

Lecture 19 — April 8th

Lecturer: Lior Pachter

Scribe/ Editor: Wenjing Zheng/ Meromit Schuster

19.1 Neighbor-joining algorithm (review)

Recall from last week, the neighbor-joining algorithm receives as input a dissimilarity map D and outputs a tree additive map. The algorithm can be outlined as follows:

1. Given a dissimilarity map D on a set X , start with a star tree T of $|X|$ vertices. Let $n = |X|$.
2. Compute

$$Q(i, j) = D(i, j) - \frac{1}{n-2} \left(\sum_{k \neq i} D(i, k) + \sum_{k \neq j} D(j, k) \right)$$

Pick a pair (a, b) , with $a \neq b$, that minimizes $Q(i, j)$.

3. Update n to $n - 1$. Update T to the subtree obtained by replacing (a, b) with a new leaf c . Update D to a dissimilarity map on $X' = (X - \{a, b\}) \cup \{c\}$ where $D(i, j)$ remains the same if $i, j \neq c$, and

$$D(c, k) = \frac{1}{2}(D(a, k) + D(b, k) - D(a, b))$$

4. Repeat steps 2 and 3 until $n = 1$.

The resulting tree gives the least square edge-lengths with respect to the given D . The square edge-lengths are given by

$$\sum_{i,j} \frac{1}{2^{p_{ij}}} (D_{ij} - D_{ij}^*)$$

where D^* is the T -additive dissimilarity map produced by the algorithm.

19.2 Tree Generalizations

Definition 19.1. A split is a partition of a set X into two blocks (subsets). To specify X , we may also call it an X -split.

Remark 1. Note that a phylogenetic tree can be represented by a set of splits.

Definition 19.2. A system of splits is a collection of splits.

Definition 19.3. A pair of X -splits $A_1|B_1, A_2|B_2$ is compatible if at least one of the intersections $A_1 \cap A_2, A_1 \cap B_2, B_1 \cap A_2,$ and $B_1 \cap B_2$ is empty.

Remark 2. A set of splits can be represented by a tree iff the splits are pairwise compatible. This is the unrooted analogy of a laminar family being equivalent to a rooted X -forest.

$$\begin{array}{cc} \text{laminar family} & \text{pairwise-compatible split system} \\ \downarrow & \downarrow \\ \text{rooted } X\text{-forest} & \text{unrooted } X\text{-tree} \end{array}$$

19.2.1 Restatement of Fundamental Theorem of Phylogenetics

Given a split $S = \{A|B\}$, let

$$\delta_S(x, y) = \begin{cases} 0 & \text{if } x \text{ and } y \text{ are in same component of } S, \\ 1 & \text{else.} \end{cases}$$

Theorem 19.4. A dissimilarity map D satisfies the (weak) four-point condition iff

$$D_{xy} = \sum_{S \in \mathbb{S}} \lambda_S \delta_S(x, y)$$

where \mathbb{S} is a pairwise compatible split system and $\lambda_S \in \mathbb{R}$ (weak 4-point condition) or $\mathbb{R}_{\geq 0}$ (4-point condition).

Definition 19.5. A circular ordering π on X is a bijection from X to the vertices of an n -cycle.

A split system is circular with respect to π if every split is of the form

$$S = \{x_{i+1} \dots x_j | x_{j+1} \dots x_i\},$$

where $i < j$.

Note that a pairwise-compatible split system is circular.

Definition 19.6. A dissimilarity map D satisfies the Kalmanson condition if there exists some circular ordering π such that $\forall i < j < k < l$,

$$\begin{aligned} D(x_i, x_j) + D(x_k, x_l) &\leq D(x_i, x_k) + D(x_j, x_l), \\ D(x_i, x_l) + D(x_j, x_k) &\leq D(x_i, x_k) + D(x_j, x_l). \end{aligned}$$

(Note that this is in essence the parallelogram law)

If D is a metric, then it is *Kalmanson* with respect to many circular orderings π .

Theorem 19.7. *Let D be a dissimilarity map. Then the following are equivalent:*

1. *There is a circular ordering π and a circular split system \mathbb{S} with respect to π such that*

$$D = \sum_{S \in \mathbb{S}} \lambda_S \delta_S, \quad \lambda_S \geq 0$$

2. *D is a metric and satisfies the Kalmanson condition.*

Definition 19.8. *If D is a dissimilarity map satisfying the above condition, it is called a circular decomposable metric.*

Proof: (Theorem 19.7) To see that (1) implies (2), it suffices to show that every δ_S satisfies the Kalmanson condition. For every $i < j < k < l$, the circular ordering π will place x_i, x_j, x_k, x_l on a rectangle. Suppose that x_i, x_k are across the diagonal from each other on this rectangle, similarly for x_j, x_l . Every $S \in \mathbb{S}$ either leaves all of them in the same component, or separates x_i, x_k and/or x_j, x_l . In the first case $\delta_S = 0$ for every combination of pairs among these four, so it trivially satisfies Kalmanson. For the latter case, $\delta_S(x_i, x_k) = 1 = \delta_S(x_j, x_l)$. So the RHS is always 2, and by definition of δ_S the LHS is at most 2. So Kalmanson holds.

