Rapport du projet de M1

Combien de temps pour faire une espèce ?

Wiam Chaoui Sophie Manuel Stéphane Sadio

2021

Table des matières

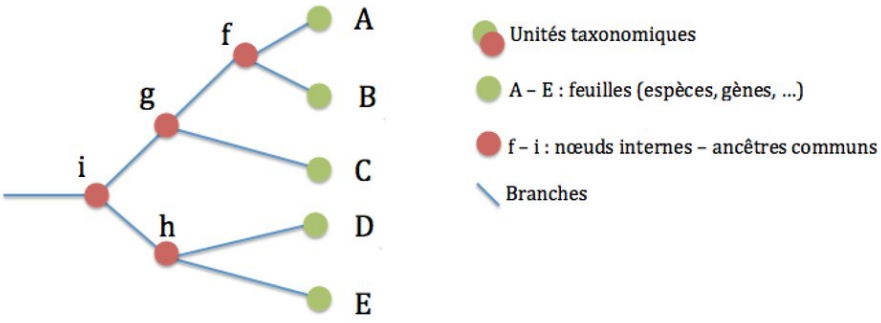
Chapter 1

Introduction

1.1 Motivation

La classification du vivant est depuis longtemps un vrai casse-tête pour les biologistes, surtout en ce qui concerne la notion d'*espèce*. De fait, il existe plusieurs définitions du mot espèce, ce qui rend encore plus compliqué un consensus. C'est pour cela que dans la suite nous ne nous étendrons pas sur cette notion et nous ne nous concentrerons que sur des espèces prédéfinies.

Les *arbres phylogénétiques* sont des outils permettant de représenter graphiquement certaines données de classification. En effet, ils présentent les relations de parenté entre *espèces*. On retrouve dessous différentes espèces actuelles, mais aussi leurs ancêtres communs (les *branchements évolutifs* qui correspondent à l'apparition d'une nouvelle homologie), ou encore la durée avant l'apparition d'une nouvelle espèce qui est donnée par la longueur des branches.



(http://zestedesavoir.com/media/galleries/2272/d1a1051e-782b-4b5d-81ac-c5641962b9c8.png.960x960_q85.jpg) Dans la suite, nous nous intéresserons aux *branchements évolutifs*. Supposons qu'un branchement évolutif apparaît après une durée

aléatoire d'une loi fixée μ indépendamment du passé et du futur évolutif des espèces.
Quelle est cette loi μ? Sa variance ? Sa moyenne ?

On observe des branchements successifs qui composent l'arbre phylogénétique et à partir de ces données quantitatives observées, on veut estimer la fonction de densité f qui donne la probabilité qu'un nouveau branchement évolutif apparaisse après un certain temps.

Formellement on a le modèle de densité suivant: soient les variables aléatoires $X_1, ..., X_n$, $n \in \mathbb{N}^*$ à valeur dans $\mathbb{R}^d$ (ici $d = 1$), indépendantes et identiquement distribuées de longueurs de branche observées. Elles ont pour fonction de densité f par rapport à la mesure de Lebesgue sur $\mathbb{R}$ supposée inconnue. Notre objectif est d'estimer cette fonction densité f sur laquelle on fait le moins d'hypothèses possibles. On fera seulement les hypothèses d'existence, de continuité et de positivité de la fonction, en servant une observation $(X_1, ...X_n)$, ce qui nous mène en statistique non-paramétrique, où le paramètre cherché est une densité de probabilité qui appartient à un espace fonctionnel infini, d'où la problématique de notre sujet.

1.2 Problématique

Comment estimer la loi de densité de la création d'une nouvelle espèce avec une méthode d'estimation non-paramétrique ?

Pour commencer, en se basant sur quelques définitions nous présenterons les méthodes d'estimations non-paramétriques et en introduisant quelques types d'estimateurs.
Par la suite, nous approfondirons les estimateurs de densité à noyau en menant une discussion sur leurs critères d'évaluation. Ainsi nous consacrerons un chapitre pour présenter des méthodes adaptatives.
Enfin, pour répondre à la problématique, nous implémenterons un estimateur à noyau adaptatif, avec la méthode de validation croisée puis l'utiliserons sur des données d'arbres phylogénétiques.