Single or multiple consensus trees a method to separate divergent genes

Alain Guénoche

CNRS, Institut de Mathématiques de Luminy
guenoche@iml.univ-mrs.fr

SeqBio 2012

Motivations

Some strains in bacteria are very dangerous (*E. Coli)*

Why ?

Because they contain abnormal genes ?

Methodology

- Compare genes in all the strains
- Establishing their own phylogeny
- Comparing the tree topologies

Pre-requisite

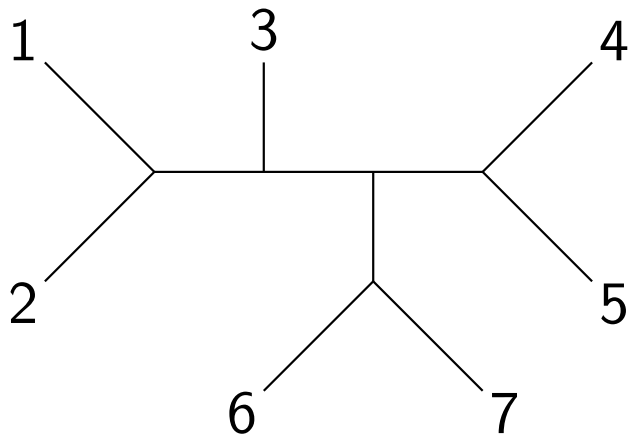
A X-tree is :

- an unrooted tree,
- X is the set of n leaves,
- nodes have degre 3,
- edges have a positive or null length.

X-tree $\Longrightarrow$ { bipartitions }

- external edges (to leaves) common to every X-tree
- internal edges (at most $n-3$) only considered

An X-tree



Bipartition set :

- 1 2 | 3 4 5 6 7
- 1 2 3 | 4 5 6 7
- 1 2 3 6 7 | 4 5
- 1 2 3 4 5 | 6 7

Consensus Tree

$\Pi = \{T_1, \ldots, T_m\}$ a *profile* of m X-trees

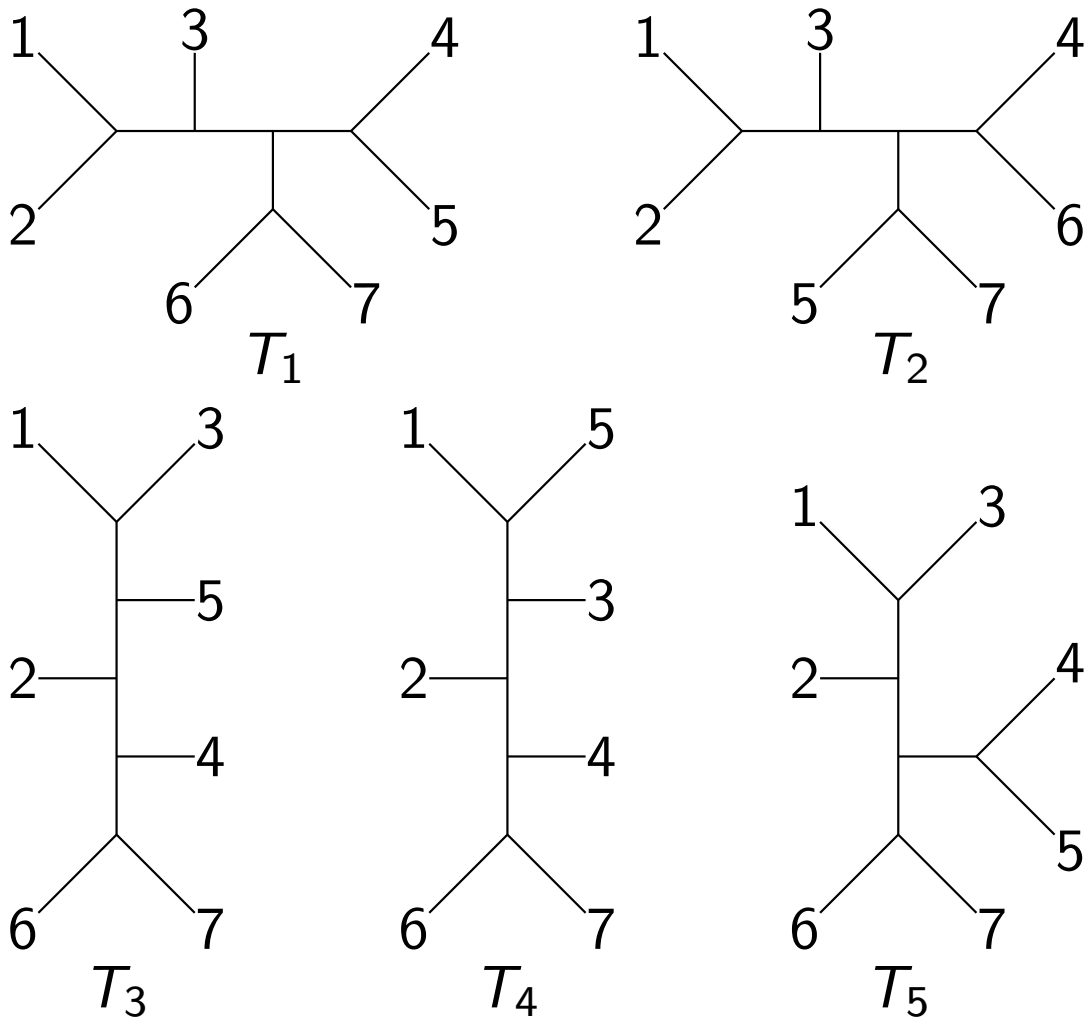
A consensus tree C is a X-tree *summarizing* Π
Several rules :

