# Identification bactérienne et phylogénie

Jean-Pierre Flandrois[1]

[1]Laboratoire de Biométrie et Biologie Evolutive
Université Claude Bernard Lyon 1 ; CNRS UMR 5558 "Microbiologie Quantitative"

21 novembre 2005

# Outlines

1 Introduction
- Survol du processus d'identification
- BLAST et la similitude

2 BIBI
- Projet Bio-Informatic Bacteria Identification
- Anomalies du processus d'identification
- Objectifs d'amélioration

3 The Tree For Bacteria Identification Project
- Objectifs T4Bi
- Principales étapes
- Automatisation de T4Bi
- Perspectives

# Identification bactérienne

- Base du diagnostic étiologique + l'industrie et l'écologie.
- Inclure une souche inconnue dans une unité taxonomique préexistante.
  - Le taxon (espèce) reconnu par l'IUMS.
  - La définition (des taxons) de l'espèce est fondée sur la phylogénie.
- Les souches d'une espèce ont un "certain niveau" de similitude phénotypique.
- Proximité habituelle de la "souche type".
- Une même espèce sous le même noeud de la phylogénie.

# Identification bactérienne

- Base du diagnostic étiologique + l'industrie et l'écologie.
- Inclure une souche inconnue dans une unité taxonomique préexistante.
  - Le taxon (espèce) reconnu par l'IUMS.
  - La définition (des taxons) de l'espèce est fondée sur la phylogénie.
- Les souches d'une espèce ont un "certain niveau" de similitude phénotypique.
- Proximité habituelle de la "souche type".
- Une même espèce sous le même noeud de la phylogénie.

# Identification bactérienne

- Base du diagnostic étiologique + l'industrie et l'écologie.
- Inclure une souche inconnue dans une unité taxonomique préexistante.
  - Le taxon (espèce) reconnu par l'IUMS.
  - La définition (des taxons) de l'espèce est fondée sur la phylogénie.
- Les souches d'une espèce ont un "certain niveau" de similitude phénotypique.
- Proximité habituelle de la "souche type".
- Une même espèce sous le même noeud de la phylogénie.

# Identification "classique"

- GenBank.
- Recherche par Blast.
- Alignement et similitudes.
  - Pas de consensus pour un niveau de similarité pour l'espèce.
  - Similarité pour le niveau espèce dépend du genre et du gène utilisé.
  - Le paramétrage de Blast est critique (Open-Extend gap parameters).

# Identification "classique"

- GenBank.
- Recherche par Blast.
- Alignement et similitudes.
  - Pas de consensus pour un niveau de similarité pour l'espèce.
  - Similarité pour le niveau espèce dépend du genre et du gène utilisé.
  - Le paramétrage de Blast est critique (Open-Extend gap parameters).

# Outlines

1 Introduction
- Survol du processus d'identification
- BLAST et la similitude

2 BIBI
- Projet Bio-Informatic Bacteria Identification
- Anomalies du processus d'identification
- Objectifs d'amélioration

3 The Tree For Bacteria Identification Project
- Objectifs T4Bi
- Principales étapes
- Automatisation de T4Bi
- Perspectives

# BIBI : Objectifs et definition

- Sous ensemble des banques publiques (BIBIDB).
  - Conformité à la Nomenclature.
  - Critère de souches types, de qualité, de longueur.
  - SSU rDNA.
- Banques spécialisées *Corynebacterinae*.
- Combiner une recherche par Blast et la phylogénie.
- Environnement orienté utilisateur non spécialisé.
  - Un ensemble d'outils améliore la décision.
  - Des liens vers des sites d'expertise.

# BIBI : Objectifs et definition

- Sous ensemble des banques publiques (BIBIDB).
  - Conformité à la Nomenclature.
  - Critère de souches types, de qualité, de longueur.
  - SSU rDNA.
- Banques spécialisées *Corynebacterinae*.
- Combiner une recherche par Blast et la phylogénie.
- Environnement orienté utilisateur non spécialisé.
  - Un ensemble d'outils améliore la décision.
  - Des liens vers des sites d'expertise.

