

Projet de programmation PI4 2021/2022 : PhyloApp on Planet 622

Proposé par Alice Rogier

alice.rogier@inserm.fr

1 Contexte

Nous sommes en 24762. Une infime partie de l'humanité a survécu dans un vaisseau et découvre une planète habitable. Vous êtes à bord de ce vaisseau et vous faites partie de l'équipe de recherche en exobiologie. On vous charge de créer une interface graphique permettant d'étudier l'évolution des espèces vivant sur votre nouvelle planète.

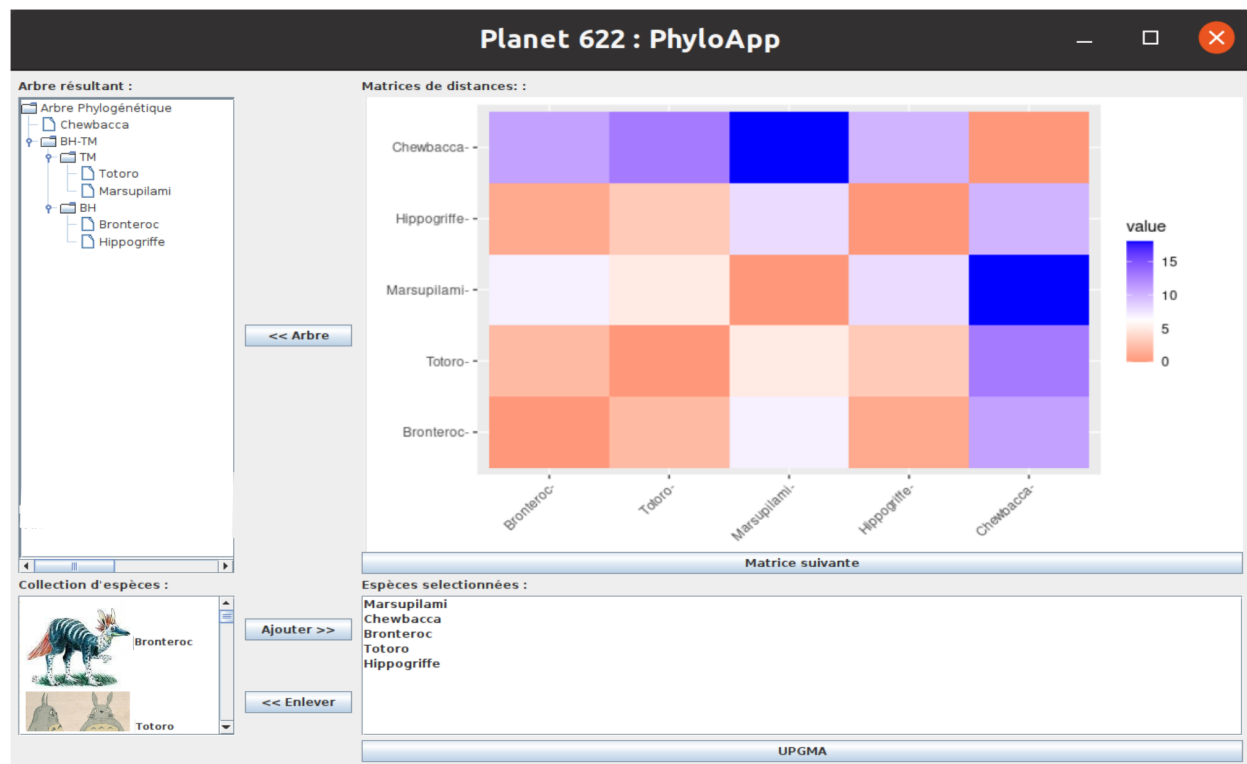


Figure 1: Proposition de présentation de l'interface

On s'appuiera sur la figure 1 pour expliquer le principe de l'algorithme UPGMA dans la section 3

