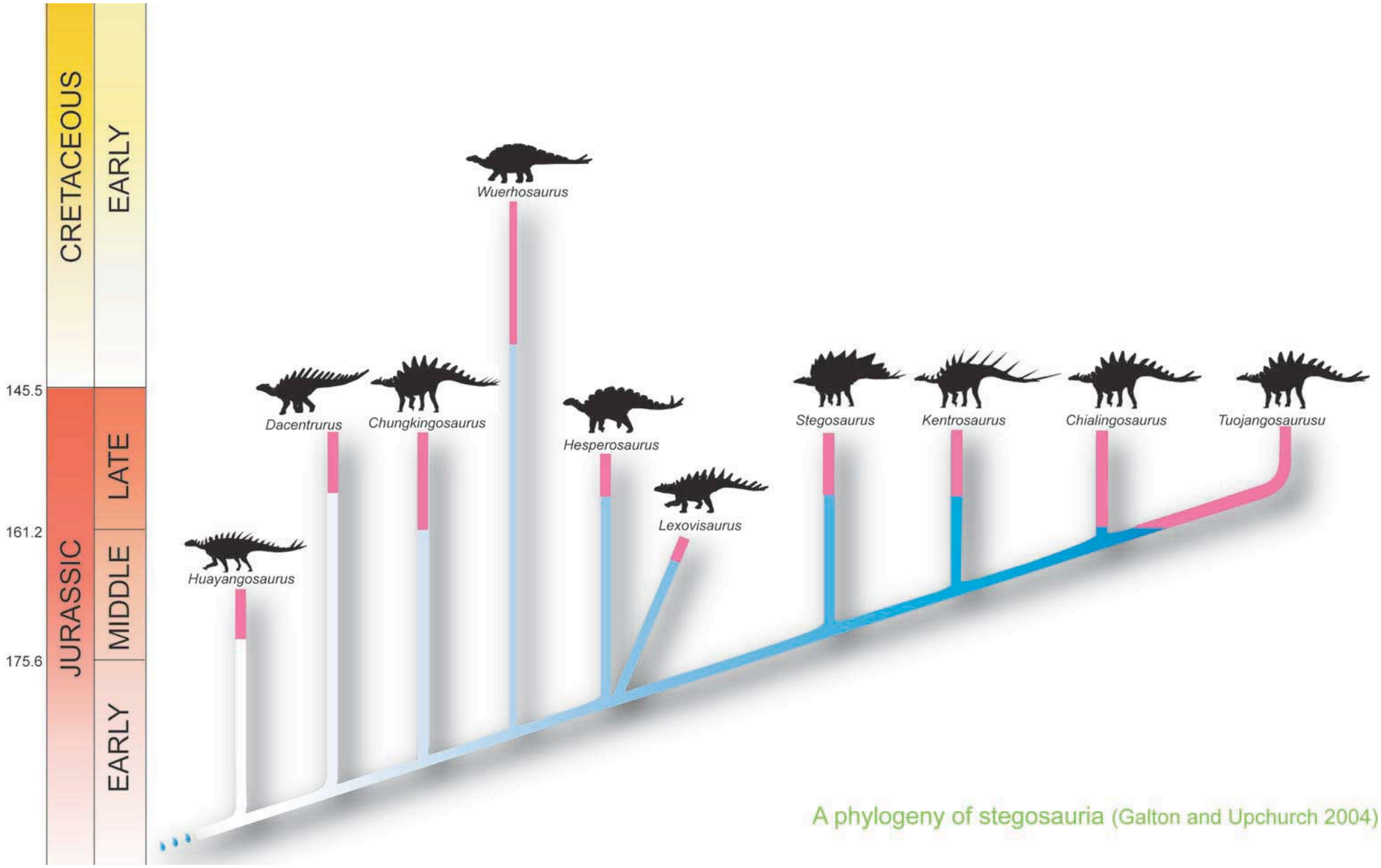
The minimum evolution problem

Daniele Catanzaro
Service Graphes et Optimisation Mathématique (G.O.M.)
Université Libre de Bruxelles

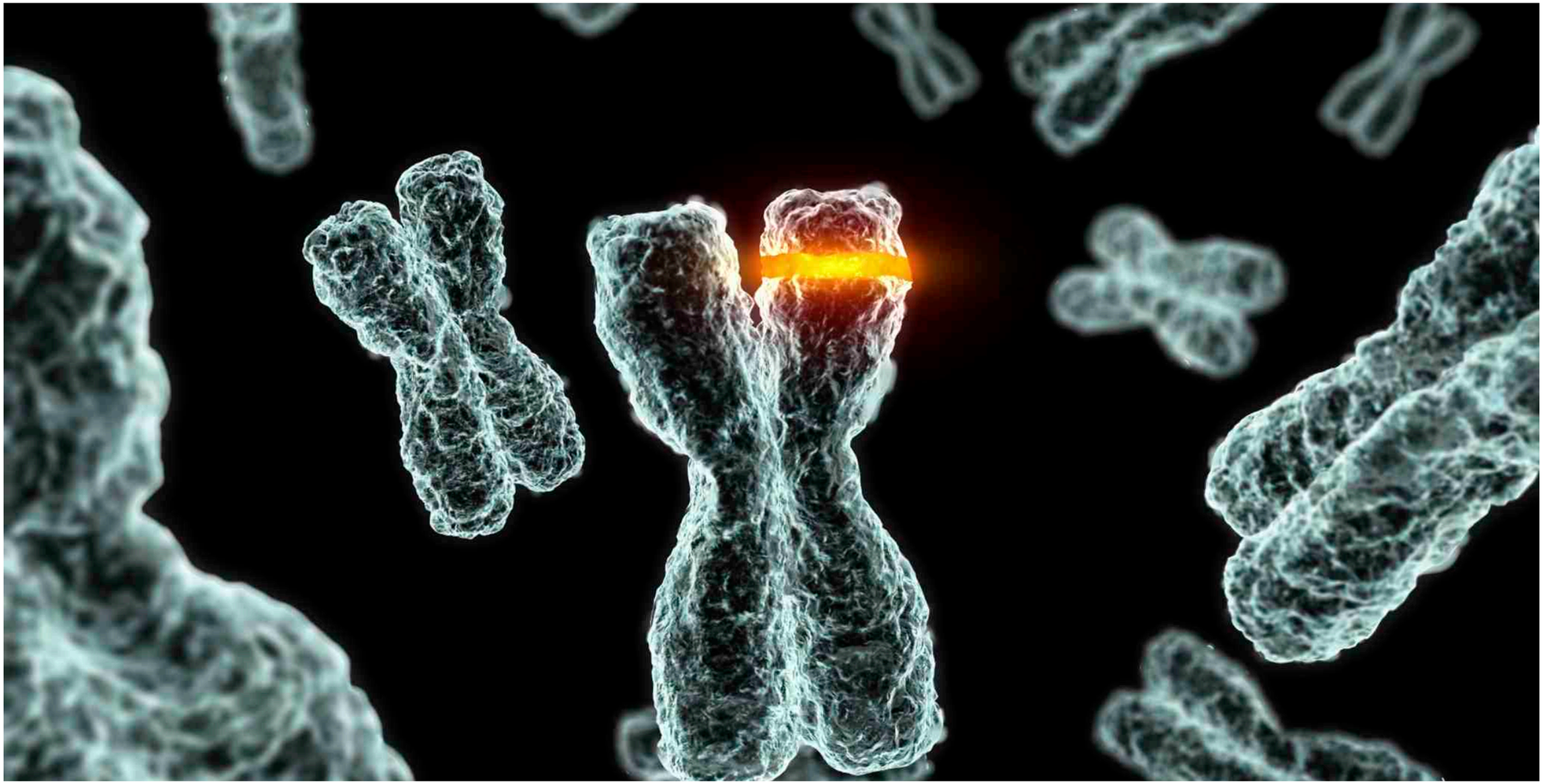


A phylogeny of stegosauria (Galton and Upchurch 2004)

From phylogenetics to molecular phylogenetics

Daniele Catanzaro

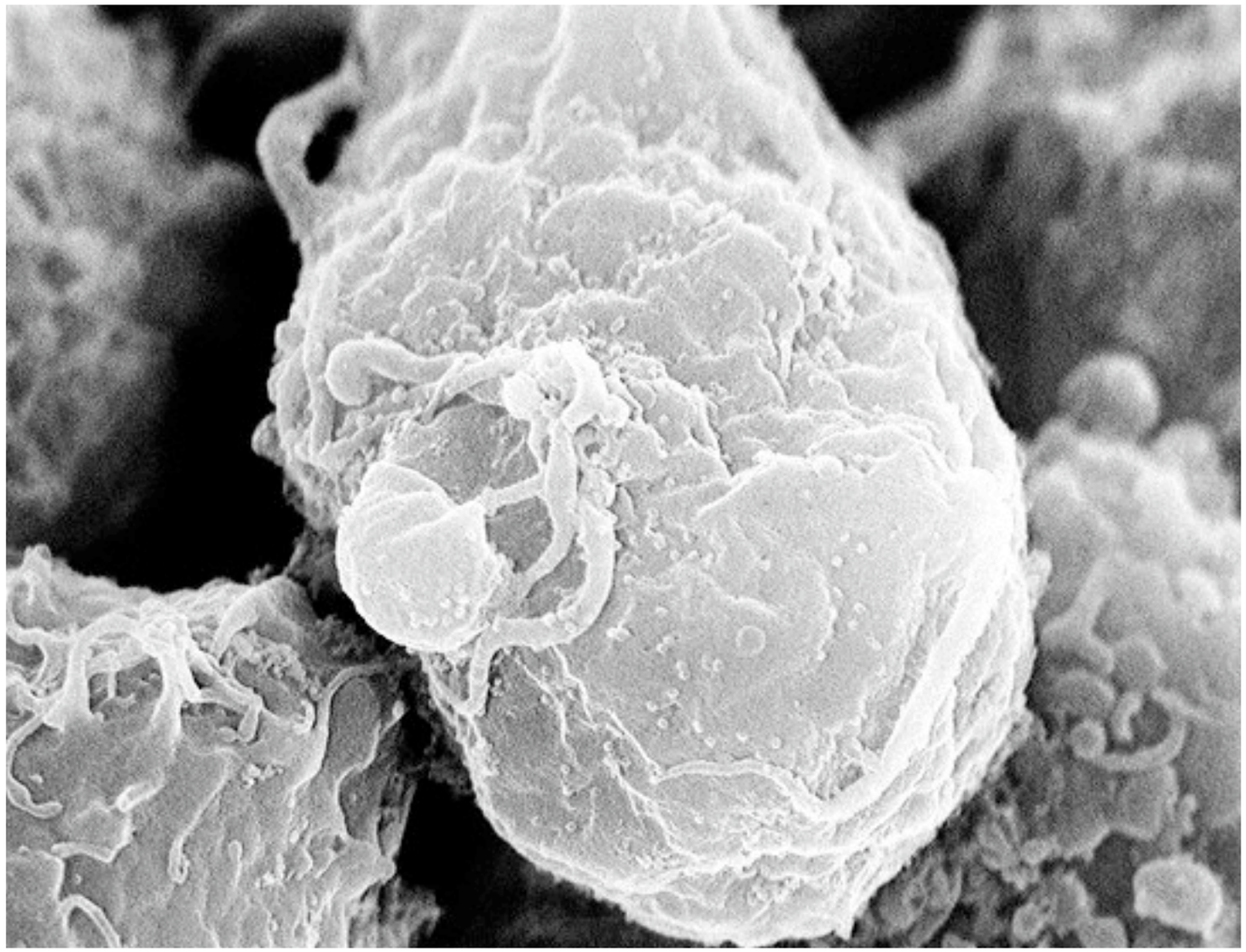
A few notes in molecular phylogenetic estimation



From phylogenetics to molecular phylogenetics

Daniele Catanzaro

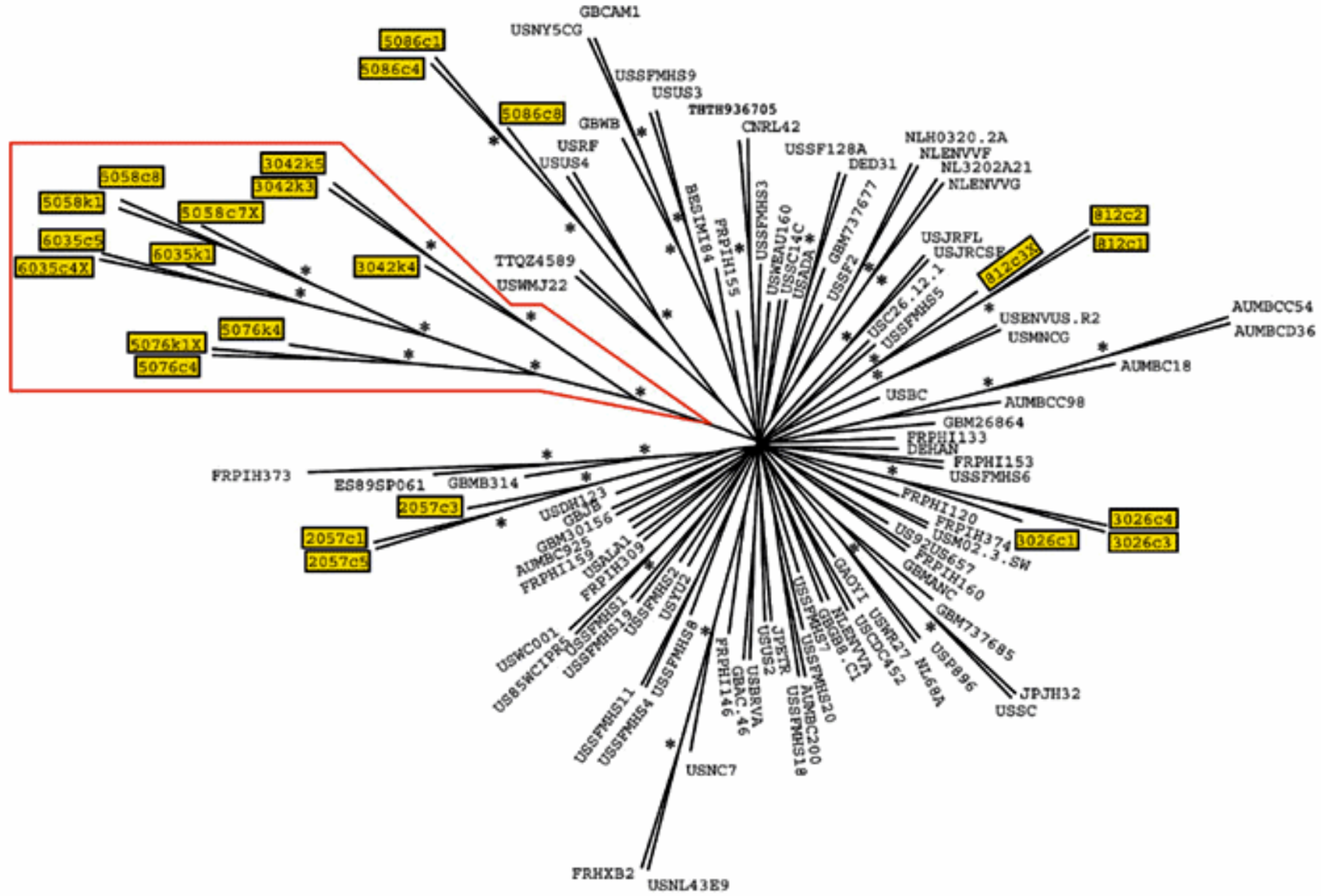
A few notes in molecular phylogenetic estimation



HIV-1 phylogeny

Daniele Catanzaro

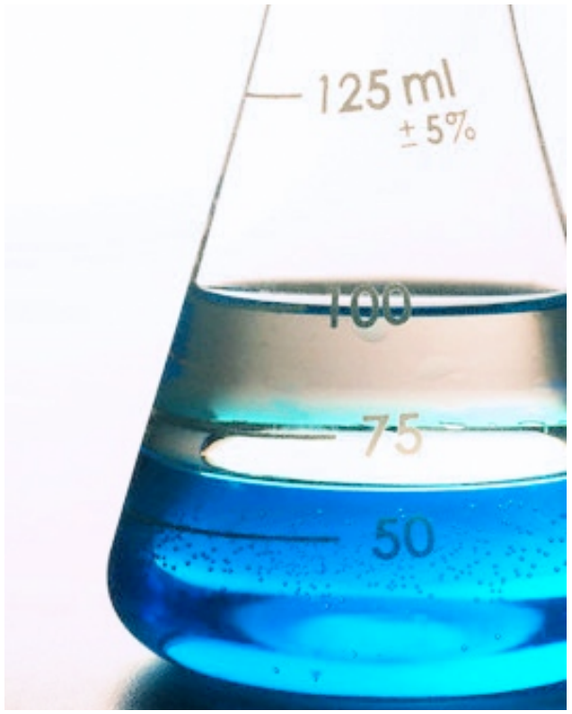
A few notes in molecular phylogenetic estimation



HIV-1 phylogeny

Daniele Catanzaro

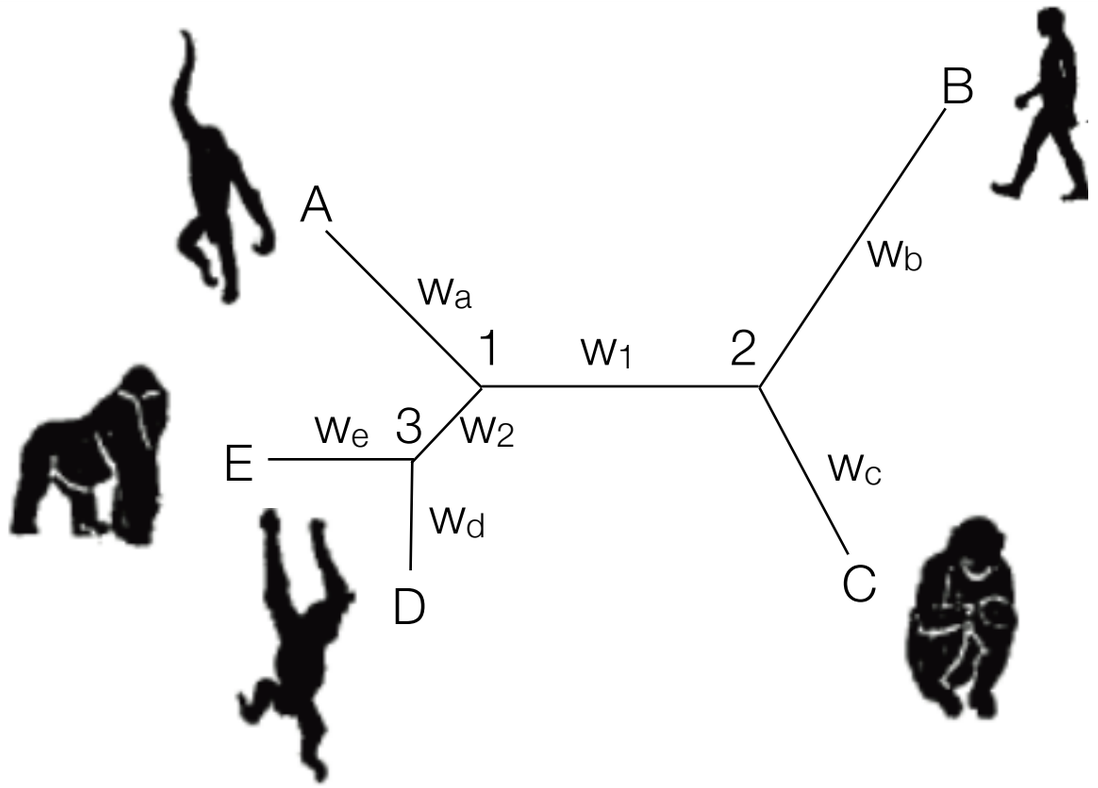
A few notes in molecular phylogenetic estimation



