TOWARDS THE OPTIMIZATION OF LARGE-SCALE PHYLOGENETIC TREES

CÁTIA VAZ and ALEXANDRE FRANCISCO

INESC-ID;

Instituto Superior de Lisboa, Instituto Politécnico de Lisboa

COMPSTAT 2022

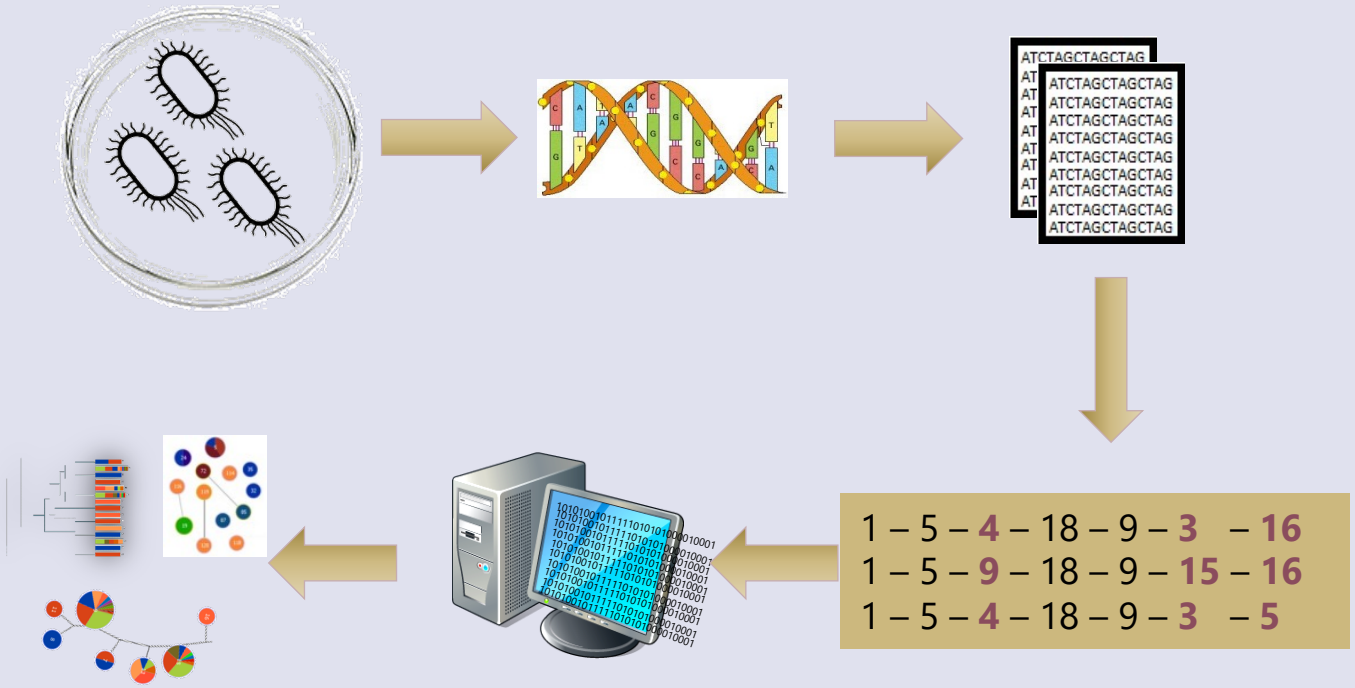
PHYLOGENETIC INFERENCE

PHYLOGENETICS is the study of the evolutionary history and relationships among individuals or groups of organisms (e.g.. species, or populations) by grouping them based on some ***(dis)similarity criterion*** that underlies some ***evolution model***.

- The result is a phylogenetic tree or network

Fundamental in epidemiological and genetic studies for the surveillance of infectious diseases and about its evolution.

PHYLOGENETIC INFERENCE



The evolutionary relationships between different species or taxa are usually inferred through known phylogenetic analysis techniques

- **Sequencing** the isolate data
- **Assembly** the sequences, comparing the draft genome to a database of gene alleles
- **Profiling** - Given the assembly results we can create an allelic profile characterizing the strain
  - Usually abstracted to categorical indexes.

Phylogenetic tree

- After profiling, the reconstruction of a phylogenetic tree can be done by several different inference methods or algorithms.

- Most of them rely on applying clustering techniques
  - the main difference among them resides on how is defined **cluster proximity (similarity)** and on which **optimization criterion** is used.
  - Both *cluster proximity (similarity)* and *optimization criteria* rely often on a **model of evolution**

- Phylogenetic trees can be mainly built using either **distance-based methods** or **character state methods**.

- **Distance-based methods** infer the relationship between individuals as the number of genetic differences between pairs of sequences

Cluster Proximity/Similarity/Dissimilarity

- Similarities are described according to several models like connectivity models, centroid models, distribution models…
- Distance based methods are based on *connectivity models*.
- ***Connectivity models*** define the similarity between elements as their distance
  - elements are more similar with nearby elements than with elements farther away)

Algorithms that implement this model differ from one another by the way distances are computed.

