Crash Course on Computational Biology for Computer Scientists – Homework

Bartek Wilczyński

November 24, 2016

The homework consists of 6 exercises. Please send your responses **to bartek@mimuw.edu.pl by December 21th**. If you want to obtain a grade N, you need to provide correct solutions of at least $N-1$ exercises. For example, for a very good grade (5) you need to solve 4 exercises. If you have questions, please either send me questions by e-mail. I will put answers to such questions to the webpage http://regulomics.mimuw.edu.pl/wp/categories/phdopen .

Exercise 1. Sequence comparison

In my lecture concerning sequence aligmnent we discussed briefly a heuristic by Carillo and Lippman that allows for faster computation of sequence alignments by using a lower bound. Can you implement a version of Smith-Waterman algorithm using such lower bound based on ungapped alignment?

Exercise 2. Multiple sequence alignments

We have discussed progressive alignment algorithm using a simple UPGMA tree based on pairwise alignments. Can you implement a simple progressive alignment algorithm for DNA sequences that uses a modified sequence comparison measure that uses the triplet code similarity? In particular, we would like to obtain a global multiple sequence alignment algorithm that allows for gaps only in multiples of 3 and scores occurence of triplets encoding the same aminoacid as a match and triplets encoding different aminoacid as a mismatch. It should also consider all 3 reading frame offsets between compared sequences.

Exercise 3. Read mapping to the genome

We have briefly discussed third generation sequencing that generates very long reads (¿1kbp) with relatively high error rates. LEt us consider a theoretical scenario with reads of length=10000 with randomly occuring errors with probability α, and the genome such that any region of length $\geq k$ is unique in the genome. How low α needs to be so that the expected proportion of uniquely mappable reads is higher than 0.99? can you suggest (not necessarily implement) an effective mapping strategy for such reads?

Exercise 4. Genome Assembly

When we discussed de novo genome assembly we mentioned deBruijn graph construction using the Eulerian path approach. I have mentioned that repetitions of k-mers in the genome create ambiguities in the resulting Euler path. Assuming a theoretical scenario, where we have an ideal k-mer spectrum of a genome, can you derive the number of possible Euler paths assuming you know the frequencies of each k-mer in the genome, but you do not know the genome sequence?

Exercise 5. HMMs and DBNs

As we discussed in lecture 5, Dynamic Bayessian Networks and Hidden Markov Models are both representation of discrete stochhastic processess. An HMM cosists of a state space, emitted symbol alphabet, transition matrix and emisson matrix. A DBN is described by a set of variables, sets of possible values for ech of the variables and conditional probability distributions. It is known that the set of processess that can be described by HMM and DBNs is the same. Can you prove that for any HMM, there exists a DBN representing the same process and vice-versa?

Exercise 6. Data storage in DNA

In lecture 6, we have discussed encoding of information in the synthesized oligonucleotide libraries. I've noted that the method described by Goldman et al (http://europepmc.org/articles/pmc3672958) was more thoroughly designed than that by Church et. al (http://science.sciencemag.org/content/337/6102/1628). Can you compare the supplementary materials to both papers and identify at least 3 potential issues that were considered by Goldman and not considered by Church?