Edges of the Balanced Minimum Evolution Polytope

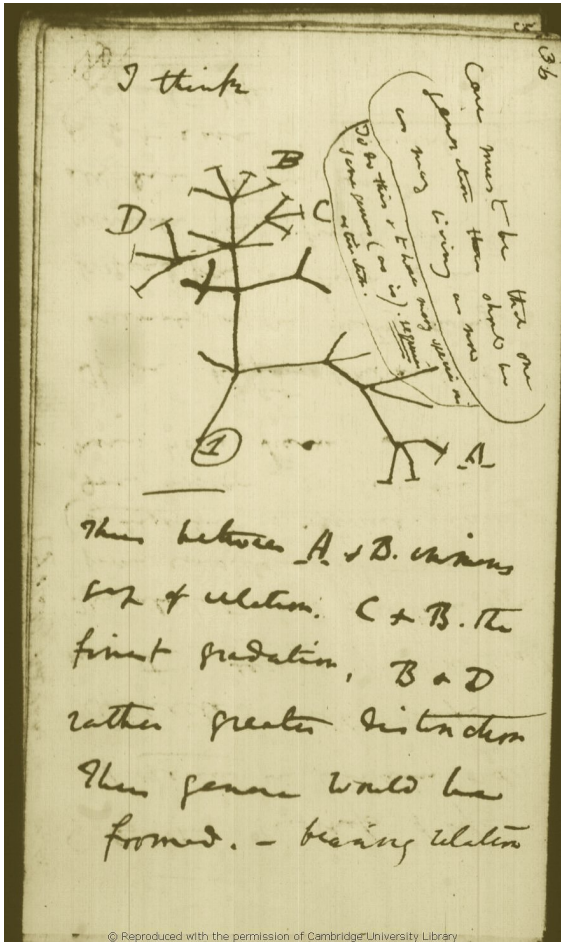
David Haws

Department of Statistics
University of Kentucky

David Haws[1], Terrel Hodge[2], Ruriko Yoshida[1]

[1]University of Kentucky, Dept. of Statistics
[2]Western Michigan University, Dept. of Mathematics

Phylogenetics



A sketch of a species tree from Darwin's early work.

[1930,1950] Phylogenies (cladograms) built based on shared morphological ancestral data (Zimmerman, Hennig)

[1960] Zuckerkandl and Pauling proposed using molecular data (DNA, protien, ...) for building phylogenies.

Pairwise Distance to Trees

Let $\mathcal{T}_n$ be the set of all unrooted binary trees on n leaves.

A clade of $T \in \mathcal{T}_n$ is a subgraph given by an internal node and all its children.

Observe alignment of DNA for n taxa, and compute pairwise distances $d(i,j)$.

Disimilarity Matrix

	A	B	C	D	E
A	0	7.1	8.9	14	13.2
B	7.1	0	7	12.1	11.3
C	8.9	7	0	11.7	10.9
D	14	12.1	11.7	0	3.6
E	13.2	11.3	10.9	3.6	0

Want tree $T \in \mathcal{T}_n$ that best explains the distances $d(i,j)$.

Follow the minimum evolution principle.

- For $T \in \mathcal{T}_n$, estimate the branch lengths from $d(i,j)$.
- Pick T with smallest tree length $:= \sum_{e \in E(T)} \text{length}(e)$.

Balanced Minimum Evolution

Given the pairwise distances $d(i,j)$ between n taxa and $T \in \mathcal{T}_n$, how does one estimate the branch lengths of T?

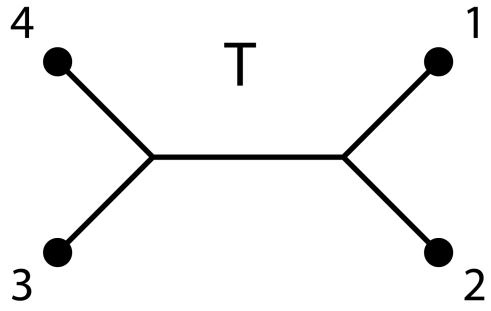
The Balanced Minimum Evolution scheme is a weighted least squares estimate of the branch lengths of T, given $d(i,j)$, that puts more confidence on the short evolutionary distances than on larger ones.

