# Algorithmes bioinformatiques pour la reconstruction d'arbres consensus et de super-arbres multiples

07 Mai 2015

**Nadia Tahiri**
Université du Québec à Montréal

UQÀM

Fonds de recherche sur la nature et les technologies Québec

I. Introduction
   1. Phylogénie
   2. Mesures de comparaison des arbres

II. Projet 1 : Classification d'arbres phylogénétiques : Consensus
   1. Problématique
   2. Algorithme
   3. Résultats préliminaires de simulation

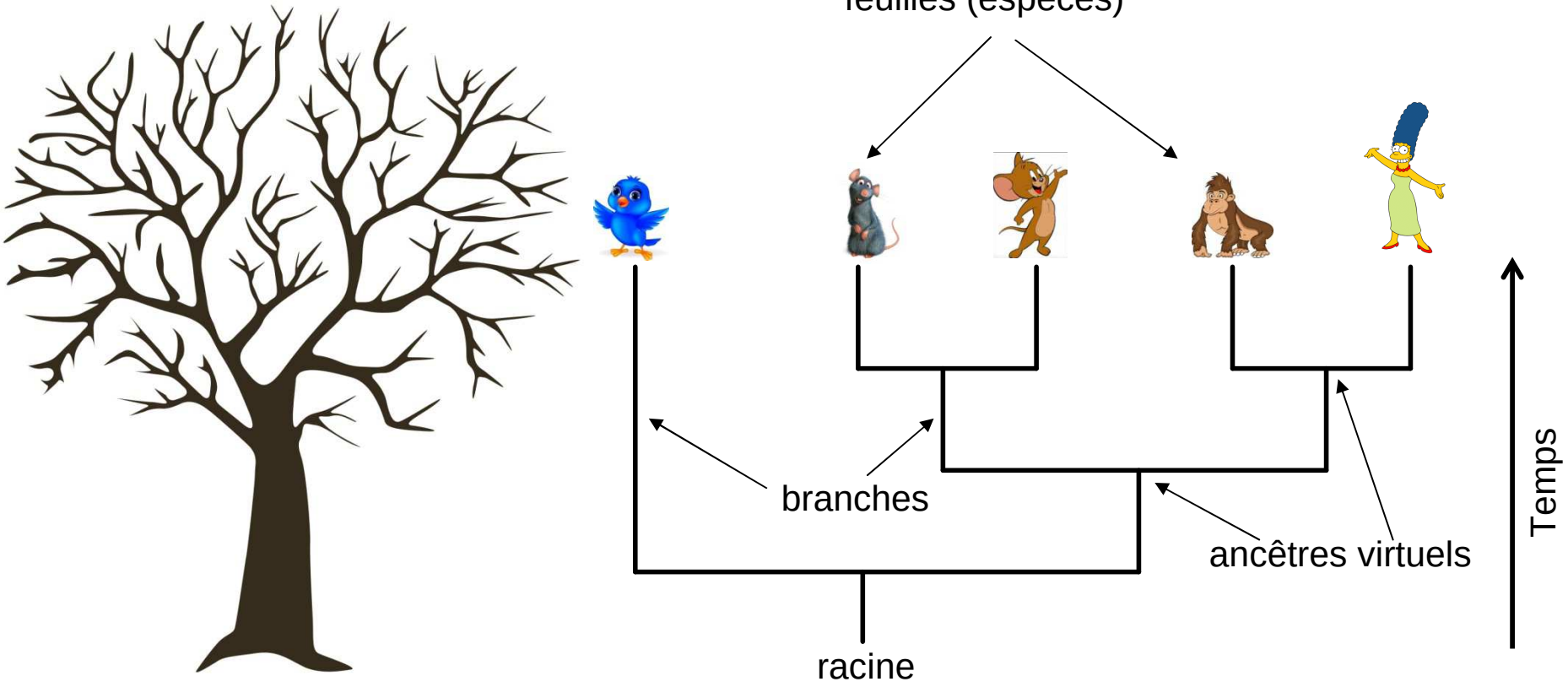
III. Projet 2 : Classification d'arbres phylogénétiques : Super-arbres
   1. Problématique
   2. Algorithme

IV. Projet 3 : Applications
   1. Données biologiques
   2. Données biolinguistiques

# Introduction

# LA PHYLOGÉNIE

**La phylogénie (ou phylogénèse) étudie la parenté entre différents êtres vivants en vue de comprendre leur évolution.**



# RECONSTRUCTION D'UN ARBRE PHYLOGÉNÉTIQUE

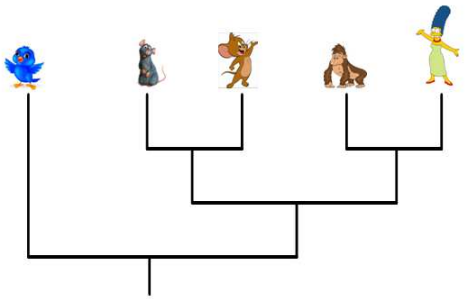
**alignement des séquences**

```
AAATGATCTGCGTCAATATTATAA
GCCTGATCCTCACTACTGTCATCTTAA
ATAGGGCCCGTATTTACCCTATAG
AACTGGTCCACCCTTATACTAAAAGACGCCTCACTAGGAAGCTAA
AACTGATCTGCTTCAATAATTTAA
```

```
AAATGATCTGCGTCAATATTA----------TAA
GCCTGATCCTCACTA----------CTGTCATCTTAA
ATA----------GGGCCCGTATTTACCCTATAG
AACTGGTCCACCCTTATACTAAAAGACGCCTCACTAGGAAGCTAA
AACTGATCTGCTTCAATAATT----------TAA
```

**calcule des distances ou des dissimilarités entre les espèces**

| | | | | | |
|---|---|---|---|---|---|
| | | | | | |
| | 0 | 4 | 2 | 4 | 4 |
| | 4 | 0 | 4 | 4 | 2 |
| | 2 | 4 | 0 | 4 | 4 |
| | 4 | 4 | 4 | 0 | 4 |
| | 4 | 2 | 4 | 4 | 0 |



**application d'une méthode de reconstruction d'arbres**

# MESURES DE COMPARAISON DES ARBRES

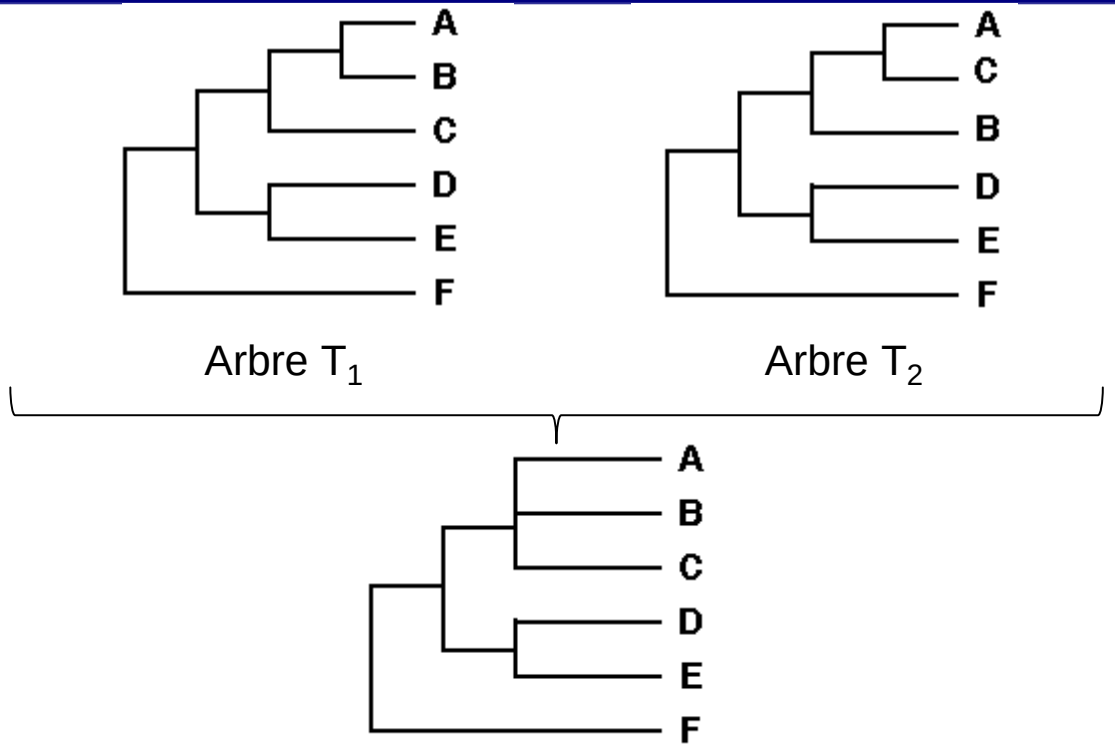
**Il existe quatre principales mesures de comparaison d'arbres phylogénétiques:**

- La distance des moindres carrés (LS) (Gauss, 1795);
- La dissimilarité de bipartitions (DB) (Boc *et al.*, 2010, Makarenkov *et al.*, 2007);
- La distance de quartets (QD) (Bryant *et al.*, 2000);
- La distance de Robinson et Foulds (RF) (Robinson et Foulds, 1981).

