Gapped alignment graph

In order to model gaps, we extend the set of arcs. It consist now of $A = A_g \cup A_p$ which represent positional constraints. Arcs in A_p represent as before *consecutivity* of characters within a same string and run from each node to its "right" neighbor, i. e., $A_p = \{(v^i_j, v^i_{j+1}) : 1 \leq i \leq k, 1 \leq j <\mid a^i \mid\}$.

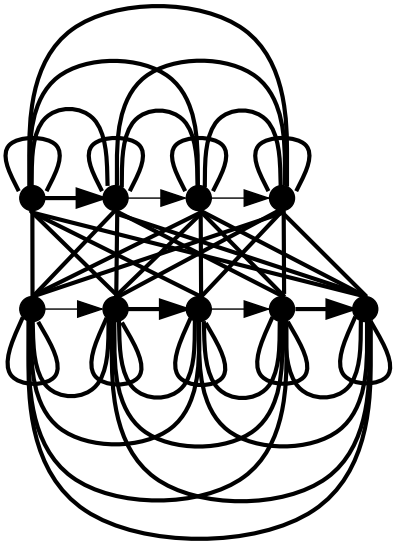
The arcs in A_g represent *gaps* in the alignment. Each substring of a string a^i can be aligned with gap characters in any other string a^j, or to put it differently, it may be the case that no character in this substring of a^i is aligned with any character in a^j.

Gapped alignment graph (2)

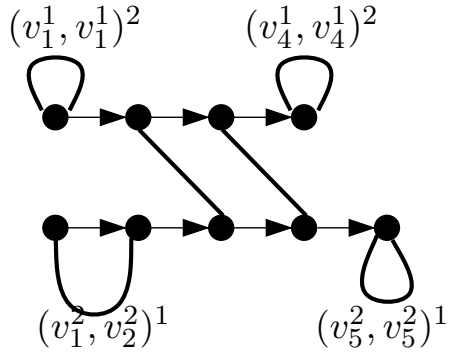
Hence we introduce for each substring of a^i from a^i_l to a^i_m and for each $1 \leq j \leq k, j \neq i$, an arc from v^i_l to v^i_m, denoted by $(v^i_l, v^i_m)^j$. In other words, there are $k-1$ arcs in A_g from v^i_l to v^i_m.

We again say that a gap arc $(v^i_l, v^i_m)^j$ is *realized* by an alignment if the substring in a^i from position l to position m is not aligned to any character in a^j, whereas both a^i_{l-1} (if $l > 1$) and a^i_{m+1} (if $m <\mid a^i \mid$) are aligned with some letter in a^j. We also say that the nodes corresponding to the substring of a^i are *spanned* by the gap arc.

Gapped alignment graph (3)



C--GT-U
-AGGTC-



The above picture shows a gapped alignment graph for two sequences and a gapped trace, realizing two alignment edges and four gap edges.

We already know what combinations of alignment edges are allowed. Now we need to impose constraints on the gap arcs.

Gapped alignment graph (4)

Let $A^{i,j} \subseteq A_g$ denote the set of all gap arcs for substrings in a^i aligned with gap characters in a^j.

Also, given two gap arcs $(v^i_l, v^i_m)^j, (v^i_p, v^i_q)^j \in A^{i,j}$, we say that they *conflict* if the substrings spanned by the arcs overlap or even touch, that is if $[l, m+1] \cap [p, q] \neq \emptyset$. There must be at least one aligned character between consecutive gap arcs. Let I the collection of all maximal sets of conflicting gap arcs.

Finally, we let

$$A^{i,j}(l \leftrightarrow m) := \{(v^i_p, v^i_q)^j : p \leq l, q \geq m\}$$

denote the set of arcs in $A^{i,j}$ spanning $v^i_l, \ldots, v^i_m$.

Gapped alignment graph (5)

In order to score the alignment, each of the edges in E and gap arcs in A_g is assigned a *weight* that corresponds to the benefit (or cost) of realizing the edge or arc. We let w_e and w_a denote respectively the weight of edge $e \in E$ and arc $a \in A_g$.

Note that arcs A_p are independent of the alignment, which specifies which edges among E and arcs among A_g are realized.

Gapped alignment graph (6)

A subgraph of the gapped alignment graph is called *gapped trace* if it corresponds to a gapped alignment. A gapped trace has to fulfill the following conditions:

1. For each pair of strings, each node is either incident to exactly one alignment edge or spanned by exactly one gap arc.
2. There must not be a critical mixed cycle in the subgraph.
3. There cannot be a pair of conflicting gap arcs for a given pair of strings.
4. Whenever we realize two edges incident with the same node, say $\{v_{l_1}^{i_1}, v_{l_2}^{i_2}\}$ and $\{v_{l_1}^{i_1}, v_{l_3}^{i_3}\}$, by transitivity we must also realize edge $\{v_{l_2}^{i_2}, v_{l_3}^{i_3}\}$.

The problem is now, given a gapped alignment graph, to find the best subgraph that fulfills all of the above conditions.

The ILP for the Gapped trace problem

We have two types of variables:

- For every edge $e = \{v_l^i, v_m^j\} \in E^{i,j}$, we define a binary variable x_e (we also write $x_{\{v_l^i, v_m^j\}}$), which indicates whether a_l^i is aligned with a_m^j or not. We call these variables the *alignment variables*.
- For every arc $a = (v_l^i, v_m^i)^j \in A^{i,j}$, representing a gap in string a^j aligned to the substring $a_{l \leftrightarrow m}^i$ of a^i, we define a binary variable y_a (we also write $y_{(v_l^i, v_m^i)^j}$). We call these variables the *gap variables*.

  For a cycle C, we denote the set of edges with C_E and the set of arcs with C_A.

The ILP for the Gapped trace problem (2)

Then the ILP formulation is:

$$\max \sum_{e \in E} w_e \cdot x_e + \sum_{a \in A_g} w_a \cdot y_a \tag{1}$$

subject to

$$\sum_{1 \le m \le |a^j|} x_{\{v_l^i, v_m^j\}} + \sum_{a \in A^{i,j}(l \leftrightarrow l)} y_a = 1, \forall\, 1 \le i, j \le k, i \ne j, 1 \le l \le |\, a^i \,| \tag{2}$$

$$\sum_{e \in C_E} x_e \le |\, C_E \,| - 1, \forall\, C \in \mathcal{C} \tag{3}$$

$$\sum_{a \in I} y_a \le 1, \forall\, I \in \mathcal{I} \tag{4}$$

$$x_{\{v_{l_1}^{i_1}, v_{l_2}^{i_2}\}} + x_{\{v_{l_1}^{i_1}, v_{l_3}^{i_3}\}} - x_{\{v_{l_2}^{i_2}, v_{l_3}^{i_3}\}} \le 1, \forall\, 1 \le i_r \le k, 1 \le l_r \le |\, a^{i_r} \,| \; (r = 1,2,3) \tag{5}$$

$$x_e, y_a \in \{0,1\}, \forall\, e \in E, a \in A_g \tag{6}$$