For $T \in \mathcal{T}_n$ and $1 \leq i,j \leq n$ define

$y_{i,j}^T :=$ # edges between i and j,

$$w_{i,j}^T := \frac{1}{2^{\mathbf{y}_{i,j}^T - 1}}$$

$$\mathbf{w}^T := (w_{1,2}^T, w_{1,3}^T, \ldots, w_{n-1,n}^T)$$



$$\mathbf{w}^T = \left(\overset{12}{\frac{1}{2}}, \overset{13}{\frac{1}{4}}, \overset{14}{\frac{1}{4}}, \overset{23}{\frac{1}{4}}, \overset{24}{\frac{1}{4}}, \overset{34}{\frac{1}{2}}\right)$$

Balanced Minimum Evolution Polytope

In the BME framework, the tree length can be expressed by Pauplin's formula

$$\sum_{e \in E(T)} \text{length}(e) = \sum_{i,j} w_{i,j}^T d(i,j).$$

Thus, given $d(i,j)$, the BME scheme is to find $T \in \mathcal{T}_n$ minimizing

$$\sum_{i,j} w_{i,j}^T d(i,j).$$

BME polytope is defined as

$$\mathcal{P}_n := \text{conv}\left(\mathbf{w}^T \mid T \in \mathcal{T}_n \right).$$

- $(2n-5)!!$ many binary trees.
- NP-hard to optimize (Day 87).

The software FastME uses tree pivot moves (NNI) to change trees and heuristically optimize Pauplin's formula. Adjacency on the BME polytope is unknown, as is the face structure.

BME Adjacency

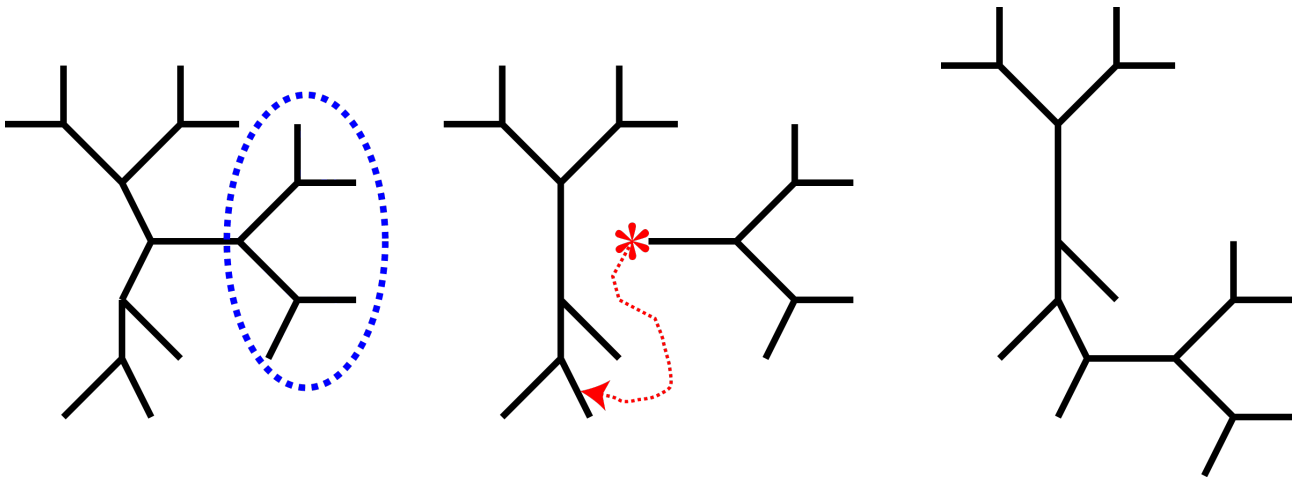
Theorem [H., Hodge, Yoshida]

Given a clade $C \subseteq T \in \mathcal{T}_n$, a $\mathbf{c} \in \mathbb{R}^{\binom{n}{2}}$ can be efficiently computed such that