- strict : (only edges common to all the trees),
- majority : (edges belonging to a majority of trees),
- extended majority : (majority edges + compatible edges)
- Nelson : (clique of compatible edges with max weight)

Two bipartitions $X_1|X_2$ et $Y_1|Y_2$ are compatible in a X-tree iff

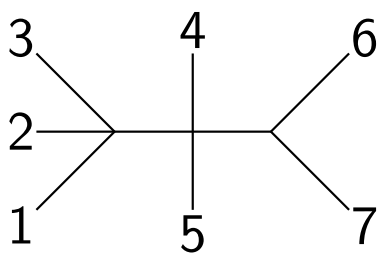
$$\emptyset \in \{X_1 \cap Y_1, X_2 \cap Y_1, X_1 \cap Y_2, X_2 \cap Y_2\}$$

Example

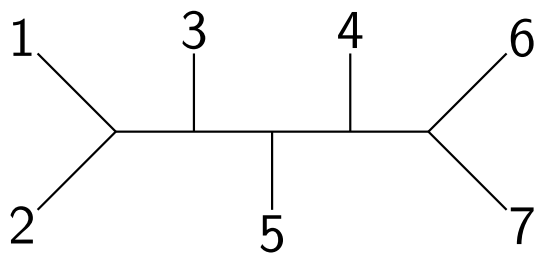


Majority and extended majority consensus

	T_1	T_2	T_3	T_4	T_5	bipartitions
1	x	x				1 2 | 3 4 5 6 7
2	x	x			x	1 2 3 | 4 5 6 7
3	x		x	x	x	1 2 3 4 5 | 6 7
4	x				x	1 2 3 6 7 | 4 5
5		x				1 2 3 4 6 | 5 7
6		x				1 2 3 5 7 | 4 6
7			x		x	1 3 | 2 4 5 6 7
8			x	x		1 3 5 | 2 4 6 7
9			x	x		1 2 3 5 | 4 6 7
10				x		1 5 | 2 3 4 6 7



C



C_E

Which consensus ?

The majority consensus is the only valid

- Computable in $O(nm)$
- Majority consensus tree C is median for the *Robinson-Foulds* distance

$$\sum_{i=1}^{m} D_{R-F}(C, T_i) \text{ minimum}$$

- the minority edges are not significant in evolution
- The Nelson consensus is NP-hard (and may contain minority edges)

The consensus tree weight

$\{P_1, \ldots P_q\}$ *majority* bipartitions

- edge weight = nb. of trees containing this edge

$$w(P_k) = |\{T_i \ \ containing \ \ P_k\}|$$

- Consensus tree weight = sum of internal edge weight

$$W(C) = \sum_{P_k \in C} w(P_k)$$

On the 5 trees in l'Example:

$$W(C) = 3 + 4 = 7$$

Unique or multiple consensus tree ?