	A	B	C
A	-	17	11
B	17	-	28
C	11	28	-

Hamming Distance

A	1 – 5 – **4** – 18 – 9 – **3** – **16**
B	1 – 5 – **9** – 18 – 9 – **15** – **16**
C	1 – 5 – **4** – 18 – 9 – **3** – **5**

Profile data

Common PHYLOGENETIC Analysis workflow

After Profiling, the phylogenetic analysis workflow using *distance based methods* can be summarized into four consecutive steps:

- distance calculation,
- distance correction,
- inference algorithm,
- local optimization steps.

- **Distance Calculation**: e.g. Hamming (normalized or unnormalized), Euclidian, etc;
- **Distance Correction**: based on some model of evolution, e.g.: Jukes-Cantor, Kimura, Felsenstein, etc.
- **Inference algorithms** and **local optimization** steps uses different different principles/criterions

  e.g.: Minimum evolution, minimum deviation, etc.

Inference algorithms/methods

Different methods can built different trees.

The strategy that each method uses to find the best tree topology can be classified as:

- exhaustive search methods
- completely bifurcating tree search methods
- stepwise clustering method

The phylogenetic tree produced by an inference method is not unique (Backeljau et al, 1996; Teixeira et al, 2015), depending on the inference method or even on the dataset input order.

Inferred trees might also not necessarily represent the best tree for the underlying evolution model

goeBURST algorithm

goeBURST follows a graph theoretic approach to better estimate a true phylogeny, reflecting the principle of minimum evolution.

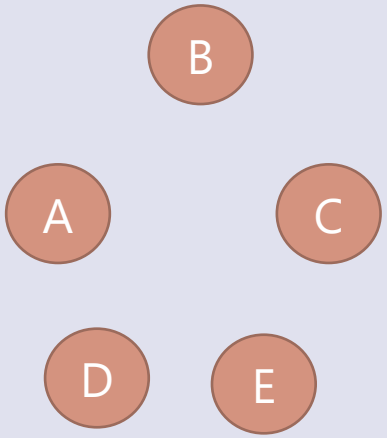
The problem is similar to the **minimum spanning tree problem**, trying to identify a subset of relationships ("links") that connects the profiles or OTUs ("points") without any cycles and with the minimum possible total link length.

It is a stepwise algorithm, which interactively process:

- The ***cluster-pair selection***: selects links connecting two OTUs with the minimum distance.
  - **Tie break rules**: select the links between OTUs with higher number of:
    1. SLVs
    2. DLVs
    3. TLVs
    4. OTU's frequency
    5. OTU's ID
- ***Cluster-pair joining***: Join those two OTUs.

goeBURST algorithm

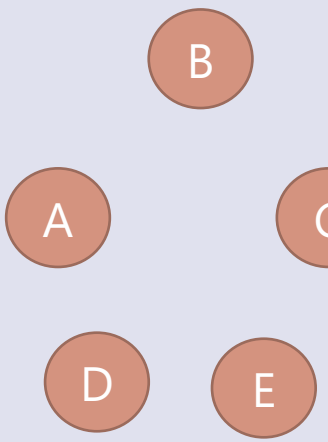
Profile	A	B	C	D	E
A	-	2	7	7	6
B	-	-	7	7	6
C	-	-	-	5	5
D	-	-	-	-	3
E	-	-	-	-	-



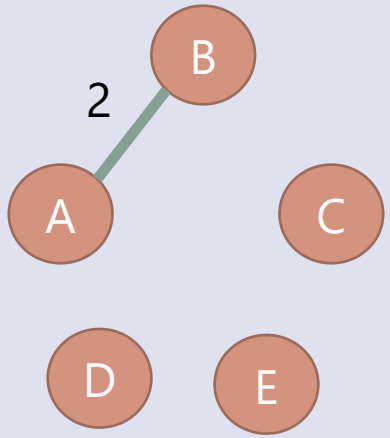
LEVEL = 1 (SLV)

goeBURST algorithm

Profile	A	B	C	D	E
A	-	2	7	7	6
B	-	-	7	7	6
C	-	-	-	5	5
D	-	-	-	-	3
E	-	-	-	-	-



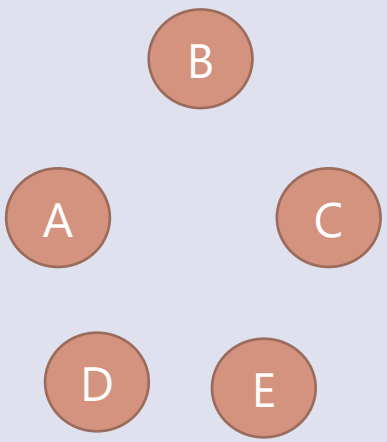
LEVEL = 1 (SLV)



