Elements of Bioinformatics (autumn 2010)

Veli Mäkinen

Exercise 5 – solutions

1. **Ultrametric condition.**

   (proof sketches)

   a) Assume that there is an ultrametric tree for the distances. Then it is easy to prove by case analysis that any way of having i,j,k in the leaves of an ultrametric tree, the two versions of the condition are identical and that they both hold (i.e. proving one direction of case b)).

   b) Because of case a), it is enough to show that if one of the versions of the condition holds, then there is an ultrametric tree. Consider the UPGMA algorithm (on min or max metric as well) for constructing the ultrametric tree. It always chooses the minimum d_{ij} among all values, so the new formulation for the three point condition says that $d_{ik} = d_{jk}$ for all k and the (shared) lowest common ancestor of d_{ik} and d_{jk} is above the lowest common ancestor of d_{ij} in any ultrametric tree. The same argument holds on all steps of the algorithm. It is then easy to see that the algorithm creates an ultrametric tree if the condition holds.

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

   The edges of the graph are `GACC->ACCA->CCAG->CAGT`, `CAGA->AGAC->GACT->ACTA->CTAG`, and `CCAG->CAGA`.

   One solution is `ACCAGACTAGT` with read `GACCAGT` having one error.

3. **Sequence assembly in practice.**

   See `http://www.cs.helsinki.fi/u/vmakinen/elements10/ex5_3`, `http://www.cs.helsinki.fi/u/vmakinen/elements10/ex5_3a_reads.fa`, `http://www.cs.helsinki.fi/u/vmakinen/elements10/ex5_3b.py`, `http://www.cs.helsinki.fi/u/vmakinen/elements10/ex5_3c.py`, and `http://www.cs.helsinki.fi/u/vmakinen/elements10/ex5_3d.py`.

4. **Gaps in assembly.**

   a) $(\frac{n-m}{m})^{nc/m}$.

   b) $\mu = n(\frac{n-m}{m})^{nc/m}$.

   c) Let X be the random variable denoting the number of positions not covered by any read. Set coverage c so that $\mu = n(\frac{n-m}{m})^{nc/m}$ is smaller than $1 - \sigma\sqrt{1/q}$, where σ is the variance of X and $q = 1 - p$. Then by Chebyshev's inequality $Pr(X \geq 1) \leq Pr(|X - \mu| \geq \sigma\sqrt{1/q}) < q$. Hence, $Pr(X < 1) > p$. The two problematic points here are that (1) we do not have closed form for the variance due to dependencies, and (2) it may not be possible to set $\mu < 1 - \sigma\sqrt{1/q}$

for the given value of p. Estimating the parameters by simulation is possible, but nontrivial because of the created circular dependency with variance and expected case.

Another way to proceed without knowing the variance is to set c so that $\mu < 1$. Then create independently $\log_2 1/p$ sets of reads. With probability $\frac{1}{2}^{\log_2 1/p} = p$ one of the read sets covers all positions. Sufficient coverage is then $c \log_2 1/p$, since one can as well merge all read sets into one assembly.

5. **Sequencing technologies I.**

   a) `TTGAAGCTGTCCTGGA`.

   b) The suffix of the read becomes random.

   c) Two adjacent colors are affected.

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

   See from previous year's solutions: `http://www.cs.helsinki.fi/bioinformatiikka/mbi/courses/08-09/memo/excercises/ex4/ex4solutions.pdf`

7. **Determining maximum product yield.**

   See from previous year's solutions: `http://www.cs.helsinki.fi/bioinformatiikka/mbi/courses/08-09/memo/excercises/ex4/ex4solutions.pdf`

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

See `http://homepages.inf.ed.ac.uk/opb/homepagefiles/OldStuff.html`.