Continuous Retractions onto Tree Metrics

MATH285K - Spring 2010 *Presenter: Shagnik Das*

Reference: [4].

1 Introduction

Given data on different taxa, our aim is to construct the evolutionary tree depicting the relationships between the different species. Distance-based methods convert the available data in a dissimilarity measure, or a (pseudo)metric on the set X of species, and then build a tree using the metric. Recalling that there is a one-to-one correspondence between X-trees and tree metrics, and that it is easy to go from one to the other (see, for instance, [1]), the problem of reconstructing trees reduces to one of mapping metrics to tree metrics.

Neighbour-Joining, a greedy recursive clustering algorithm, is a widely-used method for solving this problem. However, it is sometimes criticised for not producing mapping metrics continuously onto tree metrics. A continuous map was defined by Buneman (see [3]), but this tends to give underresolved star-like maps. In this paper, the authors Moulton and Steel introduce a refinement of Buneman's retraction that provides more meaningful trees.

During this talk, we shall first outline the key properties we would like our method to satisfy. Next, we shall define the three methods. Finally, we will consider a couple of examples comparing the methods.

2 Good Maps

For a set X of taxa, let us denote the set of metrics on X by $\mathcal{M}(X)$, and the subset of tree metrics by $\mathcal{T}(X)$. A distance-based method can then be thought of as a map $\Phi : \mathcal{M}(X) \to \mathcal{T}(X)$. We call such a map a *good map* if it has the four following properties:

- **Continuity**: We require the map Φ to be continuous, as small changes in our data should not lead to large changes in our evolutionary theory
- **Retraction on** $\mathcal{T}(X)$: If the observed data comes from a tree, the method should output the same tree.
- **Homogeneity**: If we scale the distances in our data by a scalar λ, the distances in our output should be scaled by the same factor. In other words, changing the units of measurement should not affect the evolutionary history of the taxa.
- **Equivariance**: Relabelling the taxa in our data should provide the same tree with relabelled vertices - the order of consideration of the species should not affect their evolutionary relationships.

These are clearly all desirable properties of our maps.

3 Neighbour-Joining

As mentioned before, Neighbour-Joining is a widely-used method for reconstructing phylogenetic trees. Essentially, given a dissimilarity measure, you find two taxa x and y that are close to one another, and relatively far from all other taxa. We replace this pair with a new parent taxa u, and update the metric to include distances to u. As we now have one fewer species, we can recursively generate a tree T'. Note that the base case, with two species, is obvious. We then add x and y as leaves under u to get the tree T from T'. The distances are chosen to match the average discrepancy in distances from x and distances from y.

It is a popular method because it is very quick, an intuitive approach, and produces highly resolved trees - there are no vertices of degree greater than 3. More importantly, the method is guaranteed to preserve tree metrics. However, as will be shown later in these notes, the method is *not* continuous - two dissimilarity measures that are very close to one another (with respect to an L^p norm, say) can produce very different trees.

4 Splits and Metrics

Recall that a split $\sigma = A|B$ is a partition of $X = A \cup B$ into disjoint non-empty sets. We say two splits $\sigma = A|B$ and $\sigma' = A'|B'$ are **compatible** if at least one of the sets $A \cap A'$, $A \cap B'$, $B \cap A'$ or $B \cap B'$ is empty. By the Splits Equivalence Theorem, there is a bijection between sets of pairwise compatible splits and X-trees.

To represent our weighted trees, we need to introduce a metric structure on the splits. Given a split σ, we can define a split metric δ_σ:

$$\delta_\sigma(x,y) = \begin{cases} 1 & \text{if } x \in A, y \in B \text{ or } x \in B, y \in A \\ 0 & \text{otherwise} \end{cases}$$

Then any positive linear combination $\sum_\sigma \lambda_\sigma \delta_\sigma$, $\lambda_\sigma \geq 0$, is also a metric on X. In particular, if T is an X-tree, then we can recover the corresponding tree metric by setting

$$\lambda_\sigma = \begin{cases} \omega_e & \text{if } \sigma \text{ corresponds to edge } e \\ 0 & \text{otherwise} \end{cases}$$

where ω_e is the weight of the edge e.

It is shown in [4] that the map between tree metrics and the vector of coefficients $(\lambda_\sigma)_\sigma$ is in fact a homeomorphism of metric spaces. Thus any method can be viewed as a map from metrics on X to the coefficients in the split-representation of tree metrics.

