Context-Free de Novo Genome Assembly Problem

Max A. Alekseyev

April 8, 2011

1 Preliminary definitions

A weighted directed graph $G = (V, A, \ell)$ consists of the set of vertices V, the set of (directed) arcs $A \subset V \times V$, and a function $\ell : A \to \mathbb{N}$ specifying the length (weight) of each arc. We extended function ℓ to measure the length of any path/cycle in G as the total lengths of all arcs in it.

Definition 1 *We define the (directed)* distance $d_G(u, v)$ *between two vertices* $u, v \in V$ *as the shortest length of a path (i.e., the sum of lengths of all arcs in such path) starting at* u *and ending at* v*. If there is no such path, we let* $d_G(u, v) = +\infty$*.*

Furthermore, for any two arcs $a = (u_1, u_2), b = (v_1, v_2)$*, we define the distance* $d_G(a, b) = \ell(u_1, u_2) + d_G(u_2, v_1) + \ell(v_1, v_2)$ *(i.e., the length of a shortest path starting at* a *and ending at* b*).*

Similarly, we define the distances between an arc $a = (u_1, u_2)$ *and a vertex* v*:* $d_G(a, v) = \ell(u_1, u_2) + d_G(u_2, v)$ *and* $d_G(v, a) = d_G(v, u_1) + \ell(u_1, u_2)$*.*

For the sake of simplicity, we will not distinguish a path/cycle in G from the set of vertices/arcs that it visits. Depending on the context, we will also treat a path/cycle as a linear/cyclic sequence of *vertex/arc instances*[1] visited by the path/cycle in order.

Definition 2 *For a path/cycle* X *and any two vertex/arc instances* $x, y \in X$*, we define*

- *the subpath* $\mathrm{Subpath}_X(x, y)$ *of* X*, starting at* x *and ending at* y*; if there is no such subpath,* $\mathrm{Subpath}_X(x, y) = \emptyset$*;*
- *the (directed)* X*-distance* $d_X(x, y)$ *as the length of* $\mathrm{Subpath}_X(x, y)$*; if there is no such subpath,* $d_X(x, y) = +\infty$*.*

Definition 3 *A cycle* C_2 *is called* (β, γ)-homeomorphic *to a cycle* C_1 *if for every arc* $a \in (C_2 \setminus C_1)$*, there exist two vertex instances* $u_1, u_2 \in (C_1 \cap C_2)$[2] *such that* $a \in \mathrm{Subpath}_{C_2}(u_1, u_2)$ *and*

$$\max\{d_{C_1}(u_1, u_2), d_{C_2}(u_1, u_2)\} \leq \beta,$$

$$|d_{C_1}(u_1, u_2) - d_{C_2}(u_1, u_2)| \leq \gamma.$$

[1] Notice that a vertex/arc may appear multiple times in a path/cycle. Hence, here we distinguish vertices/arcs and their instances.

[2] This is intermix of vertices and their instances. Will fix later.

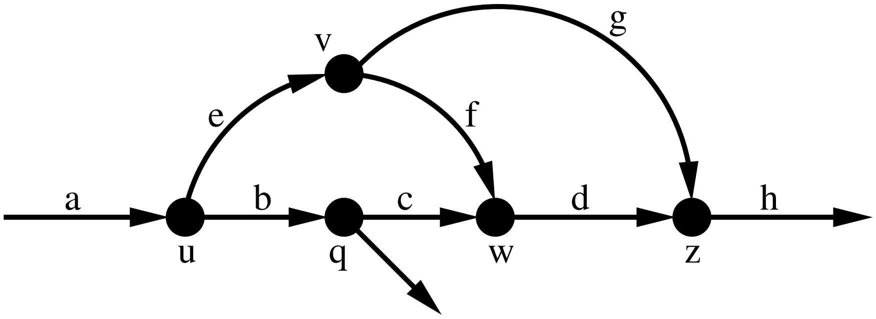

Figure 1: Example of parallel paths and bulges.

One can show[3] that (β, γ)-homeomorphism is a symmetric relationship (i.e., if C_2 is (β, γ)-homeomorphic to C_1, then C_1 is (β, γ)-homeomorphic to C_2).

2 Problem formulation

Input

- a weighted directed graph $G = (V, A, \ell)$;
- a subset of pairs of arcs $Q \subset A \times A$;
- a function $q : Q \to \mathbb{N}$ (prescribing for some pairs of arcs approximate distances between them);
- a parameter $\delta \geq 0$ (of allowed variation in distances);
- a parameter $\beta, \gamma \geq 0$ (defining parallel paths).

Output

A shortest cycle C in G such that there exists a set $\mathcal{C} \ni C$ of pairwise (β, γ)-homeomorphic cycles that traverse each arc in G at least once and for any pair of arcs $(a, b) \in Q$, there exists $C' \in \mathcal{C}$ with a subpath $\mathrm{Subpath}_{C'}(a, b)$ (for some instances of a and b) of length in the interval $d \pm \delta$.

3 Explicit description of the solution

In this section, we (try to) describe the properties of a shortest cycle C for which there exists the required set $\mathcal{C}$.

The first property guarantees that every arc not visited by C is visited by some other cycle (β, γ)-homeomorphic to C:

[3] Will prove later.

