Distance Methods for Phylogeny Estimation

Richard Desper
NCBI/NLM/NIH/DHHS
Bethesda, MD, USA

Outline

I. Metrics and Tree Metrics

II. Algebraic framework
- Split metrics – standard basis
- Split average distances – new basis

III. Common phylogeny estimation methods
- Least squares methods
- OLS Minimum Evolution
- Neighbor Joining

IV. FastNNI and FastME

Outline

IV. Balanced Minimum Evolution
- Topological averaging
- Pauplin's tree length formula

VI. Algebra of BME

VII. Consistency of Balanced Minimum Evolution

VIII. Simulations

Metrics

- A metric is a function d on pairs of objects that satisfies the following three rules:
  - $d(x,x) = 0$ for all x.
  - $d(x,y) = d(y,x) > 0$ for
  - For all $x,y,$ and $z,$

$$d(x,z) \leq d(x,y) + d(y,z)$$

- Let $[n] = \{1,2,\ldots,n\}$. Let $\mathcal{A}_n$ be the vector space generated by metrics on $[n]$. I.e., $\mathcal{A}_n$ is the space of symmetric matrices with zeros along the main diagonal.

Tree metrics

For the purposes of this discussion, we will use the word "topology" to refer to a tree without branch lengths, and "tree" will only be used for trees with lengths assigned to each branch.

Let T be a tree and l be the branch length function. For each two nodes x and y, let p_{xy} be the unique path from x to y in T.

Define $$d^T(x,y) = \sum_{e \in p_{xy}} l(e).$$

Phylogeny estimation

- Our version of the phylogeny estimation problem. Given
  - an unknown tree T_1 with leaf set $[n]$
  - a matrix Δ of estimates of D^{T_1}
- Find the tree T_2 such that D^{T_2} is a good estimator for Δ (and thus of D^{T_1}).

Algebraic structure of tree metrics

Splits

- Let $[n]$ be the leaf set of a tree T. Every edge e defines a *split* of T, $X_e|Y_e$, a bipartition of $[n]$ such that every path from X_e to Y_e includes the edge e.
- Let $\mathcal{S}(T) = \{X_e \mid Y_e : e \in E(T)\}$.

•Suppose $X \mid Y \in \Sigma(T)$ Define the *split metric* $\varepsilon^0_{X|Y}$ by $\varepsilon^0_{X|Y}(u,v) = 1$ if $\left|\{u,v\} \cap X\right| = 1$

$= 0$ otherwise

Split Metrics

- Any tree topology T is determined by the set of splits determined by its edges.
- Let $B_0(\mathrm{T}) = \{\varepsilon^0_{X|Y} : X \mid Y \in \Sigma(\mathrm{T})\}$ be the set of split metrics for the topology T.
- Let $A(\mathrm{T})$ be the vector space generated by $\mathrm{B}_0(\mathrm{T})$
- Note $\dim(\mathrm{A}(\mathrm{T})) = 2n-3$, while $\dim(A_n)=n(n-1)/2$, thus $A(T) \subset A_n$
- (It is important to note that vector spaces allow negative branch lengths, which are biologically meaningless.)

Topological matrices

We can express the tree metric d^T as a vector. Index the branches of T : $e_1, e_2, e_3, \ldots e_m$. Let $p_1, p_2, \ldots, p_{C(n,2)}$ be an enumeration of the leaf-to-leaf paths of T, where $C(n,2) = \frac{n(n-1)}{2}$

Define the matrix S by $s_{ij} = 1$ if $e_i \in p_j$,

$= 0$ otherwise.

Let L be the vector of branch lengths. Then $D^T = SL$. Equivalently, $D^T = \sum_i l(e_i)\varepsilon^0_{X_i|Y_i}$

Average Distance Functionals

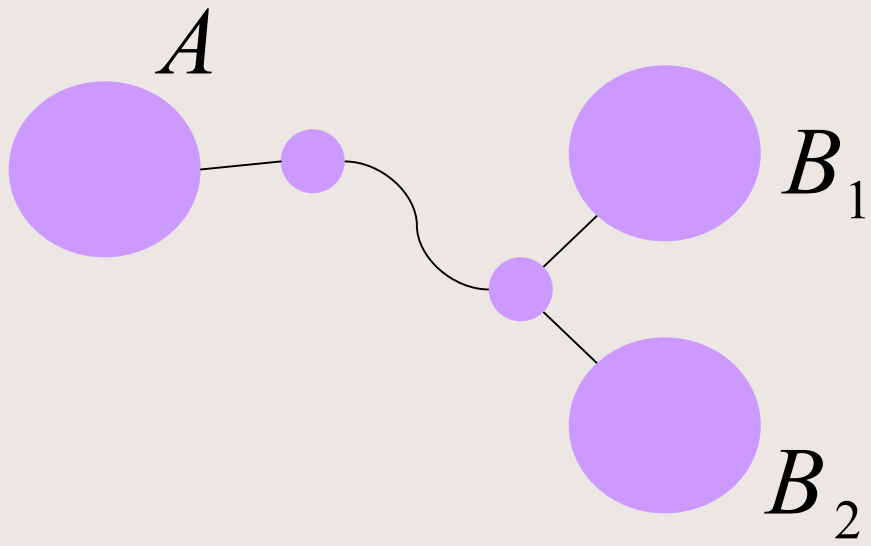
- For any A,B disjoint subsets of $[n]$, let

$$D_{A|B} = \frac{1}{|A||B|} \sum_{a \in A, b \in B} d_{ab}$$

- If we let A and B range over the subtrees of a given tree T, this quantity can be calculated recursively:
  - if $A = \{a\}$, and $B = \{b\}$, then $D_{A|B} = d_{ab}$

Weighted Average Distances

Considering A, B subtrees:



For $B = B_1 \cup B_2$,

$$\Delta_{A|B} = \frac{|B_1|}{|B_1| + |B_2|} \Delta_{A|B_1} + \frac{|B_2|}{|B_1| + |B_2|} \Delta_{A|B_2}$$