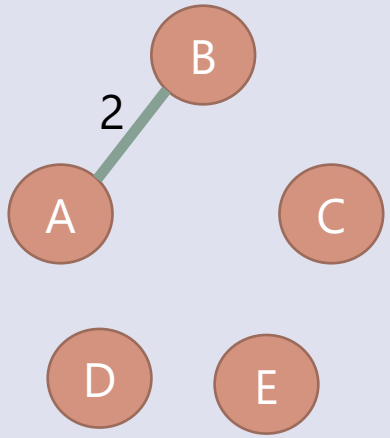
LEVEL = 2 (DLV)

goeBURST algorithm

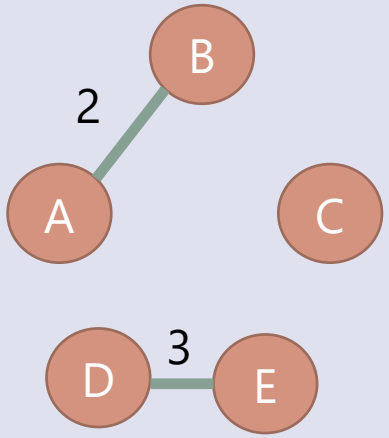
Profile	A	B	C	D	E
A	-	2	7	7	6
B	-	-	7	7	6
C	-	-	-	5	5
D	-	-	-	-	3
E	-	-	-	-	-



LEVEL = 1 (SLV)



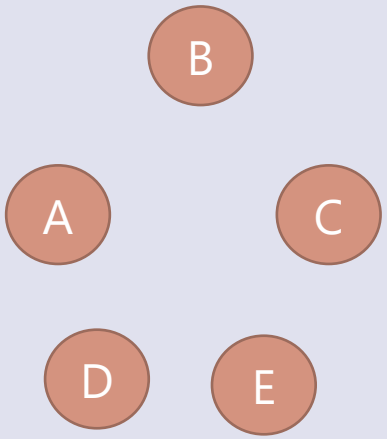
LEVEL = 2 (DLV)



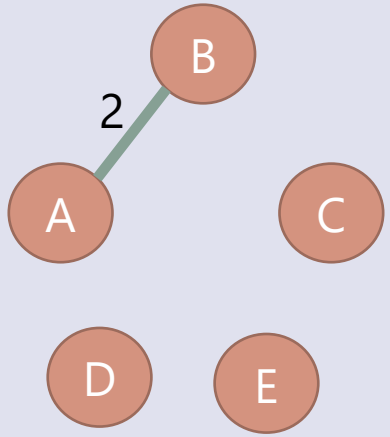
LEVEL = 3 (TLV)

goeBURST algorithm

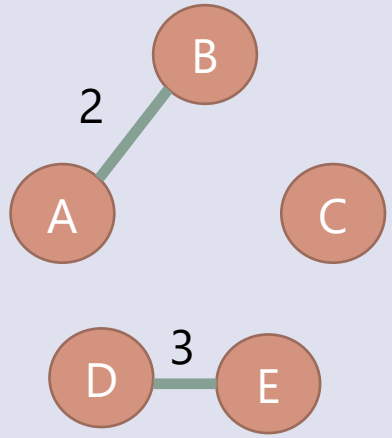
Profile	A	B	C	D	E
A	-	2	7	7	6
B	-	-	7	7	6
C	-	-	-	5	5
D	-	-	-	-	3
E	-	-	-	-	-



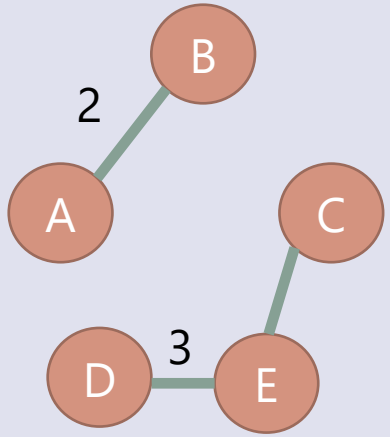
LEVEL = 1 (SLV)



LEVEL = 2 (DLV)



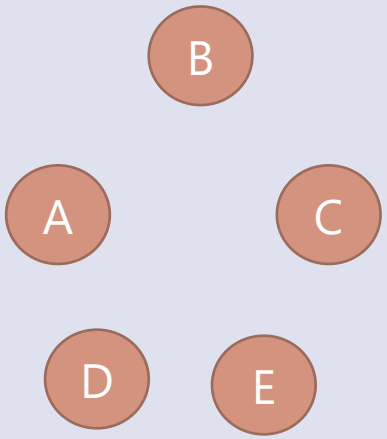
LEVEL = 3 (TLV)



