UNIVERSITÉ DU QUÉBEC À MONTRÉAL

UN NOUVEL ALGORITHME BIOINFORMATIQUE POUR RETROUVER LA RELATION ENTRE LA PHYLOGÉNIE ET LA PHYLOGÉOGRAPHIE DES ESPÈCES

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAITRISE EN INFORMATIQUE

PAR

NANCY BADRAN

MARS 2018

REMERCIEMENTS

Je tiens à exprimer mes sincères remerciements à toute personne qui a contribué de près ou de loin à mon projet de maitrise.

Je tiens à remercier chaleureusement mon directeur de recherche, M.Vladimir Makarenkov, pour son implication, son soutien, sa disponibilité et surtout ses précieux conseils. Je le remercie aussi pour la relecture de ce mémoire.

J'adresse mes profonds remerciements à ma collègue de laboratoire Nadia Tahiri, qui était toujours disponible pour l'avancement de ce projet. Je la remercie pour ses conseils, sa disponibilité et surtout sa créativité qui a contribué énormément à l'avancement de cette étude.

Un grand merci à M. Abdoulaye Baniré Diallo, pour ses conseils qui m'ont conduit à déterminer mon chemin à l'UQAM.

J'exprime aussi une profonde gratitude à ma mère Nadia, qui m'a soutenu de loin, à mon mari Ghassan, qui m'a toujours encouragé et poussé à continuer mes projets.

Enfin, à mes deux enfants, Sirine et Kazem, qui me donnent toujours le goût d'aller plus loin en leur regardant grandir année après année.

TABLE DES MATIÈRES

LISTE DE FIGURES

RÉSUMÉ

Dès l'essor de la phylogéographie, de nombreux chercheurs ont tenté d'expliquer le mécanisme qui régit la distribution géographique des espèces. Cependant, la plupart des recherches sur ce sujet souffrent du manque d'algorithmes performants.

Dans cette étude bioinformatique, nous désirons identifier des gènes ou des fragments de gènes qui déterminent la distribution géographique des espèces. L'objectif principal de cette étude consiste à identifier la relation entre la génétique et la phylogéographie des espèces à l'aide d'un nouvel algorithme que nous avons développé, puis implémenté en langage Java.

Dans le but de développer notre algorithme, nous nous sommes basés sur la théorie des graphes, les techniques de reconstruction d'arbres phylogénétiques, ainsi que sur les distances topologiques entre les arbres, comme par exemple la distance de Robinson et Foulds.

Nous avons testé notre algorithme en l'appliquant à un jeu de données de 52 mammifères carnivores localisés en Amérique du Nord.

Mots-clés : phylogéographie, arbre phylogénétique, distance topologique, algorithme bioinformatique, ADN, acide aminé.

CHAPITRE I

INTRODUCTION

Ce travail a été réalisé dans le cadre d'un projet de recherche à l'Université du Québec à Montréal au sein du laboratoire bioinformatique dirigé par le professeur Vladimir Makarenkov. Il s'agit d'une nouvelle version de l'algorithme proposé par Tahiri *et al* en 2012 (Tahiri *et al*, 2012). Cette nouvelle version a poussé le projet beaucoup plus loin, mettant en évidence de nouveaux résultats. Notre nouvel algorithme vise à répondre à la question suivante : « Est-ce que la distribution géographique des espèces est influencée par des facteurs génétiques? »

1.1 Mise en contexte

Le mécanisme responsable de la distribution géographique des espèces a toujours intéressé les écologistes. En effet, deux espèces qui sont très proches génétiquement sont-elles nécessairement proches géographiquement (c.-à-d. des habitats ayant les mêmes facteurs climatiques)? C'est à partir de ce constat que

nous avons développé notre problématique permettant ainsi une meilleure compréhension de la phylogéographie.

La phylogéographie est une science qui étudie les différents processus qui régissent la distribution des espèces (Donald et Alger, 1993). Dans notre projet, nous tenterons d'expliquer la cause de la distribution actuelle des espèces à travers leurs structures génétiques en les corrélant aux paramètres climatiques.

À travers la littérature, nous nous sommes penchés sur l'algorithme ND (Nodal distance) (Bluis et Shin, 2003) qui permet de mesurer la différence de positions relatives des espèces. Selon les auteurs, cet algorithme présente des difficultés lors de la reconstruction des arbres phylogénétiques. D'autres chercheurs ont réalisé une étude comparative des complexités algorithmiques sur un ensemble de programmes basés sur l'« Edit distance ». La plupart des travaux dans ce domaine se sont heurtés à la présence d'une borne supérieure de la chaine à analyser qu'il faut bien respecter (Hanada *et al.*, 2011). D'autres algorithmes restent seulement théoriques sans être testés sur des jeux de données réelles.

1.2 Objectifs

L'objectif principal de notre projet est de retrouver une combinaison de fragments de gènes avec un ensemble de paramètres climatiques pouvant être corrélé à la distribution géographique des espèces. Pour mener cette étude à bien, nous avons divisé cet objectif principal en plusieurs sous-objectifs qui sont comme suit :

- Choisir stratégiquement les données (c.-à-d. les espèces, la zone géographique et les gènes).

- Développer un premier algorithme permettant le prétraitement des données et l'inférence des arbres phylogénétiques et phylogéographiques.

- Développer un deuxième algorithme permettant de sélectionner les gènes qui déterminent la distribution géographique des espèces.

- Déterminer les positions d'intérêts sur chaque gène à travers une fenêtre coulissante.

1.3 Stratégie

Dans un premier temps, une revue de la littérature s'avère nécessaire pour définir les différentes terminologies et les concepts de bases tels que l'arbre phylogénétique, la phylogéographie et les outils d'inférence d'arbres phylogénétiques. Il existe plusieurs approches de reconstruction d'arbres phylogénétiques, celles qui sont basées sur les distances et d'autres basées sur les caractères. Dans notre cas, nous avons utilisé les deux approches : celles qui sont basées sur les distances en suivant un algorithme d'agglomération (*NJ*) (Saitou et Nei, 1987), et celles qui sont basées sur les caractères en utilisant *phyML* (Guindon *et al,* 2010).

Par la suite, nous avons calculé la distance topologique entre les arbres (phylogénétiques et phylogéographiques). Cette approche a permis de retrouver les paires d'arbres ayant la distance minimale.

Finalement, nous avons procédé à la localisation des positions d'intérêt sur les gènes retrouvés à l'étape précédente. Pour une meilleure interprétation des résultats, nous avons représenté ces fragments de gènes en 3D en utilisant des outils spécialisés par exemple Chimera (Pettersen *et al,* 2004). Cette représentation a un objectif principal de trouver une corrélation entre les structures protéiques de ces fragments de gènes et leur rôle dans la distribution des espèces.

1.4 Organisation du mémoire

Dans ce mémoire, nous abordons la problématique formulée au début de ce chapitre: retrouver les fragments des gènes qui expliquent la distribution géographique des espèces. Le mémoire comporte six chapitres dont le premier présente une introduction générale à notre étude. Le deuxième chapitre présente le cadre théorique du projet. Il constitue une revue de la littérature sur la terminologie dans les domaines de la phylogénie et la phylogéographie. Nous exposons aussi les techniques de reconstruction d'arbres phylogénétiques en mentionnant des exemples aidant à la compréhension des différents concepts. Dans le chapitre III, nous présentons les données sélectionnées, leurs types

(génétiques, géographiques et la liste des espèces), leur provenance et la raison de ce choix. Nous nous attardons aussi sur la phase du prétraitement des données en l'expliquant par un flux de travail. Après cet état de l'art, nous présentons notre nouvel algorithme dans le chapitre IV. Dans ce chapitre, nous présentons aussi les différentes distances topologiques entre les arbres phylogénétiques. Nous discutons également de la complexité de l'algorithme. Dans le chapitre V, nous présentons les résultats que nous avons obtenus, suivi de l'analyse et de l'interprétation de ces résultats. Finalement, le chapitre VI conclut le mémoire et propose des perspectives futures permettant d'accroître la performance de notre algorithme.

CHAPITRE II

DÉFINITIONS PRÉLIMINAIRES ET REVUE DE LA LITTÉRATURE

2.1 Introduction

Ce chapitre balisera les fondements nécessaires afin de mieux saisir la complexité et l'originalité de la problématique proposée dans le cadre de ma maitrise en informatique, à savoir la détection de la similitude entre un fragment d'un gène et un arbre de référence. Après une courte introduction, nous présenterons la phylogéographie à travers son histoire et sa définition. Nous parlerons également de l'origine des espèces, et enfin, nous clarifierons la problématique. Dans la section 3, nous rappellerons les notions relatives à l'arbre phylogénétique en y relatant son histoire, sa définition et sa terminologie. Enfin, nous évoquerons les différentes représentations de l'arbre phylogénétique et de son inférence

2. 2 La phylogéographie

2. 2.1 Historique et définition

Le terme de phylogéographie a été introduit pour la première fois en 1987 par John Avise (Avise *et al*, 1987). C'est un domaine d'études de plus en plus exploré

avec l'accroissement des données biologiques et des données climatiques (données volumineuses ou « big data ») (Marx, 2013). La phylogéographie englobe toutes les études à la base génétique, démographique et géographique qui ont mené à la distribution contemporaine des espèces. En effet, les espèces sont en échange mutuel avec leurs milieux et cherchent en permanence les habitats les plus favorables à leur survie. Les changements continus de l'environnement aboutissent en général à des variations climatiques. Ces variations poussent les communautés cladistiques à chercher d'autres abris pour assurer leur équilibre de vie ou de retrouver des moyens d'adaptation face à diverses transformations. Une telle migration engendre l'apparition ou l'extinction de certaines espèces. Dans l'ouvrage d'Avise et Nelson, les chercheurs se demandaient : « si la microévolution des espèces peut constituer une hypothèse de base pour la macroévolution entre espèces et plus précisément entre des taxons de haut niveau » (Avise et Nelson, 1989). Traditionnellement, pour retrouver les communautés cladistiques, les méthodes consistaient à aligner des séquences nucléotidiques d'un même groupe d'espèces, présentes dans une même zone géographique, et de retrouver ainsi les différences entre ces séquences (West-Eberhard, 2005). L'essor de la phylogéographie a éprouvé un grand développement par la mise en place des données moléculaires, de nouvelles méthodes d'amplification (Lamarck, 1809) et de séquençage d'ADN. Les applications bioinformatiques récentes et les nouveaux algorithmes appliqués à

ces données ont joué un rôle considérable au niveau de l'évolution des recherches en phylogéographie. La question principale qui nous intéresse est de savoir si l'arrangement actuel des espèces est corrélé à des origines génétiques de ces espèces. L'utilisation des arbres et des réseaux phylogénétiques pour modéliser la propagation d'espèces ainsi que leur histoire génétique a constitué un changement fondamental dans ce domaine.

2. 2.2 Origine des espèces

L'origine et la répartition géographique d'espèces proviennent à la fois des processus historiques qui ont marqué les migrations au cours du temps et à l'évolution génétiques menant à leur différenciation évolutive :

- La dérive génétique: consiste à l'apparition ou l'évolution d'un peuple dû à des faits aléatoires. Ce phénomène est entrainé par des modifications de la fréquence d'un allèle ou d'un génotype. Ces modifications ne sont pas liées à des processus connus comme la mutation ou la migration des gènes (Slatkin, 1987). Les effets de la dérive génétique se traduisent souvent par une perte des lignées génétiques aboutissant ainsi à une baisse dans la biodiversité[1]. Les effets de la dérive génétique deviennent de plus en plus importants quand la

[1]C'est la diversité de la vie sur terre, elle considère la diversité des écosystèmes, des espèces et des gènes dans l'espace et dans le temps.

population est petite. .Cette diminution est défavorable à l'adaptation des espèces à de nouveaux milieux de vie (Slatkin et Hudson, 1991).

- La mutation: la mutation est une modification inhabituelle que les gènes peuvent subir. Elle peut ne pas être remarquée sur une courte échelle du temps, mais selon son taux et son accumulation à travers les générations, les changements suivants peuvent survenir: l'apparition de nouvelles espèces ou l'extinction d'autres. La mutation joue un rôle important dans l'accroissement démographique des habitants (West-Eberhard, 2005).
- Le flux génétique: (c.-à-d. la migration génétique) ce mécanisme permet un échange des gènes entre les différentes populations. À long terme, le brassage génétique aboutit généralement à l'uniformisation de la génétique des espèces (Slatkin, 1987). Contrairement à la dérive génétique, ce mécanisme tend à réduire la différence génétique entre la population et à homogénéiser les fréquences allyliques.
- Les facteurs historiques : Nous citons par exemple, la migration des espèces lors des glaciations en Amérique du Nord.

Il faut prendre en considération aussi le mécanisme de migration temporelle et continue de certaines populations. Dans les deux cas, il y aura des changements génétiques engendrant une augmentation de la diversité intrapopulaire et une diminution de la diversité interpopulation.

La problématique de la phylogéographie est traitée principalement selon deux approches : 1) les méthodes hiérarchiques qui sont principalement basées sur le principe de clustering des communautés et incluent des algorithmes cladistiques tels que NCA (Gomez-Zurita *et al*, 2000) c.-à-d. Nested Cladistique Algorithme, les méthodes génétiques qui s'appuient sur la reconstruction des arbres phylogénétiques (voir la section 2.3).

2. 2.3 Problématique

Les approches phylogéographiques traitent conjointement le patrimoine génétique des espèces, ainsi que l'ensemble des paramètres climatiques. Il s'agit d'une évolution multiespèces (c.-à-d. entre espèces) (Taberlet *et al*, 1998). En comparant les flux et les divergences génétiques des inférences qui se chevauchent en lieu et en temps, la phylogéographie comparative construit une hypothèse des événements historiques et de la biodiversité (Bermingham et Moritz, 1998). Le problème posé dans notre étude est la détection de la relation historique entre les inférences génétiques des espèces et leurs distributions géographiques. Plusieurs questions peuvent être posées :

- De quelle manière les facteurs climatiques influencent-ils les structures génétiques permettant ainsi la dispersion des espèces?
- Est-ce que deux espèces proches génétiquement vont réagir de la même façon par rapport aux changements climatiques?

Les approches génétiques seules ne peuvent ni affirmer ni contester ces hypothèses. Des méthodes mathématiques, informatiques et géographiques constitueront des soutiens à la résolution de notre problème, d'où le but de notre projet qui cherche à développer un algorithme bioinformatique permettant de trouver les relations entre la phylogéographie et la génétique des espèces.

Une étape préliminaire de notre méthode est la reconstruction des arbres phylogénétiques. La suite de ce chapitre présente le concept d'arbre phylogénétique, sa définition, ses propriétés et les méthodes de sa reconstruction.

2. 3 L'arbre phylogénétique

2. 3.1 Historique

Au début du XIXe siècle, Jean-Baptiste Lamarck (Avise et Nelson, 1989) a réussi la hiérarchisation de la plupart des invertébrés constituant 80 % des animaux (Besson, 2012). Cette étude a abouti à un ouvrage de sept volumes : « Histoire naturelle des animaux sans vertèbres » (1815-1822). Mais c'est en 1809 que Lamarck (Lamarck, 1809) a exposé ses théories les plus contestées selon lesquelles représente la théorie de l'origine des différents animaux dans son œuvre « philosophie zoologique », qu'on l'appellera plus tard le transformisme.

Lamarck n'introduisait pas explicitement la notion du temps (voir Figure 2.1). Il ne croyait pas à un ancêtre commun de toutes les espèces, tandis que cela a été mentionné d'une façon explicite par Martin Ehrenberg en l'année 1838 dans la nouvelle version de « philosophie zoologique » du tome I en employant la phrase « le temps pour la formation des alternatives » (Danchin, 2011).

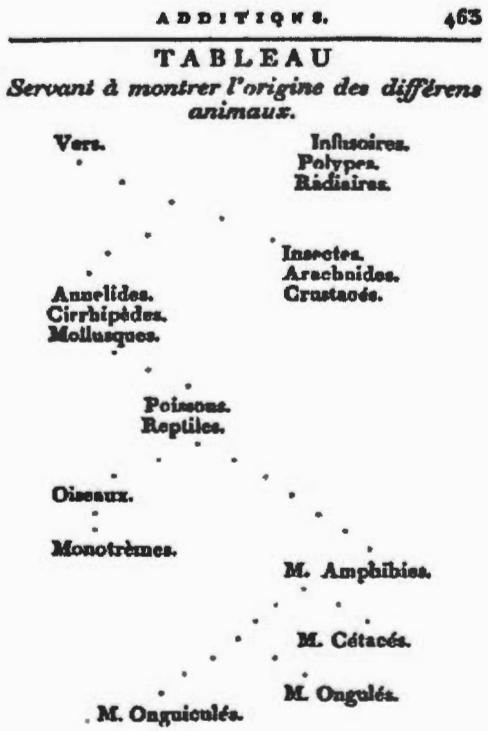


Figure 2.1 : Origine des animaux selon Lamarck (Lamarck, 1809).