Applications

medical research - epidemiology
population dynamics - drug
discovery

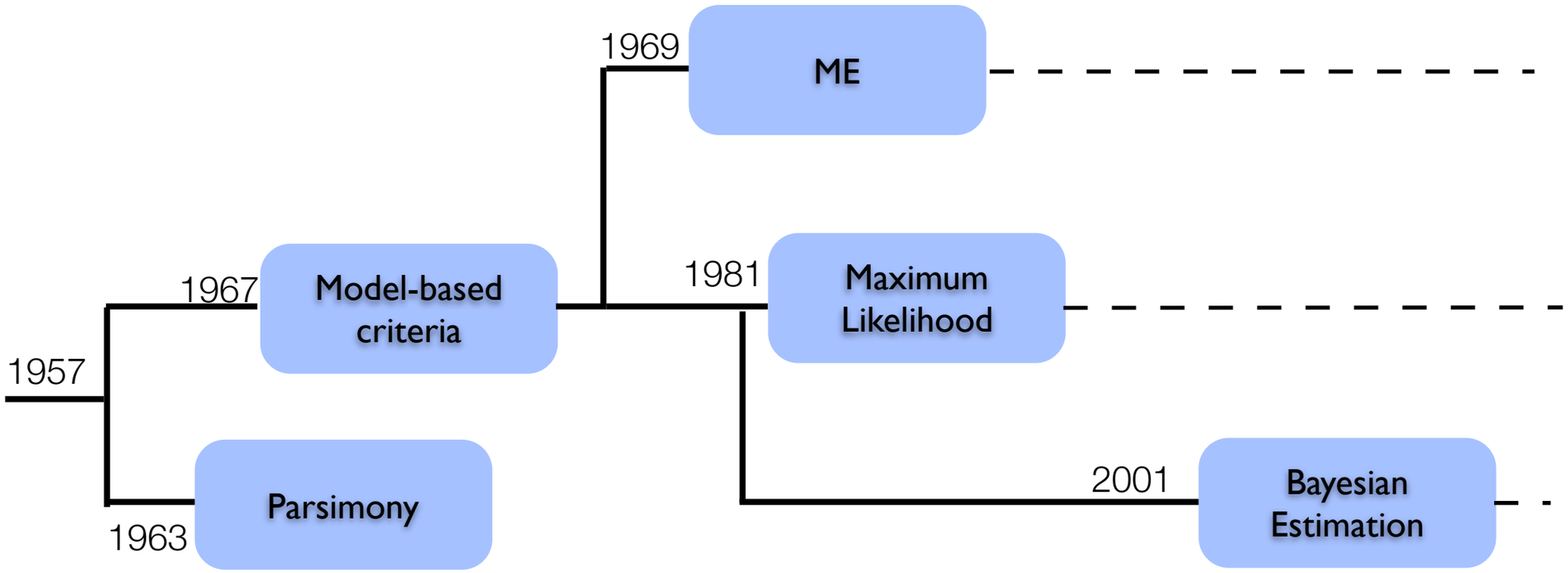
Species	Molecular Sequence
Macaca (A)	AAGCTTCATAGGAGCAACCATTCTAATAATCGCACATGGCCTTACATCATCC
Homo sapiens (B)	AAGCTTCACCGGCGCAGTCATTCTCATAATCGCCCACGGGCTTACATCCTCA
Pan (C)	AAGCTTCACCGGCGCAATTATCCTCATAATCGCCCACGGACTTACATCCTCA
Gorilla (D)	AAGCTTCACCGGCGCAGTTGTTCTTATAATTGCCCACGGACTTACATCATCA
Pongo (E)	AAGCTTCACCGGCGCAACCACCCTCATGATTGCCCATGGACTCACATCCTCC



Phylogenies

Daniele Catanzaro

A few notes in molecular phylogenetic estimation



Phylogenetic estimation models

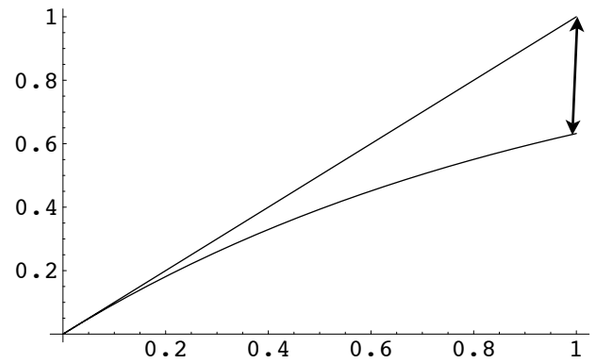
Daniele Catanzaro

The minimum evolution problem

The direct use of molecular sequences may lead to under estimation problems.

```
AAAATCTCTCTCGGTCTCACGG
AAATGTGTGTGC---CATTTTC
ATTTTCTCTCTC---CTCACGG
CCCTGTGTGTGCGGTCATTTTC
AAAAT----CTCGGTCTCACGG
```

0	7	9	8	6
7	0	5	6	7
9	5	0	3	9
8	6	3	0	1
6	7	9	1	0



Hence, some models of molecular evolution have to be taken into account in order to avoid such problems.

The sequence of a gene can be altered in a number of ways. Gene mutations have varying effects on health depending on where they occur and whether they alter the function of essential proteins. Structurally, mutations can be classified as:

Small-scale mutations, such those as affecting a small gene in one or a few nucleotides, including:

- Point mutations, Insertions /Deletions

Large-scale mutations in chromosomal structure, including:

- Gene duplications
- Deletions of large chromosomal regions.
- Chromosomal translocations
- Chromosomal inversions
- Loss of heterozygosity

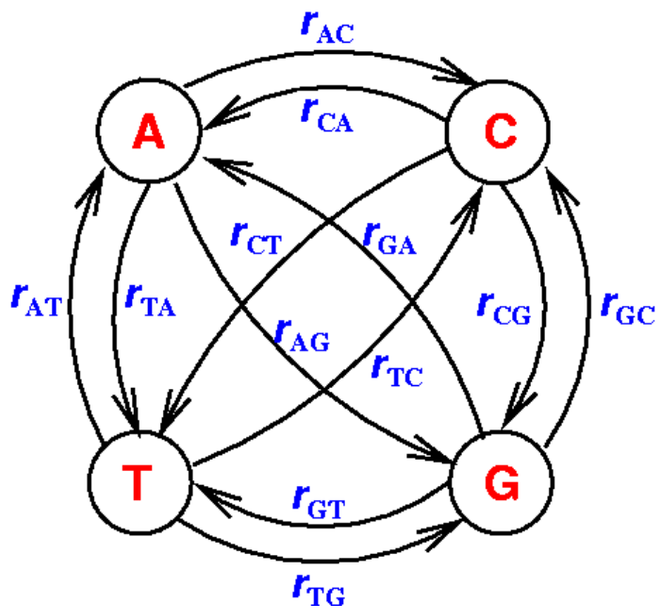
Neutrality hypothesis

Models of molecular evolution

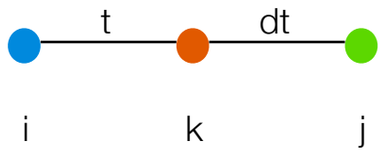
Daniele Catanzaro

The minimum evolution problem

Conservative Hypothesis



Superposition principle



$$p_{ij}(t+dt)=\Sigma p_{ik}(t)p_{kj}(dt)$$

Models of molecular evolution

Daniele Catanzaro

The minimum evolution problem

AAATCGGC

CAATCGGT

CCATCGGT

CCATCGTT

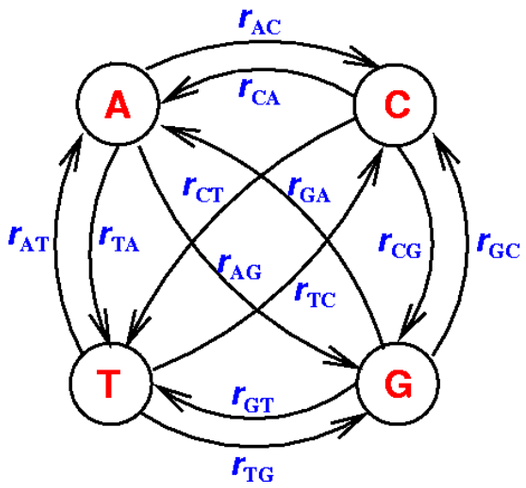
Time

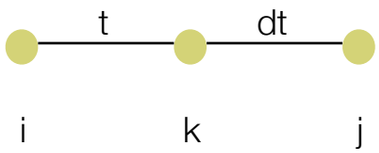
1. Neutral Selection
2. Superposition Principle
3. Conservative Hypothesis
4. Constant instantaneous rates

Models of molecular evolution

Daniele Catanzaro

The minimum evolution problem





1. Neutral Selection
2. Superposition Principle
3. Conservative Hypothesis
4. Constant instantaneous rates

$$p_{ij}(t+dt) = \sum_k p_{ik}(t)p_{kj}(dt)$$

$$p_{ij}(t+dt) - p_{ij}(t) = \sum_{k\neq j} p_{ik}(t)p_{kj}(dt) + p_{ij}(t)p_{jj}(dt) - p_{ij}(t)$$

$$\frac{p_{ij}(t+dt) - p_{ij}(t)}{dt} = \frac{\sum_{k\neq j} p_{ik}(t)p_{kj}(dt)}{dt} + p_{ij}(t)\frac{(p_{jj}(dt) - 1)}{dt}$$

$$\frac{p_{ij}(t+dt) - p_{ij}(t)}{dt} = \frac{\sum_{k\neq j} p_{ik}(t)p_{kj}(dt)}{dt} + p_{ij}(t)\frac{(1 - \sum_{k\neq j} p_{kj}(dt) - 1)}{dt}$$

$$\frac{p_{ij}(t+dt) - p_{ij}(t)}{dt} = \frac{\sum_{k\neq j} p_{ik}(t)p_{kj}(dt)}{dt} + p_{ij}(t)\frac{p_{jj}(dt)}{dt}$$

$$\dot{p}_{ij}(t) = \sum_k p_{ik}(t)r_{kj} = \sum_k r_{kj}p_{kj}(t) \quad \dot{\mathbf{P}}(t) = \mathbf{R}\mathbf{P}(t) \quad \mathbf{P}(t) = \mathbf{e}^{\mathbf{R}t} = \mathbf{\Omega}\mathbf{e}^{\mathbf{\Lambda}t}\mathbf{\Omega}^{-1}$$

Models of molecular evolution

Daniele Catanzaro

The minimum evolution problem

Models of molecular evolution are quite important. For example they can be used for computing distances.

```
AAAATCTCTCTCGGTCTCACGG
AAATGTGTGTGCACACATTTTC
```

$$\mathbf{F}^{\#} = \begin{pmatrix} f_{AA} & f_{AC} & f_{AG} & f_{AT} \\ f_{CA} & f_{CC} & f_{CG} & f_{CT} \\ f_{GA} & f_{GC} & f_{GG} & f_{GT} \\ f_{TA} & f_{TC} & f_{TG} & f_{TT} \end{pmatrix}$$

$$\mathbf{\Pi P}(t) = \mathbf{P}(t)^T \mathbf{\Pi}$$

$$\mathbf{P}(\hat{t}) = \mathbf{P} = \mathbf{\Pi}^{-1}(\mathbf{P}(\hat{t})^T \mathbf{\Pi}) = \mathbf{\Pi}^{-1}\mathbf{F}^{\#}(\hat{t}) = \mathbf{\Pi}^{-1}\mathbf{F}^{\#}$$

$$\mathbf{P}(t) = \mathbf{e}^{\mathbf{R}t} = \mathbf{\Omega}\mathbf{e}^{\mathbf{\Lambda}t}\mathbf{\Omega}^{-1}$$

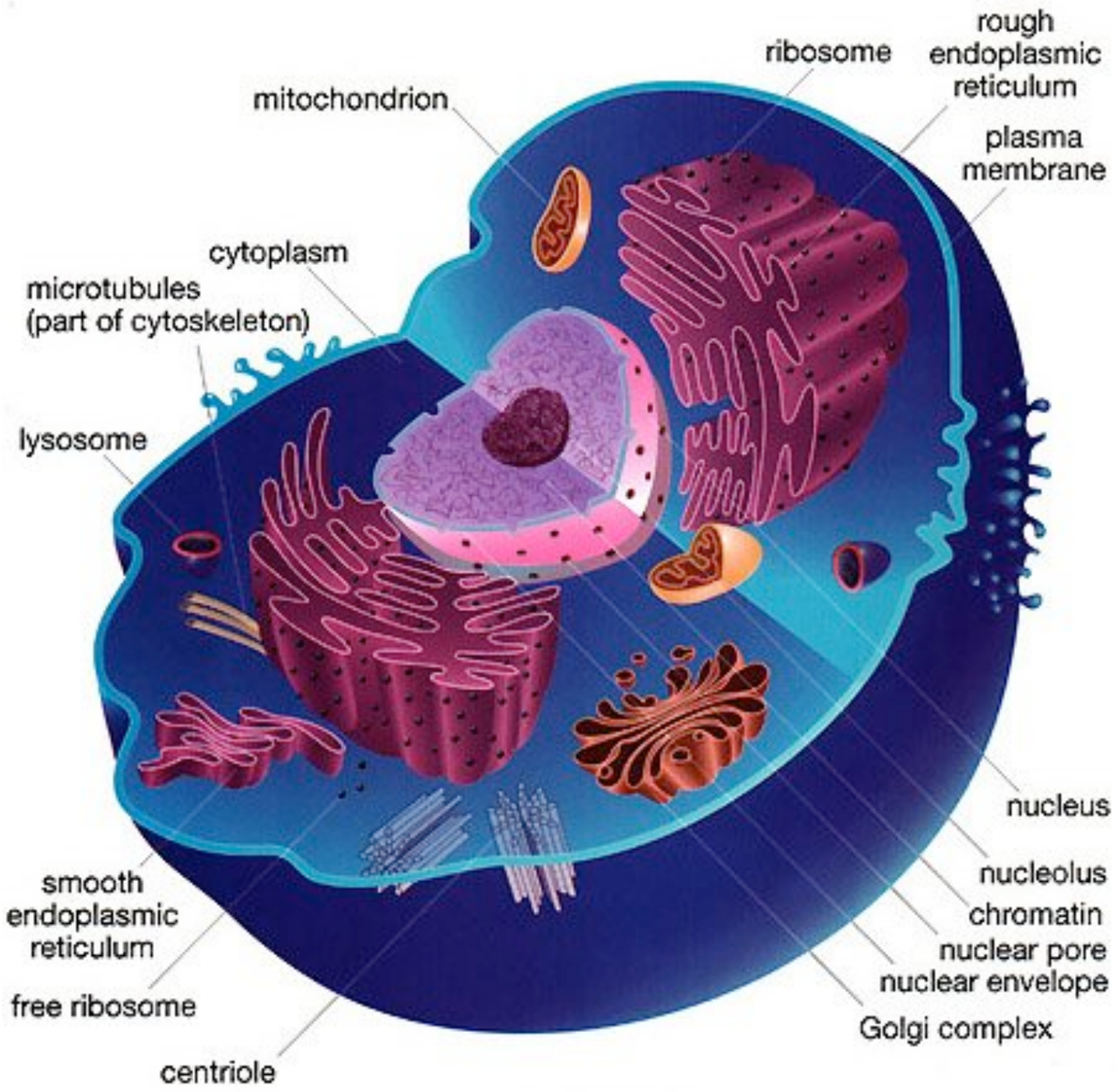
$$\hat{t} = -trace[\mathbf{\Pi log(P)}] \qquad \log(\mathbf{A}) = \mathbf{\Psi}\log(\mathbf{\Lambda})\mathbf{\Psi}$$

Lanave et al. (1984) presented a general model of DNA sequence evolution.
Tavaré (1986), Barry & Hartigan (1987), Rodriguez et al. (1990) gave a different but numerically and algebraically equivalent formulation.
Gillespie (1986), Zharkikh (1994), Waddell (1995) noted the time-reversibility (TR) of the Lanave's model; Swofford and Lewis (1997) provided a proof. Waddell & Steel (1997) summarized and extended results on the GTR model and, starting from Rodriguez's formulation, provided algorithms currently implemented in PAUP*.

Computing distances

Daniele Catanzaro

The minimum evolution problem



The cell

Daniele Catanzaro

The minimum evolution problem

		2nd base			
		U	**C**	**A**	**G**
1st base	**U**	UUU (Phe/F)Phenylalanine UUC (Phe/F)Phenylalanine UUA (Leu/L)Leucine UUG (Leu/L)Leucine	UCU (Ser/S)Serine UCC (Ser/S)Serine UCA (Ser/S)Serine UCG (Ser/S)Serine	UAU (Tyr/Y)Tyrosine UAC (Tyr/Y)Tyrosine UAA Ochre (*Stop*) UAG Amber (*Stop*)	UGU (Cys/C)Cysteine UGC (Cys/C)Cysteine UGA Opal (*Stop*) UGG (Trp/W)Tryptophan
	C	CUU (Leu/L)Leucine CUC (Leu/L)Leucine CUA (Leu/L)Leucine CUG (Leu/L)Leucine	CCU (Pro/P)Proline CCC (Pro/P)Proline CCA (Pro/P)Proline CCG (Pro/P)Proline	CAU (His/H)Histidine CAC (His/H)Histidine CAA (Gln/Q)Glutamine CAG (Gln/Q)Glutamine	CGU (Arg/R)Arginine CGC (Arg/R)Arginine CGA (Arg/R)Arginine CGG (Arg/R)Arginine
	A	AUU (Ile/I)Isoleucine AUC (Ile/I)Isoleucine AUA (Ile/I)Isoleucine AUG (Met/M)Methionine, *Start*[1]	ACU (Thr/T)Threonine ACC (Thr/T)Threonine ACA (Thr/T)Threonine ACG (Thr/T)Threonine	AAU (Asn/N)Asparagine AAC (Asn/N)Asparagine AAA (Lys/K)Lysine AAG (Lys/K)Lysine	AGU (Ser/S)Serine AGC (Ser/S)Serine AGA (Arg/R)Arginine AGG (Arg/R)Arginine
	G	GUU (Val/V)Valine GUC (Val/V)Valine GUA (Val/V)Valine GUG (Val/V)Valine	GCU (Ala/A)Alanine GCC (Ala/A)Alanine GCA (Ala/A)Alanine GCG (Ala/A)Alanine	GAU (Asp/D)Aspartic acid GAC (Asp/D)Aspartic acid GAA (Glu/E)Glutamic acid GAG (Glu/E)Glutamic acid	GGU (Gly/G)Glycine GGC (Gly/G)Glycine GGA (Gly/G)Glycine GGG (Gly/G)Glycine

Models of molecular evolution

Daniele Catanzaro

The minimum evolution problem

Amino acid	Hydrophobic effect [kcal-mol]	Molecular weight [Dalton]	Surface area [Å^2]	Side-chain volume [Å^3]
Alanine	1.0	71	115	88.6
Arginine	1.1	156	225	173.4
Asparagine	-0.1	114	160	114.1
Aspartic acid	-0.1	115	150	111.1
Cysteine	0.0	103	135	108.5
Glutamic acid	0.5	129	190	138.4
Glutamine	0.5	128	180	143.8
Glycine	0.0	57	75	60.1
Histidine	1.3	137	195	153.2
Isoleucine	2.7	113	175	166.7
Leucine	2.9	113	170	166.7
Lysine	1.9	128	200	168.6
Methionine	2.3	131	185	162.9
Phenylalanine	2.3	147	210	189.9
Proline	1.9	97	145	112.7
Serine	0.2	87	115	89
Threonine	1.1	101	140	116.1
Tryptophan	2.9	186	255	227.8
Tyrosine	1.6	163	230	193.6
Valine	2.2	99	155	140

$$h_{ij} = \begin{cases} 1 & \text{for synonymous substitutions} \\ 1/2 & \text{for replacement substitutions involving position 1 or 3} \\ 1/3 & \text{for replacement substitutions involving position 2} \\ 1/4 & \text{for replacement substitutions involving positions 1 and 3} \\ 1/6 & \text{for replacement substitutions involving positions (1 and 2) or (2 and 3)} \\ 1/8 & \text{for replacement substitutions involving positions 1, 2, 3} \\ 1/10 & \text{for replacement substitutions to stop codons} \end{cases}$$

