Distances, Tree Metrics, and Good Retractions

The goal of phylogenetic reconstruction is to infer an evolutionary tree relating species (or individual genes) from some observed data. Given a set of sequences $S = \{1, \ldots, n\}$ as raw data (genomes, partial genomes, or proteins), a pairwise distance function $d : S^2 \to \mathbb{R}_{\geq 0}$ is calculated by modeling evolutionary processes (mutations, recombinations, selections, duplications, exchanges). Then, this distance $d$ which is definite (1) and symmetric (2) is used to construct a tree defining a tree metric $d$ which satisfies the triangle inequality (3) and 4-point condition (4).

Moulton and Steel [3] focus on this second step of retraction onto a tree metric.

For the set of distances $D(S)$, the set of tree metrics $T(S) \subset D(S)$, and the permutation group $\Sigma_S$, a map $\phi : D(S) \to D(S)$ is a retraction onto $T(S)$ if $\phi$ is continuous and (5) and (6) hold. The map is also good if $\phi$ is homogeneous (7) and equivariant (8).

Buneman index, refined Buneman index, and associated trees

For any split $\sigma = \{A, B\} \in S(S)$ where $S(S)$ is the set of splits of $S$, Buneman defined a separation index $\mu_\sigma$ (10) which the authors refine to $\bar{\mu}_\sigma$ (11) via a function
Finding Good Trees from Pairwise Distances

\( \beta_q \) (9) on quartets \( q = \{a, a', b, b'\} \in Q_{\sigma} \subseteq S \) with \( \{a, a'\} \subseteq A \) and \( \{b, b'\} \subseteq B \).

\[
\beta_q = \frac{1}{2} \left( \min\{d_{ab} + d_{a'\nu}, d_{ab} + d_{a'b}\} - (d_{a'a'} + d_{b'b}) \right) \quad (9)
\]

\[
\mu_\sigma = \min_{q} \{\beta_q\} \quad (10)
\]

\[
\bar{\mu}_\sigma = \frac{1}{n-3} \sum_{i=1}^{n-3} \beta_{q_i} \text{ such that } \forall 1 \leq i \leq |Q_\sigma| \quad \beta_{q_i} \leq \beta_{q_j} \quad (11)
\]

The refined Buneman index \( \bar{\mu}_\sigma \) defines the map \( \psi : d \to \sum_{\{\sigma, \bar{\mu}_\sigma > 0\}} \bar{\mu}_\sigma \delta_\sigma \). The authors show that the set \( \{\sigma : \bar{\mu}_\sigma > 0\} \) is pairwise compatible and thus determines a unique S-tree (Corollary 5.1), so \( \psi \) satisfies (5). They also show property (6) because when \( d \) is a tree metric with weights \( w \) on the associated tree, \( \bar{\mu}_\sigma = \mu^+_e = w_e \) if edge \( e \) corresponds to split \( \sigma \) else 0. Finally, they show that the trees from \( \psi \) strictly refine those given by the Buneman index \( \mu_\sigma \).

**Proof that the refined Buneman index produces trees**

**Theorem 5.1** If \( \sigma, \sigma' \in S(S) \) are incompatible, then \( \bar{\mu}_\sigma + \bar{\mu}_{\sigma'} \leq 0 \).

**Lemma 5.1** Suppose that \( \sigma = \{A, B\} \in S(S) \) and \( \sigma' = \{A', B'\} \in S(S) \) are incompatible. Then \( |A \cap A'| \times |A \cap B'| \times |B \cap A'| \times |B \cap B'| \geq n - 3 \).

**Proof of Lemma 5.1** Define \( w = |A \cap A'|, x = |A \cap B'|, y = |B \cap A'|, \) and \( z = |B \cap B'| \). Then, \( w + x = |A| \) and \( y + z = |B| \). Additionally, \( |A| + |B| = n \), and since the splits are incompatible, \( |A|, |B| \geq 2 \). So, \( wxyz = w(|A| - w)y(|B| - y) \geq (|A| - 1)(|B| - 1) = |A||B| - |A| - |B| + 1 \geq n - 3 \). \( \square \)

**Proof of Theorem 5.1** For incompatible splits \( \sigma = \{A, B\} \in S(S) \) and \( \sigma' = \{A', B'\} \in S(S) \), choose quartets \( q = ik|jl \) and \( q' = ij|kl \) such that \( i \in A \cap A', \) \( j \in A \cap B', k \in B \cap A', \) and \( l \in B \cap B' \). By definition, \( \beta_q \leq \frac{1}{2}(d_{ij} + d_{kl} - d_{ik} - d_{jl}) \) and \( \beta_{q'} \leq \frac{1}{2}(d_{ik} + d_{jl} - d_{ij} - d_{kl}) \), so \( \beta_q + \beta_{q'} \leq 0 \). By lemma 5.1, there exist at least \( n - 3 \) choices of \( q \) and \( q' \), which get denoted as \( \bar{q}_i, \bar{q}'_i, 1 \leq i \leq n - 3 \). This makes \( \bar{\mu}_\sigma + \bar{\mu}_{\sigma'} \leq \frac{1}{n-3} \sum_{i=1}^{n-3} (\beta_{\bar{q}_i} + \beta_{\bar{q}'_i}) \leq 0 \). \( \square \)

**Further reading**

Related findings place bounds on how closely a retraction approximates the closest tree metric [1] and organize several algorithms into a structured family to show properties of the trees resulting from the methods and the computational complexities required for their construction [2].
References

