Algorithm Design - Homework 2, Exercise 4

Francesco Carpineti 1683418 - Federico Gioia 1702089 - Giuliano Abruzzo 1712313

January 18, 2019

1 Exercise 4

This is a minimum string cover problem. We can see the genes set G as a directed graph, where factorizations[1] of a string are modeled by paths in it. In particular, edges correspond to sequence intervals and nodes to positions between characters. So, as we know we are dealing with a single string s (the DNA sequence D), we can express edges and nodes as follows:

- $V_D = \{(D,0),(D,1),...,(D,n)\}$;
- $E_D = \{(D,p) \to (D,q)$: $p < q$, $(D,p) \in V_D$,$(D,q) \in V_D$, $D_{[p...q]} \in G\}$;
- $\forall g \in G$, $R(g) = \{(D,i,j) : D[i...j] = g\}$.

Now, let's define some variables useful for an ILP formulation of the problem:

- $z_{(D,i,j)}$ binary variable for each $(D,i,j) \in E_D$, where $(D,i,j) = (D,i) \to (D,j+1)$;
- w_g cost of a gene $g \in G$;
- x_g binary variable which is 1 if that $g \in G$ is present in the solution, 0 otherwise;
- $\delta^-(v)$ and $\delta^+(v)$ ingoing and outgoing edges of a node $v \in V_D$.

Our formulation is useful to find the less expensive path from $(D,0)$ to $(D,|D|)$, which is also the problem's goal, obviously. Let's see the Primal problem:

$$
\begin{aligned}
& min \sum_{g \in G} w_g x_g, \quad s.t. \\
& z_{(D,i,j)} \leq x_g && \forall g \in G, \forall (D,i,j) \in R(g) \\
\sum_{(D,i,j) \in \delta^+((D,0))} & z_{(D,i,j)} = 1 \\
\sum_{(D,i,j) \in \delta^-(v)} & z_{(D,i,j)} = \sum_{(D,i,j) \in \delta^+(v)} z_{(D,i,j)} && \forall v \in V_D \setminus (D,0),(D,n) \\
& x_g, z_{(D,i,j)} \in \{0,1\} && \forall g \in G, (D,i,j) \in E_D
\end{aligned}
$$

The first constraint is needed to guarantee that if there are several z variables referred to the same substring (so, a substring appears more than one time in a word), then the cost associated to this gene is paid only one time, also if this gene is reused to build the final string. The second constraint is referred only to gene that start the path to reach the completness of the full string, in fact this constraint says that for starting node $((D,0))$ there will be only one outgoing edge from it. The third, last constraint is linked to all nodes that are not the first or the last one in the path, and it simply says that, for all these nodes of the graph, the number of outgoing edges have to be equal to the number of ingoing edges.

In order to do a good trasformation from primal to dual, first of all we need to rebuild the primal system in order to leave in the right part of all restricts only their known terms and also we

[1] To understand what a string factorization is, see `http://ls2-www.cs.tu-dortmund.de/grav/grav_files/people/schwiegelshohn/string-cover.pdf`, page 76, section 1.2.

need to relax the problem, removing the integer conditions and adding only the non-negativity of the variables. W.r.t those condictions we reach this form of the Primal problem:

$$
\begin{aligned}
min \sum_{g \in G} w_g x_g, \quad s.t. & \\
-z_{(D,i,j)} + x_g \geq 0 & \qquad \forall g \in G, \forall (D,i,j) \in R(g) \\
\sum_{(D,i,j) \in \delta^+((D,0))} z_{(D,i,j)} = 1 & \\
\sum_{(D,i,j) \in \delta^-(v)} z_{(D,i,j)} - \sum_{(D,i,j) \in \delta^+(v)} z_{(D,i,j)} = 0 & \qquad \forall v \in V_D \setminus (D,0), (D,n) \\
x_g, z_{(D,i,j)} \geq 0 & \qquad \forall g \in G, (D,i,j) \in E_D
\end{aligned}
$$

Before doing the dual of it, we need to focus on the coefficents matrices for every single constraint of the primal problem focusing on their dimensions and their content in terms of numbers:

- I is a identity matrix $|E|$ x $|E|$ with only -1 on the diagonal;
- F is a $|E|$ x $|G|$ matrix where there is 1 if $\forall g \in G, (D,i,j) \in R(g)$, else 0;
- H is a 1 x $|E|$ matrix where there is 1 if $(D,i,j) \in \delta^+((D,0))$, else 0;
- L is a $(|D|-1)$ x $|E|$ matrix where there is 1 if $(D,i,j) \in \delta^-(v) \setminus ((D,0))$, -1 if $(D,i,j) \in \delta^+(v) \setminus ((D,0))$, else 0.

Let's see the new aspect of primal problem w.r.t these 4 matrices:

$$
\begin{aligned}
min \sum_{g \in G} w_g x_g, \quad s.t. & \\
Iz + Fx \geq 0 & \\
Hz = 1 & \\
Lz = 0 & \\
x_g, z_{(D,i,j)} \geq 0 & \qquad \forall g \in G, (D,i,j) \in E_D
\end{aligned}
$$

The dual of this primal problem is:

$$
\begin{aligned}
max \text{ b} \quad s.t. & \\
I^T u^{'} + H^T u^{''} + L^T u^{'''} \leq 0 & \\
F^T u^{'} \leq w_g & \qquad \forall g \in G \\
u_g^{'} \geq 0 & \qquad \forall g \in G
\end{aligned}
$$

b is the vector of known terms with dimension $(|E|+|D|)$ x 1 and $u^{'}, u^{''}, u^{'''}$ are the new variables of the dual problem. Bringing back the original matrices's content we reach the final version of the

dual problem:

$$
\begin{aligned}
&max\ \mathrm{b} \quad s.t. \\
&-u'_{i,j} + u'' - \sum_{(D,i,j)\in\delta^-(v)} u'''_v \leq 0 && \forall (D,i,j) \in \delta^-((D,0)) \\
&-u'_{i,j} + \sum_{(D,i,j)\in\delta^+(v)} u'''_v - \sum_{(D,i,j)\in\delta^-(v)} u'''_v \leq 0 && \forall (D,i,j) \notin \delta^-((D,0)) \\
&u'_g \leq w_g && \forall g \in G \\
&u'_g \geq 0 && \forall g \in G
\end{aligned}
$$