To see that (2) implies (1), we will use an algorithm that takes as input D , and outputs a tree and a circular ordering. This tree minimizes the lengths of D with respect to a circular ordering. This algorithm is called the neighbor-net algorithm.

First, we need some new concepts. Let $G \subseteq C_n$ be a subgraph of the n -cycle. A partial circular ordering \mathcal{C} consists of G together with a bijection from X to the set of vertices of G . Equivalently, \mathcal{C} is a partition of X into $\{C_1, \dots, C_m\}$. Let \widehat{C}_r denote the endpoints of C_r (by this definition, \widehat{C}_r has at most two elements). A weight for \mathcal{C} is a function $\mu : X \rightarrow \mathbb{R}_{\geq 0}$, such that for every r , $\sum_{i \in C_r} \mu(i) = 1$ and $\mu(i) > 0$ for $i \in \widehat{C}_r$. As an analogy to defining D in neighbor-joining, define

$$\mu(i) = \begin{cases} 1/2 & i \in \widehat{C}_r, |C_r| = 1 \\ 1 & |C_r| = 1 \\ 0 & \text{else} \end{cases}$$

The algorithm proceeds as follows: Given D as in (2), set G to be the trivial graph with no edges, i.e. disjoint union of $n = |X|$ vertices. Set $\mu(i) = 1$, for all i . Let \mathcal{C} be the unique circular ordering with respect to G , where each component is a singleton. While $|\mathcal{C}| > 1$, let

$$Q(C_r, C_s) = (|\mathcal{C}| - 2) D(C_r, C_s) - \sum_{t \neq r} D(C_r, C_t) - \sum_{t \neq s} D(C_s, C_t),$$

where

$$D(C_r, C_t) = \sum_{i \in C_r, j \in C_t} \mu(i)\mu(j)D(i, j), \text{ and}$$

$$D(x, C_r) = \sum_{i \in C_r} \mu(i)D(x, i)$$

Choose a pair C_{r^*}, C_{s^*} that minimizes Q . Then let

$$\begin{aligned} \widehat{Q}(i, j) = & \left(|\mathcal{C}| - 4 + |\widehat{C}_{r^*}| + |\widehat{C}_{s^*}| \right) D(i, j) - \sum_{t \neq r^*, s^*} (D(i, C_t) - D(j, C_t)) \\ & - \sum_{k \in C_{r^*} \cup C_{s^*} - i} D(i, k) - \sum_{k \in C_{r^*} \cup C_{s^*} - j} D(j, k). \end{aligned}$$

Choose the pair $i^* \in \widehat{C}_{r^*}, j^* \in \widehat{C}_{s^*}$ that minimizes \widehat{Q} add edge (i^*, j^*) to G , and merge C_{r^*} and C_{s^*} . Repeat until $|\mathcal{C}| = 1$. \square

Definition 19.9. Let D be a dissimilarity map, and \mathcal{C} be a circular ordering. Then the length of D with respect to the circular ordering \mathcal{C} is given by

$$\mathcal{L}(D, \mathcal{C}) := \frac{1}{|\mathcal{O}(\mathcal{C})|} \sum_{x_1, \dots, x_n \in \mathcal{O}(\mathcal{C})} \left(\frac{1}{2} \sum_{i=1}^n D(x_i, x_{i+1}) \right),$$

where $\mathcal{O}(\mathcal{C})$ is the set of circular orderings consistent with \mathcal{C} .

Figure 19.1 gives an example of this definition.

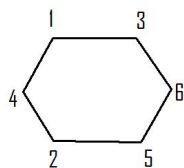


Figure 19.1. $\mathcal{L}(D, \mathcal{C}) = D(x_1, x_3) + D(x_3, x_6) + \dots + D(x_4, x_1)$.

The neighbor-net algorithm described in the above proof will output a tree with a circular ordering such that this tree minimizes the lengths of D with respect to this circular ordering.

19.3 Homework

How many circular orderings are there on a tree of n taxa?