Cependant, le terme de phylogénie fut utilisé pour la première fois par Ernst Haeckel en 1866 pour expliquer la succession des espèces animales et végétales au cours du temps. Son ouvrage principal est « Generelle Morphologie der

organismen » c.-à-d. Morphologie générale des organismes (Haeckel, 1866). La représentation de l'évolution animale de Haeckel était un arbre constitué de clades[2] et divisé verticalement, contrairement à la représentation horizontale actuelle. C'est Darwin, dans son ouvrage «Origine des espèces» (Darwin, 1859) qui était parmi les premiers à exposer le concept de phylogénie, suivi d'une première image d'un arbre phylogénétique (Figure 2.2), constituant ainsi une révolution scientifique. Darwin mentionne dans son ouvrage que des changements successifs peuvent arriver à une espèce et aboutir ainsi à la formation d'une nouvelle lignée d'espèces.

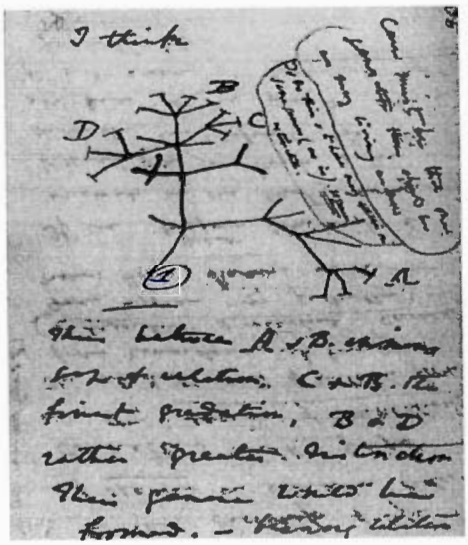


Figure 2. 2: Arbre de vie de Darwin tiré de son carnet.

[2] Un **clade** (du grec *κλάδος clados*, qui signifie « branche ») est un groupe monophylétique d'organismes, vivants ou ayant vécu, comprenant un organisme particulier et la totalité de ses descendants

Après la découverte de l'ADN, des milliers d'études ont été faites pour reconnaitre l'évolution des êtres vivants au cours du temps. Les changements des structures morphologiques successifs observés par Darwin ont trouvé leur source à travers les mutations des séquences nucléiques. Les conséquences de ces mutations peuvent mener à des spéciations dans les branches de l'arbre phylogénétique et faire apparaître une nouvelle espèce. La classification des communautés a été à la base du fondement de la notion de clades (Felsenstein, 2004).

2. 3.2 Définition et terminologie

La phylogénie est un domaine d'étude associé au développement des êtres vivants. Dans un arbre phylogénétique (ou un X-arbre) (Barthélemy et Luong, 1987), l'analyse des caractères utilise le principe de l'hérédité incluant toutes les transformations aléatoires (c.-à-d. inversion, substitution, insertion ou délétion). Du point de vue mathématique, un arbre phylogénétique est une structure de données spécifique, c.-à-d. un graphe acyclique. Ce dernier est constitué de quatre objets fondamentaux (Felsenstein, 2004).

- Les sommets: On a deux types de sommets: internes et externes. Ces derniers sont appelés aussi feuilles, ils représentent les espèces (c.-à-d. taxons) vivantes ou éteintes. Les sommets internes représentent des espèces

ancestrales. Des informations liées aux unités évolutives sont nécessaires pour l'induction de l'arbre.

- Les branches ou les liens entre les nœuds: Ce sont les segments de dépendance entre les sommets déterminant les liens de parenté selon une trace ascendante envers un ancêtre commun. Des propriétés peuvent être associées à chaque branche. Ici nous pouvons citer :
    - La *mesure* : qui est une distance évolutive, définie comme étant la durée nécessaire à l'apparition d'une nouvelle espèce ou le taux de mutations.
    - Le *poids pondéré des branches* : il est associé à chaque arrête; il définit la vitesse d'évolution ou le taux de mutation.
    - L'*orientation* : Un graphe peut être orienté, c.-à-d. ayant les branches avec des directions, d'où les notions d'ancêtre et de descendant.
- La racine : Elle représente l'ancêtre commun de toutes les espèces dans l'arbre phylogénétique. Trouver, puis placer la racine dans un arbre phylogénétique n'est pas toujours évident. En effet, la plupart des méthodes de construction d'arbres phylogénétiques produisent des arbres non enracinés. Il existe plusieurs méthodes pour placer la racine. Ces méthodes ne dépendent pas des processus d'inférence d'arbres phylogénétiques. Nous pouvons mentionner l'approche de l'outgroupe (c.-à-d. par un groupe extérieur ou le

plus éloigné phylogénétiquement) et l'approche du point médian (c.-à-d. l'enracinement du poids moyen).

- Le *degré* d'un arbre est le nombre maximum d'arrêtes adjacentes à un nœud. On distingue deux types d'arbres : *l'arbre parfaitement résolu* dans lequel les nœuds internes ont tous le degré égal à 3 et les feuilles dont le degré est 1 (voir Figure 2.3) versus *l'arbre non résolu* dont les nœuds internes peuvent avoir le degré plus grand que 3 (voir Figure 2.4).

On peut aussi caractériser les arbres en fonction de présence ou absence de la racine :

- Arbre non enraciné : est un graphe connexe sans cycle dont les arrêtes n'ont pas de direction. Dans un arbre non enraciné, il existe un seul et unique chemin menant d'un sommet à un autre. Les arrêtes représentent les liens de parenté entre les espèces. Dans un tel arbre, nous ne pouvons pas définir la relation entre un ancêtre et ses descendants.
- Arbre enraciné : est un arbre comprenant une racine qui représente l'ancêtre commun de toutes les espèces représentées. L'ajout d'une racine à un arbre phylogénétique l'enrichit en lui donnant une orientation (c.-à-d. d'un parent vers un fils) (voir Figures 2.3 et 2.4). Cependant, il est souvent difficile de déterminer l'ancêtre commun de tous les taxons (Darlu et Tassy, 1993). Il est

donc plus adéquat d'utiliser les arbres non enracinés quand les sens d'orientation sont incertains.

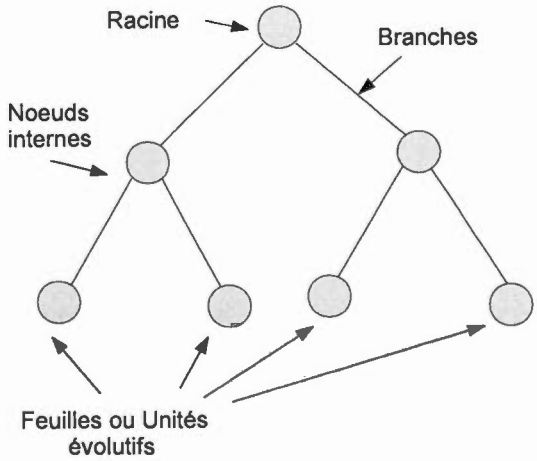


Figure 2.3: Illustration d'un arbre phylogénétique résolu et enraciné.

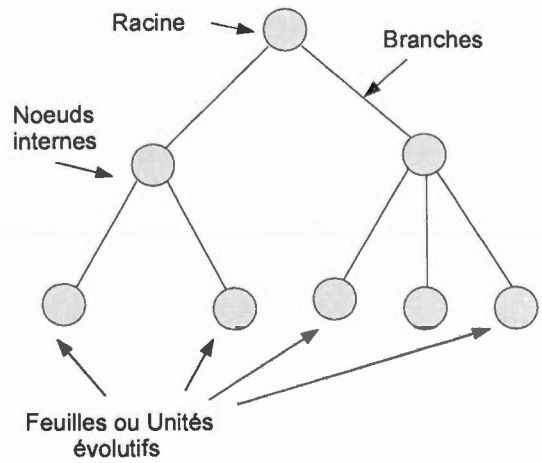


Figure 2.4: Illustration d'un arbre phylogénétique non défini et enraciné.

Étant donné n espèces, le nombre d'arbres phylogénétiques enracinés possibles peut être obtenu par la récurrence suivante (voir Équation 2.1) :

$$T_n^{enraxine} = \frac{(2n-3)!}{(2^{(n-2)})(n-3)!} \qquad (2.1)$$

Alors que le nombre d'arbres phylogénétiques non enracinés possible est obtenu par la récurrence suivante (voir Équation 2.2) :

$$T_n^{non\ enraxine} = \frac{(2n-5)!}{(2^{(n-3)})(n-3)!} \qquad (2.2)$$

2. 3.3 Représentation de l'arbre phylogénétique

Soient les deux ensembles X = $\{x_i / i= 1....n\}$ et Y = $\{y_i / i=1... n\}$, une dissimilarité entre X et Y est une mesure qui quantifie la prochetée génétique d'une espèce à une autre. Un ensemble d'espèces, ayant une mesure de dissimilarité entre ses éléments peut être représenté d'une façon générale par un X-arbre. Les sommets de cet arbre seront étiquetés par les éléments de l'ensemble, tandis que les arrêtes seront pourvues d'une pondération plus grande que 0. Cette pondération doit respecter la propriété que la somme des arrêtes constituant un chemin entre deux sommets x et y soit une bonne approximation de la valeur de dissimilarité entre ces deux sommets x et y. À travers la revue de littérature, nous distinguons plusieurs types de représentation. Les essentiels sont présentés dans le livre de Barthélemy et Guénoche (Barthélemy et Guénoche, 1991) qui ont exposé trois tracés : Tracé axial, radial et hiérarchique. À noter que

ces différentes représentations peuvent être obtenues à partir du logiciel T-Rex (Boc et Makarenkov, 2012 ; Makarenkov, 2001).

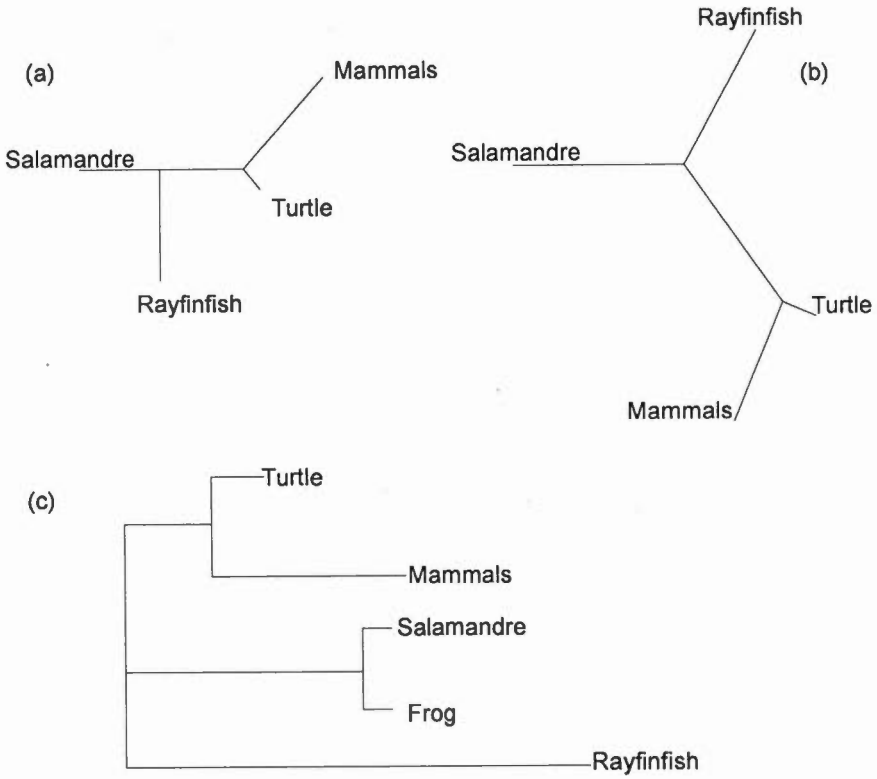


Figure 2.5: Trois tracés d'un arbre phylogénétique issus du logiciel T-Rex. (a) Axial, (b) Radial, (c) Hiérarchique horizontal.

2. 3.4 De la génétique à la phylogénétique

La reconstruction de l'arbre phylogénétique est un sujet qui s'avère aussi ardu, car la vraie relation représentée entre les espèces données est souvent inconnue. Les

scientifiques peuvent prédire des traces historiques, mais leurs scénarios doivent toujours être justifiés. Le principal défi est la mise en place d'un algorithme capable de reconstruire l'arbre phylogénétique le plus fidèle aux connaissances biologiques. Il existe plusieurs méthodes de reconstruction d'une phylogénie (Felsenstein, 2004) et la plupart de ces méthodes s'appuient sur la même intuition. Une espèce subissant de nombreux changements (c.-à-d. des mutations) peut se détacher de son espèce mère pour soit former une nouvelle soit s'éteindre.

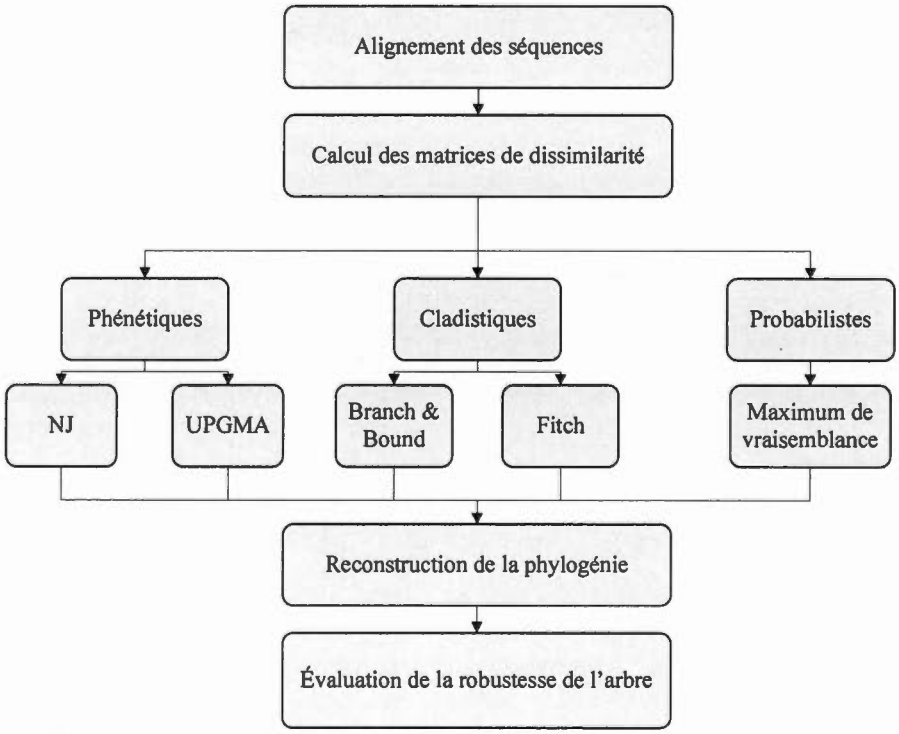


Figure 2.6: Flux de reconstruction d'une phylogénie.

La procédure de la reconstruction d'un arbre phylogénétique débute toujours par une étape primordiale : *l'alignement des séquences*. Les séquences utilisées sont des

séquences d'ADN, d'ARN ou de protéines représentatives d'un gène donné. Le processus d'alignement ne tient pas compte de la fonction d'évolution (Etien, 2006). La différence entre les lignées était auparavant de types morphologiques. C'est à la suite de la révélation des séquençages et l'alignement d'ADN que la phylogénie a été mise en place. En effet, il existe des mutations qui peuvent affecter les séquences nucléiques sans qu'elles apparaissent morphologiquement dans les espèces. Les données à séquencer proviennent généralement des bases de données en ligne, par exemple GenBank de NCBI (Benson *et al*, 2008). Après le téléchargement des données (p. ex. sous format FASTA), plusieurs logiciels sont disponibles pour effectuer le processus de l'alignement. L'alignement des séquences consiste à mettre en correspondances les sites des différentes séquences pour pouvoir comparer leurs caractères comparables. Parmi ces logiciels, le plus connu est ClustalW (Li, 2003). Les fichiers de sorties sont souvent en format *PHYLIP* qui est agréable pour la reconstruction d'arbres. Plusieurs autres logiciels sont aussi accessibles via le serveur T-Rex par exemple MUSCLE (Edgar, 2004) et MAFFT (Katoh *et al*, 2005).

La distance entre deux taxons *i* et *j* dans un arbre phylogénétique représente le nombre d'événements de mutation qui ont mené à la séparation de ces taxons. Mathématiquement dit, c'est la somme des longueurs des branches qui séparent deux espèces dans un arbre phylogénétique. Il existe trois principales approches pour la reconstruction d'arbres phylogénétiques.

