Wavefront sequence-to-graph alignment

Andrea Guarracino and Erik Garrison

February 6, 2022

Abstract

We explore an extension to the wavefront algorithm that generalizes it to work on sequence graphs. This follows the basic formulation of partial order alignment, wherein the recurrence relations defining alignment are extended to consider the topology of a target graph.

1 Segmented sequence graph

As input we take an alphabet Σ, a query sequence $q = \Sigma^n$, and a target sequence graph $G = (V, E \subseteq (V \times V), \sigma : V \to \Sigma)$, where V is the node set, E is a set of directed edges, and σ is a function that appoints one character to each node as the node label. We denote the in-neighbors of a node as $\delta_v^{in} = \{x : (x, v) \in E\}$, and the out-neighbors as $\delta_v^{out} = \{x : (v, x) \in E\}$. The segments of the graph C are linear unbranching components made of series of nodes, $C = \{c = v_j \dots v_{|j+c|} : \forall_{v_i, v_{i+1} \in c}(v_i, v_{i+1}) \in E \wedge |\delta_{v_i}^{out}| = 1 \wedge |\delta_{v_{i+1}}^{in}| = 1\}$. Segments correspond to compressible series of nodes that could be combined into a single node with a string label without disrupting the language of strings represented by the graph. A graph G with length $m = |V|$ can also be represented as a target sequence $T = \Sigma^m = t_0 \dots t_{m-1}$, where each character corresponds to a node in our graph, $\sigma(v_i) = t_i$.

2 Sequence-to-graph alignment

We define the pairwise global alignment between the query sequence $Q = q_0 \dots q_{n-1}$ and a graph $G \implies T = t_0 \dots t_{m-1}$ as the computation of the path from $(0, 0)$ to (n, m) with minimum score, allowing for matches, mismatches, and gaps, each with costs defined by a different score function. A trivial graph with a single segment $|C| = 1$ is equivalently represented as a sequence T. In this case, we could apply standard pairwise alignment method like Smith-Waterman-Gotoh (SWG) [1, 2] or the WaveFront Algorithm (WFA) [3] to derive an optimal alignment. To formulate sequence-to-graph alignment, we first define algorithms for pairwise sequence alignment, and then show how these are generalized to the case where the target graph is not a single segment.

3 Smith-Waterman-Gotoh

In SWG, we maintain three matrices, H, E, and F.[1] The value of a cell in $H_{i,j}$ represents the score of an alignment ending at the corresponding location in the query q_i and the graph t_j, while E contains scores ending with a gap extending along the target T (or deletions in Q relative to T) and F contains scores ending in a gap extending along the query Q (or insertions in Q relative to T). Our gap-affine penalty scores $p = a, x, o, e$ define the penalty for matching (a) or mismatching (x) any base of the target and query, while the gap-penalty function for a gap of n bp is expressed as $g(n) = o + n \cdot e$. A function $s(i, j)$ scores the bases in our query and target, yielding a if $q_i = t_j$ and x otherwise. Equation 1 shows the recurrence relations in standard SWG.

$$\begin{cases} E_{i,j} = \min\{H_{i,j-1} + o + e, E_{i,j-1} + e\} \\ F_{i,j} = \min\{H_{i-1,j} + o + e, F_{i-1,j} + e\} \\ H_{i,j} = \min\{E_{i,j}, F_{i,j}, H_{i-1,j-1} + s(q_{i-1}, t_{j-1})\} \end{cases} \tag{1}$$

4 Partial order alignment

We now extend these recurrence relations to consider the topology of the graph. Insertions in the query relative to the target only require that we consider scores in the same node of the graph, so the relation defining F remains unchanged. However, both H and E must consider the set of inbound nodes δ_v^{in}. Assuming that we index nodes $v \in V$ in the same order as characters in t, then $\sigma(v_j) = t[j]$ and δ_j^{in} yields the set of indexes of inbound nodes of v_j in t. Equation 2 shows the generalized recurrence relation in partial order alignment [5], or SWG-POA.