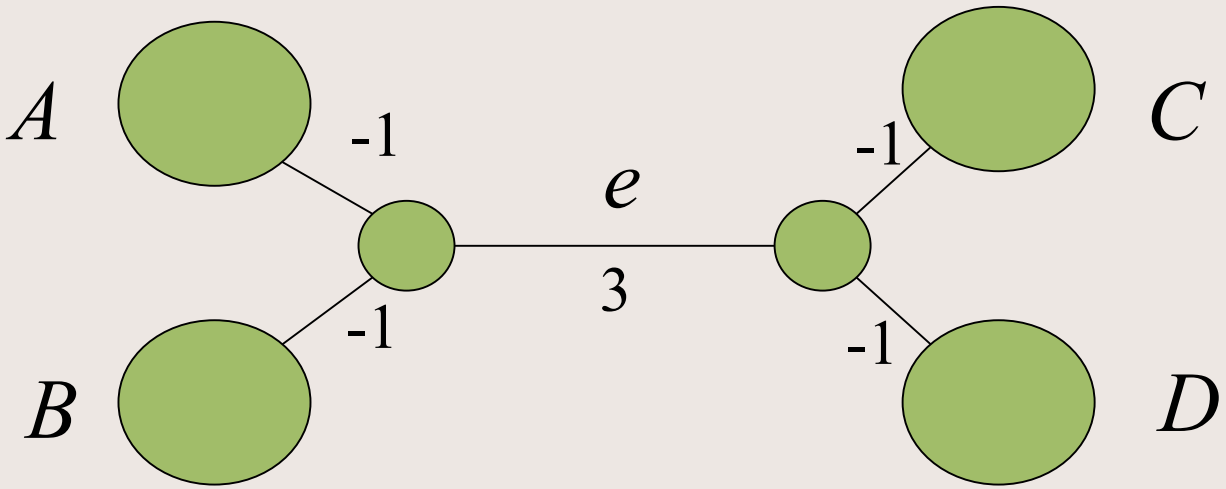
Algebra

- Let $X_i|Y_i$ be the split corresponding to the edge e_i. Suppose we choose $x \in X, y \in Y$ at random from X_i and Y_i respectively. Consider the edge e_j. Define $p_{ij} = \Pr[e_j \in p_{xy}]$
- Let $\mathbf{P} = (p_{ij})$. $\mathbf{P}$ relates the branch lengths of T to the vector $\Delta_{avg} = (\Delta_{X_i|Y_i})$
- $\mathbf{P}$ is invertible. (Desper and Vingron, 2002) Invertibility was demonstrated by showing trees $T^j = \varepsilon^1_{X_j|Y_j}$ such that $D^{T_i}_{X_j|Y_j} = 1$ if $i = j$

$= 0$ otherwise

Example

Suppose e is an internal edge separating four subtrees of the same size, with all edges in the subtrees having zero length, and other edges having lengths:



The tree above is $\varepsilon^{1}_{A\cup B|C\cup D}$

Common Phylogeny Estimation Methods

Least Squares Fitting

- The *fit* of a tree T to a matrix Δ is defined to be

$$\text{fit}(T) = \sum_{i,j} \frac{(d_{ij}^T - \delta_{ij})^2}{\sigma_{ij}^2}$$

- Least-squares fitting seeks the weighted tree (of any topology) minimizing fit(T). (Fitch and Margoliash 1967)
- If $\sigma_{ij} = 1$ for all i and j, this method is called ordinary least-squares, otherwise it is called weighted least-squares.

Average Distances and OLS

- T is OLS tree iff (Vach 1989)

$$D^{T}_{X|Y} = \Delta_{X|Y} \text{ for all } X \mid Y \in \mathcal{S}(T)$$

- This observation leads to branch length formulae for edges in terms of average distances. The formulae are used by Bryant and Waddell's OLS algorithm.

Least Squares Fitting

- Solving ordinary least squares is equivalent to minimizing $(SL-\Delta)^t(SL-\Delta)$, the solution of which is $L=(S^tS)^{-1}S^t\Delta$. (Cavalli-Sforza and Edwards 1967)
- Weighted least squares requires a diagonal matrix W of weights. In this case, the solution is $L=(S^tWS)^{-1}S^tW\Delta$
- Bryant and Waddell (1998) provided an $O(n^3)$ algorithm for solving WLS for a fixed topology.

Minimum Evolution methods

Minimum evolution methods have two steps:

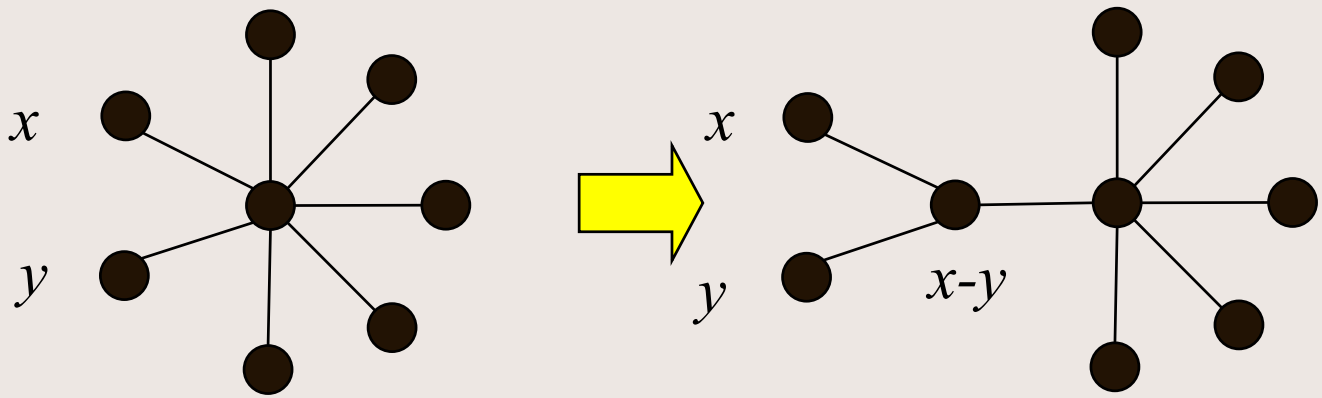
- Each* topology $\mathcal{T}$ is assigned edge lengths according to some function l, for example, the OLS function.
- We choose the topology minimizing

$$l(T) = \sum_{e \in E(T)} l(e)$$

*In practice, not all topologies are examined; rather, a heuristic is used to consider likely topologies.

Neighbor Joining

The neighbor-joining step: We join the neighbors *x* and *y*, and form the new node *x*-*y*.



This tree is assigned edge weights via OLS. NJ uses a minimum evolution criterion to select the smallest tree over all pairs (*x*,*y*).

Neighbor Joining

- The length of the tree pairing x and y is

$$\frac{1}{2(n-2)}\sum_{z\neq x,y}\left(\delta_{xz}+\delta_{yz}\right)+\frac{\delta_{xy}}{2}+\frac{1}{n-2}\sum_{w,z\neq x,y}\delta_{wz}$$

- The neighbors x and y are joined, and a new node x-y is formed. The distance from x-y to the node z is

