GENOME-SCALE ALGORITHM DESIGN
by Veli Mäkinen, Djamal Belazzougui, Fabio Cunial and Alexandru I. Tomescu
Cambridge University Press, 2nd edition, 2023
www.genome-scale.info

Exercises for Chapter 16. Transcriptomics

16.1 Work out the calculations for deriving the system of equations (16.3), using the formula

$$\begin{aligned}\frac{\partial\left(w_i - \alpha_{i,1}e_1 - \alpha_{i,2}e_2 - \cdots - \alpha_{i,k}e_k\right)^2}{\partial e_j} =& 2\left(w_i - \alpha_{i,1}e_1 - \alpha_{i,2}e_2 - \cdots - \alpha_{i,k}e_k\right) \cdot \\ & \cdot \frac{\partial\left(w_i - \alpha_{i,1}e_1 - \alpha_{i,2}e_2 - \cdots - \alpha_{i,k}e_k\right)}{\partial e_j} \\ =& -2\alpha_{i,j}\left(w_i - \alpha_{i,1}e_1 - \alpha_{i,2}e_2 - \cdots - \alpha_{i,k}e_k\right).\end{aligned}$$

16.2 Formalize the exon detection phase described using HMMs (recall Insight 14.1 and Chapter 7).

16.3 Consider the following version of Problem 16.3 in which there are no costs associated with the arcs for the DAG and in which we need to cover only a given collection of paths. Given a DAG $G = (V, E)$, a set $S \subseteq V$ of possible start vertices, a set $T \subseteq V$ of possible end vertices, and a collection $\mathcal{P}^{in} = \{P_1^{in}, \ldots, P_t^{in}\}$ of directed paths in G, find the minimum number k of paths $P_1^{sol}, \ldots, P_k^{sol}$ such that

- every P_i^{sol} starts in some vertex of S and ends in some vertex of T,
- every path $P^{in} \in \mathcal{P}^{in}$ is entirely contained in some P_i^{sol}.

Assume also that no P_i^{in} is entirely included in another P_j^{in} (otherwise P_i^{in} can be removed from $\mathcal{P}^{in}$). Show that this problem can be solved by a reduction to a minimum-cost flow problem (without having to iteratively merge paths with longest suffix-prefix overlaps). What features does the resulting flow network have? Can you apply a specialized minimum-cost flow algorithm with a better complexity for such a network?

16.4 Given an input for the problem of transcript assembly with paired-end reads, show that we can decide in polynomial time whether it admits a solution with two paths, and if so, find the two solution paths.

16.5 Using the reduction in the proof of Theorem 16.4, conclude that there exists for no $\varepsilon > 0$ an algorithm returning a solution with k paths for the problem of transcript assembly with paired-end reads, where k is greater than the optimal number of paths by a multiplicative factor $\frac{4}{3} - \varepsilon$.

16.6 Argue that the reduction of Problem 16.5 to a network flow problem with convex costs is correct.

16.7 Show that the function g defined in (16.7) is a flow on G, namely that

- for all $y \in V(G) \setminus (S \cup T)$, $\sum_{x \in N^-(y)} g(x,y) = \sum_{x \in N^+(y)} g(y,x)$;
- $\sum_{s \in S} \sum_{y \in N^+(s)} g(s,y) = \sum_{t \in T} \sum_{x \in N^-(t)} g(x,t)$.

16.8 Explain what changes need to be made to the flow network N, if in Problem 16.6 we also get in the input a coefficient $\alpha(x,y)$ for every arc (x,y) of G, and we are asked to find the paths and their expression levels which minimize

$$\sum_{(x,y) \in E} \alpha(x,y) \left| w(x,y) - \sum_{j \,:\, (x,y) \in P_j} e_j \right|.$$

16.9 Consider a variant of Problem 16.6 in which we want to minimize the absolute differences between the total coverage (instead of the average coverage) of an exon and its predicted coverage. Explain how this problem can be reduced to the one in Exercise 16.8 above, for an appropriately chosen function α.