$$\begin{cases} E_{i,j} = \min_{\forall u \in \delta_j^{in}}\{H_{i,u} + o + e, E_{i,u} + e\} \\ F_{i,j} = \min\{H_{i-1,j} + o + e, F_{i-1,j} + e\} \\ H_{i,j} = \min\left\{E_{i,j}, F_{i,j}, \min_{\forall u \in \delta_j^{in}}\{H_{i-1,u} + s(q_{i-1}, t_u)\}\right\} \end{cases} \tag{2}$$

5 Wavefront algorithm

WFA reformulates the SWG model to support the incomplete evaluation of the H matrix while maintaining the guarantee that we find the lowest-cost path from $(0, 0)$ to (n, m). Instead of filling the matrix uniformly, such as vertically or horizontally, WFA derives a series of wavefronts, each of which contains the set of cells in H reached from $(0, 0)$ with the same score. Setting the match score $a = 0$ allows these wavefronts to extend in one step through consecutive matches.

For a score s and diagonal $k = j - i$, the furthest-reaching point $\mathcal{R}_{s,k}$ indicates the cell in H that is the furthest from the beginning of diagonal k with score s.

[1] To avoid ambiguity about the polarization of insertions and deletions, we adopt the notation of [4] to formulate SWG and refer to indels specifically as gaps versus the query or target sequence.

$\widetilde{E}_{s,k}$, $\widetilde{F}_{s,k}$, and $\widetilde{H}_{s,k}$ store the offset in the diagonal to furthest-reaching point $\mathcal{R}_{s,k}$ in each of the SWG matrices. For a given score s, the s-wavefront $\mathcal{W}_s$ is the set of offsets $\widetilde{E}_{s,k}$, $\widetilde{F}_{s,k}$, and $\widetilde{H}_{s,k}$ for all k. Considering all k, we can more simply refer to the components of $\mathcal{W}_s$ as $\widetilde{E}_s$, $\widetilde{F}_s$, and $\widetilde{H}_s$. With each component of a wavefront represented as a vector of offsets centered around the main diagonal $k = 0$, we can define the highest (or rightmost) and lowest (or leftmost) diagonals in the component as $\widetilde{H}_s^{hi}$ and $\widetilde{H}_s^{lo}$.

In WFA, our goal is to derive a global alignment by computing the minimum s such that any of the furthest-reaching points in $\mathcal{W}_s$ reaches (n, m). Equation 3 shows the recurrence relations defined in WFA, which reformulate the same scoring model in SWG to consider furthest-reaching points in diagonals.

$$\begin{aligned}
\widetilde{E}_{s,k} &= \max\left\{\begin{array}{ll} \widetilde{H}_{s-o-e,k-1} & \text{(open gap in } Q\text{)} \\ \widetilde{E}_{s-e,k-1} & \text{(extend gap in } Q\text{)} \end{array}\right\} + 1 \\
\widetilde{F}_{s,k} &= \max\left\{\begin{array}{ll} \widetilde{H}_{s-o-e,k+1} & \text{(open gap in } T\text{)} \\ \widetilde{F}_{s-e,k+1} & \text{(extend gap in } T\text{)} \end{array}\right\} \\
\widetilde{H}_{s,k} &= \max\left\{\begin{array}{ll} \widetilde{H}_{s-x,k} + 1 & \text{(substitution)} \\ \widetilde{E}_{s,k} & \text{(gap in } Q\text{)} \\ \widetilde{F}_{s,k} & \text{(gap in } T\text{)} \end{array}\right\}
\end{aligned} \tag{3}$$

WFA thus computes the furthest-reaching points of $\mathcal{W}_s$ using only $\mathcal{W}_{s-o}$, $\mathcal{W}_{s-e}$, and $\mathcal{W}_{s-x}$. Note that, with $a = 0$, considering matches requires us to extend all previously computed points by as far as possible by following matching characters along the diagonal. This extension of the wavefront in the case of matches is essential to the good memory and time characteristics of WFA.

6 Wavefront partial order alignment (WFPOA)

We consider the generalization of WFA to the case where our target is the directed acyclic graph G, which we term WFPOA. In general, this extension is similar to that applied to transform SWG into SWG-POA. But, the definition of diagonals in the alignment matrix is no longer stable across the full score matricies H, E, and F. As such, we must define wavefronts in terms of the diagonals of each segment $c \in C$.

