Université de Montpellier

Master: Sciences et Numérique pour la Santé

Spécialité: Bioinformatique, Connaissances, Données

Stage de Master

Utilisation de la méthode ARt-DeCo pour réaliser l'annotation fonctionnelle

Frédéric BIGEY

2015–2016

Encadrantes de stage :

Mme Sèverine Bérard
Mme Annie Chateaux

Tutrice pédagogique :

Mme Anne-Muriel Chifolleau

Je tiens tout d'abord à remercier Mesdames Sèverine Bérard et Annie Chateau, enseignantes-chercheuses à l'Université de Montpellier, qui m'ont offert l'opportunité de réaliser ce stage, et grâce aux quelles celui-ci a pu se dérouler dans de bonnes conditions. Merci pour leurs patientes explications sur le fonctionnement d'ARt-DeCo et les discussions autour de la phylogénie.

Un grand merci à Yoann Anselmetti, doctorant à l'Institut des Sciences de l'Évolution de Montpellier, pour l'aide apportée dans la prise en main d'ARt-DeCo.

Enfin, je remercie mon employeur, l'Institut National de la Recherche Agronomique et plus particulièrement le département scientifique MICA et le centre INRA de Montpellier qui ont soutenu et financé cette année cette formation.

Sommaire

Remerciements iii

1 Introduction **1**

1.1 Contexte du stage 1

1.1.1 Présentation du Master 1

1.1.2 Le stage de Master 1

1.1.3 L'Institut des Sciences de l'Évolution de Montpellier (ISE-M) 1

2 Études préliminaires **3**

2.1 Les méthodes d'annotation 3

2.1.1 L'annotation structurale 3

2.1.2 L'annotation fonctionnelle 3

2.2 Présentation d'ART-DECO 4

2.2.1 Description des étapes de la méthode ART-DECO 4

2.3 Les méthodes de recherche d'orthologues 5

2.3.1 Les méthodes de recherche d'orthologues 5

2.3.2 `Forty-Two` 6

2.4 La base de donnée Ensembl Compara 6

2.5 Les méthodes de placements phylogénétiques 6

2.5.1 Principe du placement phylogénétique 6

2.5.2 Pplacer 7

2.6 Objectif du travail à effectuer 7

3 Développements effectués au cours du stage **9**

3.1 Importation et transformation des données d'entrée 9

3.2 À la recherche d'un groupe d'orthologie 12

3.3 Et d'une branche pour greffer la séquence 12

3.4 Mise en œuvre d'ART-DECO 13

4 Conclusion **15**

Bibliographie **17**

Chapitre 1

Introduction

1.1 Contexte du stage

1.1.1 Présentation du Master

Le Master Sciences et Numérique pour la Santé, spécialité Bioinformatique, Connaissances, Donnée, forme des étudiants issus des sciences du vivant (agronomie, biologie moléculaire, cellulaire, physiologie animale et végétale, biochimie), des sciences médicales, de l'informatique ou des mathématiques aux besoins spécifiques de la bioinformatique. Ainsi, il renforce les compétences du profil initial des étudiants dans les domaines de la modélisation du vivant, les systèmes d'information biologiques, l'algorithmique, l'analyse de séquences et de génomes, la prédictions structurales, la phylogénie, ... Il fait une large place à la formation continu en accueillant des étudiants plus âgés souhaitant acquérir ou valider des compétences dans le domaine de la bioinformatique. Pour les personnes ayant un doctorat, ce qui est mon cas, un parcours pédagogique individualisé permet de compléter le Master en une seule année universitaire.

1.1.2 Le stage de Master

Un stage d'une duré de 3 mois doit donner lieu à un rapport. Mon stage s'est déroulé au sein de l'ISE-M du 4 avril au 30 juin 2016 sous l'encadrement de Mme Sèverine Bérard et Mme Annie Chateau toutes deux enseignantes au département d'informatique de l'Université de Montpellier.

1.1.3 L'Institut des Sciences de l'Évolution de Montpellier (ISE-M)

L'institut est situé sur le campus de la faculté des sciences de l'Université de Montpellier dans les bâtiments 22 et 24. Il regroupe des chercheurs du CNRS, de l'IRD et de l'Université de Montpellier, s'intéressant à l'évolution des espèces aussi bien à l'échelle morphologique que moléculaire. Mon stage s'est déroulé au sein de l'équipe de Phylogénie et Évolution Moléculaire qui est dirigée par le Professeur Emmanuel Douzery.