# Projet 1 : Classification d'arbres phylogénétiques : Consensus

Nadia Tahiri, Matthieu Willems, Vladimir Makarenkov (2014) Classification d'arbres phylogénétiques basée sur l'algorithme des *k*-moyennes, article publié dans les actes de la conférence SFC-2014.

# ALGORITHMES D'INFÉRENCE D'ARBRES CONSENSUS




Arbre consensus (strict et majoritaire) de $T_1$ et $T_2$

Les trois principales méthodes pour l'inférence d'arbres consensus:

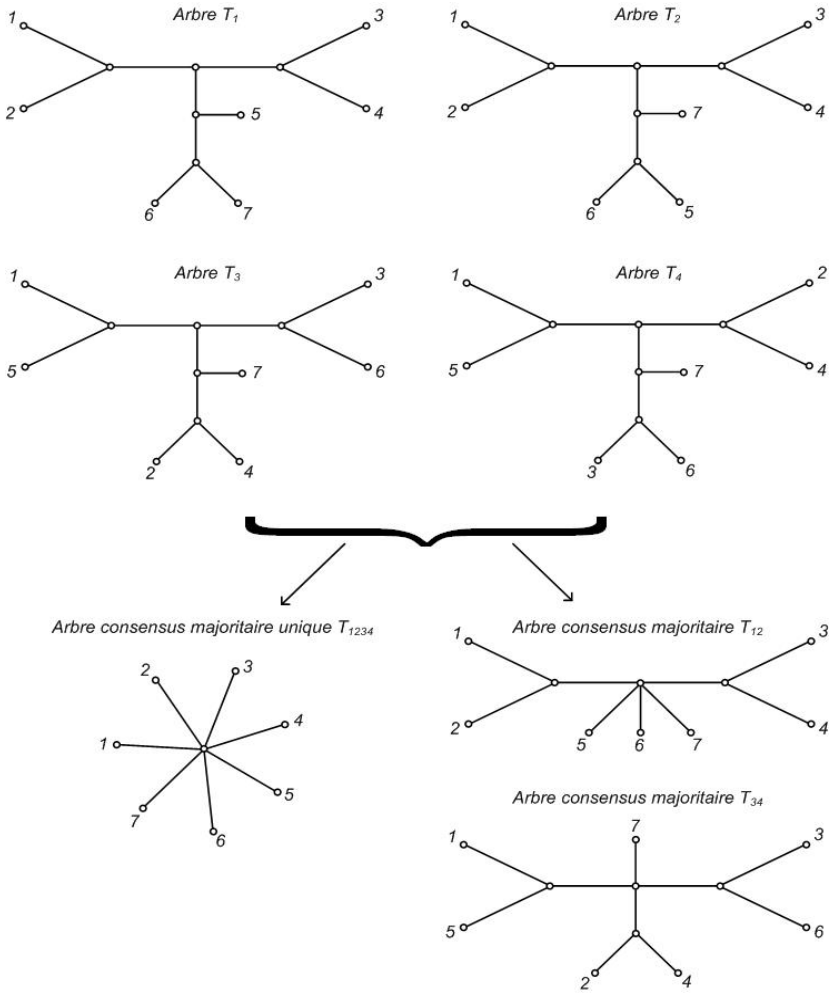
- Arbre consensus strict (Sokal et Rohlf, 1981)
- Arbre consensus majoritaire (Margush et McMorris, 1981)
- Arbre consensus majoritaire étendu (Felsenstein, 1985)

# PROBLÉMATIQUE

**<u>Idée</u>: La classification d'arbres phylogénétiques basée sur l'algorithme des *k*-moyennes permet de distinguer les familles de gènes qui ont la même histoire évolutive (*e.g.* gènes orthologues)**

- Nécessité de **<u>fusionner</u>** les arbres phylogénétiques via le projet ToL (Tree of Life) [1] (Maddison *et al.*, 2007)
- **<u>Perte</u>** d'informations
- **<u>Incohérence</u>** de la fusion des arbres phylogénétiques

**Nous proposons ici une méthode de partitionnement d'un ensemble de *n* arbres phylogénétiques qui se base sur l'algorithme des *k*-moyennes**




Quatre arbres phylogénétiques $T_1$, $T_2$, $T_3$ et $T_4$ définis sur un ensemble de 7 feuilles; leur arbre consensus majoritaire classique $T_{1234}$ et la solution à deux arbres-consensus majoritaires $T_{12}$ et $T_{34}$.

[1] http://tolweb.org/tree/

# ALGORITHME

**Nom : *Consensus-trees***

**Méthode** : utilisation de l'algorithme des *k*-moyennes pour partitionner un ensemble d'arbres phylogénétiques

**Propriété:** *arbre consensus est un arbre médian d'un groupe d'arbres dans le sens de la distance topologique de Robinson et Foulds* (Barthélemy et McMorris,1986).

**Données en entrée** :

- *n* arbres phylogénétiques définis sur le même ensemble d'espèces (*i.e.*, objets, taxa)

**Données en sortie**:

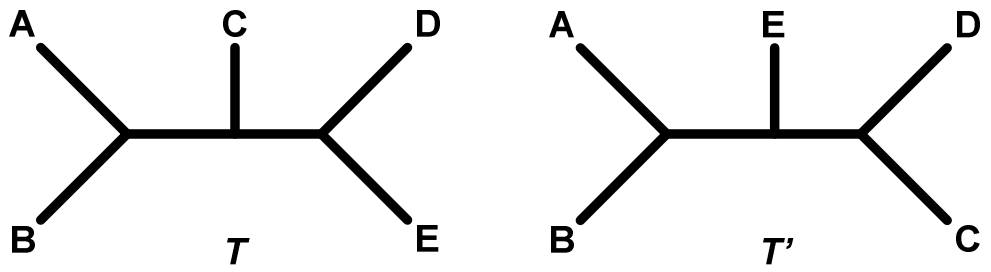
- partitionnement optimal de ces arbres en un ou plusieurs groupes;

- pour chaque groupe retrouvé:
  - ✓ la liste des arbres phylogénétiques associés
  - ✓ l'arbre-consensus de ce groupe
  - ✓ Indices utilisés:
    - ➢ *CH* (Calinski-Harabasz, 1974),
    - ➢ *W* (notre nouvelle fonction objective).

# ALGORITHME DES *K*-MOYENNES

*Description*: Permet de déterminer le partitionnement optimal des données (*i.e.*, arbres phylogénétiques dans notre cas) en *k* groupes selon un critère de similarité (MacQueen, 1968).

## Distance choisie: distance de Robinson et Foulds (1981)




La distance topologique de Robinson et Foulds entre deux arbres phylogénétiques est égale au nombre minimal d'opérations élémentaires de fusion et de séparation de noeuds, nécessaires pour transformer un arbre en un autre ($d$(T,T') = 2).

## Critères d'évaluation

1) Calinski-Harabasz (1974);

2) Fonction objective *W*.