5 The Buneman Retraction

In [3], Buneman constructs a good map, the details of which we outline here. Define a **quartet** $q = wx|yz$ to be a partial split on at most four taxa, with at most two on either side. If the quartet is exhibit in the split σ, we say that q is a **subquartet** of σ, denoted $q \subset \sigma$.

If two splits σ, σ' are incompatible, then the four sets $A \cap A'$, $A \cap B'$, $B \cap A'$ and $B \cap B'$ are all nonempty. Thus we can find a subquartet $q \subset \sigma$ containing one element from each of these sets. There

is a corresponding subquartet $q' \subset \sigma'$. We call these subquartets **witnesses**.

For a quartet $q = wx|yz$, define the **weight** β_q of the quartet as follows:

$$\beta_q = \frac{1}{2}\left[\min\left\{ d(w,y) + d(x,z), d(w,z) + d(x,y) \right\} - \Big(d(w,x) + d(y,z) \Big) \right]$$

We can now define the **Buneman index** of a split σ as $\mu_\sigma = \min_{q \subset \sigma}\{\beta_q\}$. The positive part of the Buneman index is denoted by $\mu_\sigma^+ = \max\{\mu_\sigma, 0\}$.

The key result in defining the map, as appearing in [3], is that if σ and σ' are incompatible splits, then $\mu_\sigma + \mu_{\sigma'} \leq 0$. This is because if q and q' are corresponding witnesses for the two splits, then, following easily from the definitions, $\beta_q + \beta_q' \leq 0$. As a corollary of this result, the splits with positive indices must form a pairwise compatible family of splits (and thus give rise to a unique X-tree).

The Buneman retraction is then defined as

$$\Phi^{(B)}(d) = \sum_\sigma \mu_\sigma^+(d)\delta_\sigma$$

By our earlier remarks, this maps any metric to a tree metric. In [3] it is further proven that this map is indeed a good map.

However, the drawback of this method is that, especially when there are many species being studied, there may be very few splits with positive indices. This gives rise to highly unresolved graphs, as will be demonstrated in a later example. While Buneman was happy to accept this as the price paid for a continuous map, Moulton and Steel were able to modify the indices to provide a continuous map producing more refined trees.

6 The Refined Buneman Retraction

The authors were able to improve the retraction by noting that incompatible splits have many witnesses. Suppose $\sigma = A|B$ and $\sigma = A'|B'$ are incompatible splits. Let $W_1 = A \cap A'$, $W_2 = A \cap B'$, $W_3 = B \cap A'$ and $W_4 = B \cap B'$. Then for every choice of $w \in W_1$, $x \in W_2$, $y \in W_3$ and $z \in W_4$, $q = wx|yz$ is a witness for σ while $q' = wy|xz$ is a witness for σ'. Hence the number of witness is given by $m = |W_1||W_2||W_3||W_4|$. Given that these sets partition $\{1, 2, \ldots, n\}$, and are all non-empty, it follows that m is at least $n - 3$.

This motivates the definition of the refined indices as the average weight of the $n-3$ smallest quartets:

$$\overline{\mu_\sigma} = \frac{1}{n-3}\sum_{j=1}^{n-3} \beta_{q_j}$$

where the subquartets of σ are ordered by increasing weight, so that $\beta_{q_1} \leq \beta_{q_2} \leq \ldots$

Since the weights of corresponding witness quartets for incompatible splits is non-positive, it follows that for incompatible splits, $\overline{\mu_\sigma} + \overline{\mu_{\sigma'}} \leq 0$. Hence we once again get a map from metrics to tree metrics defined by

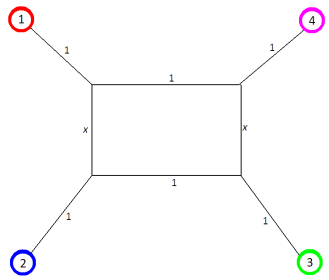
$$\Phi^{(R)}(d) = \sum_\sigma \overline{\mu_\sigma^+}(d)\delta_\sigma$$

The authors were again able to establish in [4] that this refined Buneman retraction is indeed a good map. Since $\overline{\mu_\sigma} \geq \mu_\sigma$, we tend to get more refined trees than with the original Buneman retraction.