- Les méthodes phénétiques, ou celles qui sont basées sur les distances entre les séquences, se limitent sur le calcul de la distance évolutive entre les différentes séquences puis calculent une instance d'arbre par des algorithmes d'agglomération (regroupement hiérarchique). Ces méthodes étudient les liens entre les taxons en observant leur degré de ressemblance sans tenir compte de leur processus d'évolution.
- Les méthodes cladistiques, ou celles qui sont basées sur les caractères, cherchent les caractères qui sont partagés entre les différents taxons pour chaque nœud ou ancêtre éventuel. Elles génèrent souvent tous les scénarios d'évolution possibles en remontant par le temps. Par la suite le meilleur scénario est choisi. Le maximum de parcimonie est à la base de cette approche. Ces méthodes étudient les liens de parenté entre les taxons en s'intéressant à leur degré de similitude et sont souvent basées sur la généalogie.
- Les méthodes probabilistes et bayésiennes ont été premièrement appliquées à la phylogénie moléculaire par Neyman (Neyman, 1971), puis développées par Felsenstein (Felsenstein, 1981). Elles reposent sur le fait suivant : connaissant un modèle d'évolution, on estime l'arbre phylogénétique par des méthodes statistiques comme le maximum de vraisemblance. Ces méthodes et les méthodes cladistiques sont les plus justifiées du point de vue biologique, mais

les plus couteuses au niveau du temps. Les méthodes bayésiennes font partie de cette dernière approche (voir Figure 2.6).

2. 3.4.1 Méthodes phénétiques

En premier temps, nous sommes menés à introduire quelques définitions de la théorie des graphes concernant la distance appliquée à l'arbre phylogénétique ou le X-arbre selon (Barthélemy et Luong, 1987). La distance $d(x, y)$ entre deux sommets x et y dans un X-arbre est définie comme étant la somme de toutes les longueurs d'arrêtes du chemin unique menant de x à y. Cette distance est euclidienne et vérifie les conditions suivantes :

- *La non-négativité : $d(x, y) \geq 0$;*
- *La symétrie : $d(x, y)=d(y, x)$;*
- *La distinction : $d(x, y)=0$ ssi $x=y$;*
- *L'inégalité triangulaire : $d(x, z) \leq d(x, y) + d(y, z)$*

L'étape suivante sera le calcul de la matrice de dissimilarité entre les taxons:

Définition 1.1 : *Soit $X= \{x_1, x_2, \ldots x_n\}$ un ensemble de n éléments. La dissimilarité sur X est une fonction d positive de $X \rightarrow X \times X$ telle que :*

1- $d(x_i, x_j)=d(x_j, x_i)$ et

2- $d(x_i, x_j)=d(x_j, x_i) \geq d(x_i, x_i)=0$

Définition1.2 : *Une dissimilarité d sur X, satisfait la condition des quatre points, si pour tout x, y, z, et w de X : d (x, y) +d (z, w) ≤ max {d (x, z) +d (y, w); d(x,w)+d(y, z)}.*

Théorème 1.1 : *Zaretskii (1965), Buneman (1971), Patrinos et Hakimi (1972) et Dobson (1974) ont prouvé qu'une dissimilarité qui vérifie la condition des quatre points peut être représentée par un X-arbre (arbre phylogénétique), et que cet arbre est unique.*

Pour calculer la matrice de dissimilarité, la distance est calculée en comptant le nombre de substitutions observées puis on la divise par le nombre total des nucléotides dans le site. Ce calcul est effectué entre toutes les séquences prises deux à deux. Il faut noter qu'une telle métrique est une estimation de la distance d'évolution. Cette dernière est à la base de la reconstruction de l'arbre phylogénétique. Cependant, cette distance inclut deux biais majeurs :

- Une possibilité d'avoir des mutations multiples non observables qui se sont déroulées au cours du temps.
- Les substitutions des sites ont des probabilités différentes, qui peuvent varier selon le type de données.

La correction de ces biais est faite par des méthodes dédiées à des modèles d'évolutions (voir section 2.4).

- De la matrice de dissimilarité à la phylogénie

La matrice de dissimilarité est dite aussi la matrice de distance évolutive. En partant d'une matrice dissimilarité **D** on peut toujours construire la phylogénie correspondant. Une fois cette matrice est établie, il y a plusieurs méthodes pour la création d'un arbre. Les méthodes basées sur le clustering sont très utilisées en bioinformatiques. Elles se reposent sur l'idée de regroupement des données étape par étape pour aboutir finalement à la phylogénie. Afin de reconstruire l'arbre, on emploie des techniques mathématiques : les techniques d'ajustement, l'évolution minimale et le clustering. Les méthodes d'ajustement consistent à trouver un arbre non enraciné en estimant les longueurs des branches qui donnent le meilleur ajustement à la matrice des distances observées **D** (Avise et Nelson, 1989; Darlu et Tassy, 1993), c.-à-d. à trouver la distance d'arbre la plus proche de **D**. Ce problème est NP difficile. Contrairement à cette méthode, la technique de l'évolution minimale est basée sur la somme des longueurs des branches de l'arbre qui est déterminée par la méthode des moindres carrées non pondérée. La dernière approche est celle qui est utilisée dans notre projet qui emploie l'algorithme *NJ* (Saitou et Nei, 1987). Cette dernière applique un algorithme de regroupement sur la matrice de dissimilarité pour reconstruire l'arbre. La méthode *NJ* est avantageuse, car le temps d'exécution est très rapide qui est d'ordre n^3 ou n est le nombre d'espèces. On suppose souvent que toutes les unités évolutives ont la

même distance de la racine, c.-à-d. qu'ils ont le même rapport évolutif comme c'est fait dans la méthode *UPGMA* (Yap et Nelson, 1996).

- Algorithmes se basant sur les méthodes phénétiques

Parmi ces algorithmes, deux prédominent : 1) l'algorithme de Neighbor-Joining (*NJ*) et 2) l'algorithme d'*UPGMA* « Unweighted Pair Group Method with Arithmetic mean›› .

1) Algorithme de *NJ*

L'algorithme de *NJ* c.-à-d. Neighbor joining est un algorithme itératif d'analyse de groupe (Saitou et Nei, 1987). Les méthodes qui utilisent cet algorithme sont parmi les plus rapides au niveau du temps. Il est un algorithme itératif qui commence par un arbre en forme d'étoile ayant n feuilles, ou espèces, et finit par la reconstruction d'un arbre phylogénétique non enraciné avec n feuilles et $2n$-3 branches.

Soit $D = (d_{ij})$ une matrice de dissimilarité entre n espèces telles que :

$$D = \begin{cases} D[i][j] = 0 \, pour \; tout \; 1 \le i \le n \\ D[i][j] = D[j][i] pour \; tout \; 1 \le i, j \le n \end{cases}$$

On choisit itérativement deux taxons à regrouper ensemble de sorte que la longueur totale de l'arbre soit minimale. On cherche à minimiser la valeur suivante :

$$S_0 = \frac{1}{n-1} \sum_{i=1}^{n-1} \sum_{j=i+1}^{n} d_{ij} \qquad (2.3)$$

Pour chaque couple de nœuds *i* et *j*, on calcule ainsi la somme des longueurs de branches par la formule (2.4) :

$$Sij = \frac{1}{2} D[i][j] + \frac{1}{2(n-1)} \sum_{1 \le k \le n; k \ne i,j} [D[i][k] + D[j][k] + \frac{1}{n-2} \sum_{k \le l \le n; l \ne i,j} D[k][l] \qquad (2.4)$$

Si la somme des longueurs de branches de deux nœuds *i* et *j* est minimale, alors ils seront remplacés par un nœud X qui sera leur ancêtre commun. Ainsi, on supprime les deux nœuds *i* et *j* de la matrice et on calcule la nouvelle distance de X aux autres nœuds dans la matrice par la formule (2.5). La taille de la matrice sera réduite de 1.

$$D[X][K] = \frac{1}{2}(D[i][k] + D[j][k]), k \ne i, j \qquad (2.5)$$

L'algorithme *NJ* se décompose en 4 étapes.

- Étape 1 : Calculer pour chaque couple de taxons la somme des longueurs de branches suivant la formule (2.4).
- Étape 2 : Rechercher deux espèces *i* et *j* qui minimisent la longueur totale de l'arbre si ces deux espèces sont réunies la formule (2.3).
- Étape 3 : Éliminer les deux espèces *i* et *j* de la matrice et considérer le cluster (i, j) comme inséparable.
- Étape 4 : Recommencer ces deux étapes jusqu'à ce qu'il ne reste que 3 espèces dans la matrice.

Par la suite, on illustre un exemple de l'exécution de l'algorithme (*NJ*) :

$$Dij = \begin{matrix} 1 \\ 2 \\ 3 \\ 4 \end{matrix}\begin{bmatrix} 0 & & & \\ 4 & 0 & & \\ 5 & 7 & 0 & \\ 6 & 8 & 5 & 0 \end{bmatrix}$$

Étape 1 : Calculer la distance de chaque nœud à tous les autres : S_{ij} pour toute espèce i et j en appliquant la formule (2.4) :

$$S_{12}=\frac{1}{2}D_{12}+\frac{1}{2(4-1)}[D_{12}+D_{23}+D_{14}+D_{24}]+\frac{1}{4-2}D_{34}=11$$

$$S_{13}=\frac{1}{2}D_{13}+\frac{1}{2(4-1)}[D_{12}+D_{32}+D_{14}+D_{34}]+\frac{1}{4-2}D_{24}=12$$

$$S_{14}=\frac{1}{2}D_{14}+\frac{1}{2(4-1)}[D_{12}+D_{42}+D_{13}+D_{43}]+\frac{1}{4-2}D_{32}=12$$

$$S_{23}=14$$

$$S_{24}=12$$

$$S_{34}=11$$

Le tableau suivant met en évidence les valeurs de S_{ij} :

(1,2)	(1,3)	(1,4)	(2,3)	(2,4)	(3,4)
11	12	12	14	12	11

Étape 2 : Chercher les deux espèces i et j dont la valeur de S_{ij} est minimale. Dans le tableau précédent, les espèces (1,2) peuvent être jointes, ainsi que les espèces (3,4). Ce choix a été fait, car S_{12} et S_{34} constituent les minimums parmi les valeurs de S_{ij}.

Étape 3 : Joindre les deux espèces (1,2) et les remplacer par l'espèce intermédiaire X.

Étape 4 : Joindre les espèces X, 3 et 4 dans une topologie unique possible d'arbres à 3 feuilles. Étant donné que dans cet exemple on a seulement 4 espèces, le résultat est déduit sans faire une deuxième itération entre les nouveaux sommets. Le nombre d'itérations dépend du nombre des sommets n, et égale à n-3.

La Figure 2.6 représente l'arbre de départ et l'arbre obtenu par *NJ*.

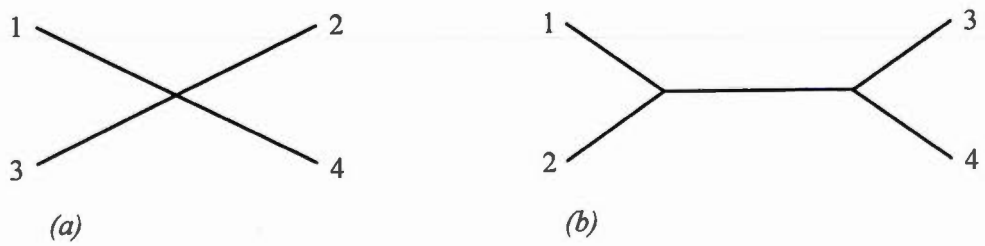


Figure 2.7 : (a) forme étoilée de départ (b) est la forme de l'arbre après application de (*NJ*).

2) Algorithme d'*UPGMA*

L'algorithme d'*UPGMA* (c.-à-d. « Unweighted Pair Group Method with Arithmetic mean››) est un algorithme de clustering hiérarchique. On utilise souvent cet algorithme si les séquences ne sont pas très différentes. C'est un algorithme séquentiel de regroupement. Il cherche les deux séquences les plus proches. *UPGMA* les considère comme un groupe, puis identifie les séquences les plus proches et ainsi de suite jusqu'au regroupement de toutes les séquences (Sneath et Sokal, 1973).

L'algorithme *UPGMA* est comme suit :

1- Trouver deux espèces *i* et *j* dont la distance d_{ij} est minimale;

2- Connecter ces deux espèces puis calculer L_i et L_j comme suit

$$L_i = L_j = \frac{d_{ij}}{2};$$

3- Recalculer la distance entre le nouveau nœud *K* qui remplace *i* et *j* et les autres. Cette distance est donnée par la formule suivante :

$$d_{ij,k} = n_i * \frac{d_{ik}}{(n_i + n_j)} + n_j * \frac{d_{jk}}{(n_i + n_j)}$$

4- Éliminer la ligne et la colonne correspondante à *i* et à *j* dans la matrice de distance *D*;

5- L'algorithme s'arrête s'il ne reste que 3 espèces dans la matrice.

Les deux algorithmes (*NJ*) et (*UPGMA*) ont une complexité algorithmique de $O(n^3)$ où n est le nombre d'espèces.

2. 3.4.2 Méthodes cladistiques

Les procédures cladistiques, ou de parcimonie sont utilisées dans le cas de petit nombre d'espèces. Elles cherchent à minimiser le nombre de mutations entre les séquences données. On s'appuie normalement dans ces méthodes sur deux hypothèses :

- L'évolution des sites est indépendante d'un site à l'autre;
- La vitesse d'évolution au cours du temps est lente et constante.

L'idée est de trouver la solution la plus parcimonieuse. Ces méthodes se basent essentiellement sur des algorithmes dynamiques. Pour chaque nœud dans l'arbre, on calcule tous les états possibles en remontant dans l'arbre vers la racine. Pour trouver l'arbre de maximum de parcimonie, l'algorithme recherche toutes les topologies possibles et choisit parmi eux la meilleure solution. Cependant, trouver l'arbre le plus parcimonieux est un problème NP-difficile (Felsenstein, 2004). Pour cela les méthodes basées sur le principe de distances sont plus rapides. Il existe plusieurs heuristiques et algorithmes qui se sont basés sur le principe de maximum de parcimonie. Mentionnons ici l'algorithme de Branch et Bound (Fukunaga et Narendra, 1975) et l'algorithme de Fitch (Fitch, 1971).

2.3.4.3 Méthodes probabilistes

Ces méthodes ont été d'abord appliquées à la phylogénie moléculaire par Neyman (Neyman, 1971), et puis développées par Felenstein (Felsenstein, 1981). Elles reposent sur le fait suivant : connaissant un modèle d'évolution, on estime l'arbre phylogénétique par des méthodes statistiques comme le maximum de vraisemblance. Ces méthodes sont caractérisées par leur convergence vers une valeur correcte (la consistance) et aussi par leur variance minimale tout autour de cette valeur (efficience). Elles sont les plus justifiées, mais les plus couteuses au niveau du temps. Toutes ces méthodes citées au-dessus sont implémentées et accessibles à partir des logiciels tels que PAUP (Swofford et Begle, 1993), *PHYLIP* (Plotree et Plotgram, 1989) et *T-Rex* (Makarenkov, 2001; Boc et Makarenkov, 2012). Notons que les méthodes de maximum de vraisemblances les plus efficaces sont : *PhyML* 3.0 (Guindon et *al*, 2010) et *RAxML* (Stamatakis, 2006). Dans notre projet, nous avons utilisé la méthode *PhyML* (voir le Chapitre IV).

2.4 Modèles d'évolutions

Une mutation est le changement de l'état des nucléotides au cours du temps. Un modèle d'évolution décrit ces changements avec des procédures probabilistes

liées au temps. C'est un processus qui permet de calculer la probabilité des mutations observables entre les différentes séquences. On suppose à la base que :

- Chaque site dans la séquence évolue d'une façon indépendante des autres.

- Les mutations ou le transfert de l'état *i* à l'état *j* suit un modèle Markovien (Bourguignon et Robelin, 2004) c.-à-d. le passage de l'état *i* à l'état *j* dépend uniquement de l'état *i* et non des états précédents.

- On suppose aussi que le système est homogène, c.-à-d. les probabilités sont fixe sur toutes les branches de la phylogénie qui représente ces séquences.

- Tous ces processus proposent des modèles réversibles dans le temps, c.-à-d. que la vitesse à laquelle l'état *i* se transforme en *j* est égale à la vitesse amenant l'état *j* à l'état *i*.

