ENS de Lyon
Licence Informatique

Spécialité Informatique Fondamentale

Rapport de stage :

Ordonner de Grandes Phylogénies

Julien BRAINE

Responsables de l'UE :

M. Colin Riba ENS Lyon

Tuteur :

M. Damien Stehlé ENS Lyon

Responsables du stage :

M. Fabio Pardi LIRMM Montpellier
Mme. Céline Scornavacca ISEM Montpellier

Soutenue le : 2 Septembre 2014

Résumé

L'utilisation d'algorithmes pour afficher des données et faciliter leur interprétation est un problème qui se pose dans de nombreuses disciplines, allant de l'affichage de graphe en informatique à l'affichage de données statistiques. Durant mon stage de Licence 3, je me suis interéssé à l'affichage d'arbres binaires dont les feuilles sont etiquetées ; l'idée consiste à grouper ensemble les feuilles similaires tout en respectant la structure de l'arbre dans l'objectif de faciliter une analyse phylogénétique ou d'expression de gènes. Dans un premier temps, j'ai cherché un algorithme rapide pour maximiser la similarité des feuilles adjacentes [Lien : Bar Joseph] et j'ai proposé un algorithme en $O(nk^2)$ où k désigne le nombres d'etiquettes différentes. Dans un deuxième temps, je me suis attaché à identifier la représentation attendue par les biologistes.

Table des matières

Introduction

L'utilisation d'algorithmes pour afficher des données et faciliter leur interprétation est un problème qui se pose dans de nombreuses disciplines, allant de l'affichage de graphe en informatique à l'affichage de données statistiques en passant par les dessins industriels. En biologie, ce problème se pose souvent, notamment en phylogénétique et en analyse d'expression de gènes.

En phylogénétique, la donnée est souvent un arbre représentant l'évolution. Chaque feuille représente un organisme échantillonné et les nœuds internes un ancêtre fictif. A chaque nœud interne correspond une spéciation, c'est à dire la division d'un organisme en plusieurs autres organismes. La méthode pour construire l'arbre à partir des séquences d'ADN donne un arbre binaire enraciné [1] [BOOK : inferring phylogenies]. Une fois l'arbre construit, il peut être intéressant de trouver les corrélations morphologiques ou géographiques avec l'évolution de l'organisme. Ainsi, afficher l'arbre de façon à faciliter son analyse est primordial.

En analyse d'expression de gène, les données sont généralement une matrice dont les lignes sont les différents gènes, et les colonnes divers échantillons. Chaque élément de la matrice contient le taux d'expression du gène de l'échantillon considéré [2]. Une analyse d'expression de gène peut avoir plusieurs objectifs. Par exemple, on peut vouloir déterminer quelles parties du corps ont les même comportements. Pour cela, on prelève divers échantillons sur le corps humain et on cherche à grouper les échantillons ayant les même profils d'expression de gène ensemble. La méthode utilisée pour classer les données consiste à faire un clustering hiérarchique [3], nous permettant d'avoir un arbre binaire, puis à ordonner cet arbre binaire [Therese Biedl].

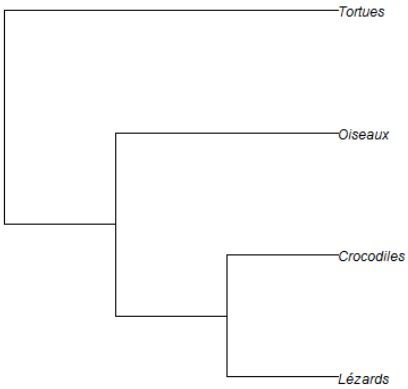
Mon stage consiste, étant donné un arbre binaire etiqueté aux feuilles et une distance entre les paires d'étiquettes, à trouver la représentation de l'arbre qui groupe ensemble les feuilles similaires.

1. Les données ADN seules ne permettent pas d'enraciné l'arbre. On utilise des meta-informations comme "'toutes les feuilles doivent être à égale distance de la racine.

2. Le taux d'expression d'un gène dans un échantillon est la quantité d'ARN ou de protéine que ce gène produit dans cet échantillon.