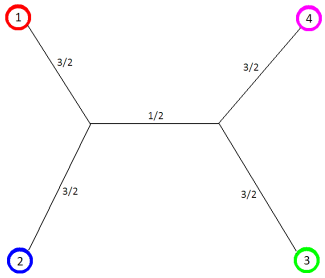
7 Examples

7.1 Example 1

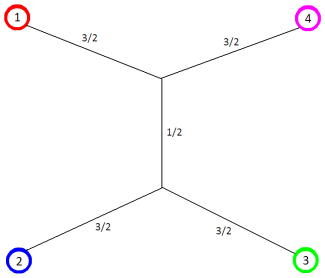
The first example shows that the Neighbour-Joining method is discontinuous. Consider the metric given by the following graph:



Depending on the value of x, we get different results from the algorithm. In particular, as x tends to 1 from above and below, we obtain the two different limits shown below:



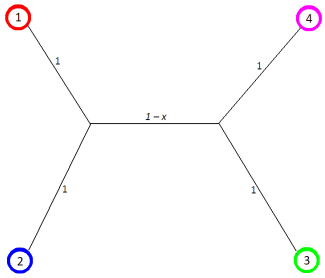
The limit tree as $x \nearrow 1$



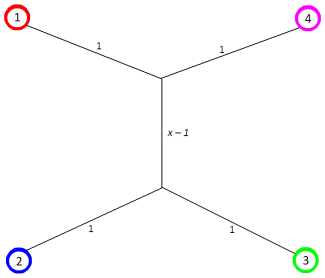
The limit tree as $x \searrow 1$

Thus we see that the map representing the Neighbour-Joining algorithm is not continuous - we can take two arbitrarily close metrics that result in trees a fixed distance apart.

On the other hand, if we apply the refined Buneman retraction to this data, we obtain:



The tree when $x < 1$

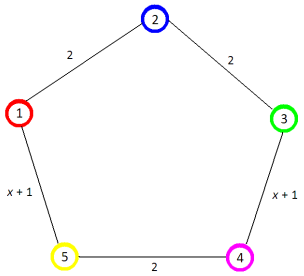


The tree when $x > 1$

In this case, note that the middle edge vanishes when $x = 1$, and so we obtain the star tree in the limit case $x = 1$. The refined Buneman retraction is indeed continuous.

7.2 Example 2

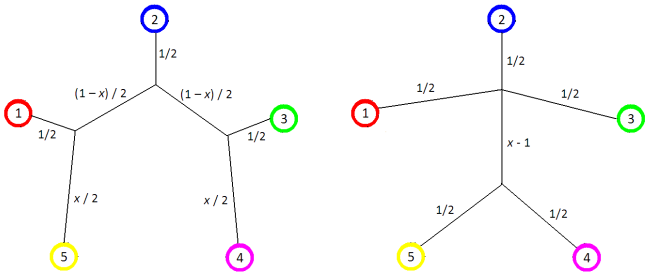
In this example we show that the refined Buneman retraction can provide meaningful trees even when the original Buneman retraction gives us no information. The input metric is derived from the following graph:



Applying the original Buneman retraction, we get the two following unresolved trees:

The tree when $x \leq 2$ The tree when $x \geq 2$

There is very little separation of the taxa, and hence little evolutionary information is gained. Applying the refined Buneman retraction offers greater insight:



The tree when $x \leq 1$ The tree when $x \geq 1$

The greater resolution is evident, and the advantage of the refined retraction is immediately apparent.

8 Conclusion

The preceding examples clearly exhibit the need for and the advantages of the refined Buneman retraction. It is interesting to note that there is a polynomial-time implementation of the algorithm, as discussed in [2]. Thus this new method is not just theoretically appeasing, but also practical.

One possible area of further study is the error-tolerance of the method. The authors show that when the input metric is sufficient close to a tree metric, the correct topology of the tree will be recovered. However, the maximum permissible error for this guarantee is the same as that for Neighbour-Joining, which is in fact optimal for distance-based methods. It would be interesting to see, though, which method is more often returns the correct tree when the errors are beyond the limits given by these theorems. The answer to this question could determine which method is better for practical use.

References

[1] H.-J. Bandelt *Recognition of Tree Metrics.* SIAM J. Discrete Math. 3, 1990.

[2] D. Bryant, V. Moulton *A polynomial time algorithm for constructing the refined Buneman tree.* Appl. Math. Lett., 1998.

[3] P. Buneman *The recovery of trees from measures of dissimilarity.* Mathematics in the Archaelogical and Historical Sciences, 1971.

[4] V. Moulton, M. Steel *Retractions of finite distance functions onto tree metrics.* Discrete Appl. Math., 1999.