# CRITÈRE DE CALINSKI-HARABASZ (1974)

## Formule de Calinski-Harabasz

$$CH = \frac{SSB}{SSW} \times \frac{(N-K)}{(K-1)}$$

*N* – nombre d'arbres phylogénétiques
*K* – nombre de groupes
SSB – indice d'évaluation intergroupe
*SSW* – indice d'évaluation intragroupe

# L'INDICE *SSB* ET L'INDICE *SSW*

## SSB – indice d'évaluation intergroupe

$$SSB = \frac{1}{N}\sum_{i=1}^{N-1}\sum_{j=i}^{N} w_{ij} \times \frac{RF(T_{ik}, T_{jk'})}{(2n_{ijk} - 6)} - SSW$$

*RF* – Distance de Robinson et Foulds (1981)
($T_{ik}$, $T_{jk'}$) – Deux arbres phylogénétiques $T_{ik}$ et $T_{jk'}$ appartenant à des classes différentes
*RF* ($T_{ik}$, $T_{jk}$) – Distance *RF* entre les arbres phylogénétiques $T_{ik}$ et $T_{jk}$
$N_k$ – Nombre d'arbres phylogénétiques dans le cluster *k*
*K* – Nombre de clusters
*N* – Nombre d'arbres phylogénétiques dans l'ensemble des jeux de données

**Limite:**
Ne permet pas de comparer la solution en un arbre-consensus unique (cas où $K = 1$) avec la solution admettant les arbres-consensus multiples (cas où $K \geq 2$).

## SSW – indice d'évaluation intragroupe

$$SSW = \sum_{k=1}^{K}\sum_{i=1}^{N_k-1}\sum_{j=i}^{N_k} w_{ij} \times \frac{RF(T_{ik}, T_{jk})}{(2n_{ijk} - 6)}$$

($T_{ik}$, $T_{jk}$) – Deux arbres phylogénétiques $T_{ik}$ et $T_{jk}$ appartenant à la même classe

# FONCTION OBJECTIVE *W*

## Fonction objective *W*

$$W(\Pi) = \frac{1}{(N-K)} \sum_{k=1}^{K} \frac{2}{N_k \times (N_k - 1)} \sum_{i=1}^{N_k - 1} \sum_{j=i+1}^{N_k} w_{ij} \times \frac{RF(T_{ik}, T_{jk})}{(2n_{ijk} - 6)} \rightarrow Min$$

*RF* – Distance de Robinson et Foulds (1981)
$T_{ik}$ et $T_{jk}$ – Deux arbres phylogénétiques $T_{ik}$ et $T_{jk}$ appartenant à la même classe *k*
*RF* ($T_{ik}$, $T_{jk}$) – Distance *RF* entre les arbres phylogénétiques $T_{ik}$ et $T_{jk}$
$N_k$ – Nombre d'arbres phylogénétiques dans la classe *k*
*K* – Nombre de classes
*N* – Nombre total d'arbres phylogénétiques considérés

**Limite:**
Ne tiens pas compte de la distance intergroupe.

# Simulations

# VALIDATION DE L'APPROCHE ET DES CRITÈRES

Plan des simulations:

- **Étape 1**: Générer *k* arbres phylogénétiques binaires aléatoires $\{T_1 \ldots T_k\}$ avec *n* feuilles chacun, en utilisant le site T-Rex[1] (Boc *et al.*, 2012), où $k = \{1 \ldots 10\}$ et $n = \{8, 16, 32, 64\}$.

- **Étape 2**: Pour chaque arbre phylogénétique $T_i$ (où $i = 1 \ldots k$), générer l'ensemble de 100 arbres appartenant à la classe *i* pour chacun des intervalles indiqués ci-dessous. Pour ce faire: nous allons générer des arbres phylogénétiques aléatoires tels que le pourcentage de similitude (mesuré à l'aide de la distance *RF*) entre eux et $T_i$ soit:

de 0 à 10% (Intervalle I),
de 10 à 25% (Intervalle II),
de 25 à 50% (Intervalle III),
de 50 à 75% (Intervalle IV).

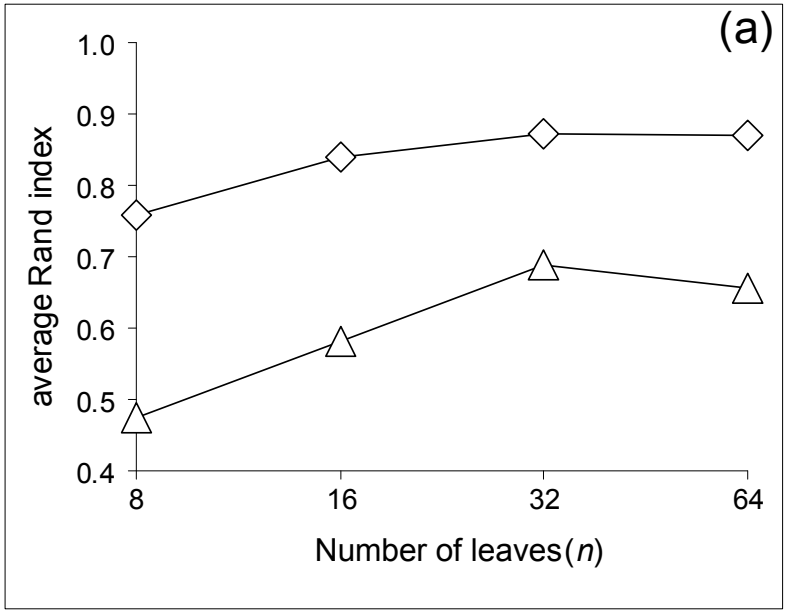
- **Étape 3**: Exécuter l'algorithme *Consensus-trees* sur les ensembles d'arbres générés avec les différents paramètres *(k, n*, Intervalle, Fonction Objective *W* et le critère *CH*).

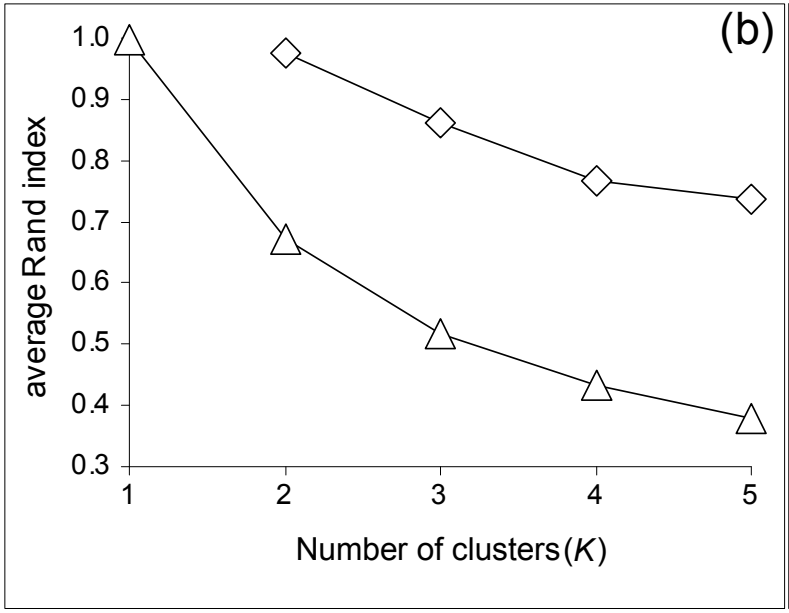
[1]http://trex.uqam.ca/index.php?action=randomtreegenerator&project=trex

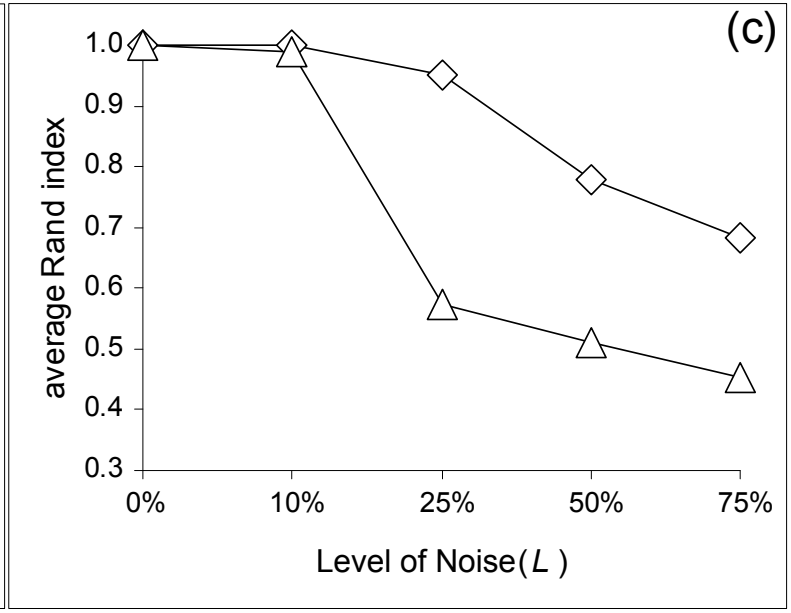
# VALIDATION DE L'APPROCHE ET DES CRITÈRES