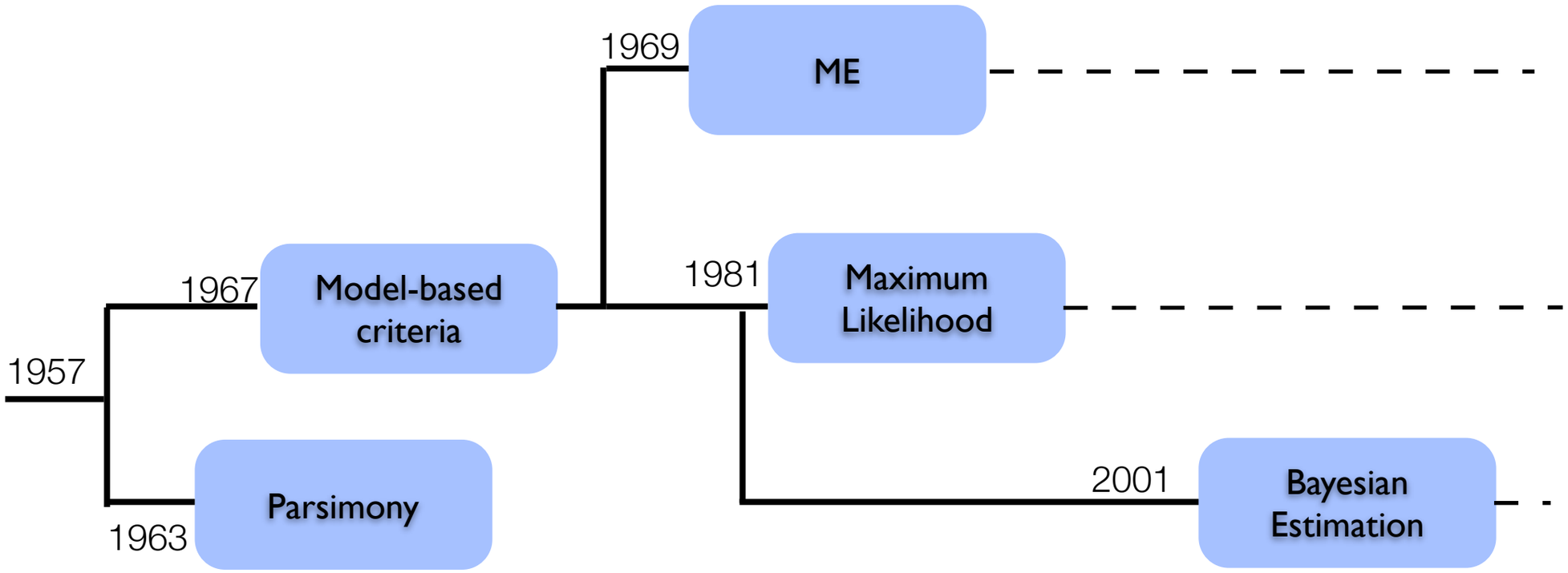
$$a_{ij} = r_{ij} h_{ij} e^{-\sum_k \omega_k |c_i^k - c_j^k|}$$

$$\dot{\mathbf{P}}(t) = \mathbf{P}(t)\mathbf{A} = \mathbf{A}\mathbf{P}(t)$$

Models of molecular evolution

Daniele Catanzaro

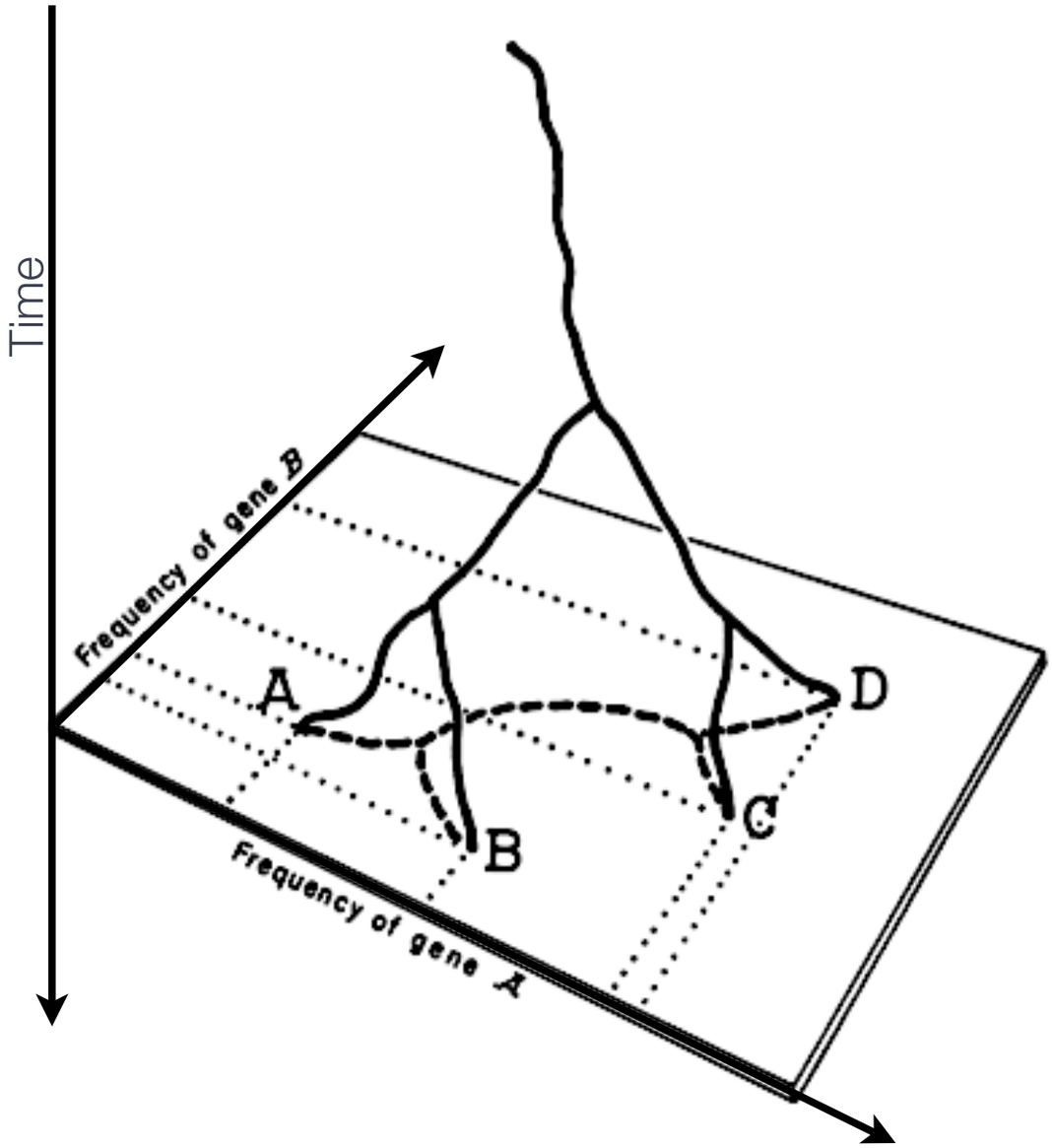
The minimum evolution problem



Phylogenetic estimation models

Daniele Catanzaro

The minimum evolution problem



Cavalli-Sforza and Edwards: A first model of evolution

Daniele Catanzaro

The minimum evolution problem

Measure of the dissimilarity between species A and B

Error relative to the approximation

$$w_{A1'} + w_{1'B} = d_{AB} + e_{AB}$$

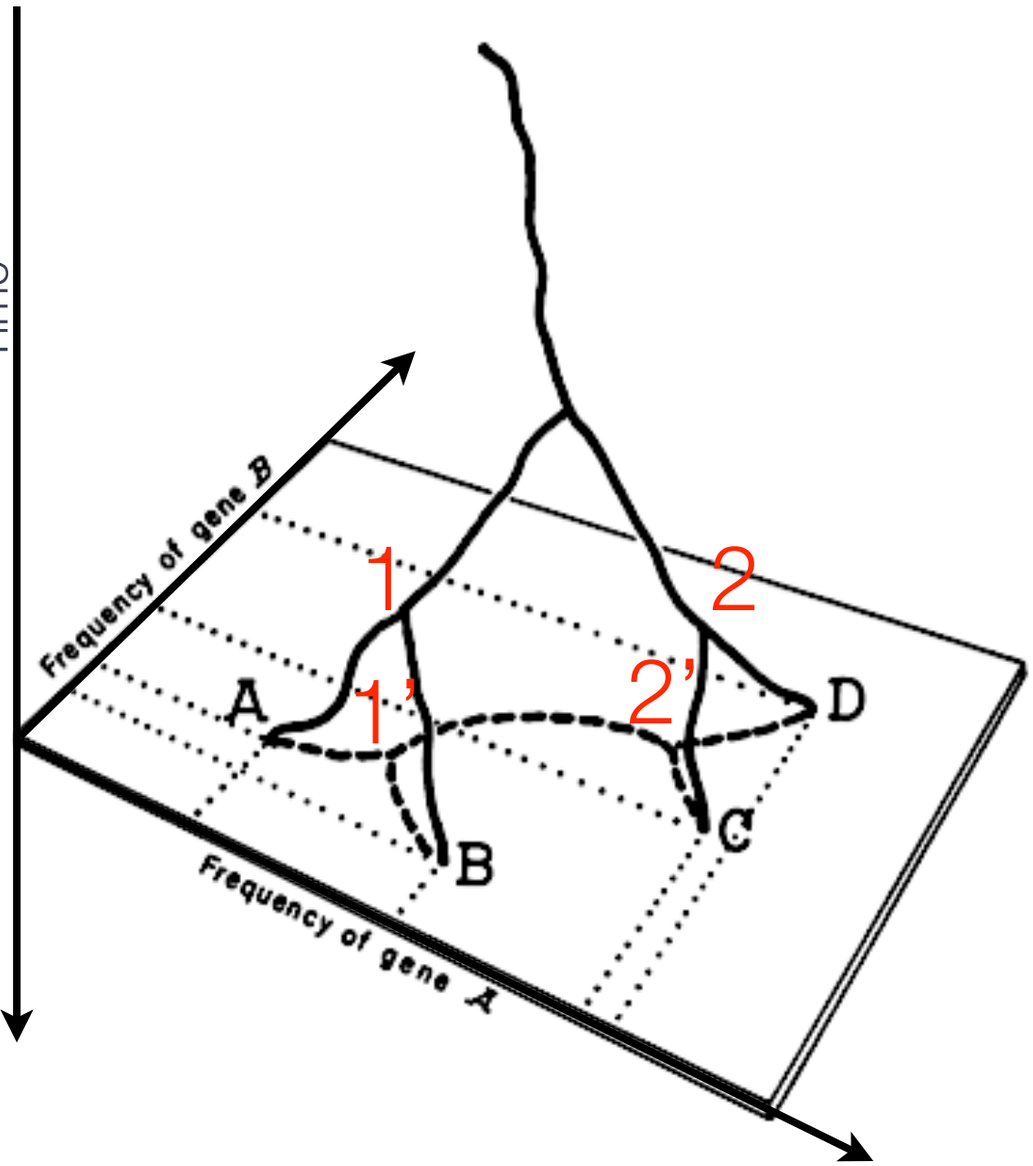
$$w_{A1'} + w_{1'2'} + w_{2'C} = d_{AC} + e_{AC}$$

$$w_{A1'} + w_{1'2'} + w_{2'D} = d_{AD} + e_{AD}$$

$$w_{B1'} + w_{1'2'} + w_{2'C} = d_{BC} + e_{BC}$$

$$w_{B1'} + w_{1'2'} + w_{2'D} = d_{BD} + e_{BD}$$

$$w_{C2'} + w_{2'D} = d_{CD} + e_{CD}$$



Cavalli-Sforza and Edwards: A first model of evolution

Daniele Catanzaro

The minimum evolution problem

vectorially: **Xw=d+e**

where

X=

	A1’	B1’	1’2’	2’C	2’D
AB	1	1	0	0	0
AC	1	0	1	1	0
AD	1	0	1	0	1
BC	0	1	1	1	0
BD	0	1	1	0	1
CD	0	0	0	1	1

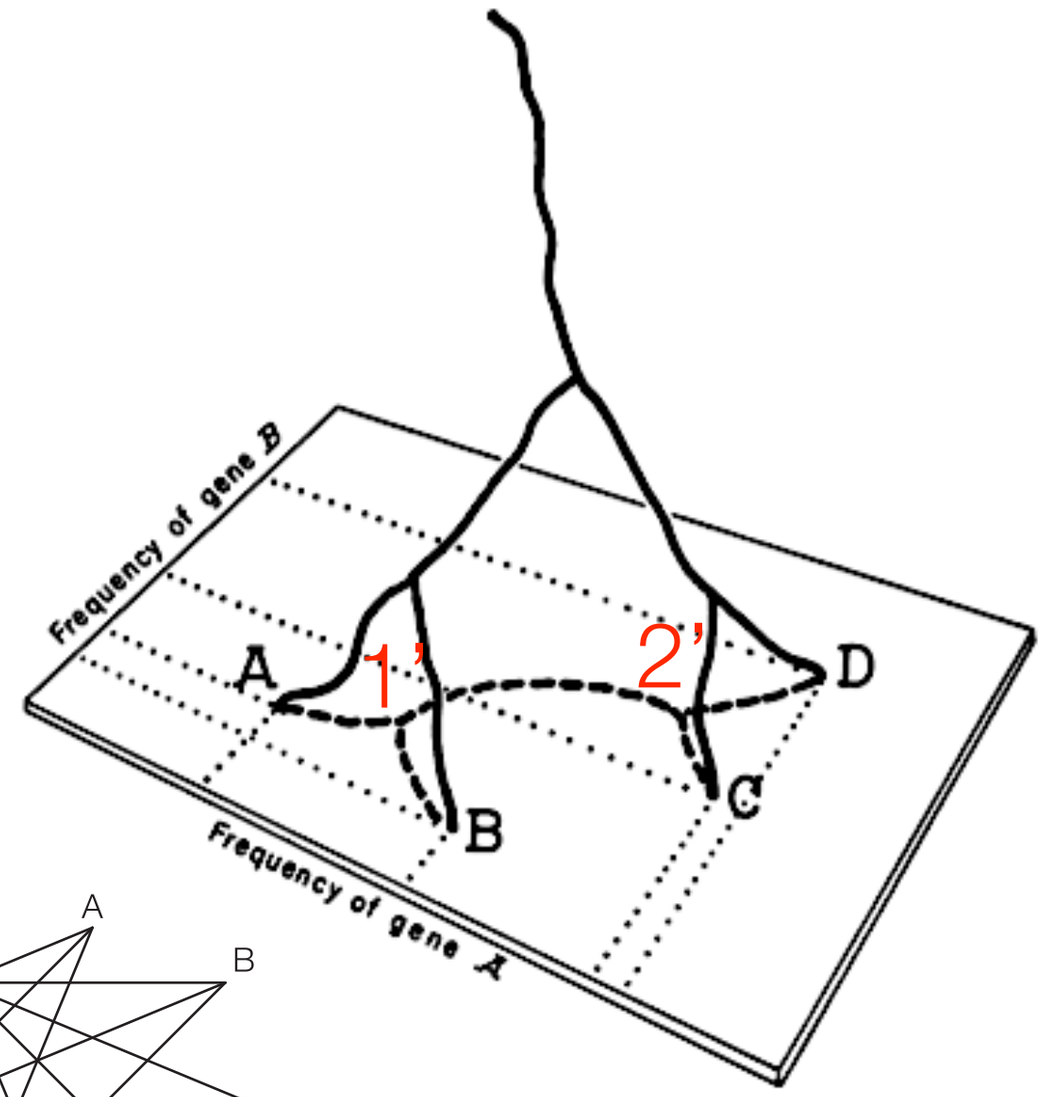
d=

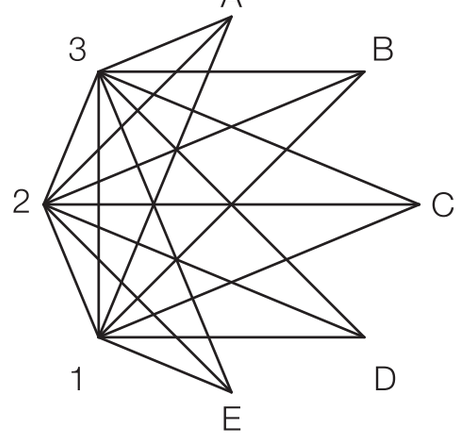
d_{AB}
d_{AC}
d_{AD}
d_{BC}
d_{BD}
d_{CD}

Hence, Cavalli-Sforza and Edwards model becomes:

Find the best **X** s.t.

$\|\mathbf{e}\|_2$ is minimized





Cavalli-Sforza and Edwards: A first model of evolution

Daniele Catanzaro

The minimum evolution problem

In other words, Cavalli-Sforza and Edwards proposed the use of the Least-Squares (LS) method, i.e., to find a phylogeny having the lowest distortion from an additive phylogeny.

$$w=(X^tX)^{-1}X^td=X^\dagger d$$

This estimation model is characterized by several drawbacks:

- Species generally do not evolve independently from each others.
- The rate of evolution may not be the same for each species.
- The additive model may provide phylogenies having negative edge weights which is a nonsense.

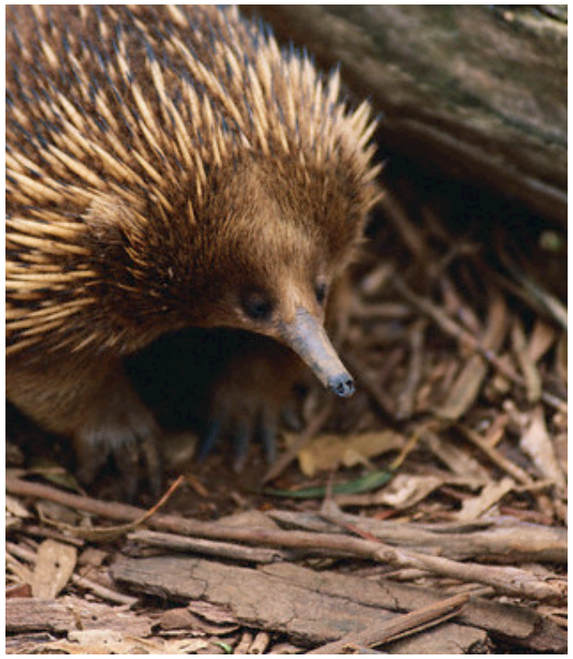
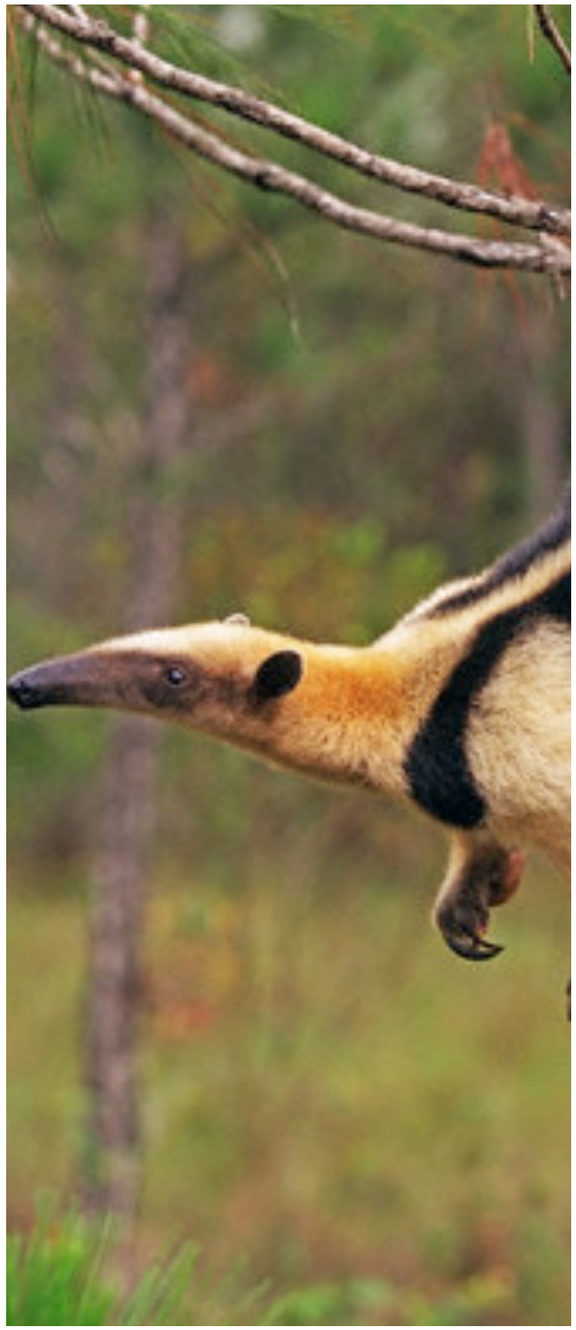
In other words, Cavalli-Sforza and Edwards proposed the use of the Least-Squares (LS) method, i.e., to find a phylogeny having the lowest distortion from an additive phylogeny.

$$w=(X^tX)^{-1}X^td=X^\dagger d$$

Some authors proposed to

- Consider the evolutionary dependencies between species (Weighted Least-Squares (WLS) and Generalized Least-Squares (GLS)).
- Consider models which allow different evolutionary rates (Minimum Evolution models, Maximum Likelihood (ML) models, Bayesian models (BM)).
- Impose the positivity constraint in order to remove negative edge weights (Projective Algorithms (PA), Minimum Distortion Algorithms (MDA), Balanced Least-Squares (BLS)).

