Phylogenetic Reconstruction and Applications

Emanuele Caglioti - Francesca Tria

"Sapienza" Università di Roma - ISI FoundationTorino

collaborations with

Vittorio Loreto - "Sapienza" Università di Roma and ISI Torino
Andrea Pagnani - ISI Foundation Torino
Simone Pompei - Politecnico Torino

Corinaldo 2010

Outline

Outline

- The problem

Outline

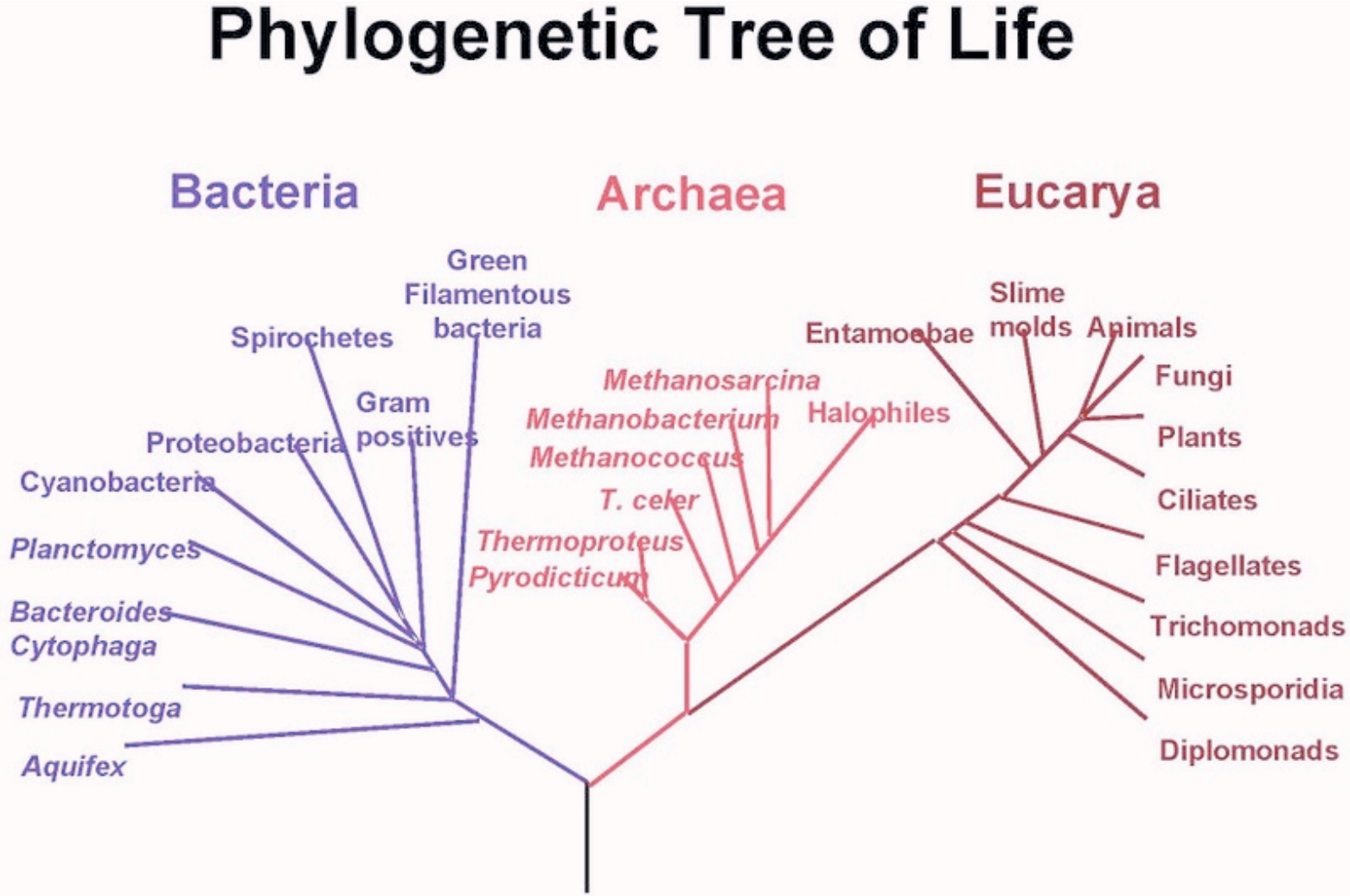
- The problem
- Additive trees

Outline

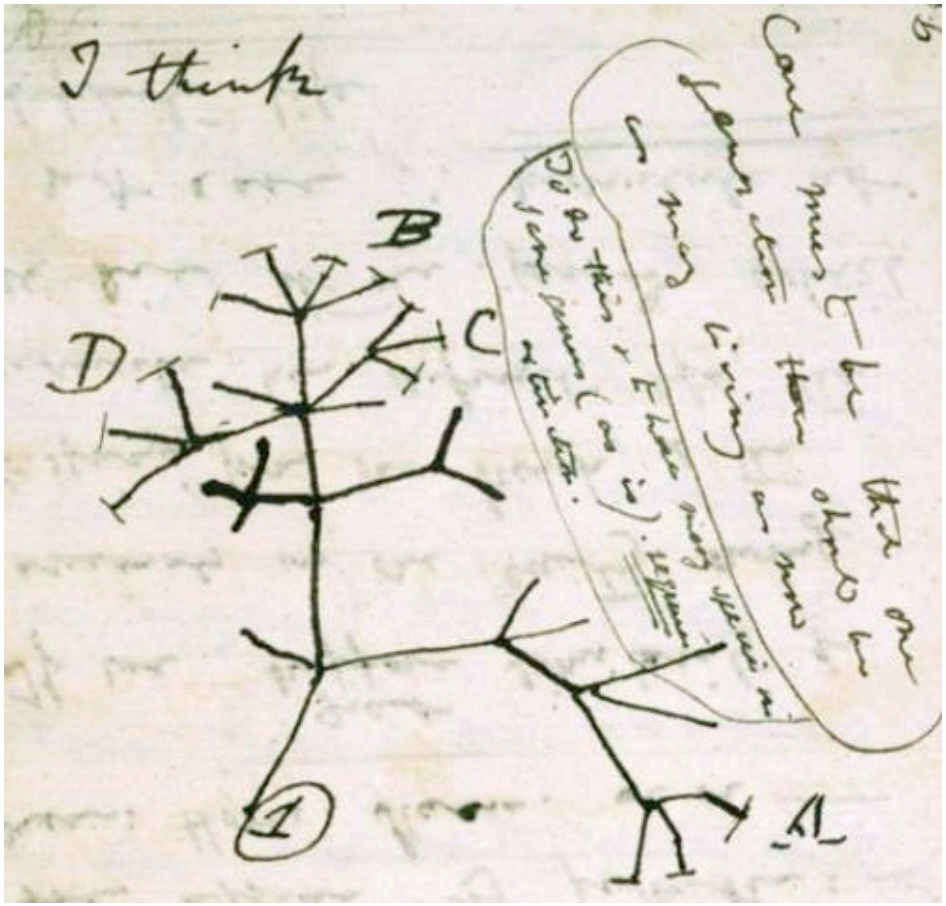
- The problem
- Additive trees
- Iterative algorithms

Outline

- The problem
- Additive trees
- Iterative algorithms
- Global (variational) algorithms


Phylogenetic Tree of Life
Bacteria
Archaea
Eucarya
Green
Filamentous
bacteria
Spirochetes
Gram
positives
Proteobacteria
Cyanobacteria
Planctomyces
Bacteroides
Cytophaga
Thermotoga
Aquifex
Methanosarcina
Methanobacterium
Methanococcus
T. celer
Halophiles
Thermoproteus
Pyrodicticum
Slime
molds
Entamoebae
Animals
Fungi
Plants
Ciliates
Flagellates
Trichomonads
Microsporidia
Diplomonads

Darwin's tree of life




nel mezzo del cammin di nostra vita
nel mezzo di cammin di nostra vita
nel mezzo del cammino di nostra vita
nel mezzo di cammin di vostra vita
nel mezzo di cammin di nostra via
nel mezzo di cammin di nostra vita
a mezzo di cammino di nostra vita
mutation
back mutation

A simple evolutionary model

A simple evolutionary model

Simulate the evolution changing randomly binary sequences with a certain mutation rate per site and branching at Poisson times

A simple evolutionary model

Simulate the evolution changing randomly binary sequences with a certain mutation rate per site and branching at Poisson times

0000000000000000000000000000000000

A simple evolutionary model

Simulate the evolution changing randomly binary sequences with a certain mutation rate per site and branching at Poisson times

0000000000000000000000000000000000
0000000010000000000000000000000000

A simple evolutionary model

Simulate the evolution changing randomly binary sequences with a certain mutation rate per site and branching at Poisson times

0000000000000000000000000000000000
0000000010000000000000000000000000
0000000010000000000001000000000000

A simple evolutionary model

Simulate the evolution changing randomly binary sequences with a certain mutation rate per site and branching at Poisson times

0000000000000000000000000000000000
0000000010000000000000000000000000
0000000010000000000000100000000000
0000000010000000000100010000000000

A simple evolutionary model

Simulate the evolution changing randomly binary sequences with a certain mutation rate per site and branching at Poisson times

0000000000000000000000000000000000
0000000010000000000000000000000000
0000000010000000000000010000000000
0000000010000000000100010000000000
0000000010000000000000010000000000

mutation

back mutation

Algorithms for Phylogenetic reconstruction

Algorithms for Phylogenetic reconstruction

Distance based

Characters based

Algorithms for Phylogenetic reconstruction

Distance based

Infer the tree using the distance matrix only

UPGMA

Neighbor Joining

Fitch

Weighbor

FASTME

Characters based

Algorithms for Phylogenetic reconstruction

Distance based

Infer the tree using the distance matrix only

UPGMA

Neighbor Joining

Fitch

Weighbor

FASTME

Characters based

Compare different sequences character by character

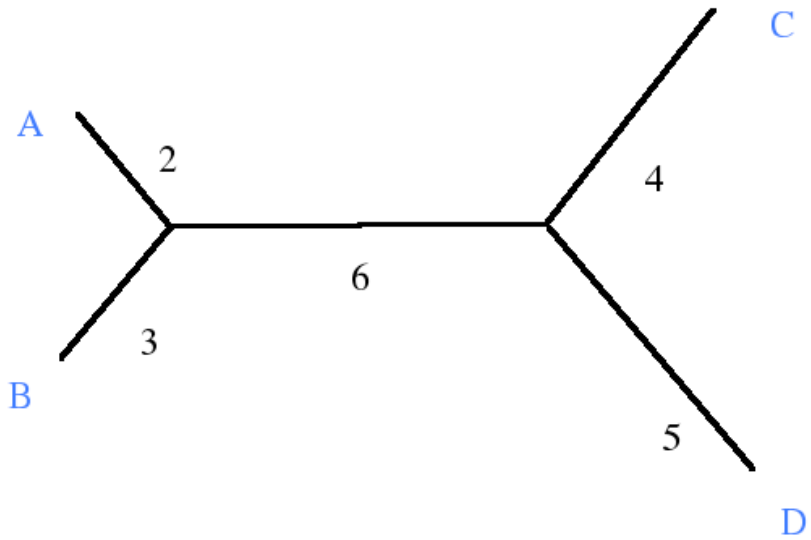
Parsimony

MrBayes

Additivity

Definition. A distance matrix is additive if there exists a tree on which, for each pair of taxa X, Y, $d_{X,Y}$ is the sum of the length of the branches connecting X and Y

Additivity



	A	B	C	D
A	0	5	12	13
B	5	0	13	14
C	12	13	0	9
D	13	14	9	0

Four Point Condition

Definition.

A distance is additive iff for any four taxa A,B,C,D it is

$$d_{A,B} + d_{C,D} < d_{A,C} + d_{B,D} = d_{A,D} + d_{B,C}$$

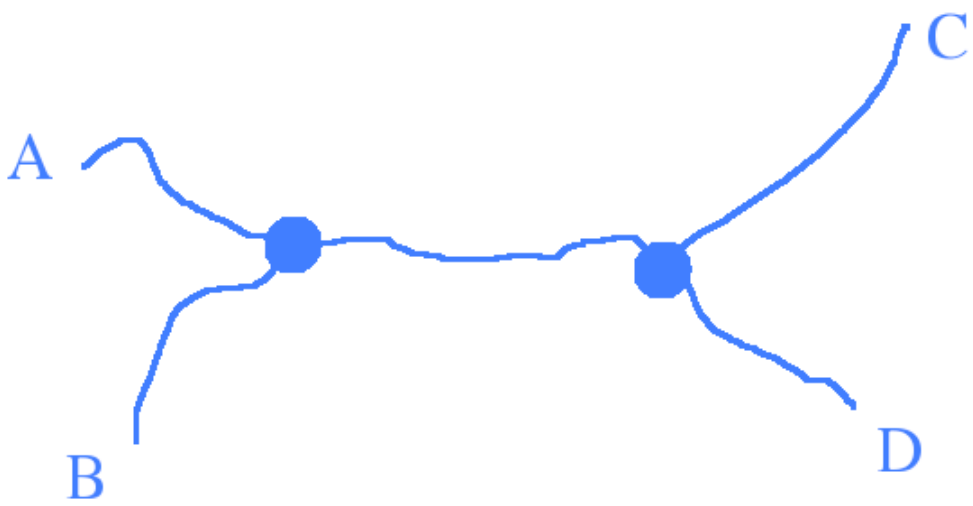
or

$$d_{A,C} + d_{B,D} < d_{A,B} + d_{C,D} = d_{A,D} + d_{B,C}$$

or

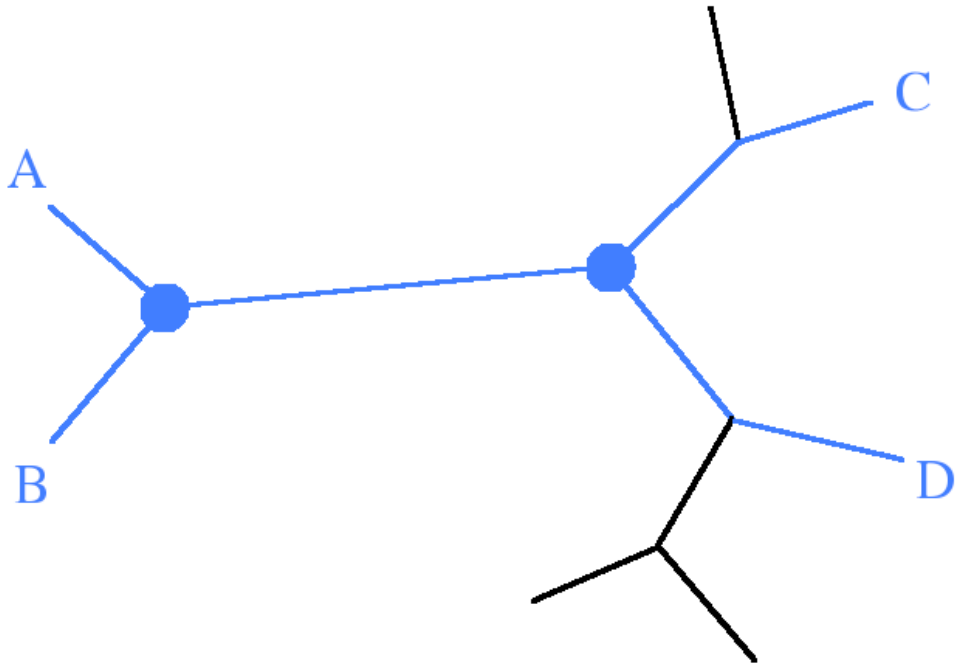
$$d_{A,D} + d_{B,C} < d_{A,C} + d_{B,D} = d_{A,B} + d_{C,D}$$

Four Point Condition



$$d_{A,B} + d_{C,D} < d_{A,C} + d_{B,D} = d_{A,D} + d_{B,C}$$

Four Point Condition



$$d_{A,B} + d_{C,D} < d_{A,C} + d_{B,D} = d_{A,D} + d_{B,C}$$

Iterative Algorithms

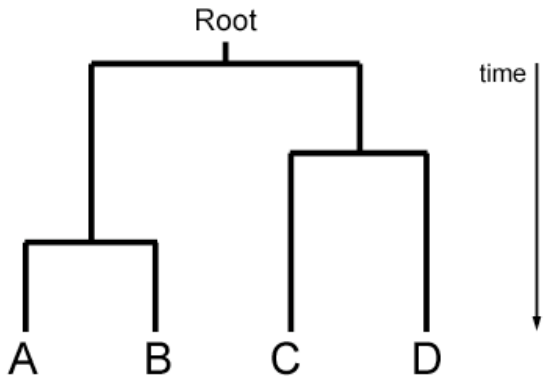
UPGMA

- connect the two nearest taxa XY
- compute the distance between the new taxa and the other taxa
- iterate till there remain only 3 taxa

Neighbor Joining

UPGMA works for ultrametric trees but not for additive trees in general

ultrametricity $\leftrightarrow$ constant evolutive speed



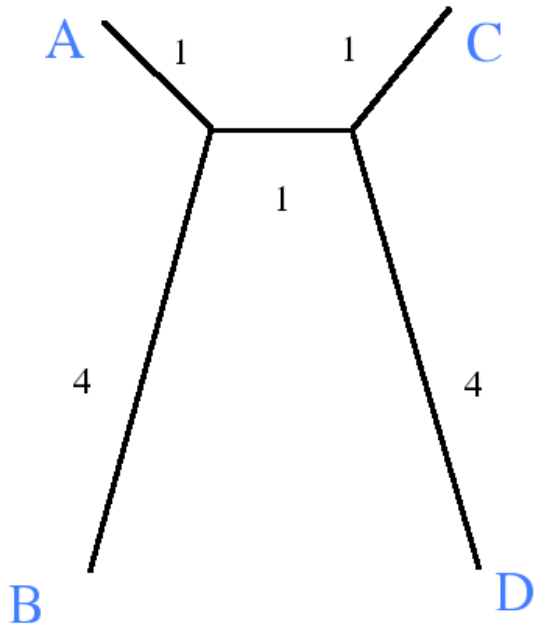
ultrametricity: for any three leaves a, b, c

$$d_{a,b} \leq \max(d_{a,c}, d_{b,c})$$

Neighbor Joining

UPGMA works for ultrametric trees but not for additive trees in general

ultrametricity $\leftrightarrow$ constant evolutive speed



$$d_{A,C} < d_{A,B}$$

Neighbor Joining

Define the matrix D as

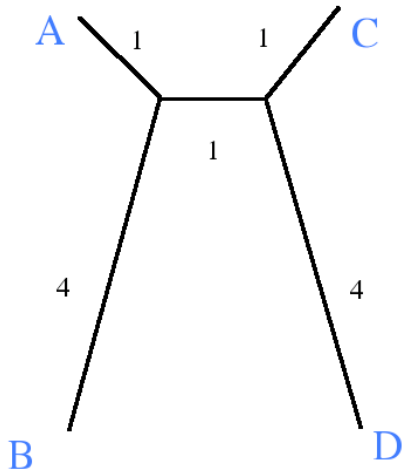
$$D_{X,Y} = d_{x,y} - r_X - r_Y$$

where

$$r_X = \frac{1}{N-2} \sum_Y d_{X,Y}$$

- connect the two taxa which minimizes $D_{X,Y}$
- compute D between the new taxa and the other taxa
- iterate till there remain only 3 taxa