Property 1 *For every arc $a = (v_1, v_2) \notin C$, there exist two vertex instances $u_1, u_2 \in C$ such that*

$$\max\{d_C(u_1, u_2), d_G(u_1, v_1) + \ell(v_1, v_2) + d_G(v_2, u_2)\} \leq \beta$$

and

$$|d_G(u_1, v_1) + \ell(v_1, v_2) + d_G(v_2, u_2) - d_C(u_1, u_2)| \leq \gamma.$$

For example, in Fig. 1 if cycle C passes through arcs a, b, c, d, h but not e, then for the arc $a = e$ candidate vertices satisfying Property 1 would be $u_1 = u$ and $u_2 = w$.

Denote by U_1 and U_2 the mappings that map an arc $a \notin C$ correspondingly to vertex instances $u_1, u_2 \in C$ (as defined in Property 1).[4]

The second property ensures that the cycle C cannot be shortened by rerouting of some of its subpath:

Property 2 *There exist no two distinct subpaths $P_1, P_2 \subset C$ both starting and ending at the same two vertices such that*

$$\max\{\ell(P_1), \ell(P_2)\} \leq \beta$$

and

$$|\ell(P_1) - \ell(P_2)| \leq \gamma.$$

For example in Fig. 1, if cycle C passes through arcs b, c as well as through arcs e, f, that would violate Property 2 unless $\max\{\ell(b)+\ell(c), \ell(e)+\ell(f)\} > \beta$ or $|\ell(b)+\ell(c)-\ell(e)-\ell(f)| > \gamma$.

The third property enforces obeying by C the prescribed distances between pairs of arcs in Q:

Property 3 *For any pair of arcs $(a, b) \in Q$,*

- *if $a, b \in C$, then there exists their instances in C with $|d_C(a, b) - q(a, b)| \leq \delta$;*
- *if $a \in C$ and $b \notin C$, then there exists an instance of a in C with $|d_C(a, U_1(b)) + d_G(U_1(b), b) - q(a, b)| \leq \delta$;*
- *if $a \notin C$ and $b \in C$, then there exists an instance of b in C with $|d_G(a, U_2(a)) + d_C(U_2(a), b) - q(a, b)| \leq \delta$;*
- *if $a, b \notin C$, then*

$$|d_G(a, U_2(a)) + d_C(U_2(a), U_1(b)) + d_G(U_1(b), b) - q(v_1, v_2)| \leq \delta.$$

[4] If there are multiple such pairs u_1, u_2 exist, we let U_1, U_2 map to an arbitrary chosen such pair.

4 Graph contraction

A graph G can be contracted in the form where there is no vertex of indegree and outdegree both equal 1.

If there is such a vertex v with a single incoming arc (u,v) and a single outgoing arc (v,w), we remove v by introducing an arc (u,w) of length $\ell(u,v)+\ell(v,w)$. Furthermore, we propagate affected distance estimates to the new edge (u,w) as follows (t is any arc):

- $q((u,w),t) \leftarrow q((u,v),t)$;
- $q(t,(u,w)) \leftarrow q(t,(u,v)) + \ell(v,w)$;
- $q((u,w),t) \leftarrow q((v,w),t) + \ell(u,v)$;
- $q(t,(u,w)) \leftarrow q(t,(v,w))$.

In reality, there may be multiple estimates for the distance between the same pair of arcs. We will keep their sum and number that will eventually allow us to compute the averaged estimate.

5 Bulges removal

A path (and, in particular, a single arc) in the graph G is called *short* if its length does not exceed β. Two paths in the graph G are called *parallel* if they start and end at the same two vertices but otherwise share no intermediate vertex.

A *bulge* in the graph G is a pair of parallel short paths, whose lengths differ by at most γ.

We notice that the solution cycle cannot contain any bulges, suggesting that they may be removed before actual finding the cycle.

Assuming that the graph G is contracted, we define a *simple bulge* as a bulge where one of the paths consists of a single arc.

A simple bulge $\{(u,v),P\}$, where (u,v) is an arc and $P=\langle a_1,a_2,\dots,a_k\rangle$ is a parallel path represented as a sequence of arcs, can be removed as follows:

- Remove the arc (u,v) from G, propagating $q((u,v),t)$ and $q(t,(u,v))$ to each arc of P. Namely, for $j=1,2,\dots,k$ we propagate

  $$q(t,a_j) \leftarrow q(t,(u,v)) - \ell(u,v) + \ell(a_1) + \ell(a_2) + \cdots + \ell(a_j) + \frac{j\cdot\Delta}{k}$$

  and

  $$q(a_j,t) \leftarrow q((u,v),t) - \ell(u,v) + \ell(a_k) + \ell(a_{k-1}) + \cdots + \ell(a_j) + \frac{(k+1-j)\cdot\Delta}{k},$$

  where $\Delta = \ell(u,v) - \ell(P)$.
- Contract G as described in previous section.

We remove simple bulges iteratively, noticing that removal one bulge may result in appearance of a new simple bulge. For example, in Fig. 1 removing the arc g from a simple bulge $\{g,\langle e,f\rangle\}$ allows contraction of arcs e,f into a single arc forming a new simple bulge.