Possible solutions

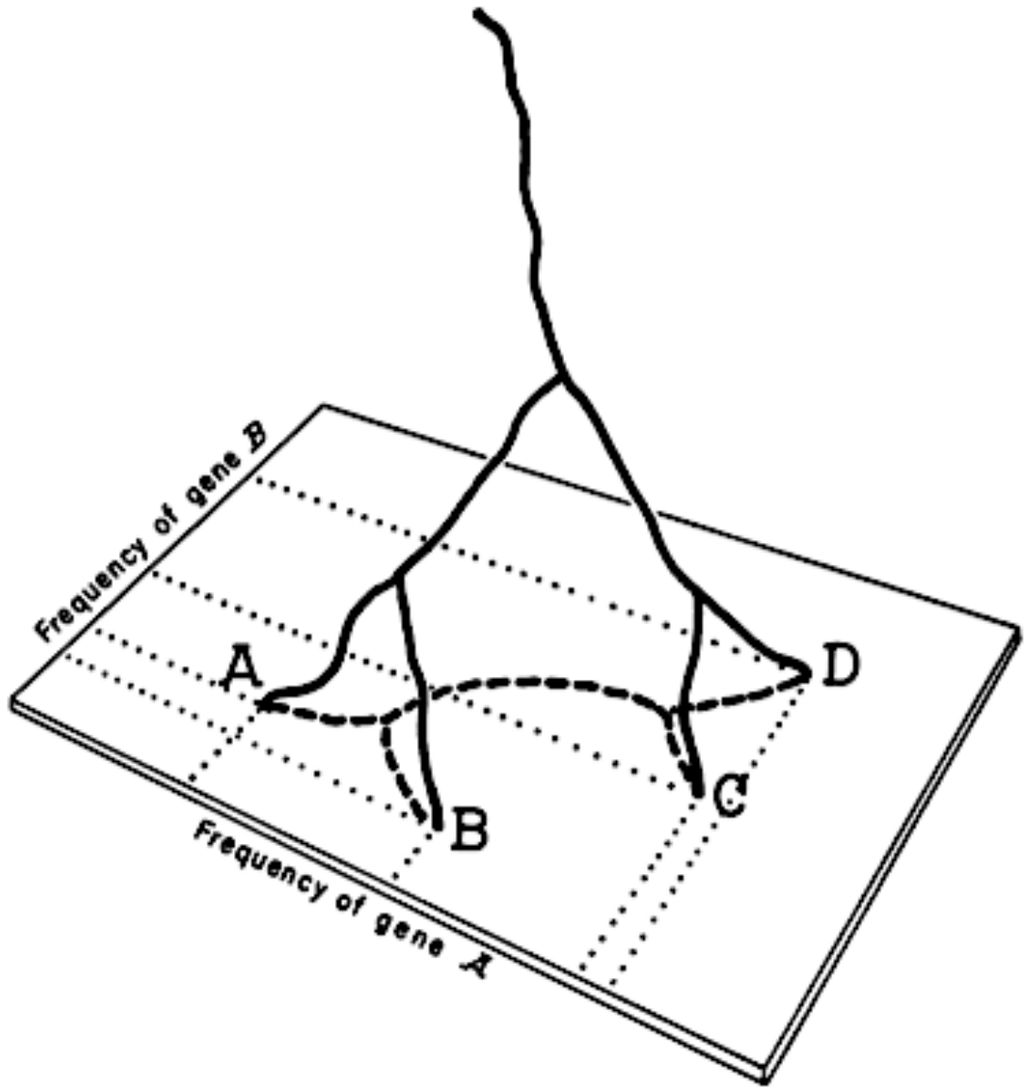





The Minimum Evolution (ME) model

Daniele Catanzaro

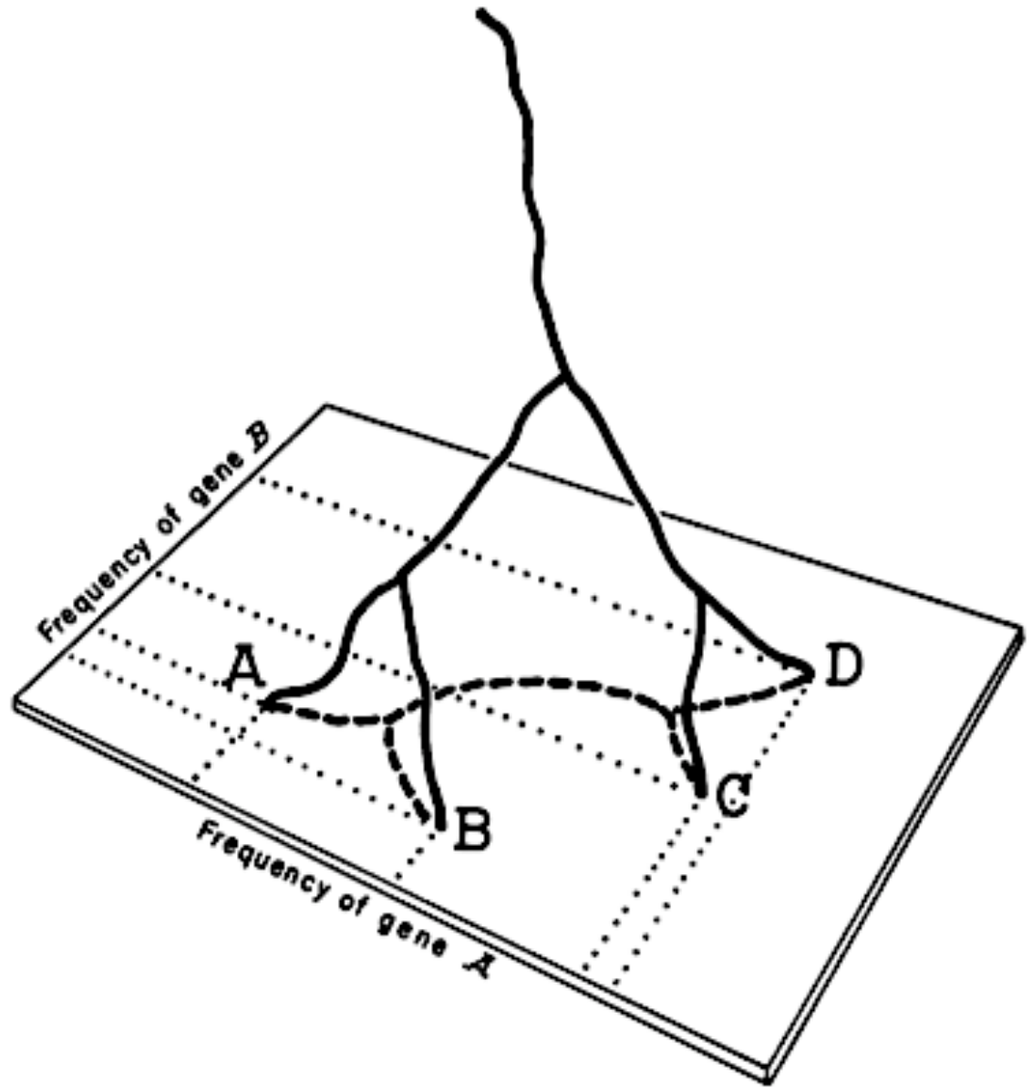
The minimum evolution problem

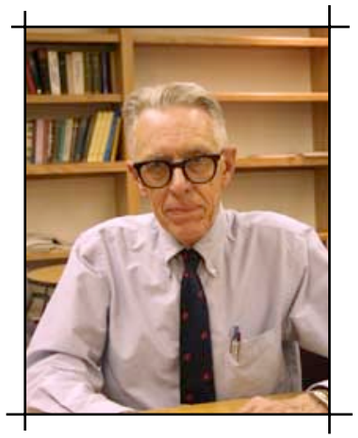


The minimum evolution problem

Daniele Catanzaro

The minimum evolution problem





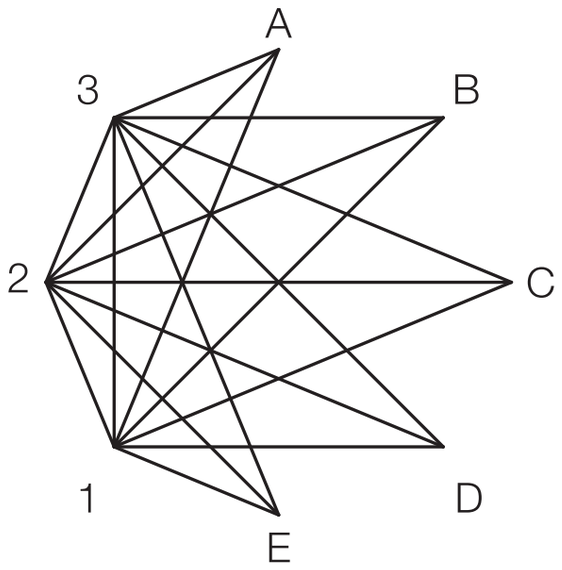
The minimum evolution problem

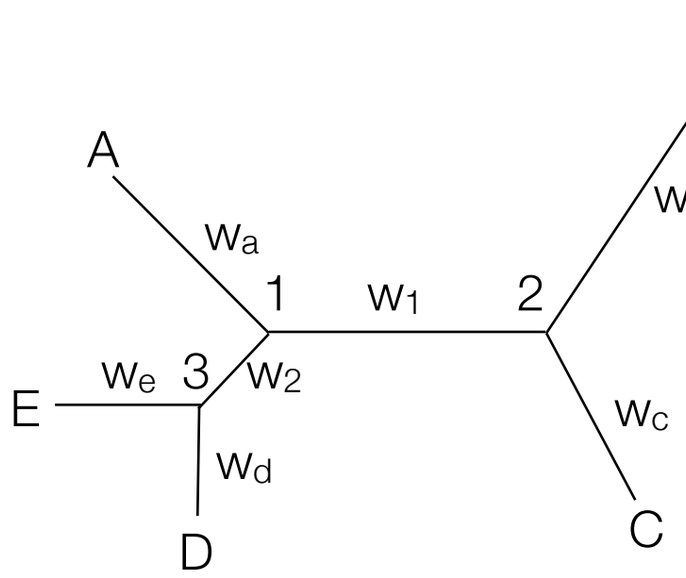
Daniele Catanzaro

The minimum evolution problem

Minimum Evolution Problem Under Linear Programming (MELP)

Given a phylogenetic graph and a distance matrix among species, find a phylogeny whose length is minimum.





$$\min_{(X,w)} \sum_{e=1}^{(2n-3)} w_e$$

$$\text{s.t. } Xw \geq D$$

$$X \in \chi$$

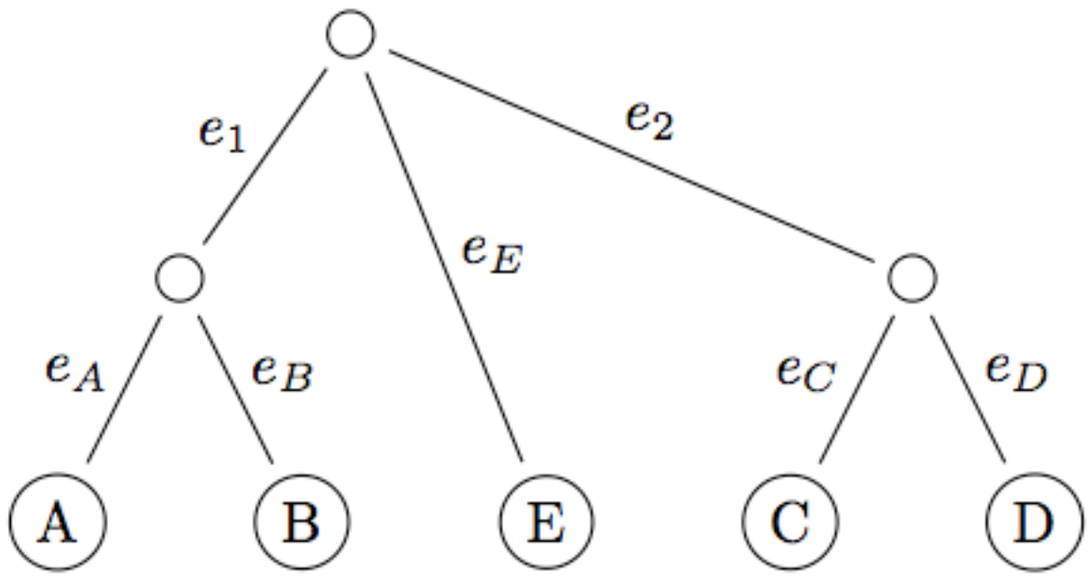
$$w \in \mathbb{R}_{0^+}^{(2n-3)}$$

MELP

Daniele Catanzaro

The minimum evolution problem

Assume the set of species Γ is lexicographically ordered, and assume without loss of generality that the rows of the EPT matrix are always ordered lexicographically on the basis of the order in Γ. Assume also that the first n columns of X correspond to the external edges of a phylogeny T and that they are sorted according to the order of the taxa at one of their extremes.



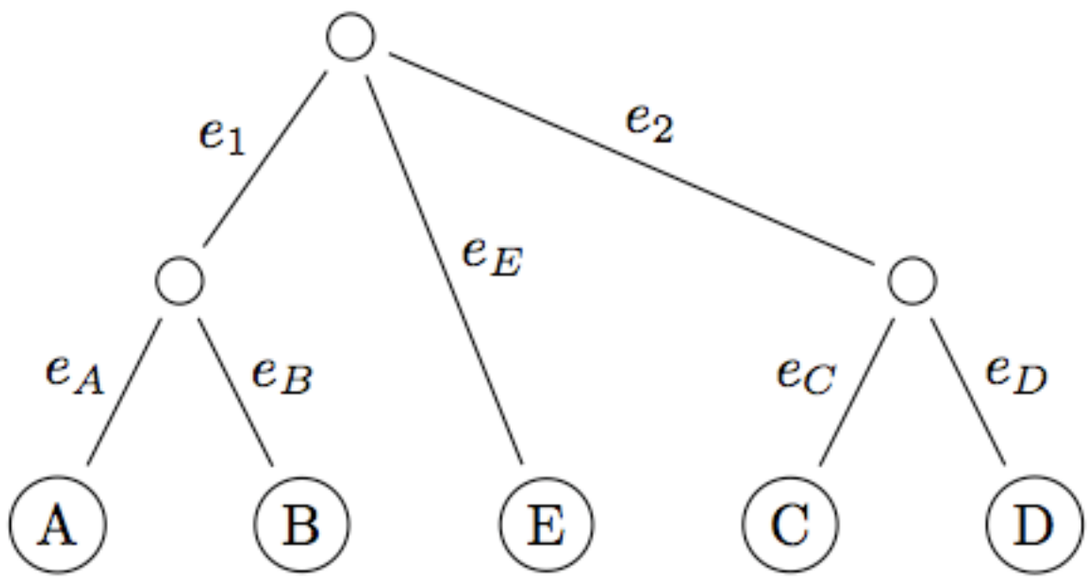
	e_A	e_B	e_C	e_D	e_E	e_1	e_2
AB	1	1	0	0	0	0	0
AC	1	0	1	0	0	1	1
AD	1	0	0	1	0	1	1
AE	1	0	0	0	1	1	0
BC	0	1	1	0	0	1	1
BD	0	1	0	1	0	1	1
BE	0	1	0	0	1	1	0
CD	0	0	1	1	0	0	0
CE	0	0	1	0	1	0	1
DE	0	0	0	1	1	0	1

The EPT model

Daniele Catanzaro

The minimum evolution problem

Finally, we assume that the remaining (n-3) columns of X, corresponding to the internal edges of T are sorted according to a relation defined in the following way. Given a generic external edge e, define dist(e) as the topological distance of e from the leaf associated with taxon A. In addition, define path(e) the first path, from a lexicographical point of view, to which e belongs. Then, we impose that the column associated with the internal edge e1 precedes the column associated with the internal edge e2 in X if one of the following two conditions holds: path(e1) lexicographically precedes path(e2) or path(e1) = path(e2) and dist(e1) < dist(e2). This order relation is complete as the lexicographic order is complete in the path set, and in a tree we cannot have path(e1) = path(e2) and dist(e1) = dist(e2).



	e_A	e_B	e_C	e_D	e_E	e_1	e_2
AB	1	1	0	0	0	0	0
AC	1	0	1	0	0	1	1
AD	1	0	0	1	0	1	1
AE	1	0	0	0	1	1	0
BC	0	1	1	0	0	1	1
BD	0	1	0	1	0	1	1
BE	0	1	0	0	1	1	0
CD	0	0	1	1	0	0	0
CE	0	0	1	0	1	0	1
DE	0	0	0	1	1	0	1

Given this order, any EPT matrix of a phylogeny can be decomposed in blocks:

	e_A	e_B	e_C	e_D	e_E	e_1	e_2
AB	1	1	0	0	0	0	0
AC	1	0	1	0	0	1	1
AD	1	0	0	1	0	1	1
AE	1	0	0	0	1	1	0
BC	0	1	1	0	0	1	1
BD	0	1	0	1	0	1	1
BE	0	1	0	0	1	1	0
CD	0	0	1	1	0	0	0
CE	0	0	1	0	1	0	1
DE	0	0	0	1	1	0	1

It is easily seen that only the red block (hereafter indicated as F) is necessary to describe a phylogeny. In fact the bleu and the green blocks can be obtained as xor of the relative yellow and red blocks.

In order to represent a phylogeny the entries of matrix F must obey to the following theorem:

Theorem 2.1. $\mathbf{F}$ *is feasible if and only if all the following conditions hold on its entries:*

1. *It does not include any of the following* 2×2 *submatrices*

$$\mathbf{M}_1 = \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix} \tag{2.7}$$

or

$$\mathbf{M}_2 = \begin{pmatrix} 0 & 1 \\ 1 & 1 \end{pmatrix} \tag{2.8}$$

where the columns and the rows whose intersection defines $\mathbf{M}_1$ *and* $\mathbf{M}_2$ *are not necessarily adjacent in* $\mathbf{F}$.

2. *Any of its column has at least two entries equal to 1.*

3. *If columns* r *and* s*, with* $r < s$*, have an entry equal to 1 in the same row then the number of entries equal to 1 in column* r *is larger than the number of entries equal to 1 in* s.

The xor conditions can be expressed as follows:

$$
\begin{aligned}
x_{ij,e} &\leq x_{Ai,e} + x_{Aj,e} \\
x_{ij,e} &\leq 2 - x_{Ai,e} - x_{Aj,e} \\
x_{ij,e} &\geq x_{Ai,e} - x_{Aj,e} \\
x_{ij,e} &\geq -x_{Ai,e} + x_{Aj,e}.
\end{aligned}
$$

The first condition of Theorem 1 can be expressed as follows:

$$
\begin{aligned}
x_{Ai,f} + x_{Aj,f} + x_{Ai,e} - x_{Aj,e} &\leq 2 \\
x_{Ai,f} + x_{Aj,f} - x_{Ai,e} + x_{Aj,e} &\leq 2.
\end{aligned}
$$

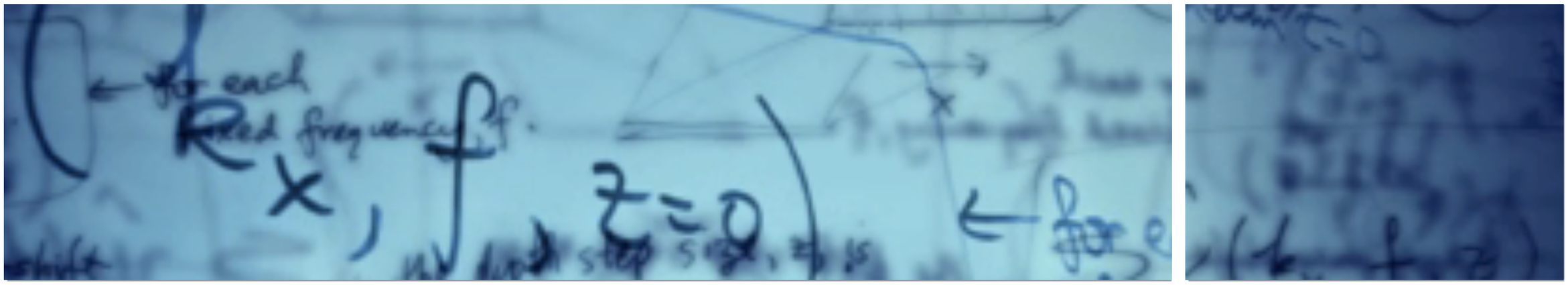
The second condition of Theorem 1 can be expressed as follows:

$$
\sum_{i \in V_e \setminus A} x_{Ai,e} \geq 2
$$

The third condition of Theorem 1 can be expressed as follows:

$$
\sum_{i \in V_e \setminus A} x_{Ai,e} \leq n - 2
$$

$$
\sum_{i \in V_e \setminus A} x_{Ai,s} \leq \sum_{i \in V_e \setminus A} x_{Ai,r} - 1 + (n-1)(2 - x_{Aj,r} - x_{Aj,s})
$$



$$\min \; z = \sum_e w_e$$

linearizing constraints

$$v_{ij,e} \leq d_{ij} x_{ij,e}$$
$$v_{ij,e} \leq w_e$$

Biological constraints

$$\sum_e v_{ije} \geq d_{ij}$$

XOR constraints

$$\begin{aligned} x_{ij,e} &\leq x_{Ai,e} + x_{Aj,e} \\ x_{ij,e} &\leq 2 - x_{Ai,e} - x_{Aj,e} \\ x_{ij,e} &\geq x_{Ai,e} - x_{Aj,e} \\ x_{ij,e} &\geq -x_{Ai,e} + x_{Aj,e}. \end{aligned}$$