$$\delta_{x-yz}=\frac{\delta_{xz}+\delta_{yz}}{2}$$

FastME algorithms

OLS version

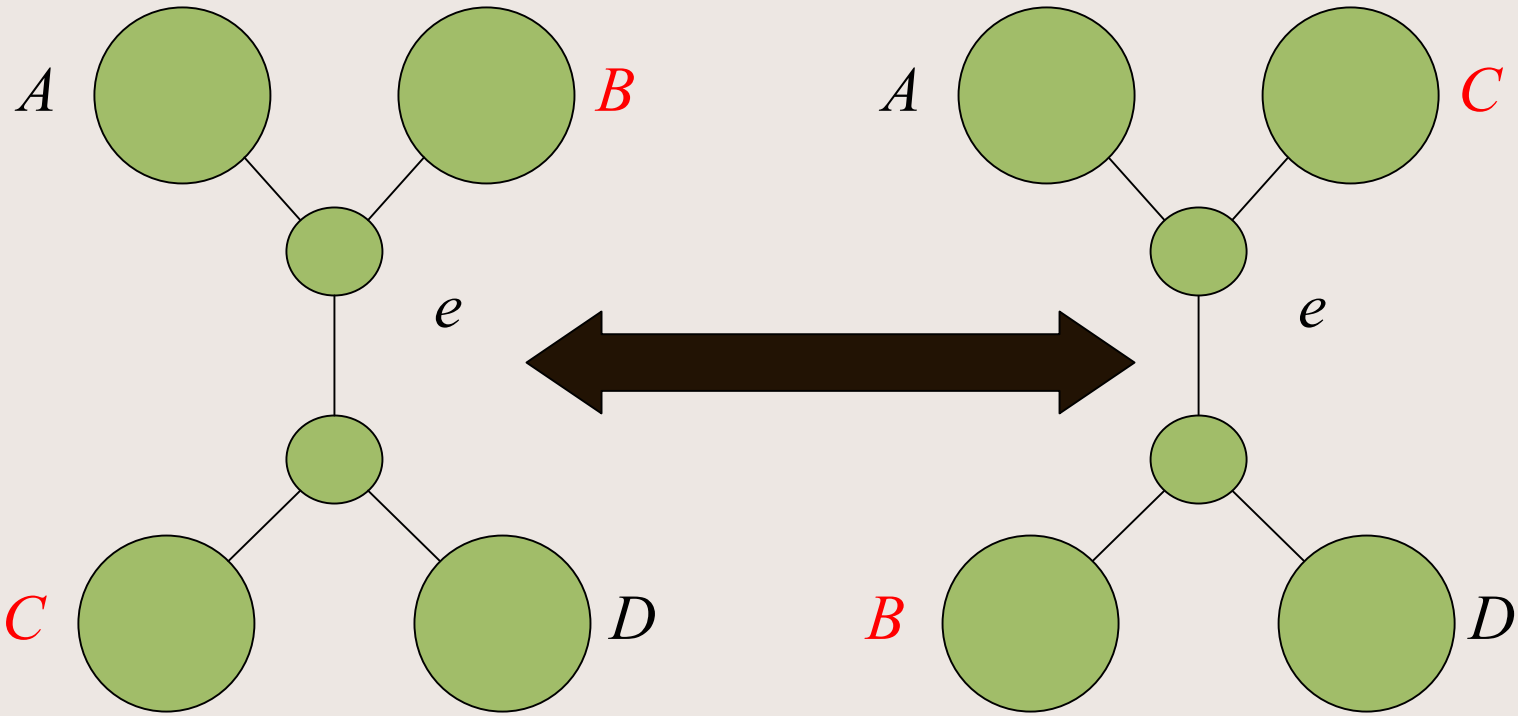
Fast ME algorithms

- The tree length formula depends only on a relatively small number of average distances.
- Small topological changes in a test topology lead to a change in the tree length expressible as a linear sum of a constant number of average distances.
- Maintaining a matrix of appropriate average distances allows for quick calculation of tree lengths for a large number of topologies.

FastNNI

- Input matrix Δ, tree topology $\mathcal{T}$.
- To search the space of topologies, we'll keep in memory :
  - Number of taxa of each subtree
  - Matrix of average distances $\Delta_{X|Y}$ for X, Y disjoint subtrees
- We update the matrix in an efficient manner if/when we select select a new topology.

Tree Swapping by NNI



NNI swapping is a basic step in topology searching

Tree Length after NNI

Given $\mathcal{T} \rightarrow \mathcal{T}'$ the tree swap in prior slide, l the edge length function, T, T' the OLS trees:

$$(1) \quad l(T) - l(T') = \frac{1}{2}\begin{bmatrix} (\lambda - 1)(\Delta_{A|C} + \Delta_{B|D}) \\ -(\lambda' - 1)(\Delta_{A|B} + \Delta_{C|D}) \\ -(\lambda - \lambda')(\Delta_{A|D} + \Delta_{B|C}) \end{bmatrix}$$

where λ and λ' are constants depending on the topologies. (Desper and Gascuel 2002)

OLS:FastNNI

1. Pre-compute average distances between non-intersecting sub-trees. ($O(n^2)$ computations)
2. Loop over all internal edges, select the best swap using Equation (1). ($O(n)$)
3. If no swap improves length of the tree, stop and return the tree, else perform the best swap and update the matrix of average distances and repeat Step 2. ($O(n)$ per swap; there is only one new split.)

Thus, if we require p swaps, the total complexity of FASTNNI is $O(n^2 + pn)$.

FastNNI – Pros and Cons

- Using NNIs leads to a fast algorithm ($O(n^2)$) (Greedy Minimum Evolution) for building an initial topology.
- Even with NNI postprocessing, GME + FastNNI is faster than Neighbor-joining
- Unfortunately, Gascuel (2000) showed that the minimum evolution approach using OLS branch lengths is inferior to NJ in estimating tree topologies.

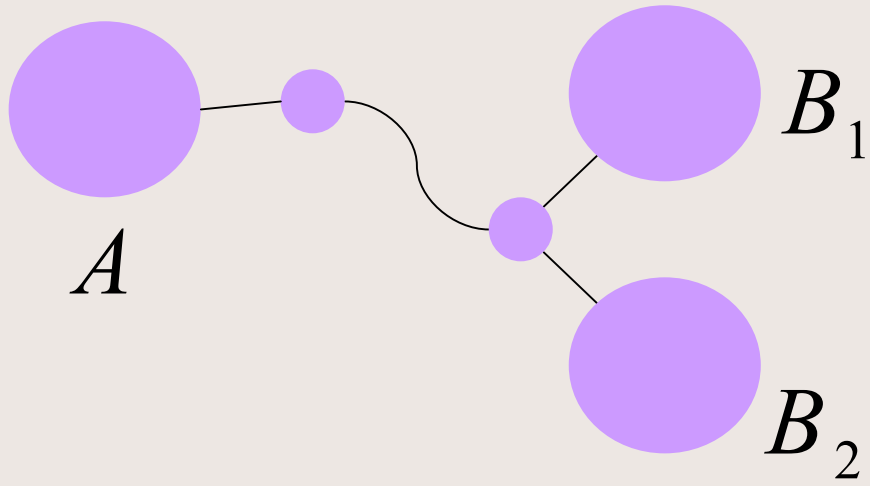
Balanced Minimum Evolution

Balanced Average Distance Functionals

- OLS averages are insensitive to topology: a leaf topologically distant is as important to the calculation of an average as one nearby.
- We'll define "balanced" averages to allow the topology to affect the calculation of average distances. (Pauplin 2000)
- Let Δ be a metric. As A and B range over the subtrees of a given tree T, we'll define $\Delta^T_{A|B}$ recursively:
  - if $A = \{a\}$, and $B = \{b\}$, then $\Delta^T_{A|B} = \delta_{ab}$

Balanced Average Distances

For $B = B_1 \cup B_2$, subtrees of T, we'll define



$$\Delta^{\mathrm{T}}_{A|B} = \frac{1}{2}\Delta^{\mathrm{T}}_{A|B_1} + \frac{1}{2}\Delta^{\mathrm{T}}_{A|B_2}$$

Balanced Averages

- Given Δ and the topology $\mathcal{T}$, we'll select the branch lengths of T to satisfy a Vach-like set of equalities: $D^{\mathrm{T}}_{A|B} = \Delta^{\mathrm{T}}_{A|B}$ for all $A \mid B \in \Sigma(\mathrm{T})$.
- These weights can be found (proof omitted) by solving $L_T = (S_T^t W S_T)^{-1} S_T^t W \Delta$ where the weights are determined by $w_{(ij)} = 2^{1-p^T(i,j)}$,with $p^T(i,j)$ is the topological length of the path in T from i to j.
- As with the OLS tree, each branch length can be expressed as a simply linear sum of average distances. (Simply use $\lambda = \lambda' = ½$ in OLS formulae).

