0.1 Steiner consensus strings

Recall that $D(S_i, S_j)$ denotes the weighted edit distance of strings S_i and S_j.

Definition Given a set of strings $\mathcal{S}$, and given another string S, the *consensus error* of a string S *relative* to $\mathcal{S}$ is $E(S) = \sum_{S_i \in \mathcal{S}} D(S, S_i)$. Note that S need not be from $\mathcal{S}$.

Definition Given a set of strings $\mathcal{S}$, an optimal *Steiner string* S^* for $\mathcal{S}$ is a string that *minimizes* the consensus error $E(S^*)$ over all possible strings.

Note that S^* need not be from $\mathcal{S}$, and generally will not be. We will usually refer to a "Steiner consensus string" as a "Steiner string".

The Steiner string S^* attempts to capture and reflect in a single string, the common characteristics of the set of strings $\mathcal{S}$. There is no known efficient method to find S^*, but there is an approximation method which finds a string S such that the ratio $E(S)/E(S^*)$ is never more than two.

Lemma 0.1 *Let $\mathcal{S}$ have k strings. Then there exists a string $\overline{S} \in \mathcal{S}$ such that $E(\overline{S})/E(S^*) \leq 2 - 2/k < 2$.*

Proof Let $\overline{S}$ be any string in $\mathcal{S}$. By the property of weighted edit distance, $D(\overline{S}, S_i) \leq D(\overline{S}, S^*) + D(S^*, S_i)$ for any S_i. So $E(\overline{S}) = \sum_{S_i \in \mathcal{S}} D(\overline{S}, S_i) \leq \sum_{S_i \neq \overline{S}} (D(\overline{S}, S^*) + D(S^*, S_i)) = (k-2)D(\overline{S}, S^*) + E(S^*)$.

Now pick $\overline{S}$ to be a string in $\mathcal{S}$ which is *closest* to the optimal Steiner string S^*. That is, choose $\overline{S}$ so that $D(\overline{S}, S^*)$ is less than or equal to $D(S_i, S^*)$, for any $S_i \in \mathcal{S}$. (Of course, $\overline{S}$ is not known constructively since S^* is not known, but $\overline{S}$ does exist.) Then $E(S^*) = \sum_{S_i \in \mathcal{S}} D(S^*, S_i) \geq kD(\overline{S}, S^*)$. Therefore $E(\overline{S})/E(S^*) \leq ((k-2)D(\overline{S}, S^*)/kD(\overline{S}, S^*)) + 1 = (k-2)/k + 1 = 2 - 2/k < 2$. □

Now define the *center string* S_c is a string in $\mathcal{S}$ which minimizes $\sum_{S_i \in \mathcal{S}} (S_c, S_i)$ over all strings in $\mathcal{S}$. Clearly, we can efficiently find S_c.

Theorem 0.1 $E(S_c)/E(S^*) \leq 2 - 2/k$.

Proof The theorem follows immediately from Lemma 0.1 and the fact that $E(S_c) \leq E(\overline{S})$ by definition. □