Étude préliminaire de l'évolution de l'indice Rand moyen:

(a) en fonction du nombre de feuilles (*n*);
(b) en fonction du nombre de partitions (*k*);
(c) en fonction du pourcentage de similitudes (*L*) entre les arbres phylogénétiques (Intervalle)
pour les deux critères:
◊ Calinski-Harabasz; △ *W.*



# Projet 2 : Classification d’arbres phylogénétiques : Super-arbres

# ALGORITHMES D'INFÉRENCE D'UN SUPER-ARBRE

**Les méthodes des super-arbres réconcilient des arbres phylogénétiques définis sur des ensembles de taxons différents, mais partiellement chevauchants.**

Principales méthodes d'inférence de super-arbres:

- Dans le passé (Gordon, 1986)

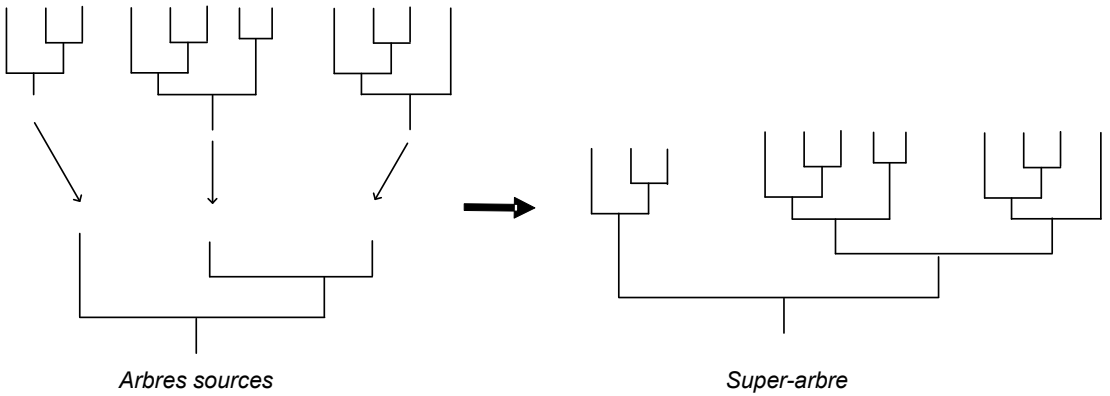



Illustration d'une reconstruction dans le passé (Bininda-Edmonds, 2004).

# ALGORITHMES D'INFÉRENCE D'UN SUPER-ARBRE

**Les méthodes des super-arbres réconcilient des arbres phylogénétiques définis sur des ensembles de taxons différents, mais partiellement chevauchants.**

Principales méthodes d'inférence de super-arbres:

- Dans le présent : MPR (Ragan, 1992; Doyle, 1992; Baum, 1992)

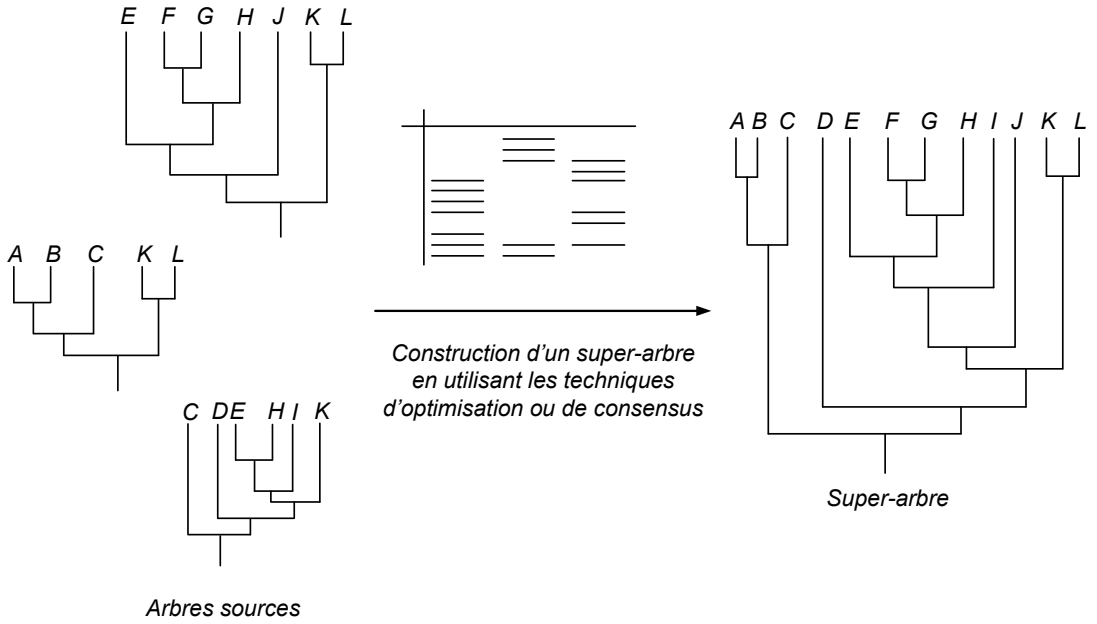



Illustration d'une reconstruction dans le présent (Bininda-Edmonds, 2004).

# PROBLÉMATIQUE

**Idée: Généraliser la classification d'arbres phylogénétiques ayant le même ensemble de taxons.**

**Motivation: Contribuer au projet ToF**[1] (Maddison *et al.*, 2007)

- ❑ Nécessité de **fusionner** les arbres phylogénétiques via le projet ToL (Tree of Life);
- ❑ **Perte** d'informations et mauvais recoupement des feuilles;
- ❑ **Incohérence** de la fusion des arbres phylogénétiques.

**Nous proposons ici une méthode de partitionnement d'un ensemble de *n* arbres phylogénétiques ayant des ensembles de feuilles différents, qui se base sur l'algorithme des *k*-moyennes.**

[1] http://tolweb.org/tree/

# ALGORITHME

**Nom : *Super-trees***

**Méthode** : utilisation de l'algorithme des *k*-moyennes pour classifier les arbres phylogénétiques

**Données en entrée** :

- *n* arbres phylogénétiques définis sur des ensembles différents d'espèces (*i.e.*, objets, taxa), mais chevauchants

Particularité: Il faudra filtrer les ensembles d'espèces

Difficulté: Définir un seuil minimum de feuilles chevauchantes entre les arbres phylogénétiques

**Données en sortie**:

- partitionnement optimal de ces arbres en un ou plusieurs groupes;
- pour chaque groupe retrouvé:
  - ✓ la liste des arbres phylogénétiques associés;
  - ✓ l'arbre-consensus de ce groupe
  - ✓ Indices utilisés: *W'* et *W''*

# FONCTIONS OBJECTIVES *W' ET W''*

$$W'(\Pi) = \left(\frac{1}{(N-K)} \times \sum_{k=1}^{K} \left(\left(\frac{2}{N_k \times (N_k - 1)}\right) \times \sum_{i=1}^{N_k} RF(C_k^{MR}, T_{ik})\right)\right) \rightarrow \min$$

*K* – nombre de groupes
$N_k$ – nombre d'arbres phylogénétiques dans le groupe *k*
$C_k^{MR}$ – super-arbre majoritaire (*i.e.*, l'arbre centroïde) du groupe *k* obtenu par la règle majoritaire (*MR*)
$RF(C_k^{MR}, T_{ik})$ – distance *RF* entre $C_k^{MR}$ et $T_{ik}$

$$W''(\Pi) = \left(\frac{1}{(N-K)} \times \sum_{k=1}^{K} \left(\left(\frac{2}{N_k \times (N_k - 1)}\right) \times \sum_{i=1}^{N_k - 1} \sum_{j=i+1}^{N_k} \frac{RF^2(T'_{ik}, T'_{jk})}{(2n_{ijk} - 6)}\right)\right) \rightarrow Min$$

$T'_{ik}$ et $T'_{ik}$ - arbres phylogénétiques *i* et *j* du groupe *k* réduits aux feuilles communes
$RF(T'_{ik}, T'_{jk})$ – distance *RF* entre $T'_{ik}$ et $T'_{jk}$
$n_{ijk}$ – nombre de feuilles communes aux arbres $T'_{ik}$ et $T'_{jk}$.