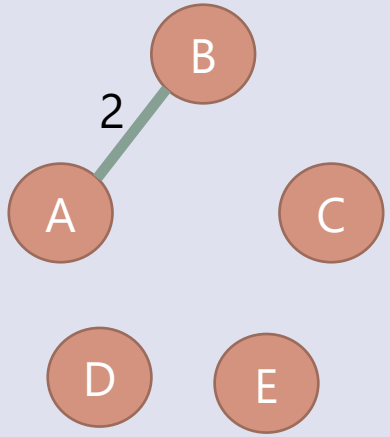
LEVEL = 5,
assuming #E>#D

goeBURST algorithm

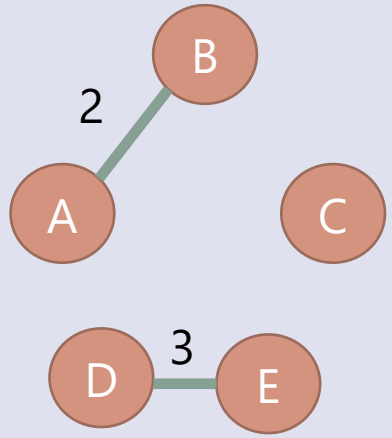
Profile	A	B	C	D	E
A	-	2	7	7	6
B	-	-	7	7	6
C	-	-	-	5	5
D	-	-	-	-	3
E	-	-	-	-	-



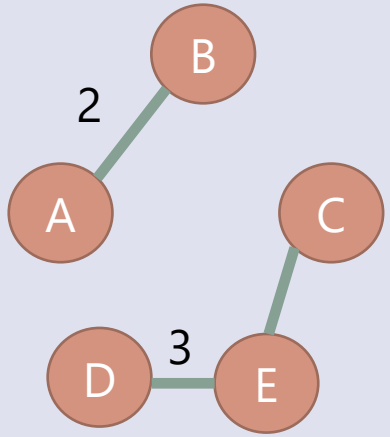
LEVEL = 1 (SLV)



LEVEL = 2 (DLV)



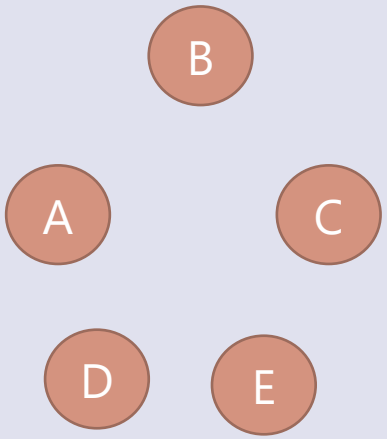
LEVEL = 3 (TLV)



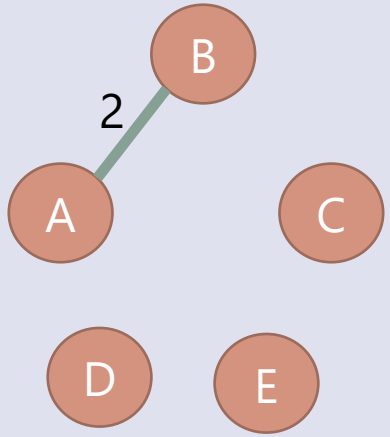
LEVEL = 5,
assuming #E>#D

goeBURST algorithm

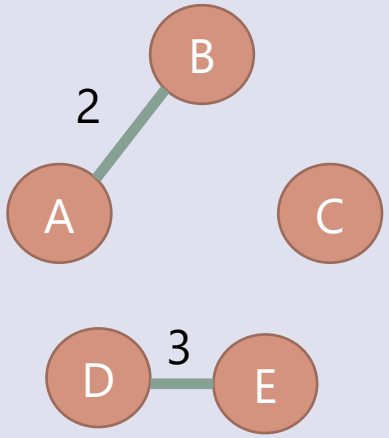
Profile	A	B	C	D	E
A	-	2	7	7	6
B	-	-	7	7	6
C	-	-	-	5	5
D	-	-	-	-	3
E	-	-	-	-	-



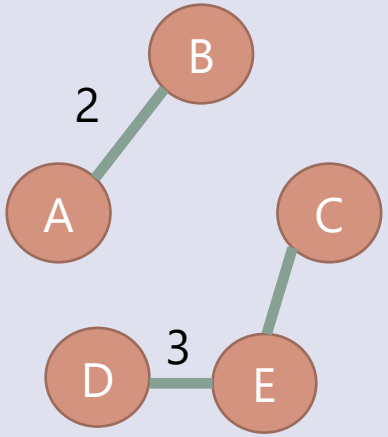
LEVEL = 1 (SLV)



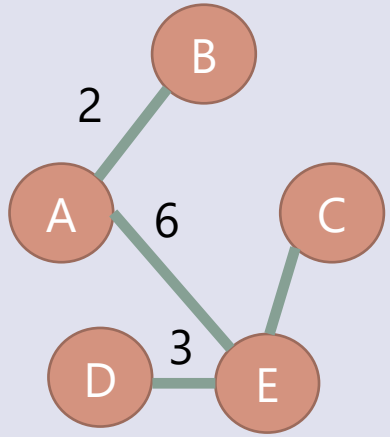
LEVEL = 2 (DLV)



LEVEL = 3 (TLV)



LEVEL = 5, assuming #E>#D



LEVEL = 6, assuming ID tiebreak

Local Branch Recrafting

The phylogenetic tree does not necessarily represent true phylogenetic relationships between sequences

- allelic distances do not always correlate with divergence time.

To correct, we may apply a ***Local Branch Recrafting (LBR)***

Input: A phylogenetic tree **T** over a set of elements **S**.