16.10 Adapt the reduction to a minimum-cost flow problem from Section 16.3 to solve the following problem. Given a DAG $G = (V,E)$, a set $S \subseteq V$ of possible start vertices, a set $T \subseteq V$ of possible end vertices, and a weight function $w : E \to \mathbb{Q}_+$, find a collection of paths $P_1, \ldots, P_k$ in G, and their corresponding expression levels $e_1, \ldots, e_k$, such that

- every P_i starts in some vertex of S and ends in some vertex of T,

and the following function is minimized:

$$\sum_{(x,y) \in E} \left| w(x,y) - \sum_{j \,:\, (x,y) \in P_j} e_j \right| + \lambda \sum_{i=1}^{k} e_i.$$

What can you say about the problem in which we use the squared difference, instead of the absolute value of the difference?

16.11 Show that the following problem is NP-hard, for any fixed $\beta \geq 1$. Given a DAG $G = (V,E)$, a set $S \subseteq V$ of possible start vertices, a set $T \subseteq V$ of possible end vertices, a weight function $w : E \to \mathbb{Q}_+$, and an integer k, find a collection of paths $P_1, \ldots, P_k$ in G, and their corresponding expression levels $e_1, \ldots, e_k$, such that

- every P_i starts in some vertex of S and ends in some vertex of T,

and the following function is minimized:

$$\sum_{(x,y) \in E} \left| w(x,y) - \sum_{j \,:\, (x,y) \in P_j} e_j \right|^{\beta}.$$

Hint. Use Theorem 5.3 or its proof.

16.12 Show that, given $k \geq 1$ and given $e_1, \ldots, e_k$, the following problem is solvable in time $O(n^{2k+2})$. Given a DAG $G = (V,E)$, a set $S \subseteq V$ of possible start vertices, a set $T \subseteq V$ of possible end vertices, weight function $w : E \to \mathbb{Q}_+$, and an integer k, find a collection of paths $P_1, \ldots, P_k$ in G, such that:

- every P_i starts in some vertex of S, and ends in some vertex of T,

and the following function is minimized

$$\sum_{(x,y)\in E} \left| w(x,y) - \sum_{j\,:\,(x,y)\in P_j} e_j \right|. \tag{1}$$

Hint. Use dynamic programming and consider the optimal k paths ending in every tuple of k vertices.

16.13 Show that, given an error function $\delta : \mathbb{Q}_+ \times \mathbb{Q}_+ \to \mathbb{Q}_+$, the problem considered in Exercise 16.12 in which the objective function the above equation (1) is replaced with

$$\sum_{(x,y)\in E} \delta\left(w(x,y), \sum_{j\,:\,(x,y)\in P_j} e_j \right) \tag{2}$$

is solvable with the same time complexity $O(n^{2k+2})$ (assuming δ is computable in time $O(1)$).

16.14 Consider the following problem of *transcript assembly and expression estimation with outliers*, in which the weights of the arcs not belonging to any solution path do not contribute in the objective function. Given $k \geq 1$, a DAG $G = (V, E)$, a set $S \subseteq V$ of possible start vertices, a set $T \subseteq V$ of possible end vertices, a weight function $w : E \to \mathbb{Q}_+$, and an integer k, find a collection of paths $P_1, \ldots, P_k$ in G, such that

- every P_i starts in some vertex of S and ends in some vertex of T,

and the following function is minimized:

$$\sum_{(x,y)\text{: exists } P_i \text{ with } (x,y)\in P_i} \left| w(x,y) - \sum_{j\,:\,(x,y)\in P_j} e_j \right|.$$

Show that this problem is NP-hard. *Hint.* Use Theorem 5.3 or its proof.

16.15 Reduce the problem of transcript assembly and expression estimation with outliers from Exercise 16.14 to the generalized problem from Exercise 16.13, by showing an appropriate error function δ.

16.16 Prove that the co-linear chaining algorithm works correctly even when there are tuples containing other tuples in T or in R, that is, tuples of type (x, y, c, d) and (x', y', c', d') such that either $x < x' \leq y' < y$ or $c < c' \leq d' < d$ (or both).

16.17 Modify the co-linear chaining algorithm to solve the following variations of the ordered coverage problem.

a) Find the maximum ordered coverage of R such that all the tuples involved in the coverage must overlap in R.

b) Find the maximum ordered coverage of R such that the distance in R between two consecutive tuples involved in the coverage is at most a given threshold value α.

c) Find the maximum ordered coverage of R such that the distance in T between two consecutive tuples involved in the coverage is at most a given threshold value β.

16.18 Co-linear chaining gives a rough alignment between the transcript and the genome as a list of subregion correspondences. Consider how this rough alignment can be fine-grained into a sequence level alignment.