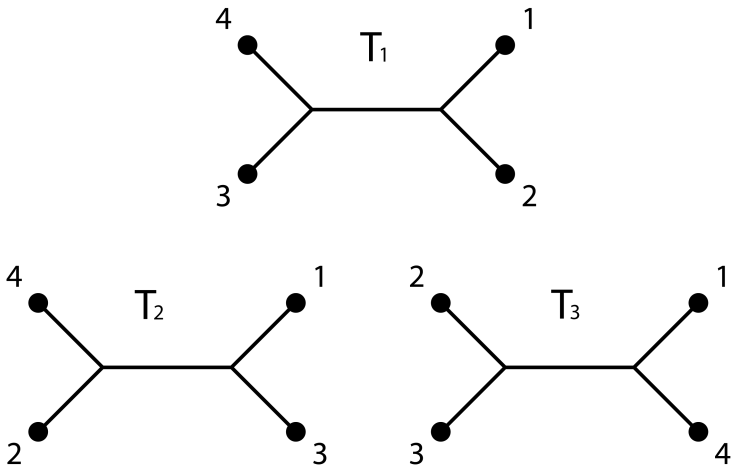
$$\text{argmax}_{T \in \mathcal{T}_n} \mathbf{w}^T \cdot c = \{ T \in \mathcal{T}_n \mid C \subseteq T \}.$$

Theorem [H., Hodge, Yoshida]

If $T', T'' \in \mathcal{T}_n$ are adjacent by a subtree-prune-regraft(SPR) move then T' and T'' are adjacent on the BME polytope $\mathcal{P}_n$.



Cherry Forcing Example



$$\mathbf{w}^{T_1} = \left(\overset{12}{\frac{1}{2}}, \overset{13}{\frac{1}{4}}, \overset{14}{\frac{1}{4}}, \overset{23}{\frac{1}{4}}, \overset{24}{\frac{1}{4}}, \overset{34}{\frac{1}{2}}\right)$$

$$\mathbf{w}^{T_2} = \left(\overset{12}{\frac{1}{4}}, \overset{13}{\frac{1}{2}}, \overset{14}{\frac{1}{4}}, \overset{23}{\frac{1}{4}}, \overset{24}{\frac{1}{2}}, \overset{34}{\frac{1}{4}}\right)$$

$$\mathbf{w}^{T_3} = \left(\overset{12}{\frac{1}{4}}, \overset{13}{\frac{1}{4}}, \overset{14}{\frac{1}{2}}, \overset{23}{\frac{1}{2}}, \overset{24}{\frac{1}{4}}, \overset{34}{\frac{1}{4}}\right)$$

T_1 is uniquely determined by specifying either $\{1,2\}$ or $\{3,4\}$ as a cherry. Let $\mathbf{c} = (1,0,0,0,0,0)$ and note

$$\mathbf{w}^{T_1} \cdot c > \mathbf{w}^{T_2} \cdot c = \mathbf{w}^{T_3} \cdot c.$$

Idea: Given a tree $T \in \mathcal{T}_n$, iteratively force "cherries" of T by settings corresponding entries of $\mathbf{c}$ to smaller and smaller values, similar to a moment curve.