Chapitre 2

Études préliminaires

2.1 Les méthodes d'annotation

L'annotation est l'étape permettant d'identifier sur le génome les différents éléments (ARN, gène, transposon,...) et à leurs attribuer une fonction. L'annotation des génomes est devenue un enjeu de plus en plus important avec l'explosion des projets de séquençage. Mais obtenir une annotation de qualité repose encore sur un processus qui comprend beaucoup d'étapes manuelles.

Traditionnellement l'annotation est constituée de deux étapes principales :

2.1.1 L'annotation structurale

Elle consiste à localiser les divers éléments du génome tels les régions codant les protéines, les ARN non codant (ARNt, ARNr) ainsi que les éléments répétés (transposons, séquence microsatellite,...). Pour les séquences codantes, l'étape suivante est de prédire la structure en intron/exon et d'inférer la séquence de la protéine correspondante. Cette étape est très importante car il serait préjudiciable de rater un domaine dans l'étape ultérieure de prédiction de fonction.

Les outils d'annotation de génomes eukaryotes ont généralement recours à GLIMMER [1] ou AUGUSTUS [2] pour prédire les séquences codantes. Les autres séquences codant pour des ARN sont prédites par des outils tels que tRNAscan-SE [3] pour les ARNt et RNAmmer [4] pour les ARNr.

2.1.2 L'annotation fonctionnelle

Elle consiste à déterminer la fonction biologique. Tout commence par la recherche de gènes homologues au sein des bases de données publiques (UnitProt) ou à l'aide de motifs construits à partir de familles de gènes (méthodes utilisant des motifs ou des signatures ...)

Lorsque des similarités sont observées entre protéines, cela peut indiquer qu'elles sont orthologues, c'est-à-dire que leurs gènes descendent d'une copie ancestrale par spécialisation. Les séquences peuvent provenir d'un événement de duplication (paralogues) ou par transfert horizontal (xénologues). Ces événements évolutifs peuvent être étudiés à l'aide des phylogénies des gènes et des espèces.

Deux approches peuvent être utilisées pour annoter un génome :

- l'annotation dite *de novo* qui utilise l'homologie de séquences avec des séquences présentes dans des bases de données. Cette stratégie est particulièrement adaptée pour annoter des génomes d'organismes inconnus ou phylogénétiquement éloignés. Ces outils sont adaptés aux analyses métagénomiques comportant différentes espèces.
- le transfert automatisé d'annotations par comparaison avec un ou plusieurs génomes de référence déjà annotés provenant d'espèces phylogénétiquement proches. Les résultats obtenus dépendront donc fortement du choix du ou des génomes de références et de la qualité de leurs annotations.

2.2 Présentation d'ARt-DeCo

DeCo propose une méthode de reconstruction de la structure des génomes ancestraux dans un contexte phylogénétique [5]. Cette méthode tient compte de trois événements évolutifs au niveau des gènes : spéciation, duplication et perte. DeCo est une méthode efficace car c'est une méthode exacte et de faible complexité. à détailler...

Elle implémente un algorithme d'optimisation reposant sur le critère du maximum de parcimonie. Il utilise une méthode de programmation dynamique avec l'objectif de minimiser le coût de construction des arbres des adjacences. à définir...

ARt-DeCo [6] a été développé à partir de DeCo afin de proposer une méthode permettant d'améliorer le *scaffolding* à l'issu de l'assemblage d'un génome. ARt-DeCo utilise le contexte phylogénétique afin d'imputer la présence de cassures au niveau du génome assemblé. ARt-DeCo comme DeCo ont été écrits en langage C++, à l'aide de la librairie Bio++ [7].

2.2.1 Description des étapes de la méthode ARt-DeCo

Le programme est modulaire, chaque module permettant de réaliser une étape de l'analyse :

- La phase de création des fichiers d'entrée (module `Step0_dataset`) ;
- La phase de réconciliation (module `Step1_reconciliation`) ;
- La phase de création des classes d'équivalences (module `Step2`) ;
- La phase de création des arbres d'adjacences (module `Step3_proba`) ;
- La phase de synthèse de résultats (module `Step4`).

