

   On large ℓ, $O(|\Sigma|^\ell)$ may be the dominating term. Is there any way of doing the same computation in $O(|s|)$ time?

3. **Analyzing Random Projections algorithm.**

   a) Assume k, s, and ℓ are fixed and your task is to be an *adversary* trying to find a motif finding problem instance where the Random Projections algorithm fails in finding the optimal solution. What is the maximum score of a planted motif (as a function of k, s, t, n and ℓ) that the algorithm has no chances in finding.

   b) Assume now that only s and ℓ are fixed. How would you design the planted motif so that it is hardest to find with any k.

   c) Based on the above, derive an analytic formula for the probability that one out of m random projections succeeds in finding the correct k-projection of any planted motif.

   d) Derive an analytic upper bound for the probability that some projection finds a random motif.

4. **Protein sequencing.**

   Let T be the *theoretical mass spectrum* of a peptide $P = p_1 p_2 \cdots p_n$ (short sequence of amico acids), consisting of the masses of its prefixes and suffixes computed from its molecular formula. *Tandem mass spectrometry (MS/MS)* can estimate the same mass spectrum, but unfortunately, checking whether the measured spectrum M is identical to T is not enough to identify the peptide: While breaking the peptide at each possible bond, some small parts from both fragments of the molecule may be lost. Also the experimental spectrum may be completely missing some masses and may be containing some masses from background chemical noise.

   For simplicity, let us assume that only the masses of prefixes are included in M and T, and that the masses can be ordered from smallest to largest corresponding to the length of the prefix. How would you define a good distance measure between M and T. How can it be computed? Can you extend your distance measure

definition to include simultaneously prefix and suffix masses? Can you extend the algorithm for this case?

5. **Microarrays and probe selection.**

   Microarrays were designed for sequencing by hybridization problem, yet, they are now commonly used for measuring *gene expression*: A *probe* (DNA fragment of length ℓ) is designed for each gene such that the complementary DNA of its mRNA product hybridizes to the probe. Thousands of copies of the same probe are are attached to the spot reserved for that gene in the microarray. Expression level of a gene can be estimated by observing how much cDNA product is hybridized to that spot. *Probe selection problem* is to design a probe for each gene such that only the mRNA product of that gene hybridizes to the probe. Cross-hybridizations can occur when there is similar sequence in some other gene's DNA. How would you solve the probe selection problem given the genome sequence and the genes of interest?