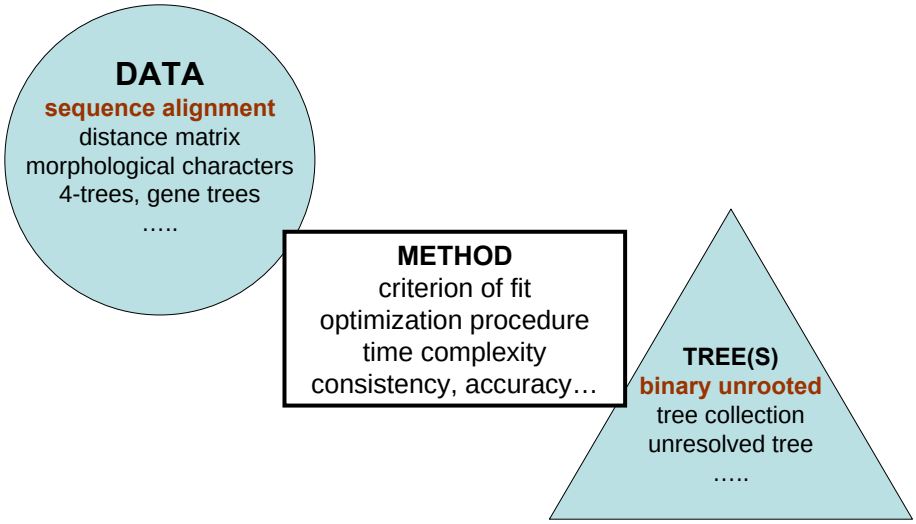
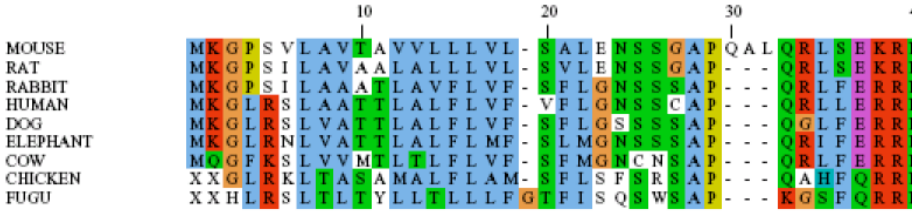
Arbres formels et Arbre(s) de la Vie

- A bit of history and biology
- Definitions
- Numbers
- Topological distances
- Consensus
- Random models
- **Algorithms to build trees**

Basic principles



Multiple alignments



- Protein alignment (more and more used)
- Computed thanks to score matrices that reflect the biochemical properties of amino acids
- Popular programs: Clustal, Muscle, Mafft, Tcoffee

Criterion of fit

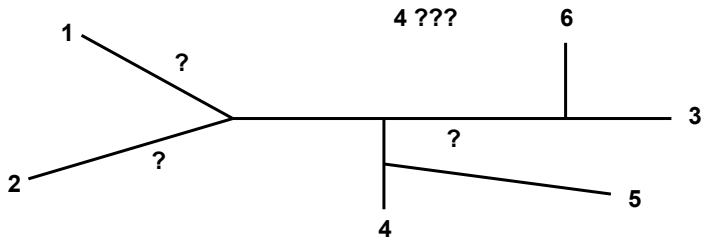
- Measures the fitness of the tree under construction to the data
- Number of compatible morphological characters
- Number of compatible 4-trees
- Parsimony with (DNA) sequence alignments
- Least square and tree length with distances
- ……

Optimization procedure

- An exponential number of trees
- Most (if not all) criteria impose an exponential computing time to find the best tree(s)
- We have to use heuristic procedures, to find nearly optimal trees in acceptable run times
- Solution 1: greedy tree construction
- Solution 2: iterative tree amelioration using topological moves (NNI, SPR, TBR…)
- Both solutions are usually combined

Greedy tree construction: iterative taxon addition

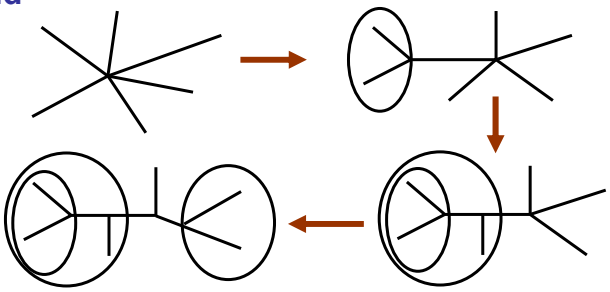
- Taxa are inserted one after the other in a growing tree
- The insertion branch maximizes the criterion value