Initialization: Initialize the set **E** with the edges of the tree **T.**

Loop: While |E| > 0 do:

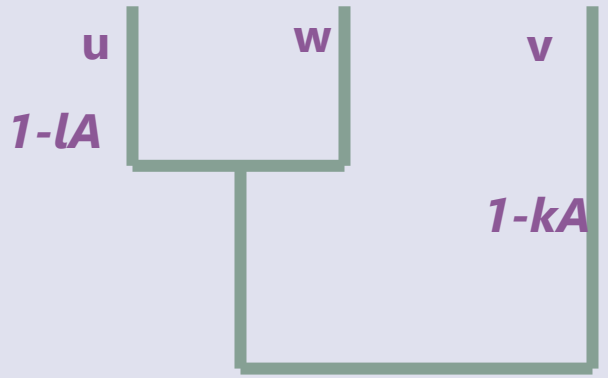
1. **Selection:** Select an edge **(u → v)** of the set **E** and remove it from the tree **T**, dividing it into two sub-trees **Tu** (containing u) and **Tv** (containing v).
2. **Joining:** Find two vertices **w** and **z** that best connect the two sub-trees by an edge **(w → z).**
3. **Reduction:** Remove the edge **(u → v)** from E and add the edge **(w → z)** to

1. **Finalization:** Return the tree **T**.

Local Branch Recrafting – Best connection?

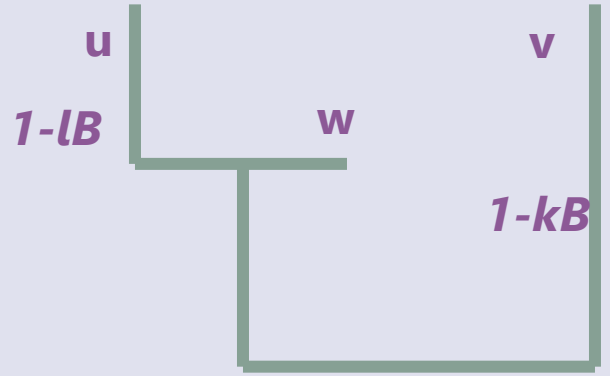
1. **Joining:** For each edge **(u → v)** Find two vertices **w** and **z** that **best connect** the two sub-trees **Tu** and **Tv** by an edge **(w → z).**

Contemporary model (MA): nodes *u* and *w* are contemporary sisters that diverged from a hypothetical common ancestor.



lAxlA-> invariant alleles proportion/probability between u and w;
kAxkA -> invariant alleles proportion between (u,w) and v

Ancestor-Descendent model(MB): node *w* is the direct ancestor of *u*.



lBxlB -> invariant alleles proportion between u and w;
kBxKB -> invariant alleles proportion between (u,w) and v

Local Branch Recrafting – Best connection?

- **1-lAxlA -> variant** alleles proportion/probability between u and w;
- **1-lAxKA** -> **variant** alleles proportion/probability between (u,w) and v;

Best connection?

- **Step 1:** Calculate the likelihoods of a contemporary model versus an ancestor-descendent model

Given L, the set of distint alelles

$$\underset{0\leq lA\leq 1,0\leq kA\leq 1}{\operatorname{argmax}} \log P(MA \mid lA, kA) = \underset{0\leq lA\leq 1,0\leq kA\leq 1x}{\operatorname{argmax}} \log P(u \rightarrow w \mid lA)\, P(u \rightarrow v \mid lA, kA) P(w \rightarrow v \mid lA, kA)$$

Local Branch Recrafting – Best connection?

- **(1-lA)xlA -> variant** alleles probability between u and w;
- **(1-lA)xkA** -> **variant** alleles propability between (u,w) and v;

- L, the set of distint alelles

log (a*b)=log(a)+log(b)

- Distances are normalized

$$\underset{0\le lA\le 1,0\le kA\le 1}{\operatorname{argmax}} \log P(MA \mid lA, kA) = \underset{0\le lA\le 1,0\le kA\le 1x}{\operatorname{argmax}} \log P(u \to w \mid lA)\, P(u \to v \mid lA, kA) P(w \to v \mid lA, kA)$$

$$= \underset{0\le lA\le 1,0\le kA\le 1x}{\operatorname{argmax.}} \log P(u \to w \mid lA) + \log P(u \to v \mid lA, kA) + \log P(w \to v \mid lA, kA)$$

The likelihood of a branch is:
(branch length) (number of different alleles) * (1-branch length) (number of identical alleles)

Thus, for instance, branch u → w

$|L|d(u,v) \log(1-lA^2) + |L|(1-d(u,v)) \log(1-lA^2)$

Local Branch Recrafting – Best connection?