Saitou and Nei 1987



$$r_A = r_C = \frac{5+3+6}{2} = 7, \qquad r_B = r_D = \frac{5+6+9}{2} = 10$$

$$d = \begin{pmatrix} 0 & 5 & 3 & 6 \\ 5 & 0 & 6 & 9 \\ 3 & 6 & 0 & 5 \\ 6 & 9 & 5 & 0 \end{pmatrix} \qquad D = \begin{pmatrix} 0 & -12 & -11 & -11 \\ -12 & 0 & -11 & -11 \\ -11 & -11 & 0 & -12 \\ -11 & -11 & -12 & 0 \end{pmatrix}$$

Properties of NJ

- Theorem (Saitou and Nei)
  If the tree is additive then NJ reconstruct the correct tree.
  Main ingredient: if X.Y minimizes D then X,Y is a cherry

Properties of NJ

- Theorem (Saitou and Nei)
  If the tree is additive then NJ reconstruct the correct tree.
  Main ingredient: if X.Y minimizes D then X,Y is a cherry
- Stability Theorem (K. Atteson 1997) If the distance estimates are at most half of the minimal edge length of the tree away from their true value then Neighbor-Joining will reconstruct the correct tree

Properties of NJ

- Theorem (Saitou and Nei)
  If the tree is additive then NJ reconstruct the correct tree.
  Main ingredient: if X.Y minimizes D then X,Y is a cherry
- Stability Theorem (K. Atteson 1997) If the distance estimates are at most half of the minimal edge length of the tree away from their true value then Neighbor-Joining will reconstruct the correct tree
- Neighbor Joining runs in $O(N^3)$ time

Pauplin Formula

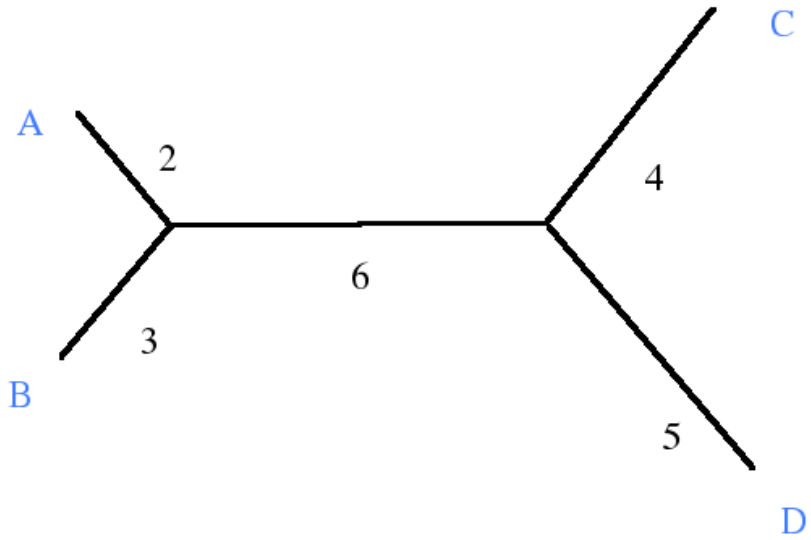
If a distance is additive than the total length of the corresponding tree is given by

$$P = \sum_{\{X,Y\}} 2^{-t_{X,Y}} d_{X,Y}$$

where $t_{X,Y}$ is the number of nodes between X and Y

Pauplin Formula

$$L = \frac{1}{2}d_{A,B} + \frac{1}{4}d_{A,C} + \frac{1}{4}d_{A,D} + ... = 20$$



Y. Pauplin 2000

Proof of Pauplin formula

Let L be the length of the tree: i.e.

$$L = \sum_{k:k\ edge} l_k$$

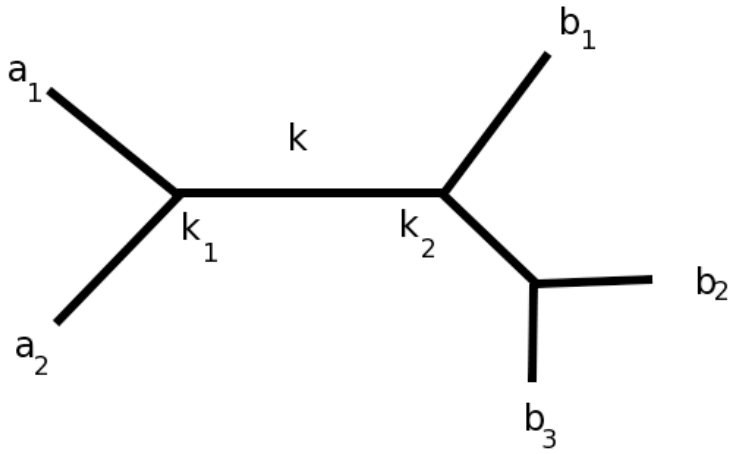
where we denoted with l_k the length of the edge k.
The Pauplin expressione P can be written as

$$\begin{aligned} P = & \sum_{a,b} 2^{-t_{a,b}} d_{a,b} \\ = & \sum_{a,b} \sum_{k \in w_{a,b}} 2^{-t_{a,b}} l_k \\ = & \sum_k l_k \sum_{a,b} 1_{k \in w_{a,b}} 2^{-t_{a,b}} \end{aligned}$$

where $\sum_{a,b}$ denotes the sum on distinct pair of leaves a, b and where we denoted with $W_{a,b}$ the set of edges connecting the leaf a with the leaf b.
Given an edge k, let us denote with k_1 and k_2 the two nodes connected by k, and let us denote with T_1 and T_2 the two trees which has roots in k_1 and k_2 , respectively.
We can notice that

$$\sum_{a,b} 1_{k \in W_{a,b}} 2^{-t_{a,b}} = \sum_{a \in T_1} \sum_{b \in T_2} 2^{-t_{a,b}}$$

in fact all the pairs which contribute to the above sum have one leaf in T_1 and the other in T_2.



Now we can notice that for any $a \in T_1$, and $b \in T_2$ we can write

$$t_{a,b} = z_a + z_b$$

where z_a is the the number of branches in the path between a and k_1 and where z_b is the number of branches in the path between b and k_2.

Therefore

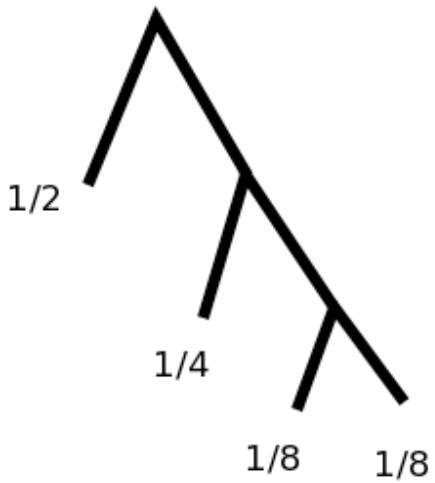
$$\sum_{a\in T_1}\sum_{b\in T_2} 2^{-t_{a,b}} = \sum_{a\in T_1}\sum_{b\in T_2} 2^{-z_a-z_b} = \sum_{a\in T_1} 2^{-z_a}\sum_{b\in T_2} 2^{-z_b} = 1\cdot 1 = 1 \tag{1}$$

by Kraft equality.

Kraft Equality

In a binary tree let z_k be the depth of the leaf k. Then

$$\sum_k 2^{-z_k} = 1$$



Balanced Minimum Evolution Principle

If a distance is additive the right tree T is the one that minimizes

$$L_T = \sum_{\{X,Y\}} 2^{-t_{X,Y}} d_{X,Y}$$

R. Desper, O.Gascuel 2003

A BME Algorithm

A BME Algorithm

Define an Energy:

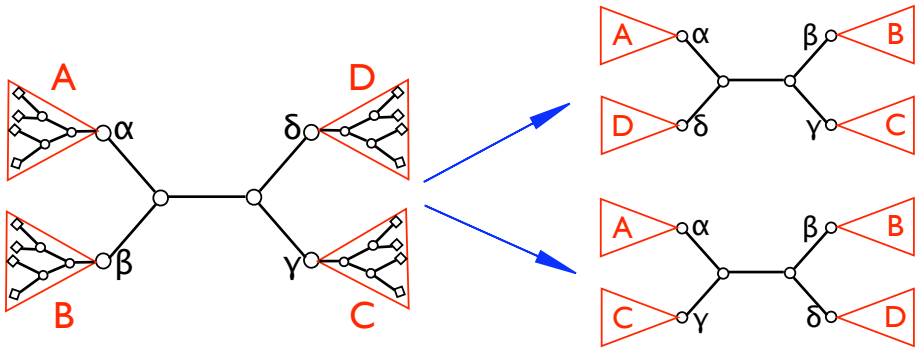
$$E = \sum_{a,b} 2^{-t_{a,b}} d_{a,b}$$

A BME Algorithm

Define an Energy:

$$E = \sum_{a,b} 2^{-t_{a,b}} d_{a,b}$$

Define an elementary move

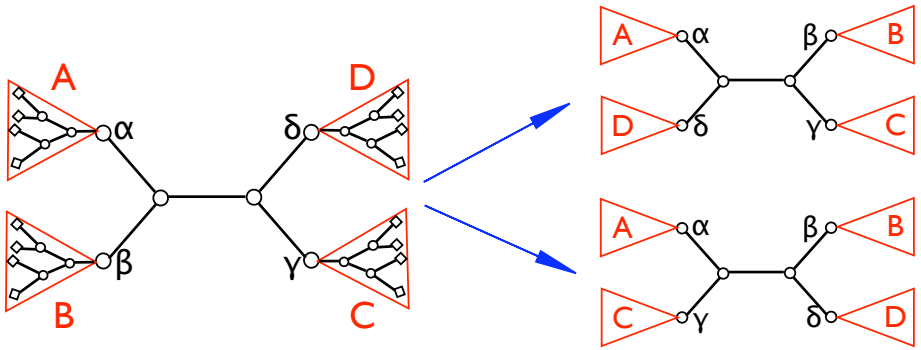


A BME Algorithm

Define an Energy:

$$E = \sum_{a,b} 2^{-t_{a,b}} d_{a,b}$$

Define an elementary move



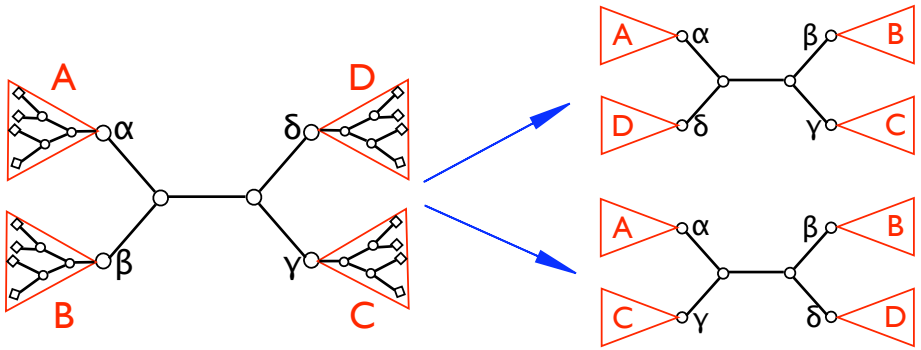
- start from a reasonable tree (NJ tree)

A BME Algorithm

Define an Energy:

$$E = \sum_{a,b} 2^{-t_{a,b}} d_{a,b}$$

Define an elementary move



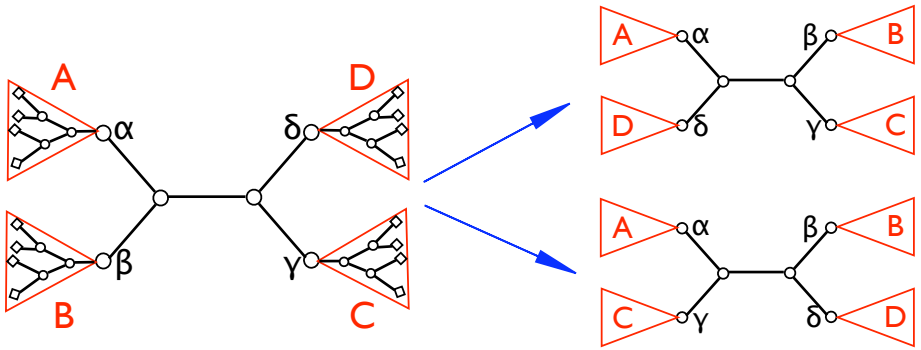
- start from a reasonable tree (NJ tree)
- extract a link

A BME Algorithm

Define an Energy:

$$E = \sum_{a,b} 2^{-t_{a,b}} d_{a,b}$$

Define an elementary move



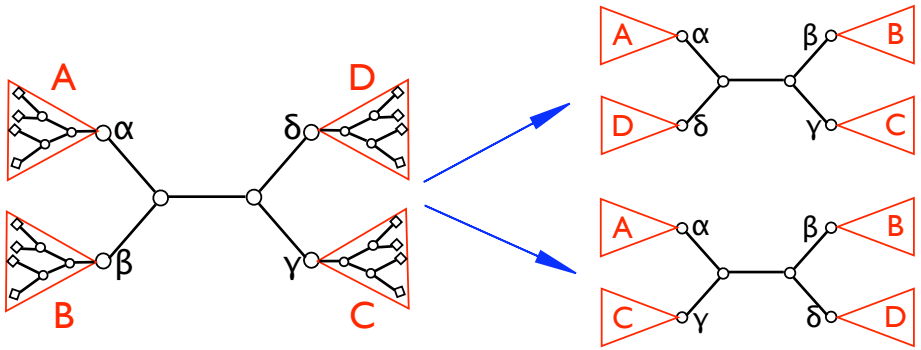
- start from a reasonable tree (NJ tree)
- extract a link
- extract an elementary move

A BME Algorithm

Define an Energy:

$$E = \sum_{a,b} 2^{-t_{a,b}} d_{a,b}$$

Define an elementary move



- start from a reasonable tree (NJ tree)
- extract a link
- extract an elementary move
- accept the move if $\Delta E < 0$

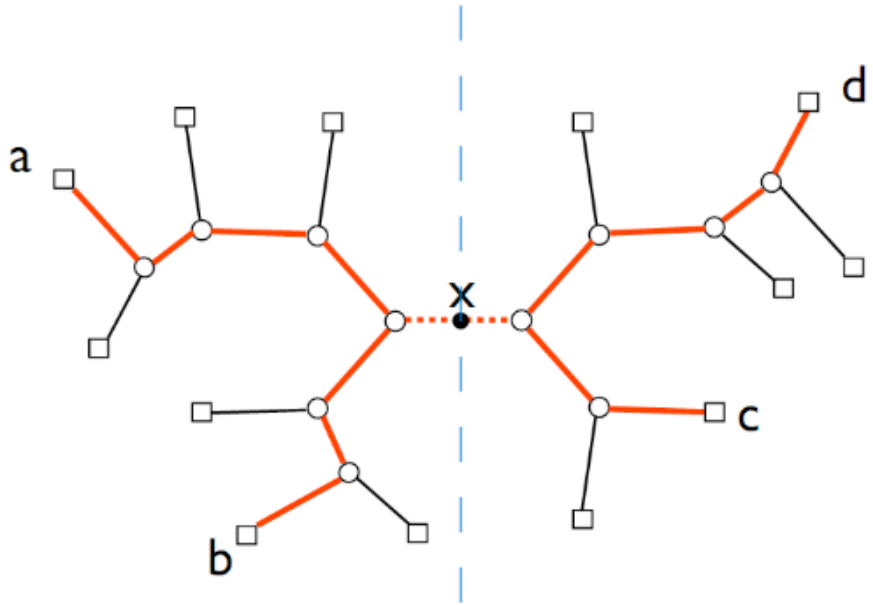
FASTME algorithm

$$\sum_{\{X,Y\}} 2^{-t(X,Y)} d(X,Y)$$

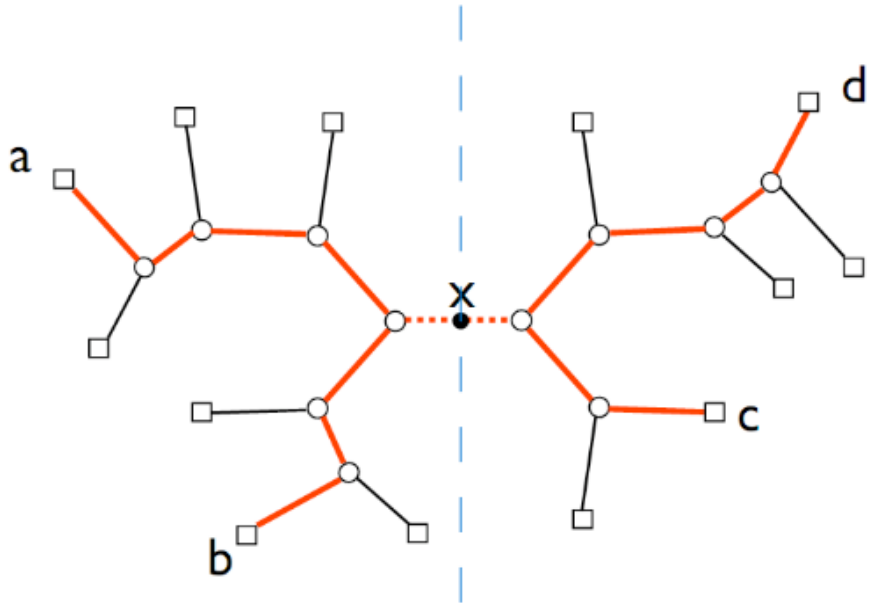
The algorithm FASTME starts form a reasonable tree (NJ tree) and then makes suitable elementary moves to minimize the formula above