- The resulting tree depends on taxon ordering
- Some programs provide a jumbling option to use several random orders

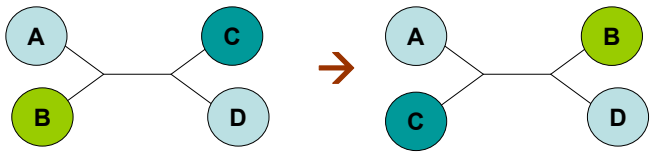
Greedy tree construction: neighbor joining

- We start from a star tree and iteratively join taxa-subtree pairs until we obtain a fully resolved tree
- O(n^2) choices at each step, selected using criterion at hand



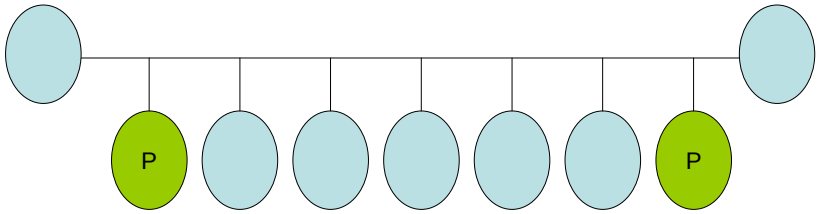
Tree amelioration using topological moves: NNIs

- We start from a (reasonnable) complete binary tree
- Two NNIs (nearest neighbor interchanges) per internal branch, O(n) total
- For each, we compute the criterion value and compare to the current value
- When an improving move is found, it is achieved and tree search is continued; else we return the current tree



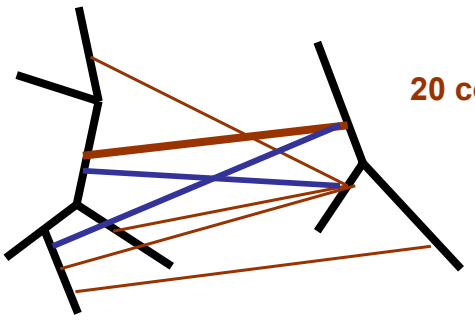
- NNIs are fast, but limited in some cases

Subtree pruning and regraft (SPR)



- O(n) SPRs per subtree, O(n^2) in total
- NNIs are special SPRs
- Any SPR can be achieved using a series of NNIs
- Usually sufficient for thorough tree space exploration

Tree bisection and reconnection (TBR)



20 combinations

- O(n^2) possibilities per split, O(n^3) in total
- SPRs are special TBRs
- 1 TBR = 2 SPRs
- Heavy search, but usefull for parsimony

Time complexity

- Running times depend on the size of the data: *n* taxa, *s* sites, σ: alphabet size …
- O(*f*(*parameters*)) basically means that the running time is proportionnal to *f*(*parameters*)
- For example, (original) NJ is in O(n^3), computing all distances is in O(sn^2), and thus the complete tree building procedure in O($n^3 + sn^2$)
- With NP-hard problems (e.g. MP, ML…) finding the best tree(s) has worst case running times in O(e^n)
- O(n^2) → 100.000 taxa tree, O(n^3) → 10.000, O(n^4) → 500, O(n^5) → 100…

Olivier Gascuel – Arbres formels et Arbre de la Vie – Conférence ENS Cachan, septembre 2011

Consistency, accuracy

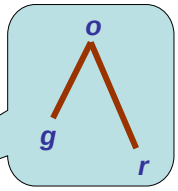
- Being fast is not sufficient…
- Consistency: garanty to reconstruct the correct tree with perfect data (not sufficient, but necessary)
- Accuracy: measured using computer simulations

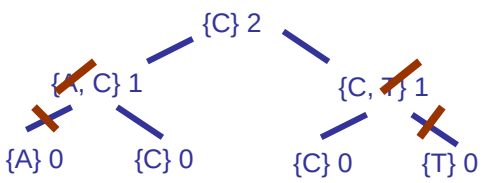
Parsimony criterion and methods

- We assume that (multiple) substitutions are relatively rare and uniformly distributed among sites and tree branches
- We then search for the most parsimonious tree, i.e. the tree requiring the smallest number of substitutions to explain the sequences