2 Construction d'arbre phylogénétique avec la méthode UPGMA

2.1 Arbre Phylogénétique

Pour comprendre l'évolution des espèces, les biologistes construisent des arbres phylogénétiques. Ces arbres montrent les relations de parenté d'un groupe d'espèces donné. Si deux espèces (feuilles de l'arbre) descendent d'un même noeud, alors elles sont proches entre elles. Elles ont le même ancêtre commun (noeud qui les relie). Pour construire les arbres phylogénétiques, les biologistes comparent des séquences de gènes très conservés en les alignant. Un gène très conservé est un gène qui code pour une protéine dont l'activité est essentielle et présente chez toutes les espèces animales. Cette protéine étant essentielle, la séquence des gènes les codant va très peu se modifier au cours du temps. Les mutations dites silencieuses (qui ne modifient pas la fonction de la protéine) vont s'accumuler chez des espèces différentes au cours du temps. Ainsi, plus les espèces sont éloignées en terme d'évolution, plus les séquences de ces gènes conservés sont différentes. La similarité des séquences de ces gènes permet de déduire les liens de parenté entre les espèces et de construire l'arbre. Plusieurs techniques existent, ici on se focalise sur la méthode UPGMA (Unweighted Pair Group Method with Arithmetic Mean).

2.2 Algorithme UPGMA

On veut créer un arbre phylogénétique schématisant les liens de parenté de n gènes conservés g_0, \dots, g_{n-1} . Pour cela, on calcule d'abord la matrice des distances M entre ces gènes (grâce à de l'alignement de séquence). Chaque cellule $M[i][j]$ contient un nombre décimal qui correspond à la distance euclidienne entre les gènes g_i et g_j (M est donc une matrice symétrique, avec sa diagonale nulle).

2.2.1 Idée

On construit un arbre phylogénétique de façon itérative :

- Au départ, les feuilles de l'arbre sont les n gènes.
- À chaque étape de l'algorithme, on regroupe les deux gènes (ou groupe de gènes) g_i et g_j les plus proches afin de former un nouveau groupe de gènes. Le nouveau groupe est représenté comme un noeud de l'arbre dont les fils sont g_i et g_j .
- On itère ce processus jusqu'à ce qu'il ne reste plus que deux groupes de gènes. On les rassemble pour former la racine de l'arbre.

Cette vidéo explique très bien le principe de l'algorithme, je vous conseille de la regarder : <https://www.canal-u.tv/chaines/inria/5-arbres-phylogenetiques/54-1-algorithme-upgma>

2.2.2 Détail

Soit n le nombre de gènes (=le nombre d'espèces).
Soit M la matrice de distances de dimension $n \times n$.

Tant qu'il reste plus de deux groupes de gènes (= que la dimension de la matrice est supérieure à 2*2):

- Trouver les indices min_i et min_j qui correspondent au minimum de la matrice M . Créer un nouveau noeud à l'arbre qui a pour enfant les gènes ou groupes de gènes associés à min_i et min_j .
- Mettre à jour la matrice M c'est-à-dire :
 - créer une matrice temporaire M_{tmp} de dimension $n - 1 * n - 1$
 - remplir M_{tmp} : M_{tmp} est égale à M sans les colonnes et les lignes d'indices min_i et min_j . La dernière ligne et colonne M_{tmp} reflète la distance entre le groupement de gène min_i et min_j avec les autres gènes ou groupe de gènes. Elles sont égales à la moyenne arithmétique des deux distances des éléments à regrouper avec les autres gènes/groupes de gènes.
 - $M = M_{tmp}$

Pour mieux comprendre l'algorithme, lisez l'exemple dans la section suivante.

3 Exemple

Pour simplifier, on représentera les séquences des gènes par des entiers strictement positifs. Sur la planète 622, on a séquencé un gène homologue à 5 espèces :

Espèce	Séquence du gène
Bronteroc (b)	14
Totoro (t)	12
Marsupilami (m)	7
Hippogriffe (h)	15
Chewbacca (c)	25

Dans un premier temps, on calcule la matrice des distances euclidiennes de ces gènes.