O. Gascuel and M. Steel 2006

Quartet Based Algortihms



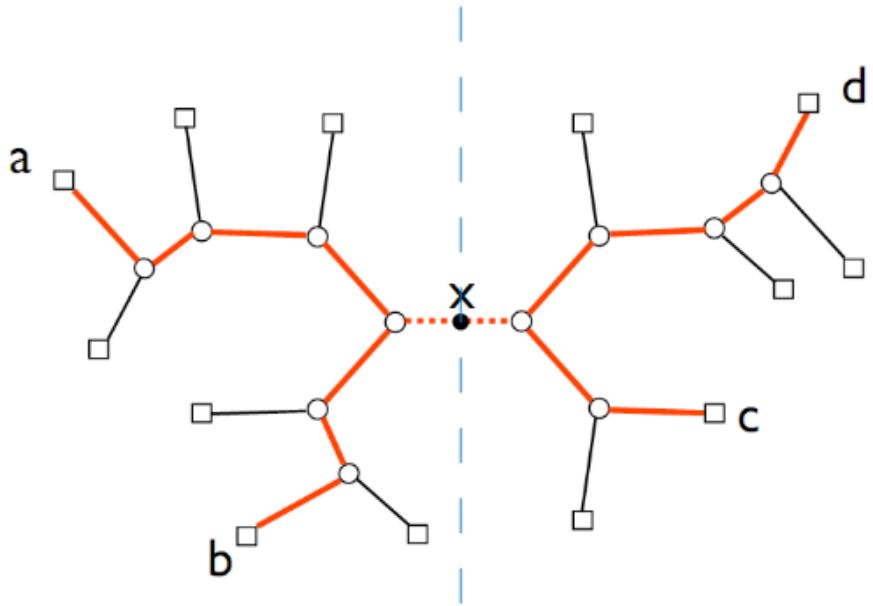
Quartet Based Algortihms



strong four points condition

$$d_{A,B} + d_{C,D} < d_{A,C} + d_{B,D} = d_{A,D} + d_{B,C}$$

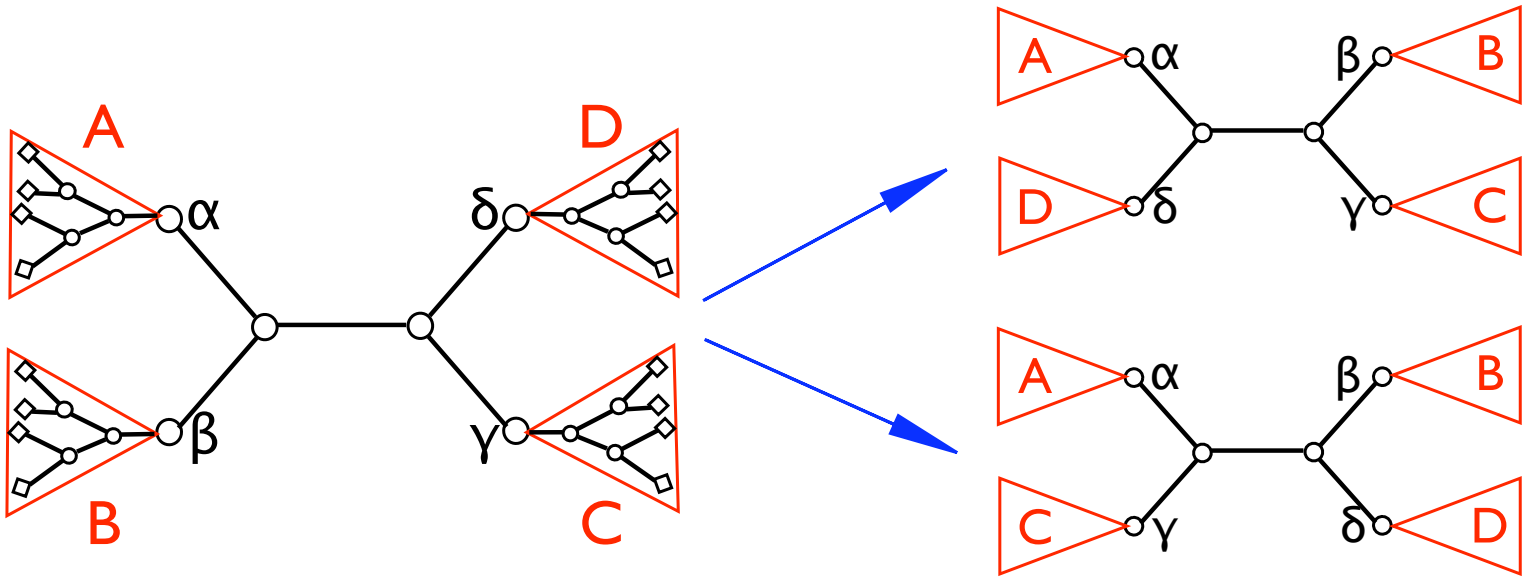
Quartet Based Algortihms



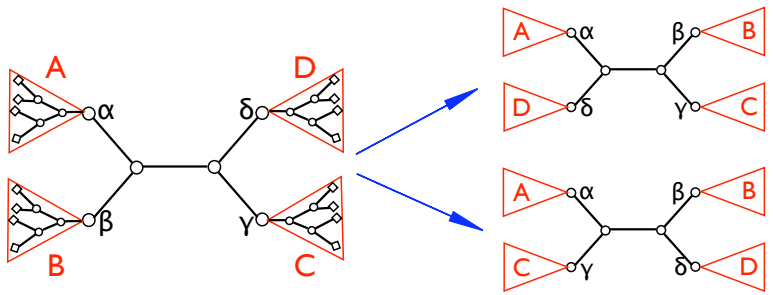
soft four points condition

$$d_{A,B} + d_{C,D} < min(d_{A,C} + d_{B,D}, d_{A,D} + d_{B,C})$$

A Quarted Based Algorithms: Elementary Move



A Quarted Based Algorithms: Elementary Move



for any $a \in A, b \in B, c \in C, d \in D$

define $D_1 = d_{a,b} + d_{c,d}, D_2 = d_{a,c} + d_{c,d}, D_3 = d_{a,d} + d_{b,c}$

define quartet frustration as

$$f_{(a,b)(c,d)} = max(0, D_1 - min(D_2, D_3))$$

A Quartet Based Algorithm

define configuration energy

$$E_{(A,B),(C,D)} = \sum_{a\in A, b\in B, c\in C, d\in D} 2^{-n(a,\alpha)-n(b,\beta)-n(c,\gamma)-n(d,\delta)} f_{(a,b),(c,d)}$$

A Quartet Based Algorithm

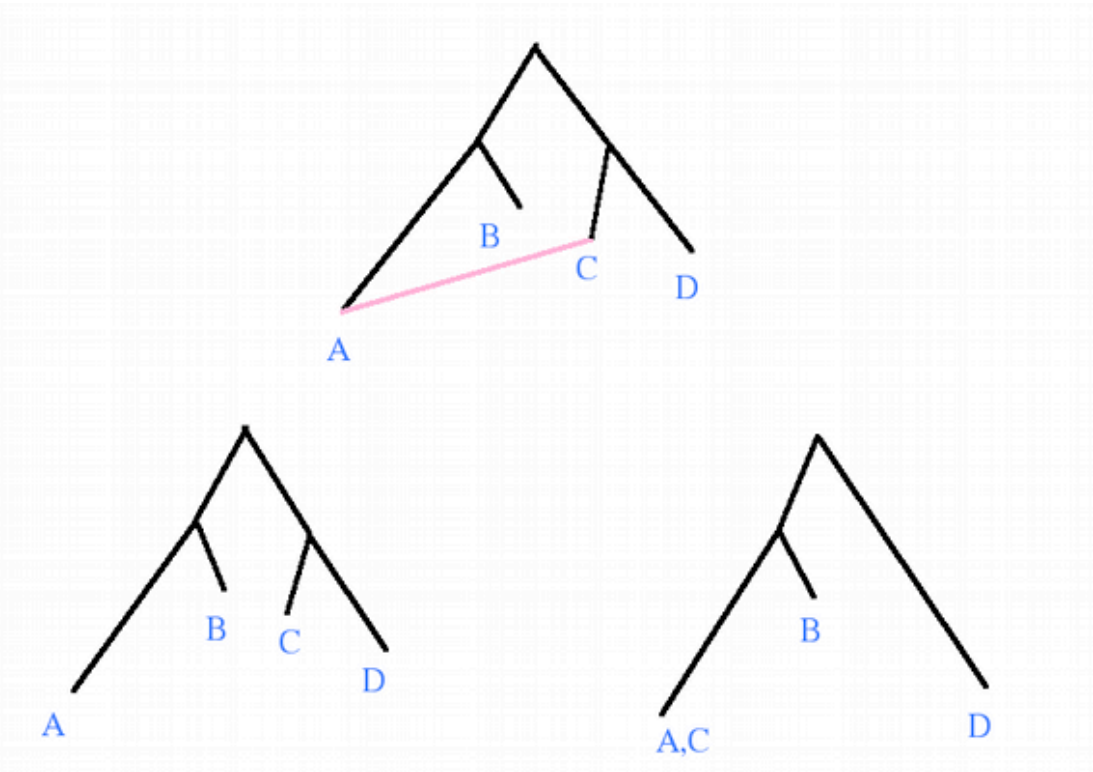
define configuration energy

$$E_{(A,B),(C,D)} = \sum_{a\in A, b\in B, c\in C, d\in D} 2^{-n(a,\alpha)-n(b,\beta)-n(c,\gamma)-n(d,\delta)} f_{(a,b),(c,d)}$$

ΔE is the variation of a functional and the functional is the Pauplin length L_T

Horizontal Transfer

An entire part of sequence A is copied in sequence C
As a result the distance matrix is the convex combination of two additive matrices, then it is not an additive matrix



Phylogeny reconstruction with applications in linguistics and biology

Emanule Caglioti and Francesca Tria

Sapienza Università di Roma Fondazione ISI, Torino

joint work with

Vittorio Loreto Sapienza Università di Roma & ISI

Andrea Pagnani Fondazione ISI, Torino

Simone Pompei Fondazione ISI, Torino

Outline

Outline

- Algorithms

Outline

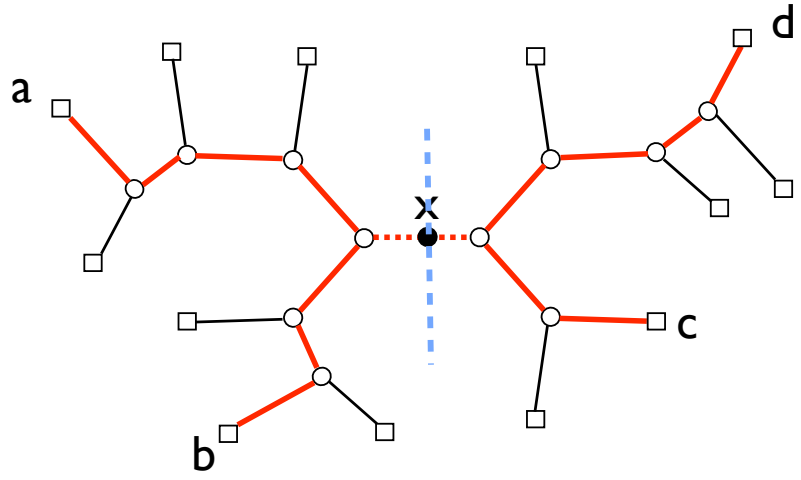
- Algorithms
- An application in linguistics

Outline

- Algorithms
- An application in linguistics
- A biology related problem

A recall

Soft four points condition

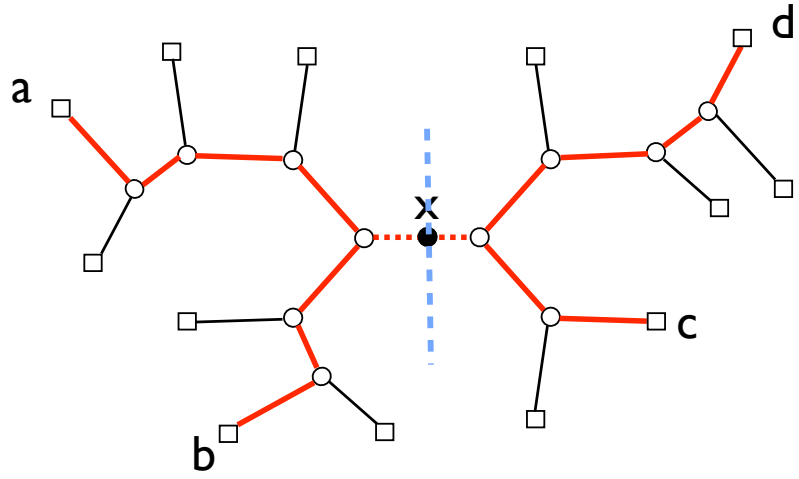


$$D_1 = d_{a,b} + d_{c,d}, D_2 = d_{a,c} + d_{c,d}, D_3 = d_{a,d} + d_{b,c}$$

$$D_1 = \min(D_1, D_2, D_3)$$

A recall

Soft four points condition

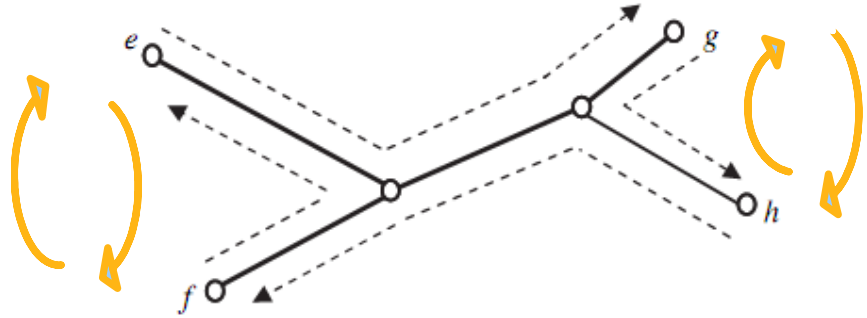


$$D_1 = d_{a,b} + d_{c,d}, D_2 = d_{a,c} + d_{c,d}, D_3 = d_{a,d} + d_{b,c}$$

$$D_1 = \min(D_1, D_2, D_3)$$

Pauplin's formula

$$L_P = \sum_{a<b} 2^{-t(a,b)} \mathcal{D}(a,b)$$

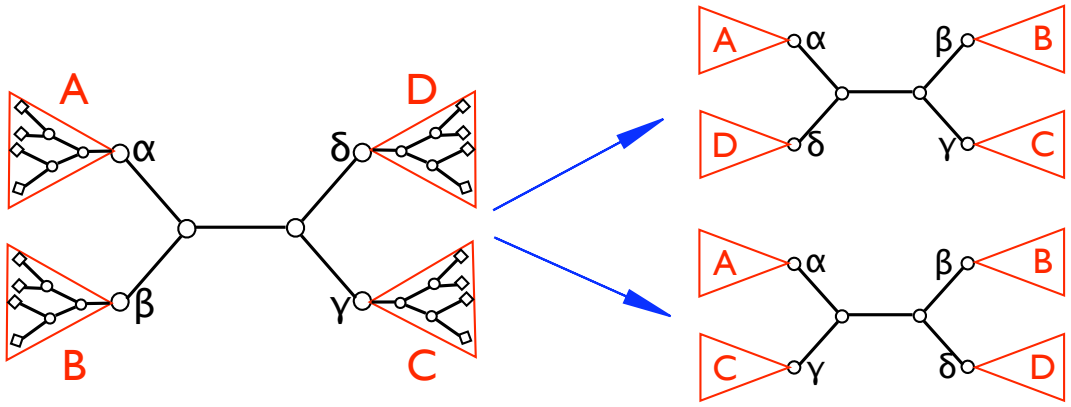


O. Gascuel and M. Steel 2006

$$l = \frac{1}{2}[d(e,g) + d(g,h) + d(h,f) + d(f,e)]$$

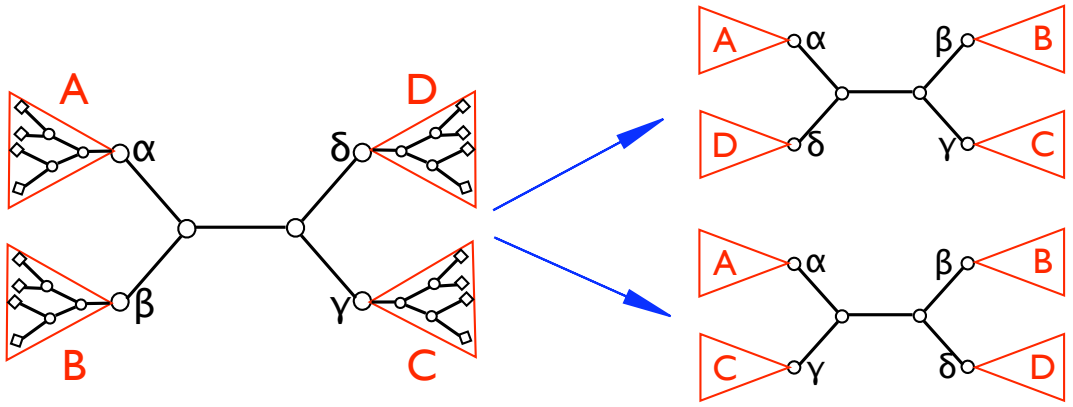
$$L_P = \frac{1}{2}[d(e,f) + d(g,h)] + \frac{1}{4}[d(e,g) + d(e,h) + d(f,h) + d(f,g)] \qquad L_P = l$$

Stochastic local search algorithms (SLS)



elementary moves (NNI)

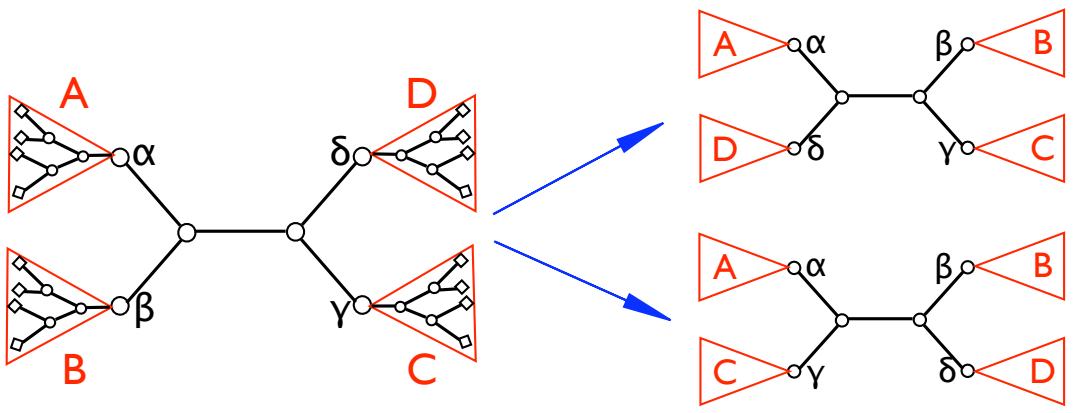
Stochastic local search algorithms (SLS)



elementary moves (NNI)

1. extract a link
2. extract an elementary move
3. accept the move with probability $e^{-\beta \Delta E}$

Stochastic local search algorithms (SLS)



elementary moves (NNI)

1. extract a link
2. extract an elementary move
3. accept the move with probability $e^{-\beta \Delta E}$

simulated annealing-like procedure:

β (*inverse temperature*) increases with time

zero temperature procedure:

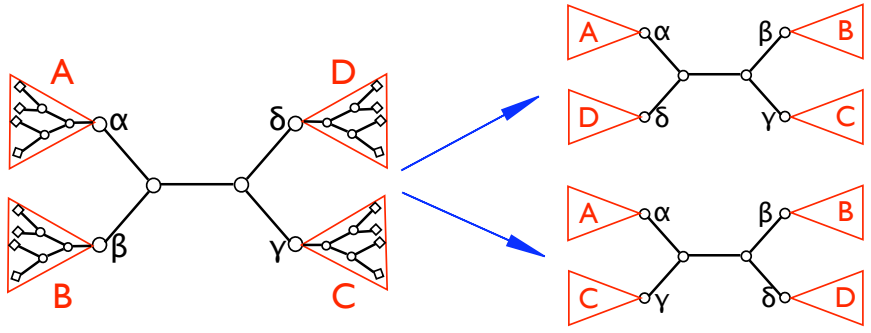
$\beta = +\infty \longleftrightarrow$ accept the move iff $\Delta E < 0$

Local (configurational) energy definition based on the soft four points condition