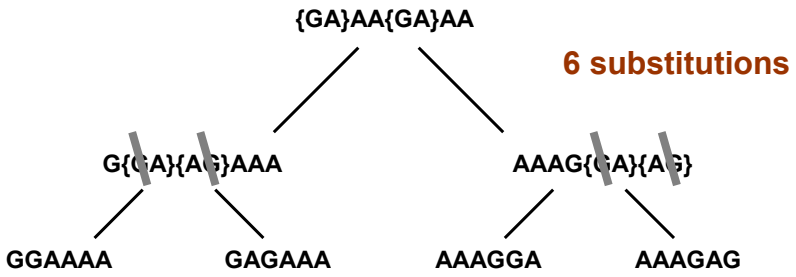
Computing the parsimony-score of a tree

- Computed separately for every site and recursively
- Does not depend on the root position
- For each site, each node *o* is associated to a set of possibilities *So* and a parsimony value *Po*
- If *o* = leaf, then $So = \{Xo\}$ and $Po = 0$
  Elseif $Sg \cap Sr = \emptyset$, then $So = Sg \cup Sr$ and $Po = Pg + Pr + 1$
  else $So = Sg \cap Sr$ and $Po = Pg + Pr$





Computing the parsimony-score of a tree

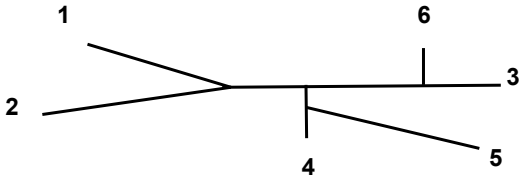


This computation requires O(*ns*) time

Heuristic approaches

Iterative taxon addition

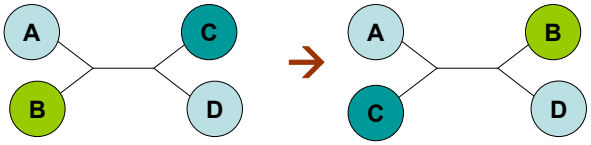
- When adding taxon *j* we have (2*j* – 5) possibilities
- The time complexity is $\sum_{j=4}^{n} sj(2j-5) = O(sn^3)$
- Can be implemented in O(sn^2)



Heuristic approaches

NNIs

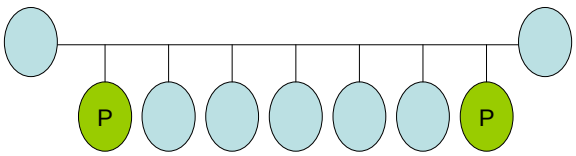
- O(*n*) possible NNIs
- Finding the best one requires O(*n* x *sn*) time
- Can be implemented in O(*sn*)
- In total O(*ksn*), with *k* iterations (~*n*)



Heuristic approaches

SPRs

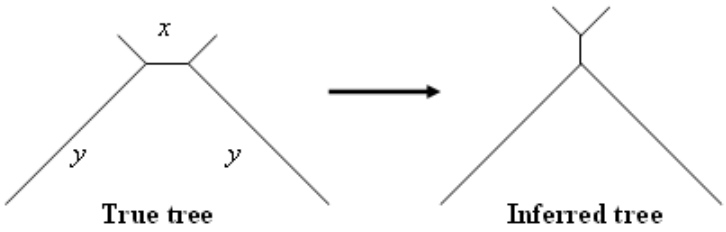
- O(n^2) possible SPRs
- Finding the best one requires O(n^2 x *sn*) time
- Can be implemented in O(sn^2)
- In total O(ksn^2) ($k < n$)

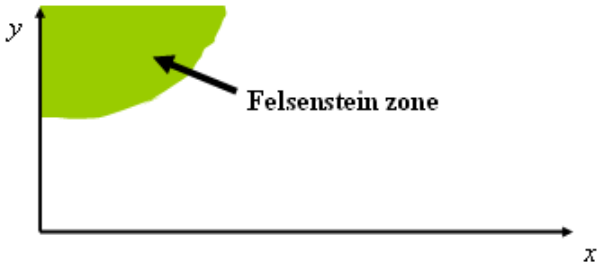


Parsimony, summary, properties

- Thorough heuristics (TBR, ratchet, tree merging)
- Relatively fast (say $O(sn^3)$), efficient implementations (bit vectors, TNT…), making it possible to deal with very large alignments
- Accurate when the basic assumptions are fulfilled, i.e. when substitutions are rare and uniformly distributed among sites and branches
- But inconsistent and subject to long-branch attraction