Balanced NNI

1. Calculate balanced averages of all pairs of sub-trees. ($O(n^2)$)
2. Calculate improvement for each swap using

$$(2)\quad l(T) - l(T') = \frac{1}{2}\left(\Delta^{T}_{A|B} + \Delta^{T}_{C|D} - \Delta^{T}_{A|C} - \Delta^{T}_{B|D}\right)$$

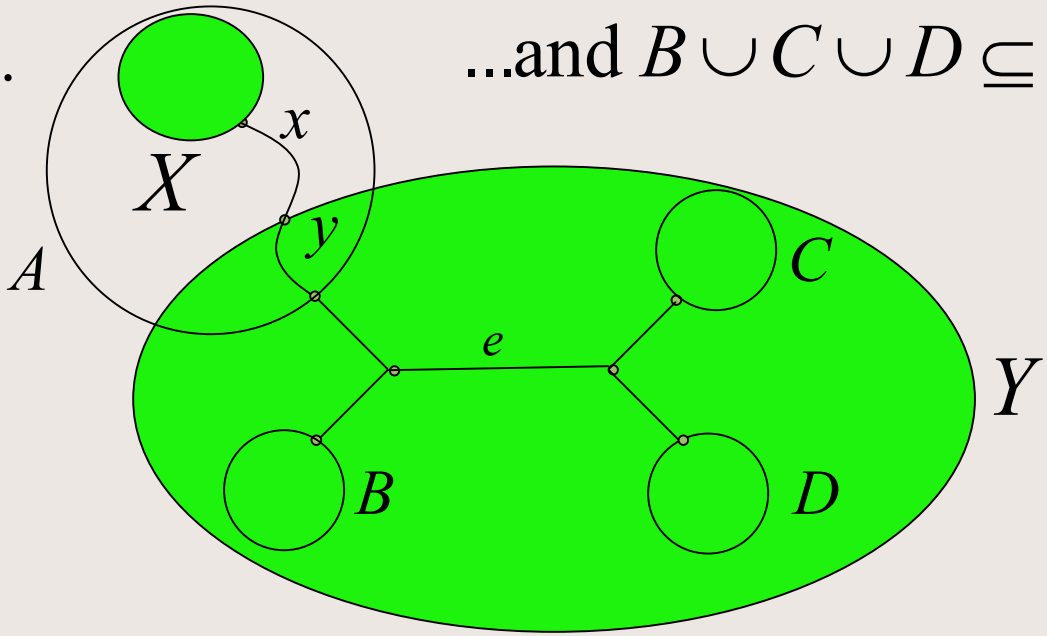
3. If no tree swap improves length of the tree, stop and return tree, else update matrix of average distances and repeat Step 2. ($O(n\ \mathrm{diam}(T))$ per swap)

The average complexity, when performing p swaps, is $O(n^2 + pn\ \mathrm{diam}(T))$.

Updating Subtree Averages

Here, $X \subseteq A$...

...and $B \cup C \cup D \subseteq Y$



If we perform the B-C tree swap, then we must recalculate $\Delta^{T}_{X|Y}$

Q: How many recalculations? A: O(n diam(T))

If T is generated randomly, the expected value of diam(T) can range from $O(\log n)$ to $O\left(\sqrt{n}\right)$

BME: its algebra and WLS connections

Joint work with Olivier Gascuel

BME=BLS

- BME is a weighted least squares approach with

$$\sigma_{ij} = c\,2^{p^T(i,j)}.$$

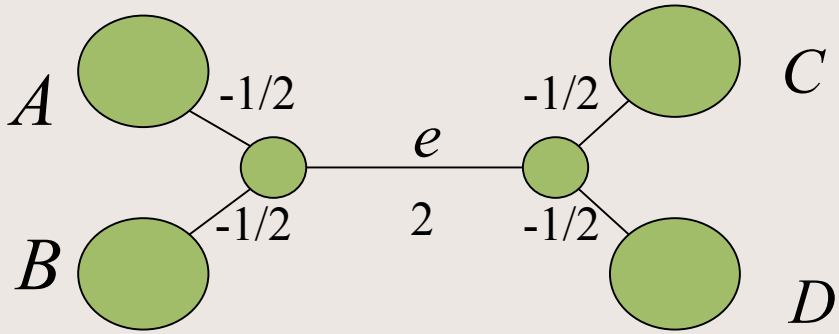
- Standard models of evolution (e.g. Kimura) yield a variance on the estimates of evolutionary distances:

$$\sigma_{ij} \propto e^{d_{ij}}$$

- Presuming evolutionary distances are proportional to topological distances, the BME approach yields a better approximation to variances of evolutionary distances than usual WLS methods.

The Balanced Dual Basis

- As with the OLS setting, we can find basis vectors dual to balanced average distance functionals.
- With branch lengths:



$$D^{\mathrm{T}}_{X|Y} = 1 \text{ if } X \,|\, Y = A \cup B \,|\, C \cup D$$

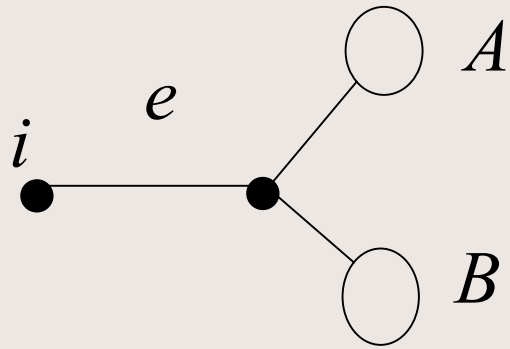
$$= 0 \text{ otherwise, for } X \,|\, Y \in \Sigma(\mathrm{T})$$

The Balanced Dual Basis

- For an external edge e, set $l(e) = 3/2$, $l(f) = -1/2$ for f incident to e, and $l(g)=0$ for all other edges g.
- Again, if $X \mid Y \in \Sigma(\mathrm{T})$

$$D^{T}_{X|Y} = 1 \text{ if } X \mid Y = i \mid A \cup B$$

$$= 0 \text{ otherwise}$$



- Let B_e be the tree with lengths described above or on the previous slide, for any edge e

Pauplin's Formula

- Let T be a weighted tree of topology $\mathcal{T}$ and Δ be a metric. Pauplin's formula for the length of T is $l(T) = \sum_{i<j} 2^{1-p^T(i,j)} \delta_{ij}$.