# BIBI : Objectifs et definition

- Sous ensemble des banques publiques (BIBIDB).
  - Conformité à la Nomenclature.
  - Critère de souches types, de qualité, de longueur.
  - SSU rDNA.
- Banques spécialisées *Corynebacterinae*.
- Combiner une recherche par Blast et la phylogénie.
- Environnement orienté utilisateur non spécialisé.
  - Un ensemble d'outils améliore la décision.
  - Des liens vers des sites d'expertise.

# BIBI : Objectifs et definition

- Sous ensemble des banques publiques (BIBIDB).
  - Conformité à la Nomenclature.
  - Critère de souches types, de qualité, de longueur.
  - SSU rDNA.
- Banques spécialisées *Corynebacterinae*.
- Combiner une recherche par Blast et la phylogénie.
- Environnement orienté utilisateur non spécialisé.
  - Un ensemble d'outils améliore la décision.
  - Des liens vers des sites d'expertise.

# BIBI : Objectifs et definition

- Sous ensemble des banques publiques (BIBIDB).
  - Conformité à la Nomenclature.
  - Critère de souches types, de qualité, de longueur.
  - SSU rDNA.
- Banques spécialisées *Corynebacterinae*.
- Combiner une recherche par Blast et la phylogénie.
- Environnement orienté utilisateur non spécialisé.
  - Un ensemble d'outils améliore la décision.
  - Des liens vers des sites d'expertise.

# BIBI workflow

- Entrer la séquence inconnue, choisir la banque.
- Sélection des séquences candidates par Blast.
- Alignement des sequences par Mabios (blocs).
- Résultats tabulés, arbre NJ et alignement.
- Outils de décision et de recherche d'information.

# BIBI workflow

- Entrer la séquence inconnue, choisir la banque.
- Sélection des séquences candidates par Blast.
- Alignement des sequences par Mabios (blocs).
- Résultats tabulés, arbre NJ et alignement.
- Outils de décision et de recherche d'information.

# BIBI workflow

- Entrer la séquence inconnue, choisir la banque.
- Sélection des séquences candidates par Blast.
- Alignement des sequences par Mabios (blocs).
- Résultats tabulés, arbre NJ et alignement.
- Outils de décision et de recherche d'information.

# BIBI output

**Sequence features**

| Sequence size | A | C | T | G | N | GC% |
|---|---|---|---|---|---|---|
| 472 | 109 | 113 | 87 | 163 | 0 | 58.47 |

**Realignment without checked sequences** >>

**Identification result**

| Distance | Level | GenBank | Sequence name | LBSN | sp.rep | seqC | size | #N | simil | QI | remove |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0000 | 0 | AJ409154 | Nocardia sp. partial 16S rRNA gene, strain | Info | ND | ND | 472 | 0 | 100 | 100.00 | 1 |
| 0.0000 | 0 | AJ298932 | Nocardia salmonicida partial 16S rRNA gene, | Info | ND | ND | 472 | 0 | 100 | 100.00 | 2 |
| 0.0090 | 1 | AF277210 | Nocardia sp. R7 16S ribosomal RNA gene, | Info | ND | ND | 438 | 1 | 98 | 98.28 | 5 |
| 0.0120 | 1 | AF430063 | Nocardia sp. DSM 6249 16S ribosomal RNA gene, | Info | ND | ND | 453 | 0 | 98 | 98.80 | 3 |
| 0.0120 | 1 | AF277216 | Nocardia sp. R26 16S ribosomal RNA gene, | Info | ND | ND | 435 | 0 | 98 | 98.16 | 6 |
| 0.0120 | 1 | AF277217 | Nocardia sp. R47 16S ribosomal RNA gene, | Info | ND | ND | 430 | 0 | 98 | 97.98 | 8 |
| 0.0120 | 2 | AF430053 | Nocardia fluminea strain DSM 44489 16S | Info | ND | ND | 453 | 3 | 98 | 98.69 | 4 |
| 0.0160 | 2 | AF277206 | Nocardia sp. W138 16S ribosomal RNA gene, | Info | ND | ND | 439 | 0 | 98 | 98.24 | 7 |
| 0.0160 | 2 | AF277204 | Nocardia fluminea 16S ribosomal RNA gene, | Info | ND | ND | 436 | 0 | 98 | 98.13 | 14 |

# Anomalies relevées

## Observations d'anomalies

- Positionnement, erreurs de dénomination.
- Incongruences phylogénie-taxinomie (nomenclature).

