Finding Good Trees from Pairwise Distances

MATH 833 - Fall 2012 *Presenter: Mark Chapman*

Reference: Retractions of finite distance functions onto tree metrics [3].

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.

$$\forall i, j \in S \quad d_{ij} = 0 \Leftrightarrow i = j \tag{1}$$
$$\forall i, j \in S \quad d_{ij} = d_{ji} \tag{2}$$
$$\forall i, j, k \in S \quad d_{ik} \leq d_{ij} + d_{jk} \tag{3}$$
$$\forall i, j, k, l \in S \quad d_{ij} + d_{kl} \leq \max\{d_{ik} + d_{jl}, d_{il} + d_{jk}\} \tag{4}$$

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

$$\forall d \in \mathcal{D}(S) \quad \phi(d) \in \mathcal{T}(S) \tag{5}$$
$$\forall d \in \mathcal{T}(S) \quad \phi(d) = d \tag{6}$$
$$\forall d \in \mathcal{D}(S)\ \forall \lambda > 0 \quad \phi(\lambda d) = \lambda \phi(d) \tag{7}$$
$$\forall \tau \in \Sigma_S \quad \phi(d^\tau) = \phi(d)^\tau \text{ where } (d^\tau)_{ij} = d_{\tau(i)\tau(j)} \tag{8}$$

Buneman index, refined Buneman index, and associated trees

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

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

$$\beta_q = \frac{1}{2}(\min\{d_{ab} + d_{a'b'}, d_{ab'} + d_{a'b}\} - (d_{aa'} + d_{bb'})) \quad (9)$$

$$\mu_\sigma = \min_q\{\beta_q\} \quad (10)$$

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

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 ψ satisifies (5). They also show property (6) because when d is a tree metric with weights w on the associated tree, $\bar{\mu}_\sigma = \mu^+_\sigma = w_e$ if edge e corresponds to split σ else 0. Finally, they show that the trees from ψ strictly refine those given by the Buneman index μ_σ.

Proof that the refined Buneman index produces trees

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

Lemma 5.1 Suppose that $\sigma = \{A, B\} \in \mathcal{S}(S)$ and $\sigma' = \{A', B'\} \in \mathcal{S}(S)$ 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, $wxyz = w(|A| - w)y(|B| - y) \geq (|A| - 1)(|B| - 1) = |A||B| - |A| - |B| + 1 \geq n - 3$. □

Proof of Theorem 5.1 For incompatible splits $\sigma = \{A, B\} \in \mathcal{S}(S)$ and $\sigma' = \{A', B'\} \in \mathcal{S}(S)$, 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 $\hat{q}_i, \hat{q}'_i, 1 \leq i \leq n - 3$. This makes $\bar{\mu}_\sigma + \bar{\mu}_{\sigma'} \leq \frac{1}{n-3}\sum_{i=1}^{n-3}(\beta_{\hat{q}_i} + \beta_{\hat{q}'_i}) \leq 0$. □

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

[1] Richa Agarwala, Vineet Bafna, Martin Farach, Mike Paterson, and Mikkel Thorup. On the approximability of numerical taxonomy (fitting distances by tree metrics). *SIAM Journal on Computing*, 28:1073–1085, 1999.

[2] David Bryant and Vincent Berry. A structured family of clustering and tree construction methods. *Advances in Applied Mathematics*, 27:705–732, 2001.

[3] Vincent Moulton and Mike Steel. Retractions of finite distance functions onto tree metrics. *Discrete Applied Mathematics*, 91:215–233, 1999.