La phase de création des fichiers d'entrée

Cette étape permet de se procurer les données à partir de la banque Ensembl. Le module `Step0_dataset` se charge de télécharger les fichiers nécessaires à partir d'une liste d'espèces. Dans un second temps, les fichiers récupérés servent à produire les fichiers nécessaires à l'analyse.

Au total quatre fichiers nécessaires aux étapes suivantes sont disponibles :

- Le fichier gène–espèce :
  Ce fichier contient à chaque ligne l'identifiant d'un gène et son espèce associée ;
- Le fichier contenant les arbres de gène :
  Ce fichier contient tous les arbres de gène, un par famille, au format newick par exemple ;
- Le fichier de l'arbre des espèces au format newick ;

- Le fichier des adjacences :
  Ce fichier indique à chaque ligne un couple de gènes voisins sur un génome.

Dans notre cas, cette étape ne sera pas utile, les trois fichiers devant être générés de façon différente (*cf* partie 3).

La phase de réconciliation

Cette étape consiste à réconcilier les arbres de gènes avec la topologie de l'arbre des espèces. en utilisant la réconciliation parcimonieuse LCA élaborée par [8].

La phase de création des classes d'équivalences

Une fois l'attribution des événements de spéciations, duplications et pertes de gènes aux arbres de gènes réconciliés, le module `Step2` crée des classes d'équivalence d'adjacences à partir des arbres de gènes et de la liste d'adjacence. Une classe d'équivalence d'adjacences permet de grouper les adjacences qui ont potentiellement une adjacence ancestrale commune.

La phase de création des arbres d'adjacences

Pour chaque classe d'équivalence le module `Step3_proba` calcule une matrice de scores pour chaque couple de gènes des arbres de gènes. Pour chaque couple, deux coûts minimum sont calculés suivant les deux histoires évolutives possibles : gènes adjacents ou non adjacents. Après avoir calculé cette matrice de scores, l'algorithme effectue une étape de *backtracking* qui parcourt la matrice afin de construire la forêt d'arbres d'adjacence, correspondant à l'histoire des adjacences qui minimise le critère de coût.

Les fichiers de sortie

À l'issue des calculs les informations suivantes sont disponibles dans les fichiers de sortie :

- Les adjacences ancestrales par espèces,
- Les gènes qui se sont dupliqués ou perdus ensemble,
- Les arbres d'adjacences,
- Les degrés des gènes ancestraux (nombre de voisins),
- Taille des arbres d'adjacences.

2.3 Les méthodes de recherche d'orthologues

2.3.1 Les méthodes de recherche d'orthologues

Une des applications principale de l'orthologie est la propagation d'annotation fonctionnelle, car les orthologues sont sensés avoir très souvent une fonction semblable. En fait, cet usage est si répandu que beaucoup d'auteurs utilisent abusivement le terme orthologue pour désigner des gènes ayant une fonction conservée. Les méthodes permettant de déterminer les relations d'orthologie se divisent en trois familles : celles basées sur les arbres, ceux faisant appel aux graphes ou chacun des deux (méthodes hybrides). D'un point de vue pratique, cette classification

distingue également les méthodes *ab initio* de celles faisant appel à des groupes d'orthologie précalculés afin d'imputer de nouveaux orthologues.

Les méthodes utilisant les graphes

Principe assez simple basé sur la comparaison des séquences deux à deux (BLAST, ...) puis une étape de clustering reposant sur les relations entre gènes. Les méthodes les plus connues sont : InParanoid [9], OrthoMCL [10], Forty-Two [11], ...

Les méthodes utilisant les arbres phylogénétiques

Ces méthodes sont plus lourdes en calcul car elles reposent sur la construction d'arbres phylogénétiques et la réconciliation avec l'arbre des espèces. Elles sont souvent considérées comme plus précises que les méthodes de comparaison, mais cette affirmation n'a jamais été démontrée formellement. Nous pouvons citer : TreeBeST [12], TreeFam [13], ...