# VALIDATION DE L'APPROCHE ET DES CRITÈRES

Plan des simulations:

- **Étape 1**: Générer *k* arbres phylogénétiques binaires aléatoires {$T_1 \ldots T_k$}, ayant de $n_1$ à $n_2$ feuilles chacun (et au moins *n* feuilles communes) , en utilisant le site T-Rex[1] (Boc *et al.*, 2012), où $k = \{1 \ldots 10\}$ et $n = \{8, 16, 32, 64\}$.

- **Étape 2**: Pour chaque arbre phylogénétique $T_i$ (où $i = 1 \ldots k$), générer l'ensemble de 100 arbres appartenant à la classe *i* pour chacun des intervalles indiqués ci-dessous. Pour ce faire: nous allons générer des arbres phylogénétiques aléatoires tels que le pourcentage de similitude (mesuré à l'aide de la distance *RF*) entre eux et $T_i$ soit:
 de 0 à 10% (Intervalle I),
 de 10 à 25% (Intervalle II),
 de 25 à 50% (Intervalle III),
 de 50 à 75% (Intervalle IV).

- **Étape 3**: Exécuter l'algorithme *Super-trees* sur les ensembles d'arbres générés avec les différents paramètres (*k*, *n*, $n_1$, $n_2$, Intervalle, Fonc. Obj. = *W'', CH, BH, LogSS et Silhouette*).

[1]http://trex.uqam.ca/index.php?action=randomtreegenerator&project=trex

# Projet 3 : Applications

# CLASSIFICATION DES PROTÉINES RIBOSOMALES DES ARCHAEBACTÉRIES (MATTE-TAILLIEZ *ET AL.*, 2002 )

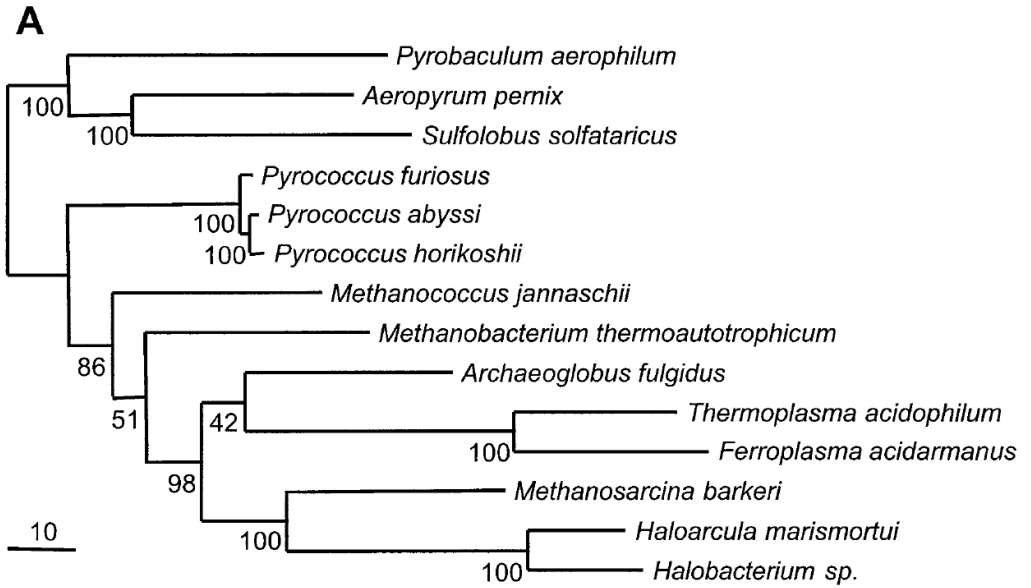
Données:

- Soit 49 protéines ribosomales de 14 archéobactéries (étudiées initialement par Matte-Tailliez *et al.*, 2002).

Motivations:

- Trouver les protéines des 14 archaebactéries partageant la même histoire évolutive.

- Détecter les gènes qui ont subi les mêmes transferts horizontaux (HGT).

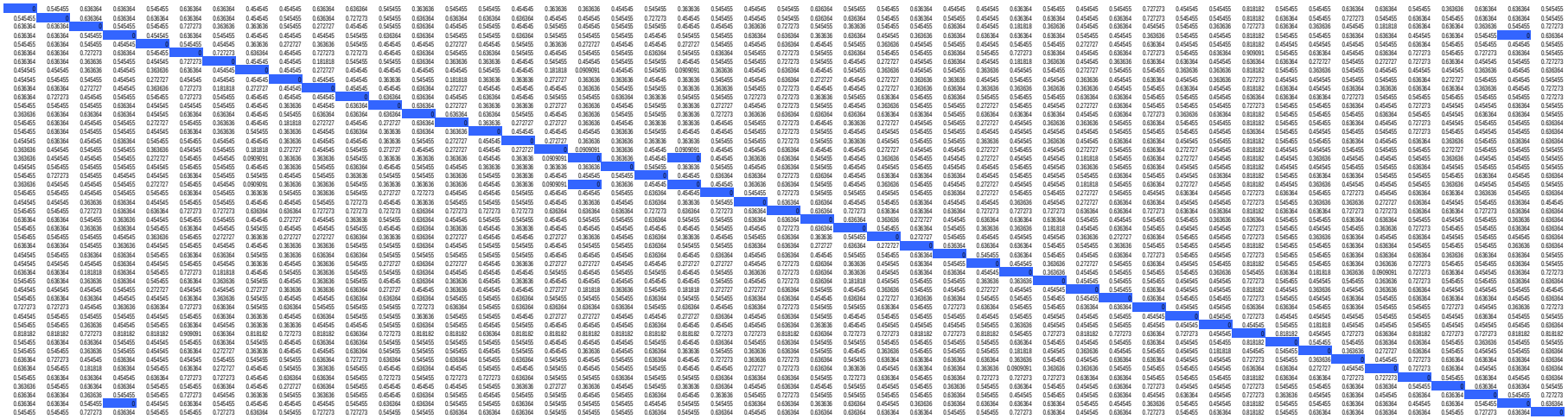



L'arbre d'espèces pour 14 archéobactéries
*(Matte-Tailliez et al., Mol. Biol. Evol., 2002).*

# RÉSULTATS PRÉLIMINAIRES DES ARCHAEBACTÉRIES (MATTE-TAILLIEZ *ET AL.*, 2002 )

Algorithme : ***Consensus-trees***

Matrice des distances RF normalisées:



Résultats:

| K | Score W | Score CH |
|---|---|---|
| 1 | 0.006339 | - |
| 2 | 0.012502 | 52.832736 |
| 3 | 0.015264 | 29.699751 |
| 4 | 0.02287 | 18.689989 |
| 5 | 0.023623 | 16.376206 |
| 6 | 0.027507 | 13.608494 |
| 7 | 0.02834 | 12.573911 |
| 8 | 0.025445 | 13.896887 |
| 9 | 0.045315 | 7.588349 |
| 10 | 0.043689 | 7.584642 |

Meilleur *K*:

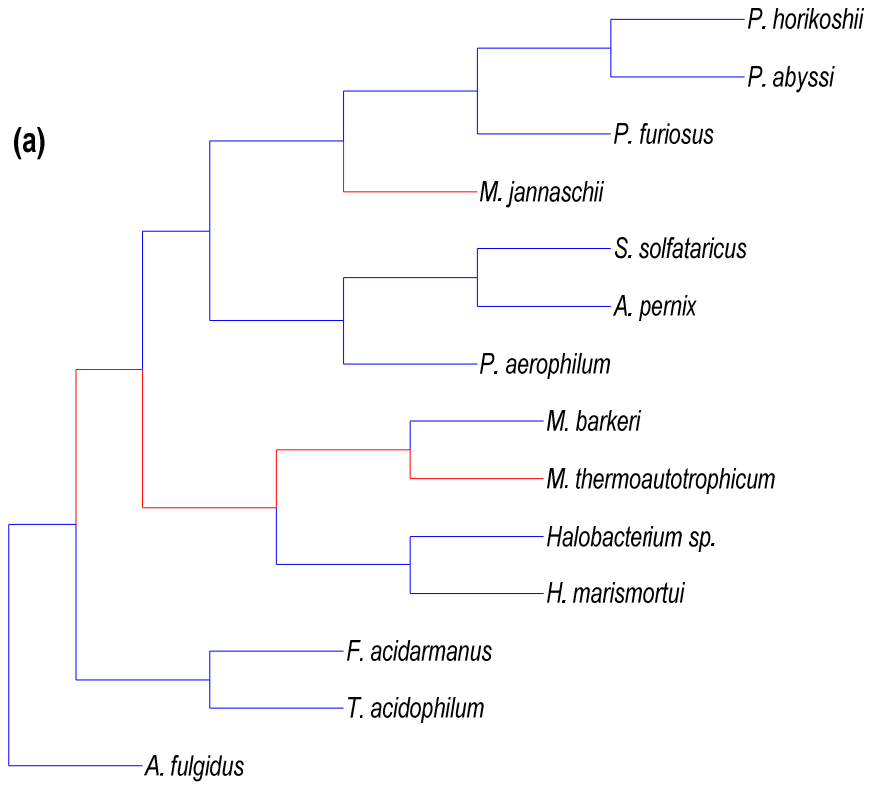
- *K* = 1; pour *W*
- *K* = 2; pour *CH*

Limites:

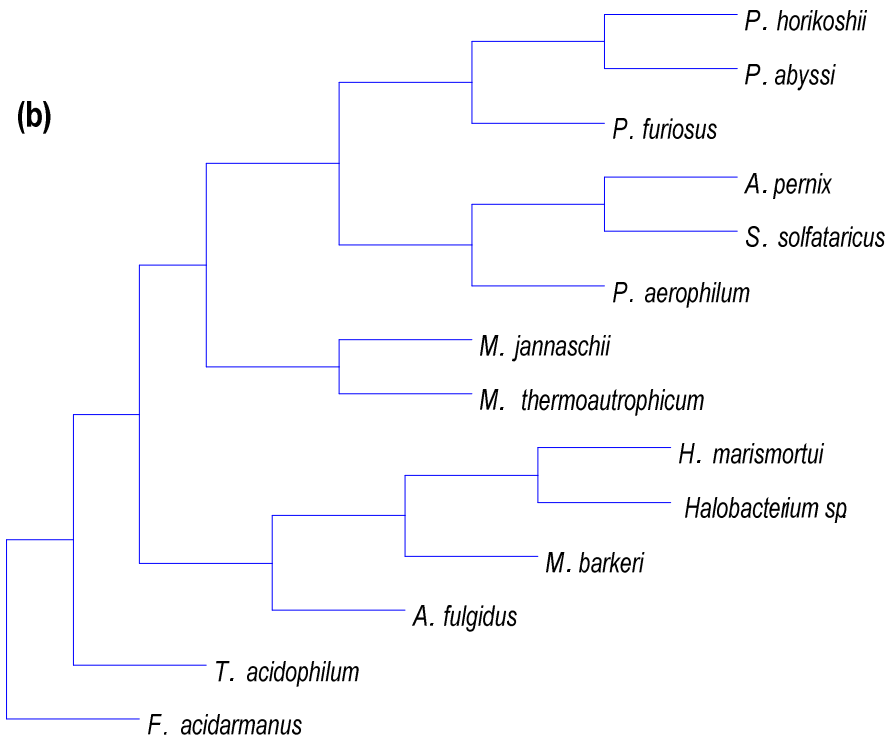
- ✓Score *W* : le minimum local
- ✓Score *CH* : le maximum local et pas de possibilité d'une solution unique

