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 December 17, 2010**, 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 two biallelic loci in a population of diploid chromosomes (i.e., population of haplotype pairs), the first locus with alleles A and a, and the second locus with alleles B and b. Let p_{AB} be the population frequency of the haplotype AB. Let $p_{A/B}$ be the population frequency of the "cross-haplotype" A/B (that is, A occuring in one chromosome and B in the other chromosome of the diploid). Show that

$$p_{AB} + p_{A/B} = 2p_{AABB} + p_{AABb} + p_{AaBB} + \frac{1}{2}p_{AaBb},$$

   where, in general, p_{uvxy} is the population frequency of haplotype pairs that have exactly the four alleles u, v, x, and y, that is, p_{AABb} is the total population frequency of the pairs (AB, Ab) and (Ab, AB).

   Grading: reasonable partial solutions with correct arguments yield 1–2 points; correct proof yields 3 points.

2. Suppose the haplotype frequencies p_h for $h \in H_0 = \{0,1\}^m$ and the genotyping error models $\tau_s(g_s; g'_s)$ for $g_s, g'_s \in \{0,1,2\}$, $s = 1,2,\ldots,m$, are fixed and given. Assume HWE (as usual). Describe a method that given a trio genotype $\mathbf{g} = (g^{\mathrm{m}}, g^{\mathrm{f}}, g^{\mathrm{c}})$ outputs a most probable haplotype pair for each member of the trio; for instance, for the child, it outputs a pair $(i,j) \in H_0 \times H_0$ such that

$$\Pr(\text{the child's haplotype pair is } (i,j) | \mathbf{g})$$

   is maximized; analogously for the mother and the father. Analyze the time complexity of your method.

   Grading: reasonable and correct math expressions yield +1 point; correct analysis of the time complexity yields +1 point; correct and sufficient arguments for the time bound $O(m4^m)$ yield +1 point.

3. Present a pseudo code for an algorithm that for a given list of n haplotypes over m SNPs with no missing data, outputs a tagging set according to the Patil et al. definition with $\beta = 0.80$. 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. When the number of (biallelic) markers m is large, the number of potential haplotypes, 2^m, may prohibit the application of the presented techniques for haplotype inference. There is a popular heuristic approach, called *partition ligation*, to address this issue. The idea is to infer the possible candidate haplotypes separately for the first and the last $m/2$ markers; usually the number of haplotypes with nonzero frequency estimates is much below $2^{m/2}$ for each of the two parts, say ℓ and r for the first and the last part, respectively. Then the haplotype frequencies over the complete set of m markers is inferred under the supposition that non-zero frequencies are hold only by haplotypes that belong to the $k := \ell r \ll 2^m$ possible combinations of the already inferred shorter haplotypes.

   Consider unrelated genotype and the model extended by a deletion haplotype. Show that each iteration in the EM algorithm for estimating the frequencies of the k possible haplotypes can be computed in time $O(k(\ell+r)m)$ (per observed unrelated genotype).

   Grading: sufficient argumentation for the time bound $O(k^2m)$ yields 1 point; sufficient argumentation for the bound $O(k(\ell + r)m)$ yields 3 points.

5. Read the article

   - Stefansson H et al., A common inversion under selection in Europeans. Nature Genetics 37, 129 - 137 (2005)

   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) Which one of the two haplotype lineages H1 and H2 is more homogeneous and which one more diverse?

   (b) What observations or experiments speak for non-neutrality (i.e., positive or negative selection) of the inversion polymorphism? (Use at most 100 words in your answer.)

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

6. Read the article

   - Sindi SS, Raphael BJ, Identification and frequency estimation of inversion polymorphisms from haplotype data. J Comput Biol 17, 517 531 (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) What is the key difference or advancement of the proposed method as compared to an earlier method by Bansal et al. (2007) cited in the article?

   (b) How good is the method at estimating the population frequency of an inversion? (Describe strengths and weaknesses.)

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