2.3.2 `Forty-Two`

`Forty-Two` est un outil développé par Denis Baurain (Faculté des Sciences de Liège, Belgique) sur lequel un utilisateur de l'institut avait acquit une certaine expérience. Il a été développé afin d'ajouter et d'aligner des séquences à des alignements multiples.

2.4 La base de donnée Ensembl Compara

La base de données Ensembl Compara est une ressource unique pour les étude de génomique comparative [14]. Elle met à disposition des alignements, des arbres de gène, des groupes d'orthologues, des informations sur la synthénie. La version 31 disponible au début du stage rassemble 589 espèces de champignons et levures (Ensembl Fungi, release 31, téléchargement `ftp://ftp.ensemblgenomes.org/pub/fungi/release-31/emf/ensembl-compara/`)

2.5 Les méthodes de placements phylogénétiques

2.5.1 Principe du placement phylogénétique

Sachant le groupe d'orthologues connu, comment placer une séquence sur l'arbre phylogénétique correspondant ? On rencontre souvent cette question en métagénomique lorsque l'on souhaite inférer l'origine phylogénétique d'une séquence (ARNr 18 S, 26 S). Les méthodes placements phylogénétiques reposent actuellement sur le principe du maximum de vraisemblance. Des heuristiques sont nécessaires car le problème est NP-dur et le temps de calcul devient prohibitif sur des jeux de données comprenant un grand nombre de taxons. Dans notre cas, ce n'est vraiment pas un problème car la grande majorité des arbres de références possèdent moins de 10 feuilles.

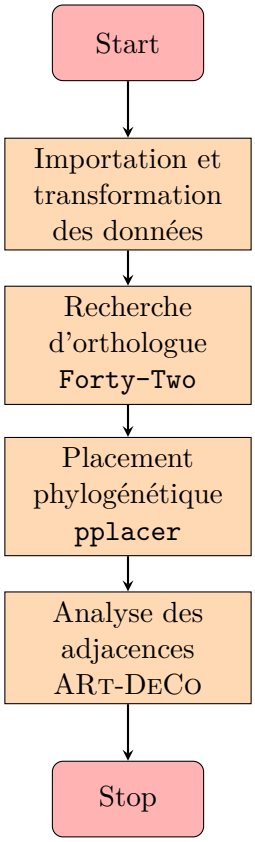

FIGURE 2.1 – Les étapes successives du projet

2.5.2 Pplacer

Pplacer est un programme de placement phylogénétique de séquences sur un arbre phylogénétique de référence. La séquence cible à placer est apportée dans un alignement multiple. Utilisant le principe du maximum de vraisemblance (ML), `pplacer` essaie de trouver l'emplacement et la longueur de branche qui maximisent la vraisemblance de l'arbre. En mode bayésien, `pplacer` est capable de calculer la probabilité postérieure de chaque placements, ce qui permet d'en évaluer l'incertitude. En outre, `pplacer` permet de travailler en temps et en empreinte mémoire linéaires [15].

2.6 Objectif du travail à effectuer

Au cours de ce projet nous nous proposons de construire un pipeline permettant de tester ART-DECO dans le cadre d'un transfert d'annotation. Les objectifs après l'étude préliminaire sont résumés à la FIGURE 2.1.

Chapitre 3

Développements effectués au cours du stage

Les briques logiciels étant à présent sélectionnées, il reste à construire le pipeline puis le tester.

3.1 Importation et transformation des données d'entrée

Sélection des données d'entrée

L'étude des données Ensembl accessibles a révélé que les informations étaient éparpillées dans différents fichiers (alignements multiples, arbres phylogénétiques, ...). En outre, les séquences étaient tantôt identifiées par leur gène (alignements) tantôt par leur protéine (arbres phylogénétiques).

Nous avons constaté que seul les arbres phylogénétiques disponibles au format `PhyloXML` [16] permettraient d'obtenir l'ensemble des informations nécessaires : arbre phylogénétique, alignement multiple des séquences et identifiant du groupe d'orthologie. Une archive contenant l'ensemble des arbres a été téléchargées à partir du serveur d' Ensembl Fungi (release 31) à l'adresse suivante : `ftp://ftp.ensemblgenomes.org/pub/fungi/release-31/emf/ensembl-compara/homologies/Compara.phyloxml_aa_trees.31.tar.gz`

Pour la phase de mise au point, nous avons sélectionner 18 espèces de levures sur les 589 espèces de levures et champignons que comporte la base Ensembl Fungi (Table 3.1). L'arbre des espèces est présenté à la Figure 3.1. La levure *Saccharomyces eubayanus* sera utilisé comme génome de test.