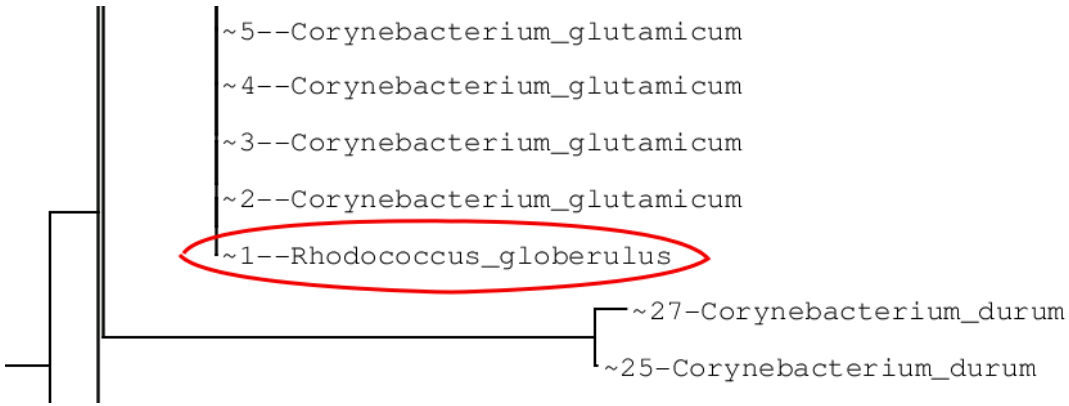
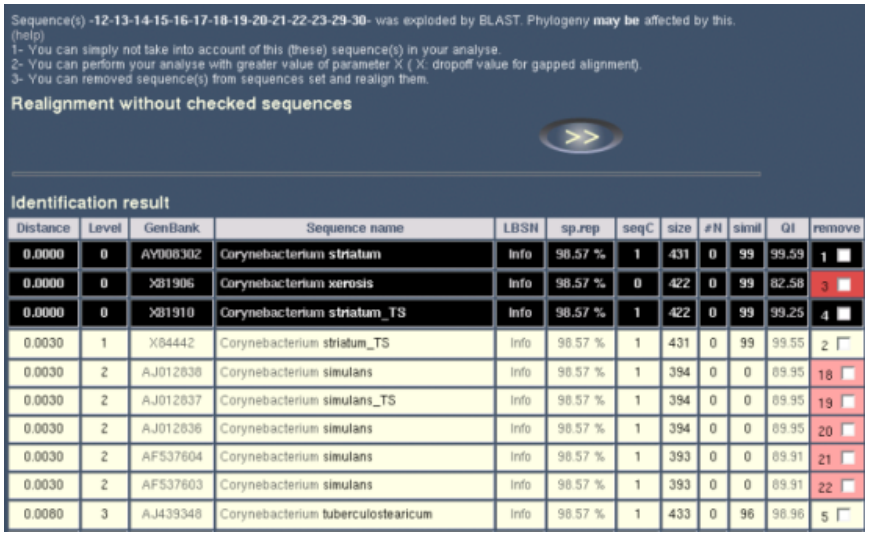