Parsimony, the Felsenstein zone



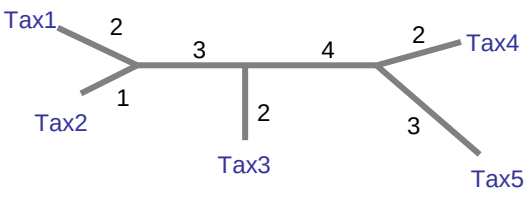


Distance methods

- A set of aligned sequences
- Gapped regions are removed
- Estimation of all pairwise distances between sequences
- We obtain a distance matrix $\Delta = (\delta_{ij})$
- A distance-based algorithm is run to infer a tree

The rationale of distance methods

- The number of substitutions separating any taxon pair defines an additive distance, which is the evolutionary distance
- Knowing this distance, tree reconstruction would be easy



$$\begin{pmatrix} 0 & 3 & 7 & 11 & 12 \\ 3 & 0 & 6 & 10 & 11 \\ 7 & 6 & 0 & 8 & 9 \\ 11 & 10 & 8 & 0 & 5 \\ 12 & 11 & 9 & 5 & 0 \end{pmatrix}$$

The rationale of distance methods

- But hidden substitutions occurred and the number of substitutions is higher than the number of observed differences
- The evolutionary distance (number of substitutions per site) is estimated thanks to models (e.g. JC69) using analytical formulae or ML
- Standard alignments have a high tree signal:

$$\frac{\sum\left(\delta_{ij} - t_{ij}\right)^2}{\sum\left(\delta_{ij} - \bar{\delta}\right)^2} \approx 0$$ (typically 1-2%)

Distance estimation

Jukes and Cantor model

$$\pi_A = \pi_C = \pi_G = \pi_T = \frac{1}{4}$$

$$P(t) = e^{Qt}$$

$$Q = \begin{pmatrix} -3r & r & r & r \\ r & -3r & r & r \\ r & r & -3r & r \\ r & r & r & -3r \end{pmatrix}$$

$$p_{change}(t) = \frac{3}{4}\left(1 - e^{-4rt}\right) = \frac{3}{4}\left(1 - e^{-\frac{4}{3}d}\right)$$

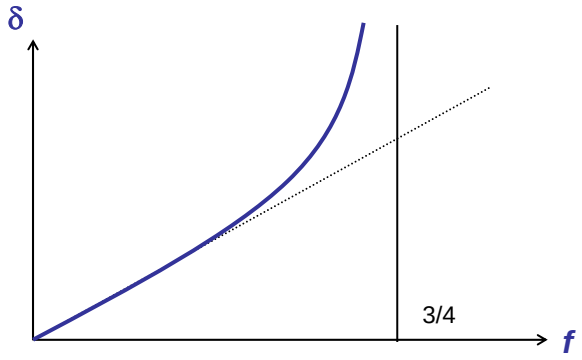
$$d = \sum_{A,T,G,C} \frac{1}{4}\left(rt + rt + rt\right) = 3rt$$

$$\delta = -\frac{3}{4}\ln\left(1 - \frac{4}{3}f\right)$$

Distance estimate

Frequency of differences

Distance estimation: Jukes and Cantor (1969)



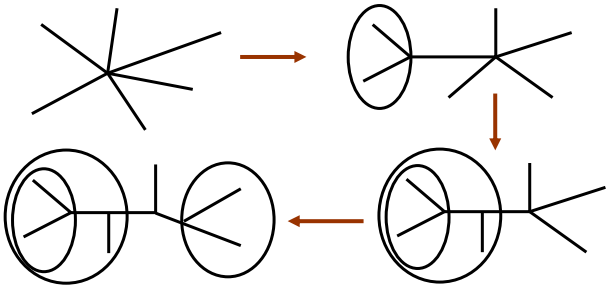
$$\delta = -\frac{3}{4}\ln\left(1-\frac{4}{3}f\right)$$

ADDTREE (Sattath and Tversky 1977)

Continuation of UPGMA (Sokal and Michener, 1958) which assumes a molecular clock and performs poorly with sequence data

Agglomerative scheme, i.e. iteratively:

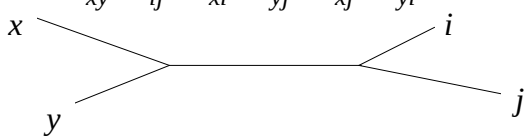
- **select** two taxa according to a criterion *Q*
- **estimate** the two corresponding branch lengths and store the structure
- **reduce** the distance matrix and replace both taxa by a unique taxon