$$\operatorname*{argmax}_{0\le lA\le 1,0\le kA\le 1} \log P(MA \mid lA, kA) = \dots$$

$$= \operatorname*{argmax}_{0\le lA\le 1,0\le kA\le 1} \log P(u \to w \mid lA) + \log P(u \to v \mid lA, kA) + \log P(w \to v \mid lA, kA)$$

$$= \operatorname*{argmax}_{0\le lA\le 1,0\le kA\le 1} \begin{aligned} &|L| d(u \to w) \log (1 - lA^2) + |L|(1 - d(u \to w)) \log (lA^2) + \\ &|L| d(u \to v) \log (1 - lAkA) + |L|(1 - d(u \to v)) \log (lAkA) + \\ &|L| d(w \to v) \log (1 - lAkA) + |L|(1 - d(w \to v)) \log (lAkA) \end{aligned}$$

Similarly, in Model B

$$\operatorname*{argmax}_{0\le lB\le 1,0\le kB\le 1} \log P(MB \mid lB, kB) = \operatorname*{argmax}_{0\le lB\le 1,0\le kB\le 1} \log P(w \to u \mid lB) P(u \to v \mid lB, kB) P(w \to v \mid lB, kB)$$

$$= \operatorname*{argmax}_{0\le lB\le 1,0\le kB\le 1} \begin{aligned} &|L| d(w \to u) \log (1 - lB) + |L|(1 - d(w \to u)) \log (lB) + \\ &|L| d(u \to v) \log (1 - lBkB) + |L|(1 - d(u \to v)) \log (lBkB) + \\ &|L| d(w \to v) \log (1 - kB) + |L|(1 - d(w \to v)) \log (kB) \end{aligned}$$

Local Branch Recrafting – Best connection?

The solution to parameters is:

$$lA = \sqrt{1 - d(u \rightarrow w)}$$

$$kA = \frac{1 - (1/2)(d(u \rightarrow v) + d(w \rightarrow v))}{lA}$$

$$lB = 1 + \frac{x d(w \rightarrow u)}{d(u \rightarrow v) - 2x}$$

$$kB = 1 + \frac{x\, d(w \rightarrow v)}{d(u \rightarrow v) - 2x}$$

with

$$x = 1 - \frac{(1 - d(w \rightarrow u))(1 - d(w \rightarrow v)) + (1 - d(u \rightarrow v))}{2}$$

Best connection?

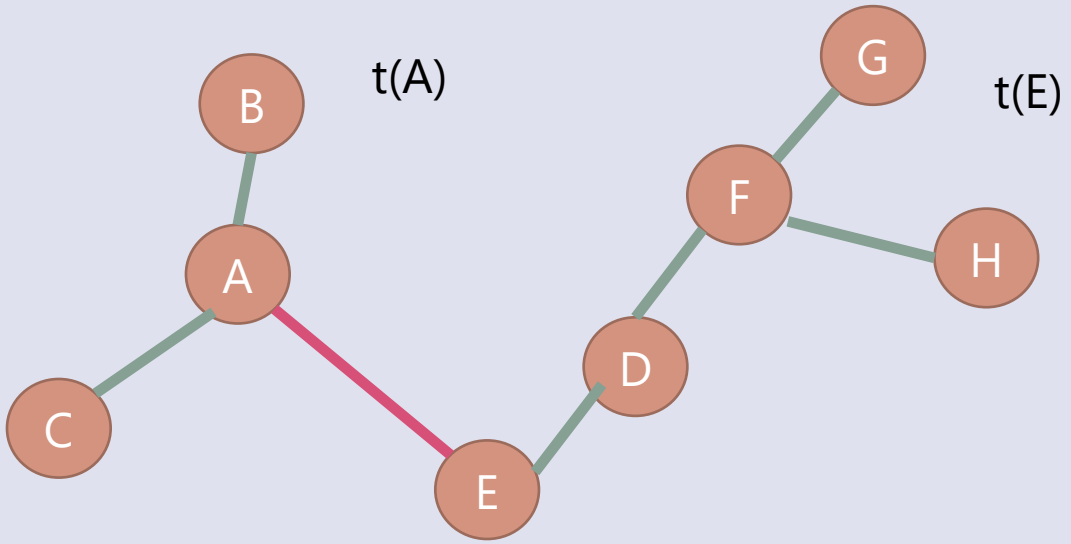
- **Step 1:** Calculate the likelihoods of a contemporary model versus an ancestor-descendent model
- **Step 2: the** joining criterion consists of finding two nodes that have the minimum harmonic distance if the contemporary model has a higher or equal likelihood to the ancestor-descendent model

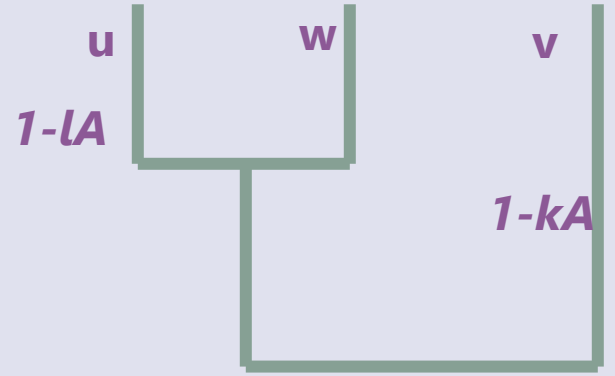
Best connection

Input: Initial edge **(u → v)** , Tree **t(u)∈F**, harmonic tiebreaker **ht**
Output: New edge **(u′ → v′)**
1: Initialize **u′ = u**
2: for each node **w ∈ TargetNodes(t(u′))** do
    3: P(MA), P(MB) = **ModelSelection(**d(u′ → w), d(w → u′), d(u′ → v), d(w → v)**)**
    4: if **P(MA) ≥ P(MB)** and **ht(u′) > ht(w)** then
        5: u′ = w
    6: else if **P(MA)<P(MB)** and **d(u′ → v) >= d(w → v)** then
        7: u′ = w
 8: Repeat **(1–7)** on **v** to obtain **v′**
9: Return **(u′ → v′)**