FIG.: Erreur nomenclature

# Gestion des anomalies

## Glanage des erreurs par les biologistes

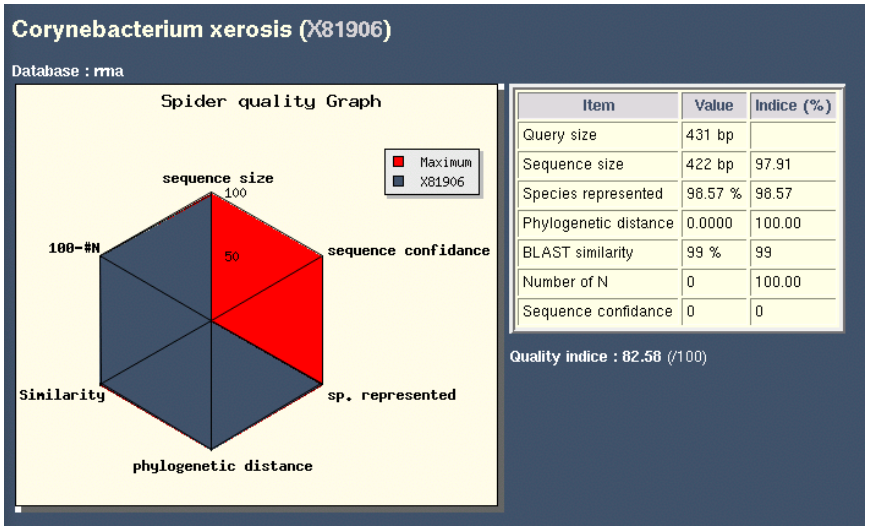

Sequence(s) -12-13-14-15-16-17-18-19-20-21-22-23-29-30- was exploded by BLAST. Phylogeny **may be** affected by this.
(help)
1- You can simply not take into account of this (these) sequence(s) in your analyse.
2- You can perform your analyse with greater value of parameter X ( X: dropoff value for gapped alignment).
3- You can removed sequence(s) from sequences set and realign them.

**Realignment without checked sequences**

**Identification result**

| Distance | Level | GenBank | Sequence name | LBSN | sp.rep | seqC | size | #N | simil | QI | remove |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 0.0000 | 0 | AY008302 | Corynebacterium **striatum** | Info | 98.57 % | 1 | 431 | 0 | 99 | 99.59 | 1 |
| 0.0000 | 0 | X81906 | Corynebacterium **xerosis** | Info | 98.57 % | 0 | 422 | 0 | 99 | 82.58 | 3 |
| 0.0000 | 0 | X81910 | Corynebacterium **striatum_TS** | Info | 98.57 % | 1 | 422 | 0 | 99 | 99.25 | 4 |
| 0.0030 | 1 | X84442 | Corynebacterium **striatum_TS** | Info | 98.57 % | 1 | 431 | 0 | 99 | 99.55 | 2 |
| 0.0030 | 2 | AJ012838 | Corynebacterium **simulans** | Info | 98.57 % | 1 | 394 | 0 | 0 | 89.95 | 18 |
| 0.0030 | 2 | AJ012837 | Corynebacterium **simulans_TS** | Info | 98.57 % | 1 | 394 | 0 | 0 | 89.95 | 19 |
| 0.0030 | 2 | AJ012836 | Corynebacterium **simulans** | Info | 98.57 % | 1 | 394 | 0 | 0 | 89.95 | 20 |
| 0.0030 | 2 | AF537604 | Corynebacterium **simulans** | Info | 98.57 % | 1 | 393 | 0 | 0 | 89.91 | 21 |
| 0.0030 | 2 | AF537603 | Corynebacterium **simulans** | Info | 98.57 % | 1 | 393 | 0 | 0 | 89.91 | 22 |
| 0.0080 | 3 | AJ439348 | Corynebacterium **tuberculostearicum** | Info | 98.57 % | 1 | 433 | 0 | 96 | 98.96 | 5 |

FIG.: Indication dans la table




| Item | Value | Indice (%) |
|---|---|---|
| Query size | 431 bp | |
| Sequence size | 422 bp | 97.91 |
| Species represented | 98.57 % | 98.57 |
| Phylogenetic distance | 0.0000 | 100.00 |
| BLAST similarity | 99 % | 99 |
| Number of N | 0 | 100.00 |
| Sequence confidance | 0 | 0 |

**Quality indice : 82.58** (/100)

FIG.: Indicateurs qualité

# Objectifs d'amélioration

## Constats :

- Les bases de données génomiques ne sont pas "propres".
- Ceci perturbe le diagnostic d'espèce.
- Le processus s'amplifie par propagation des erreurs.

## Remédiation :

- Ne pas compter sur le glanage de l'information.
- Nécessité de détecter ces erreurs.