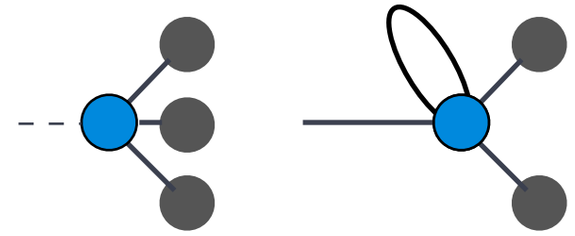
Anti-cycle constraints

$$\begin{aligned} x_{Ai,f} + x_{Aj,f} + x_{Ai,e} - x_{Aj,e} &\leq 2 \\ x_{Ai,f} + x_{Aj,f} - x_{Ai,e} + x_{Aj,e} &\leq 2. \end{aligned}$$

Degree constraints

$$\sum_{i \in V_e \setminus A} x_{Ai,e} \leq (n-2)$$

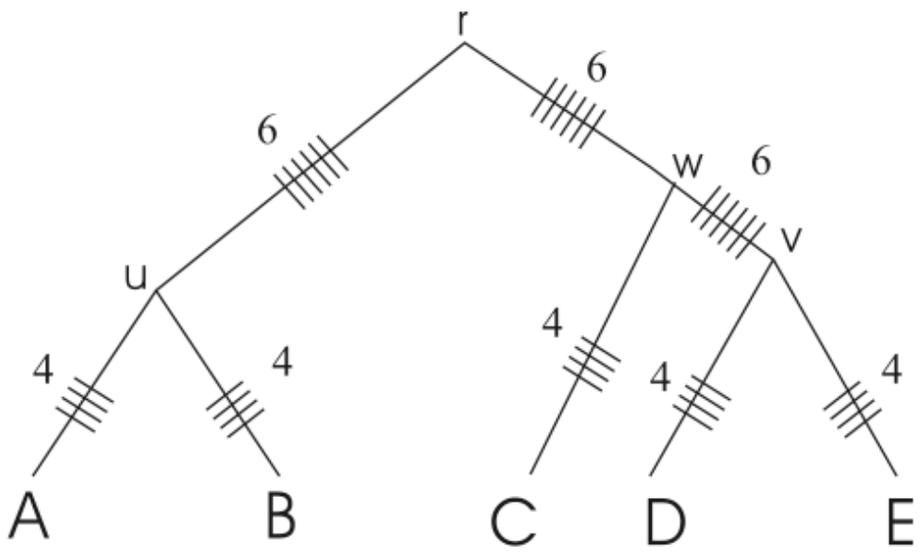
$$\sum_{i \in V_e \setminus A} x_{Ai,s} \leq \sum_{i \in V_e \setminus A} x_{Ai,r} - 1 + (n-1)(2 - x_{Aj,r} - x_{Aj,s})$$



The EPT model

Daniele Catanzaro

The minimum evolution problem



Measure of the dissimilarity between species A and B

Error relative to the approximation

$$w_{Au} + w_{Bu} = d_{AB} + e_{AB}$$

$$w_{Au} + w_{ur} + w_{rw} + w_{wC} = d_{AC} + e_{AC}$$

$$w_{Au} + w_{ur} + w_{rw} + w_{wv} + w_{vD} = d_{AD} + e_{AD}$$

$$w_{Au} + w_{ur} + w_{rw} + w_{wv} + w_{vE} = d_{AE} + e_{AE}$$

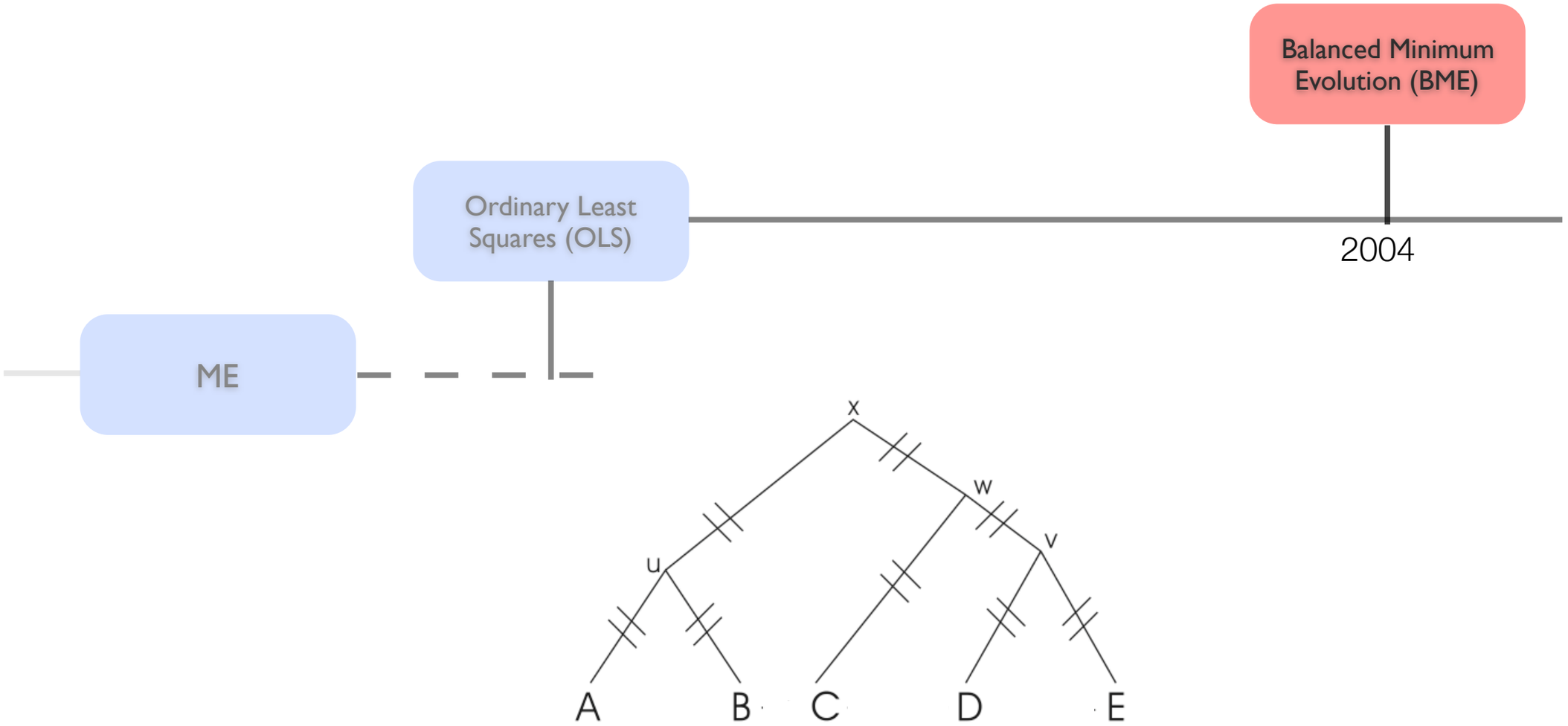
$$w_{Bu} + w_{ur} + w_{rw} + w_{wC} = d_{BC} + e_{BC}$$

...

A different version of MEP: BME

Daniele Catanzaro

The minimum evolution problem



The balanced minimum evolution criterion of phylogenetic estimation

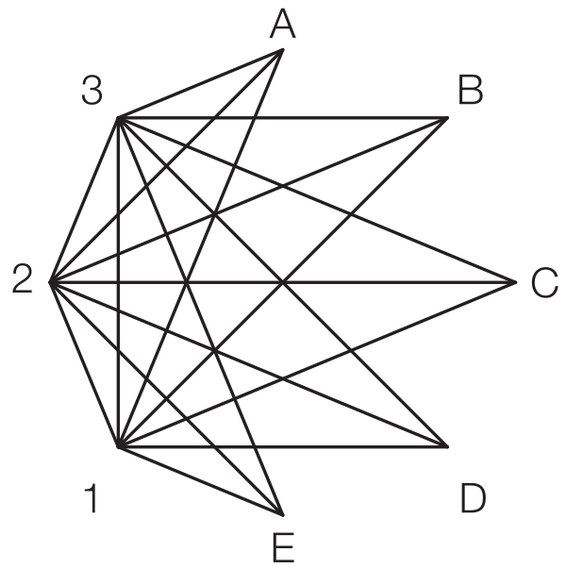
Daniele Catanzaro

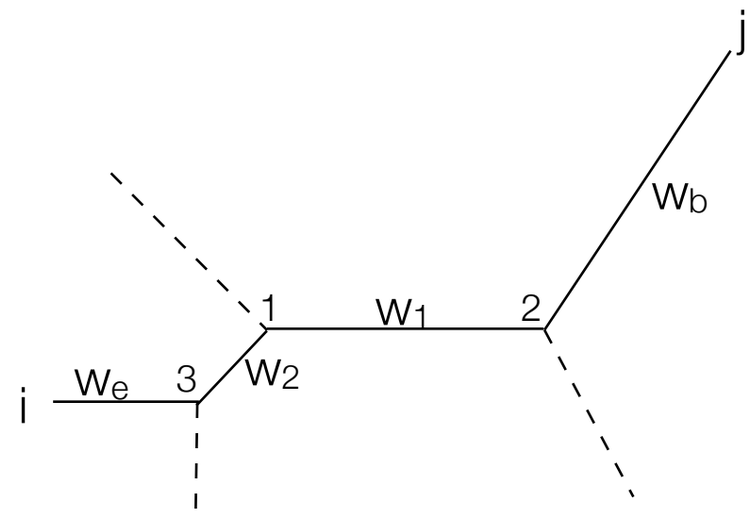
The minimum evolution problem

A minimal length phylogeny provides a lower bound on the overall amount of mutation events occurred along evolution of the set of species analyzed.

The balanced minimum evolution criterion is a variation of ME in which the length of a phylogeny is computed as:

$$L = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{d_{ij}}{2^{\tau_{ij}}}$$



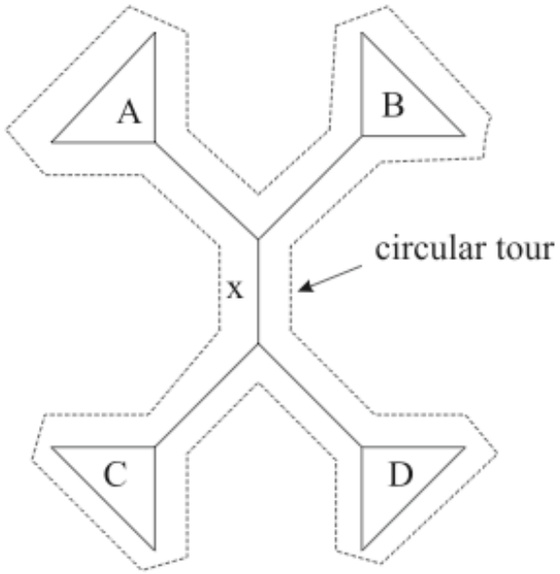


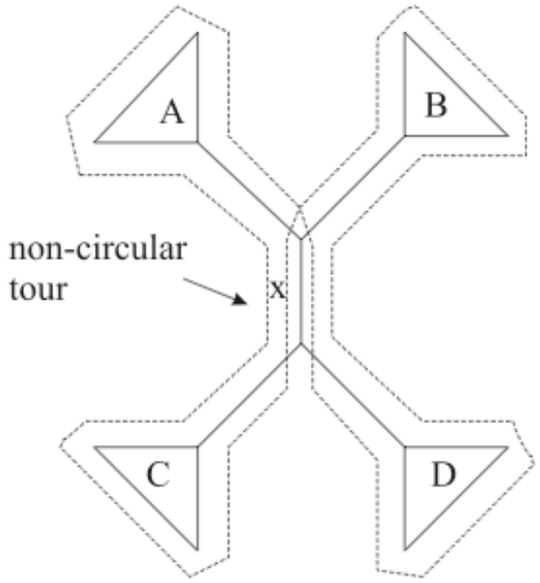
Fundamentals of the balanced minimum evolution criterion

Daniele Catanzaro

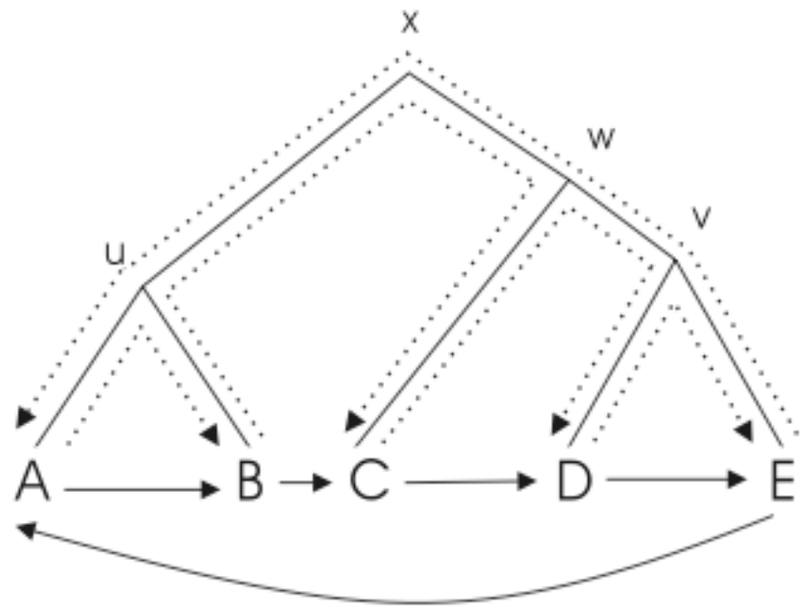
The minimum evolution problem

The phylogeny length under BME is equivalent to the average of the circular orders associated to a given phylogeny.





$$\mathbf{L} = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{\mathbf{d}_{ij}}{\mathbf{2}^{\tau_{ij}}}$$



$$L(T) = \frac{1}{|o(T)|} \sum_{\pi \in o(T)} \frac{1}{2} \sum_{i=1}^{n} d_{x_i, x_{i+1}}$$

sum of the edge weight
belonging to the path from leaves x_i to x_{i+1}

Combinatorial interpretation of BME

Daniele Catanzaro

The minimum evolution problem

The problem of finding a phylogeny which satisfies the balanced minimum evolution criterion is known as Balanced Minimum Evolution Problem (BME) and consists of minimizing the function

$$L = \sum_{i=1}^{n} \sum_{j=1}^{n} \frac{d_{ij}}{2^{\tau_{ij}}}$$

with the constraint that {τij} form a phylogeny.

BME is in P if

$$d_{zi} + d_{kj} \leq \max\{d_{zj} + d_{ik}, d_{kz} + d_{ij}\}$$

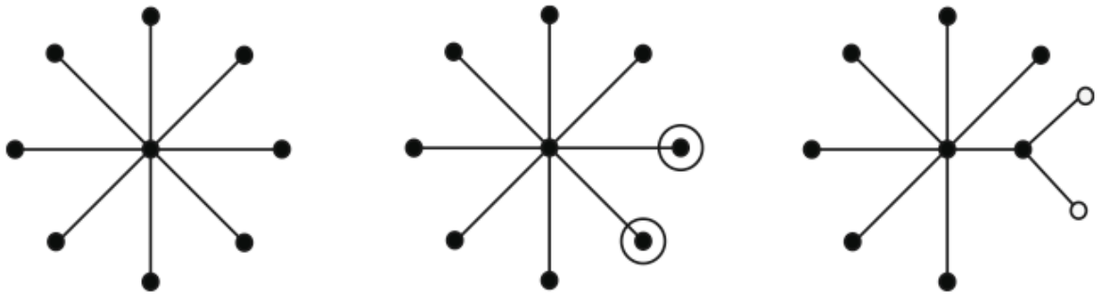
$$d_{ij} \leq \max\{d_{ik}, d_{kj}\}$$

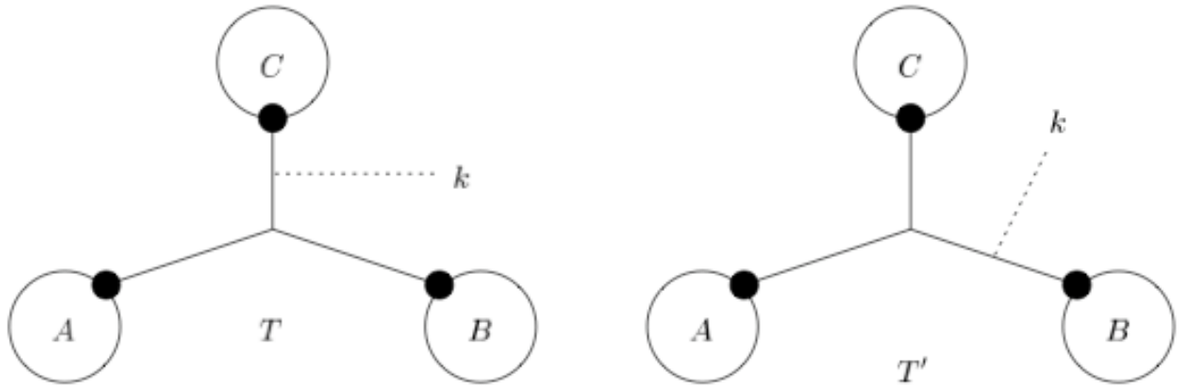
However in the most general case the complexity of BME is unknown.

The Balanced Minimum Evolution Problem (BME)

Daniele Catanzaro

The minimum evolution problem





Approaches to solution: Heuristics

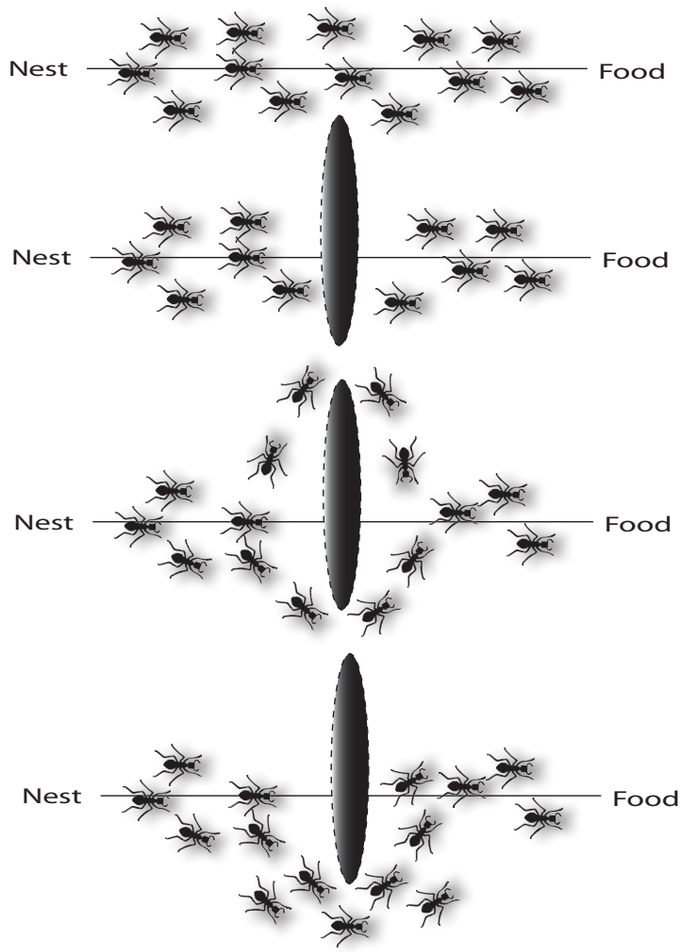
Daniele Catanzaro

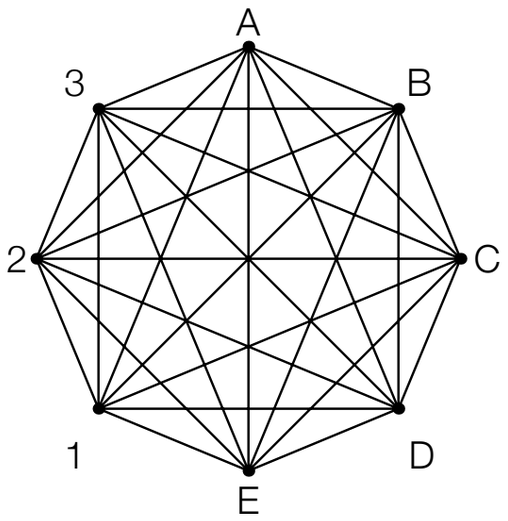
The minimum evolution problem

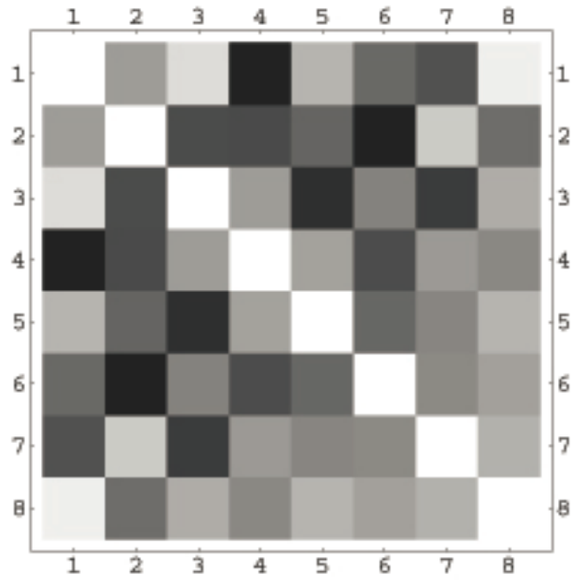
Approximate algorithm for MEP

Daniele Catanzaro

The minimum evolution problem





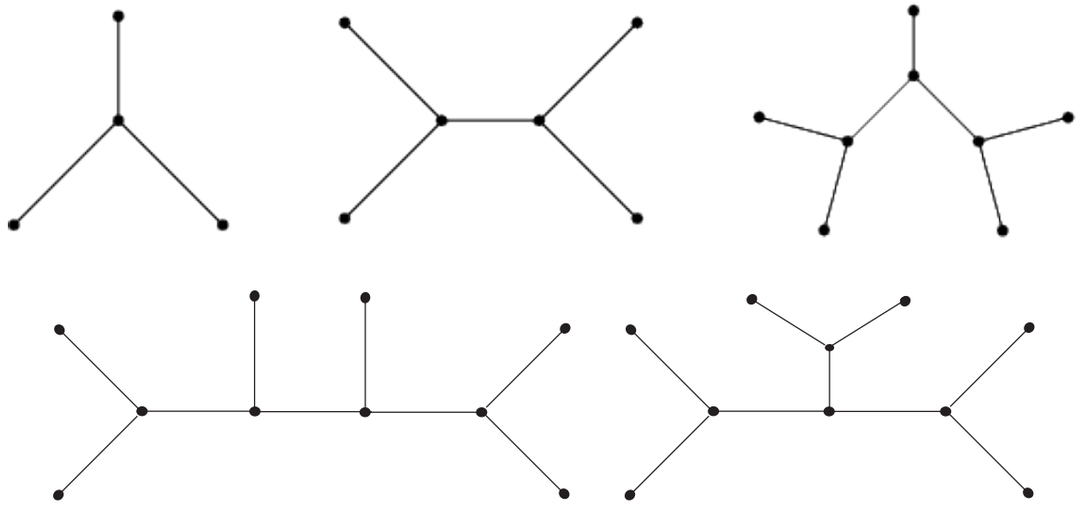


$$p_{ij} = \frac{\alpha\tau_{ij} + (1-\alpha)\eta_{ij}}{\sum_{q\in\Gamma\setminus\Gamma_{\mathcal{G}_k}} \alpha\tau_{iq} + (1-\alpha)\eta_{iq}}$$