- Let us decompose D^T according to the dual basis:

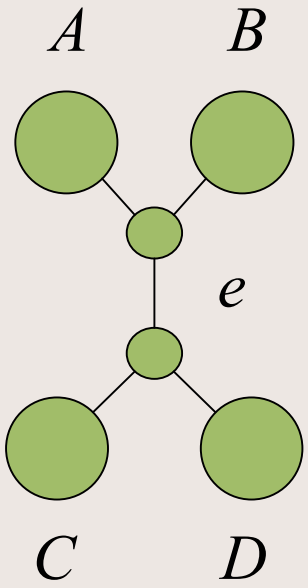
$$D^T = \sum_{X_e|Y_e \in \mathcal{S}(\mathcal{T})} D^T_{X_e|Y_e} \varepsilon^{\mathcal{T}}_{X_e|Y_e},$$

Proof of Pauplin's formula

- By linearity, $l(T) = \sum_{X_e|Y_e \in \mathcal{S}(T)} D^T_{X_e|Y_e} l(B_e).$
- Observe $l(\mathrm{B}_e) = 0$ for e internal, and $l(B_e) =$ ½ for e external. Thus

$$l(T) = \frac{1}{2} \sum_{i \in [n]} D^T_{i|V\setminus i} = \sum_{1 \le i < j \le n} 2^{1-p^T(i,j)} \delta_{ij}.$$

Positive Branch Lengths after BNNI



$$l(e) = \frac{1}{2}\left[\frac{1}{2}(\Delta_{A|C} + \Delta_{B|D} + \Delta_{A|D} + \Delta_{B|C}) - (\Delta_{A|B} + \Delta_{C|D})\right]$$

We do not perform the $B \leftrightarrow C$ switch because

$$l(T) - l(T') = \frac{1}{2}\left(\Delta^{\mathcal{T}}_{A|B} + \Delta^{\mathcal{T}}_{C|D} - \Delta^{\mathcal{T}}_{A|C} - \Delta^{\mathcal{T}}_{B|D}\right) < 0,$$

i.e.

$$\Delta_{A|C} + \Delta_{B|D} > \Delta_{A|B} + \Delta_{C|D}.$$

Similarly,

$$\Delta_{A|D} + \Delta_{B|C} > \Delta_{A|B} + \Delta_{C|D}.$$

Thus

$$l(e) > 0$$

Consistency of BME

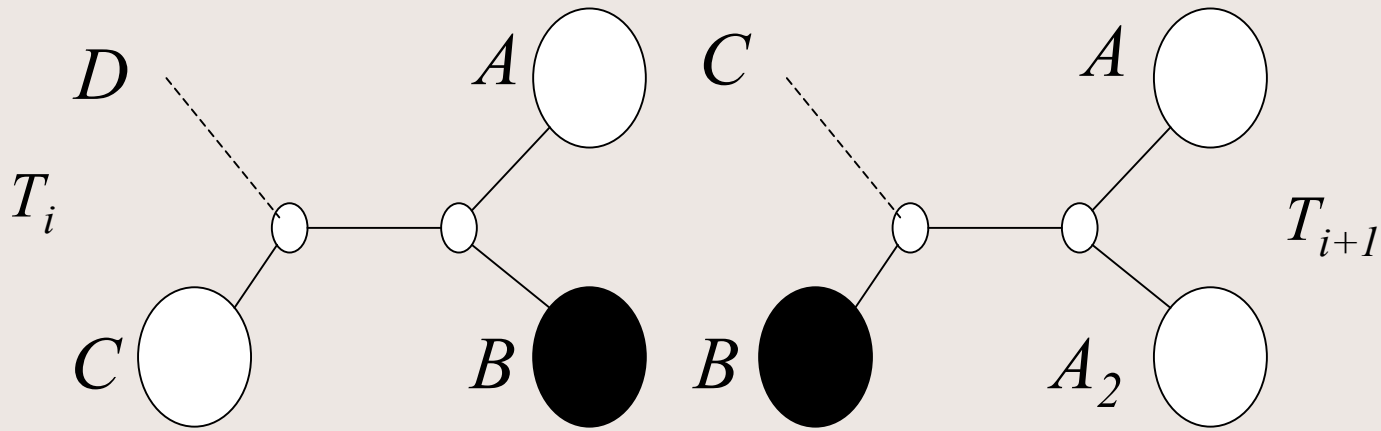
Modeled after OLS/ME
proof of Rzhetsky and Nei (1993)

Balanced ME consistency

- Basic idea: let l be the tree length function on the space of topologies. We find a sequence of topologies, $T=T_0,\ T_1,\ \ldots\ T_k=S$ such that
  - Each T_{i+1} can be reached from T_i via one of two simple topological transformations
  - $l(T_i) > l(T_{i+1})$ for all i.

Type I transformation

Color the leaves black or white according to the split metric given by *S*. A Type I transformation uses a NNI to form a larger monochromatic cluster.

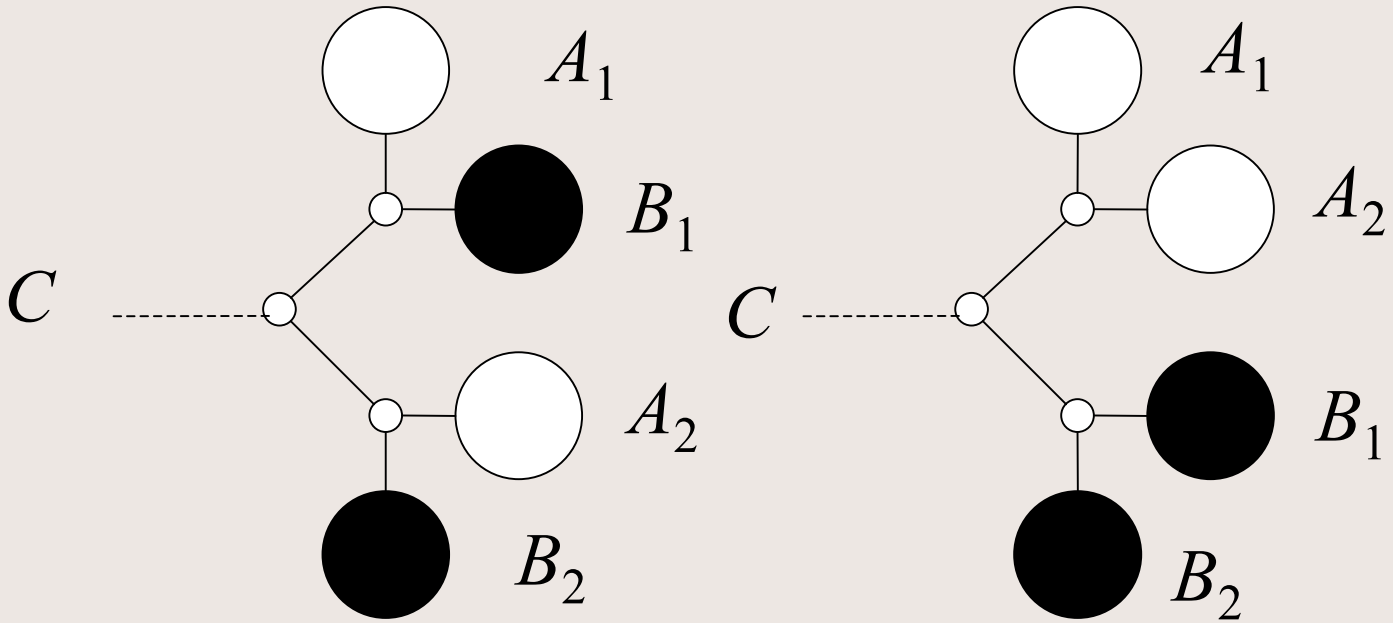


This transformation reduces the size of the tree under *l*

$$l(T_{i+1}) - l(T_i) = \frac{1}{4}\left(\Delta^{T_i}_{B|C} - 1 - \Delta^{T_i}_{A_1|C}\right) < 0$$

