1 Derivation of EAD recurrences given in the main text Section 3.2

Beyond the notations introduced in section 3.2 of the main text, the derivation below uses the following additional ones:

$\mathbf{A}$ the **set** of all possible alignments between $\langle S, T\rangle$.

$\mathbf{A}_{(i,j)}$ the **set** of all possible alignments of their prefixes $\langle S_{1\ldots i}, T_{1\ldots j}\rangle$ of the sequences.

$\mathbf{A}^{\mathtt{m}}_{(i,j)}$ the **subset** of all alignments of prefixes that end in a `match(m)` state at cell (i, j).

$\mathbf{A}^{\mathtt{i}}_{(i,j)}$ the **subset** of all alignments of prefixes that end in a `insert(i)` state at cell (i, j).

$\mathbf{A}^{\mathtt{d}}_{(i,j)}$ the **subset** of all alignments of prefixes that end in a `delete(d)` state at cell (i, j).

$\mathcal{A}^{\mathtt{m}}_{(i,j)}$ any **alignment** of prefixes that ends in a `match(m)` state at (i, j).

$\mathcal{A}^{\mathtt{i}}_{(i,j)}$ any **alignment** of prefixes that ends in a `insert(i)` state at (i, j).

$\mathcal{A}^{\mathtt{d}}_{(i,j)}$ any **alignment** of prefixes that ends in a `delete(d)` state at (i, j).

$\mathcal{A}^{\mathtt{m}|\mathtt{m}}_{(i,j)}$ any **alignment** of prefixes that ends in a `match(m)` state at (i, j) given a `match(m)` state at $(i-1, j-1)$. (Similar notation for all 9 possible transitions going between any two states of {`match`, `insert`, `delete`}.)

$\Pr(\mathtt{m}|\mathtt{m})$ the transition probability of going into a `match` given a previous `match` state. (Similar notation for all 9 possible transitions going between any two states of {`match`, `insert`, `delete`}.)

$\Pr(\langle s_i, t_j\rangle)$ the joint probability of matching a pair of amino acids, $s_i \in S$ and $t_j \in T$.

Derivation

Starting with recurrence (6) in the main text, by the definition of $EAD^{\mathtt{m}}(i, j)$, we have:

$$EAD^{\mathtt{m}}(i,j) = \sum_{\forall \mathcal{A}^{\mathtt{m}}_{(i,j)} \in \mathbf{A}^{\mathtt{m}}_{(i,j)}} \Pr(\mathcal{A}^{\mathtt{m}}_{(i,j)}, \langle S_{1\ldots i}, T_{1\ldots j}\rangle) \times \text{distance}(\mathcal{A}^{\mathtt{m}}_{(i,j)}, \mathcal{A}_{\text{ref}}), \quad (1)$$

But all alignments $\mathbf{A}^{\mathtt{m}}_{(i,j)}$ that end in a `match (m)` state at (i, j) are derived by extending all alignments arriving at the cell $(i-1, j-1)$ in any of the three alignment states ({`match`, `insert`, `delete`}), that is the set of alignments $\mathbf{A}_{(i-1,j-1)} = \mathbf{A}^{\mathtt{m}}_{(i-1,j-1)} \cup \mathbf{A}^{\mathtt{i}}_{(i-1,j-1)} \cup \mathbf{A}^{\mathtt{d}}_{(i-1,j-1)}$, by a pair of matched amino acids corresponding to the cell (i, j), that is, $\langle s_i, t_j\rangle$.

Therefore, Equation 1, can be decomposed based on the above observation as:

$$\begin{aligned} EAD^{\mathtt{m}}(i,j) &= \sum_{\forall \mathcal{A}^{\mathtt{m}|\mathtt{m}}_{(i,j)} \in \mathbf{A}^{\mathtt{m}}_{(i,j)}} \Pr(\mathcal{A}^{\mathtt{m}|\mathtt{m}}_{(i,j)}, \langle S_{1\ldots i}, T_{1\ldots j}\rangle) \times \text{distance}(\mathcal{A}^{\mathtt{m}|\mathtt{m}}_{(i,j)}, \mathcal{A}_{\text{ref}}) \\ &+ \sum_{\forall \mathcal{A}^{\mathtt{m}|\mathtt{i}}_{(i,j)} \in \mathbf{A}^{\mathtt{m}}_{(i,j)}} \Pr(\mathcal{A}^{\mathtt{m}|\mathtt{i}}_{(i,j)}, \langle S_{1\ldots i}, T_{1\ldots j}\rangle) \times \text{distance}(\mathcal{A}^{\mathtt{m}|\mathtt{i}}_{(i,j)}, \mathcal{A}_{\text{ref}}) \\ &+ \sum_{\forall \mathcal{A}^{\mathtt{m}|\mathtt{d}}_{(i,j)} \in \mathbf{A}^{\mathtt{m}}_{(i,j)}} \Pr(\mathcal{A}^{\mathtt{m}|\mathtt{d}}_{(i,j)}, \langle S_{1\ldots i}, T_{1\ldots j}\rangle) \times \text{distance}(\mathcal{A}^{\mathtt{m}|\mathtt{d}}_{(i,j)}, \mathcal{A}_{\text{ref}}) \end{aligned} \quad (2)$$

where the component joint probability terms in the r.h.s of Equation 2 are equivalent to:

$$\begin{aligned} \Pr(\mathcal{A}^{\mathtt{m}|\mathtt{m}}_{(i,j)}, \langle S_{1\ldots i}, T_{1\ldots j}\rangle) &= \Pr(\mathcal{A}^{\mathtt{m}}_{(i-1,j-1)}, \langle S_{1\ldots i-1}, T_{1\ldots j-1}\rangle) \times \Pr(\mathtt{m}|\mathtt{m}) \times \Pr(\langle s_i, t_j\rangle) \\ \Pr(\mathcal{A}^{\mathtt{m}|\mathtt{i}}_{(i,j)}, \langle S_{1\ldots i}, T_{1\ldots j}\rangle) &= \Pr(\mathcal{A}^{\mathtt{i}}_{(i-1,j-1)}, \langle S_{1\ldots i-1}, T_{1\ldots j-1}\rangle) \times \Pr(\mathtt{m}|\mathtt{i}) \times \Pr(\langle s_i, t_j\rangle) \\ \Pr(\mathcal{A}^{\mathtt{m}|\mathtt{d}}_{(i,j)}, \langle S_{1\ldots i}, T_{1\ldots j}\rangle) &= \Pr(\mathcal{A}^{\mathtt{d}}_{(i-1,j-1)}, \langle S_{1\ldots i-1}, T_{1\ldots j-1}\rangle) \times \Pr(\mathtt{m}|\mathtt{d}) \times \Pr(\langle s_i, t_j\rangle) \end{aligned}$$

Further, the component distance terms in the r.h.s of Equation 2 can be expanded as:

$$\begin{aligned} \text{distance}(\mathcal{A}^{\mathtt{m}|\mathtt{m}}_{(i,j)}, \mathcal{A}_{\text{ref}}) &= \text{distance}(\mathcal{A}^{\mathtt{m}}_{(i-1,j-1)}, \mathcal{A}_{\text{ref}}) + \delta(i+j-1) + \delta(i+j) \\ \text{distance}(\mathcal{A}^{\mathtt{m}|\mathtt{i}}_{(i,j)}, \mathcal{A}_{\text{ref}}) &= \text{distance}(\mathcal{A}^{\mathtt{i}}_{(i-1,j-1)}, \mathcal{A}_{\text{ref}}) + \delta(i+j-1) + \delta(i+j) \\ \text{distance}(\mathcal{A}^{\mathtt{m}|\mathtt{d}}_{(i,j)}, \mathcal{A}_{\text{ref}}) &= \text{distance}(\mathcal{A}^{\mathtt{d}}_{(i-1,j-1)}, \mathcal{A}_{\text{ref}}) + \delta(i+j-1) + \delta(i+j) \end{aligned}$$

This holds because any alignment ending in a `match` state at (i, j) must arrive from $(i-1, j-1)$, and in doing so, will cross two skew-diagonals (see Figure 1. Also cf. Figure 1 in the main text)

1. one skew-diagonal indexed by $i + j - 1$ and
2. the other skew-diagonal indexed by $i + j$.