# Outlines

1 Introduction
- Survol du processus d'identification
- BLAST et la similitude

2 BIBI
- Projet Bio-Informatic Bacteria Identification
- Anomalies du processus d'identification
- Objectifs d'amélioration

3 The Tree For Bacteria Identification Project
- Objectifs T4Bi
- Principales étapes
- Automatisation de T4Bi
- Perspectives

# Objectifs de T4Bi

## Comment reconnaître des séquences mal nommées

Utilisation des incongruences entre la taxinomie et la phylogénie

Création de T4BI : Tree For Bacteria Identification

- Génération d'arbres phylogénétiques pour l'ensemble des genres et des espèces pour un fragment de gène donné en un temps réduit.
- Visualisation et détection des anomalies.

# Principales étapes

- Création de banques de données :
  - Taxinomiques, utilisation de NCBI, DSMZ, LBSN.
  - Génomiques,TaxoBacGen (ACNUC) → base de données par genre.
- Traitements des séquences :
  - Sélection du fragment (ex : SSU rDNA, fragment 4-6).
  - Suppression des séquences de mauvaise qualité, trop courtes, identiques.
  - Alignement des séquences par espèces, par genre.
- Construction des arbres :
  - Arbres avec genres voisins ("cousins").
- Interface web : scripts CGI.

# Principales étapes

- Création de banques de données :
  - Taxinomiques, utilisation de NCBI, DSMZ, LBSN.
  - Génomiques,TaxoBacGen (ACNUC) → base de données par genre.
- Traitements des séquences :
  - Sélection du fragment (ex : SSU rDNA, fragment 4-6).
  - Suppression des séquences de mauvaise qualité, trop courtes, identiques.
  - Alignement des séquences par espèces, par genre.
- Construction des arbres :
  - Arbres avec genres voisins ("cousins").
- Interface web : scripts CGI.

# Principales étapes

- Création de banques de données :
  - Taxinomiques, utilisation de NCBI, DSMZ, LBSN.
  - Génomiques,TaxoBacGen (ACNUC) → base de données par genre.
- Traitements des séquences :
  - Sélection du fragment (ex : SSU rDNA, fragment 4-6).
  - Suppression des séquences de mauvaise qualité, trop courtes, identiques.
  - Alignement des séquences par espèces, par genre.
- Construction des arbres :
  - Arbres avec genres voisins ("cousins").
- Interface web : scripts CGI.

# Visualisation des arbres

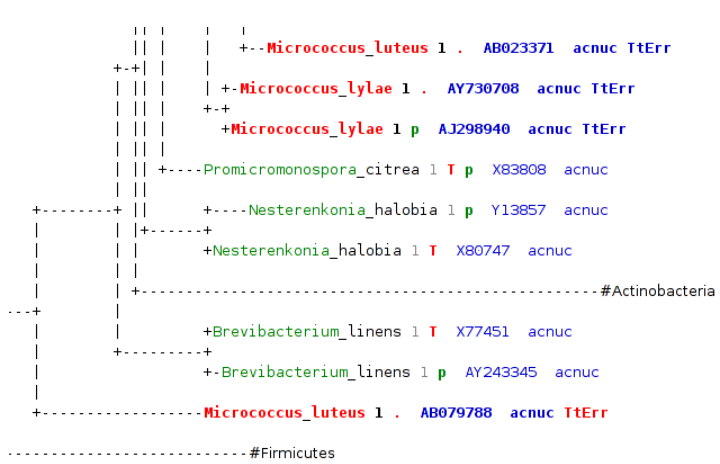
## Visualisation des arbres : NJplot ou scripts CGI


FIG.: Arbres avec genres voisins.

On recherche les mélanges de couleur.

# Expertise et typologie

- Erreur de famille.




- Erreur de genre.

# Résultats de T4Bi

## Analyse visuelle par expert

- Génération d'un ensemble de données à partir de 276 genres.
  - Etude des cinq fragments SSU rDNA.
  - 65 séquences détectées, estimation taux d'incohérence 2.64%
- Différence entre fragments.
  - Mauvaise qualité des séquences : présence de N, chimères.
  - Arbres différents selon le fragment étudié.
  - Incertitude de nomenclature.
- Conclusion
  - Détection manuelle fastidieuse, nécessité d'automatiser.
  - L'automatisation doit retrouver toutes les incertitudes.