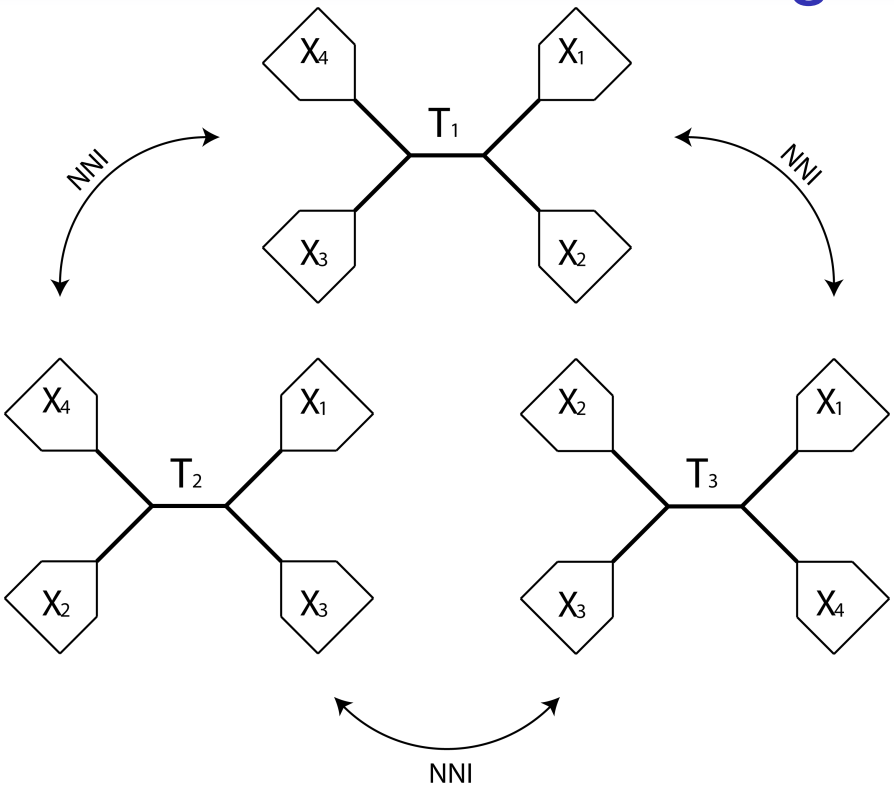
Cherry Forcing Algorithm

Let $\widehat{T} \in \mathcal{T}_n$ and $K > 0$. (I will do example for n=7 on the board.)

Let $T' := \widehat{T}$.
Let $\widetilde{K} := K$.
Let $\mathbf{c} := \mathbf{0} \in \mathbb{R}^{\binom{n}{2}}$.
repeat
    Pick a cherry $\{k, l\}$ of T'.
    Let i be a shallowest leaf in subtree k of $\widehat{T}$.
    Let j be a shallowest leaf in subtree l of $\widehat{T}$.
    Let $\mathbf{c}_{ij} := \widetilde{K}$.
    Let $\widetilde{K} := \frac{1}{2}\mathbf{w}_{ij}^{\widehat{T}}\widetilde{K}$.
    Let $T' :=$ binary tree of T' where leaves k and l are amalgamated to one leaf.
until T' is the star tree on three leaves.
return c and $\widetilde{K}$.

▸ Example

Nearest Neighbor Interchange

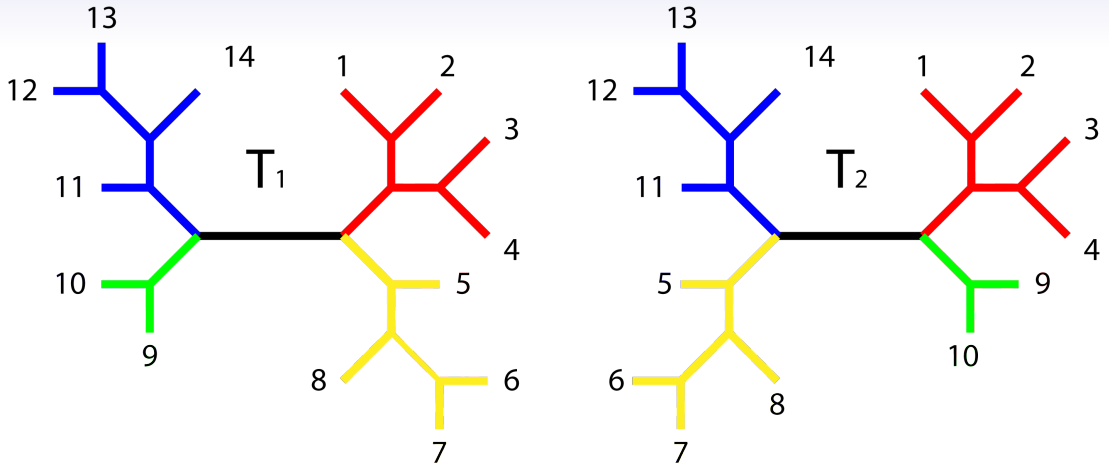


Example of an NNI move where X_1, X_2, X_3, and X_4 are subgraphs (clades) of trees T_1, T_2, and T_3.

NNI $\subset$ SPR.

NNI Adjacent $\Longrightarrow$ BME adjacent.