Local (configurational) energy definition based on the soft four points condition

for any

$$a \in A, b \in B, c \in C, d \in D$$



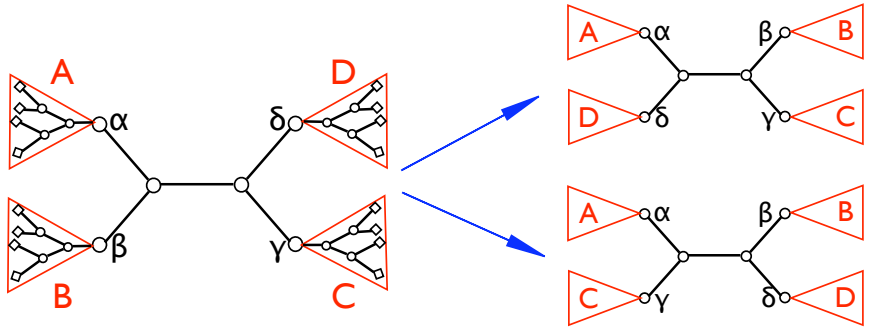
define

$$D_1 = d_{a,b} + d_{c,d}, D_2 = d_{a,c} + d_{c,d}, D_3 = d_{a,d} + d_{b,c}$$

Local (configurational) energy definition based on the soft four points condition

for any

$$a \in A, b \in B, c \in C, d \in D$$

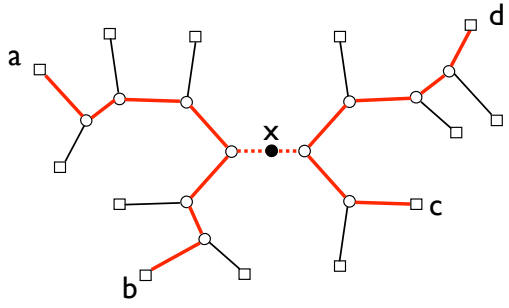


define

$$D_1 = d_{a,b} + d_{c,d}, D_2 = d_{a,c} + d_{c,d}, D_3 = d_{a,d} + d_{b,c}$$

soft four points condition

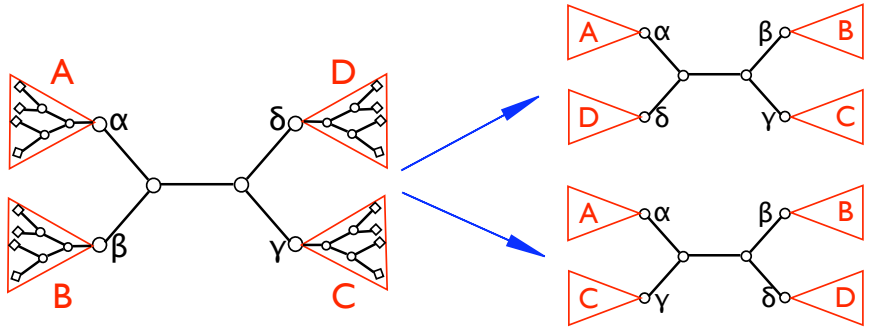
$$D_1 = \min(D_1, D_2, D_3)$$



Local (configurational) energy definition based on the soft four points condition

for any

$$a \in A, b \in B, c \in C, d \in D$$



define

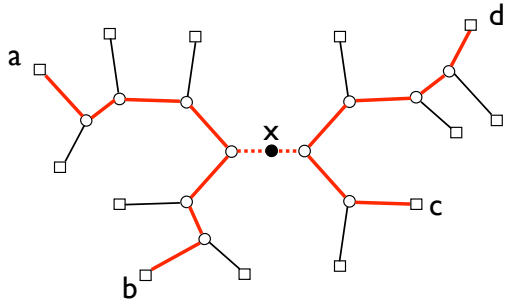
$$D_1 = d_{a,b} + d_{c,d}, D_2 = d_{a,c} + d_{c,d}, D_3 = d_{a,d} + d_{b,c}$$

soft four points condition

$$D_1 = \min(D_1, D_2, D_3)$$

quartet frustration

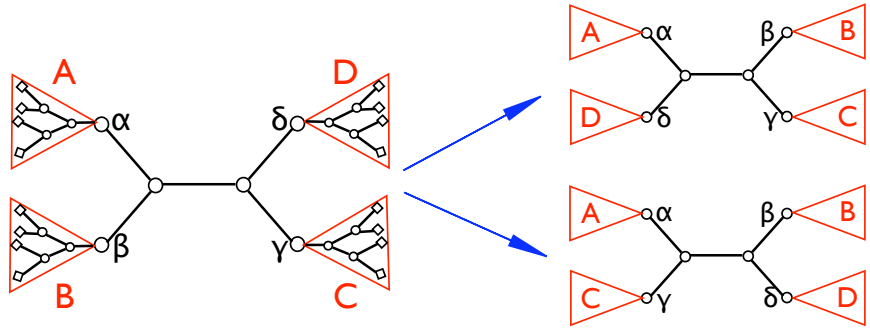
$$f_{(a,b)(c,d)} = max(0, D_1 - min(D_2, D_3))$$



Local (configurational) energy definition based on the soft four points condition

for any

$$a \in A, b \in B, c \in C, d \in D$$

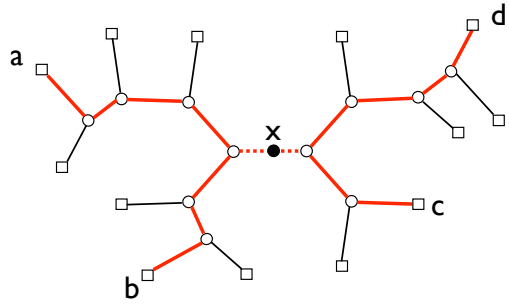


define

$$D_1 = d_{a,b} + d_{c,d}, D_2 = d_{a,c} + d_{c,d}, D_3 = d_{a,d} + d_{b,c}$$

soft four points condition

$$D_1 = \min(D_1, D_2, D_3)$$



quartet frustration

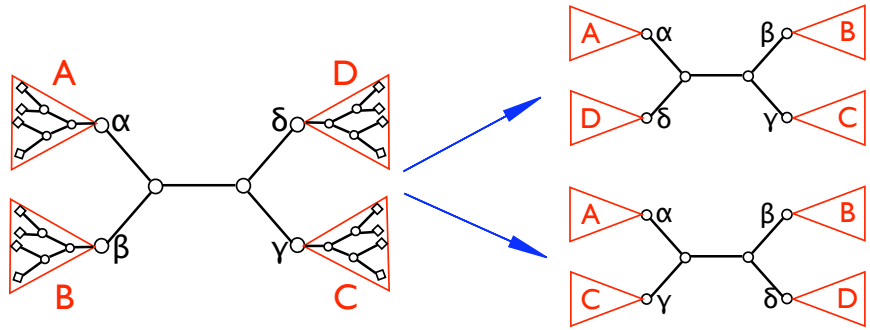
$$f_{(a,b)(c,d)} = max(0, D_1 - min(D_2, D_3))$$

1. $$E_{(A,B),(C,D)} = \sum_{a \in A, b \in B, c \in C, d \in D} 2^{-n(a,\alpha)-n(b,\beta)-n(c,\gamma)-n(d,\delta)} f_{(a,b),(c,d)}$$

Local (configurational) energy definition based on the soft four points condition

for any

$$a \in A, b \in B, c \in C, d \in D$$

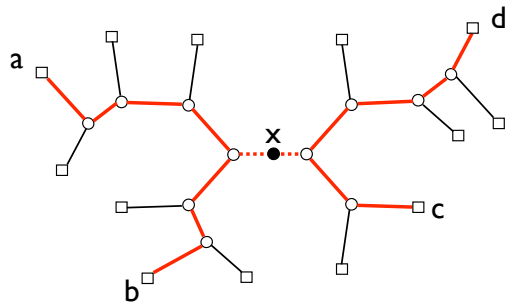


define

$$D_1 = d_{a,b} + d_{c,d}, D_2 = d_{a,c} + d_{c,d}, D_3 = d_{a,d} + d_{b,c}$$

soft four points condition

$$D_1 = \min(D_1, D_2, D_3)$$



quartet frustration

$$f_{(a,b)(c,d)} = max(0, D_1 - min(D_2, D_3))$$

1. $$E_{(A,B),(C,D)} = \sum_{a \in A, b \in B, c \in C, d \in D} 2^{-n(a,\alpha)-n(b,\beta)-n(c,\gamma)-n(d,\delta)} f_{(a,b),(c,d)}$$

$$\Delta E = 4\Delta L_P$$

1. $E_{(A,B),(C,D)} = \sum_{a\in A, b\in B, c\in C, d\in D} 2^{-n(a,\alpha)-n(b,\beta)-n(c,\gamma)-n(d,\delta)} f_{(a,b),(c,d)}$

1. $$E_{(A,B),(C,D)} = \sum_{a \in A, b \in B, c \in C, d \in D} 2^{-n(a,\alpha)-n(b,\beta)-n(c,\gamma)-n(d,\delta)} f_{(a,b),(c,d)}$$

2. $$E_{(A,B),(C,D)} = \sum_{a \in A, b \in B, c \in C, d \in D} \frac{2^{-n(a,\alpha)-n(b,\beta)-n(c,\gamma)-n(d,\delta)} f_{(a,b),(c,d)}}{(D_1 + min(D_2, D_3))^K}$$

$$K > 0$$

Motivation: longer distances have larger fluctuations

1. $$E_{(A,B),(C,D)} = \sum_{a\in A, b\in B, c\in C, d\in D} 2^{-n(a,\alpha)-n(b,\beta)-n(c,\gamma)-n(d,\delta)} f_{(a,b),(c,d)}$$

2. $$E_{(A,B),(C,D)} = \sum_{a\in A, b\in B, c\in C, d\in D} \frac{2^{-n(a,\alpha)-n(b,\beta)-n(c,\gamma)-n(d,\delta)} f_{(a,b),(c,d)}}{(D_1 + min(D_2, D_3))^K}$$

$$K > 0$$

Motivation: longer distances have larger fluctuations

drawback: ΔE is not the variation of any functional

1. $$E_{(A,B),(C,D)} = \sum_{a\in A, b\in B, c\in C, d\in D} 2^{-n(a,\alpha)-n(b,\beta)-n(c,\gamma)-n(d,\delta)} f_{(a,b),(c,d)}$$

2. $$E_{(A,B),(C,D)} = \sum_{a\in A, b\in B, c\in C, d\in D} \frac{2^{-n(a,\alpha)-n(b,\beta)-n(c,\gamma)-n(d,\delta)} f_{(a,b),(c,d)}}{(D_1 + min(D_2, D_3))^K}$$

$$K > 0$$

Motivation: longer distances have larger fluctuations

drawback: ΔE is not the variation of any functional

3. $$E_{(A,B),(C,D)} = \sum_{a\in A, b\in B, c\in C, d\in D} \frac{f_{(a,b),(c,d)}}{(D_1 + min(D_2, D_3))^K}$$

ΔE is the variation of a functional
(same expression with the sum over all quadruplets a,b,c,d)

1. $$E_{(A,B),(C,D)} = \sum_{a\in A, b\in B, c\in C, d\in D} 2^{-n(a,\alpha)-n(b,\beta)-n(c,\gamma)-n(d,\delta)} f_{(a,b),(c,d)}$$

2. $$E_{(A,B),(C,D)} = \sum_{a\in A, b\in B, c\in C, d\in D} \frac{2^{-n(a,\alpha)-n(b,\beta)-n(c,\gamma)-n(d,\delta)} f_{(a,b),(c,d)}}{(D_1 + min(D_2, D_3))^K}$$

$$K > 0$$

Motivation: longer distances have larger fluctuations

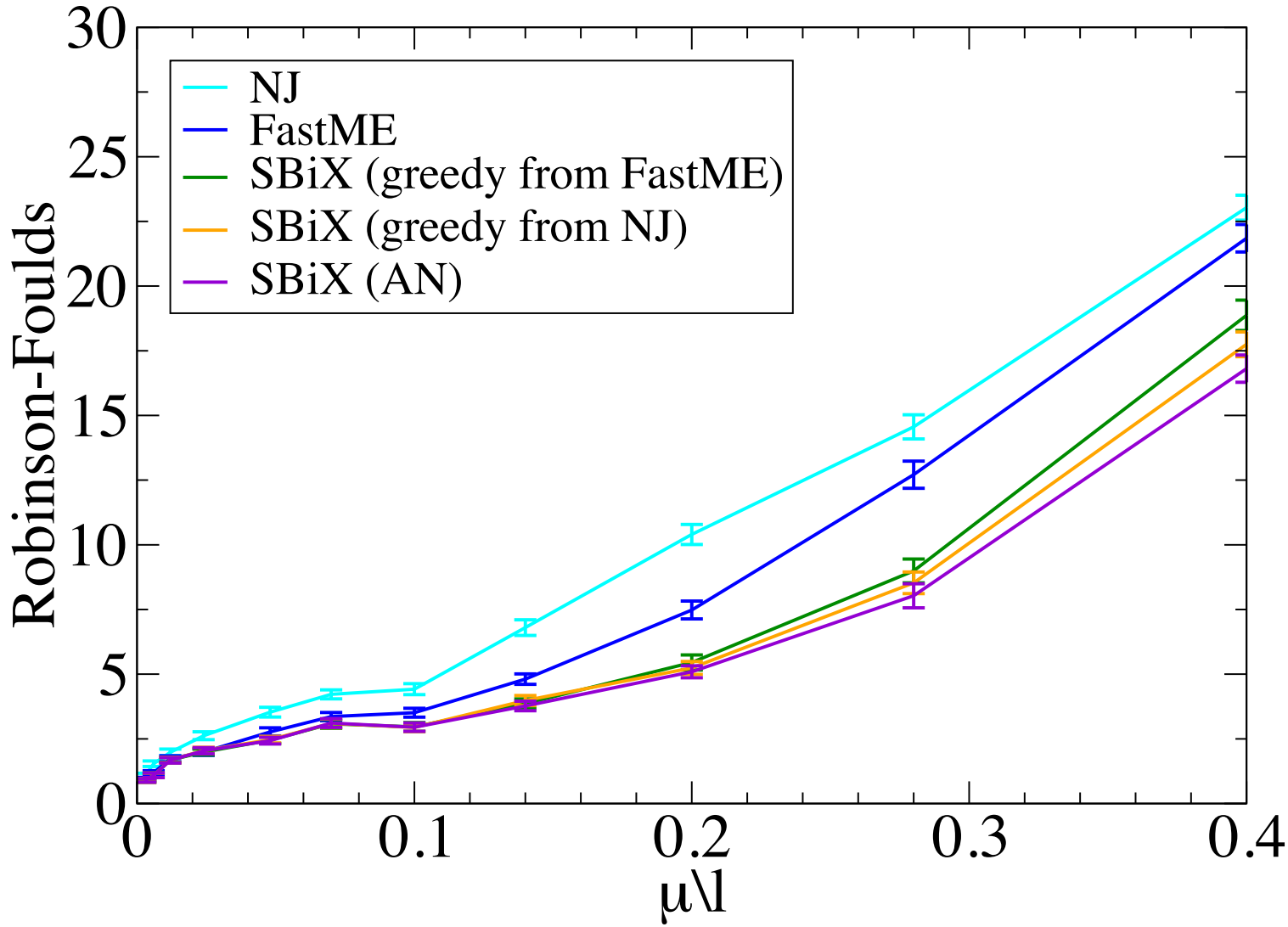
drawback: ΔE is not the variation of any functional

3. $$E_{(A,B),(C,D)} = \sum_{a\in A, b\in B, c\in C, d\in D} \frac{f_{(a,b),(c,d)}}{(D_1 + min(D_2, D_3))^K}$$

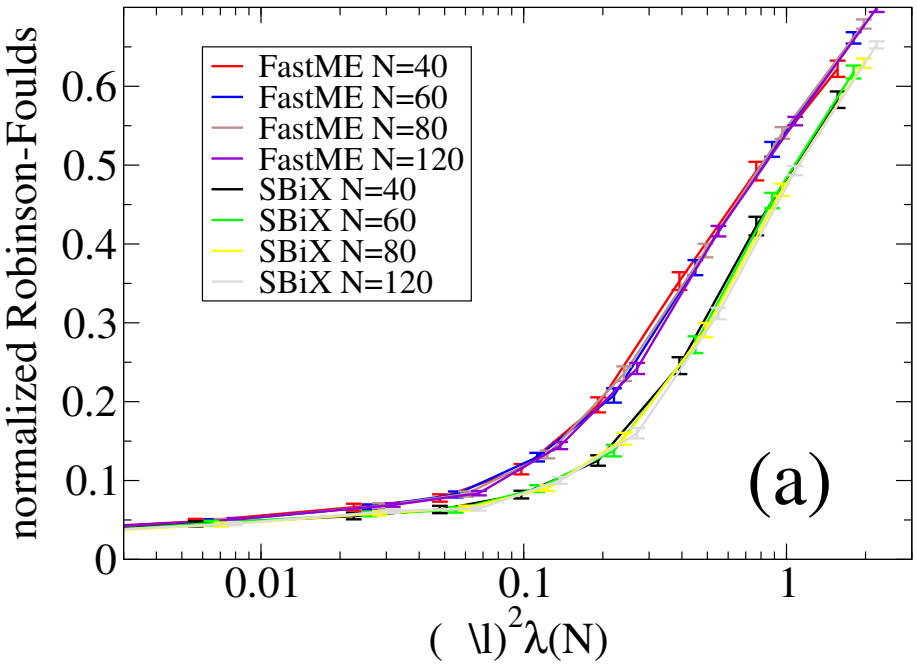
ΔE is the variation of a functional
(same expression with the sum over all quadruplets a,b,c,d)

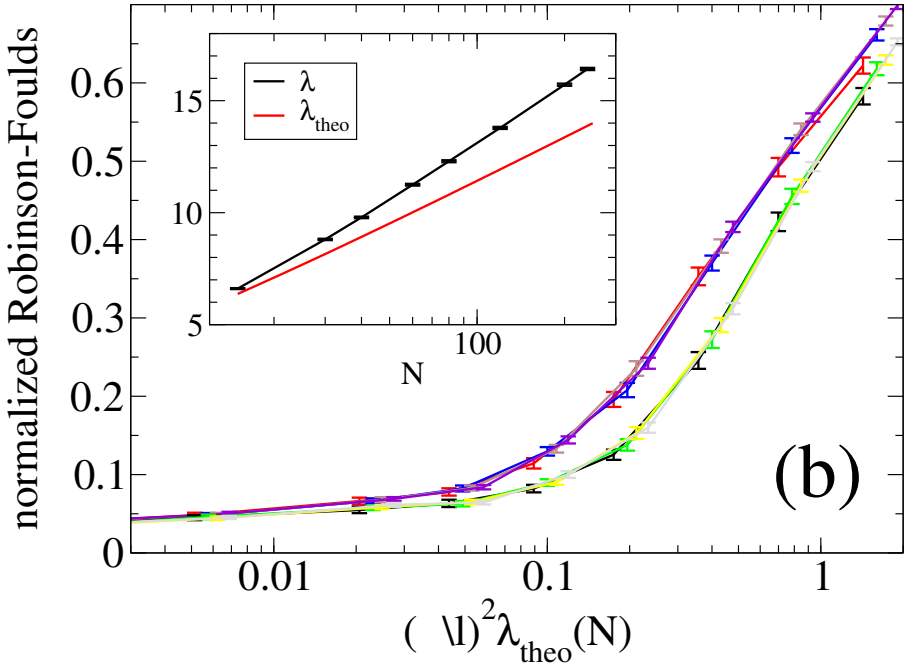
Why choosing 2.? It works better

Performances of distance-based reconstruction algorithms as a function of mutation rate per site



Scaling behavior with the number of taxa N





dependence on $\mu^2\lambda(N)$

$\lambda(N)$ mean distance between leaves

$$\lambda_{\text{theo}}(N) = \frac{2N(\log_2 N + 1) - 4N + 2}{N - 1}$$ for completely balanced trees

Practical drawback of SBiX: computational complexity *O*(N^4)

still good for large phylogenies (~ 1000 taxa)

but not for very large (>10000 taxa)

Practical drawback of SBiX:
computational complexity *O*(N^4)
still good for large phylogenies (~ 1000 taxa)
but not for very large (>10000 taxa)

Why is this important?

Practical drawback of SBiX:
computational complexity *O*(N^4)
still good for large phylogenies (~ 1000 taxa)
but not for very large (>10000 taxa)

Why is this important?

e.g. Protein interaction networks
Virus Phylogeny
Tree of life

Practical drawback of SBiX:
computational complexity *O*(N^4)
still good for large phylogenies (~ 1000 taxa)
but not for very large (>10000 taxa)

Why is this important?