# Détection Automatique
## Approche Générale

- Automatisation mise au point sur fragment 46 pour *Actinobacteria* (165 genres).
- Mise en évidence des positionnements douteux, confirmation par un biologiste.
- Algorithme basé sur la topologie des arbres et les distances en noeuds.
- Utilisation des arbres avec cousins :
  - Souches du genre étudié : souches A.
  - Souches des genres cousins : souches nonA.

# Détection Automatique
## Conception

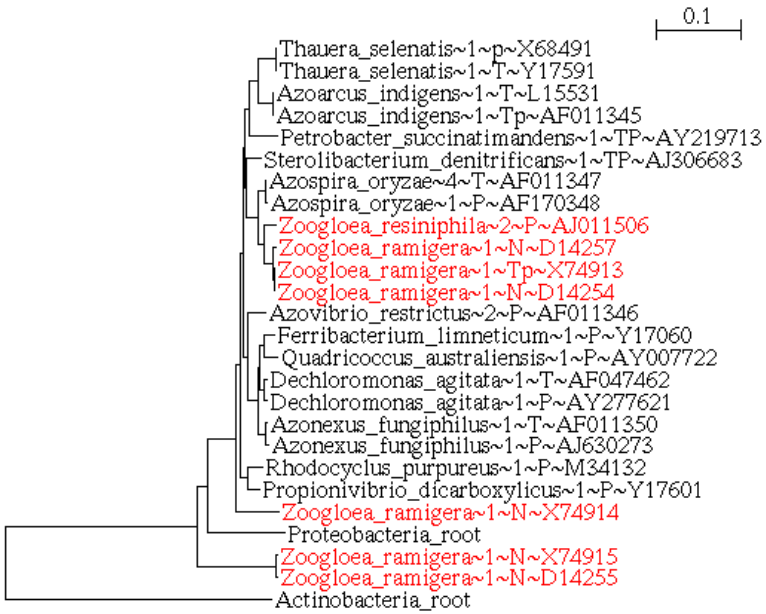
### *Classe Python Tree.*

- Parcours des arbres.
- Calcul des distances.

### *Classe Python TreeModif.*

- Compaction des arbres.
- Recherche des erreurs.
- Retour arbre de départ.

Cas du genre *Zoogloea*.




Arbre de départ du genre.

# Détection Automatique
## Conception

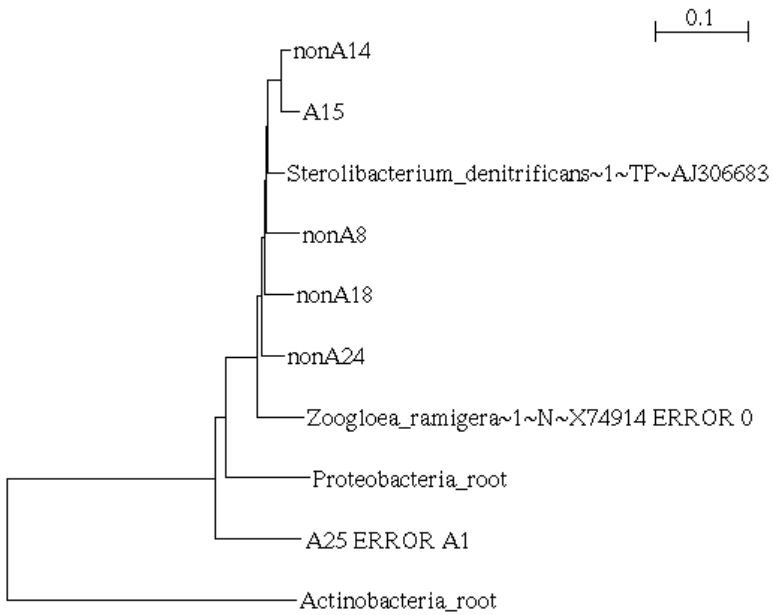
### *Classe Python Tree.*

- Parcours des arbres.
- Calcul des distances.