Ant philosophy

Daniele Catanzaro

The minimum evolution problem



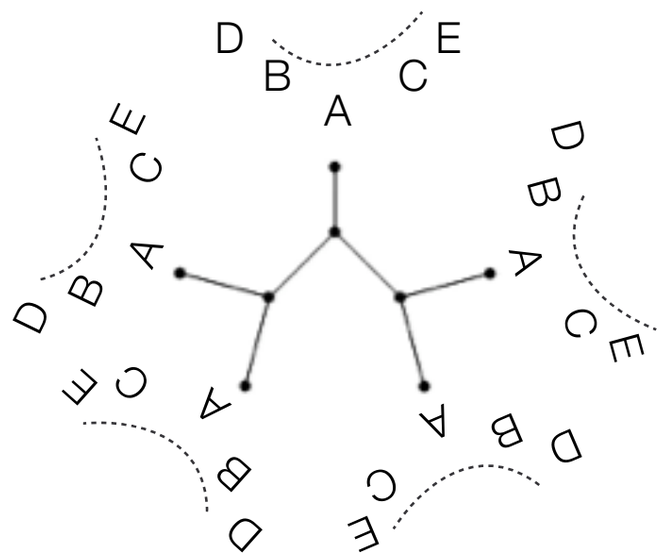
Leaves	NI-Shapes	Shapes
3	1	3
4	1	15
5	1	105
6	2	945
7	2	10395
8	3	135135
9	4	2027025
10	11	34459425
15	265	1012
20	11020	1021
30	14502229	1038
40	11077270355	1057

The total number of possible trees with n leaves is (2n-3)!!
The number of non-isomorphic shapes increases more slowly!
A possible strategy to solve the optimization problem could be:
(i) to enumerate all the possible non-isomorphic shapes (P# problem); (ii) to find an optimal assignment for each non-isomorphic shape (NP-hard problem).

Non-isomorphic generation

Daniele Catanzaro

The minimum evolution problem



	v_1	v_2	v_3	v_4	v_5	v_6	v_7
AB	1	1	0	0	0	0	1
AC	1	0	1	0	0	1	0
AD	1	0	0	1	0	0	1
AE	1	0	0	0	1	1	0
BC	0	1	1	0	0	1	1
BD	0	1	0	1	0	0	0
BE	0	1	0	0	1	1	1
CD	0	0	1	1	0	1	1
CE	0	0	1	0	1	0	0
DE	0	0	0	1	1	1	1

= **X**

It is possible to change the assignment of k leaves on the tree by simply swapping the k corresponding rows of **X**.

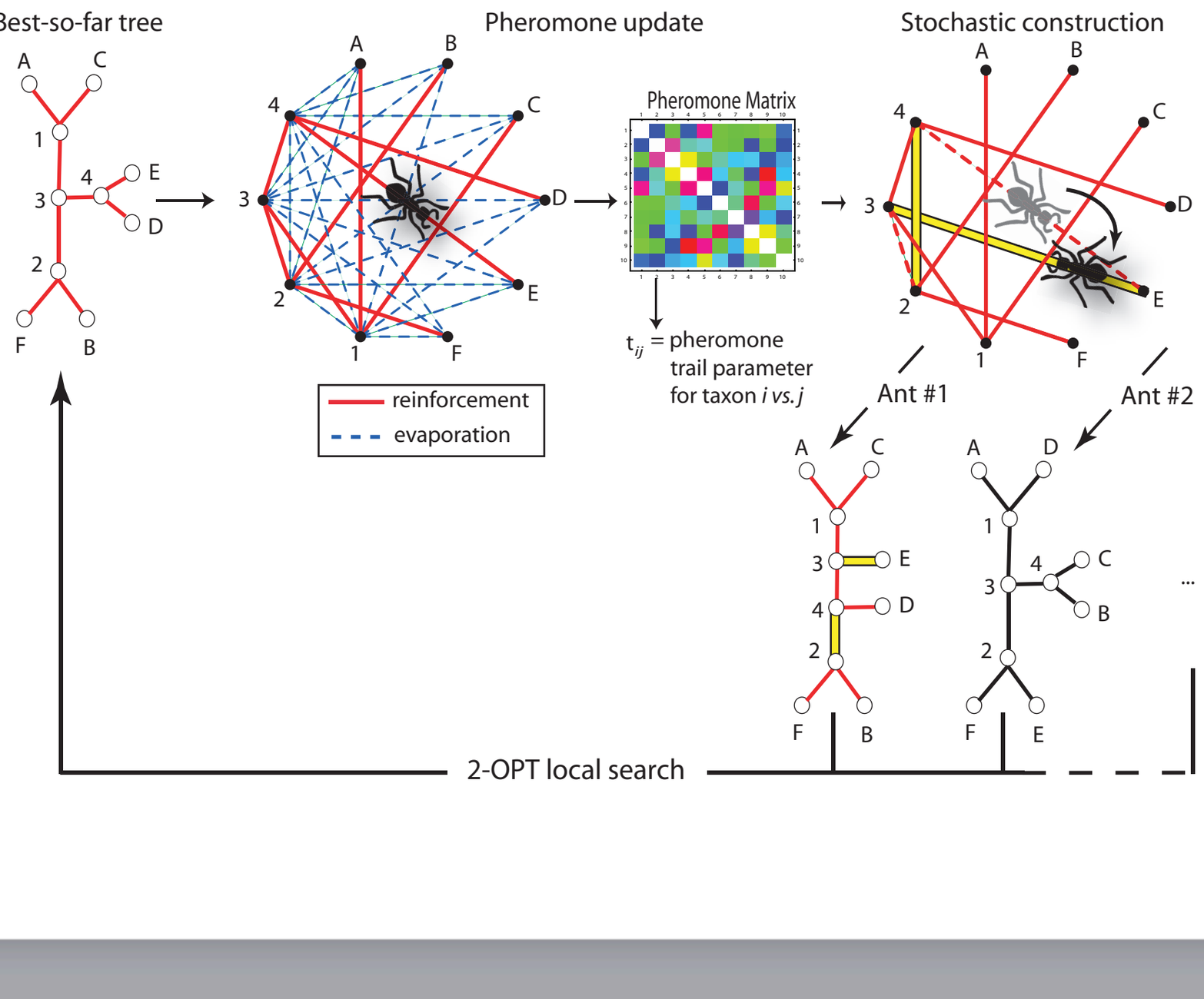
The k-OPT local search is therefore easy and fast.

Contra: it seems not possible to get an a-priori cost of each swap; consequently each swap requires a complete evaluation of the entire tree (heavy).

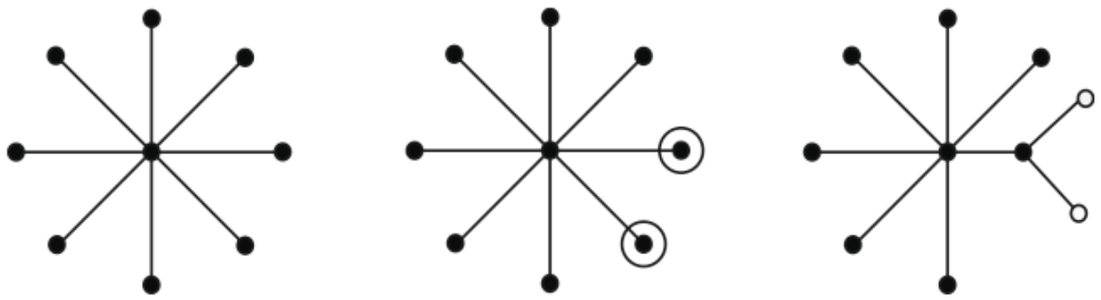
Species assingment

Daniele Catanzaro

The minimum evolution problem



Ant algorithm

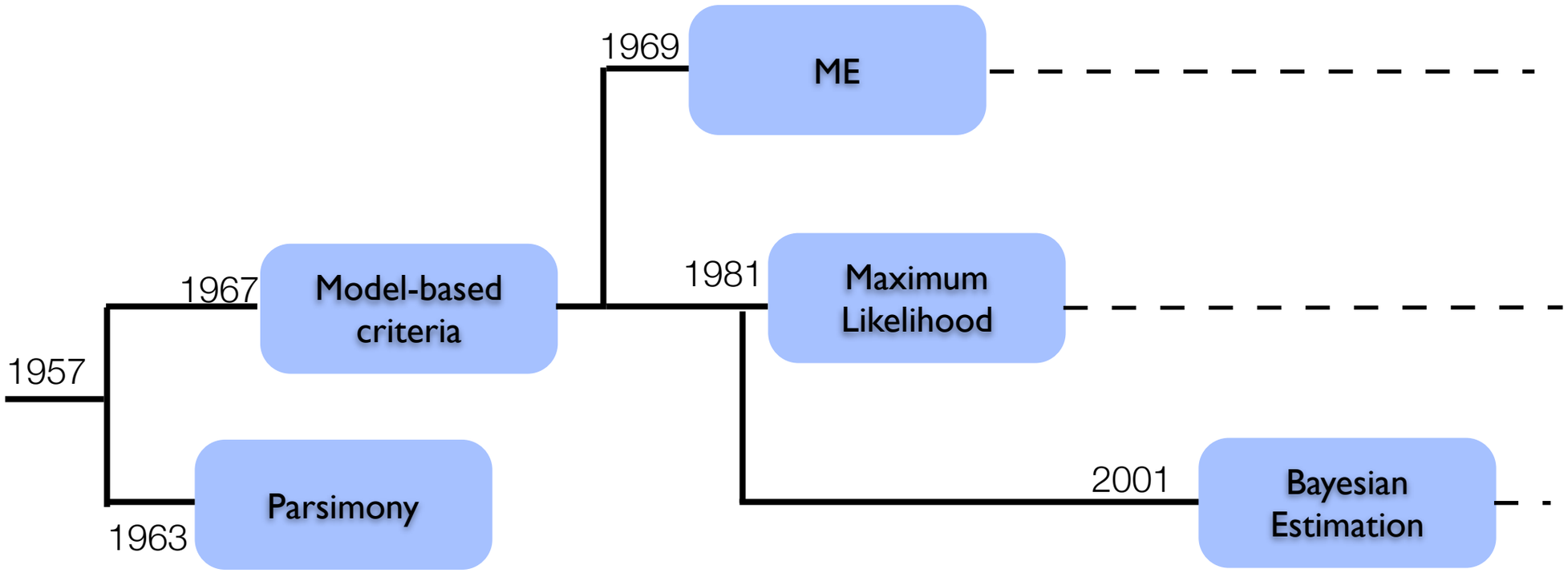


$$p_{ij} = \frac{\alpha\tau_{ij} + (1-\alpha)\eta_{ij}}{\sum_{q \in \Gamma \setminus \Gamma_{\mathcal{G}_k}} \alpha\tau_{qj} + (1-\alpha)\eta_{qj}}$$

$$\eta_{ij} = (\Delta_{ij} - u_i - u_j)^{-1}$$

$$\tau_{ij} \leftarrow (1-\rho)\tau_{ij} + \epsilon_{ij}$$

$$\epsilon_{ij} = \begin{cases} \kappa\rho / l_{\mathcal{G}_{|\Gamma|}}, & \text{if } w_i \text{ is adjacent to } w_j \text{ in solution } s^{best}; \\ 0, & \text{otherwise.} \end{cases}$$

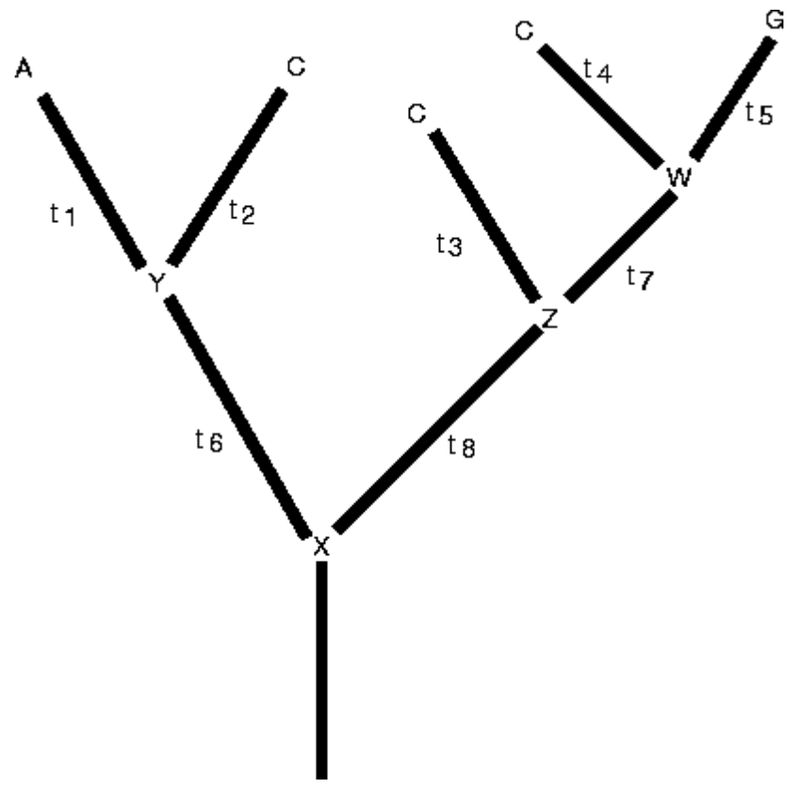


Phylogenetic estimation models

Daniele Catanzaro

The minimum evolution problem

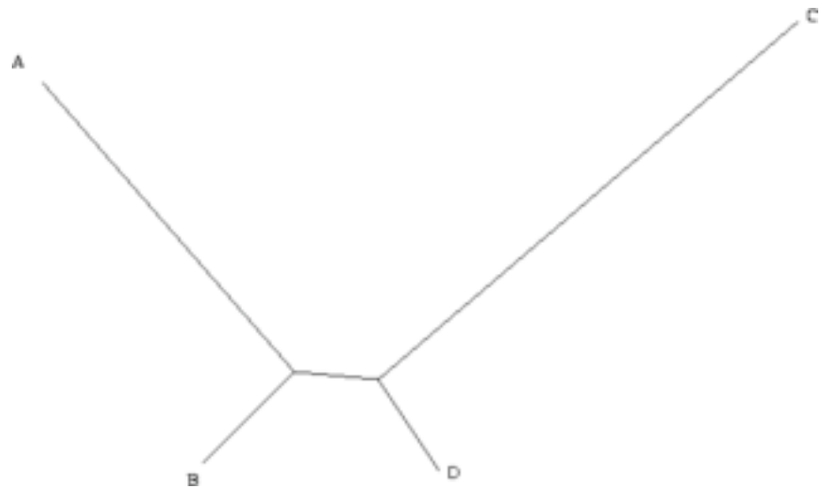
The likelihood criterion states that under many plausible explanations of an observed phenomenon, the one with the highest probability of occurring should be preferred. Hence, under the likelihood criterion, a phylogeny is defined to be optimal (or the most likely) if it has the highest probability of explaining the observed taxa.



The likelihood criterion of phylogenetic estimation

Daniele Catanzaro
A few notes in molecular phylogenetic estimation

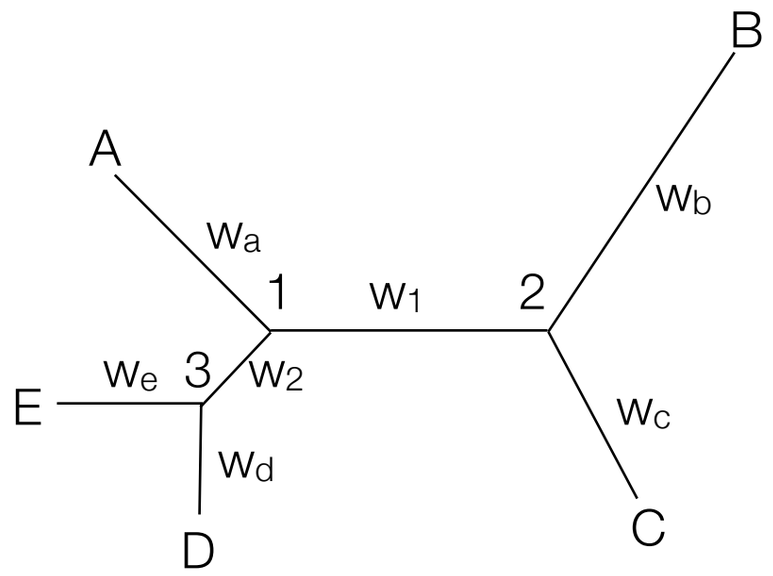
The likelihood criterion is introduced by Joe Felsenstein in 1981, as an attempt at solving some “distortion” problems of Cavalli-Sforza and Edwards model. In fact, it is possible to prove that in presence of high mutation rates (evidenced e.g., by a high divergence of the molecular sequences) or convergent/divergent evolution, Cavalli-Sforza and Edwards model leads to edge weights having values quite different from the true phylogeny. This phenomenon is known as “long branch attraction”.