Let T_1 and T_2 be as above. Use the Cherry Forcing Algorithm (CFA) four times, with the output $\widetilde{K}$ as input K for the next run, to find $\mathbf{c}$ such that subgraphs X_1, X_2, X_3, X_4 are fixed when optimizing $\mathbf{w}^T \cdot \mathbf{c}$. Finally, pick the deepest leaf i in X_1 and deepest leaf j in X_4. Set $c_{i,j} = -\widetilde{K}$, where $\widetilde{K}$ is the output of the last run of the CFA.



Consider T_1 and T_2 above and $K = 1$. Then we set $\mathbf{c} \in \mathbb{R}^{\binom{14}{2}}$, where

$\mathbf{c}_{1,2} = 1$, $\mathbf{c}_{3,4} = \frac{1}{2^2}$, $\mathbf{c}_{1,3} = \frac{1}{2^4}$,

$\mathbf{c}_{6,7} = \frac{1}{2^8}$, $\mathbf{c}_{6,8} = \frac{1}{2^{10}}$, $\mathbf{c}_{5,8} = \frac{1}{2^{13}}$,

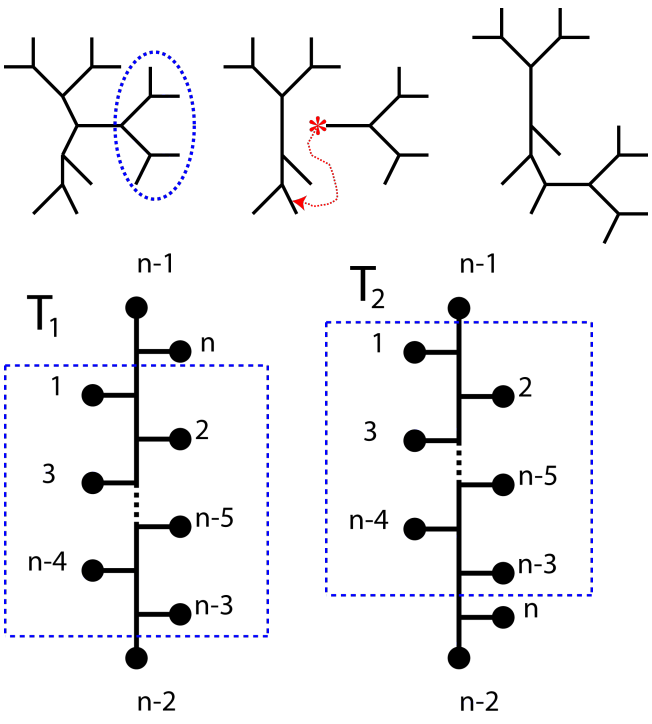
$\mathbf{c}_{9,10} = \frac{1}{2^{16}}$,

$\mathbf{c}_{12,13} = \frac{1}{2^{18}}$, $\mathbf{c}_{12,14} = \frac{1}{2^{20}}$, $\mathbf{c}_{11,14} = \frac{1}{2^{23}}$,

and $\mathbf{c}_{1,12} = -\frac{1}{2^{26}}$.

For $\mathbf{c}$ above, only T_1 and T_2 maximize $\mathbf{w}^T \cdot c$.

SPR $\Longrightarrow$ BME adjacency

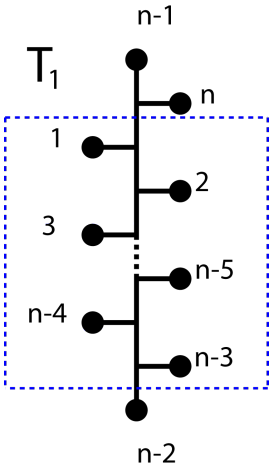


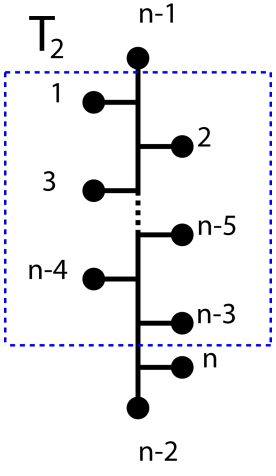
For any $T', T'' \in \mathcal{T}_n$ adjacent by an SPR move, there is a set of clades common to both T' and T''. Using the CFA, the common clades can be fixed. Hence, any SPR move can be reduced to the trees T_1 and T_2 on the left.