ADDTREE

- **The four-point condition:**

$$t_{xy} + t_{ij} \leq t_{xi} + t_{yj} = t_{xj} + t_{yi}$$



- ***x* and *y* are "neighbors" for *i* and *j* when:**

$$\delta_{xy} + \delta_{ij} \leq \min\left\{\delta_{xi} + \delta_{yj},\ \delta_{xj} + \delta_{yi}\right\}$$

- **Pair selection criterion**

$$Q_{xy} = \sum_{i,j} H\left(\delta_{xi} + \delta_{yj} - \delta_{xy} - \delta_{ij}\right) H\left(\delta_{xj} + \delta_{yi} - \delta_{xy} - \delta_{ij}\right)$$

$$H(z) = 1 \text{ when } z > 1, \text{ else } H(z) = 0$$

ADDTREE time complexity

- **O(*n*) steps**
- **O(n^2) pairs of taxa**
- **each requires O(n^2) calculations**
- **computing times in O(n^5) with naive implementation**
- **can be lowered to O(n^4)**
- **able to deal with a few hundreds taxa**

Least-squares criteria

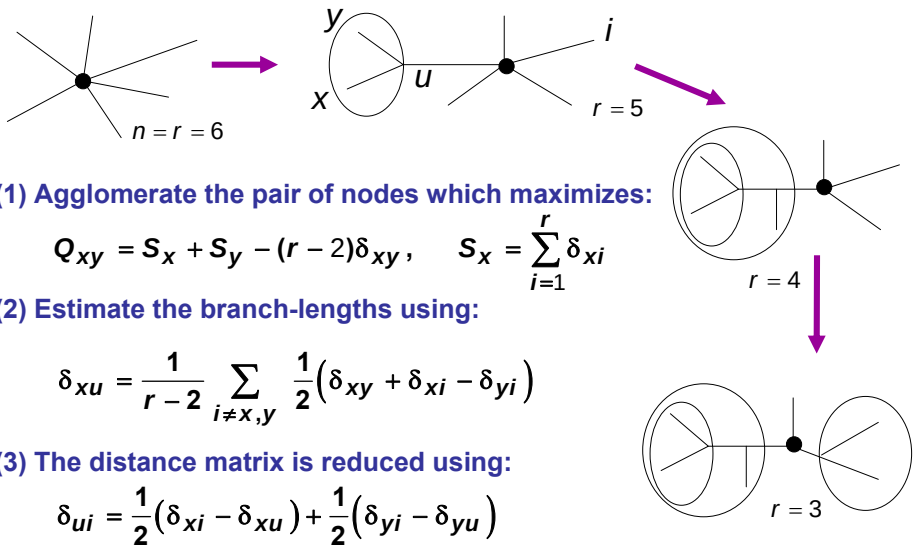
We minimize the difference between the distance matrix (δ_{ij}) and the tree distances (t_{ij}):

$$\textbf{OLS}: \sum_{i<j} \left(\delta_{ij} - t_{ij}\right)^2 \qquad \textbf{WLS}: \sum_{i<j} \frac{1}{Var\left(\delta_{ij}\right)} \left(\delta_{ij} - t_{ij}\right)^2$$

$$\textbf{Fitch\&Margoliash}: \sum_{i<j} \frac{1}{\delta_{ij}^2} \left(\delta_{ij} - t_{ij}\right)^2$$

Heavy calculations !

NJ algorithm (Saitou et Nei 1987)



(1) Agglomerate the pair of nodes which maximizes:

$$Q_{xy} = S_x + S_y - (r-2)\delta_{xy}, \qquad S_x = \sum_{i=1}^{r} \delta_{xi}$$

(2) Estimate the branch-lengths using:

$$\delta_{xu} = \frac{1}{r-2} \sum_{i \neq x,y} \frac{1}{2}\left(\delta_{xy} + \delta_{xi} - \delta_{yi}\right)$$

(3) The distance matrix is reduced using:

$$\delta_{ui} = \frac{1}{2}\left(\delta_{xi} - \delta_{xu}\right) + \frac{1}{2}\left(\delta_{yi} - \delta_{yu}\right)$$

NJ time complexity

- O(n) steps
- precomputing of the S_X sums in O(n^2)
- O(n^2) pairs of taxa
- each requires O(1) calculations
- computing times in O(n^3) with naïve implementation
- can be lowered to O(n^2)
- able to deal with thousands of taxa
- the limiting step is the computation of the distances in O(sn^2)

Strong consistency result (1996)