Modification des arbres phylogénétiques

Chaque arbres phylogénétiques au format `PhyloXML` est traité afin de ne conserver que les séquences provenant des 18 espèces de levures. L'arbre résultant est sauvegardé au format Newick (appelé également format New Hampshire). L'alignement multiple des séquences correspondantes est quand à lui sauvegardé dans le format Fasta. Ce traitement est schématisé à la Figure 3.2.

TABLE 3.1 – Les espèces de levures utilisées lors de la phase de mise au point

Ashbya gossypii
Brettanomyces bruxellensis AWRI1499
Candida albicans SC5314
Candida dubliniensis CD36
Candida glabrata
Candida tropicalis MYA-3404
Debaryomyces hansenii CBS767
Kazachstania naganishii CBS 8797
Kluyveromyces lactis
Lachancea thermotolerans CBS 6340
Lodderomyces elongisporus NRRL YB-4239
Saccharomyces cerevisiae
Saccharomyces kudriavzevii IFO 1802
Torulaspora delbrueckii
Vanderwaltozyma polyspora DSM 70294
Yarrowia lipolytica
Zygosaccharomyces bailii ISA1307
Zygosaccharomyces rouxii

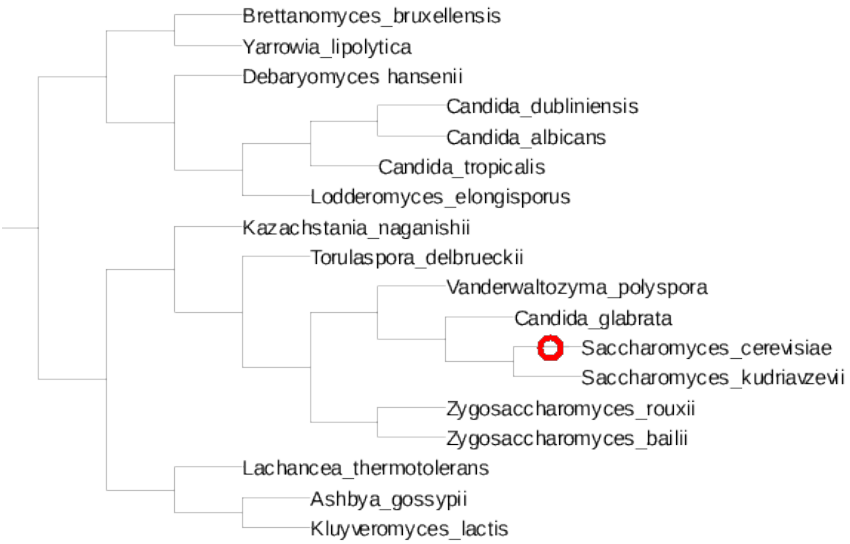

FIGURE 3.1 – Topologie de l'arbre phylogénétique des 18 espèces de levures utilisées pour l'étude. Le cercle rouge indique le branchement de la souche test *Saccharomyces eubayanus*