A Type II transformation

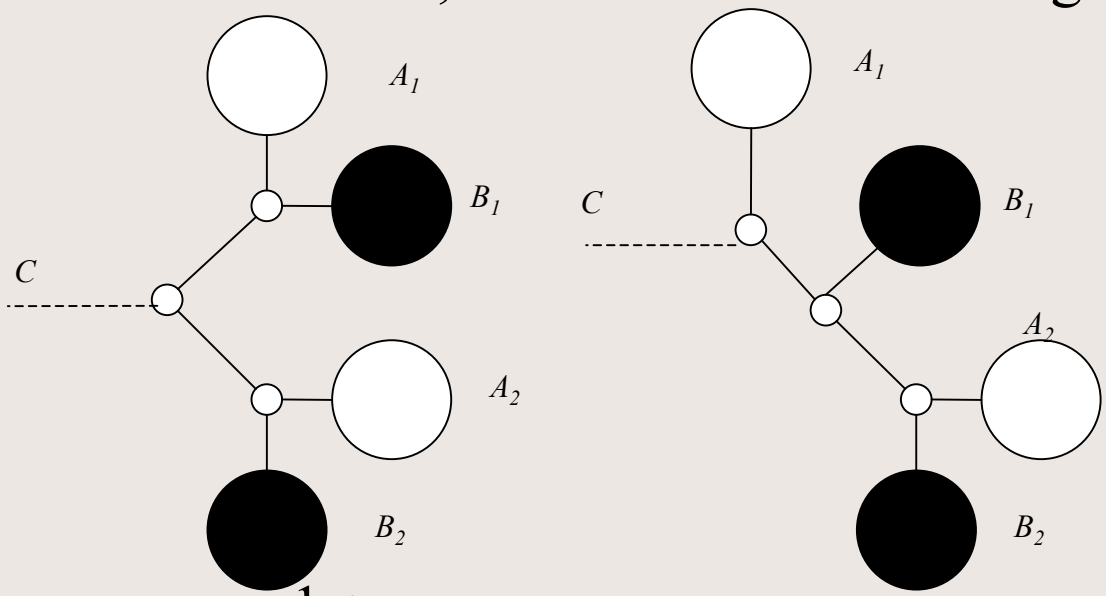
A Type II transformation uses two NNIs to form two monochromatic subtrees



This transformation also reduces the value of the size of the tree under l...

Decomposing a Type II transformation

We use two NNIs to perform a Type II transformation. Let T^i be the tree on the left, T^{*I} be the tree on the right.



$$l(T^{*i}) - l(T^i) = \frac{1}{4}\left(D^{T_i}_{A_1|C} + D^{T_i}_{B_1|A_2\cup B_2} - D^{T_i}_{A_1|B_1} - D^{T_i}_{A_2\cup B_2|C}\right)$$

$$= \frac{1}{4}(p_c - 1),$$

where p_c is the relative weight of black nodes within C.

Simulations

Using Aldous topology generation and covarion model for rate variation

Simulations

- Simulated 5000 trees with 100 taxa each.
- Generated using Aldous distribution on trees, a distribution that includes a Yule-Harding distribution at one extreme and a uniform distribution at the other, with a paramter β determining range between -1.5 and 0.
- Branch lengths determined by a standard coalescent model, and perturbed from ultrametric by multiplying by exponential r.v.
- For each tree, we generated DNA sequences 600 base pairs long. Covarion model for rate variation.
- Used `dnadist` to calculate Jin-Nei maximum likelihood distances for each set of sequences, yielding 5000 matrices.

New results: error functions

We also consider related topological error functions that distinguish the very short edges that are not realistically recoverable. For any $\delta > 0$, and T, T', define

$$e_1(T,T',\delta) = |\{e \in E(T') : l(e) > \delta, X_e \mid Y_e \notin \Sigma(\mathrm{T})\}|$$

$$e_1(T,T',\delta) = |\{e \in E(T) : l(e) > \delta, X_e \mid Y_e \notin \Sigma(\mathrm{T}')\}|$$

With 600 bases in each sequence, we use $\delta = \frac{1}{1200}$

Summary results

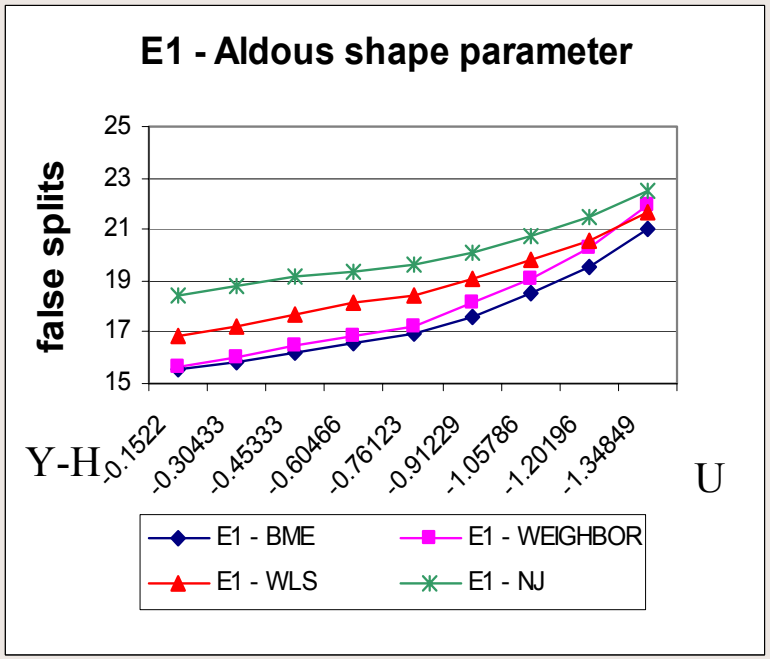
algorithm	RF	e_1	e_2	r_{alg}	r_{obs}
BME	58.06	17.65	9.25	80.25	71.85
Weighbor	61.50	18.10	11.59	78.36	71.85
WLS	62.08	18.91	11.28	79.48	71.85
NJ	64.99	20.09	14.49	77.44	71.85

