6.5 Multiple Alignment

6.5.1 Feasible Lower Bound

Consider strings $S_1, \ldots, S_k$. As we have seen in Chap. 2, multiple alignment becomes easy whenever we have a tree structure on string set $S_1, \ldots, S_k$ at hand such that only pairs of string connected in the tree count for overall score (instead of all pairs as in sum-of-pairs scoring).

For this section, let us switch from score maximization to distance minimization. This is only a minor modification, but has the advantage that in terms of distance functions notions such as the triangle inequality make sense, which would be rather unusual for scoring functions. Thus we assume that we have a symmetric distance measure $d(x, y) = d(y, x)$ on pairs of characters, and $d(x, -) = d(-, x)$ on pairs of characters and spacing symbol. Further assume that $d(x, x) = 0$. Usually we also assume triangle inequality, that is $d(u, v) \leq d(u, w) + d(w, v)$. If in addition $d(u, v) > 0$ for any two different objects u and v holds, we call d a metric.

Given a distance measure d on pairs of characters and spacing symbol, we define the distance for an alignment T_1, T_2 of common length n as follows.

$$d^*(T_1, T_2) = \sum_{p=1}^{n} d(T_1(p), T_2(p)) \tag{6.13}$$

Given strings S_1 and S_2 we define their minimum distance value taken over all alignments T_1, T_2 of S_1 with S_2.

$$d_{\text{opt}}(S_1, S_2) = \min_{\text{alignments } T_1, T_2} d^*(T_1, T_2) \tag{6.14}$$

As in Sect. 6.3.3 we make use of center string S_{center} for string set $S_1, \ldots, S_k$ defined by the following equation:

$$\sum_{i=1}^{k} d_{\text{opt}}(S_{\text{center}}, S_i) = \min_{a=1,\ldots,k} \sum_{i=1}^{k} d_{\text{opt}}(S_a, S_i) \tag{6.15}$$

Lemma 6.12. ***Feasible Lower Bound***
Let $T_1, \ldots, T_k$ be an optimal alignment of $S_1, \ldots, S_k$, and S_{center} be a center string for $S_1, \ldots, S_k$. Then the following holds.

$$\frac{k}{2} \sum_{i=1}^{k} d_{\text{opt}}(S_i, S_{\text{center}}) \leq d^*(T_1, \ldots, T_k)$$