2. 4.1 Taux de substitution

On considère Σ = {A, C, G, T} l'alphabet de la grammaire nucléotique tel que : A= Adénine, C = Cytosine, T = Thymine, G = Guanine. On représente les deux nucléotides A et G par R et les deux autres C et T par Y. Les substitutions R <-> R sont des transitions tandis que les substitutions de la forme R<-> Y sont des transversions. Le modèle de substitution est représenté par une matrice (4*4) contenant le rapport des transitions instantanées :

$$Q_{ij} = \{c_{ij}\}$$

Tel que c_{ij} est le taux de changement de la vitesse à laquelle l'échantillon *i* devient l'échantillon *j*. Ce rapport aussi est connu par la distance évolutive et est calculé suivant la formule (2.6) de Kimura (Kimura, 1980).

$$d = -\frac{1}{2}\ln[(1-2p-q)\sqrt{1-2q} \qquad (2.6)$$

Où $\{\frac{p \text{ corresponda la dist evolutive propre a une transition}}{q \text{ corresponda la dist evolutive propre a une transversion}}$

Concernant les séquences protéiques, d'autres contraintes s'ajoutent :

- Chaque emplacement d'un acide aminé a une probabilité donnée de modification. Les probabilités de remplacement de chaque acide aminé par un autre sont corrélées selon les différentes matrices suivantes (PAM, JTT et WAG) (Kosiol et Goldman, 2005; Whelan et Goldman, 2001).

- La distance évolutive est calculée dans ce cas en fonction de nombre *p* de différences observées entre les deux séquences. Cette approximation est soulignée par la formule (2.7) de Kimura protéine (Kimura et Ohta, 1972).

$$d = -\ln(1-p-0.2p^2) \qquad (2.7)$$

Notons que dans les séquences codantes, les transitions sont plus abondantes que les transversions ce qui peut influencer leurs modèles d'évolution. En effet, les matrices de substitutions confirment que la probabilité propre à des transitions est dominante par rapport à celle des transversions. Dans la plupart des cas, ces séquences correspondent à des acides aminés conservateurs (Barrett *et al.*, 1991).

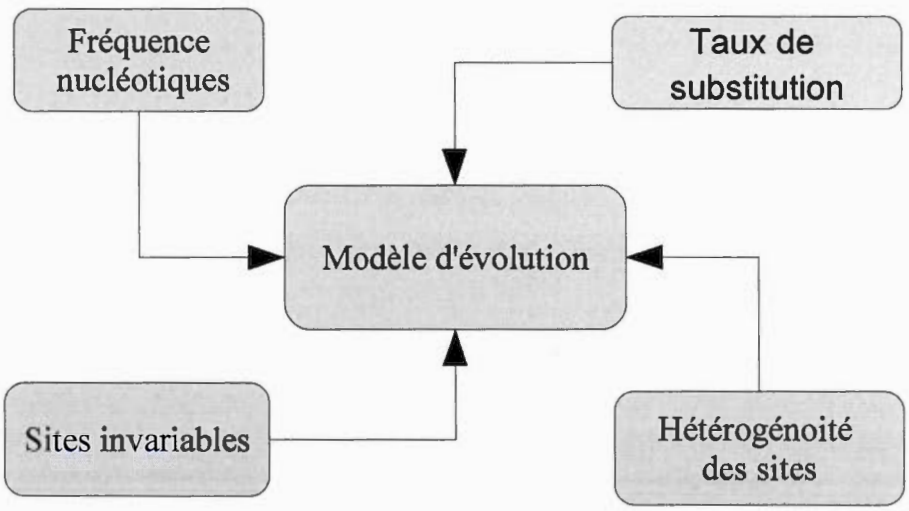


Figure 2.8: les principaux constituants d'un modèle dévolution des séquences nucléotidiques (Eddy, 1996).

2.4.2 Les différents modèles d'évolution

Les premiers modèles d'évolution réversibles étaient homogènes en temps et en sites, c.-à-d. que les nucléotides évoluaient indépendamment les unes des autres et qu'elles ont la même probabilité d'occurrence. On prend comme exemple celui de

Jukes et Cantor 1969 (Jukes et Cantor, 1969). La matrice génératrice de ce modèle est la suivante :

$$Q_{JC69} = \begin{matrix} A \\ C \\ G \\ T \end{matrix} \begin{bmatrix} - & 1 & 1 & 1 \\ 1 & - & 1 & 1 \\ 1 & 1 & - & 1 \\ 1 & 1 & 1 & - \end{bmatrix}$$

En suivant la loi de poisson (Tavaré, 1986; Delsuc et Douzery, 2004), à partir de cette matrice on pourrait estimer l'histoire évolutive d'un nucléotide en parcourant l'arbre jusqu'à sa racine. Bien que le modèle JC69 fût satisfaisant à l'époque or les chercheurs ont constaté que la probabilité d'avoir des transitions durant la procédure de l'évolution est beaucoup plus élevée que d'avoir des transversions d'où il était nécessaire de prendre en considération cette différence.

Pour ajuster ce biais, Kimura (Kimura, 1980) a proposé un modèle en ajoutant un paramètre K qui désigne la proportion des transitions par rapport à celle des transversions. D'où la matrice génératrice de ce modèle :

$$Q_{K80} = \begin{bmatrix} - & 1 & k & 1 \\ 1 & - & 1 & k \\ k & 1 & - & 1 \\ 1 & k & 1 & - \end{bmatrix}$$

En recherchant de plus en plus sur des séquences nucléotiques, les scientifiques ont découvert que la probabilité d'occurrence des nucléotides dans les séquences n'est pas la même : les mitochondries des métazoaires sont plus riches en A et en

T. De ce fait, Felsenstein (Felsenstein, 1981) propose un autre modèle de substitution qui corrige le biais précédent (modèle F81). Ce processus considère que les nucléotides A, C, G et T ont des fréquences d'équilibre (π_i) différentes qui définissent les transitions observées. Il y a aussi des modèles qui font la combinaison de deux autres F81 et K80 en supposant que les branches sont non uniformes en temps et en site et que le taux des transitions se diffère du taux de transversion. C'est le cas du modèle HKY85 (Posada et Crandall, 2001). Finalement le modèle général (GTR) qui est réversible et homogène. (Voir Figure 2.9).

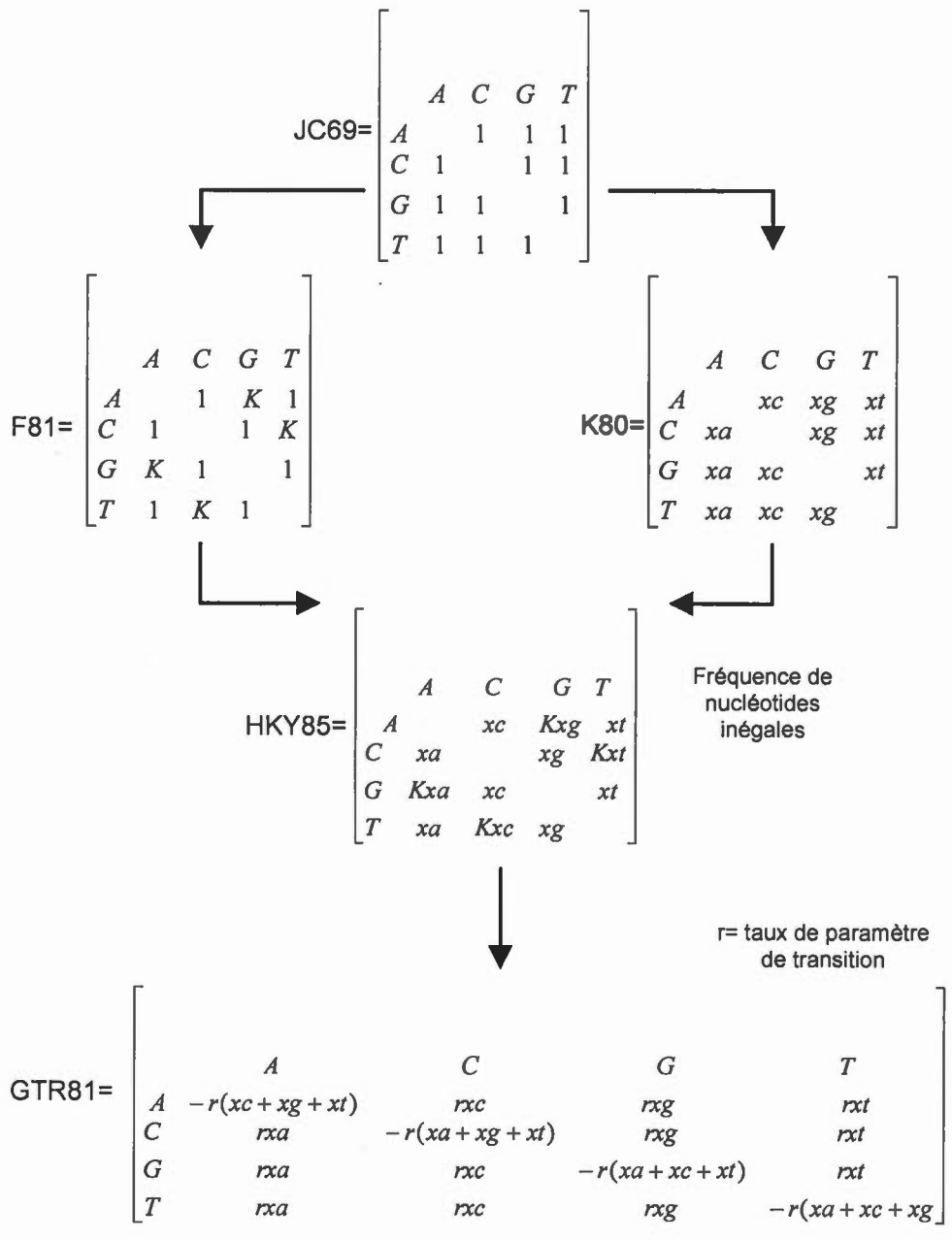


Figure 2.9: Les différents modèles d'évolution. Diagramme inspiré de (Mullahy, 1986).

2. 5 Conclusion

Dans ce chapitre, nous avons introduit les deux principaux axes de notre projet de maîtrise d'informatique. En redonnant une définition claire de la phylogéographie, ainsi que sa terminologie, nous avons décrit ces principaux enjeux et aussi son potentiel. De plus, nous avons présenté les principes de base relatifs à l'analyse phylogénétique à travers des définitions de base, la terminologie et surtout les différentes méthodes d'inférence d'arbres phylogénétiques. Après ce premier chapitre introductif, nous décrirons plus en détail les jeux de données que nous analyserons, notre algorithme et les résultats donnés par cet algorithme.

CHAPITRE III

DESCRIPTION DES DONNÉES ET LEUR PRÉTRAITEMENT

3. 1 Introduction

Afin de tester et de valider notre algorithme (voir chapitre IV pour la description de cet algorithme), nous décrirons dans ce chapitre les données sélectionnées. Ces données sont à la fois génétiques et géographiques. Elles sont complètes dans le sens qu'elles contiennent la liste des espèces, leurs distributions dans les zones géographiques, ainsi que leurs séquences génétiques. Dans un premier temps, nous déterminerons et sélectionnerons un ensemble d'espèces choisies. Nous donnerons les critères que nous allons utiliser dans le chapitre IV ainsi que leurs descriptions. Puis, nous présenterons en détail les différentes étapes de prétraitement de données (c.-à-d. leur récupération et leur préparation afin d'obtenir les arbres phylogéographiques). Enfin, nous présenterons les données génétiques utilisées, en expliquant la particularité de chaque gène choisi.

3. 2 La liste des espèces

Dans notre projet, nous avons choisi un ensemble d'espèces appartenant au groupe des mammifères carnivores. Nous exposerons la liste des données en relatant d'abord la notion et l'origine de ce concept fondamental soit « *l'espèce* ».

3.2. 1 La notion de l'espèce

L'espèce est choisie comme un item de taxonomie. Cette unité de base a été définie par Mayr (Mayr, 1942) comme étant « un groupe de populations naturelles réellement ou potentiellement interfécondes, isolé de tout autre groupement analogue ». Cette notion est totalement différente de la notion « typique » (c.-à-d. de type), cette dernière ne reconnait que les caractères morphologiques. D'autres chercheurs affirment qu'une espèce, et malgré le flux génétique entre les générations au cours du temps, maintient une identité génétique propre (Cracraft, 1983). C'est sur cette identité que repose la phylogénétique des espèces. Du point de vue écologique, une espèce est tout organisme adapté à sa niche écologique (Donald et Alger, 1993).

3. 2.2 Choix des espèces

Nous avons choisi notre ensemble des espèces pour plusieurs raisons. Tout d'abord, par le fait que l'étude de Tahiri *et al* (2012) a déjà exposé ces données ce qui aidera à comparer notre approche à la leur. De plus, ces données ont des

caractéristiques diversifiées, à la fois génétiquement que géographiquement. Par exemple, leur spectre alimentaire large et varié, leur capacité d'explorer de nouveaux milieux et d'y reconstruire leurs niches écologiques[3] et enfin leurs fonctions importantes dans l'écosystème (par exemple les cervidés nettoient la forêt en se nourrissant des pouces d'arbre, les prédateurs régularisent la population animale) (Sakai *et al*, 2001).

Les données utilisées étaient dans leur état brut, mais ne répondaient pas exactement à notre problématique. Afin de les modéliser, nous avons procédé à un prétraitement des données. Toutefois, la provenance conservait un certain degré de confiance qui les rend assez fiables. Nous avons analysé :

1. Une base de données provenant de l'université de McGill et construite à partir des informations diffusées par des sondes étalées à travers le monde. Ces sondes transmettent des renseignements sur la présence ou l'absence des carnivores dans un milieu environnemental quelconque.

[3] . La niche écologique est un des concepts théoriques de l'écologie. Il traduit à la fois : la *« position »* occupée par un organisme et la *« somme des conditions »* nécessaires à une population viable de cet organisme.

2. Un autre fichier de l'université de McGill montrant la distribution des espèces sur des zones géographiques en fonction des paramètres environnementaux (par exemple les précipitations moyennes et la température moyenne).

La liste des espèces était la suivante :

Nom de l'espèce	Nombre d'espèces de la même famille	Nom de la famille carnivore
Ursus maritimus *Ursus arctos* *Ursus americanus*	3	*Ursidae*
Odobenus rosmarus	1	*Odobenidae*
Phoca groenlandica *Phoca fasciata* *Phoca largha* *Phoca vitulina* *Phoca hispida = Pusa hispida* *Halichoerus grypus* *Cystophora cristata* *Mirounga angustirostris* *Erignathus barbatus*	9	*Phocidae*
Callorhinus ursinus *Eumetopias jubatus* *Zalophus californianus* *Arctocephalus townsendi*	4	*Ortaiidae*
Bassariscus astutus *Nasua narica* *Procyon lotor*	3	*Procyonidae*
Martes americana *Martes pennanti* *Mustela nivalis* *Mustela erminea* *Mustela frenata*	16	

Mustela vison *Mustela nigripes* *Lontra canadensis = lutra canadensis* *Enhydra lutris* *Gulo gulo* *Taxidea taxus* *Mephitis mephitis* *Mephis macroura* *Spilogale putorius* Suite		
Spilogale pygmaea = Spilogale gracilis *Conepatus mesoleucus = Conepatus leuconotus*		*Mustelidae*
Canis lupus *Canis rufus* *Canis latrans* *Urocyon cinereorgenteus* *Urocyon littoralis* *Vulpes vulpes* *Alopex lagopus* *Vulpes macrotis* *Vulpes velox*	9	*Canidae*
Lynx canadensis *Lynx rufus Panthera onca* *Puma yaguarondi = Herpailurus yaguarondi* *Puma concolor* *Leopardus pardalis* *Leopardus weidii*	7	*Felidae*

Tableau 3.1 La liste de 52 espèces carnivores considérées dans le projet.

3.2. 3 Description des familles de groupe des carnivores utilisées :

Dans cette section, nous décrirons brièvement les différentes familles d'espèces choisies. Dans l'ensemble définitions qui suivent, nous ferons référence à l'encyclopédie « *Mammals species of the world* » :

Les Canidaes : Ils appartiennent à la famille de classe mammifère, ordre carnivore, sous ordre caniformia. Ils sont caractérisés par leurs nombreuses molaires et leurs griffes non rétractiles comme le renard.

Les Felidaies : Ces espèces sont classées sous la famille mammifère carnivore de sous-ordre feliformia. La plupart des espèces de cette famille possèdent 30 dents, des griffes rétractiles et s'appuient sur leurs doigts en marchant, par exemple le lynx canadien.

Les Mystelidaes : Le mot d'origine est mustela en latin (c.-à-d. belette), cette famille est de sous ordre caniformia, les espèces appartenant à ce classement se caractérisent par des glandes très développées qui sécrètent une odeur musquée en cas d'attaque, par exemple le Furet.

Les Ursidaes : Ils sont aussi de sous-ordre caniformia. C'est la famille des ours qui sont caractérisés par de grands corps, long museau et un pelage dense. Les espèces appartiennent toutes au groupe des carnivores à l'exception du grand

panda qui se nourrisse essentiellement de bambous. Le grand Panda n'est pas considéré dans notre jeu de donnée.