Let

- $P_\Pi = \{\Pi_1, \ldots, \Pi_k\}$ a partition of Π in k classes,
- $\{m_1, \ldots, m_k\}$ nb. of elements
- $\{C_1, \ldots, C_k\}$ the consensus trees of sub-profiles

The generalized score of P_Π, denoted $\mathcal{W}^k(P_\Pi)$ is the sum of consensus tree weight of a class multiplied by its nb. of elements :

$$\mathcal{W}^k(P_\Pi) = \sum_{i=1}^{k} m_i \times W_{\Pi_i}(C_i).$$

m_i trees support C_i with a high or low weight

Problem

To find a partition of Π maximizing the generalized score

$$\max_{P_\Pi=\{\Pi_1,\ldots,\Pi_k\}\in\mathcal{P}(\Pi)} \mathcal{W}^k(P_\Pi)$$

Double optimization, over the nb. of classes (k) and over the set of partitions in k classes

Extreme values : $\mathcal{W}^1(P_1)$ and $\mathcal{W}^n(P_n)$

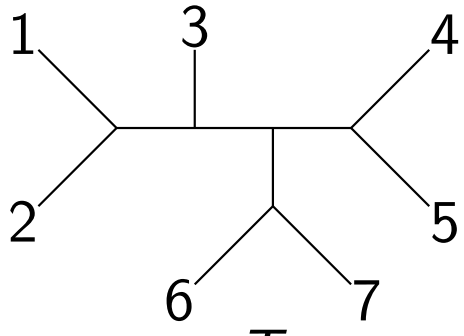
1 classe — atomic partition

Proposition

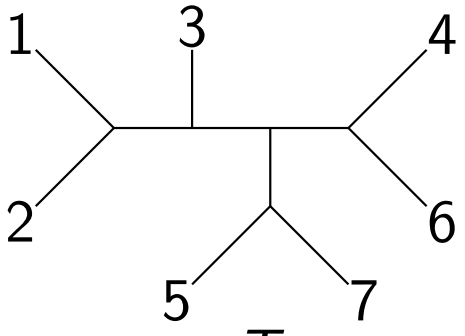
Two X-arbres make a single consensus iff the share more than half internal edges ($|C|$)

$$\mathcal{W}^1 = 2 \times |C| > \mathcal{W}^2 = n_1 + n_2$$

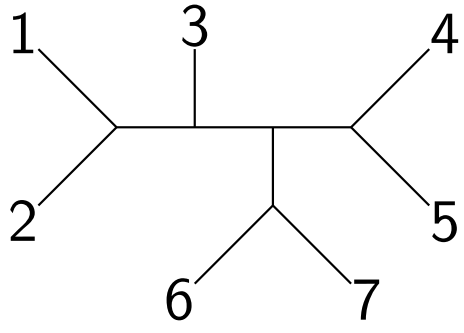
Consensus C_1



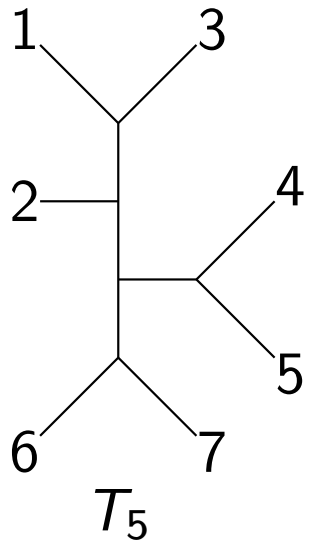
T_1



T_2

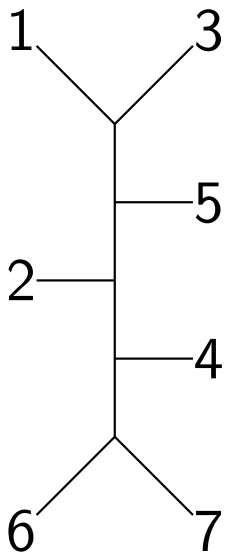


$W(C_1) = 9$

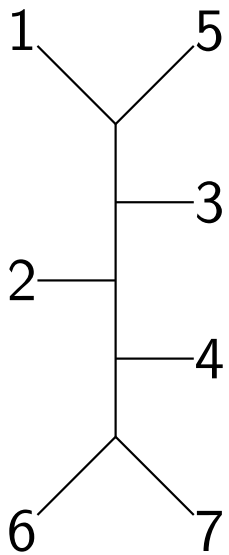


T_5

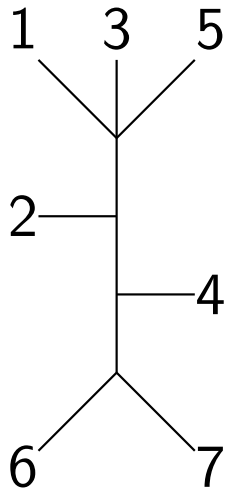
Consensus C_2



T_3



T_4



$W(C_2) = 6$

There is multiple consensus

Homogeneous Profile

$\Rightarrow$ Single consensus

$$\mathcal{W}^1 = m \times W_\Pi(C)$$

Each tree (n_i internal edges) is its own consensus

$\Rightarrow$ Atomic consensus

$$\mathcal{W}^m = \sum_{i=1}^{m} n_i \leq m \times (n-3)$$

But :

