GENOME-SCALE ALGORITHM DESIGN
by Veli Mäkinen, Djamal Belazzougui, Fabio Cunial and Alexandru I. Tomescu
Cambridge University Press, 2015
www.genome-scale.info

Exercises for Chapter 14. Genomics

14.1 Calculate the probability of an SNP given a read pileup taking into account the measurement errors.

14.2 *Variant annotation* is the process of assigning a function for each detected variant. For example, a 1-base deletion inside an exon creates a frame-shift and may cause an abnormal protein product to be translated. Consider different variants that might appear and think about why some variants can be called *silent mutations*. Browse the literature to find out what *nonsense mutations*, *missense mutations*, and *regulatory variants* are.

14.3 A greedy algorithm to solve Problem 14.1 is to start with empty E', choose $h \in H$ with most incoming edges not in E', add those unmatched edges to E', and iterate the process until all $r \in R$ are incident to an edge in E'. Show that this can produce a solution whose size is $O(|H|)$ times the optimal solution size.

14.4 Consider the following algorithm for the minimum-cost set cover problem. Start with an empty C and iteratively add to C the most cost-effective set H_i, that is, the set H_i maximizing the ratio between $c(H_i)$ and the number of un-covered elements of $\{r_1, \ldots, r_m\}$ that H_i covers. This is repeated until all elements are covered by some set in C. Show that this algorithm always produces a solution whose cost is $O(\log m)$ times the cost of an optimal solution.

14.5 Show that Problem 14.3 can be reduced to the particular shortest-path problem on DAGs from Exercise 4.16.

14.6 Under what assumptions does the shortest-detour algorithm of Section 6.1.2 applied to aligning haploid prediction to labeled DAG of diploid ground-truth give $O(dn)$ running time, where d is the resulting edit distance and n is the maximum of the DAG size and haploid length?

14.7 Complete the proof of Theorem 14.3, by proving that if A and B are two haplotypes obtained with $|V||E| - 2t$ bit flips from M, then the set $C = \{i \mid A[2i-1] = A[2i] = 0\} \subseteq V$ has the property that there are t edges between C and $V \setminus C$. Obtain this proof by showing the following intermediary properties:

- $A[2i-1] = A[2i]$, for all $i \in \{1, \ldots, |V|\}$;
- $B[2i-1] = B[2i]$, for all $i \in \{1, \ldots, |V|\}$;
- A is the bitwise complement of B.

14.8 Show that the algorithm given for the haplotype assembly under minimum error correction problem can be optimized to run in $O((2^K + r)s)$ time. How could you get the time complexity down to $O(2^K s)$?

14.9 Show how the algorithm given for the haplotype assembly under minimum error correction problem can be implemented to use $O(2^K\sqrt{s})$ memory, while maintaining the same asymptotic time complexity. *Hint.* Use the same idea as in Exercise 7.2.

14.10 Let T denote the maximum number of bit flips needed in any column of the matrix M in an optimal solution to Problem 14.4. Show that the algorithm given for this problem can be modified to run in time $O((K^T + r)s)$, and even in time $O(K^T s)$ (see also Exercise 14.8 above).

14.11 Consider the problem of haplotype assembly under minimum error correction problem in which each cell of M has also a cost of flipping, and one is looking for a partition of the rows of M that is compatible with two haplotypes and minimizes the sum of the costs of all bit flips. What changes to the algorithm we presented are needed in order to solve this more general problem with the same time complexity?