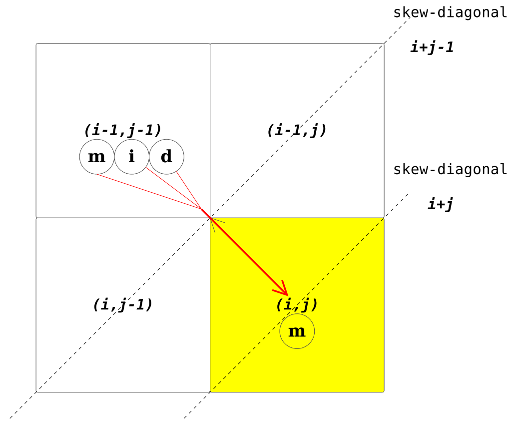

Figure 1: All alignments ending in a `match` state at (i, j) cross two skew-diagonals, along which their additional distances has to be accounted for during dynamic programming

Thus, for all alignments going from $(i-1, j-1)$ to (i, j), the component distance terms at $(i-1, j-1)$ get augmented by a $\delta(i+j-1) + \delta(i+j)$, accounting for their widths/slacks with respect to the reference alignment $\mathcal{A}_{\text{ref}}$ along the above two skew-diagonals.

Substituting the expanding the component joint probability and distance terms shown above into Equation 2, after rearranging yields:

$$\begin{aligned}
EAD^{\mathtt{m}}(i,j) \;=\;& \underbrace{\sum_{\forall \mathcal{A}^{\mathtt{m}}_{(i-1,j-1)} \in \mathbf{A}^{\mathtt{m}}_{(i-1,j-1)}} \Pr(\mathcal{A}^{\mathtt{m}}_{(i-1,j-1)}, \langle S_{1\ldots i-1}, T_{1\ldots j-1}\rangle) \times \text{distance}(\mathcal{A}^{\mathtt{m}}_{(i-1,j-1)}, \mathcal{A}_{\text{ref}}) \times \Pr(\mathtt{m}|\mathtt{m})}_{EAD^{\mathtt{m}}(i-1,j-1)} \times \Pr(\langle s_i, t_j\rangle) \\
&+ \sum_{\forall \mathcal{A}^{\mathtt{m}}_{(i-1,j-1)} \in \mathbf{A}^{\mathtt{m}}_{(i-1,j-1)}} \Pr(\mathcal{A}^{\mathtt{m}}_{(i-1,j-1)}, \langle S_{1\ldots i-1}, T_{1\ldots j-1}\rangle) \times \Pr(\mathtt{m}|\mathtt{m}) \times \Pr(\langle s_i, t_j\rangle) \times [\delta(i+j-1)+\delta(i+j)] \\
+\;& \underbrace{\sum_{\forall \mathcal{A}^{\mathtt{i}}_{(i-1,j-1)} \in \mathbf{A}^{\mathtt{i}}_{(i-1,j-1)}} \Pr(\mathcal{A}^{\mathtt{i}}_{(i-1,j-1)}, \langle S_{1\ldots i-1}, T_{1\ldots j-1}\rangle) \times \text{distance}(\mathcal{A}^{\mathtt{i}}_{(i-1,j-1)}, \mathcal{A}_{\text{ref}}) \times \Pr(\mathtt{m}|\mathtt{i})}_{EAD^{\mathtt{i}}(i-1,j-1)} \times \Pr(\langle s_i, t_j\rangle) \\
&+ \sum_{\forall \mathcal{A}^{\mathtt{i}}_{(i-1,j-1)} \in \mathbf{A}^{\mathtt{i}}_{(i-1,j-1)}} \Pr(\mathcal{A}^{\mathtt{i}}_{(i-1,j-1)}, \langle S_{1\ldots i-1}, T_{1\ldots j-1}\rangle) \times \Pr(\mathtt{m}|\mathtt{i}) \times \Pr(\langle s_i, t_j\rangle) \times [\delta(i+j-1)+\delta(i+j)] \\
+\;& \underbrace{\sum_{\forall \mathcal{A}^{\mathtt{d}}_{(i-1,j-1)} \in \mathbf{A}^{\mathtt{d}}_{(i-1,j-1)}} \Pr(\mathcal{A}^{\mathtt{d}}_{(i-1,j-1)}, \langle S_{1\ldots i-1}, T_{1\ldots j-1}\rangle) \times \text{distance}(\mathcal{A}^{\mathtt{d}}_{(i-1,j-1)}, \mathcal{A}_{\text{ref}}) \times \Pr(\mathtt{m}|\mathtt{d})}_{EAD^{\mathtt{d}}(i-1,j-1)} \times \Pr(\langle s_i, t_j\rangle) \\
&+ \sum_{\forall \mathcal{A}^{\mathtt{d}}_{(i-1,j-1)} \in \mathbf{A}^{\mathtt{d}}_{(i-1,j-1)}} \Pr(\mathcal{A}^{\mathtt{d}}_{(i-1,j-1)}, \langle S_{1\ldots i-1}, T_{1\ldots j-1}\rangle) \times \Pr(\mathtt{m}|\mathtt{d}) \times \Pr(\langle s_i, t_j\rangle) \times [\delta(i+j-1)+\delta(i+j)]
\end{aligned} \tag{3}$$

