Universität Bielefeld
Technische Fakultät

AG Genominformatik
Prof. Dr. Jens Stoye
Michel T. Henrichs

Algorithms in Genome Research
Winter 2020/2021

Exercises

Number 10, Discussion: 2021 January 29

1. Give general formulas for the following questions. If this is difficult, enumerate the solutions for small examples.

   (a) For n biallelic sites, represented by the columns of a $k \times n$ binary haplotype matrix, how many *different* haplotype sequences (rows) are at most possible?

   (b) If the haplotypes come in blocks of 10 sites each, how does this decrease the number of *different* haplotype vectors?

   (c) For l founder sequences and m recombination hot spots, how many haplotype vectors are possible (under the assumption that recombinations only occur at hot spots)?

   (d) For k haplotype sequences, what is the maximum number $n_{\max}(k)$ of *different* configurations at segregating sites such that the four-gametes test does not fail?

2. Molecular haplotyping modeled by the (weighted) minimum error correction problem focuses on assembling SNP haplotypes from reads of a sequenced genome. To fully characterize an individual genome, however, haplotyping must produce exhaustive lists of both SNPs and non-SNPs, that is, larger variants. Discuss any ideas how non–SNP variants could be integrated in the analysis.

3. In class we have discussed the Haplotype Assembly Problem and its solution by an Integer Linear Program (ILP). Details and some more discussion can be found in the textbook "Integer Linear Programming in Computational and Systems Biology" by Dan Gusfield (2019), Section 20.2. There you find also the following two exercises (Exercise 20.2.5):

   (a) Explain how to modify the ILP formulation to the haplotype assembly problem, when a column could contain more than two characters.

   (b) What changes are needed in the ILP if the reads can be of unequal length?