# RÉSULTATS PRÉLIMINAIRES POUR LES ARCHAEBACTÉRIES (MATTE-TAILLIEZ *ET AL.*, 2002 )

Arbres consensus selon le critère *CH*



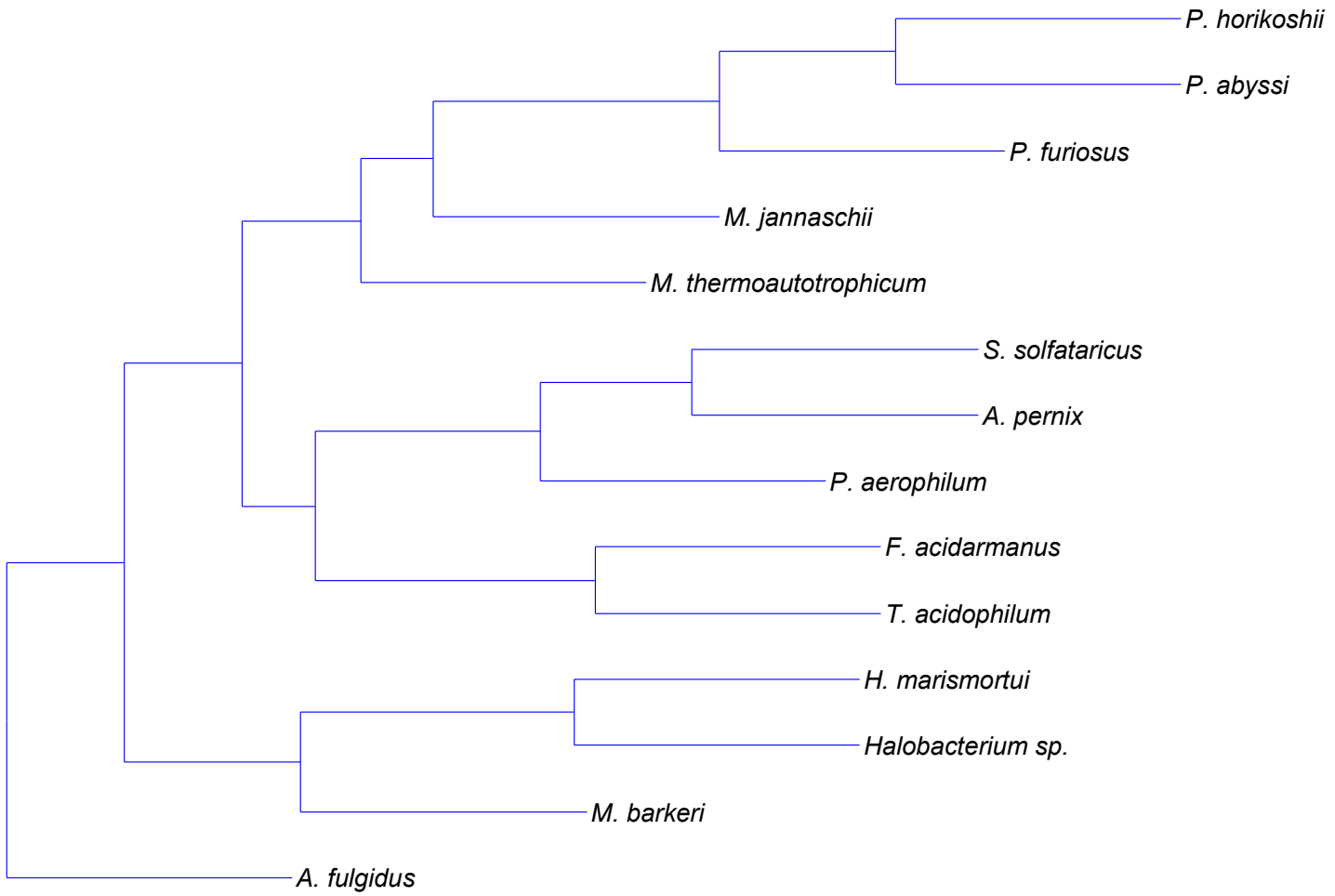

Arbre consensus 1




Arbre consensus 2

# RÉSULTATS PRÉLIMINAIRES POUR LES ARCHAEBACTÉRIES (MATTE-TAILLIEZ *ET AL.*, 2002 )

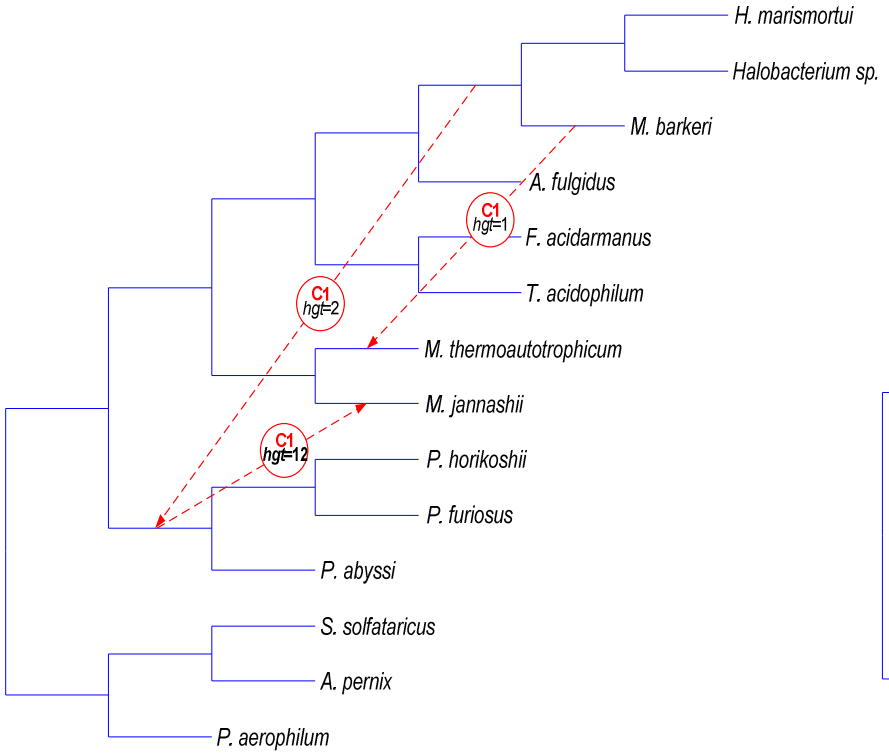
Arbre consensus selon la fonction objective *W*



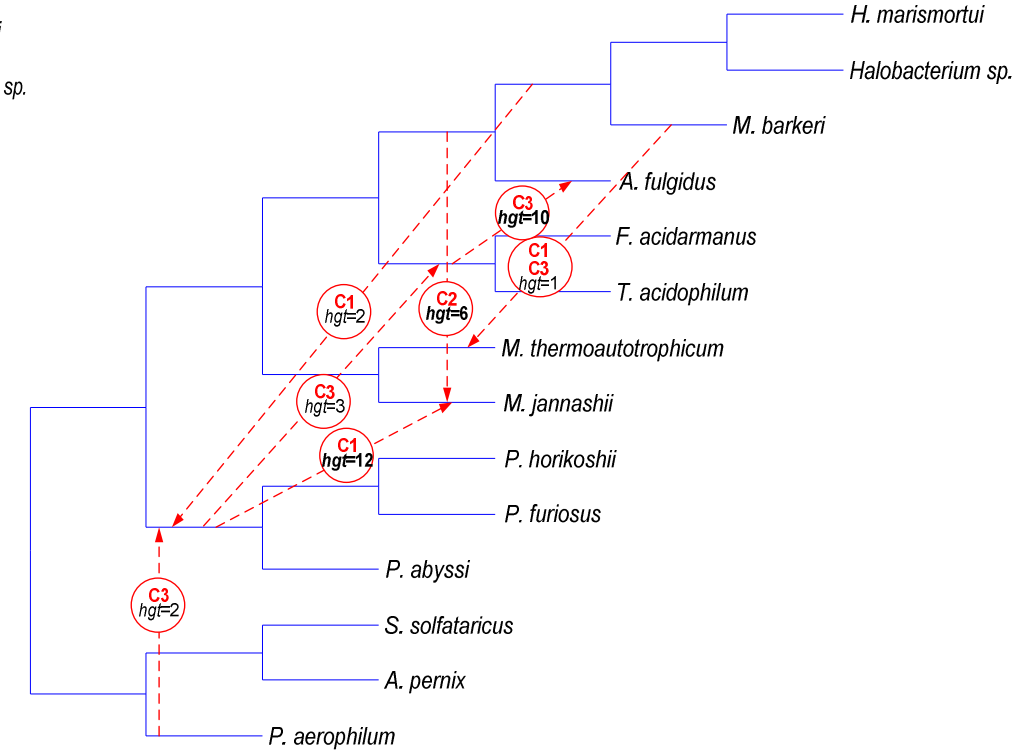

Arbre consensus unique

# RÉSULTATS PRÉLIMINAIRES POUR LES ARCHAEBACTÉRIES (MATTE-TAILLIEZ *ET AL.*, 2002 )

Un scénario de transferts horizontaux de gènes (HGT) pour différents *K:*




*K*=2




*K*=3

# AUTRES DONNÉES BIOLOGIQUES À ÉTUDIER

| Groupe biologique étudié | Nombre d'espèces | Nombre d'arbres phylogénétiques |
|---|---|---|
| Papilionidés (Gepts *et al.*, 2005) | 558 | 19 |
| Marsupiaux (Cardillo *et al.*, 2004) | 267 | 158 |
| Mammifères placentaires (Beck *et al.*, 2006) | 116 | 726 |
| Oiseaux de mer (Kennedy et Page, 2002) | 122 | 7 |

# ÉVOLUTION DES LANGUES INDO-EUROPÉENNES (IE)

La base de données:

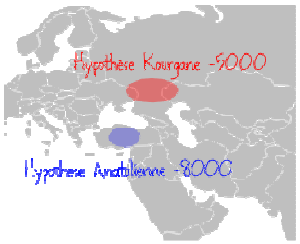
- Organisée par Dyen *et al.* (1997) et améliorée par Boc *et al.* (2010).

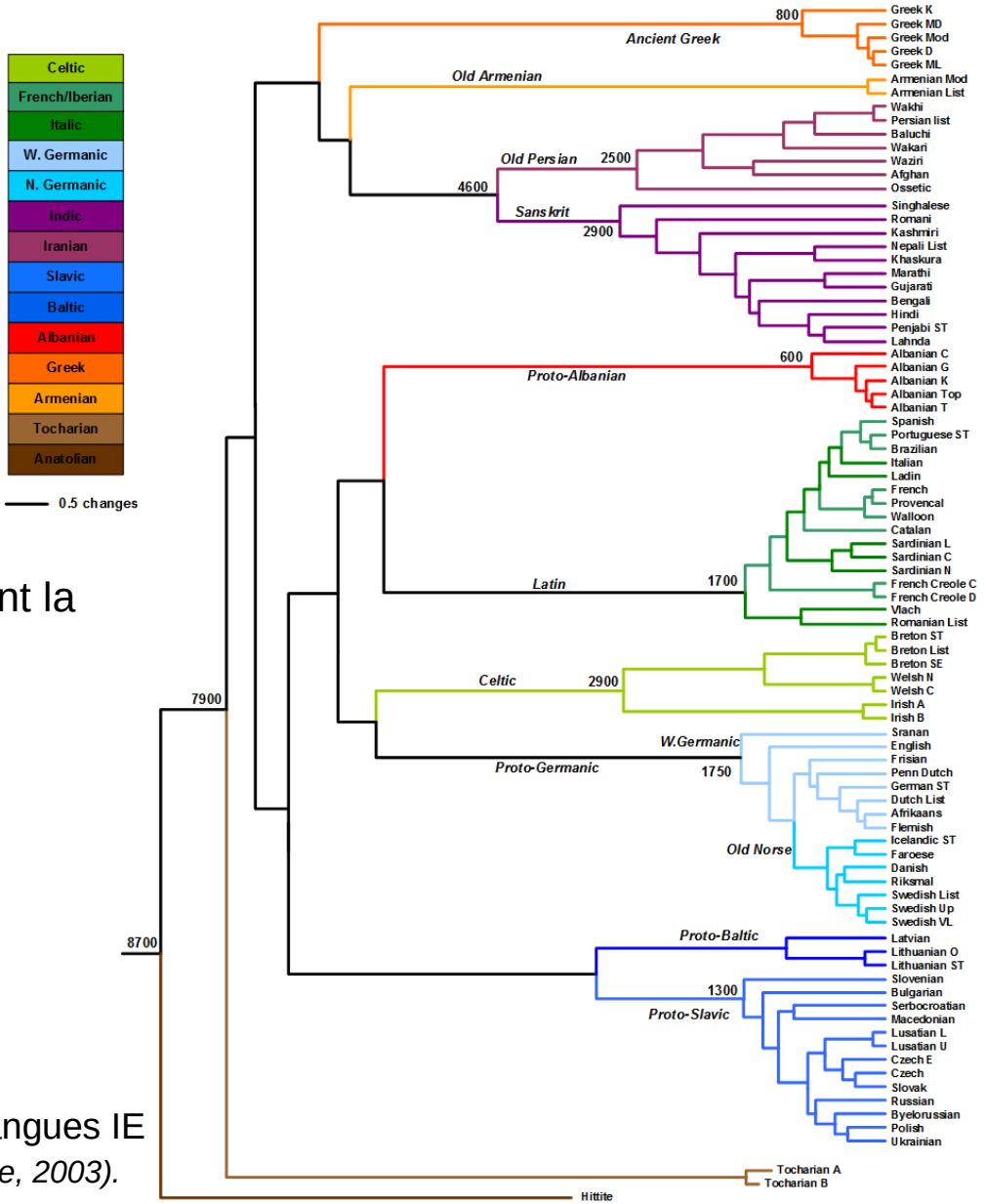
- Regroupée en 200 mots de la liste Swadesh, traduite dans 87 langues et structurée en 1315 cognats.