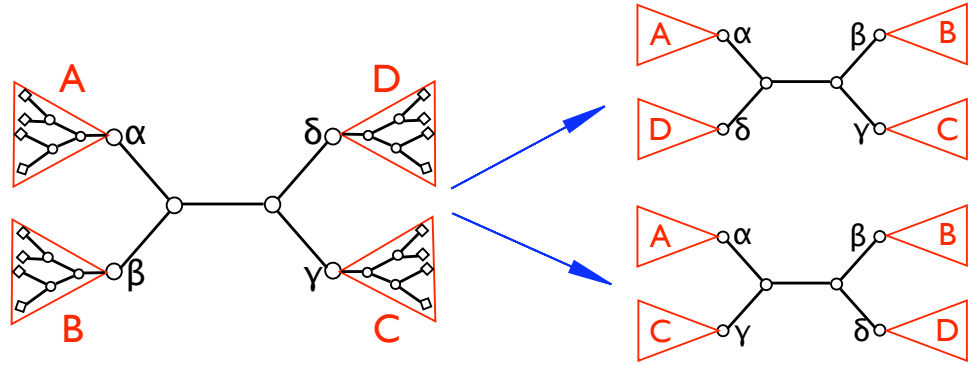
e.g. Protein interaction networks
Virus Phylogeny
Tree of life

In general, areas where the phylogeny
is important for prediction

Subtree representatives

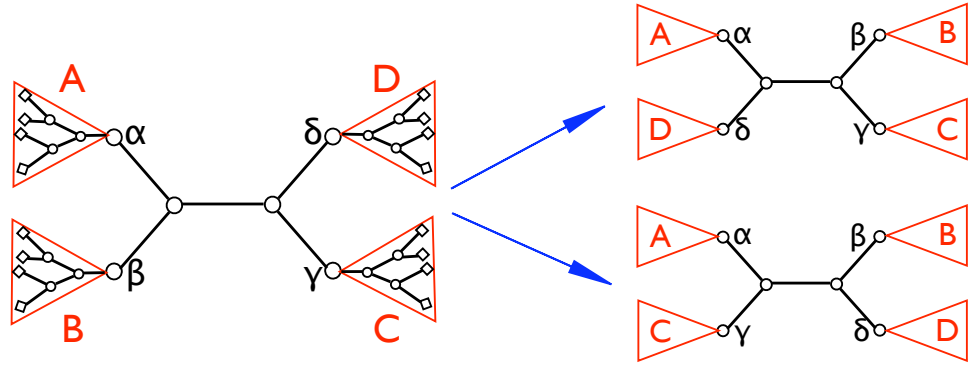
Subtree representatives

$$\mathcal{M}_1 = M_{\alpha,\beta} + M_{\gamma,\delta}$$
$$\mathcal{M}_2 = M_{\alpha,\gamma} + M_{\beta,\delta}$$
$$\mathcal{M}_3 = M_{\alpha,\delta} + M_{\beta,\gamma}$$



Subtree representatives

$$\mathcal{M}_1 = M_{\alpha,\beta} + M_{\gamma,\delta}$$
$$\mathcal{M}_2 = M_{\alpha,\gamma} + M_{\beta,\delta}$$
$$\mathcal{M}_3 = M_{\alpha,\delta} + M_{\beta,\gamma}$$

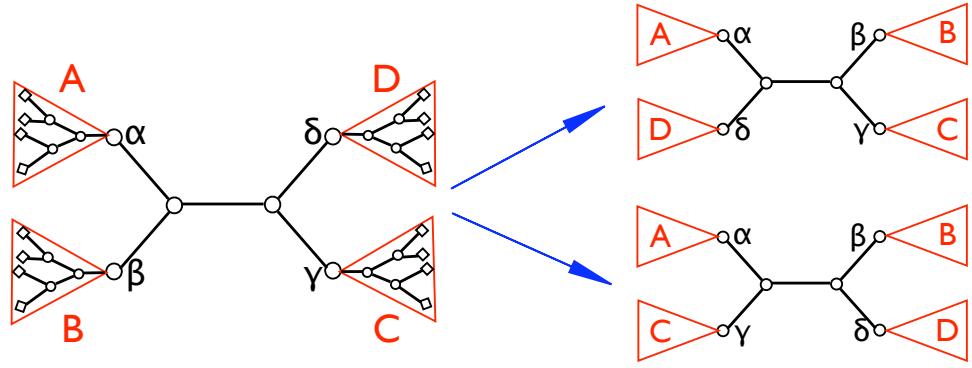


$$E_{((A,B),(C,D))} = \max(0, \mathcal{M}_1 - \min(\mathcal{M}_2, \mathcal{M}_3))$$

How is M defined?

Subtree representatives

$$\mathcal{M}_1 = M_{\alpha,\beta} + M_{\gamma,\delta}$$
$$\mathcal{M}_2 = M_{\alpha,\gamma} + M_{\beta,\delta}$$
$$\mathcal{M}_3 = M_{\alpha,\delta} + M_{\beta,\gamma}$$



$$E_{((A,B),(C,D))} = \max(0, \mathcal{M}_1 - \min(\mathcal{M}_2, \mathcal{M}_3))$$

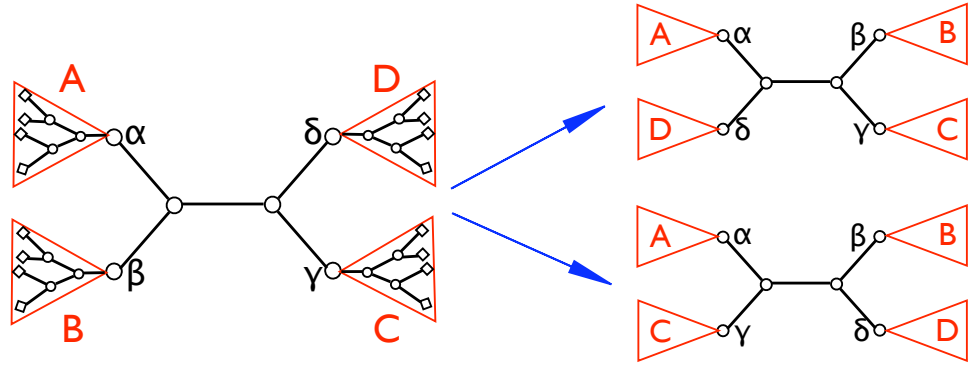
How is M defined?

If $\quad M_{\alpha,\beta} = \sum_{a \in A, b \in B} \mathcal{D}(a,b)\, 2^{-t(a,\alpha)-t(b,\beta)}$

$$\Delta E = 4L_P$$

Subtree representatives

$$\mathcal{M}_1 = M_{\alpha,\beta} + M_{\gamma,\delta}$$
$$\mathcal{M}_2 = M_{\alpha,\gamma} + M_{\beta,\delta}$$
$$\mathcal{M}_3 = M_{\alpha,\delta} + M_{\beta,\gamma}$$



$$E_{((A,B),(C,D))} = \max(0, \mathcal{M}_1 - \min(\mathcal{M}_2, \mathcal{M}_3))$$

How is M defined?

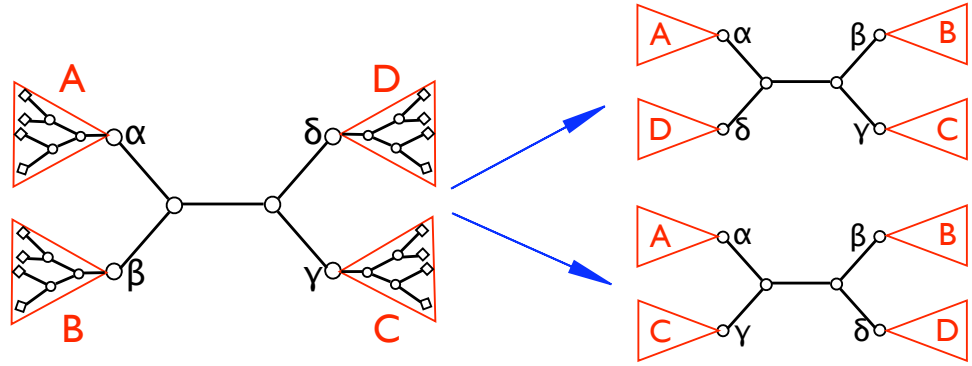
If $M_{\alpha,\beta} = \sum_{a \in A, b \in B} \mathcal{D}(a,b)\, 2^{-t(a,\alpha)-t(b,\beta)}$

$\Delta E = 4L_P$

Introduce the idea that longer distances have larger fluctuations

Subtree representatives

$$\mathcal{M}_1 = M_{\alpha,\beta} + M_{\gamma,\delta}$$
$$\mathcal{M}_2 = M_{\alpha,\gamma} + M_{\beta,\delta}$$
$$\mathcal{M}_3 = M_{\alpha,\delta} + M_{\beta,\gamma}$$



$$E_{((A,B),(C,D))} = \max(0, \mathcal{M}_1 - \min(\mathcal{M}_2, \mathcal{M}_3))$$

How is M defined?

If $$M_{\alpha,\beta} = \sum_{a\in A, b\in B} \mathcal{D}(a,b)\, 2^{-t(a,\alpha)-t(b,\beta)}$$

$$\Delta E = 4L_P$$

Introduce the idea that longer distances have larger fluctuations

for each leaf *a* we define a *length*

$$l_a = \frac{1}{N_B}\sum_{b\in B}\mathcal{D}(a,b) + \frac{1}{N_C}\sum_{c\in C}\mathcal{D}(a,c) + \frac{1}{N_D}\sum_{d\in D}\mathcal{D}(a,d)$$

the idea is filtering leaves' contributions using this length

Introduce a weight $p_a = f\left(l^A_{\min}/l_a\right)$

Introduce a weight $p_a = f\left(l^A_{\min}/l_a\right)$

simple case $p_a = \theta\left(l^A_{\min}/l_a - l_t\right)$

Introduce a weight $p_a = f\left(l^A_{\min}/l_a\right)$

simple case $p_a = \theta\left(l^A_{\min}/l_a - l_t\right)$

A better approach is to use a smooth weight, e.g.

$p_a = \left(l^A_{\min}/l_a\right)^k \quad k > 0$

Introduce a weight $p_a = f\left(l^A_{\min}/l_a\right)$

simple case $p_a = \theta\left(l^A_{\min}/l_a - l_t\right)$

A better approach is to use a smooth weight, e.g.

$p_a = \left(l^A_{\min}/l_a\right)^k \quad k > 0$

then $M_{\alpha,\beta} = \sum_{a\in A, b\in B} \mathcal{D}(a,b) w_a w_b p_a p_b$

with w *effective* Pauplin's weights

Introduce a weight $p_a = f\left(l^A_{\min}/l_a\right)$

simple case $p_a = \theta\left(l^A_{\min}/l_a - l_t\right)$

A better approach is to use a smooth weight, e.g.

$p_a = \left(l^A_{\min}/l_a\right)^k \quad k > 0$

then $M_{\alpha,\beta} = \sum_{a\in A, b\in B} \mathcal{D}(a,b) w_a w_b p_a p_b$

with w *effective* Pauplin's weights

Identify the weight p_a with the probability of considering the leaf *a* when defining the subtree representative

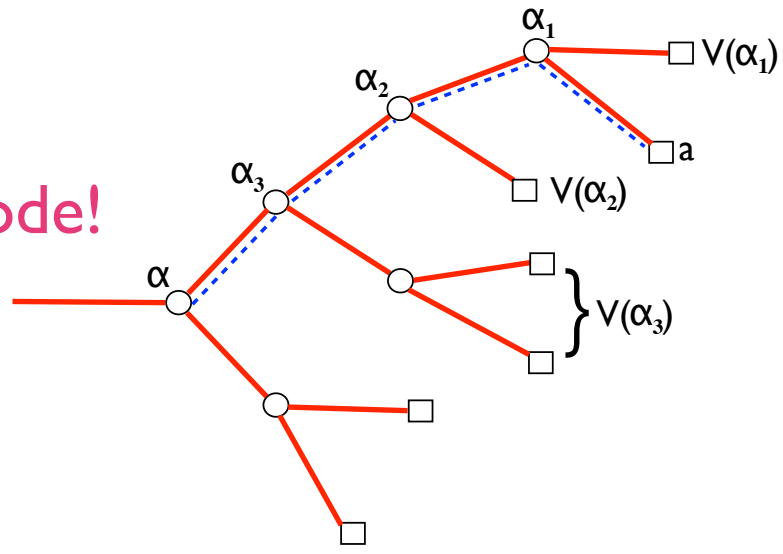
Compute the effective Pauplin weights
by using a suitable partition function

$$w_a = \sum_{\{\sigma(\alpha_i)\}} \mathrm{P}(\alpha_1, \alpha_2, \alpha_3) 2^{-\sigma(\alpha_1)-\sigma(\alpha_2)-\sigma(\alpha_3)}$$

$$\sigma(\alpha_i) \in \{0, 1\}$$

but the probability factorizes on each node!

$$w_a = \frac{1}{2^{n(\mathcal{P}_a)}} \prod_{\alpha_i \in \mathcal{P}_a} \left(1 + \prod_{a_i \in V(\alpha_i)} (1 - p_{a_i}) \right)$$



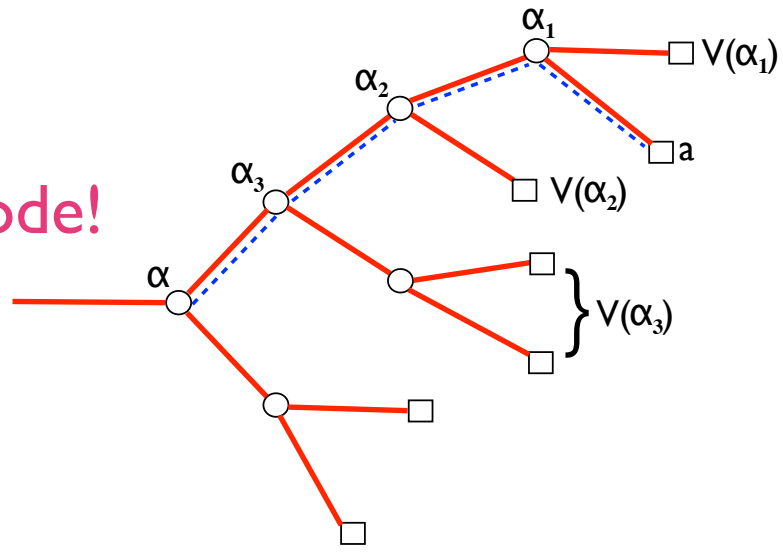
$$w_a = \sum_{\{\sigma(\alpha_i)\}} \mathrm{P}(\alpha_1, \alpha_2, \alpha_3) 2^{-\sigma(\alpha_1)-\sigma(\alpha_2)-\sigma(\alpha_3)}$$

$$\sigma(\alpha_i) \in \{0, 1\}$$

but the probability factorizes on each node!

$$w_a = \frac{1}{2^{n(\mathcal{P}_a)}} \prod_{\alpha_i \in \mathcal{P}_a} \left(1 + \prod_{a_i \in V(\alpha_i)} (1 - p_{a_i}) \right)$$

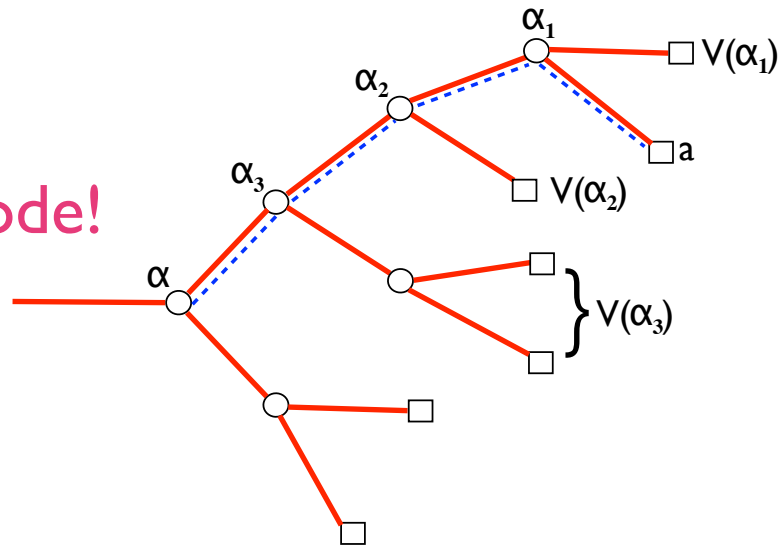
number of nodes in the path from *a* to α



$$w_a = \sum_{\{\sigma(\alpha_i)\}} \mathrm{P}(\alpha_1, \alpha_2, \alpha_3) 2^{-\sigma(\alpha_1)-\sigma(\alpha_2)-\sigma(\alpha_3)}$$

$$\sigma(\alpha_i) \in \{0, 1\}$$

but the probability factorizes on each node!



$$w_a = \frac{1}{2^{n(\mathcal{P}_a)}} \prod_{\alpha_i \in \mathcal{P}_a} \left(1 + \prod_{a_i \in V(\alpha_i)} (1 - p_{a_i}) \right)$$

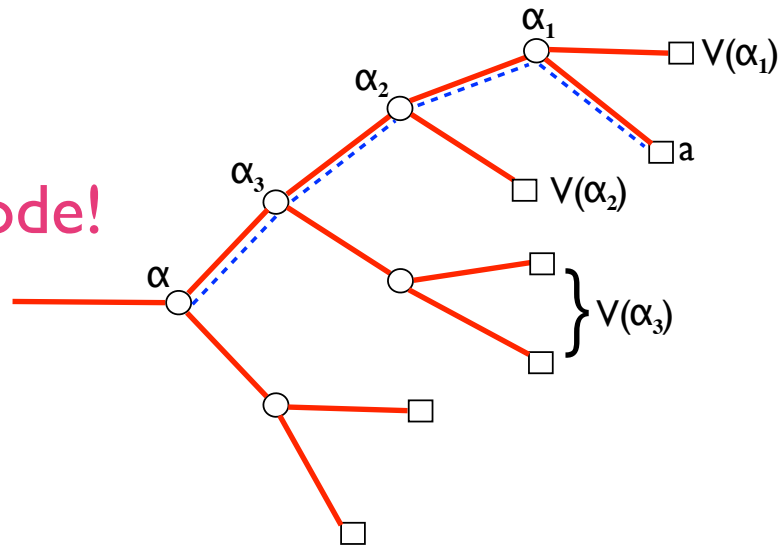
probability that the node α_i doesn't exist

number of nodes in the path from *a* to α

$$w_a = \sum_{\{\sigma(\alpha_i)\}} \mathrm{P}(\alpha_1, \alpha_2, \alpha_3) 2^{-\sigma(\alpha_1)-\sigma(\alpha_2)-\sigma(\alpha_3)}$$

$$\sigma(\alpha_i) \in \{0, 1\}$$

but the probability factorizes on each node!