harmonic mean gives less weight to vertices that are close to the vertex of interest than to those far away,

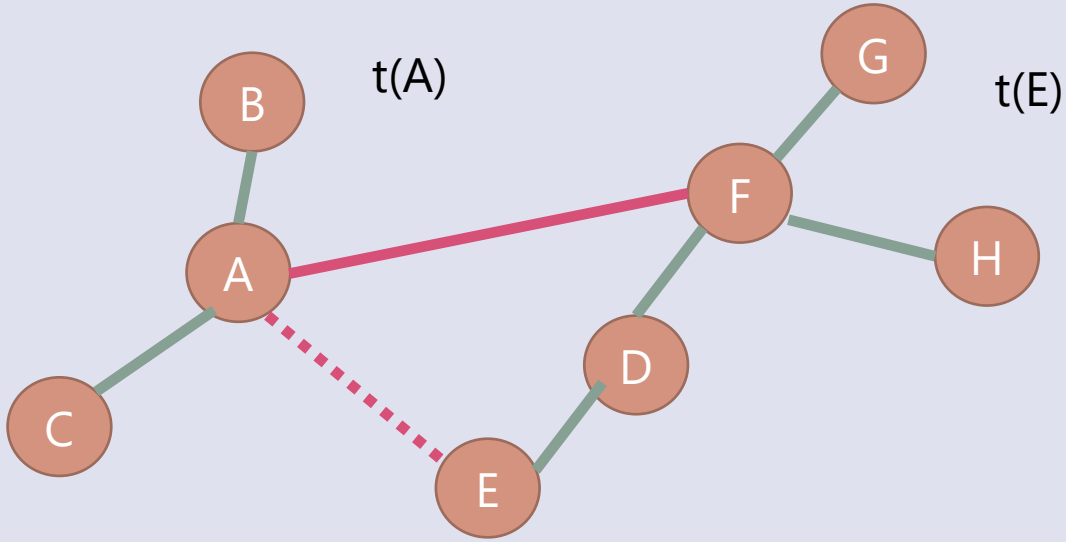
Example

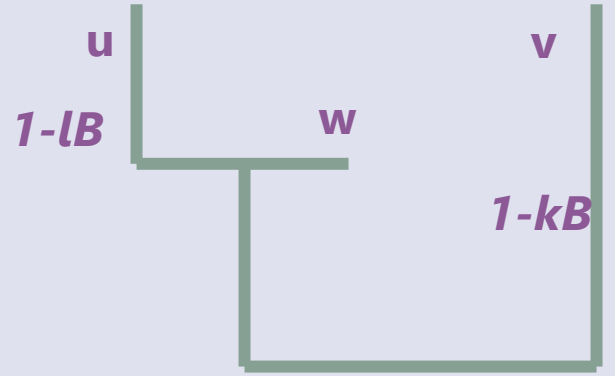




Model A

(A→E) is compared with branch (A→F), where node F has the lowest harmonic average distance to other nodes and Model A has higher probability

