Elements of Bioinformatics (autumn 2010)

Lecturers: Veli Mäkinen, Lars Paulin, Esa Pitkänen

Exercise 5

Thu 9.12, 10-12, B222, Veli Mäkinen

Choose any 5 assignments from below (each assignment gives 1 point, 5 points is maximum for each week; you can do more for better learning, of course).

1. **Ultrametric condition.**

   Consider the *three-point condition*: A symmetric distance matrix $D = \{d_{ij} \mid 1 \leq i \leq n, 1 \leq j \leq n\}$ corresponds to an *ultrametric tree* if and only if $d_{ij} \leq \max(d_{ik}, d_{kj}\}$ for all i, j, k. An ultrametric tree for D is an *edge-weighted tree* (positive weight associated to each edge) such that the sum of weights in the path from leaf i to node v and from leaf j to v are both $\frac{1}{2}d_{ij}$, where v is the lowest common ancestor of i and j. (Notice this is an alternative but semantically identical definition to what was used in the lectures).

   a) Prove that the three-point condition can identically be stated as follows: two of the three values d_{ij}, d_{ik}, and d_{kj} are equal and one is smaller than the others.

   b) Prove that the condition holds.

2. **Sequence assembly with virtual SBH approach.**

   Draw the virtual sequencing by hybridization (SBH) graph (page 45 on Mon 29.11 lectures) with $k = 5$ on the following set of reads:

   ```
   GACCAGT
   CAGACTA
   ACCAGAC
   AGACTAG
   CCAGACT
   AGACTAG
   ```

   Can you deduce the original sequence by analyzing the graph? *Hint. One of the reads has one measurement error, others are correct. The experiment happens to have full coverage (there is at least one read starting from every position of the original sequence), and all reads are from the same strand.*

3. **Sequence assembly in practice.**

   Try out Velvet software[1]

[1] Installed on `users` at directory `/home/tkt_mbie/software/velvet_1.0.15`. Check syntax of `velveth` and `velvetg`; run these programs in this order.

a) to assemble the reads of previous assignment.

b) to assemble simulated reads of length 50 generated with $20x$ average coverage from any DNA sequence of length 10000 (add repetitions to the DNA sequence to make a challenging set of reads, and remember that reads should be from both strands).

c) to assemble simulated *paired end* reads of length $50 + 50$ with exact span 1000 generated with $20x$ coverage from the same DNA sequence you used for case b). Note: both ends of paired end reads are from the same strand.

d) same as c), but simulate paired end reads with *average* span 1000.

For generation of reads given an average coverage c, just read the substrings of length m from nc/m random positions of the sequence of length n. For average span s, you may use e.g. uniform distribution in a reasonable size interval whose midpoint is s.

4. **Gaps in assembly.**

Consider the read simulation of previous assignment. Derive the formulas for the following cases (making some assumptions and approximations on the distribution when necessary).

a) What is the probability that a given position in the sequence is not covered by any read?

b) What is the expected number of positions not covered by any reads?

c) What the coverage should be so that with probability p there are no positions that are not covered by any read?

5. **Sequencing technologies I.**

SOLiD uses a two-base coding defined by the matrix (row=first base, column=second base):

```
  A C G T
A 0 1 2 3
C 1 0 3 2
G 2 3 0 1
T 3 2 1 0
```

Consider a *colour-space* SOLiD read:

```
T012023211202102
```

a) Decode the above read.

b) What happens if there is a sequencing error at i-th position (substitution or indel of number)?

c) What happens if there is a SNP at i-th position (that is, reference DNA has different base in the position from where the read is originating in the donor DNA)?

6. **Specifying a metabolic network model.**

   Consider the metabolic network defined by the following reaction equations:

$$
\begin{aligned}
r_1 &: \quad A \to C \\
r_2 &: \quad B + C \to E \\
r_3 &: \quad D + E \to F \\
r_4 &: \quad E \to G \\
r_5 &: \quad 2F \to H + J \\
r_6 &: \quad B + G \to I \\
r_7 &: \quad I \to J
\end{aligned}
$$

   Add exchange reactions r_8, r_9, r_{10}, r_{11} and r_{12} to metabolites A, B, D, H and J, respectively, so that non-zero flux through system becomes possible. Construct the stoichiometric matrix $S = (s_{ij})$ corresponding to reactions $r_1, \ldots, r_7$ and the exchange reactions $r_8, \ldots, r_{12}$.

7. **Determining maximum product yield.**

   (a) Give the task of maximizing the production of metabolite J in the above metabolic network as a flux balance analysis problem. You can assume the following:

   - uptake of B is constrained to $0 \leq v_9 \leq 1$ and
   - other exchange fluxes are unconstrained.

   (b) Solve the problem by hand or by using any linear problem solver (e.g., MATLAB linprog or lp_solve). What is the maximum production rate of J? What are the fluxes corresponding to this rate? Did you get a unique solution? Why/why not?

****Voluntary exercise: Additive condition (2 extra points)****

Prove that the four-point condition is both necessary and sufficient for a matrix of distances to correspond to an additive tree.