$$w_a = \frac{1}{2^{n(\mathcal{P}_a)}} \prod_{\alpha_i \in \mathcal{P}_a} \left(1 + \prod_{a_i \in V(\alpha_i)} (1 - p_{a_i}) \right)$$

probability that the node α_i doesn't exist

number of nodes in the path from *a* to α

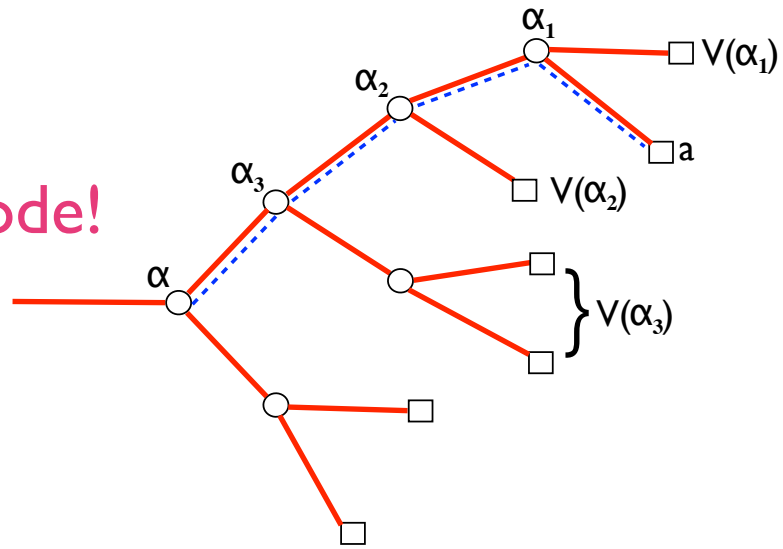
drawback: ΔE is not the variation of any functional

⟶ weights depend on the chosen edge!

$$w_a = \sum_{\{\sigma(\alpha_i)\}} \mathrm{P}(\alpha_1, \alpha_2, \alpha_3) 2^{-\sigma(\alpha_1)-\sigma(\alpha_2)-\sigma(\alpha_3)}$$

$$\sigma(\alpha_i) \in \{0, 1\}$$

but the probability factorizes on each node!



$$w_a = \frac{1}{2^{n(\mathcal{P}_a)}} \prod_{\alpha_i \in \mathcal{P}_a} \left(1 + \prod_{a_i \in V(\alpha_i)} (1 - p_{a_i}) \right)$$

probability that the node α_i doesn't exist

number of nodes in the path from *a* to α

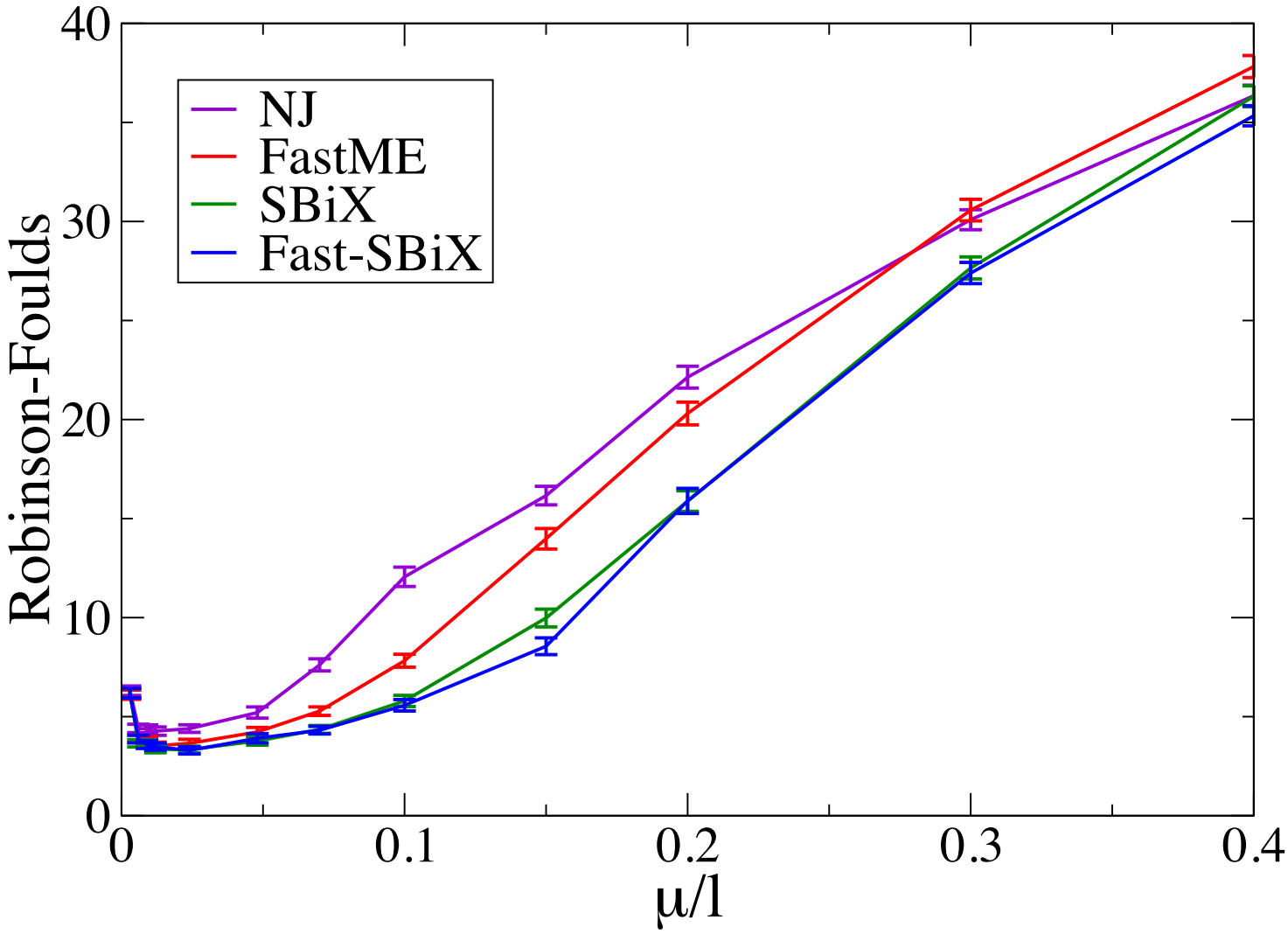
drawback: ΔE is not the variation of any functional

→ weights depend on the chosen edge!

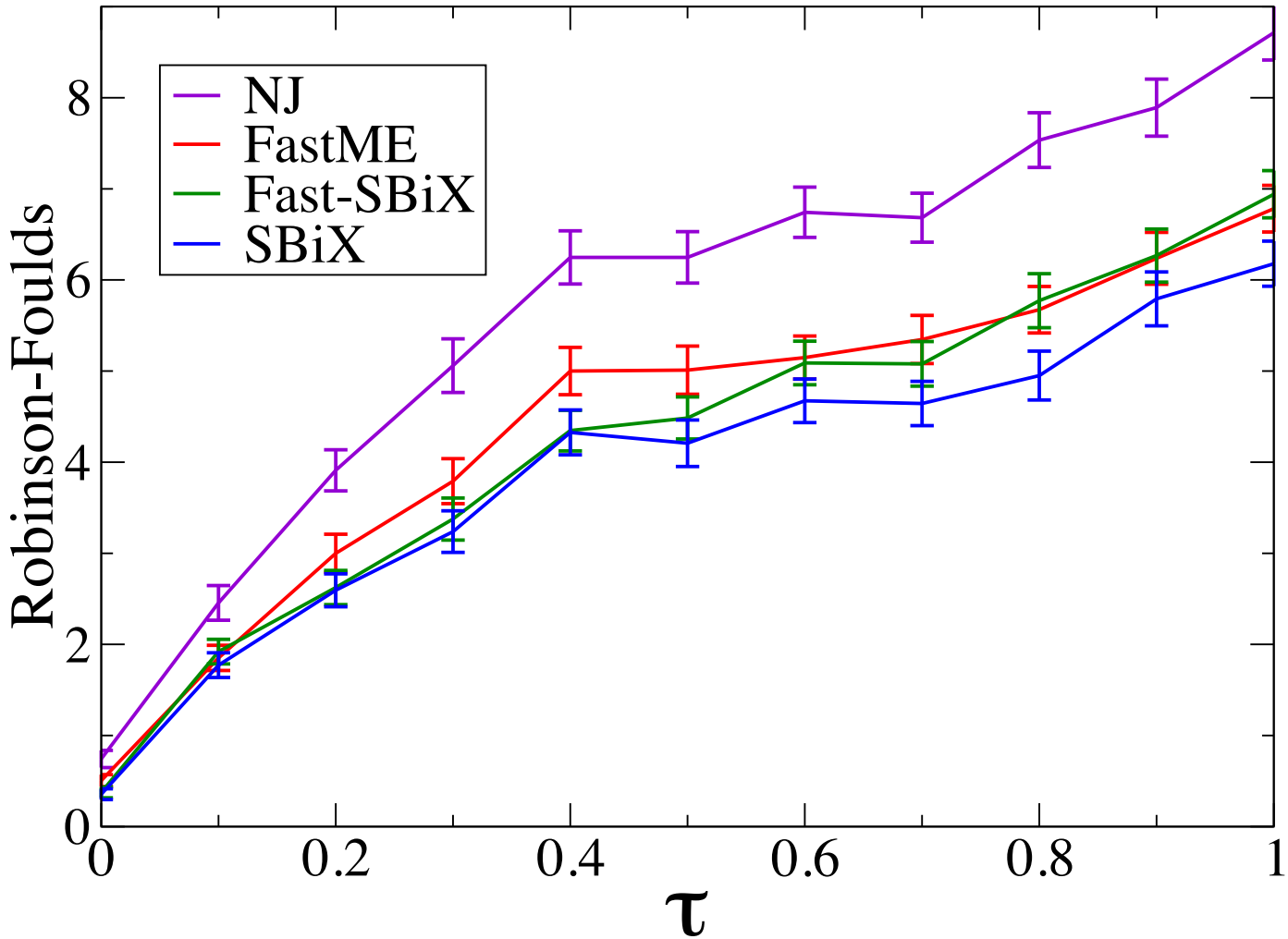
positive facts:

1. It works
2. computational complexity *O*(N^2 log(N))
3. general method to weigh taxa

Performances of distance-based reconstruction algorithms as a function of mutation rate per site



Performances of distance-based reconstruction algorithms as a function of the horizontal transfer rate



The benchmarking problem

Copystree

Copystree

Is a web experiment -game- aimed to provide
a completely controlled and model-free phylogeny

Copystree

Is a web experiment -game- aimed to provide
a completely controlled and model-free phylogeny

Players (copysts) have few minutes to copy a given text.
Texts evolve by:

- copying
- degradation

Copystree

Is a web experiment -game- aimed to provide
a completely controlled and model-free phylogeny

Players (copysts) have few minutes to copy a given text.
Texts evolve by:

- copying ⟵ human ability
- degradation ⟵ parameters

Copystree

Is a web experiment -game- aimed to provide
a completely controlled and model-free phylogeny

Players (copysts) have few minutes to copy a given text.
Texts evolve by:

- copying ⟵ human ability
- degradation ⟵ parameters

O dolce amor de caritate, pregote che me te lassi tenere, se te piace che da ti non me deggia mai partire. Lo dolce amore che me se è dato, madonna, puoi che me l'hai prestato, non me lo tollere questa fiata puoi che l'agio desiderato; puoi che ce so' venuta e che me te si' dato, non me voglio più partire. O dolce amore smesurato che te si' umiliato ed a questa misera te si dato. O dolce matre non me lo retollere, lo voglio tenere che me confuorti e che ma mea mente allustri. Puoi che m'hai reconsolata ed ame allustrata, no lo retollere se te piace. O dolce amore de grande confuorto che resusciti chi è muorto. O dolce amor de veritate che dai lume alli accecati, illuminame se te piace.

Copystree

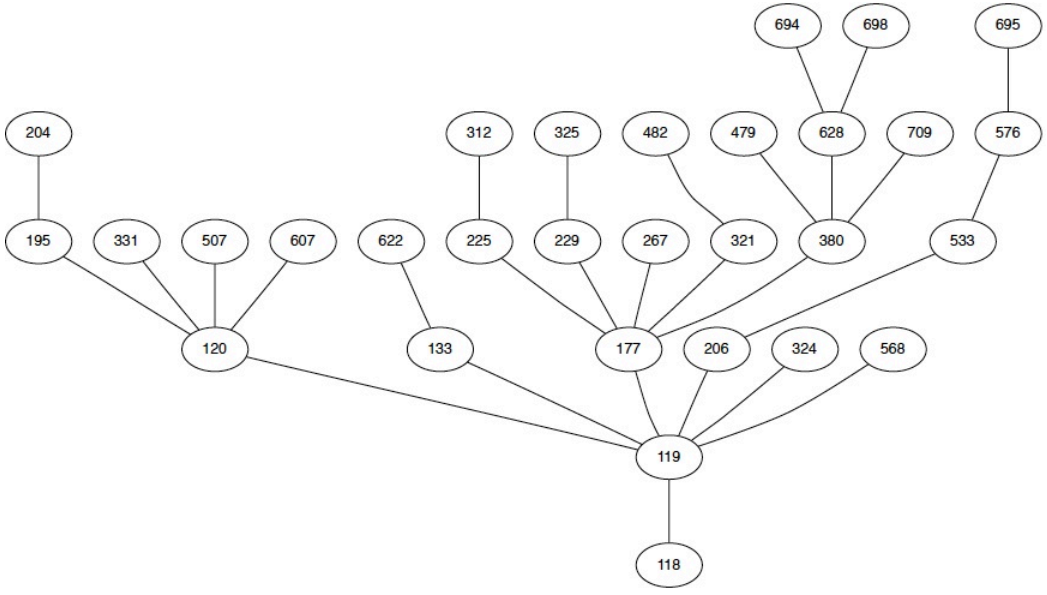
Is a web experiment -game- aimed to provide
a completely controlled and model-free phylogeny

Players (copysts) have few minutes to copy a given text.
Texts evolve by:

- copying ⟵ human ability
- degradation ⟵ parameters

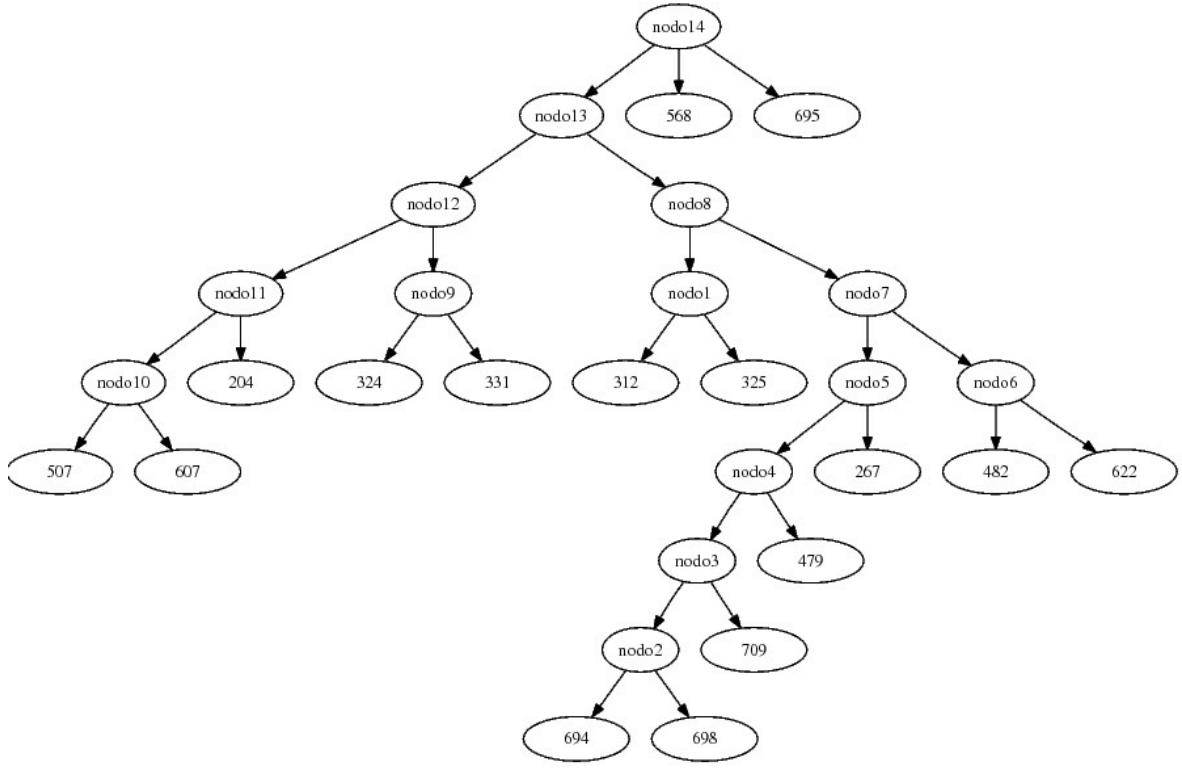
TIME

O dolce amor de caritate, pregote che me te lassi tenere, se te piace che da ti non me deggia mai partire. Lo dolce amore ch[illegible]e se è dato, madonna, puoi che me l'hai prestato, non me lo tollere questa fiata puoi che l'agio desider[illegible]uoi che ce so' venuta e che [illegible]i' dato, non me voglio più partire. O dolce [illegible] smesurato che te si' umiliat[illegible] questa misera te si da[illegible] O dolce matre non me lo retollere, lo voglio tenere che me confuorti e che ma[illegible]nea mente allustri. Puoi che m'hai reconsolata ed ame allustrata, no lo retollere se te piace. O dolce amore de grande confuorto che resusciti chi è muorto. O dolce amor de veritate che dai lume alli accecati, illuminame se te piace.



real phylogeny

reconstructed phylogeny



An application in linguistics: The tree of languages

Swadesh lists

Lists of words representing a language

Swadesh lists

Lists of words representing a language

- First attempt: cognate words → distance 0/1
- phonological characters

Swadesh lists

Lists of words representing a language

- First attempt: cognate words → distance 0/1
- phonological characters

I	ei	io
You	yu	tu
We	wi	noi
Ear	ir	oreky~o
Eye	ei	oky~o

→ Levenshtein (edit) distance:
number of insertions, deletions or substitutions
to go from a word to another

Swadesh lists

Lists of words representing a language

- First attempt: cognate words → distance 0/1
- phonological characters

I	ei	io
You	yu	tu
We	wi	noi
Ear	ir	oreky~o
Eye	ei	oky~o

distance between languages:
average Levenshtein distance
between homologous words

→ Levenshtein (edit) distance:
number of insertions, deletions or substitutions
to go from a word to another

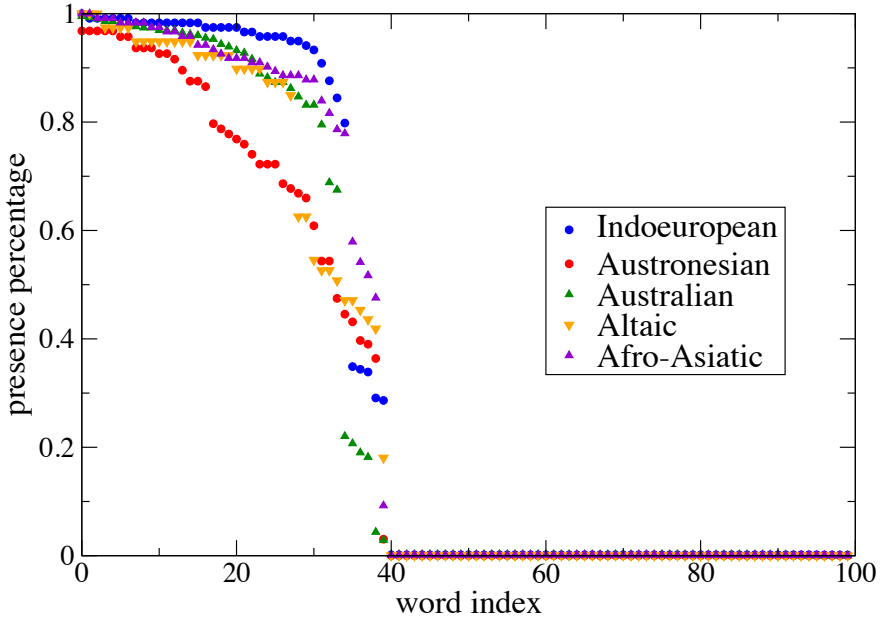
ASJP (*Automated Similarity Judgment Program)* database

50 language families
languages per family varing from *O*(10) to *O*(100)
each language list 100 homologous words (but incomplete!)