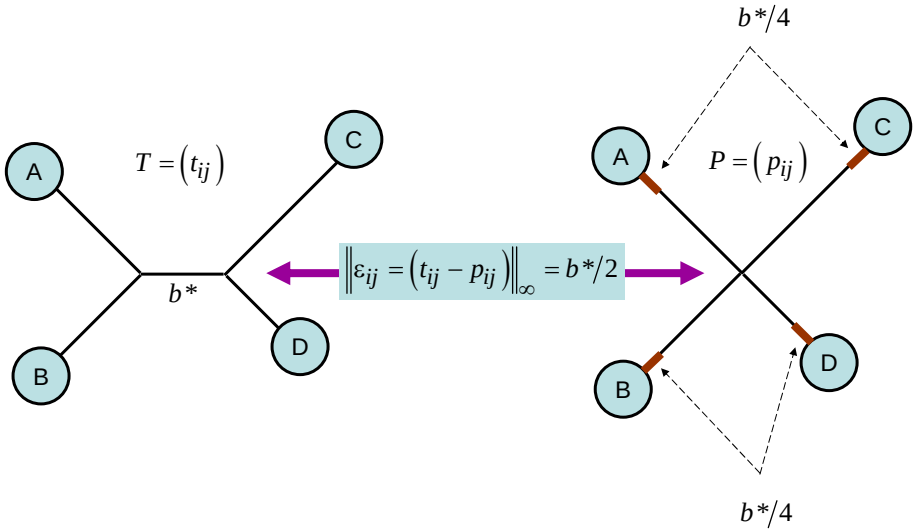
Atteson, K. The performance of the neighbor-joining methods of phylogenetic reconstruction, *Algorithmica,* 1999.

- Let $\left(\delta_{ij}\right) = \left(t_{ij}\right) + \left(\varepsilon_{ij}\right)$, i.e. estimated distances are equal to true distances plus noise.
- Let $\left\|\varepsilon\right\|_{\infty} = Max\left|\varepsilon_{ij}\right| < b^*/2$, where b^* is the shortest internal branch of T

Then: NJ reconstructs the correct tree topology from $\left(\delta_{ij}\right)$

Moreover: $b^*/2$ is optimal (no method can perform any better)

1/2 is the optimal safety radius



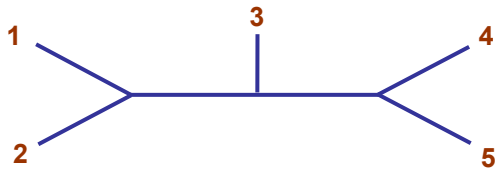
Tree length criteria

Minimum Evolution (ME): we minimize the tree length, estimated using some LS approach (parsimony flavor).

Balanced Minimum Evolution (BME), Pauplin formula:

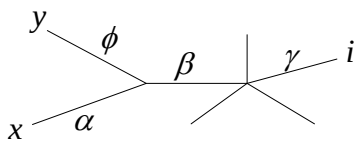
$$BL(T,\Delta) = \sum_{i<j} \frac{1}{2^{\tau_{ij}-1}} \delta_{ij}$$

where τ_{ij} is the topological distance between *i* and *j*



$$\frac{1}{2}(\delta_{12}+\delta_{45})+\frac{1}{4}(\delta_{13}+\delta_{23}+\delta_{34}+\delta_{35})+\frac{1}{8}(\delta_{14}+\delta_{15}+\delta_{24}+\delta_{25})$$

LS tree length estimation

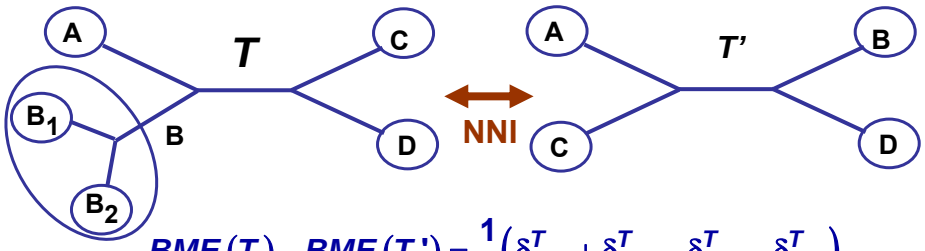


- The topology of tree T is known, and we aim to estimate its branch-lengths given evolutionary distances $\left(\delta_{ij}\right)$
- Any t_{ij} is equal to a sum of branch-lengths, e.g. $t_{xi} = \alpha + \beta + \gamma$
- We have to minimize a quadratic criterion with shape

$$\sum w_{ij}\left[\left(\alpha+\beta+\gamma\right)-\delta_{xi}\right]^2$$

• The unique solution is obtained by solving a linear system (easy, but branches may be negative). The tree length estimate is the sum of branch length estimates.