The likelihood criterion of phylogenetic estimation

Daniele Catanzaro

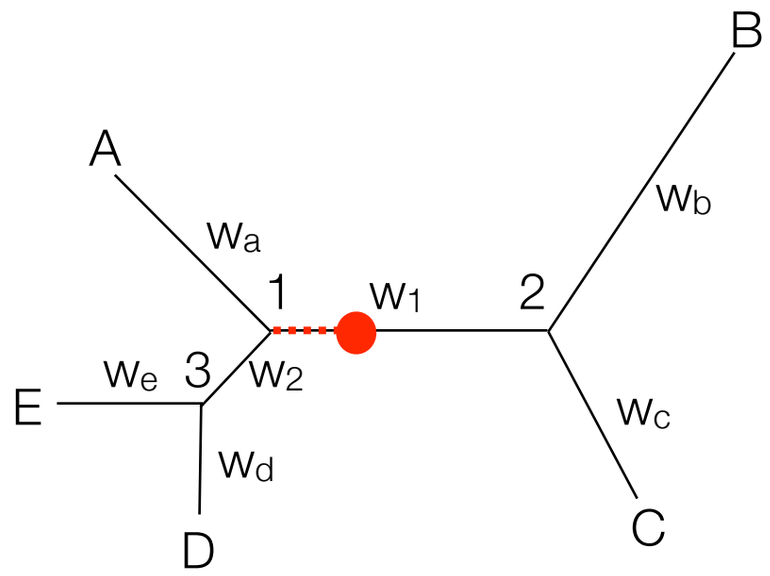
A few notes in molecular phylogenetic estimation



Rooted phylogenies

Daniele Catanzaro

A few notes in molecular phylogenetic estimation

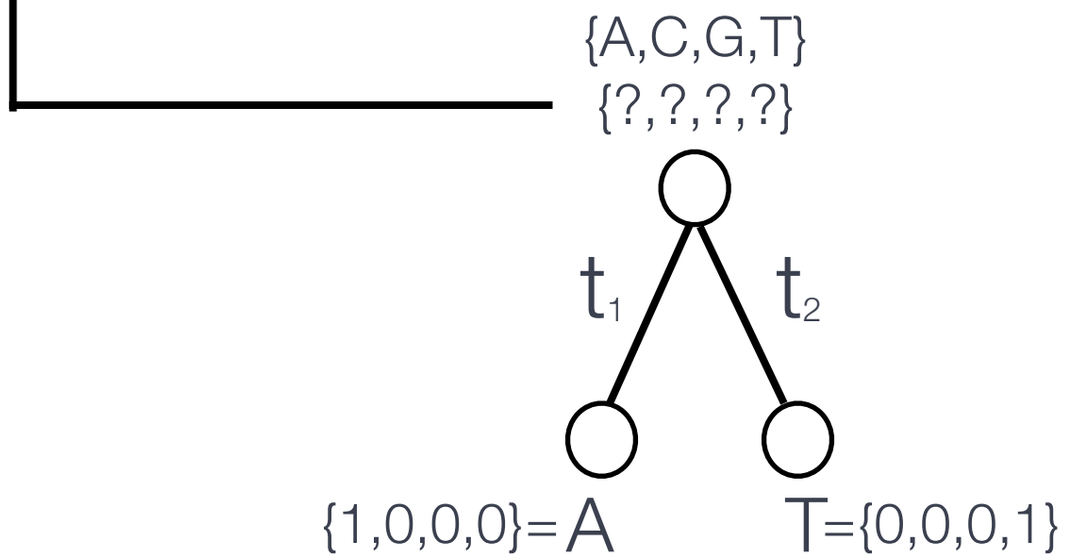


Rooted phylogenies

Daniele Catanzaro

A few notes in molecular phylogenetic estimation

$$S(j) = \left(\sum_{i=\{A,C,G,T\}} child_1(i)\, P_{ij}(t_1) \right) \left(\sum_{i=\{A,C,G,T\}} child_2(i)\, P_{ij}(t_2) \right)$$



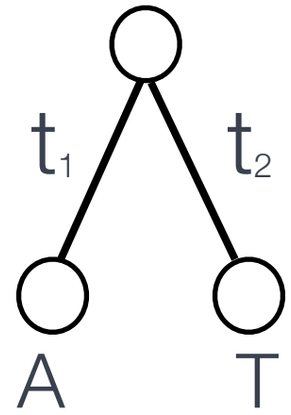
Likelihood score of a phylogeny

Daniele Catanzaro

A few notes in molecular phylogenetic estimation

$$L = \sum_{j=1}^{4} S\,(j)\ \mathbf{frequency}\ (j)$$

{0.1,0.2,0.4,0.4}



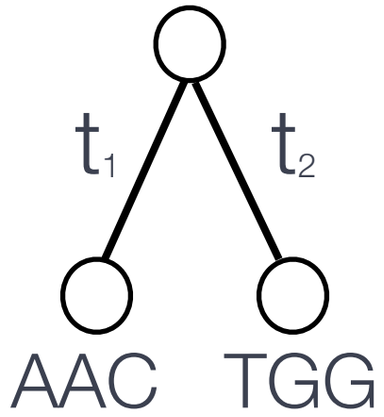
Likelihood score of a phylogeny

Daniele Catanzaro

A few notes in molecular phylogenetic estimation

$$L = \prod_{c=1}^{3} \sum_{j=1}^{4} S\,(j)\ \mathbf{frequency}\,(j)$$

{{0.1,0.2,0.4,0.4}, {0.3,0.12,0.2,0.1}, {0.5,0.1,0,0.1}}



Likelihood score of a phylogeny

Daniele Catanzaro

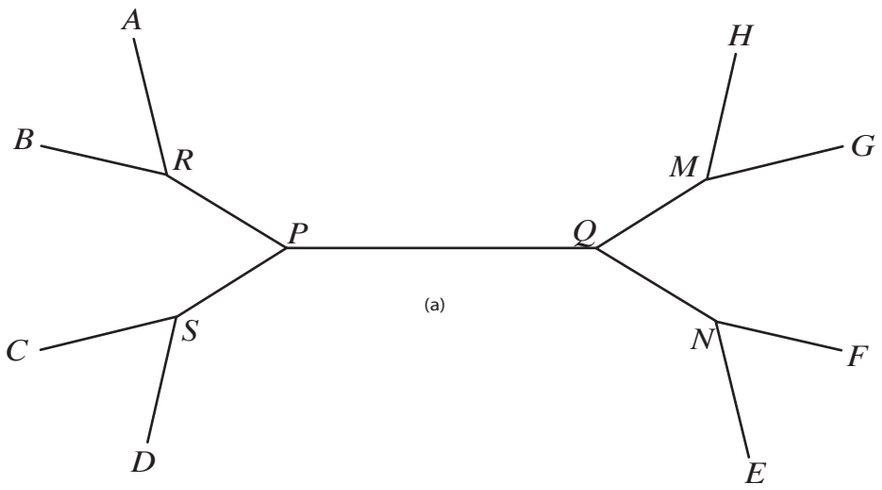
A few notes in molecular phylogenetic estimation

$$\underset{T,t,P}{\text{maximize}}\, L\,(T,\, t,\, P)$$

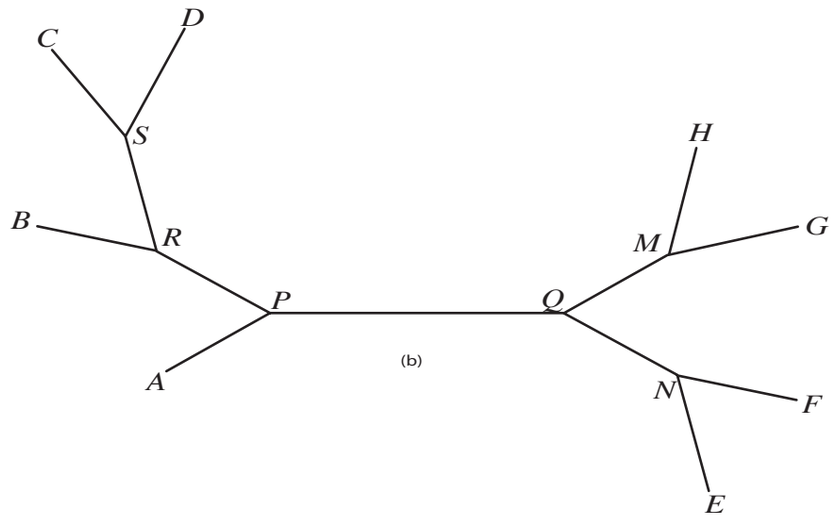
The estimation problem

Daniele Catanzaro

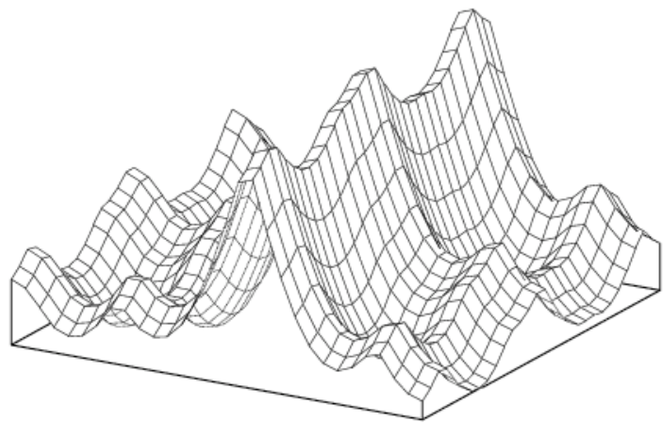
A few notes in molecular phylogenetic estimation



(a)



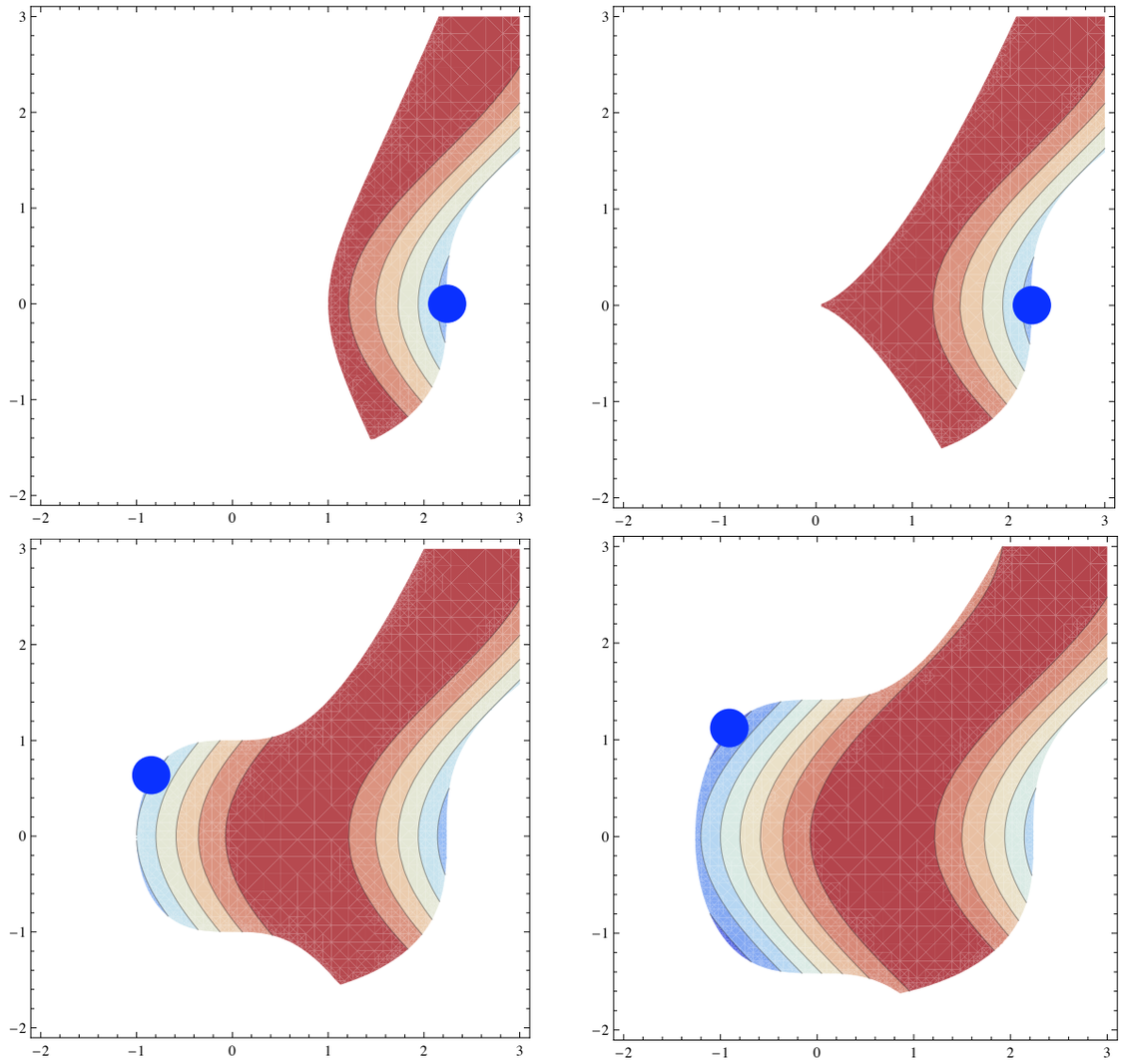
(b)



In the literature...

Daniele Catanzaro

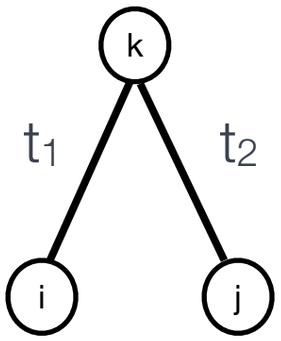
A few notes in molecular phylogenetic estimation



Very Large-Scale Neighborhood (VLSN) techniques

Daniele Catanzaro

A few notes in molecular phylogenetic estimation



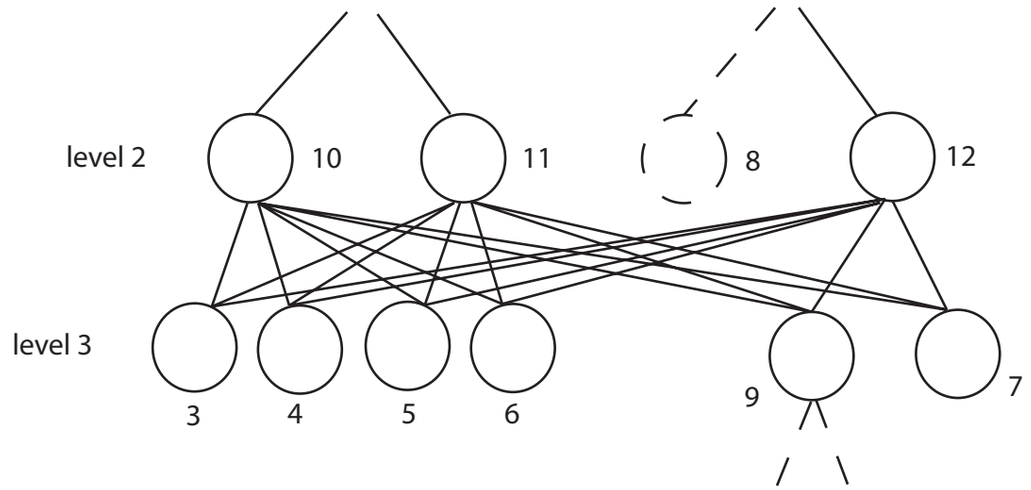
$$c_{ij,k} = \|S_k - S_j\| + \|S_k - S_i\|$$

$$c_{ij,k} = L(S_k|S_i, S_j, w_{ki}, w_{kj})$$

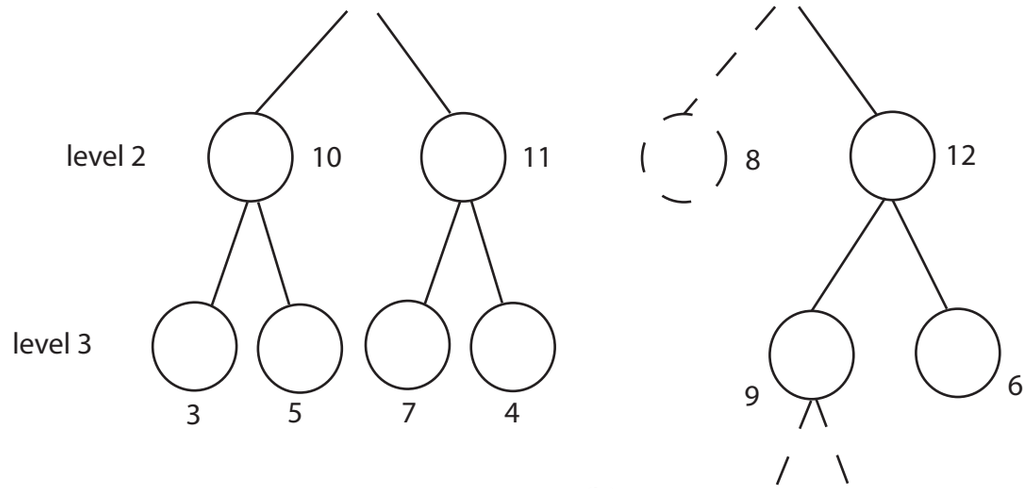
VLSN techniques for phylogeny estimation

Daniele Catanzaro

A few notes in molecular phylogenetic estimation



(a)

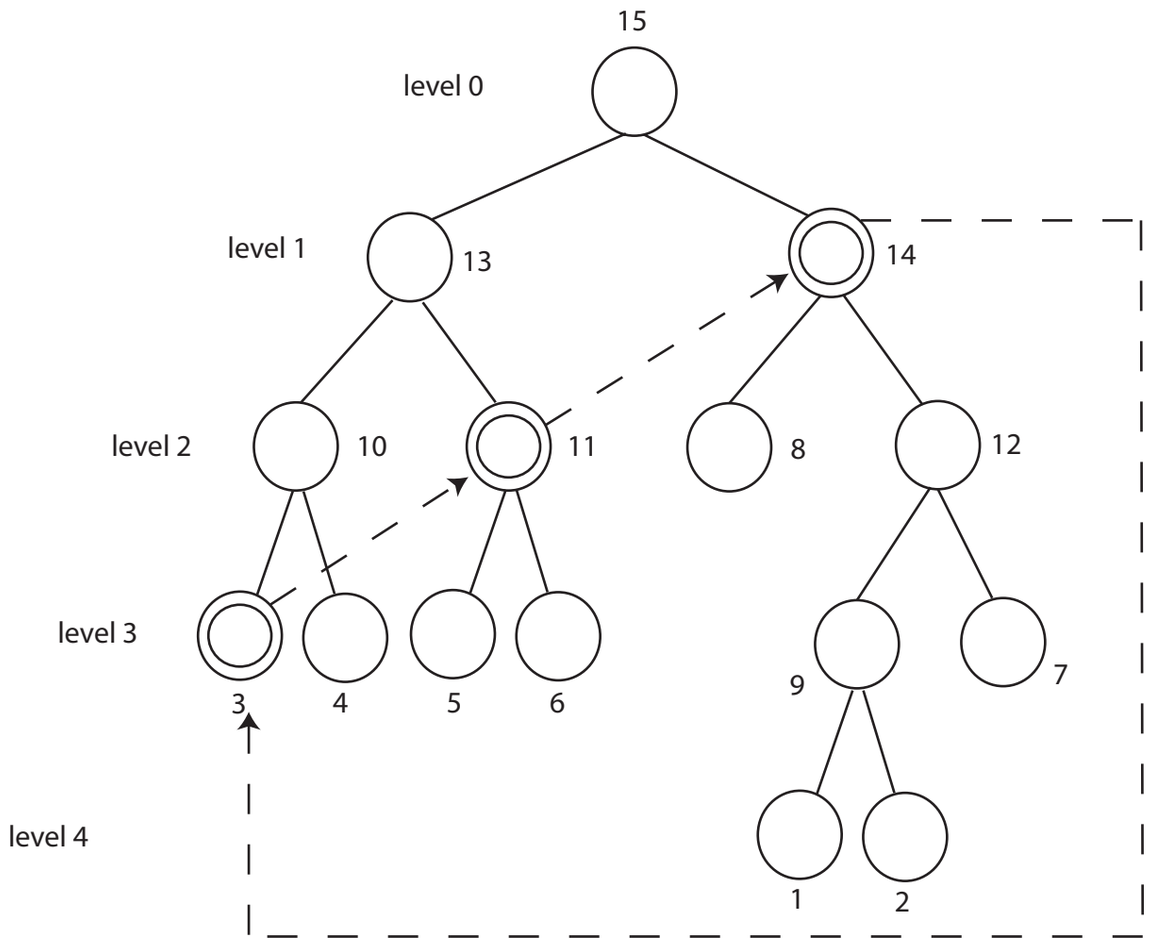


(b)