3. Un clustering hiérarchique peut être ascendant ou descendant. Lorqu'il est ascendant, chaque donnée est initialement dans un cluster différent, puis à chaque étape on groupe les paires de données les plus proche. Lorsqu'il est descendant, toutes les données sont dans un même cluster, puis on divise le cluster à chaque étape.

Exemple de données : un arbre phylogenetique et une matrice de distance



	Lézards	Crocodiles	Oiseaux	Tortues
Lézards	0	1	10	1
Crocodiles	1	0	10	2
Oiseaux	10	10	0	10
Tortues	1	2	10	0

Une représentation d'un arbre ne doit pas changer la nature de l'arbre. C'est à dire que les représentations d'un arbre sont engendrées par les permutations des fils droits et fils gauches à chaque noeud. Il y a donc 2^{n-1} représentations d'un même arbre, où n désigne le nombre de feuilles. Ensuite, ce qui fait qu'une représentation est meilleure qu'une autre, c'est l'ordre des etiquettes sur les feuilles puisque l'objectif est de regrouper les feuilles similaires ensemble.

Un sous problème est de se restreindre au cas des "'couleurs"'. C'est à dire qu'on a un nombre fini d'étiquettes n'ayant rien en commun ; c'est en particulier possible dans le cas de données phylogénétiques où chaque "'couleur"' peut être, par exemple, un continent. La notion de distance entre les etiquettes est alors binaire : 1 si les deux etiquettes sont différentes et 0 sinon.

Mon stage s'est déroulé en deux temps. J'ai d'abord chercher un algorithme qui minimise la somme des distances entre les feuilles adjacentes dans le cas général et dans le cas des couleurs.Puis j'ai cherché une définition formelle de ce qu'est la meilleure représentation dans le cas des couleurs.

Quelques Notations

Definition : un arbre binaire enraciné avec feuilles étiquetées est un triplet(Squelette, id, data) où squellette est un arbre binaire enraciné sous forme de graphe orienté décrit par des ensembles d'adjacence, id une bijection de l'ensemble des noeuds dans [1, nbNoeuds] et data Definition d'un arbre binaire
Definition d'une représentation d'un arbre binaire
Definition d'une methode
Definition de closest
Definition du cas des couleurs

Chapitre 1

L'approche Closest

1.1 La Cas Général

1.1.1 Etat de l'art

Presentation de l'algorithme de Bar-Joseph Presentation de l'amélioration de Therese Biedl Presentation de l'algorithme sur les arbres complets

1.1.2 Quelques idées

Presentation de l'algorithme du catapillar
Generalisation de l'algorithme sur les arbres complets
Essaie de mélange
Echec sur l'arbre catapillar dont les feuilles sont des arbres complets

1.2 Le cas des couleurs

1.2.1 Un cas très différent

Le cas de l'arbre étoile Bornes de complexités

1.2.2 Un algorithme simple

Presentation de l'algorithme de Therese Biedle dans le cas des couleurs

1.2.3 Analyse de complexité

Demonstration du $O(nk^2)$

1.2.4 Limites

Exemples ou Closest ne donne pas le resultats voulu

Chapitre 2

De nouvelles approches

2.1 Des propriétés nécessaires

Nouvelle idée, étendre la relation d'ordre à toutes les permutations

2.1.1 Quelques notations

Definition de X
Equivalence distance/relation

2.1.2 Les propriétés

Propriété changement de lettre
Propriété decoupage + rev + deplacement + changement d'ordre
Propriété globale faible
Propriété globale forte
Propriété sorted
Propriété scale

2.1.3 Des propriétés non-suffisantes

Beaucoup d'approche subsistent. Exemples

2.2 Des recherches par des exemples

Tableau des approches/propriétés/contre-exemple Le contre exemple général ?

2.3 Integrale/Breaks

2.3.1 Les différents cas

Les approches $integrale(i-j)^\alpha$ Le cas $\alpha = 1$
Le cas $\alpha > 0$
Le cas $\alpha = 0$
Le cas $0 > \alpha > -2$ et $\alpha \neq -1$
Le cas $\alpha = -1$
Le cas $\alpha < -2$

2.3.2 Impression experimental

Ce que l'on prefere

Conclusion et Perspective

Generalisation de $integrale(i-j)^{\alpha}$ en perspective.