$$\mathcal{W}^1 = 5 \times 7 = 35 > \mathcal{W}^5 = 5 \times 4 = 20$$

$$\mathcal{W}^2 = 3 \times 9 + 2 \times 6 = 39$$

Method 1

Similarity indices on X-trees

- *Robinson-Foulds* similarity

$$S(T_i, T_j) = \frac{2 \times |\{a \in T_i \cap T_j\}|}{|T_i| + |T_j|}.$$

- quadruple similarity $|\{x, y, z, t\}|$
  $+1$ if identical topologies; $+1/2$ only one resolved topology

Average Linkage Hierarchy

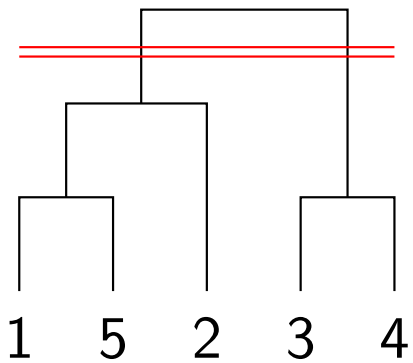
- Hierarchy of partitions (from P_0)
- Consensus tree of the new class
- Generalized score value

Example

From profile Π in Example 1

S	T_1	T_2	T_3	T_4
T_2	4			
T_3	2	0		
T_4	2	0	6	
T_5	6	2	4	2

Robinson-Foulds similarity



Dendrogram

$\mathcal{W}^5 = 20$, $\mathcal{W}^4 = 24$, $\mathcal{W}^3 = 28$, $\mathcal{W}^2 = 39$ and $\mathcal{W}^1 = 35$

Method 2

- Join the 2 classes maximizing the generalized score
- Consensus tree of this new class

	T_1	T_2	T_3	T_4
T_2	20			
T_3	16	12		
T_4	16	12	24	
T_5	24	16	20	16

	$T_{1,5}$	T_2	T_3
T_2	35		
T_3	32	16	
T_4	26	16	28

	$T_{1,2,5}$	T_3
T_3	28	
T_4	28	39

Nb. of common majority edges

Validation on random trees

Two tests :

- Random topologies $\rightarrow$ Atomic consensus

$$\mathcal{W}^1 = 0 \quad \text{and} \quad \mathcal{W}^n \text{ Maximum}$$

- 3 random topologies $\rightarrow$ 15 noisy trees (swapping leaves)

$$\mathcal{W}^3 \text{ Maximum}$$

- 30 trees from one random rooted topology $|T| = 16$
  - one 1000 bp random sequence evolving along the tree
  - substitution rate from root to leaves : 0.25
  - 16 aligned sequences
  - Kimura distance (K_{2p}) + NJ $\rightarrow$ T_k

$$\mathcal{W}^1 \text{ Maximum}$$

Validation on homogeneous trees

BROWN, J.R., DOUADY, C.J., ITALIA, M.J., MARSHALL, W.E., STANHOPE, M.J. (2001) Universal trees based on large combined protein sequence data sets. *Nat Genet*, 28, 281–285.

Here we use large combined alignments of 23 orthologous proteins conserved across 45 species from all domains to construct highly robust universal trees. Although individual protein trees are variable in their support of domain integrity, trees based on combined protein data sets strongly support separate monophyletic domains ... (after) elimination of 9 proteins, which were likely candidates for horizontal gene transfer.

	BiP	Maj	$W(C)$	$\mathcal{W}^1$	$\mathcal{W}^2$	$\mathcal{W}^{23}$
	333	23	430	9890	8673	964
Theoretical max	966	42	529			

There is a single consensus

Validation on bootstrap trees

SCHUBERT, S., DARLU, P., CLERMONT, O. et al. (2009), Role of intraspecies recombination in the spread of pathogenicity islands within the *Escherichia coli* species, *PLoSpathogens*, (5(1)e1000257).