Fast topological moves with BME



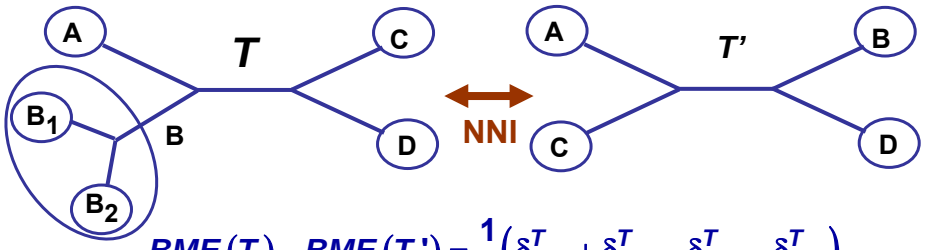
$$BME(T) - BME(T') = \frac{1}{4}\left(\delta^T_{AB} + \delta^T_{CD} - \delta^T_{AC} - \delta^T_{BD}\right)$$

where δ^T_{AB} is the balanced average distance between A and B, which is recursively defined by:

$\delta^T_{AB} = \delta_{AB}$ if A and B are tree leaves, else

$= \frac{1}{2}\left(\delta^T_{AB_1} + \delta^T_{AB_2}\right)$

Fast topological moves with BME

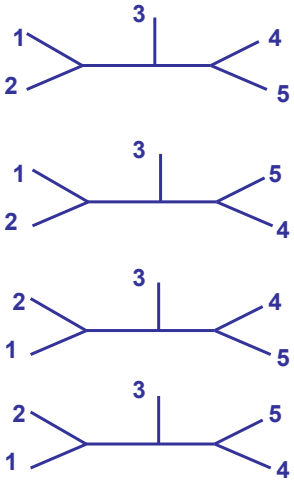


$$BME(T) - BME(T') = \frac{1}{4}\left(\delta^T_{AB} + \delta^T_{CD} - \delta^T_{AC} - \delta^T_{BD}\right)$$

- We first compute all balanced average distances in $O(n^2)$
- Then any topological move is evaluated in constant time
- Algorithms (iterative taxon insertion, NNIs, SPRs) require $O(n^3)$ or less, *i.e.* are at least as fast as NJ (standard implementation) and distances estimation

Desper R., OG, *J. Computational Biology* 2002.

BME and circular orderings



$$\delta_{13} + \delta_{34} + \delta_{45} + \delta_{52} + \delta_{21}$$

$$+$$

$$\delta_{13} + \delta_{35} + \delta_{54} + \delta_{42} + \delta_{21}$$

$$+$$

$$\delta_{12} + \delta_{23} + \delta_{34} + \delta_{45} + \delta_{51}$$

$$+$$

$$\delta_{12} + \delta_{23} + \delta_{35} + \delta_{54} + \delta_{41}$$

$$\frac{1}{2}(\delta_{12} + \delta_{54}) + \frac{1}{4}(\delta_{13} + \delta_{23} + \delta_{34} + \delta_{35}) + \frac{1}{8}(\delta_{14} + \delta_{15} + \delta_{24} + \delta_{25})$$

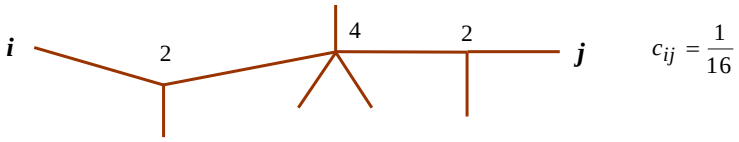
Combinatorial interpretation using circular orderings

- Pauplin's formula precisely corresponds to the average of simple tour estimates, over all possible circular orderings.

- It generalizes to non-binary trees:

$$BL\left(T,\left(\delta_{ij}\right)\right)=\sum_{i<j}c_{ij}\delta_{ij}$$

Consider the directed path from *i* to *j*, and for each interior node count how many outgoing branches are encountered—multiply these numbers together and divide 1 by the result, this gives c_{ij}.



$$c_{ij}=\frac{1}{16}$$

Semple, C. and Steel, M. *Advances in Applied Mathematics* 2004.