By grouping all even terms on the r.h.s. of Equation 3 together, we get the recurrence:

$$\begin{aligned}
EAD^{\mathtt{m}}(i,j) \;=\;& EAD^{\mathtt{m}}(i-1,j-1) \times \Pr(\mathtt{m}|\mathtt{m}) \times \Pr(\langle s_i, t_j\rangle) \\
+\;& EAD^{\mathtt{i}}(i-1,j-1) \times \Pr(\mathtt{m}|\mathtt{i}) \times \Pr(\langle s_i, t_j\rangle) \\
+\;& EAD^{\mathtt{d}}(i-1,j-1) \times \Pr(\mathtt{m}|\mathtt{d}) \times \Pr(\langle s_i, t_j\rangle) \\
+\;& \Bigg(\sum_{\forall \mathcal{A}^{\mathtt{m}}_{(i-1,j-1)} \in \mathbf{A}^{\mathtt{m}}_{(i-1,j-1)}} \Pr(\mathcal{A}^{\mathtt{m}}_{(i-1,j-1)}, \langle S_{1\ldots i-1}, T_{1\ldots j-1}\rangle) \times \Pr(\mathtt{m}|\mathtt{m}) \times \Pr(\langle s_i, t_j\rangle) \\
&+ \sum_{\forall \mathcal{A}^{\mathtt{i}}_{(i-1,j-1)} \in \mathbf{A}^{\mathtt{i}}_{(i-1,j-1)}} \Pr(\mathcal{A}^{\mathtt{i}}_{(i-1,j-1)}, \langle S_{1\ldots i-1}, T_{1\ldots j-1}\rangle) \times \Pr(\mathtt{m}|\mathtt{i}) \times \Pr(\langle s_i, t_j\rangle) \\
&+ \sum_{\forall \mathcal{A}^{\mathtt{d}}_{(i-1,j-1)} \in \mathbf{A}^{\mathtt{d}}_{(i-1,j-1)}} \Pr(\mathcal{A}^{\mathtt{d}}_{(i-1,j-1)}, \langle S_{1\ldots i-1}, T_{1\ldots j-1}\rangle) \times \Pr(\mathtt{m}|\mathtt{d}) \times \Pr(\langle s_i, t_j\rangle) \Bigg) \\
&\times [\delta(i+j-1)+\delta(i+j)]
\end{aligned} \tag{4}$$

But the last term on the r.h.s. is the marginal probability over all alignments ending in a `match` at (i, j), resulting in the final form of recurrence (6) used in the main text:

$$\begin{aligned}
EAD^{\mathtt{m}}(i,j) \;=\;& EAD^{\mathtt{m}}(i-1,j-1) \times \Pr(\mathtt{m}|\mathtt{m}) \times \Pr(\langle s_i, t_j\rangle) \\
+\;& EAD^{\mathtt{i}}(i-1,j-1) \times \Pr(\mathtt{m}|\mathtt{i}) \times \Pr(\langle s_i, t_j\rangle) \\
+\;& EAD^{\mathtt{d}}(i-1,j-1) \times \Pr(\mathtt{m}|\mathtt{d}) \times \Pr(\langle s_i, t_j\rangle) \\
+\;& \Pr_{\text{marginal}}(\langle S_{1\ldots i}, T_{1\ldots j}\rangle \,|\mathtt{match@}(i,j)) \times [\delta(i+j-1)+\delta(i+j)]
\end{aligned} \tag{5}$$

Recurrences (7) and (8) in the main text follow identical lines of derivations, with the only difference that they account for all alignments coming into (i, j) in a `insert`(i) and `delete`(d) states, respectively. Also, all such alignment transitions only cross a single skew-diagonal, indexed by $i+j$, therefore those recurrences will contain only the $\delta(i+j)$ term.