9 genes in 30 *Escherichia coli* strains
500 bootstrap trees per gene

	BiP	*Maj*	*W*	$\mathcal{W}^1$	*NbClas*	$\mathcal{W}_{next}$
UR	8	7	2623	1311500	2(2)	1304768
trpB	28	15	6248	3124000	2(1)	3114271
trpA	45	9	3824	1912000	3(1,1)	1900390
putP	57	17	6608	3304000	2(80)	2508400
polB	119	14	5331	2665500	2(3)	2639187
icd	69	15	5681	2840500	2(4)	2929008
HPI	76	13	4971	2485500	2(2)	2467626
pabB	57	8	3667	1833500	2(1)	1827846
DR	12	8	2685	1342500	2(2)	1335146

Validation on divergent trees ; previous method

DARLU, P. and GUENOCHE, A. (2011), The *TreeOfTrees* method to evaluate the congruence between gene trees, *J. of Classification*, 28(3), 390-403

Input : A set of aligned gene sequences or a set of boostrapped genes trees

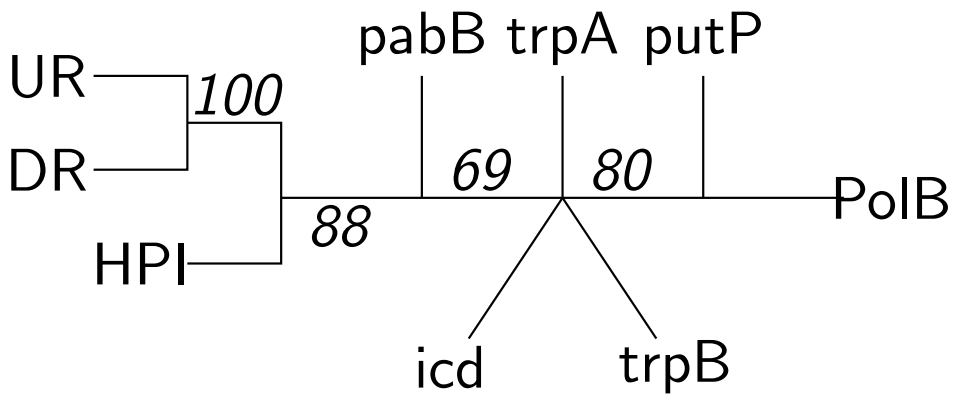
	G_1	G_2	...	G_m	X-Tree comparison	Distance on genes	NJ	Gene-Tree
bo	T_1^1	T_2^1	...	T_m^1	$\rightarrow$	Δ_1	$\rightarrow$	T^1
os	T_1^2	T_2^2	...	T_m^2	$\rightarrow$	Δ_2	$\rightarrow$	T^2
tr					...			
ap	T_1^{100}	T_2^{100}	...	T_m^{100}	$\rightarrow$	Δ_{100}	$\rightarrow$	T^{100}
								$\mathcal{T}$

Output : $\mathcal{T}$ the consensus tree of gene trees

- with robustness values (on the internal edges)
- which could separate groups of genes (but not a isolated gene)

The *TreeOfTrees* tree

- 6 housekeeping genes (*icd*, *pabB*, *polB*, *putP*, *trpA*, *trpB*),
- 3 other genes, *HPI*, *DR* and *UR*, (Hight Pathogenicity Island and its Downstream and Upstream regions)
  Highly suspected to come from LGT



Validation on divergent trees : the consensus method

The 9 consensus trees on E. coli make profile Π
Similarity

- Robinson-Foulds
- Quadruple

NbClas	1	2	3	4	5	6	7	8	9
R-F	144	150	174	147	154	139	120	130	140
Quad	144	150	135	159	169	136	146	129	140
Greedy	144	168	182	147	160	145	155	130	140

Best generalized scores for all the number of classes

$$\mathcal{W}(\{HPI, UR, DR\}, \{pabB, trpA, trpB, icdetPolB\}, \{putP\}) = 182$$

Conclusion

- An efficient, simple method
- to decide if there is an atomic consensus or not ($\mathcal{W}^m$ maximum)
- to define a single or multiple consensus
- to detect divergent genes.
- Optimality is not sure, but ...

$$\mathcal{W}^k(P) > \mathcal{W}^1 \Rightarrow \Pi \text{ non homogeneous}$$