Les Phosidaes : Cette famille de groupe ne possède que 18 espèces actuellement encore vivantes, par exemple les phoques et les éléphants de mer. À noter que cette famille a perdu des pigments rétiniens au cours de leur évolution, cette perte est due à une adaptation avec leur milieu aqua-terrestre (Peichl *et al*, 2001).

Les Ortidaies : Les espèces de ce groupe appartiennent au sous ordre conformai. Ces espèces partagent leur vie entre la mer et la terre ferme, leur permettant d'être de redoutables prédateurs à la chasse que d'avoir d'excellente condition de natation. Par exemple le lion de mer.

Les Prociyonidae : Cette famille a été classée après l'apparition de la phylogénie moléculaire. Elles appartiennent au sous ordre de caniformia. Le raton laveur fait partie de cette famille.

Les Odobenidae : Mammifère carnivore de sous ordre conformai, cette famille possède une seule espèce vivante c'est le morse.

Afin d'explorer et de réaliser une connaissance plus approfondie de nos données, nous avons réalisé un heatmap (voir Figure 3.1). À partir de la matrice de dissimilarité (voir Figure 5.2) entre chaque paire d'espèces (voir le tableau 3.1 pour plus de détail concernant la liste des espèces), nous avons pu mettre en

corrélation les espèces partageant le même milieu géographique. Pour réaliser cette étape, nous avons utilisé heatmap.2 de la librairie 'gplots' dans R (version 3.2.2) (Ihaka et Gentleman, 1996). Cet heatmap met en correspondance l'intensité des distances selon une gamme de couleurs. Le rouge représente les espèces qui sont toujours ensemble, ce qui explique la diagonale (car il s'agit d'une espèce avec elle-même). Le jaune (ou des teintes pâles) indique les espèces qui sont très fréquemment ensemble inversement à la couleur bleue (ou des teintes foncées) qui présente les espèces qui sont disjointes.

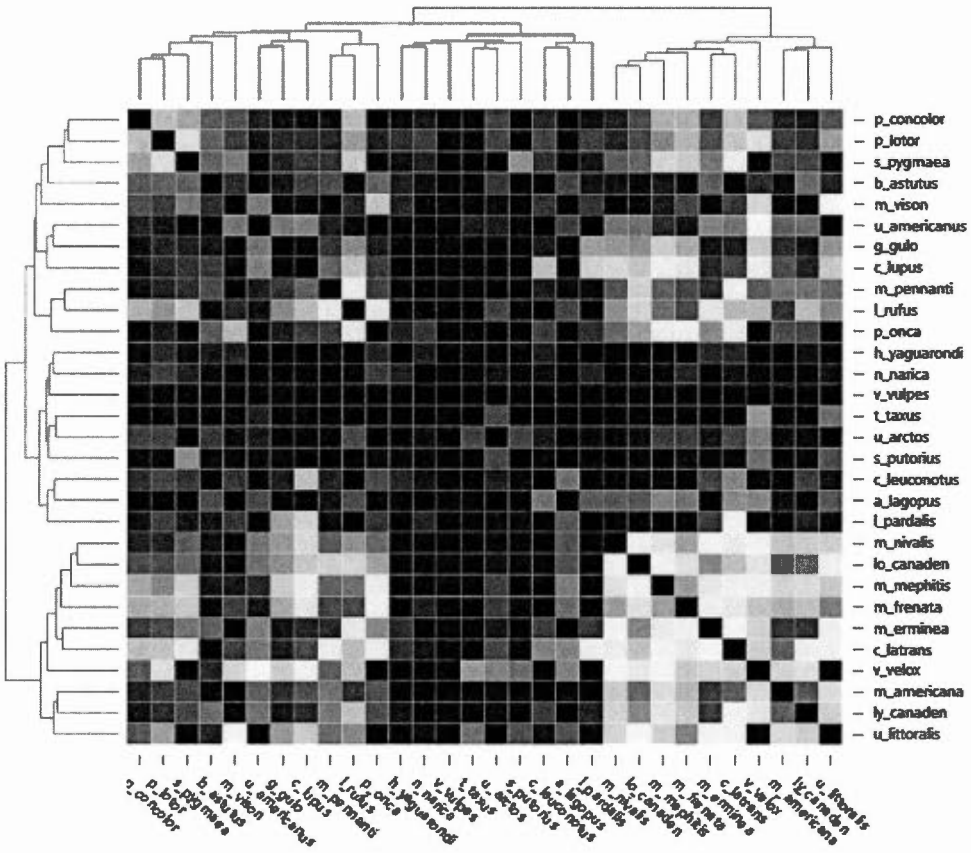


Figure 3.1 : Heatmap des 32 espèces étudiées dans notre travail.

Nous constatons clairement en observant la Figure 3.1 que trois groupes d'espèces suivent le même modèle selon la condition climatique température moyenne. Le premier groupe est constitué de *p. concolor, p. lotor, s. pygmaea, b. astutus, m. vision, u. amaricanus, g. gulo, c. lupus, m. pennanti et l. rufus*. Le deuxième groupe est constitué de *m. pennanti, l. rufus, p. onca, h. yaguarondi, n. narica, v. vulpes, t. taxus, u. arctos, s. putorius, c. leuconotus et a. lagopus*. Le troisième

groupe est constitué de *l. pardalis, m. nivalis, lo. canaden, m. mephtis, m. frenata, m. erminea, c. latrans, v. velos, m. americana, ly. canaden, u. littolis.*

3.3 Données géographiques

La zone géographique prise en considération dans notre étude est l'Amérique du Nord. Cette zone comprend les États-Unis d'Amérique, le Canada et le Mexique. Elle est située dans l'hémisphère nord-ouest de la terre. Les motifs justifiant le choix de cette zone faunique pour réaliser cette étude sont multiples : en effet, selon la CCE (*Commission de coopération environnementale)*[4] cette région forme un cadre nommé « *région écologique de l'Amérique du Nord* » et constitue pour les chercheurs, les universitaires et les organismes gouvernementaux une base de recherches géologique, génétique et géographique.

Chacune des parties échantillonnées possède ses propres caractéristiques. D'après les géologues, l'Amérique du Nord peut être divisée en 15 écorégions. Une écorégion est distinguée par un système d'homogénéité au niveau d'écosystème, des paramètres climatiques et de la qualité du sol. La diversité écologique de cette zone a constitué un avantage majeur. C'est pour toutes ces raisons que notre étude portera sur ces données géographiques. La physiologie et le mode de reproduction

[4] http://www.cec.org/

des espèces sont directement affectés par les paramètres climatiques qui règnent dans une zone géographique donnée; citons par exemple, la température et le régime hydrique. La superficie étudiée (c.-à-d. l'Amérique du Nord) marque une variété radicale au niveau de la température et de la précipitation (c.-à-d. régime hydrique). Cette superficie est de 24 930 333 km^2. Notons que sur cette énorme étendue de terre, les mammifères carnivores ont construit leurs niches écologiques et se sont bien adaptés à leurs milieux (Kays et Wilson, 2009).

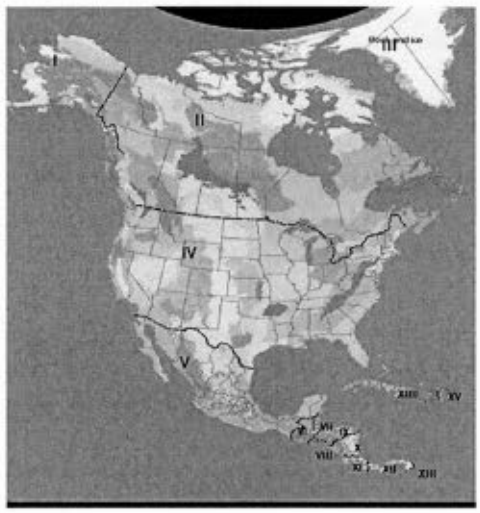


Figure 3.2: La subdivision de la zone étudiée en 15 écorégions. Source https://naturalhistory.si.edu/mna.

3.3. 1 Prétraitement des données phylogéographiques

Avant d'inférer les arbres phylogéographiques, nous avons développé un algorithme correspondant à une étape fondamentale et obligatoire, à savoir la

préparation des données. Cette étape a abouti à la reconstruction de 30 arbres phylogéographiques décrivant la distribution des espèces dans les 15 écorégions, selon deux paramètres climatiques : la température moyenne et les précipitations moyennes (voir Figure 3.2).

À l'aide d'un script développé en langage Perl (version 5.10.1), nous avons reproduit 30 arbres phylogéographiques. Dans ce chapitre, nous concentrerons à la phase de la reconstruction des arbres phylogéographiques, alors nous parlerons de la reconstruction des arbres phylogénétiques dans le chapitre suivant. La procédure complète est décrite dans le flux de travail (c.-à-d. workflow) suivant (Figure 3.3) :

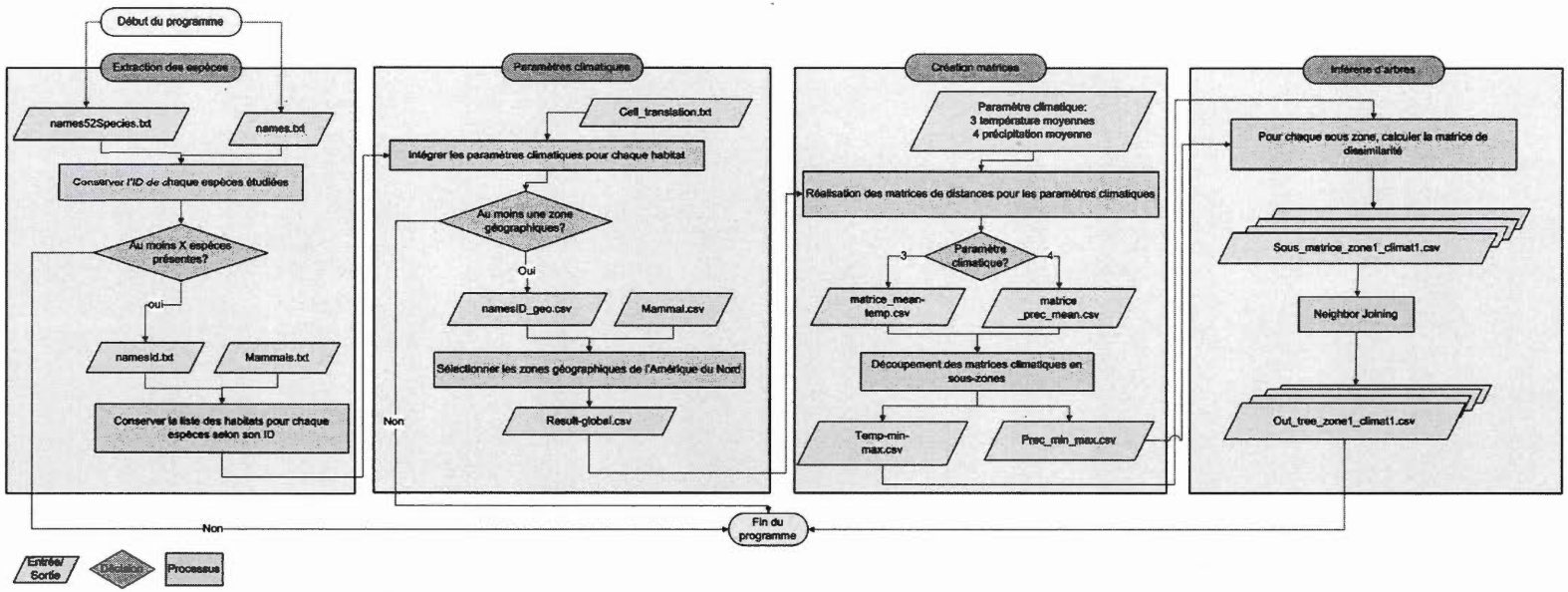


Figure 3.3 : Flux de travail montrant le prétraitement des données phylogéographiques.

3.3. 2 Description du flux de travail

Algorithme 3.1 : Extraction des espèces

Entrée:
2 fichiers:
Names52Species.txt[1]
names.txt[2]
Sortie:
1 fichier:
NamesID.txt[3]
Variables :
Booléen valide=vrai
Int nb_espece// nombre total d'espèces
nb_espece=0// initialisation du nombre d'espèces

Pour chaque esp_1 dans names.txt **Faire**
Pour chaque esp_2 dans names52Species.txt **Faire**
Si esp_1==esp_2 **alors**
Conserver (esp_2; ID_2)
nb_espece++
Fin Si
Fin Pour
Fin Pour
Si nb_espece < 3 **Alors**
Pour Chaque (esp_2; ID_2) dans Mammal.txt[4]
Conserver (esp_2; ID, présence/absence) dans NamesID.txt[3]
Retour valide
Sinon valide = Faux

Nous avons utilisé les fichiers suivants :

Description des fichiers :

1. *Names52Species.txt* : Ce fichier contient la liste des 52 espèces qu'on a choisi d'étudier dans l'Amérique du Nord.

2. *Names.txt* contient les données sources (c.-à-d. la liste de toutes les espèces et leur identifiant ID).

3. *NamesID.txt* : ce fichier englobe les espèces communes (30 mammifères) avec leur ID.

4. *Mammal.txt* : ce fichier contient la liste d'espèces avec leur présence/absence dans toutes les zones géographiques (c.-à-d. à travers le globe). Étant donné que notre étude se porte sur l'Amérique du Nord, alors nous avons filtré dans ce fichier la zone limitée par les deux variables X (longitude =-178; latitude=8) et Y (longitude=-44; latitude=74).

Algorithme 3.2 : Paramètres climatiques

Entrée :

3 fichiers :

NamesID.txt[3]

cell_translation.txt// *liste de toutes les zones géographiques dans notre base de données, soit 16 933 zones.*

Mammal.csv// *liste de différents paramètres géographiques (c.-à-d. latitude, longitude ainsi que différents paramètres climatiques tels que température moyenne et précipitation moyenne*)

Sortie :

1 fichier : Result_global.csv

Variables :

Int nb_espèce// nombre total d'espèces

nb_espèce=0// initialisation de nombre d'espèces

FIIE NamesIDgeo[5]//fichier temporaire

Pour Chaque (esp_1; ID_1; présence/absence) dans NamesID.txt **Faire**

Pour Chaque (esp_2; ID_2; présence/absence) dans cell_translation.txt **Faire**

Si ID_1==ID_2 **Alors**

Conserver (esp_1; ID_1; zone) dans NamesIDgeo;

nb_espece++

Fin Si

Fin Pour

Fin Pour

Pour Chaque (esp_1; ID_1; zone) dans NamesIDgeo **Faire**

Pour Chaque (esp_2; ID_2; zone) dans Mammal.csv[6] **Faire**

Si ID_1==ID_2 **Alors**

Conserver (esp_1; ID_1; zone; liste de paramètres climatiques) dans Result_gobal.csv[7]

nb_espece++

Fin Si

Fin Pour

Fin Pour

Description des fichiers (suite) :

5-NamesIDgeo.csv : Ce fichier contient la liste de 30 espèces considérées avec leur présence ou absence dans les zones géographiques étudiées.

6-Mammal.csv : Ce fichier inclut les paramètres climatiques en fonction de secteursgéographiques. On l'a utilisé pour incorporer les paramètres climatiques.

7-Result_global.csv : Contient toutes les informations nécessaires pour le calcul des matrices de distance. Il résulte de deux premiers algorithmes.

Algorithme 3.3 : Reconstruction des arbres phylogéographiques.

Entrée :

1 fichier :
Result_global.csv[7]

Sortie :

30 arbres phylogéographiques

Variables :

Int choix
Int nb_zone_total =15 //15 écorégions
Int nb_espece_ensemble// nombre de fois pour que 2 espèces soient ensemble
Int nb_zone_ensemble// nombre de zones qui contiennent les mêmes espèces

//Calcul des matrices de dissimilarité de la précipitation moyenne dans les zones géographiques

Si (Choix==2) **Alors**
Pour Chaque couple (esp_i, esp_j) **Faire**
Pour (K=0; k<=nb_zone_total; K++)
Si [(esp_i, esp_j); k] ==1; *//les deux espèces se trouvent ensemble dans la même zone K,*

nb_espece_ensemble++
nb_zone_ensemble++
Fin Si
Fin Pour

$$d(esp_i; esp_j) = 1 - \frac{nb\_espece\_ensemble}{nb\_zone\_ensemble}$$

Conserver d (esp_i; esp_j) dans Sous-matriceK //sous-matrice dans la zone K

//Appliquer l'algorithme NJ à sous-matrice pour construire l'arbre phylogéographique correspondant
Sous-matrice < — (*NJ*)
Conserver arbreK//Arbre dans la zone K
Fin Pour

//Calcul des matrices de dissimilarité de la température moyenne dans les zones géographiques

Sinon Si (Choix ==3) **Alors**
Pour Chaque couple (esp_i; esp_j) **Faire**
Pour (K=0; k<=nb_zone_total; K++)
Si [(esp_i; esp_j); k] ==1; *//les deux espèces se trouvent ensemble dans la même zone K,*
nb_espece_ensemble++
nb_zone_ensemble++
Fin Si
Fin Pour

$$d(esp_i; esp_j) = 1 - \frac{nb\_espece\_ensemble}{nb\_zone\_ensemble}$$

Conserver d (esp_i; esp_j) dans Sous-matriceK //sous-matrice dans la zone K

//Appliquer l'algorithme NJ a sous-matrice pour construire l'arbre phylogéographique correspondant
Sous-matrice < — (*NJ*)
Conserver arbreK//Arbre dans la zone K
Fin Pour

3.4 Données génétiques

3.4. 1 Acquisition des séquences protéiques

Après le choix des espèces, nous avons récupéré leurs séquences nucléotiques depuis la base de données GenBank de NCBI (Benson *et al*, 2008). Dans cette base de données, nous trouverons plusieurs versions des séquences nucléotidiques décrivant le même gène. À l'aide d'un script en BioPerl, nous avons saisi la séquence la plus longue pour chaque gène. Notre choix pour les séquences protéiques se base sur la stabilité par rapport aux séquences nucléotidiques. Nous avons aligné ces fragments de séquences en utilisant l'outil ClustalW (Li, 2003). Le tableau 3.2 représente la liste des protéines sélectionnées dans GenBank.