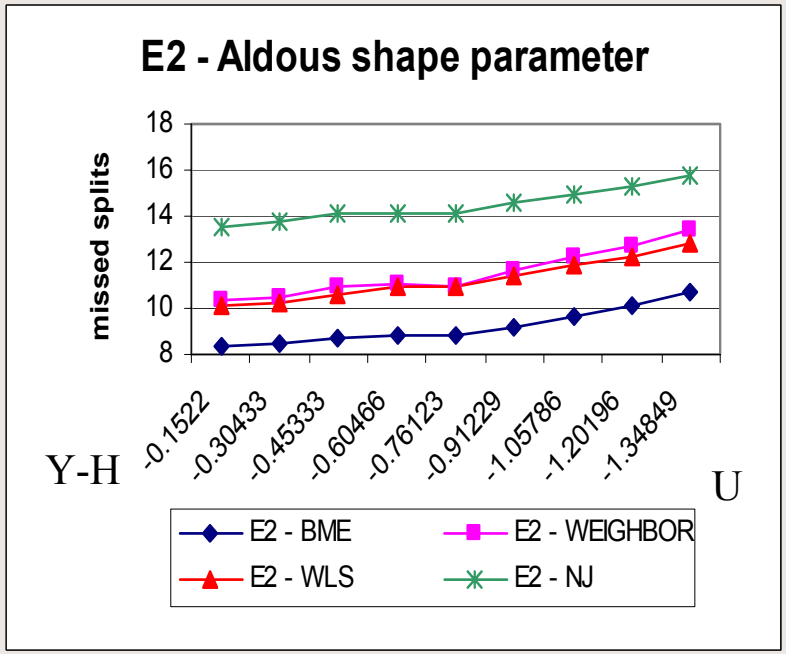
RF is Robinson-Foulds sum of missed and false splits.
r_{alg} and r_{obs} refer to the number of edges longer than δ in the algorithm tree and true tree,respectively

Interval tests

- For each of seven parameters, we sorted tests according to parameter value.
- From sorted lists, we constructed 9 subsets of the data, corresponding to the intervals of the form $[500k+1, 500k + 1000]$, for $0 \leq k \leq 8$
- For each sub-interval, we calculate error and resolution statistics.

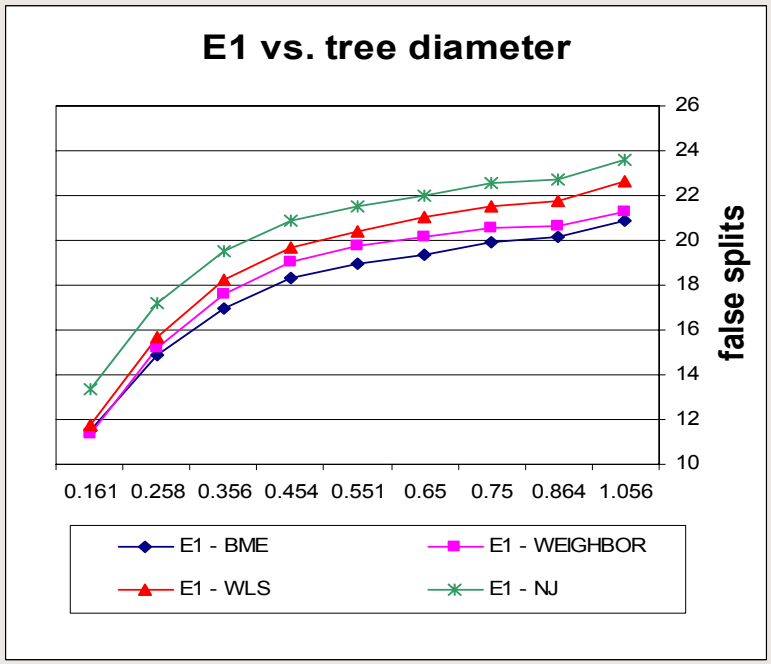
Error functions vs. Beta parameter

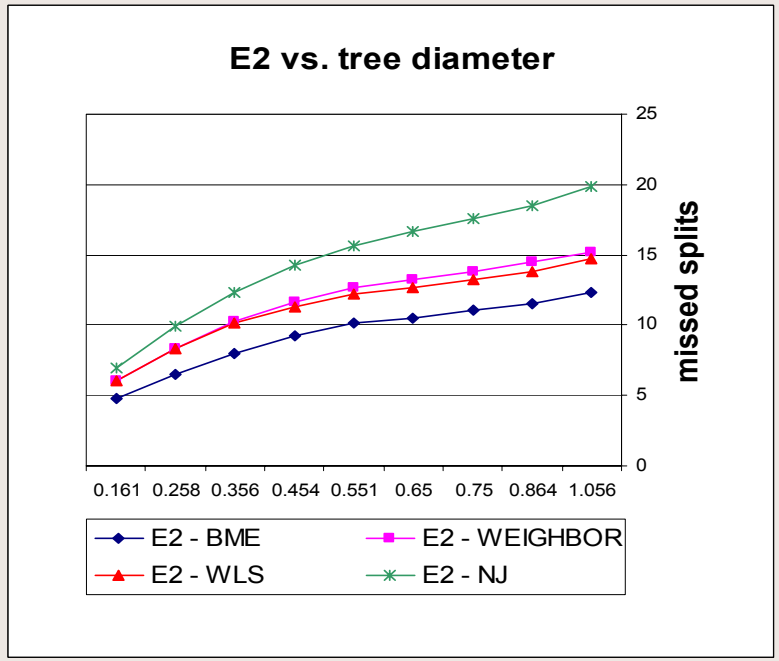




Errors increase as topology distribution moves from Yule-Harding to uniform.