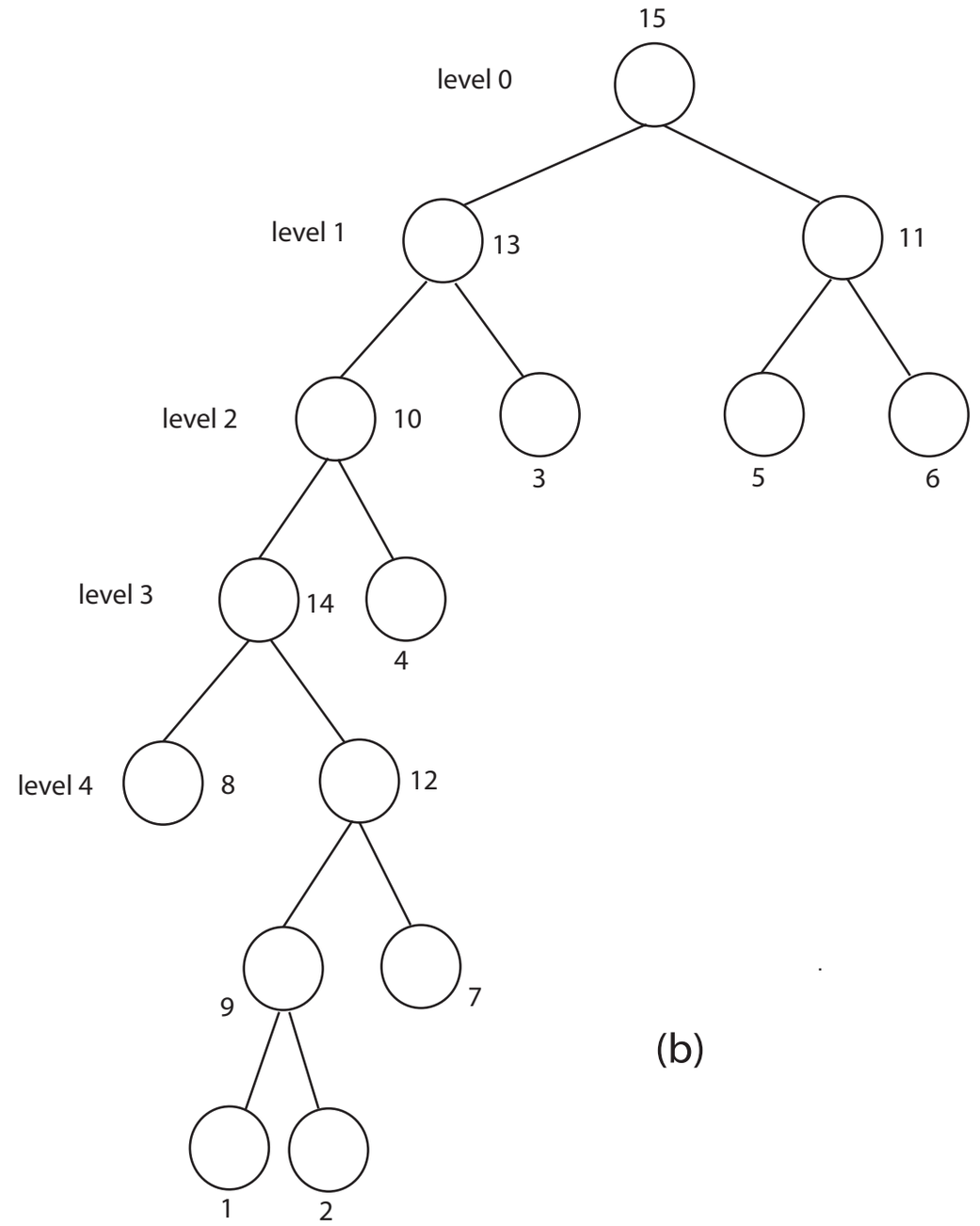
Minimum Cost Assignment Neighborhood

Daniele Catanzaro

A few notes in molecular phylogenetic estimation



(a)

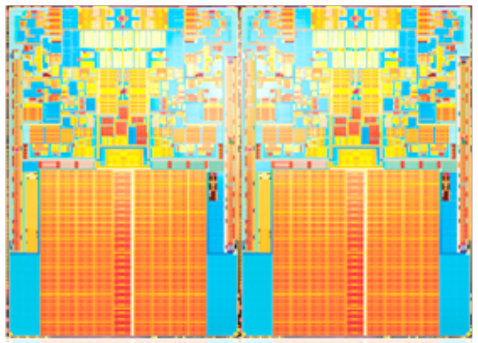


(b)

Minimum Cost Cycle Neighborhood

Daniele Catanzaro

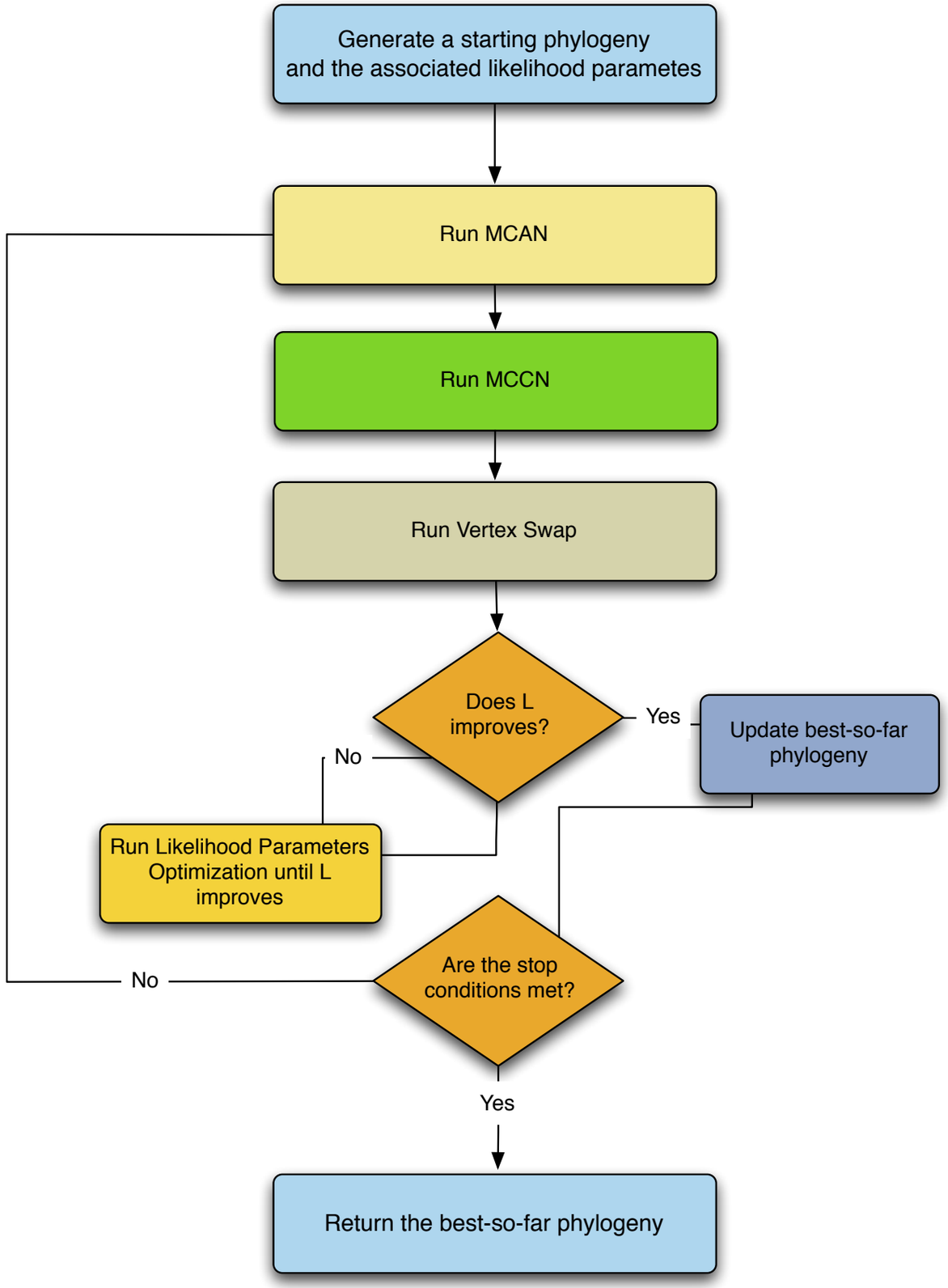
A few notes in molecular phylogenetic estimation

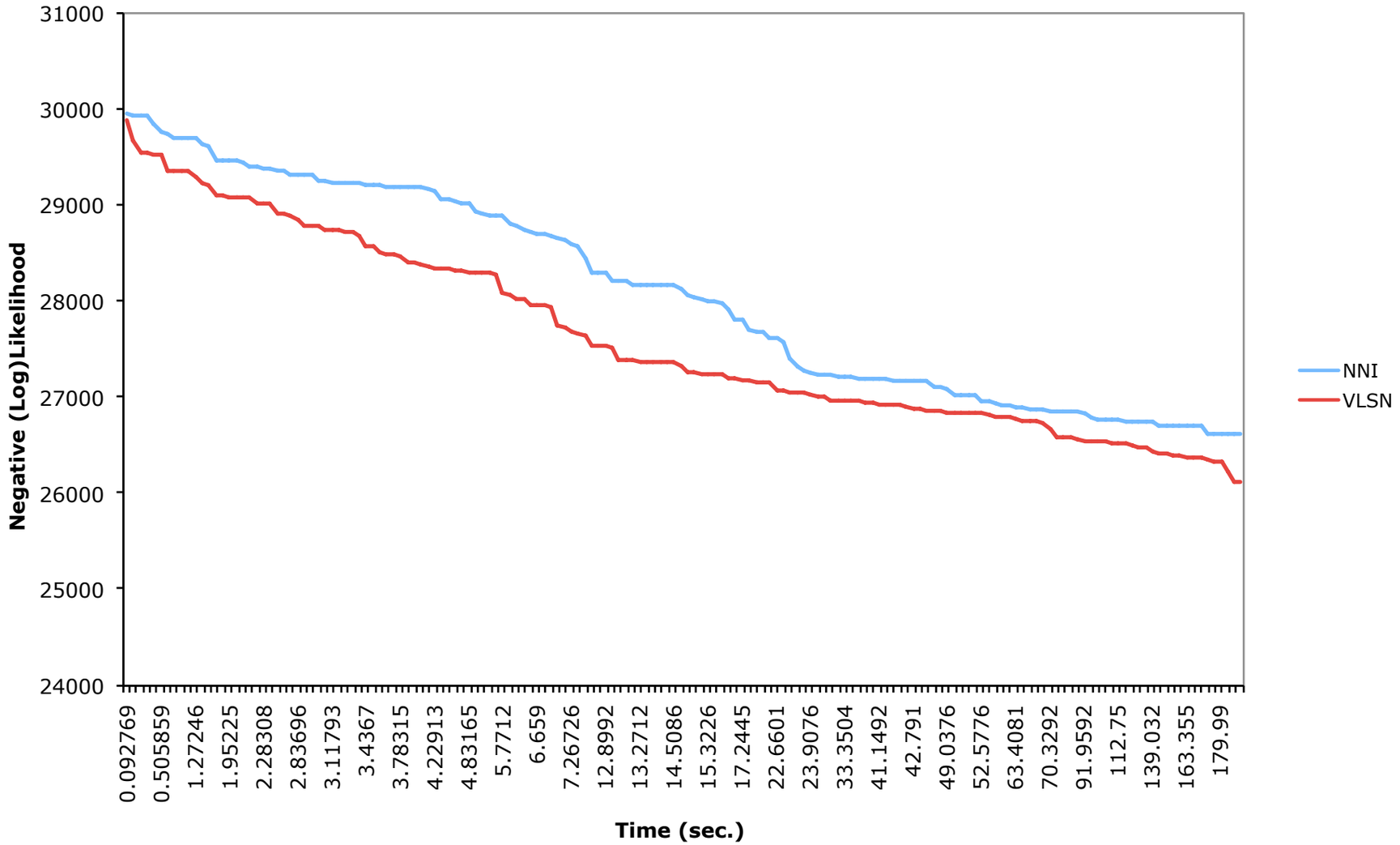


The algorithm

Estimating phylogenies under maximum likelihood: A very large-scale neighborhood approach

D. Catanzaro, R. Pesenti, and M. C. Milinkovitch - Université Libre de Bruxelles





Numerical results

Daniele Catanzaro

A few notes in molecular phylogenetic estimation

Instance: Rana64/1976

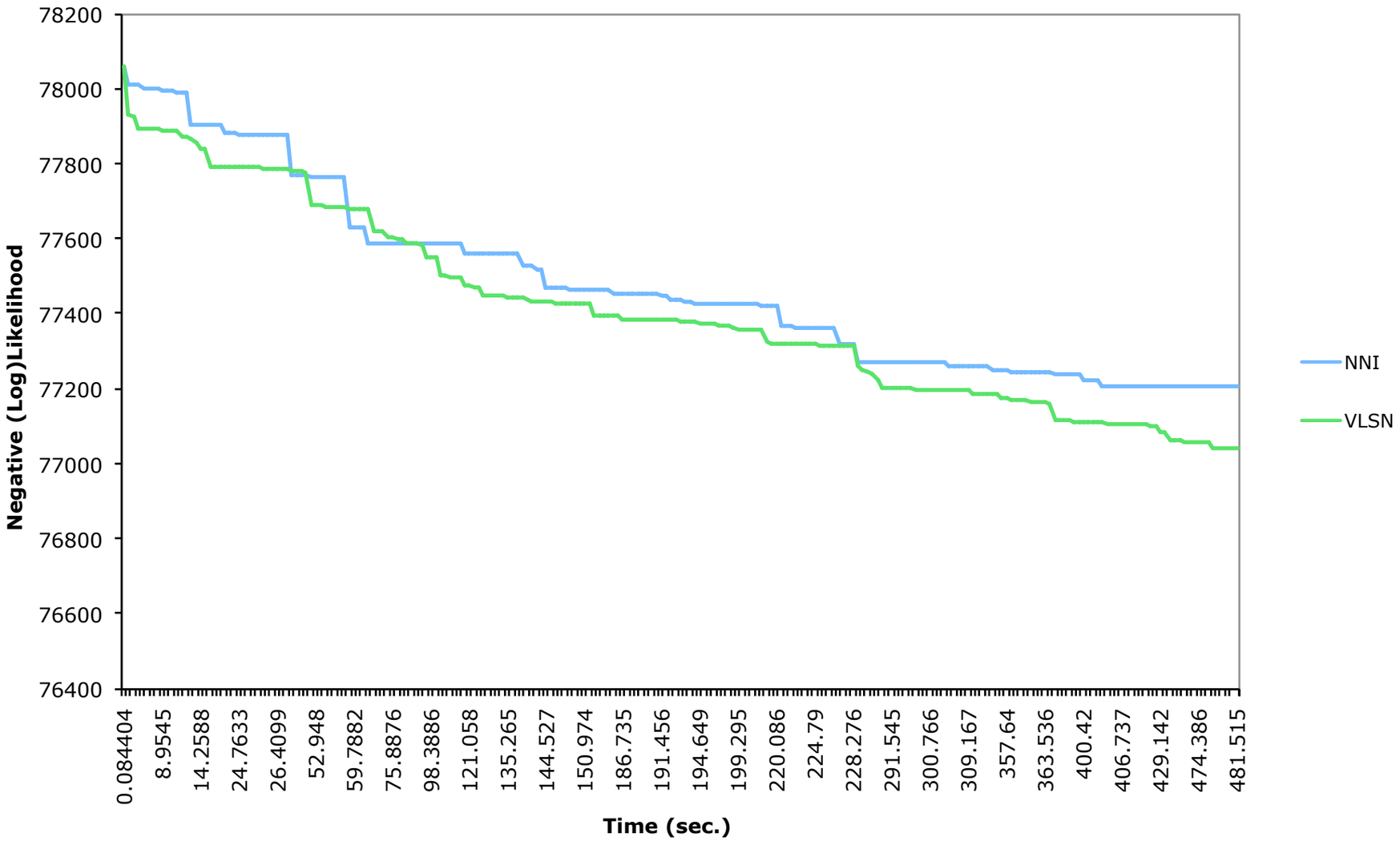
Neighborhood	Time (sec.)	Number of calls
NNI [1]	6.2±1	328±3
VLSN	3.1±0.12	38±1

[1]Paup implementation.

Numerical results

Daniele Catanzaro

A few notes in molecular phylogenetic estimation



Numerical results

Daniele Catanzaro

A few notes in molecular phylogenetic estimation

Instance: Zilla300/1428

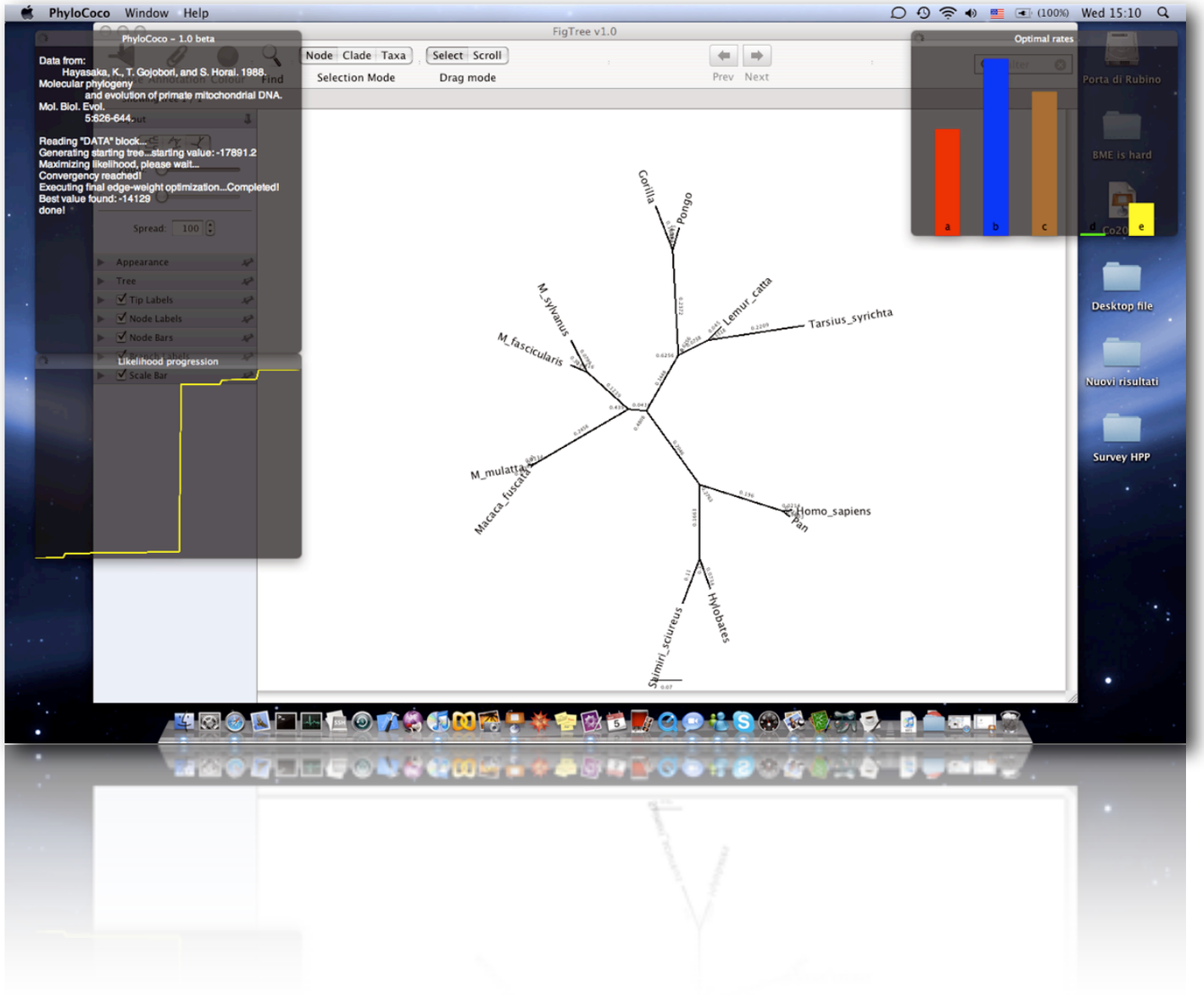
Neighborhood	Time (sec.)	Number of calls
NNI [1]	n.a.	>149±1
VLSN	413±1	116±4

[1]Paup implementation.

Numerical results

Daniele Catanzaro

A few notes in molecular phylogenetic estimation



PhyloCoco

Daniele Catanzaro

A few notes in molecular phylogenetic estimation

Phylogenetic estimation is one of the most important problem in computational biology. It is a flourishing area of interaction between molecular biology, operations research, computer science, and physics.

The day by day growing amount of molecular data stored in public databases forces to search for:

- Ad hoc models of molecular evolution
- Ad hoc models of phylogenetic estimation
- Optimization algorithms to select phylogenies among possible alternatives

Here we have presented a first introduction to phylogenetics. The most relevant issues used in tackling real-world sized problems have been outlined, as have the most interesting refinements deserving further research effort.

Conclusion

Phylogenetics

Daniele Catanzaro

A few notes in molecular phylogenetic estimation

The minimum evolution problem

Daniele Catanzaro
Service Graphes et Optimisation Mathématique (G.O.M.)
Université Libre de Bruxelles