Adenosine A3 receptor
Apolipoprotein B
ATP synthase F0 subunit 6
ATP synthase F0 subunit 8
Brain derived neurotrophic factor
Breast cancer susceptibility protein 1
Cytochrome Oxidase Subunit I
Growth hormone receptor
NADH dehydrogenase subunit 1
NADH dehydrogenase subunit 2
NADH dehydrogenase subunit 4
NADH dehydrogenase subunit 4L
NADH dehydrogenase subunit 5

NADH dehydrogenase subunit 6
Nicotinic cholinergic receptor alpha polypeptide 1 precursor
Prepronociceptin
Recombination activating protein 1
Retinoid Binding Protein
Rhodopsin
Sex determining region Y protein
Von Willebrand factor

Tableau 3.2 Liste des protéines sélectionnées dans GenBank.

3.4. 2 Description des séquences protéiques utilisées :

Dans cette section, nous dériverons les protéines utilisées ci-dessus.

- Le récepteur de l'adénosine A3 :

Le récepteur de l'adénosine A3 est lié aux récepteurs des protéines G et réparti dans plusieurs tissus de l'organisme. Les chercheurs ont remarqué que son degré d'influence s'accentue en cas d'inflammation surtout dans les tissus pulmonaires, les testicules et dans le foie. Il inhibe la production d'AMPc (adénosine 5'—monophosphate cyclique) conduisant ainsi à la libération des stocks calciques (Tahiri *et al*, 2012). Il y transmet aussi une fonction soutenue de cardioprotectrice au cours de l'ischémie cardiaque.

- L'apolipoprotéine B :

L'apolipoprotéine B est la constituante principale des LDL (c.-à-d. lipoprotéine de basse densité) et vLDL (c.-à-d. lipoprotéine de très basse densité). Des mutations génétiques (insertion et délétion) dans la partie codante de l'apo B modifient les concentrations lipidiques et augmentent par la suite le risque cardiovasculaire dans la population générale (Ingelsson *et al*, 2007; Pischon *et al*, 2005).

- Les protéines (ATP-6 et ATP-8)

Les protéines ATP-6 et ATP-8 constituent deux sous-unités du complexe transmembranaire de type F :

- L'ATP synthase F0 subunit 6 (ATP-6) également connu comme le complexe V qui est constitué de 14 sous unités nucléaires et 2 mitochondriales. Il s'agit d'un élément essentiel de la chaîne à protons. Son rôle est observé dans la translocation de protons à travers la membrane.
- L'ATP synthase F0 subunit 8 (ATP-8) est une sous-unité de l'ATP mitochondriale. Elle semble être une composante intégrale de la tige dans le stator mitochondrial de la levure F-ATPases (Tahiri *et al*, 2012).

- Le facteur du cerveau dérivé neurotrophique (en anglais brain derived neurotrophic factor, BDNF) est une protéine codée par le gène BDNF chez les humains. Il est trouvé dans des plusieurs tissus cellulaires et non uniquement dans les cerveaux localisés essentiellement dans l'hippocampe, le cortex cérébral et les tissus rétinaux. Le BDNF agit sur certains neurones du système nerveux central et du système nerveux périphérique. Il est impliqué dans la survie des neurones existants et encourage la croissance et la différenciation de nouveaux neurones et des synapses (Yamada et Nabeschima, 2003; Bekinschtein *et al,* 2008).
- Breast cancer susceptibility protein 1 (BRCA1). La protéine BRCA1 est une protéine est découvert en 1990[5] par Mary claire King. Cette protéine impliquée dans plusieurs cancers héréditaires comme le cancer du sein, des ovaires et de la prostate. Elle joue un rôle dans le maintien de la stabilité génomique et dans la réparation de l'ADN (Park *et al*, 2000).
- Le cytochrome oxydase de la sous-unité 1 du complexe IV (CO-1). La protéine CO-1 est une oxydoréductase membranaire. Elle joue un rôle initial dans la chaîne respiratoire et s'avère essentielle dans la fonction vitale du métabolisme. Il s'agit d'un grand complexe d'enzymes transmembranaires

[5] https://fr.wikipedia.org/wiki/BRCA1

trouvé dans les bactéries et les mitochondries (Liu *et al*, 2002; Tahiri *et al*, 2012).

- Le récepteur de l'hormone de croissance (en anglais growth hormone receptor ou GHR). La protéine GHR codée par le gène GRH. Il s'agit d'un polypeptide de 44 acides aminés. Cette protéine stimule la libération de la somatotrophine (STH) qui à son tour incite la croissance et la génération des cellules chez les mammifères.
- Les 6 protéines suivantes (NADH1, NADH2, NADH4, NADH4L, NADH5 et NADH6).

  L'ensemble de 6 protéines appartiennent au complexe de nicotinamide adénine dinucléotide (NADH, NAD). Le NAD est une coenzyme d'oxydoréduction, il est présent dans toutes les cellules vivantes. Ce coenzyme peut se présenter sous deux formes : NAD+ est un agent d'oxydation et NADH un agent de réduction. Cependant, la fonction principale de NADH est le transfert des électrons. Il s'agit du complexe I intervenant dans la chaîne de transport d'électrons. Ce complexe se situe dans la membrane interne des mitochondries, dont le rôle est la production d'énergie (Yusnita *et al*, 2010).
- Les protéines de rétinoïde de liaisons (RBP).

  Les protéines de rétinoïde de liaisons (RBP) appartiennent à la une famille de protéines ayant des fonctions diverses. Ce sont de protéines de transport du

plasma sanguin et du cytosol. L'ensemble de ces protéines est susceptible de se lier au rétinol, au rétinal et à l'acide rétinoïque afin de contrôler leurs effets dans l'organisme. L'acide rétinoïque et le rétinol jouent des rôles cruciaux dans la modulation de l'expression des gènes et le développement global de l'embryon[6].

- La rhodopsine, également connue sous le nom de pourpre rétinien.

  C'est un pigment biologique protéique photosensible dans les cellules photoréceptrices de la rétine. Elle est responsable de la sensibilité de l'œil à la lumière.

- Sex determining region Y protein (SRY).

  La protéine SRY définit le sexe chez les mammifères est localisée sur la branche courte du chromosome Y. connue sous le nom « Sry » pour le mammifère et SRY pour les humains. Son rôle principal s'avère crucial dans le développement du phénotype male. Des mutations sur cette protéine peuvent entraîner des femelles XY ou bien une translocation d'une partie du chromosome Y contenant la protéine SRY sur le chromosome X, provoquant ainsi le syndrome XX mâle (Iliopoulos *et al*, 2004).

- Facteur de Von Willebrand est une glycoprotéine du sang impliqué dans l'hémostase. Elle est impliquée dans un grand nombre d'autres maladies, par

[6] https://fr.wikipedia.org/wiki/Prot%C3%A9ine_de_liaison_du_r%C3%A9tinol

exemple le syndrome de Heyde, le syndrome d'hémolytique et d'urémique (Sadler, 1998).

La Figure 3.4 représente un heatmap obtenu à partir de la matrice de la distance *RF* normalisée entre les deux ensembles de données (voir tableau 5.6). Nous avons pu mettre en corrélation les protéines et leur évolution en fonction des différents paramètres climatiques. Pour réaliser cette étape, nous avons obtenu un heatmap en utilisant heatmap.2 de la librairie 'gplots' dans R (version 3.2.2) (voir Figure 3.4). Cet heatmap met en correspondance l'intensité des distances selon une gamme de couleurs. Nous constatons clairement que la protéine SRY se démarque des autres par son profil typique en relations aux différentes conditions climatiques. Les protéines rhodopsin et Polypetide 1 précurseur sont sensibles aux mêmes environnements. Enfin les autres protéines sont homogènes dans leurs évolutions climatiques.

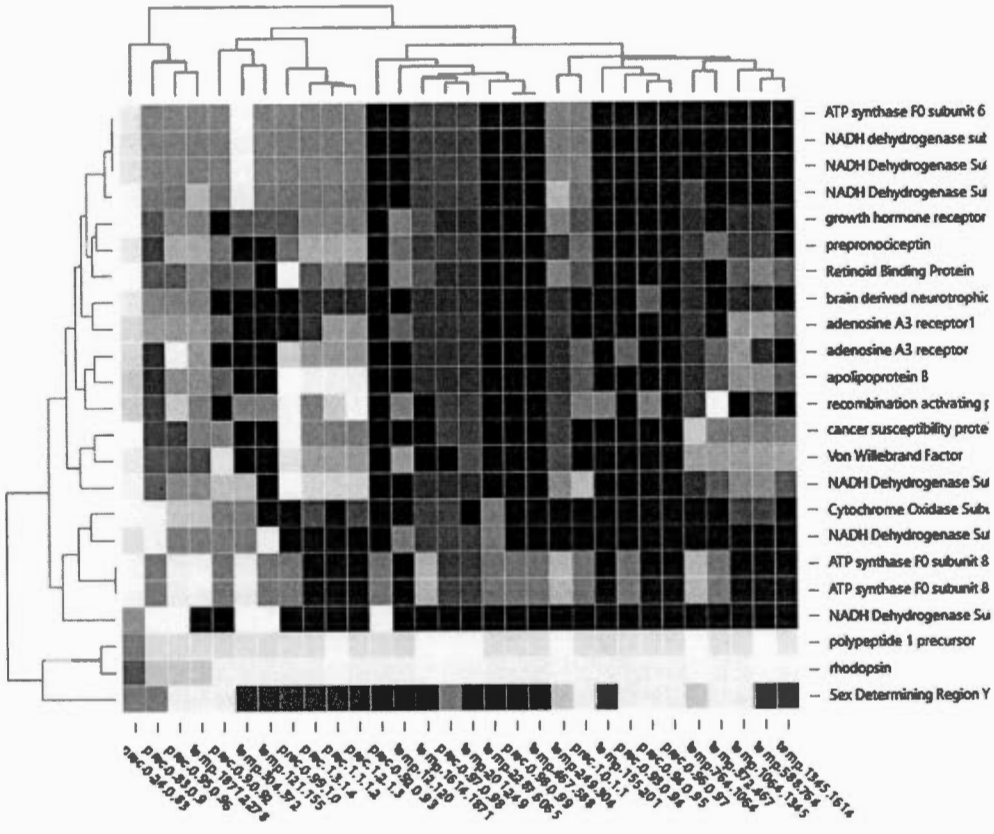


Figure 3.4 : Heatmap du patterne des protéines étudiées en relation avec les conditions climatiques.

3.5 Conclusion

Dans ce chapitre, nous avons présenté les données utilisées par notre algorithme. Nous avons parlé en détail de trois types de données : à savoir, les données phylogéographiques, les données géographiques et les données génétiques. Nous avons indiqué leur provenance, leur préparation, ainsi que leur prétraitement. Nous avons présenté aussi en détail nos algorithmes et étapes de prétraitement de

données considérées comme étant une introduction à l'algorithme général. Le chapitre suivant explique en détail notre objectif principal, soit de « développer un algorithme pour retrouver les relations entre la génétique et la phylogéographique des espèces ».

CHAPITRE IV

MÉTHODOLOGIE ET ALGORITHME

4. 1 Introduction

Après une présentation approfondie des concepts fondamentaux permettant une meilleure compréhension du sujet de ce mémoire (tel que l'arbre phylogénétique et l'arbre phylogéographique) et une description détaillée des données choisies, nous traiterons dans ce chapitre la problématique initialement prévue (voir chapitre I) à savoir : la méthode et les étapes de développement de l'algorithme pour retrouver les relations entre la phylogénie et la phylogéographie des espèces. Nous expliquerons la méthodologie, l'architecture et les éléments qui ont servi au développement de notre objectif. Nous fragmenterons la phase du prétraitement des données, en balisant les étapes de la reconstruction des arbres phylogénétiques. Nous expliquerons les différentes étapes pour l'obtention des matrices de dissimilarité en utilisant certains paquets du programme *PHYLIP* (Plotree et Plotgram, 1989). Nous indiquerons clairement les différents modèles d'évolution, nous permettant de retrouver les arbres en format Newick. Nous

décrirons l'algorithme que nous avons utilisé ainsi que notre méthodologie. Nous démontrerons que notre algorithme a une complexité moindre que celui adopté par Tahiri (Tahiri *et al*, 2012) tout en obtenant de bons résultats. Enfin, nous parlerons des différentes métriques dédiées à la comparaison d'arbres phylogénétiques.

4. 2 Prétraitement des données

Dans le chapitre précédent, nous avons collecté et préparé les données afin de valider la performance de notre algorithme. Toutefois, la reconstruction des arbres phylogénétiques est basée sur la théorie des graphes (voir le chapitre II). Dans la première version de l'algorithme, Tahiri *et al* (2012) ont mis en évidence deux protéines : **SRY** et **NADH-6.** Ces deux protéines sont pertinentes d'un point de vue de l'influence de l'environnement sur l'activation de leurs gènes respectifs. Dans cette version, nous avons considéré ces deux protéines et examiné des nouvelles.

4. 2.1 Inférence d'arbres phylogéographiques

La reconstruction d'arbres phylogéographiques est une étape préliminaire à notre projet. En effet, ce sera une des entrées principales à l'algorithme. Dans le

chapitre III, nous avons présenté les données géographiques, en nous attardant sur la source des données, la zone géographique adoptée et la variation des paramètres climatiques dans cette zone. Dans ce chapitre, nous expliquerons la phase de l'inférence des arbres, tout en nous basant sur le fichier résultant de l'étape précédente soit **result-global.csv** (voir chapitre III). Cette phase est considérée comme une introduction au développement de l'algorithme. Le fichier **result-global.csv** contient plusieurs informations dont : la présence/absence des espèces dans les différentes régions, la longitude, la latitude, la température moyenne et la précipitation moyenne. Notons que ce fichier est un sous ensemble des données contenu dans mammal.txt[7]. Plusieurs observations ont été faites à la suite de cette étape :

- Si on se base sur les données dans le fichier **result-global.csv**, les données ne permettent pas d'obtenir des arbres de référence résolus (2590 arbres de référence pour 2590 points géographiques).
- La distribution des espèces dans les régions n'est pas réaliste : chaque point géographique constitue l'intersection de la longitude et la latitude en indiquant la présence ou l'absence d'une espèce vis-à-vis de ce repère et est très peu employé dans la traçabilité des espèces.

[7] Le fichier mammal.txt contient la base des répartitions des espèces à travers le globe. Nous avons filtré ce dernier selon la zone qui nous intéresse, et aussi selon les espèces étudiées. Le résultat de ce filtre sera indiqué dans result-global.csv

D'où la subdivision de la zone géographique en 15 écorégions choisie de telle sorte que chaque écorégion contient le même nombre de points, soit 171 points géographiques pour chaque écorégion. Une telle subdivision a abouti à une redéfinition de la zone géographique en analysant la distribution géographique des espèces par sous-zone c.-à-d. des zones ayant un paramètre climatique similaire (température et précipitations moyennes). Ainsi, sur le fichier **result-global.csv** nous avons créé deux divisions : une verticale suivant les paramètres climatiques et une autre horizontale suivant les zones géographiques. À la sortie, nous avons obtenu 30 fichiers déterminant la distribution des espèces dans les écorégions selon les deux paramètres climatiques : température et précipitation moyennes (voir Figue 4.1).