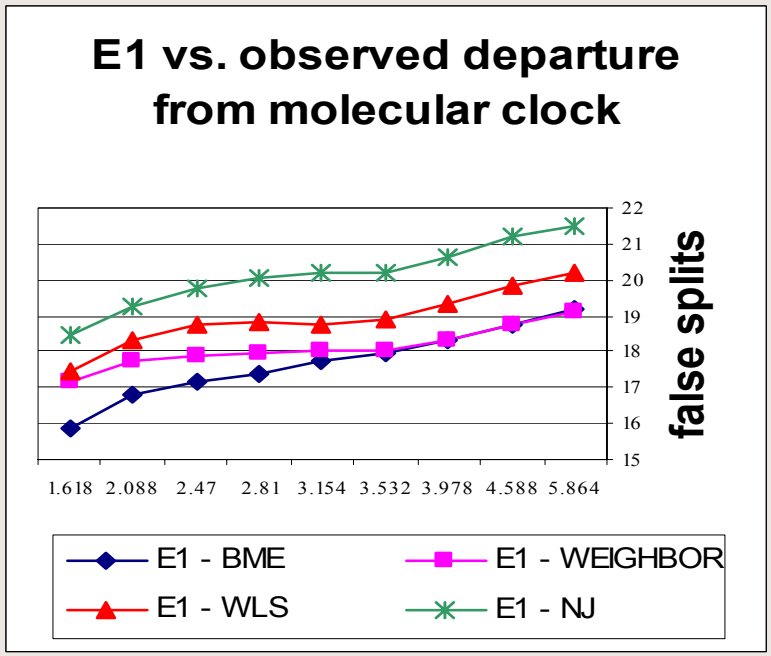
Error functions vs. tree diameter

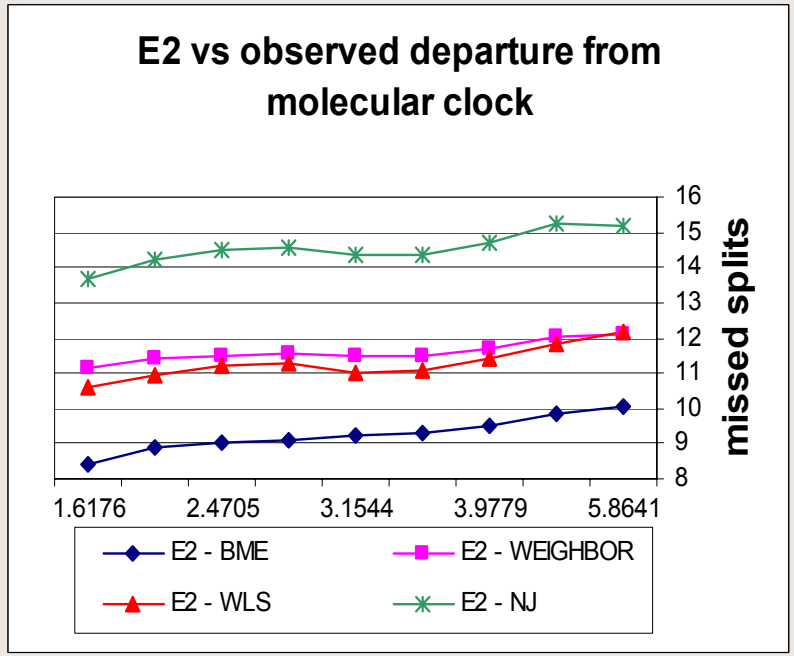




Errors increase with tree diameter.

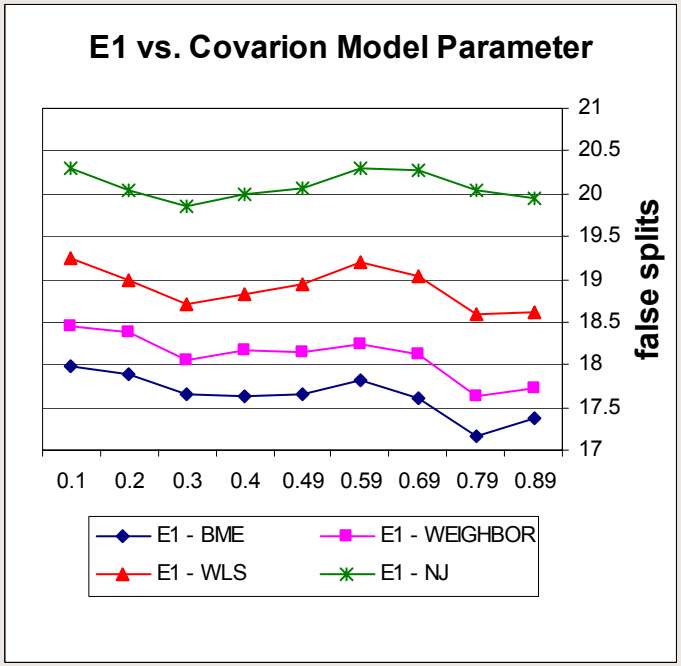
Error functions vs. departure from molecular clock

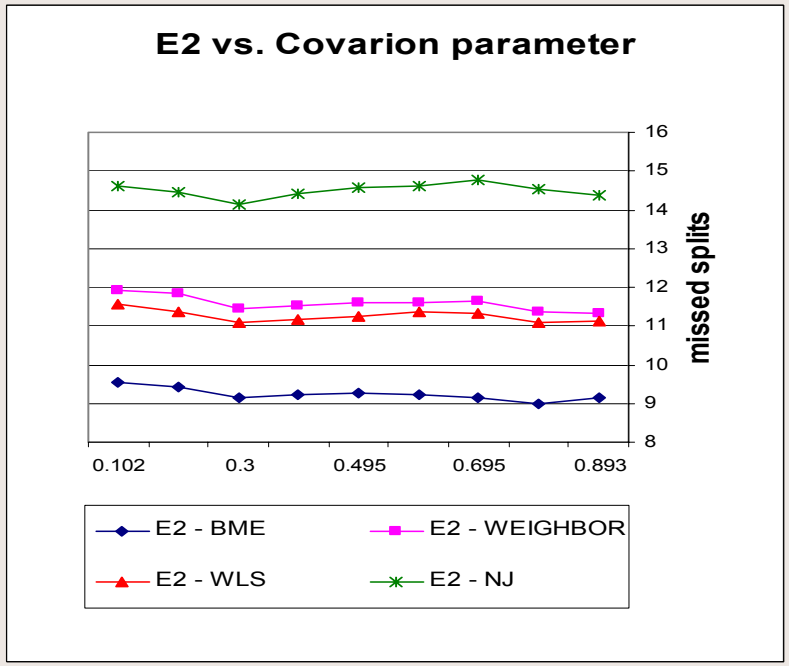




Errors increase with departure from molecular clock.

Error functions vs. covarion parameter





Change in the covarion parameter has little effect

Computational Times

in (MM:SS)

	24 Taxa	96 Taxa	1000 Taxa	4000 Taxa
GME + BNNI	0.0263	0.0842	11.3390	06:02.1
HGT/FP	0.0252	0.1349	13.8080	03:33.1
NJ/BIONJ	0.0630	0.1628	21.2500	20:55.9
WEIGHBOR	0.4244	26.8818		
FITCH	4.3745			

Computations done on Sun Enterprise E4500/E5500 running Solaris 8 on 10 400-Mhz processors with 7 Gb memory.

Conclusions

- BME + BNNI runs in $O((n^2 + pn)\,\mathrm{diam}(T))$, outputs trees better than FITCH, Weighbor, or NJ.
- BNNI outputs tree without negative branch lengths.
- BME approach shown to be consistent.
- All tested methods saw errors increase as shape parameter moved toward uniform distribution.
- All tested methods saw errors increase with increase in divergence from molecular clock, and with tree diameter.
- Changes in covarion parameter had negligible effect.
- FASTME software available at http://www.ncbi.nlm.nih.gov/CBBResearch/Desper/FastME.html or http://www.lirmm.fr/~w3ifa/MAAS/.

References

- Bryant, D., and Wadddell, P. 1998. Rapid evaluation of least-squares and minimum-evolution criteria on phylogenetic trees. Mol. Biol. Evol. **15**:1346-1359.
- Desper, R., and Gascuel, O. 2002. Fast and accurate phylogeny reconstruction algorithms based on the minimum evolution principle. J. Comp. Biol. **9**:687-705.
- Desper, R., and Vingron, M. 2002. J. Classification. **19**:87-112.
- Fitch, W.M., and Margoliash, E. 1967. Construction of phylogenetic trees. Science **155**:279-284.
- Pauplin, Y. 2000. Direct calculation of a tree length using a distance matrix. J. Mol. Evol. **51**:41-47.
- Rzhetsky, A., and Nei, M. 1993. Theoretical foundation of the minimum-evolution method of phylogenetic inference. Mol. Biol. Evol. **10**:1073-1095.
- Vach, W. 1989. Least squares optimization of additive trees. Pp. 230-238 in O. Opitz, etd. Conceptual and numerical analysis of data. Springer-Verlag, Berlin.