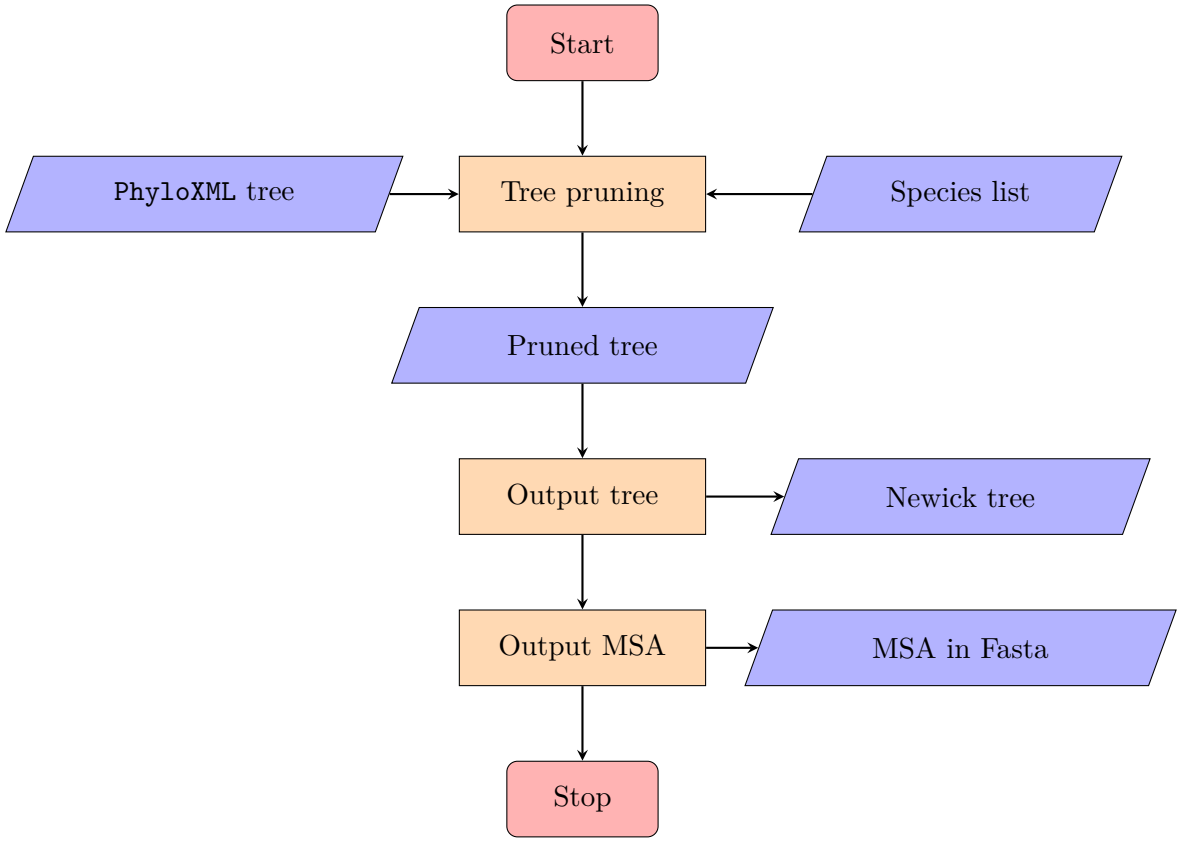

FIGURE 3.2 – Traitement des arbres phylogénétiques

Résultat

Le bilan de cette étape est présenté à la TABLE 3.2. Nous obtenons 10377 arbres dont 58% correspondent à des groupes d'orthologues à au moins 3 séquences. Les 42% restant comportant moins de 3 séquences seront traités différemment car ils ne nécessitent pas d'étape de placement phylogénétique.

TABLE 3.2 – Bilan des arbres générés à partir des fichiers `PhyloXML`

	Nombre d'arbres
téléchargés	123206
après transformation	10377
dont nombre d'arbres à	
1 séquences	3274
2 séquences	1104
$\geq$ 3 séquences	5999

TABLE 3.3 – Répartition des groupes d'orthologie en fonction du nombre de séquences accueillies

Nombre de séquences	Nombre de groupe
0	186
1	588
2	11
3	1

3.2 À la recherche d'un groupe d'orthologie

Mise en œuvre de `Forty-Two`

Dans un premier temps, `Forty-Two` nécessite que les identifiant des séquences soient représentés sous une forme bien précise dont voici un exemple :

```
>Saccharomyces_cerevisiae@YOR203
```

Dans un second temps, nous sommes attachés à paramétrer le fichier de configuration de `Forty-Two`. Ce dernier, au format `YAML`, permet de préciser l'emplacement des séquences de références. Lors de cette étape, nous en profitons pour produire les fichiers d'index nécessaires au fonctionnement de `BLAST`.

Dans un dernier temps, nous avons procédé à la conversion des MSA produits, `Forty-Two` utilisant son propre format d'alignements.