ASJP (*Automated Similarity Judgment Program*) database

50 language families
languages per family varing from *O*(10) to *O*(100)
each language list 100 homologous words (but incomplete!)

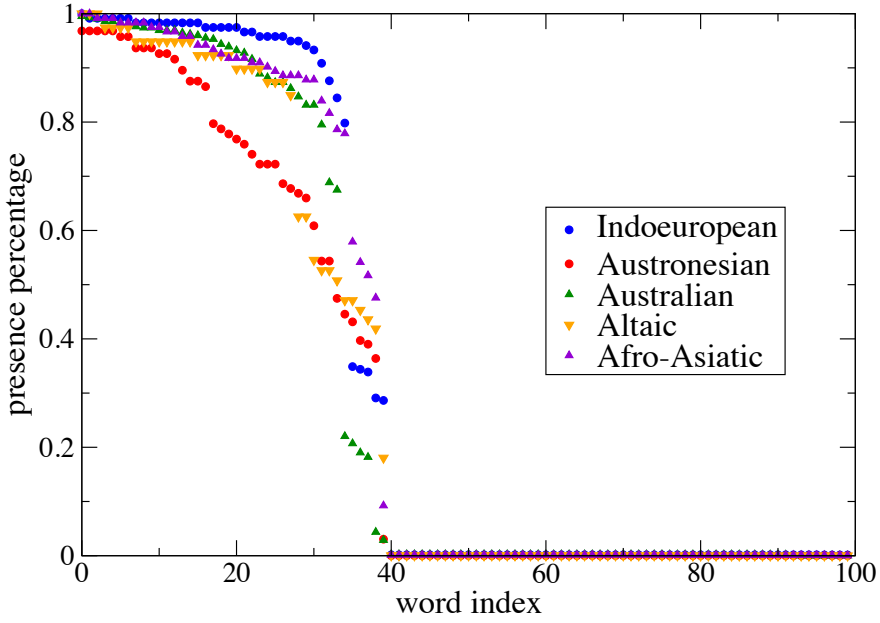
40 words common
to almost all languages
in each family

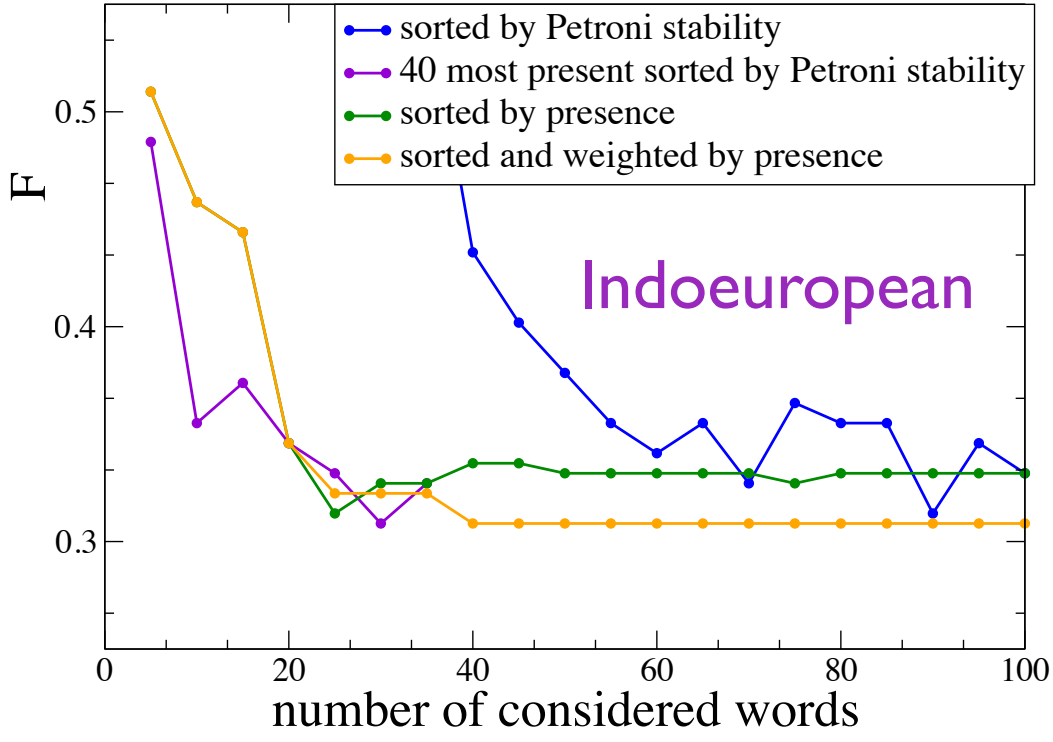


ASJP (*Automated Similarity Judgment Program*) database

50 language families
languages per family varing from *O*(10) to *O*(100)
each language list 100 homologous words (but incomplete!)

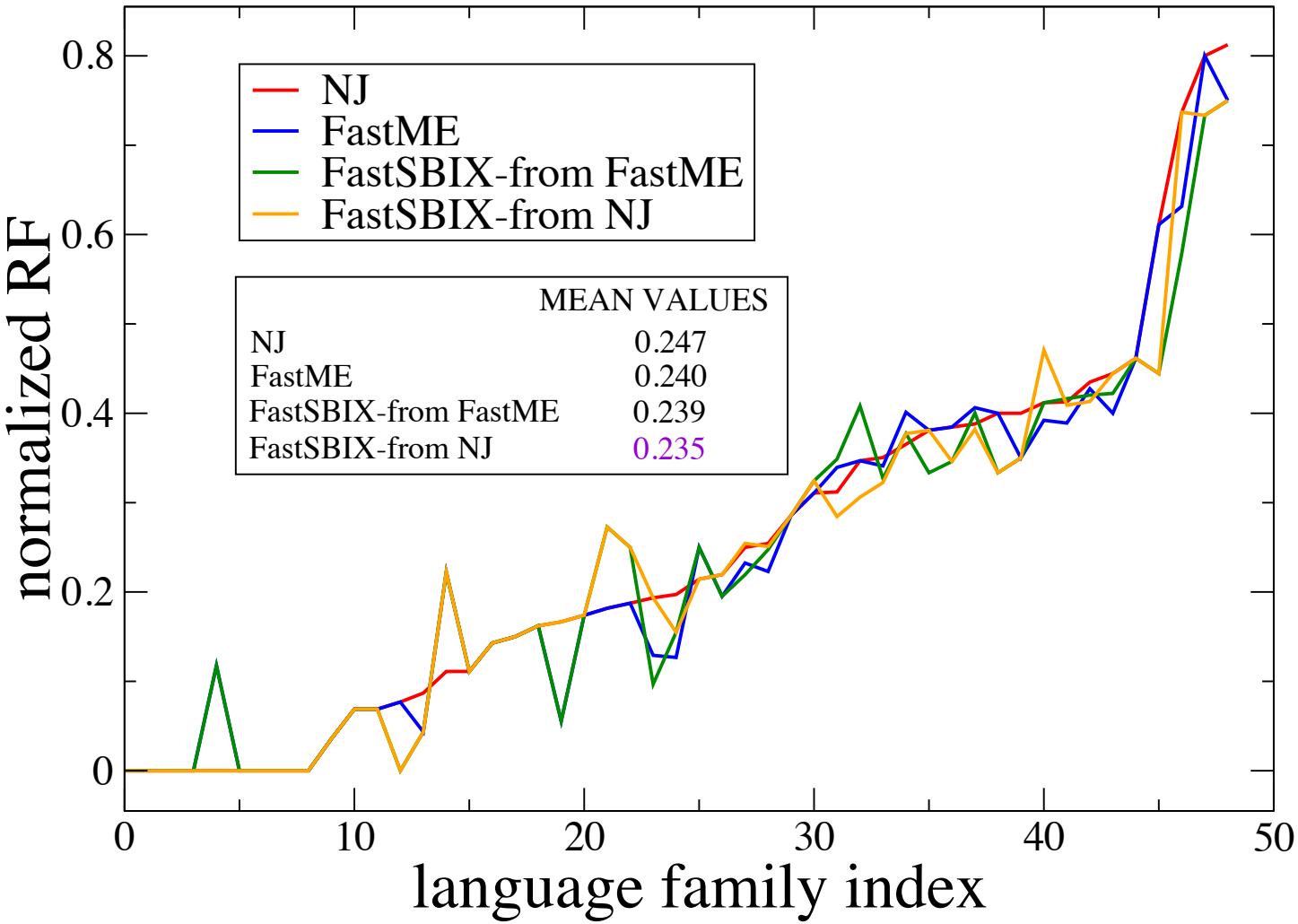
40 words common
to almost all languages
in each family





need of complete dataset!

Comparison with Ethnologue (experts) classification



No much difference in performances between reconstruction algorithms: too much noise or too short and/or incomplete lists?

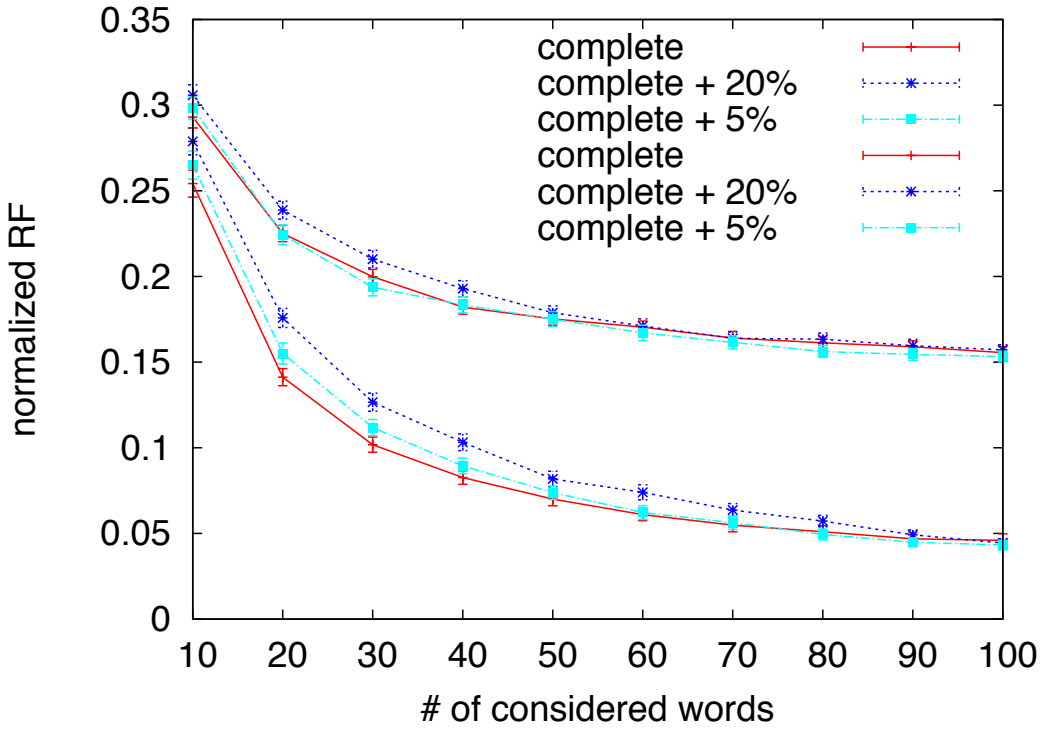
Artificial lists of words

evolving through mutation, deletion and insertion

Artificial lists of words

evolving through mutation, deletion and insertion

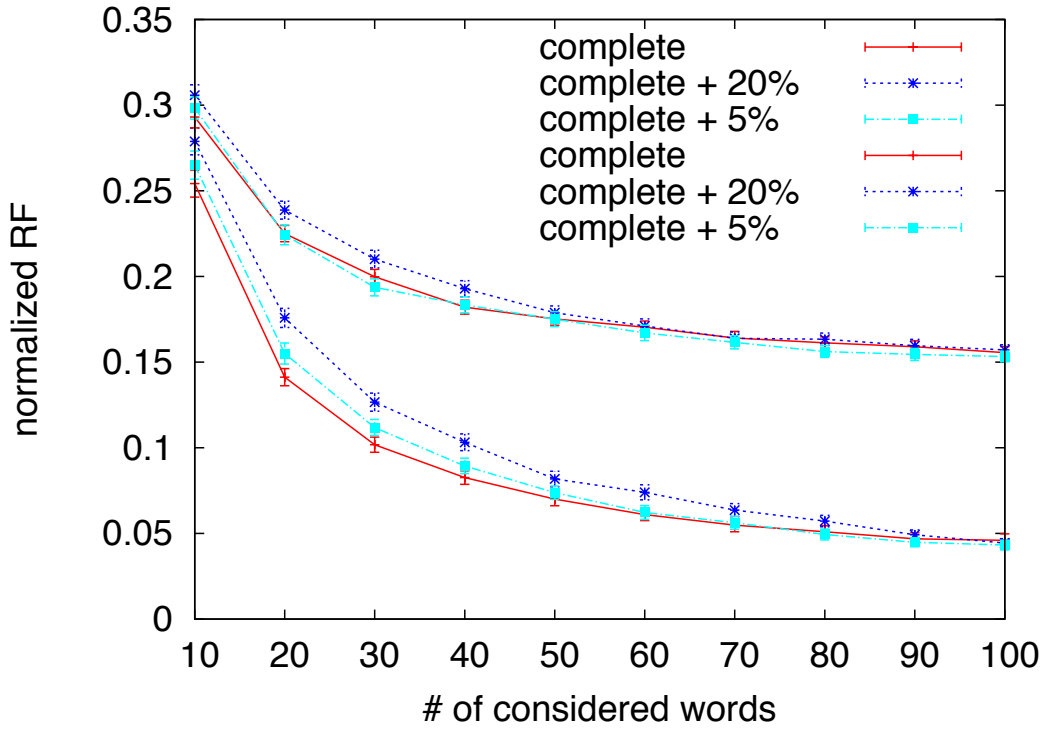
N=100 languages, lists of 100 words
common to all languages
+100 words of which 80%
is randomly and independently
deleted from each language lists

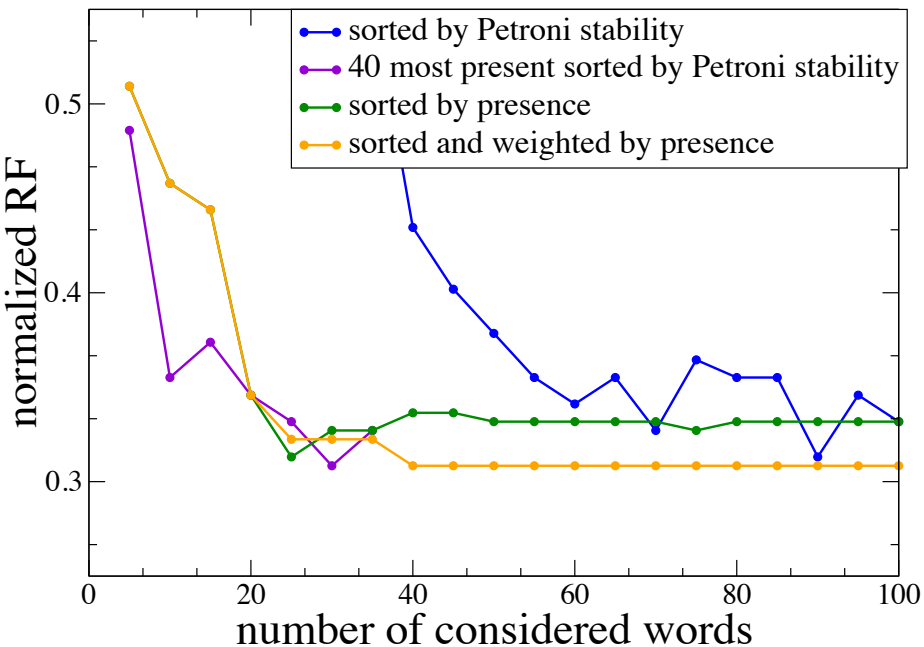


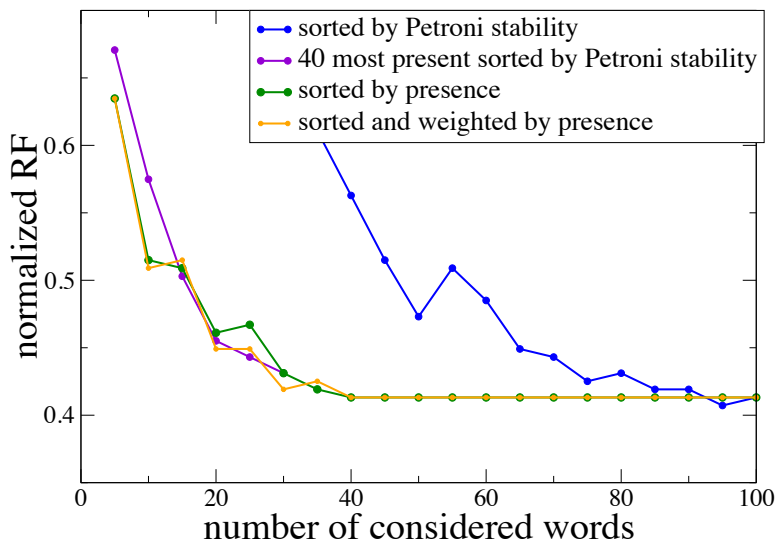
Artificial lists of words

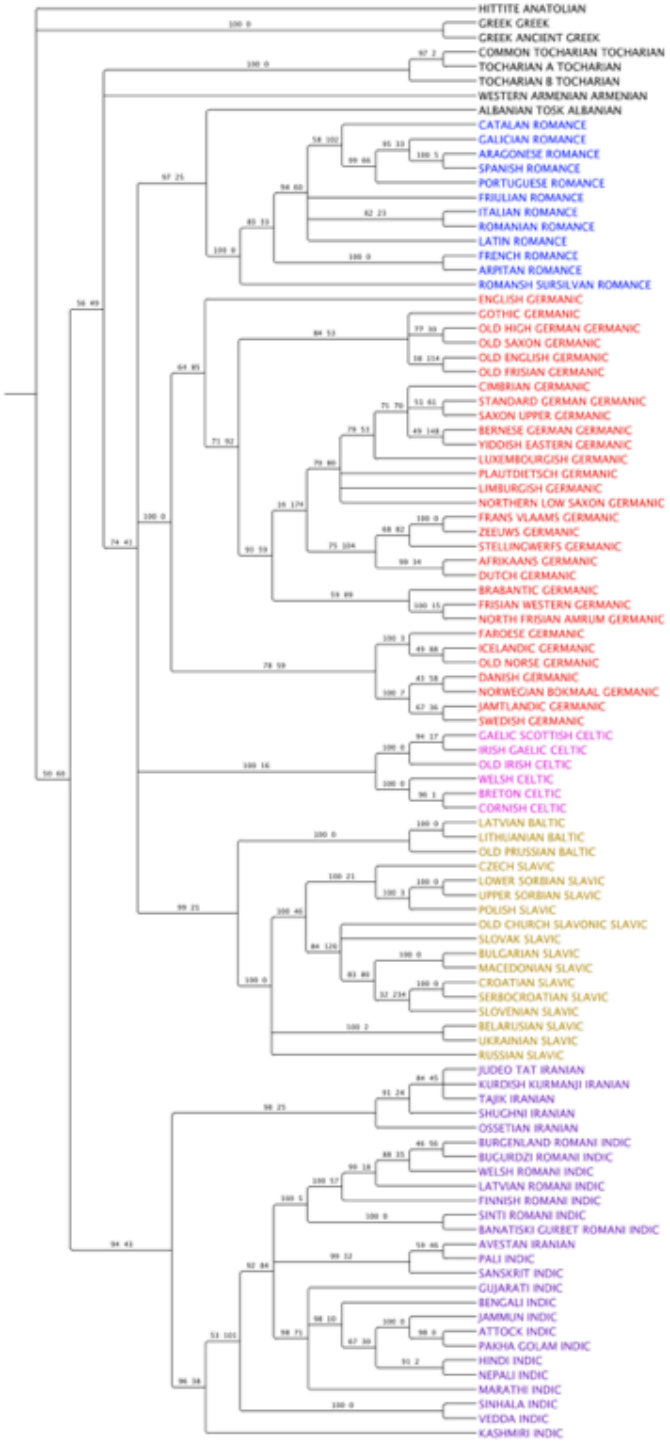
evolving through mutation, deletion and insertion