### *Classe Python TreeModif.*

- Compaction des arbres.
- Recherche des erreurs.
- Retour arbre de départ.

Cas du genre *Zoogloea*.




Arbre compacté.

# Détection Automatique
## Conception

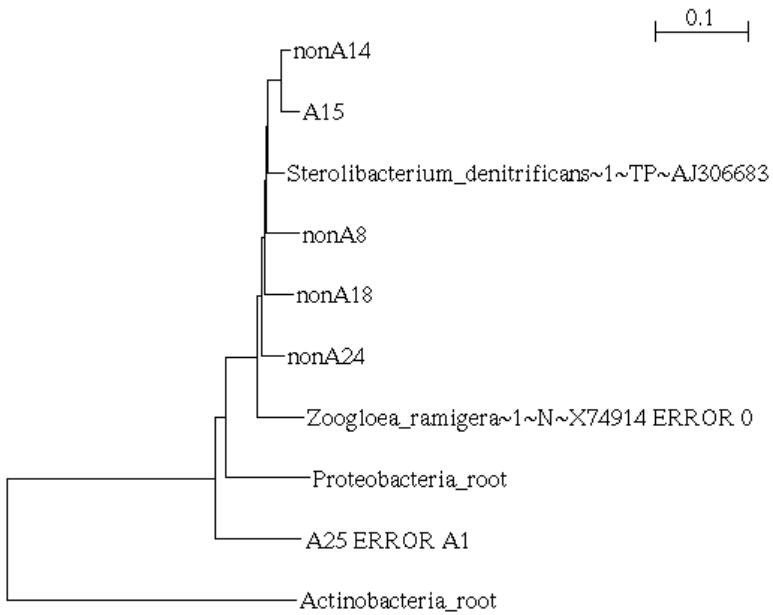
### *Classe Python Tree.*

- Parcours des arbres.
- Calcul des distances.

### *Classe Python TreeModif.*

- Compaction des arbres.
- Recherche des erreurs.
- Retour arbre de départ.

Cas du genre *Zoogloea*.




Arbre compacté.

# Détection Automatique
## Conception

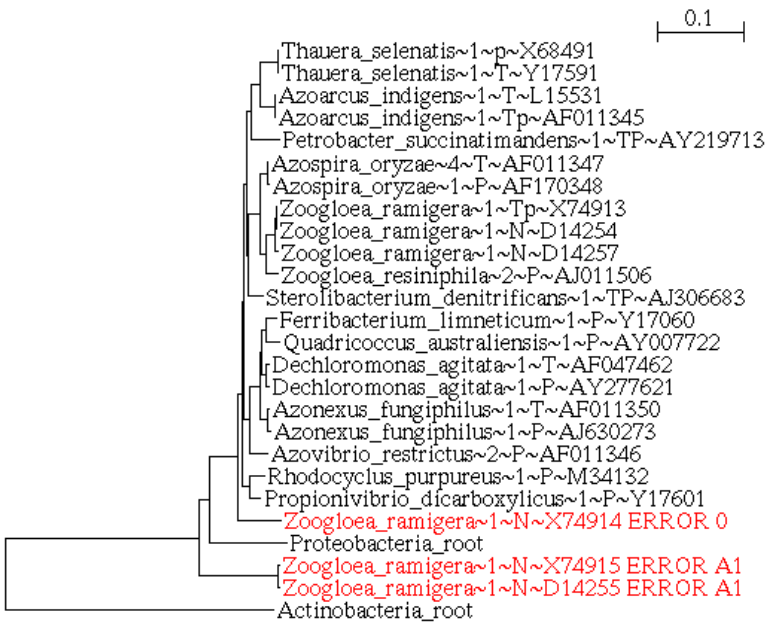
### *Classe Python Tree.*

- Parcours des arbres.
- Calcul des distances.

### *Classe Python TreeModif.*

- Compaction des arbres.
- Recherche des erreurs.
- Retour arbre de départ.

Cas du genre *Zoogloea*.




Arbre final avec anomalies.