Résultats

Afin d'accélérer l'obtention de résultats, nous avons réaliser les tests sur un jeux réduit à 786 groupes d'orthologie. Le temps de traitement est de 9 heures. Nous avons rencontré plusieurs cas de figures : des groupes sans séquence, une séquence par groupe, plusieurs séquences par groupe. La répartition est présentée à la TABLE 3.3. Nous constatons qu'un quart (186) des groupes d'orthologie n'ont pas accueilli de séquences. Ceci est normal sachant que certaines familles de gènes ne se retrouvent pas dans tout les taxons.

3.3 Et d'une branche pour greffer la séquence

Mise en œuvre de `pplacer`

Afin de faire fonctionner `pplacer`, plusieurs étapes sont nécessaires afin d'adapter nos données au format d'entré du programme :

étape 1 : Convert reference MSA into `PHYLIP` format (custom script : `fasta2phylip.sh`)

étape 2 : Creating `RAxML` information file from tree and alignment (`raxmlHPC`)

étape 3 : Create package file (`*.pkg`) containing tree and alignment (`taxit create`)

étape 4 : Reformat definition line of `Forty-Two` output file (`sed`)

étape 5 : Place query sequence on tree using `pplacer` (`v1.1.alpha17`)

étape 6 : Make one tree for each query sequence (`guppy sing`)

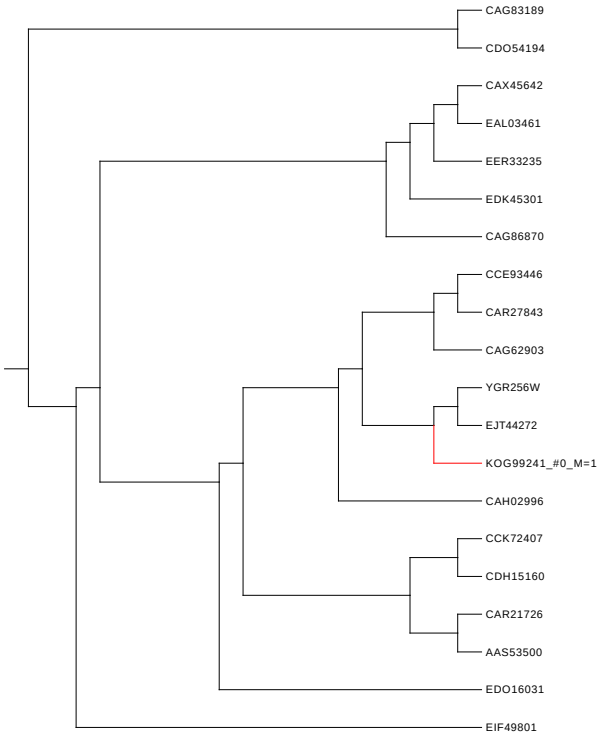

FIGURE 3.3 – Placement de la séquence `KOG99241` de *Saccharomyces eubayanus* sur le groupe d'orthologie `EFGT00050000001269` d'Ensembl Fungi

Résultats

`pplacer` a mis environ 1 heure pour traiter les 600 arbres phylogénétiques obtenus par `Forty-Two`. Un exemple de placement est présenté à la FIGURE 3.3.

3.4 Mise en œuvre d'ARt-DeCo

Génération du jeu de données

ART-DECO nécessite les 4 fichiers suivants dont le contenu a été décrit précédemment (voir section 2.2.1) :

- Le fichier gène–espèce : `species_gene.tab` ;
- Le fichier contenant les arbres de gène : `all_tree.nwk` ;
- Le fichier de l'arbre des espèces au format newick : `species_tree.nwk` ;
- Le fichier des adjacences : `adjacencies.tab` ;

Les fichiers gène–espèce et d'adjacences sont générés grâce à un script permettant d'extraire les informations du fichier `Compara.newick_trees.31.emf.gz` téléchargé d'Ensembl Fungi (`ftp://ftp.ensemblgenomes.org/pub/fungi/current/emf/ensembl-compara/homologies/Compara.newick_trees.31.emf.gz`). Le fichier contenant les arbres de gène est construit par concaténation de arbres obtenus par `pplacer`. Quand à l'arbre des espèces il est construit manuellement à partir des données issues de la littérature.

Résultats

Par manque de temps, je n'ai pas été en mesure de faire tourner ART-DECO sur le jeux de test préparé.