N=100 languages, lists of 100 words
common to all languages
+100 words of which 80%
is randomly and independently
deleted from each language lists









INDO-EUROPEAN

Romance

Germanic

Celtic

Balto-Slavic

Indo-Iranian

An application in biology: Influenza virus evolution

Extract information about the phylogenetic process from phylogenetic tree shape

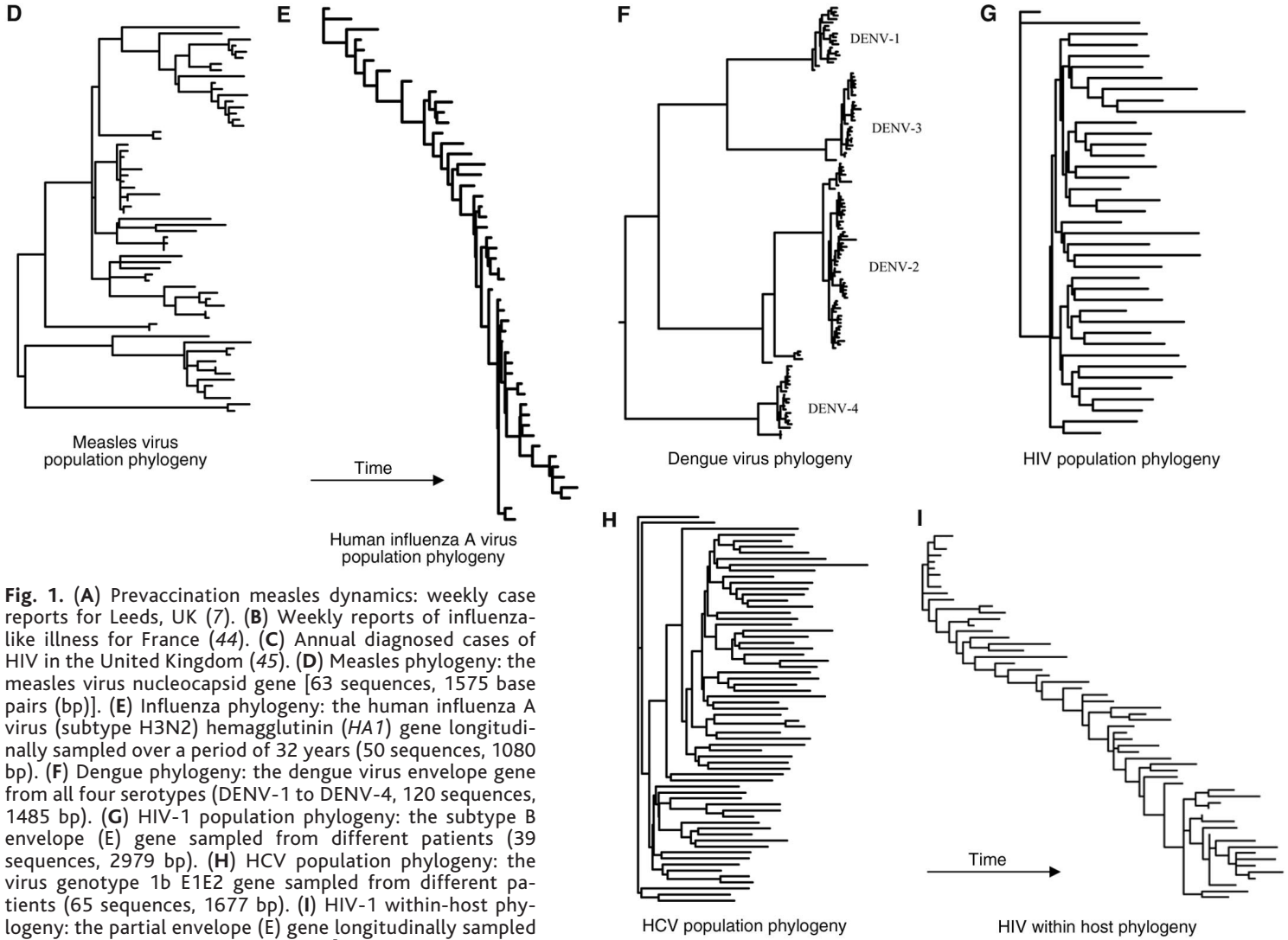
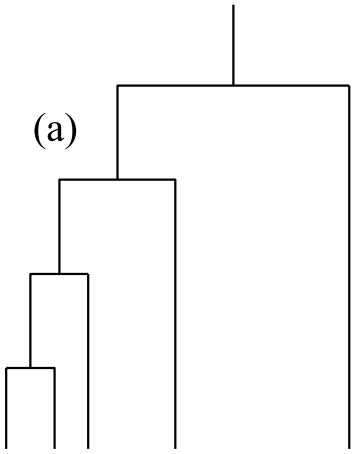


Fig. 1. (**A**) Prevaccination measles dynamics: weekly case reports for Leeds, UK (*7*). (**B**) Weekly reports of influenza-like illness for France (*44*). (**C**) Annual diagnosed cases of HIV in the United Kingdom (*45*). (**D**) Measles phylogeny: the measles virus nucleocapsid gene [63 sequences, 1575 base pairs (bp)]. (**E**) Influenza phylogeny: the human influenza A virus (subtype H3N2) hemagglutinin (*HA1*) gene longitudinally sampled over a period of 32 years (50 sequences, 1080 bp). (**F**) Dengue phylogeny: the dengue virus envelope gene from all four serotypes (DENV-1 to DENV-4, 120 sequences, 1485 bp). (**G**) HIV-1 population phylogeny: the subtype B envelope (E) gene sampled from different patients (39 sequences, 2979 bp). (**H**) HCV population phylogeny: the virus genotype 1b E1E2 gene sampled from different patients (65 sequences, 1677 bp). (**I**) HIV-1 within-host phylogeny: the partial envelope (E) gene longitudinally sampled from a single patient over 5.8 years [58 sequences, 627 bp; patient 6 from (*26*)]. All sequences were collected from GenBank and trees were constructed with maximum likelihood in PAUP* (*46*). Horizontal branch lengths are proportional to substitutions per site. Further details are available from the authors on request.

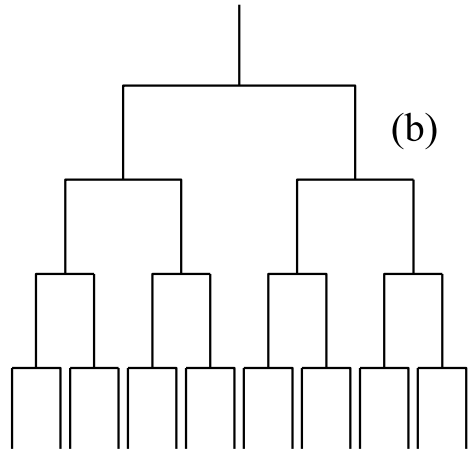
Balance/unbalance measures

A_i number of taxa diversifying from node *i*, including itself

$$C_i = \sum_j A_j$$

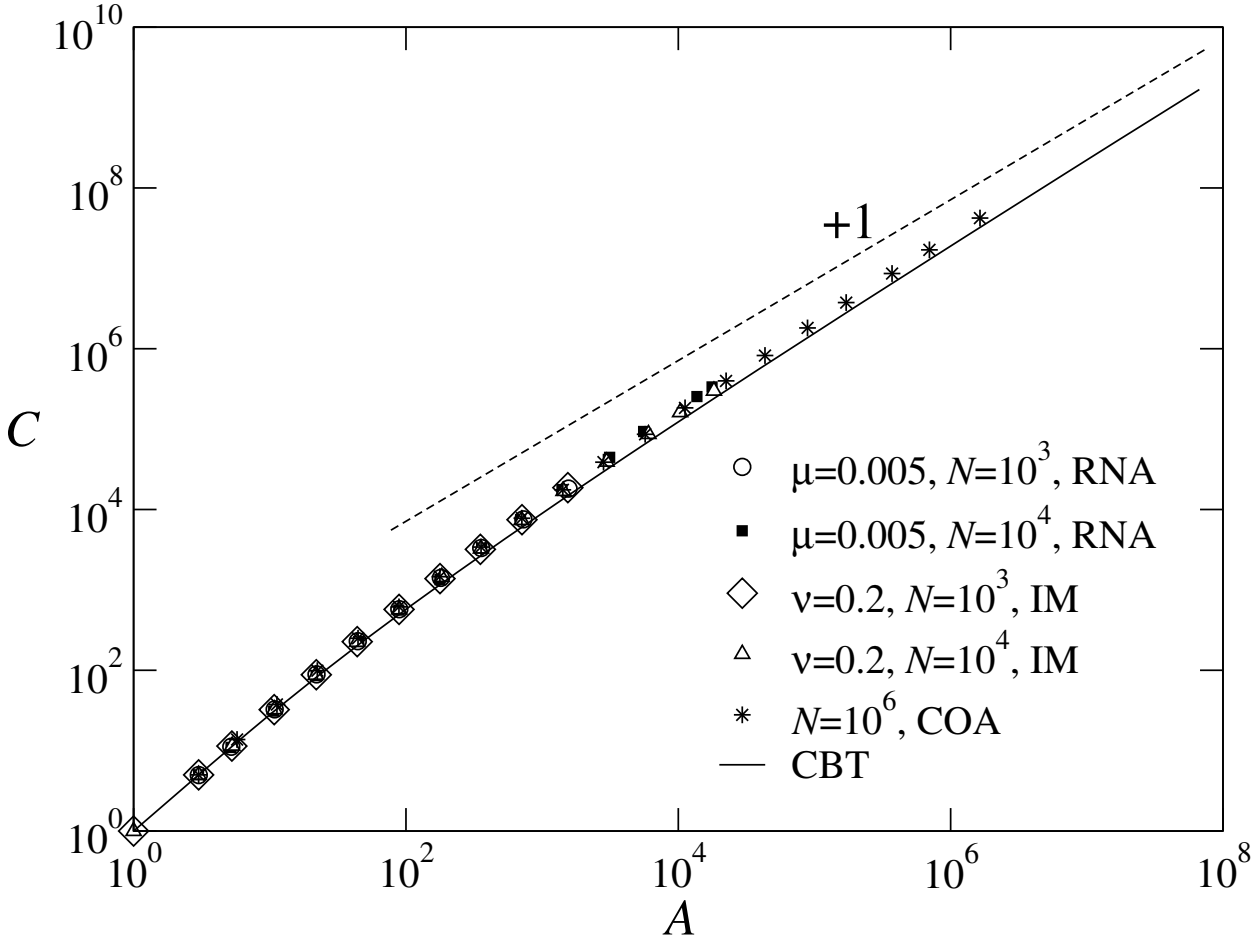


$N_u(l)$	$A_u(l)$	$C_u(l)$	l	$N_b(l)$	$A_b(l)$	$C_b(l)$
1	9	29	5	$n/16$	31	129
1	7	19	4	$n/8$	15	49
1	5	11	3	$n/4$	7	17
1	3	5	2	$n/2$	3	5
n	1	1	1	n	1	1



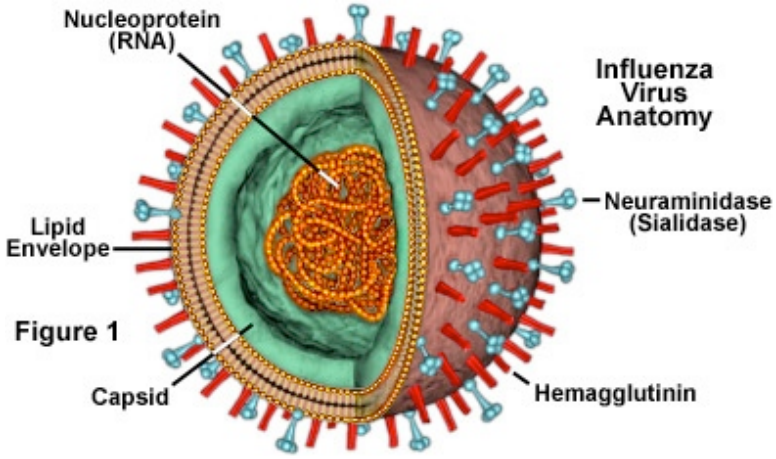
Asymptotic behavior

$$P(A) \propto A^{-\alpha} \quad P(C) \propto C^{-\gamma} \quad C(A) \propto A^{\eta}, \;\; \eta = \frac{1-\alpha}{1-\gamma}$$

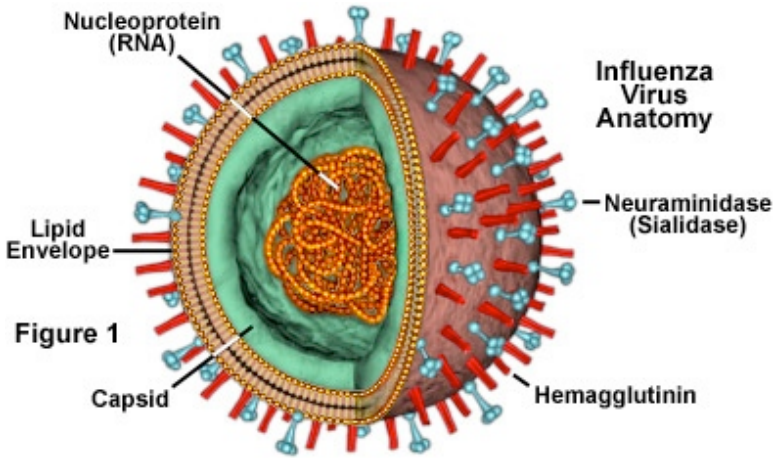


M. Stich and S.C. Manrubia:
Eur. Phys. J. B **70**, 583–592 (2009)

Almost all evolutive process produce asymptotically balanced tress



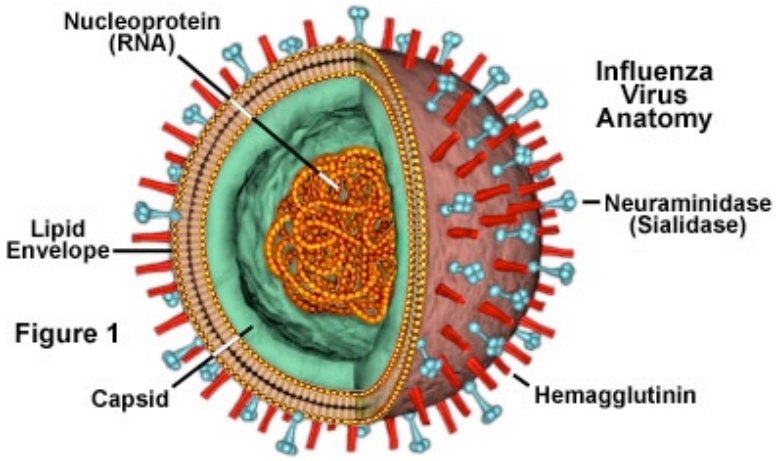
Influenza A virus
RNA in 8 segments
(~10000 nucleotides)



Influenza A virus
RNA in 8 segments
(~10000 nucleotides)

HA (Hemagglutinin) and
(NA) Neuraminidase
are the surface proteins
responsible for the interaction
with host immune system
(~1000 nucleotides each)

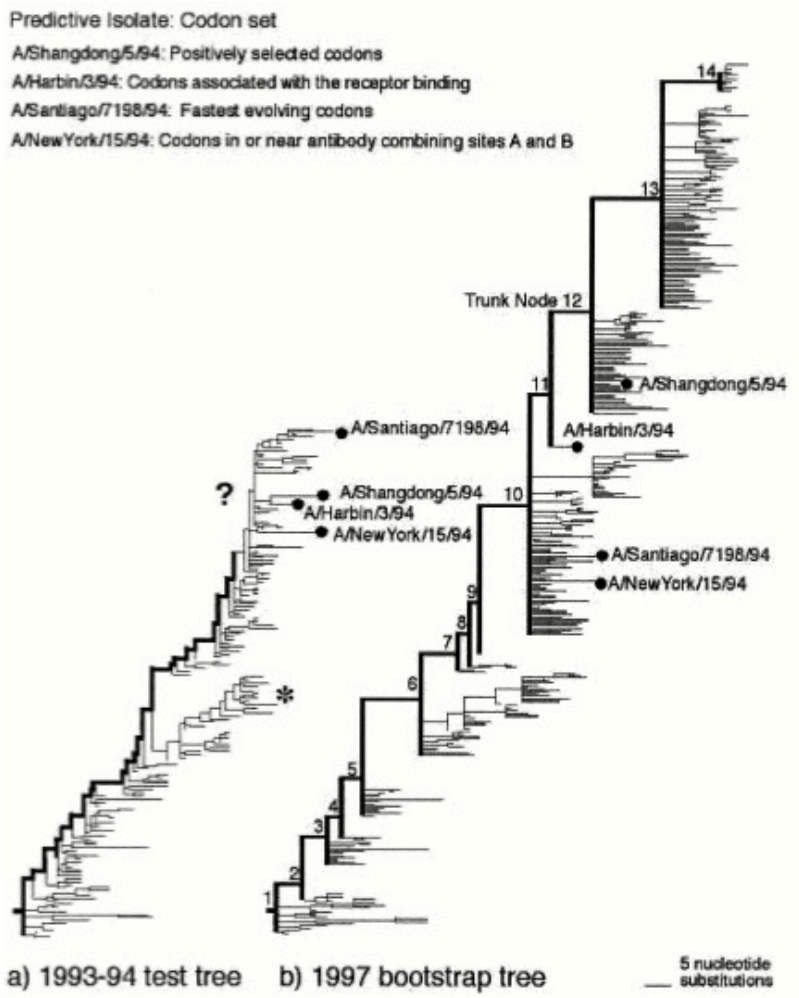
e.g. H3N2 (from 1968)



Influenza A virus RNA in 8 segments (~10000 nucleotides)

HA (Hemagglutinin) and (NA) Neuraminidase are the surface proteins responsible for the interaction with host immune system (~1000 nucleotides each)

e.g. H3N2 (from 1968)



Open question:

Which evolutive process produces
the comb-like (or unbalanced) tree?
virus - host immune system interaction?

Open question:

Which evolutive process produces
the comb-like (or unbalanced) tree?
virus - host immune system interaction?

More general open questions:

taking into account deviations from additivity
in algorithms (horizontal transfer)

exploiting additional information
(internal nodes)

modeling evolutionary processes leading to
real phylogenies

Thank you