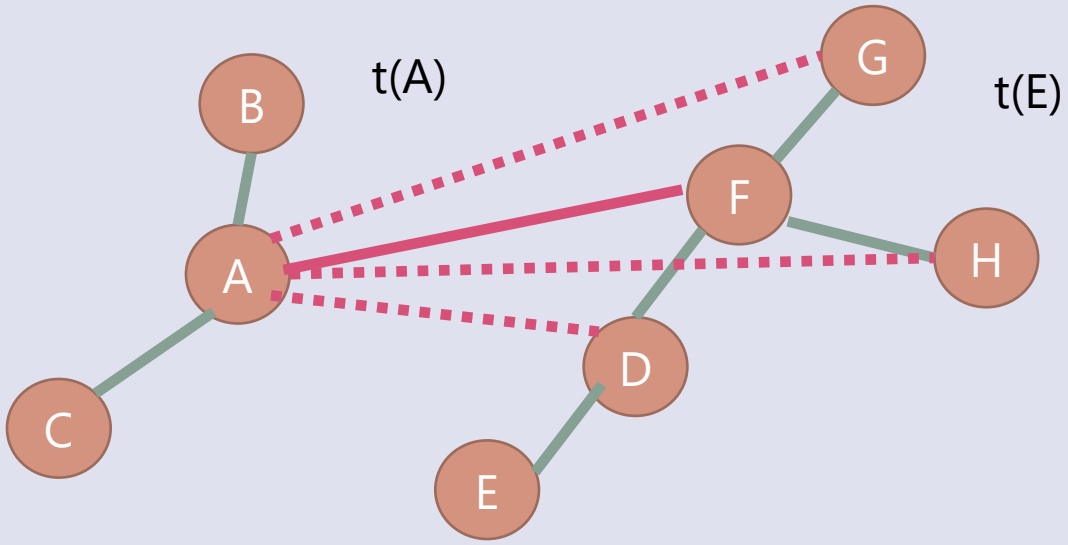


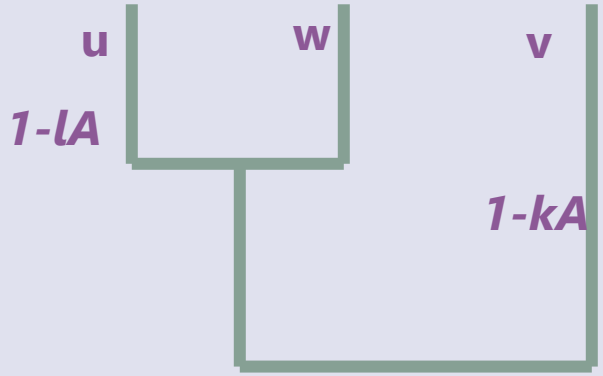


Model B

Example

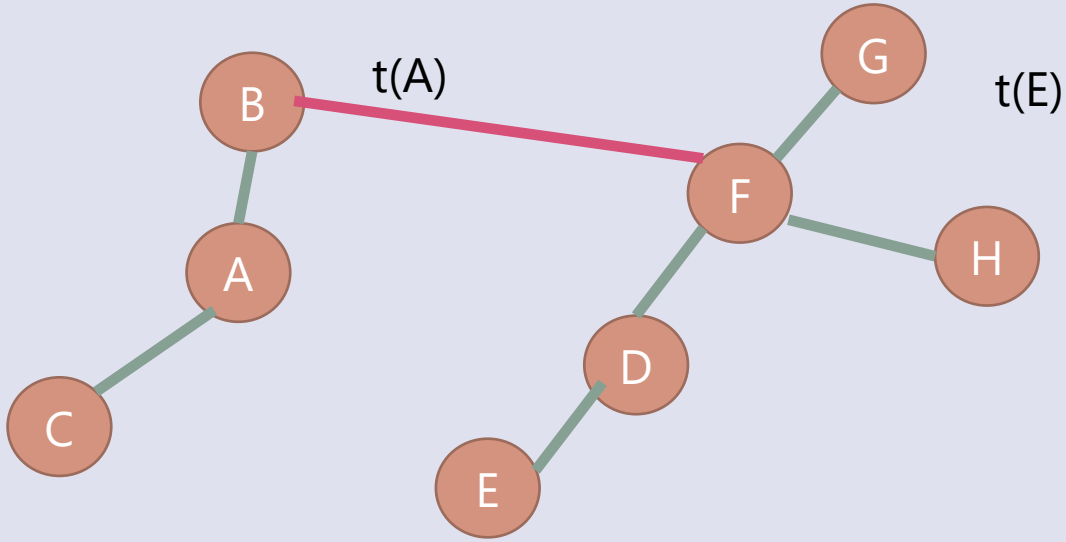
(A→F) is compared with all the nodes that are directly connected with F, but (A→F) is still the most probable branch

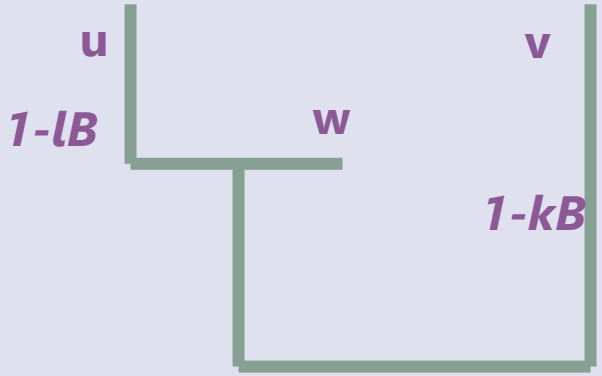




Model A

The same process is performed for tree t(A), which results in (B→F) becoming the most probable branch.





Model B

Final Remarks

- The proposed local optimization was implemented (in Java) in a generic way such that it can be applied to the output of goeBURST algorithm, or to the output of any other algorithm that produces a phylogenetic tree.
- In the case of goeBURST algorithm, we were able to incorporate this optimization in the comparison criteria, and it does not affect the running time of the algorithm.
- This method was tested with data from EnteroBase, and results are promising with trees being more meaningful from the biological point of view.

Acknowledgments

1. This work was supported by:
    - A research project from Polytechnic Institute of Lisbon - IPL/2021/DIVA ISEL
    - A research project from Fundação para a Ciência e Tecnologia(FCT) with reference NGPHYLO PTDC/CCI-BIO/29676/2017.