Motivations:

- Trouver des groupes de langues partageant la même histoire évolutive.

- Mettre en avant l'origine des langues IE
  - ❖Hypothèse Kourgane
  - ❖Hypothèse Anatolienne
  - ❖ou une nouvelle hypothèse

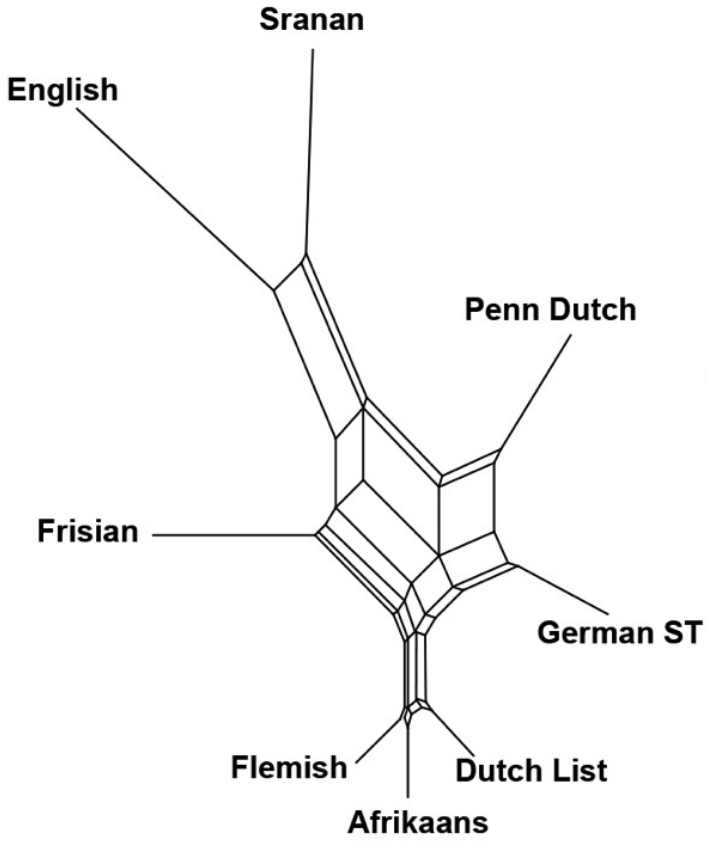


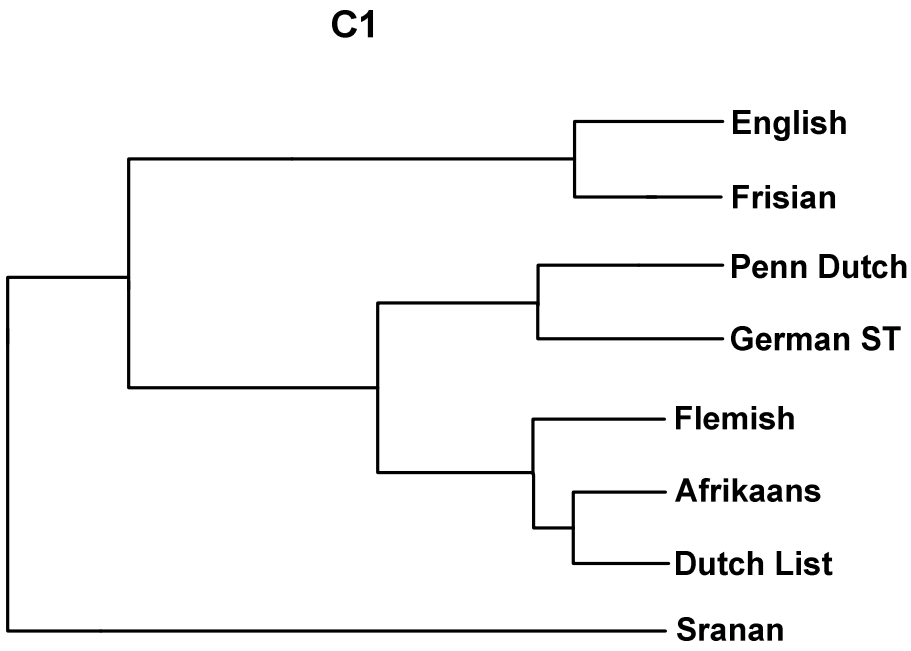

L'arbre d'évolution des langues IE
*(Gray et Atkinson, Nature, 2003).*

http://sciencetonnante.wordpress.com/2012/09/24/quelle-est-lorigine-des-langues-indo-europeennes/.

# RÉSULTATS PRÉLIMINAIRES POUR LES LANGUES IE




Split-graphe pour huit langues ouest-germaniques (Willems *et al.*)




Super-arbre *C1* par le critère *CH*

# RÉFÉRENCES


- Makarenkov, V. (2001) T-Rex: reconstructing and visualizing phylogenetic trees and reticulation networks. Bioinformatics, 17, 664-668.
- Makarenkov, V. et Legendre. P. (2001) Optimal variable weighting for ultrametric and additive trees and K-means partitioning: Methods and software. *Journal of Classification* 18(2) : 245-271.
- Margush, T. et McMorris, F. R. (1981). Consensusn-trees. *Bulletin of Mathematical Biology*, 43(2), 239-244.
- Matte-Tailliez, O., Brochier,C., Forterre, P. et Philippe,H. (2002) Archaeal phylogeny based on ribosomal proteins. *Mol. Biol. Evol.*, 19, 631-639.
- Kennedy, M. et Page, R. D. (2002). Seabird supertrees: combining partial estimates of procellariiform phylogeny. *The Auk*, 119(1), 88-108.
- Ragan, M. A. (1992). Matrix representation in reconstructing phylogenetic relationships among the eukaryotes. *Biosystems*, 28(1), 47-55.
- Robinson, D.R. et Foulds, L.R. (1981) Comparison of phylogenetic trees. *Mathematical Biosciences*, 53, 131-147.
- Sokol, R. R. et Rohlf, F. J. (1981). Biometry: the principles and practice of statistics. *Biological research*. WH Freeman Inc, San Francisco, 849.
- Stamatakis, A., Hoover, P. et Rougemont, J. (2008). A rapid bootstrap algorithm for the RAxML web servers. *Systematic biology*, 57(5), 758-771.
- Tahiri, N., Willems, M., Makarenkov, V. (2014) Classification d'arbres phylogénétiques basée sur l'algorithme des k-moyennes, Actes de *SFC-2014*.

**Merci de votre attention !!!**

# REMERCIEMENTS

Mon directeur de thèse : Vladimir Makarenkov

Recherche étudiants-es pour :

- ❑ Maîtrise
- ❑ Doctorat
- ❑ Postdoc

**Les fonds FQRNT pour le financement de ce projet.**