For a score s and segment diagonal $k = j - i$, the furthest-reaching point $\mathcal{R}_{s,k,c}$ indicates the cell in H that is the furthest from the beginning of diagonal k in segment c with score s. $\widetilde{E}_{s,k,c}$, $\widetilde{F}_{s,k,c}$, and $\widetilde{H}_{s,k,c}$ store the offset in the diagonal to furthest-reaching point $\mathcal{R}_{s,k,c}$ in each of the SWG-POA matrices. For a given score s, the s-wavefront $\mathcal{W}_s$ is the set of offsets $\widetilde{E}_{s,k,c}$, $\widetilde{F}_{s,k,c}$, and $\widetilde{H}_{s,k,c}$ for all k on all c. Considering all k on all c, we can more simply refer to the components of $\mathcal{W}_s$ as $\widetilde{E}_s$, $\widetilde{F}_s$, and $\widetilde{H}_s$.

In WFPOA, our goal is to derive a global alignment by computing the minimum s such that any of the furthest-reaching points in $\mathcal{W}_s$ reaches the end

of any of the tails of the graph $V_t = \{v : \delta_v^{out} = \emptyset\ (n', v)\}$, with $n' < n$, or any of the set of points $\{\forall_{v \in V}(n, v)\}$. Equation 4 shows the recurrence relations defined in WFPOA, which reformulate the same scoring model in SWG-POA to consider furthest-reaching points in diagonals on specific segments. We depend on function $\phi(k, c, c^{\prec})$ which computes the diagonal on the inbound segment $c^{\prec}$ corresponding to the diagonal k on c.

$$
\begin{aligned}
\widetilde{E}_{s,k,c} &= \max_{\forall c^{\prec} \in \delta_c^{in}} \left\{ \begin{array}{ll} \widetilde{H}_{s-o-e,\phi(k-1,c,c^{\prec}),c^{\prec}} & \text{(open gap in } Q\text{)} \\ \widetilde{E}_{s-e,\phi(k-1,c,c^{\prec}),c^{\prec}} & \text{(extend gap in } Q\text{)} \end{array} \right\} + 1 \\
\widetilde{F}_{s,k,c} &= \max \left\{ \begin{array}{ll} \widetilde{H}_{s-o-e,k+1,c} & \text{(open gap in } T\text{)} \\ \widetilde{F}_{s-e,k+1,c} & \text{(extend gap in } T\text{)} \end{array} \right\} \\
\widetilde{H}_{s,k,c} &= \max \left\{ \begin{array}{ll} \max_{\forall c^{\prec} \in \delta_c^{in}} \{\widetilde{H}_{s-x,\phi(k,c,c^{\prec}),c^{\prec}} + 1\} & \text{(substitution)} \\ \widetilde{E}_{s,k,c} & \text{(gap in } Q\text{)} \\ \widetilde{F}_{s,k,c} & \text{(gap in } T\text{)} \end{array} \right\}
\end{aligned}
\tag{4}
$$

We also need to extend the concept of distance along a diagonal to record the minimum and maximum possible diagonal traversal distances for each of our furthest-reaching points.

References

[1] Temple F Smith and Michael S Waterman. Comparison of biosequences. *Advances in Applied Mathematics*, 2(4):482–489, 1981.

[2] Osamu Gotoh. An improved algorithm for matching biological sequences. *Journal of Molecular Biology*, 162(3):705–708, 1982.

[3] Santiago Marco-Sola, Juan Carlos Moure, Miquel Moreto, and Antonio Espinosa. Fast gap-affine pairwise alignment using the wavefront algorithm. *Bioinformatics*.

[4] M. Farrar. Striped smith-waterman speeds database searches six times over other simd implementations. *Bioinformatics*, 23(2):156–161, Nov 2006.

[5] C. Lee, C. Grasso, and M. F. Sharlow. Multiple sequence alignment using partial order graphs. *Bioinformatics*, 18(3):452–464, Mar 2002.