A generalized version of the CFA can be roughly described as a hierarchy of rules that say either two leaves $\{i, j\}$ are

- as close as possible, or
- as far away as possible.

SPR $\Longrightarrow$ BME adjacency





T_1 and T_2 will be the only optimums of $\mathbf{w}^T \cdot \mathbf{c}$ if, in this order, $\mathbf{c}$ is chosen such that

- leaves $n-1$ and $n-2$ are as far apart as possible,
- leaf n is as close as possible to $n-1$ or $n-2$ (equally weighted),
- leaves 1 and $n-3$ are as far apart as possible,
- leaves 1 and $n-4$ are as far apart as possible,
- $\vdots$
- leaves 1 and 2 are as far apart as possible,
- leaf pairs $\{n-2, n-3\}$ and $\{1, n-1\}$ are as close as possible (equally weighted).

Future Work

Can we extend to describe faces of BME? E.g, the three trees adjacent by an NNI move make a triangular two-face.

What other adjacencies are there? Investigating non-circular adjacency.

Is tree-bisect-regraft also a BME edge?

What does the Neighbor Joining algorithm do on CFA vectors? Known that NJ is a greedy optimization over BME.

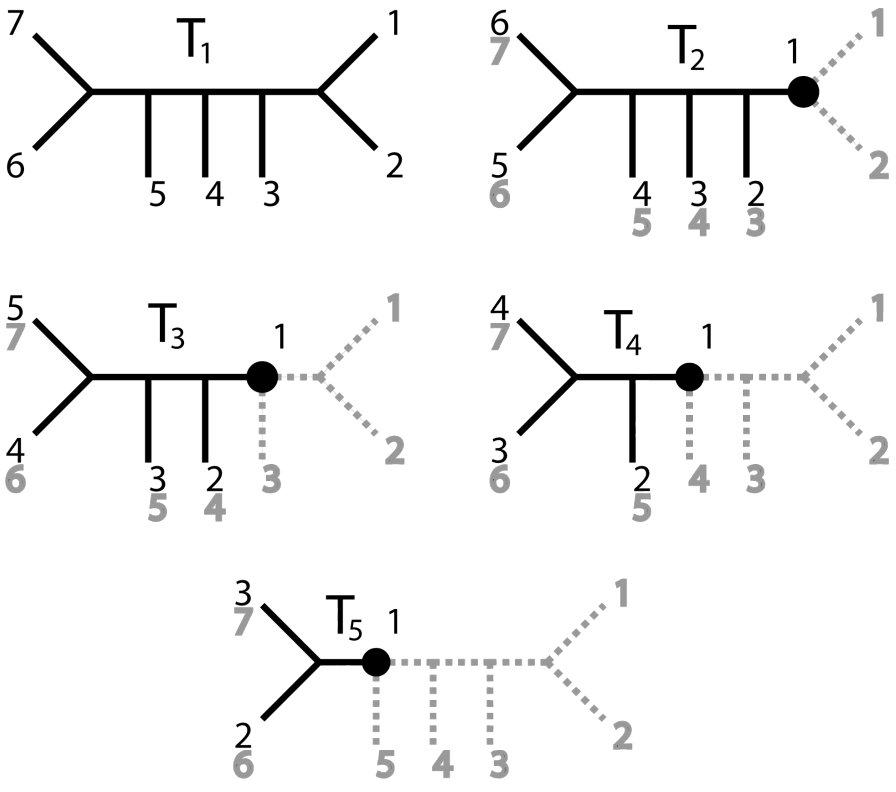
Thank you for your attention.

Edges of the Balanced Minimum Evolution Polytope

David Haws

Department of Statistics
University of Kentucky

Cherry Forcing Algorithm in Action



Cherry Forcing Algorithm would output

$$\mathbf{c} \in \mathbb{R}^{\binom{7}{2}}$$

where

$$c_{1,2} = 1,$$

$$c_{1,3} = \frac{1}{4},$$

$$c_{3,4} = \frac{1}{32},$$

$$c_{4,5} = \frac{1}{128},$$

and otherwise $c_{i,j} = 0$.

▸ CFA