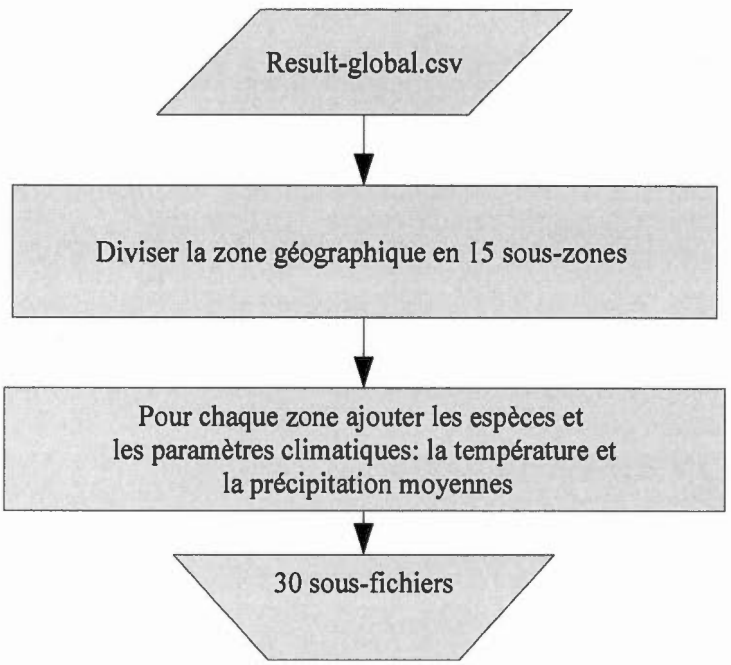


Figure 4.1 : Division de la zone considérée en 15 écorégions. Les 30 sous-fichiers résultants contiennent chacun la distribution des espèces selon un paramètre climatique choisi.

4. 2.2 Matrices de dissimilarité pour les données géographiques

Afin de déterminer les matrices de dissimilarité pour les données géographiques, le calcul a été basé sur le fait que si deux espèces existent dans la même zone géographique alors la distance entre ces deux espèces sera égale à 0, sinon cette distance est égale à 1. L'algorithme correspondant à cette étape est présenté ci-dessous.

Algorithme 4.1 : Calcul des sous-matrices de dissimilarité de chaque écorégion.

Entrée :

30 fichiers :

FILE prec_zon$_1$.csv// *fichier contenant la distribution des espèces dans la zone$_1$ selon le paramètre climatique précipitation moyenne*

FILE prec_zon$_2$.csv// *fichier contenant la distribution des espèces dans la zone$_2$ selon le paramètre climatique précipitation moyenne*

.....

FILE prec_zon$_{15}$.csv// *fichier contenant la distribution des espèces dans la zone$_{15}$ selon le paramètre climatique précipitation moyenne*

FILE temp_zon$_1$.csv// *fichier contenant la distribution des espèces dans la zone$_1$ selon le paramètre climatique température moyenne*

FILE temp_zon$_2$.csv// *fichier contenant la distribution des espèces dans la zone$_2$ selon le paramètre climatique température moyenne*

.....

FILE temp_zon$_{15}$.csv// *fichier contenant la distribution des espèces dans la zone$_{15}$ selon le paramètre climatique température moyenne*

Sortie :

30 matrices de dissimilarité :

mat_prec_zon$_1$.csv,
mat_prec_zon$_2$.csv,
.....
mat_prec_zon$_{15}$.csv,
mat_temp_zon$_1$.csv,
mat_temp_zon$_2$.csv,
.....
mat_temp_zon$_{15}$.csv

Variables :

Int nb_espece_ensemble=0// *initialisation de nombre d'espèces ensemble se trouvant dans la même zone géographique et conditionnée par le même paramètre climatique*

Int nb_zone_ensemble=0// *initialisation de nombre de zones contenant deux ou plus espèces ensemble*

Pour chaque fichier_d'entrée **Faire**
Pour Chaque couple (esp_i; esp_j) **dans** prec_zon_1.csv **Faire**
Pour (K=0; k<=nb_zone_total; K++)
Si [(esp_i; esp_j); k]==0; *//les deux espèces se trouvent ensemble dans la même zone K,*
nb_espece_ensemble++
nb_zone_ensemble++
Fin Si
Fin Pour

$$d(esp_i; esp_j) = 1 - \frac{nb\_espece\_ensemble}{nb\_zone\_ensemble}$$

Conserver d (esp_i; esp_j) dans mat_prec_zon_1.//Sous-matrice dans la $zone_1$
Fin Pour
Fin Pour

4. 2.3 Reconstruction des arbres phylogéographiques en format Newick

Une fois que les matrices de dissimilarité sont prêtes, les arbres phylogéographiques sont reconstruits en faisant appel à l'algorithme (*NJ*) de Saitou et Nei, (1987) du paquet NINJA (Wheeler, 2009). Les arbres résultants sont en format Newick[8].

4. 3 Reconstruction des arbres phylogénétiques

À l'aide d'un programme implémenté en Java (version 7), nous avons reconstruit les arbres phylogénétiques.

[8] Arbre de la forme (B, (A, C, E), D); avec A, B, C, D et E sont les feuilles, cet arbre se termine par (;) toujours.

La procédure tout entière et les programmes externes utilisés sont expliqués dans les sections suivantes.

4. 3.1 Paquet *PHYLIP*

Le paquet *PHYLIP* (*PHYLogeny Inference Package*) (Plotree et Plotgram, 1989) est un ensemble d'algorithmes phylogénétiques implémentés en langage C. Ces algorithmes servent essentiellement à inférer des arbres phylogénétiques suivant plusieurs procédures. On y trouve de nombreuses procédures telles que de *parcimonie*, de *likelihood,* les *matrices de distance*, le *boostraping* et l'arbre *consensus*. Dans notre projet, nous avons employé les méthodes suivantes : *Seqboot, ProDist, DnaDist, Neighbor* et *Consense* de *PHYLIP* (voir Figure 4.3).

- *Seqboot* : ce programme permet de réechantillonner un alignement de séquences multiples (*ASM*). L'incorporation de Seqboot à notre programme nous a permis de générer 100 reliquats de l'ensemble des *ASM.*
- *ProDist* / *DnaDist* : Nous avons fait appel à ces programmes après l'obtention des 100 reliquats de l'*ASM.* En effet, à partir d'*ASM,* ces deux programmes vont calculer les distances d'éloignement de chaque paire de séquence d'*ASM* et ceci pour 100 *ASM.* Le programme *ProDist* est spécifique aux séquences protéiques alors que le programme *DnaDist* est spécifique aux des séquences nucléiques. Notons l'importance de choisir adéquatement le modèle d'évaluation.. Dans cette étude, nous avons utilisé le modèle *kimura-protéine*

pour les séquences protéiques et *kimura -2 paramètres* pour les séquences nucléotidiques.

- *Neighbor* : est un programme qui permet la reconstruction des arbres phylogénétiques en s'appuyant sur l'algorithme Neighbor Joining (*NJ*) par agglomération successive des lignées

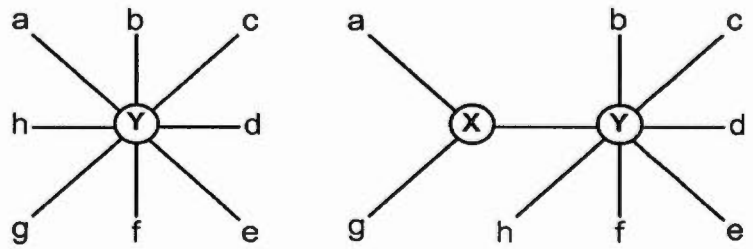


Figure 4.2 : Illustration d'une étape de l'algorithme (*NJ*) (Warren *et al.*,2008).

- *Consense* : À partir des 100 arbres phylogénétiques obtenus à partir des 100 variantes de l'*ASM* d'un même gène défini pour un ensemble d'espèces, *Consense* construit un arbre consensus représentant l'évolution de chaque gène (Sadler, 1989).

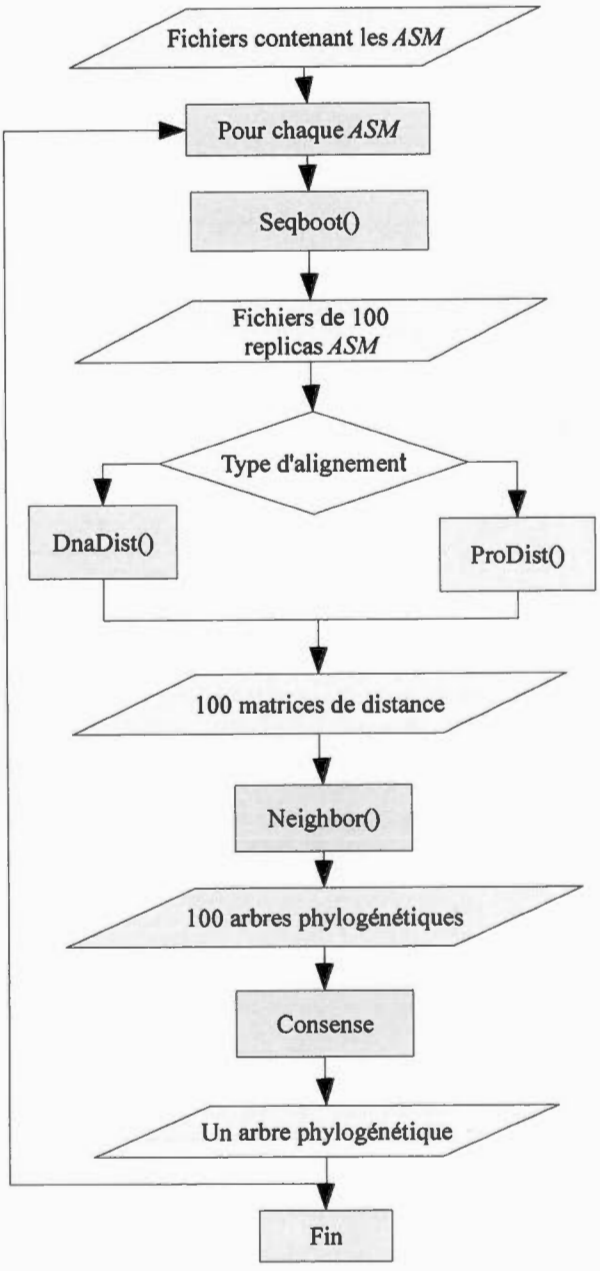


Figure 4.3 : Flux de travail illustrant la reconstruction des arbres phylogénétiques à partir d'alignement de séquences multiples.

4. 4 Description de l'algorithme

Dans cette partie, nous présenterons la méthodologie adoptée dans l'algorithme en décrivant ses différentes phases.

4. 4.1 Méthodologie

Notre algorithme est divisé en trois principales phases :

4. 4.1.1 *Validation des paramètres d'entrée*

Dans un premier temps, une étape s'avère indispensable, la validation des données d'entrée. On dispose plusieurs variables à valider au début de l'algorithme :

- L'ensemble des arbres phylogénétiques ($T_{génétique}$).
- L'ensemble des arbres phylogéographiques ($T_{géographique}$).
- Données génomiques : Alignements multiples de séquences de types protéiques ou nucléotidiques.
- Une valeur entière *TF* qui représente la taille de la fenêtre coulissante avec $TF \geq 1$.
- Une variable *P* décrivant le pas d'avancement sur la fenêtre coulissante de l'alignement multiple des séquences avec $P \geq 1$.
- Une variable *TA* désignant la taille complète de l'alignement multiple tel que $TA \geq 1$.

Notons que pour chaque paire d'arbres (génétique et géographique) doit avoir un ensemble d'espèces chevauchantes de plus de trois.

4.4.1.2 *Calcul de la distance entre les arbres génétiques et les arbres géographiques*

Une fois que la validation des paramètres d'entrées est faite, nous calculerons la distance topologique de Robinson et Foulds (Robinson et Foulds, 1981) entre chaque paire d'arbres comprenant un arbre phylogénétique et un arbre géographique. La distance résultante est récupérée dans une matrice ***D*** tel que :

$D_{ij} = D[i][j]$ *est la distance RF entre* $T_{\text{génétiques}}$ *et* $T_{\text{géographiques}}$

Notons par ailleurs que nous avons normalisé la distance *RF* pour faciliter l'étape de comparaison d'arbres. Nous avons aussi utilisé d'autres types de distance, telle que la distance de bipartition. Une fois que la matrice de distance entre les deux types d'arbres est calculée, nous avons cherché le minimum de la distance *RF* normalisée pour chaque paramètre climatique correspondant à chaque gène (voir Algorithme 4.2).

Algorithme 4.2 : Trouver les inférences à minimum de divergence

---

Entrée :

$T_{génétique}$ *//ensemble d'arbres génétiques*

$T_{géographique}$ *//ensemble d'arbres géographiques*

Sortie :

Liste de gènes d'intérêt,

Variables :

Matrice **D**

//Trouver la paire (Ti, Tj) de distances RF minimales

Pour (i= 1; i≤ nb d'arbres dans $T_{géographique}$; i++) **Faire**

Pour (j= 1; j≤ nb d'arbres dans $T_{géographique}$; j++) **Faire**

//Calculer le nombre d'espèces communes entre Ti et Tj

X = nb_espece_commune;

Si X≥3 **Faire**

//Calcul de la distance RF normalisée

$$Dij=\frac{\left|\sum T_i-\sum T_j\right|+\left|\sum T_j-\sum T_i\right|}{2X-6}$$

Conserver Dij dans **D**

Fin Pour

Fin Pour

//Trouver le minimum dans chaque ligne de **D**

Pour (k= 1; i≤X; K++) **Faire**

RF_{min} = min (D_{11}, D_{12}.....D_{1k})

Trouver le gène correspondant à RF_{min}

Conserver le gène dans la liste des gènes d'intérêt

Fin Pour

4. 4.1.3 *Traitement des gènes trouvés*

Notre but était de trouver les positions qui se montraient les plus liés à des facteurs géographiques et climatiques. Nous avons utilisé l'approche des fenêtres coulissantes. La fenêtre de taille *TF* est fixée sur l'*ASM* du gène en question, en considérant que $TF \leq TA$ tel que *TA* est la taille totale de l'*ASM*. Cette fenêtre débute à la position 1 et se déplace d'un pas *P*. À travers cette stratégie, nous avons pu aussi accélérer les complexités temporelle et spatiale vis-à-vis de l'algorithme développé par Tahiri *et al* (2012). Ces optimisations se manifestent par l'exécution conditionnée du programme *PhyML* et permettent de :

- Réduire le nombre des gènes à analyser en éliminant ceux qui ne sont pas informatifs (voir section 4.4.1.2).
- Adopter une valeur seuil de la distance normalisée *RF* entre l'arbre géographique et l'arbre génétique. Si la distance calculée est plus petite que la valeur seuil, le cas est rejeté.
- Les fenêtres coulissantes sont traitées d'une manière séquentielle, mais vu que le traitement de chaque fenêtre est indépendant de celui des autres, notre algorithme peut être exécuté en parallèle en utilisant des serveurs tels que Spark Apache.[9]

[9] http://spark.apache.org/docs

En effet, *PhyML* permet la reconstruction des arbres phylogénétiques en s'appuyant sur les méthodes de vraisemblances. Ces méthodes sont très couteuses en ce qui concerne le temps d'exécution. La phase du prétraitement des données qui concerne la reconstruction des arbres phylogénétiques et phylogéographiques, se base sur les méthodes de distance et qui sont généralement très rapides. La partie « *traitement de chaque gène* » est précédée par une autre validation de données d'entrées :

- *nbTA* (nombre d'alignement) >0.
- *TF* (taille de la fenêtre) ≥ 1.
- *TA-TF*>0 (*TA*= taille de l'alignement)
- Valeur de seuil *RF normalisée* comprise entre 0 % et 100 %.
- *P* (pas d'avancement) ≥1.

