Instructions and questions

Choose and answer at most **five** from the following list of six questions. Each question is worth of 3 points, the total maximum being 15 points. Your may write you answers in English, Finnish, or Swedish. Using ordinary paper and pen/pencil is permitted but not required. Submit your solutions to the instructor **by March 14, 2011**, either by email at `mikko.koivisto@cs.helsinki.fi` (attached in plain txt, ps, pdf, or doc format) or by regular mail to `Mikko Koivisto/Dept of CS/HIIT` (e.g., via the janitor of the Exactum building). Remember to include your name and student number into the submitted documents.

Note: group work and plagiarism are strictly not permitted.

1. Consider n genotypes over two loci—one with alleles A and a and the other with alleles B and b—sampled independently from the population.

   (a) Derive the test statistic of Fisher's exact test for genotypic linkage equilibrium at the two loci in terms of the observed genotype counts, n_{uvxy}, $uv \in \{AA, Aa, aa\}$, $xy \in \{BB, Bb, bb\}$.

   (b) How fast can you evaluate the test statistic?

   Grading: correct derivation in (a) yields +1 point; in (b) a slow but correct algorithm and analysis yields +1, whereas a fast algorithm and correct analysis yield +2 points.

2. Extend the likelihood model given for unrelated genotypes and trios to quartets, that is, mother–father–daughter–son tuples $\mathbf{g} = (g^{\mathrm{m}}, g^{\mathrm{f}}, g^{\mathrm{d}}, g^{\mathrm{s}})$:

   (a) Give an expression for the likelihood $L(p; \mathbf{g}) = \Pr(\mathbf{g}; p)$ in terms of the haplotype frequencies p_h and the SNP-wise error models $\tau_s(g_s; g'_s)$.

   (b) How fast can you evaluate the likelihood given the haplotype frequencies and error models?

   Grading: correct expression in (a) yields +1 point; in (b) a slow but correct algorithm and analysis yields +1, whereas a fast algorithm and correct analysis yield +2 points.

3. Present a pseudo code for an algorithm that for a given list of n haplotypes over m SNPs, possibly with missing data, outputs whether or not the haplotypes satisfy the Patil et al. coverage criterion of haplotype blocks; you may fix $\beta = 0.80$. before applying the coverage criterion, the algorithm should remove all haplotypes that are ambiguous in the original data. What is the asymptotic running time of your algorithm in terms of n and m?

   Grading: basic understanding of the key concepts demonstrated yields +1 point; correct algorithm with sufficient description yields +1 point; correct and rather tight analysis of the runtime yields +1 point.

4. Examine the following two trio genotypes (mother, father, child from top to bottom) over 9 SNPs. Which of the six genotypes most probably carry a deletion of some length, if any? Which markers are spanned by the deletion? Give detailed arguments that support your guesses.

   012201201
   211101221
   121202211

   101200221
   221012100
   211002101

   Grading: reasonable guesses yield 1 point; correct argumentation adds 1–2 points.

5. Read the article

   - O'Reilly PF et al., invertFREGENE: software for simulating inversions in population genetic data. Bioinformatics 26, 838–840 (2010)

   and answer the questions below. A link to the article can be found on the web page of the course; if you do not manage to download the article, contact the instructor for getting a copy.

   (a) Why is the software useful?

   (b) How do the authors support their claim that it simulates "realistic" inversions?

   Grading: correct and relevant answers yield (a) 1–2 points, (b) +1 point.

6. Read the article

   - Kidd JM et al., Mapping and sequencing of structural variation from eight human genomes. Nature 453, 56–64 (2008)

   and summarize three main conclusions (not just the results mentiond in the abstract) of the paper. A link to the article can be found on the web page of the course; if you do not manage to download the article, contact the instructor for getting a copy.

   Grading: 1 point per relevant conclusion.