BME and weighted least-squares:

BME tree length is obtained by minimizing

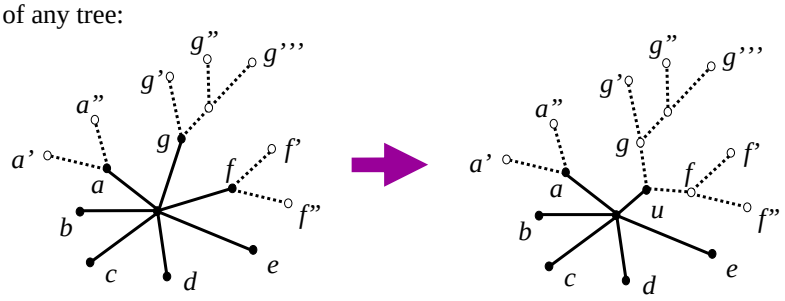
$$\textbf{WLS}: \sum_{i<j}\frac{1}{2^{\tau_{ij}}}\left(\delta_{ij}-t_{ij}\right)^2$$

which is exponentially decreasing in τ and put less weight on long (topological) distances

Desper R., OG, *Molecular Biology and Evolution* 2004.

NJ revealed

• Using generalized Pauplin's formula, we are able to estimate the length of any tree:



• NJ pair selection criterion is identical to the length difference between both trees.

• **That is, NJ does optimize balanced minimum evolution!**
(in a greedy way)

Desper & O.G., *Mathematics of Evolution & Phylogeny* 2005.
O.G., Steel M., Neighbor Joining Revealed, *Molecular Biology and Evolution*, 2006.

Olivier Gascuel – Arbres formels et Arbre de la Vie – Conférence ENS Cachan, septembre 2011

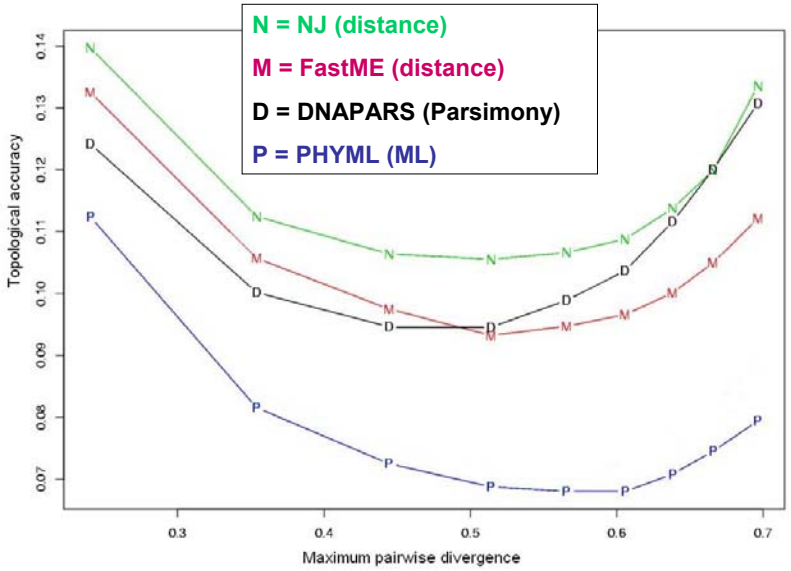
BME is unique

• BME principle has optimal safety radius (1/2)

• BME tree length is the unique formula having this optimality property, among all linear tree length formulae

Pardi F, Guillemont S & O.G., *Bulletin of Mathematical Biology* 2010.
Pardi F, in preparation.

Methods comparison: topological accuracy



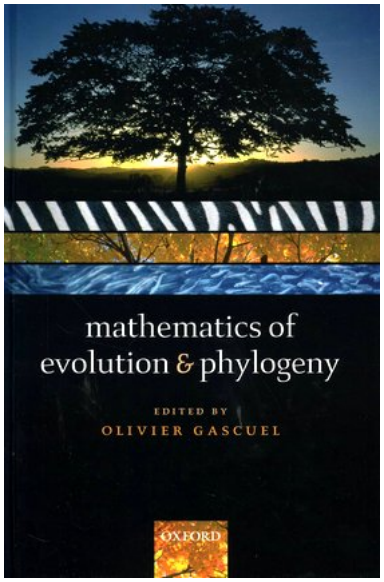
Methods comparison: computing times

Number of taxa	40	100	250
DNADIST	0.09	0.65	25
FASTME	0.008	0.055	0.34
NJ	0.0045	0.035	0.25
BIONJ	0.0052	0.055	0.60
FITCH	6	335	43,200
DNAPARS	6	230	30,000
TNT	5	13	330
DNAML	26	186	6,000
PHYML	7.5	20	390

(in seconds)

Phylogenetics
Charles Semple and Mike Steel
Oxford University Press

2005



2007