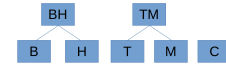
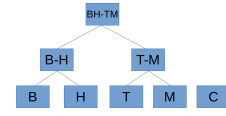
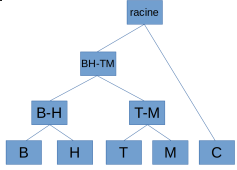
$$d(g_i, g_j) = \sqrt{(g_i - g_j)^2} \quad (1)$$

Donc par exemple $d(b, m) = \sqrt{(14 - 7)^2} = 7$

À chaque itération, on trouve le minimum de la matrice. On crée un nouveau groupement d'espèces avec ce minimum. Dans l'exemple, à la première itération, on regroupe le bronteroc et l'hippogriffe. On calcule ensuite les moyennes arithmétiques pour trouver les distances de ce groupement avec chaque espèce. Par exemple :

$$d(b - h, m) = (d(b, m) + d(h, m))/2 = (7 + 8)/2 = 7.5 \quad (2)$$

Voici le déroulement de l'algorithme :

Matrice	Arbre	Arbre Expression
$ \begin{array}{c} b \\ t \\ m \\ h \\ c \end{array} \begin{pmatrix} & b & t & m & h & c \\ 0 & 2 & 7 & 1 & 11 \\ 2 & 0 & 5 & 3 & 13 \\ 7 & 5 & 0 & 8 & 18 \\ 1 & 3 & 8 & 0 & 10 \\ 11 & 13 & 18 & 10 & 0 \end{pmatrix} $		“”
$ \begin{array}{c} t \\ m \\ c \\ b-h \end{array} \begin{bmatrix} & t & m & c & b-h \\ 0 & 5 & 13 & 5 \\ 5 & 0 & 18 & 7.5 \\ 13 & 18 & 0 & 10.5 \\ 5 & 7.5 & 10.5 & 0 \end{bmatrix} $		“(bh)”
$ \begin{array}{c} c \\ b-h \\ t-m \end{array} \begin{bmatrix} & c & b-h & t-m \\ 0 & 10.5 & 15.5 \\ 10.5 & 0 & 6.25 \\ 15.5 & 6.25 & 0 \end{bmatrix} $		“(bh)(tm)”
$ \begin{array}{c} c \\ bh-tm \end{array} \begin{bmatrix} & c & bh-tm \\ 0 & 13 \\ 13 & 0 \end{bmatrix} $		“((bh)(tm))”
		“(((bh)(tm))c)”

4 Précisions sur les fonctionnalités

Je suis consciente que ce sujet est difficile, et je serai indulgente sur les fonctionnalités. Il doit y avoir une liste d’espèces que l’on peut sélectionner pour faire un arbre. Dans l’exemple d’interface de la figure 1, il y a une liste d’espèces “Collection d’espèces” que l’on peut ajouter à la liste “Sélection d’espèces” pour construire l’arbre. Il y a un bouton “UPGMA” qui permet d’exécuter l’algorithme.

Ce que je vous demande, c’est qu’il y ait une manière de voir comment l’algorithme regroupe au fur et à mesure les espèces (cf figure 2). Dans l’exemple, après avoir cliqué sur le bouton “UPGMA”, la première matrice des distances s’affiche en haut à droite. Si l’on clique sur le bouton ‘matrice suivante’, une fenêtre s’affiche avec la matrice résultante de la seconde itération et un bouton pour afficher la suivante, et ainsi de suite.

Il n’est pas obligatoire de colorer la matrice. Vous pouvez très bien afficher les matrices sous forme de texte. De plus, ce n’est pas obligatoirement la matrice qui doit être affichée. Vous pouvez par exemple afficher un message qui dit “Le Bronteroc et l’Hippogriffe ont été regroupés”. Ou alors vous pouvez afficher les modifications de l’expression de l’arbre (cf troisième colonne du tableau ci-dessus).

Il doit y avoir un endroit (dans l’exemple de la figure 1 c’est en haut à gauche), où l’on voit l’arbre final. Cet arbre peut très bien être affiché sous forme d’expression d’arbre (cf troisième



Figure 2: Affichage de la matrice de la seconde itération.

colonne du tableau ci-dessus). L'implémentation d'un JTree (qui est difficile), n'est pas du tout obligatoire.

Enfin c'est un sujet où vous pouvez être très créatif. Vous pouvez par exemple faire une option pour que l'utilisateur ajoute une espèce à la collection d'espèces disponibles, faire une fiche descriptive des espèces, ... Si vous avancez vite vous pouvez même lier votre projet au projet "Alignement de séquences" !