Chapitre 4

Conclusion

Le bilan de ce stage est très positif car j'ai pu mettre en pratique les connaissances acquises en cours de durant cette année de formation.

Il reste encore un gros travail de test à chaque étapes du pipeline :

- validation de l'ajout des séquences cibles dans les groupes d'orthologie Ensembl (étapte `Forty-Two`)
- validation du branchement des séquences dans les arbres phylogénétiques (étape `pplacer`)
- validation du fonctionnement d'ARt-DeCo sur un génome complet

Bibliographie

[1] Steven L. Salzberg, Arthur L. Delcher, Simon Kasif, and Owen White. Microbial gene identification using interpolated Markov models. *Nucleic Acids Research*, 26(2) :544–548, January 1998.

[2] Mario Stanke, Rasmus Steinkamp, Stephan Waack, and Burkhard Morgenstern. AUGUSTUS : a web server for gene finding in eukaryotes. *Nucleic Acids Research*, 32(Web Server issue) :W309–W312, July 2004.

[3] T M Lowe and S R Eddy. tRNAscan-SE : a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Research*, 25(5) :955–964, March 1997.

[4] Karin Lagesen, Peter Hallin, Einar Andreas Rødland, Hans-Henrik Stærfeldt, Torbjørn Rognes, and David W. Ussery. RNAmmer : consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, 35(9) :3100–3108, May 2007.

[5] Sèverine Bérard, Coralie Gallien, Bastien Boussau, Gergely J. Szöllősi, Vincent Daubin, and Eric Tannier. Evolution of gene neighborhoods within reconciled phylogenies. *Bioinformatics*, 28(18) :i382–i388, September 2012.

[6] Yoann Anselmetti, Vincent Berry, Cedric Chauve, Annie Chateau, Eric Tannier, and Sèverine Bérard. Ancestral gene synteny reconstruction improves extant species scaffolding. *BMC Genomics*, 16(Suppl 10) :S11, October 2015.

[7] Julien Dutheil, Sylvain Gaillard, Eric Bazin, Sylvain Glémin, Vincent Ranwez, Nicolas Galtier, and Khalid Belkhir. Bio++ : a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics. *BMC Bioinformatics*, 7 :188, 2006.

[8] Morris Goodman, John Czelusniak, G. William Moore, A. E. Romero-Herrera, and Genji Matsuda. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Zoology*, pages 132–163, 1979.

[9] Maido Remm, Christian E. V. Storm, and Erik L. L. Sonnhammer. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons1. *Journal of Molecular Biology*, 314(5) :1041–1052, December 2001.

[10] Li Li, Christian J. Stoeckert, and David S. Roos. OrthoMCL : Identification of Ortholog Groups for Eukaryotic Genomes. *Genome Research*, 13(9) :2178–2189, January 2003.

[11] Denis Baurain. User guide for 42, November 2015.

[12] Albert J. Vilella, Jessica Severin, Abel Ureta-Vidal, Li Heng, Richard Durbin, and Ewan Birney. EnsemblCompara GeneTrees : Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, 19(2) :327–335, January 2009.

[13] Heng Li, Avril Coghlan, Jue Ruan, Lachlan James Coin, Jean-Karim Hériché, Lara Osmotherly, Ruiqiang Li, Tao Liu, Zhang Zhang, Lars Bolund, Gane Ka-Shu Wong, Weimou Zheng, Paramvir Dehal, Jun Wang, and Richard Durbin. TreeFam : a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Research*, 34(suppl 1) :D572–D580, January 2006.

[14] Arnold Kuzniar, Roeland C. H. J. van Ham, Sándor Pongor, and Jack A. M. Leunissen. The quest for orthologs : finding the corresponding gene across genomes. *Trends in Genetics*, 24(11) :539–551, November 2008.

[15] Frederick A. Matsen, Robin B. Kodner, and Virginia E. Armbrust. pplacer : linear time maximum-likelihood and Bayesian phylogenetic placement of sequences onto a fixed reference tree. *BMC Bioinformatics*, 11 :538, 2010.

[16] Mira V. Han and Christian M. Zmasek. phyloXML : XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, 10 :356, 2009.

Composé avec LaTeX