# Détection Automatique

## Survol de l'algorithme

Principes :

- Traitement des souches A isolées.
- Traitement des différents groupes de A.
- Distances A-A, A-NonA, A-Racine etc.
- Favorisation de la souche type.
- 5 types d'incongruences taximomie-phylogénie.
- Traitement spécifiques si arbres de mauvaise qualité.
- Indicateur de pertinence des arbres.

Optimisation de l'algorithme sur *Actinobacteria*, minimisation de **Erreurs non détectées/Erreurs Totales** et **Erreurs Totales** acceptable.

# Détection Automatique

## Confrontation à l'expertise

- Application de l'algorithme sur l'ensemble des arbres du fragment 46 de T4Bi : 1010 genres, 4275 espèces, 9129 souches.
- 409 anomalies de positionnement (4.5% des souches).
- L'expert valide et affecte à 4 types :
  - Famille erronée.
  - Genre erroné.
  - Mauvaise qualité de séquence.
  - Anomalie réelle autre cause (taxinomie imparfaite).
- Non validés : cas indécidables selon l'expert.

# Détection Automatique

## Comparaison avec expertise

### Incongruences confirmées par l'expert

| | Détection automatique | % accord |
|---|---|---|
| **0** | 311 | **76** |
| **0T** | 17 | **35.3** |
| **A1** | 13 | **100** |
| **A2** | 68 | **64.7** |
| TOTAL | 409 | **73.1** |

# Perspectives

- Attribution d'un coefficient de confiance aux séquences nucléotidiques, afin de l'utiliser dans BIBI pour l'identification.
- Introduction de connaissance médicale ou biologique.
- Possibilités d'exploiter encore les arbres phylogénétiques :
- Utilisation des classes *Tree* et *TreeModif*, pour d'autres applications (identification et détection automatisée des chimères).

# PhyID-CD

## Phylogenetic Identification Chimera Detection

- Base de séquences préalignées et sélectionnées (T4Bi).
- Alignement par profil de l'inconnu (Muscle) et arbre NJ.
- Identification et detection des séquences chimériques automatisée.
- Ouvert à l'introduction de bases expérimentales.
- Application en bactériologie et virologie (Hépatite C, VIH)

# PhyID-CD

## Phylogenetic Identification Chimera Detection

- Base de séquences préalignées et sélectionnées (T4Bi).
- Alignement par profil de l'inconnu (Muscle) et arbre NJ.
- Identification et detection des séquences chimériques automatisée.
- Ouvert à l'introduction de bases expérimentales.
- Application en bactériologie et virologie (Hépatite C, VIH)

# PhyID-CD

## Phylogenetic Identification Chimera Detection

- Base de séquences préalignées et sélectionnées (T4Bi).
- Alignement par profil de l'inconnu (Muscle) et arbre NJ.
- Identification et detection des séquences chimériques automatisée.
- Ouvert à l'introduction de bases expérimentales.
- Application en bactériologie et virologie (Hépatite C, VIH)

# PhyID-CD

## Phylogenetic Identification Chimera Detection

- Base de séquences préalignées et sélectionnées (T4Bi).
- Alignement par profil de l'inconnu (Muscle) et arbre NJ.
- Identification et detection des séquences chimériques automatisée.
- Ouvert à l'introduction de bases expérimentales.
- Application en bactériologie et virologie (Hépatite C, VIH)

# PhyID-CD

## Phylogenetic Identification Chimera Detection

- Base de séquences préalignées et sélectionnées (T4Bi).
- Alignement par profil de l'inconnu (Muscle) et arbre NJ.
- Identification et detection des séquences chimériques automatisée.
- Ouvert à l'introduction de bases expérimentales.
- Application en bactériologie et virologie (Hépatite C, VIH)

# The BiBi, T4Bi and PhyID group

- BiBi
  - Gregory Devulder
  - Florent Baty,
  - Ghislaine Fardel,
- T4Bi
  - Gregory Devulder
  - Florence Cavalli,
  - Emmanuelle Dantony,
  - Sophie Mignard,
- PhyId-CD
  - Stephane Velay,
  - Antony Cros,
  - Sophie Mignard,
  - Anna-Laura Erbino,
  - Ghislaine Fardel,
- et ... Manolo Gouy (NJplotext), Guy Perrière (BIBI)