Algorithme 4.3 : Traitement des gènes résultants

Entrée :

Liste de gène d'intérêt

Sortie :

Position d'intérêt sur chaque gène,

Variables :

Int nbTA//*nombre d'alignement*
Int TF//*taille de la fenêtre*
Int P //*pas d'avancement*

//Trouver la(s) position(s) significative(s) sur chaque gène trouvé dans la phase précédente

Pour chaque gène dans la liste de gène d'intérêt **Faire**

Pour (i= 1; i≤ TA, i+P) **Faire**

Conserver gène dans liste de gène d'intérêt

Positionner la fenêtre à la position i de l'alignement

Récupérer l'*ASM* de la fenêtre de taille *TF*

Reconstruire l'inférence en appliquant *PhyML*

Calculer $RF_{normalisée}$ entre l'inférence et l'arbre géographique donné

Sauvegarder la valeur de distance dans la matrice de position

Trouver le minimum de $RF_{normalisée}$ et récupérer la position du gène correspondante à la fenêtre

Fin Pour

Fin Pour

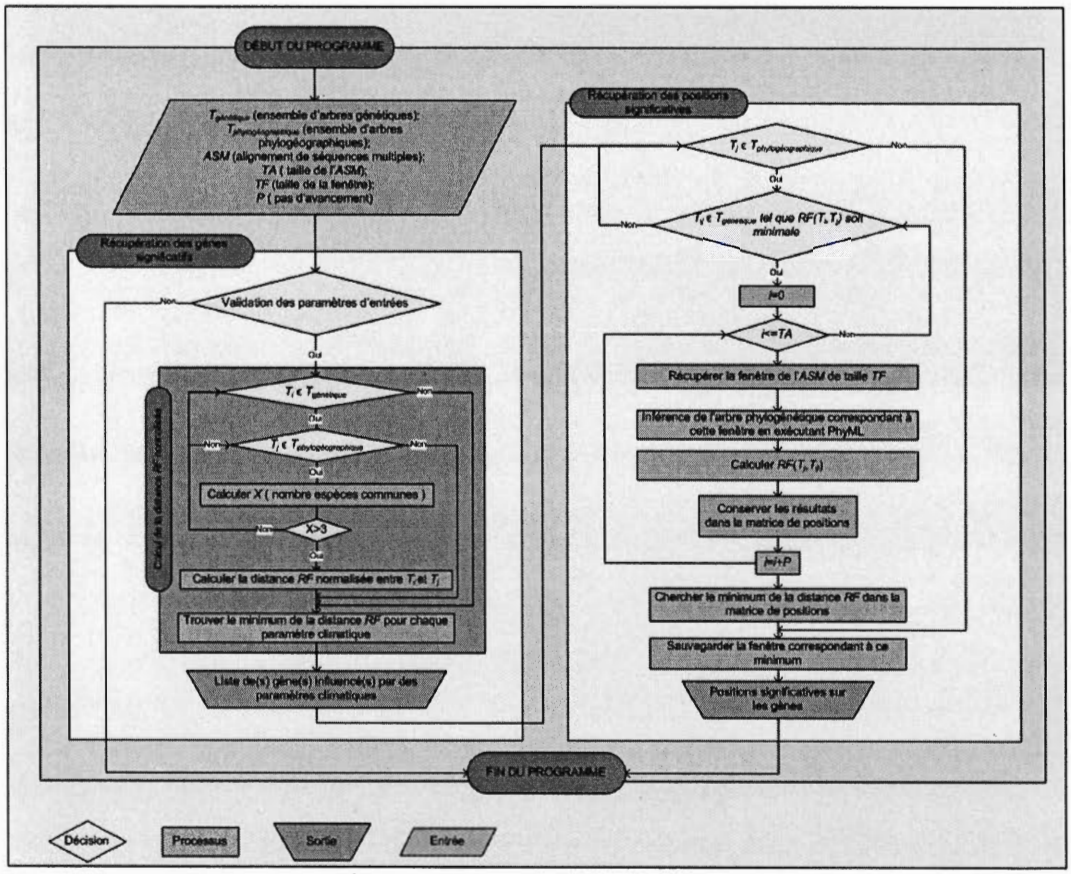


Figure 4.4 : Algorithme global développé en Java pour trouver la position génétique liée à des paramètres géographiques.

4. 5 Métriques entre arbres phylogénétiques

La comparaison entre deux arbres phylogénétiques définis pour le même groupe d'espèces est toujours une question d'actualité. Cette comparaison vise à trouver soit la similarité (arbre consensus) ou soit la dissimilarité entre les phylogénies (distance). Dans notre étude, nous avons utilisé plusieurs distances : Robinson-

Foulds (Robinson et Foulds, 1981), la distance de la bipartition (Boc *et al*, 2010) et la distance de quartet (Mailund et Pederson, 2004).

4. 5.1 La distance de Robinson et Foulds

C'est une des distances les plus utilisées au niveau de la comparaison des topologies de deux arbres additifs (Robinsons et Foulds, 1981; Makarenkov et Lapointe, 2004). En effet si T_1 et T_2 sont deux X-arbres : la distance *RF* consiste au nombre minimal d'opérations élémentaires permettant de transformer T_1 en T_2.

Les opérations peuvent être des fusions ou des disjonctions au niveau des nœuds. Robinson et Foulds ont montré qu'une telle métrique est égale au nombre de bipartitions (Felsenstein, 1981) présentes dans un arbre, mais absentes dans l'autre. Si n est le nombre d'espèces alors $2n$-6 est le nombre maximal que peut atteindre la distance *RF* entre les deux arbres binaires comparés. On définit la distance *RF* comme suit :

Définition 4.1 : La distance *RF* entre deux arbres phylogénétiques T_1 et T_2 est le nombre de bipartitions $\sum T_1$ non triviales de T_1 qui ne sont pas dans T_2 plus le nombre de bipartitions $\sum T_2$ non triviales de T_2 qui ne sont pas dans T_1. Cette distance est calculée généralement par la formule suivante :

$$RF(T_1, T_2) = \left|\sum T_1 - \sum T_2\right| + \left|\sum T_2 - \sum T_1\right| \qquad (4.3)$$

Définition 4.2 : La bipartition induite de l'arête (a, b) d'un arbre phylogénétique T est construite en retirant l'arête (a,b) générant ainsi deux sous-arbres phylogénétiques T_a et T_b.

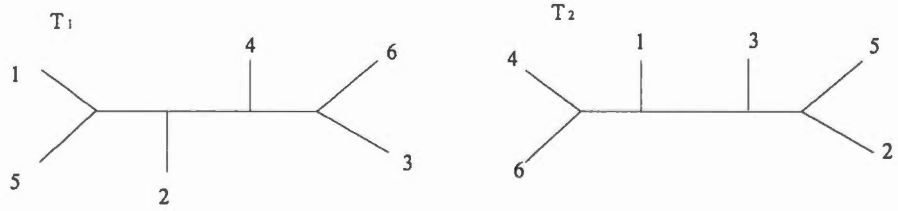
La distance *RF* est calculable en temps linéaire et le nombre de feuilles peut être approximé en temps sous-linéaire (Pattengale *et al*, 2007).

L'utilisateur de notre logiciel peut indiquer la valeur seuil de la distance *RF*. Ce critère permet à l'utilisateur de filtrer les arbres et de ne pas sélectionner les arbres ayant une distance *RF* supérieure à *RF*-seuil.

Nous normaliserons la distance *RF* (Équation 4.3) par $2n$-6 (la distance maximale entre deux arbres) où n est le nombre de feuilles. Notons que nous appliquerons cette distance uniquement dans le cas ou que les arbres soient définis sur le même ensemble d'espèces

Exemple de calcul de la distance *RF* :

Soient les deux arbres phylogénétiques T_1 et T_2



Selon la formule (4.3) on obtient :

$$RF(T_1,T_2)=\left|\sum T_1-\sum T_2\right|+\left|\sum T_2-\sum T_1\right|=3+3=6.$$

Notons également que cette distance est égale au nombre d'opérations élémentaires transformant T_1 en T_2.

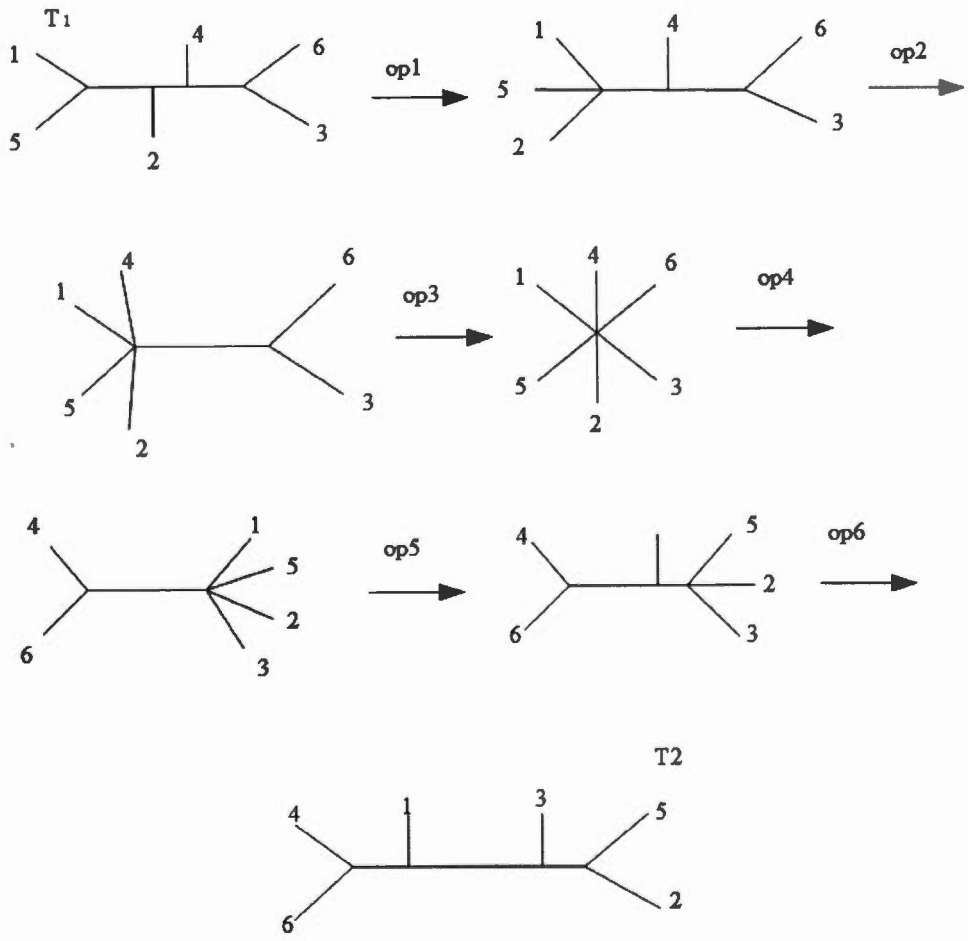


Figure 4.5 : Les opérations transformant l'arbre T_1 en l'arbre T_2. Il a fallu 6 opérations élémentaires permettant cette transformation.

4. 5.2 La dissimilarité de bipartition

Soient deux arbres binaires T et T', ayant le même ensemble de feuilles. Un vecteur de bipartitions est un vecteur binaire inféré pour T ou pour T'. Pour chaque arbre binaire, on construit une matrice dans laquelle chaque ligne représente un vecteur de bipartition liée à une branche. On aura ainsi les deux matrices BT et BT' (voir la Figure 4.8).

La dissimilarité de bipartition *bd* (Boc *et al*, 2010) est calculée par la formule suivante :

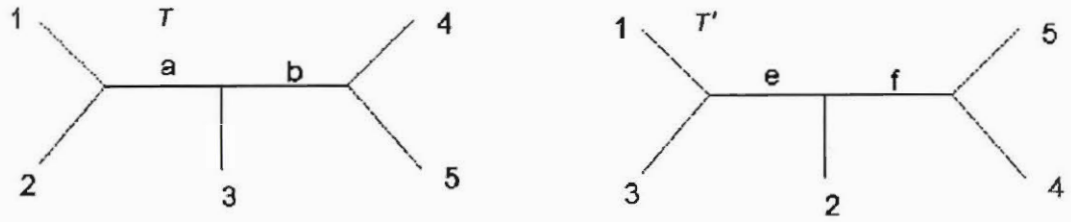
$$\frac{(\sum_{a\in BT} Min_{b\in BT'}(Min(d(a,b);d(a,\bar{b})) + \sum_{b\in BT'} Min_{a\in BT}(Mind(b,a);d(b,\bar{a})))}{2} \quad (4.4)$$

Où :

- $d(a, b)$ est la distance de Hamming entre deux vecteurs de bipartition a et b, a et b appartenant à BT et BT' respectivement.
- $\bar{a}$ et $\bar{b}$ sont les deux vecteurs complémentaires de a et b.

Exemple de calcul de la dissimilarité de bipartition :

On considère les deux arbres T et T' suivants :



Pour chacun de ces deux arbres, on obtient les deux tables de bipartition BT et BT'.

	1	2	3	4	5
a	0	0	1	1	1
b	1	1	1	0	0

	1	2	3	4	5
a	0	1	0	1	1
b	1	1	0	0	1

Figure 4.6 : Tableaux de bipartitions *BT* et *BT'*.

En appliquant la formule (4.1) on trouve que $bd = ((1+2) + (1+1))/2 = 5/2=2,5$.

4. 5.3 La distance de quartet :

La distance quartet est une des distances les plus utilisées pour comparer deux arbres phylogénétiques. Considérons un arbre phylogénétique *T*. Une topologie de quartet est le sous-arbre de *T* induit par quatre espèces.

La distance de quartet est calculée par le nombre d'ensembles de quatre espèces induisant des topologies différentes dans les deux arbres à comparer. Considérons les deux arbres binaires *T* et *T'* et l'ensemble S d'espèces communes à *T* et *T'*. La distance de quartet correspond au nombre de quartets $\{a, b, c, d\} \subseteq S$ ayant des topologies différentes dans *T* et *T'* (Buneman, 1971). Puisqu'il y a différents ensembles de quartets, la distance peut être calculée en comptant le nombre total

de quartets moins le nombre de topologies identiques. Pour avoir le nombre des topologies identiques, l'algorithme calcule le nombre de topologies orientées multiplié par deux.

$$QD=\binom{n}{4} - (\text{nb de topologies de quartets orientées } *2)$$

4. 6 Complexité de l'algorithme

La complexité de notre algorithme dépend de plusieurs facteurs à la fois : du nombre des gènes traités, du nombre de fenêtres dans chaque gène et de la complexité des programmes externes utilisés.

Les programmes externes utilisés sont les suivants :

- Paquet *PHYLIP* : O (*PHYLIP*) = O ($r.n3+ r.\ n.TA$) (Felenstein, 2002)
- *PhyML* : O (*PhyML*) = O ($e.\ n.TA$) (Guindon *et al*, 2010)
- Calcul de la distance $RF = O(n^2)$ (Makarenkov et Leclerc, 2000)

Où :

- n : nombre d'espèces
- r : nombre de reliquats
- TA : taille de l'alignement de séquences multiples
- e : nombre d'étapes de raffinage réalisé par l'algorithme *PhyML*

La complexité de notre algorithme dépend aussi de K qui représente le nombre des gènes traités.

4.7 Conclusion

Dans ce chapitre nous avons présenté l'algorithme permettant de trouver les relations entre les arbres phylogénétiques et phylogéographiques. Nous avons détaillé les différentes étapes, y compris les étapes de la préparation des données, d'inférence des arbres, du calcul des distances, du choix des distances minimums permettant de déterminer les gènes cibles. Nous avons utilisé trois types de distances d'arbres phylogénétiques, à savoir : la distance topologique de Robinson et Foulds, la dissimilarité de bipartition et la distance de quartet. Plusieurs notions à améliorer ou à ajouter seront discutées dans le chapitre Conclusions et Perspectives.

CHAPITRE V

PRÉSENTATION DES RÉSULTATS

5. 1 Introduction

Nous présenterons dans ce chapitre l'ensemble des résultats obtenus à la suite de l'application de l'algorithme du chapitre IV sur notre jeu de données. En premier lieu, nous parlerons de la liste des espèces et présenterons les matrices de dissimilarités, ainsi que les arbres phylogénétiques et phylogéographiques. En second lieu, nous exposerons les fragments des gènes qui se montrent les plus liés aux paramètres climatiques. Dans la partie discussion, nous présenterons les résultats en montrant des illustrations et des graphiques. Nous finaliserons ce chapitre par une analyse des résultats obtenus.

5. 2 Arbres phylogéographiques et phylogénétiques

La collecte et la préparation des données ont constitué un enjeu majeur dans notre recherche. Afin d'arriver à structurer une base de données majoritaire pour tester l'algorithme, on a du passer par plusieurs étapes : 1) Filtrer les données selon la zone géographique choisie; 2) Ajouter la liste d'espèces présentes dans cette zone;

3) Extraire les paramètres climatiques régnants dans cette zone; 4) Préparer des données afin de reconstruire les arbres phylogéographiques et phylogénétiques. Un programme en langage Perl a été développé afin de reconstruire l'ensemble d'arbres phylogéographiques. Dans la partie suivante, nous exposerons une suite des résultats qui s'avèrent les plus pertinents.

5. 2.1 Données phylogéographiques

5. 2.1.1 Liste des espèces

ID	Nom de l'espèce
117	Alopexlagopus
305	bassariscusastutus
443	Canislatrans
444	canis lupus
674	conepatusleuconotus
1443	gulogulo
1485	herpailurusyaguarondi
1756	leoparduspardalis
1832	lontracanadensis
1856	lynx canadensis
1859	lynx rufus
1961	martes americana
2049	mephitismephitis
2343	mustelaerminea
2346	mustelafrenata
2350	Mustelanivalis
2355	Mustelavison
2506	Nasuanarica
2834	Pantheraonca
3193	Procyonlotor
3357	Pumaconcolor
3839	Spilogaleputorius
3840	Spilogalepygmaea
4065	Taxideataxus
4191	Urocyonlittoralis
4206	Ursusamericanus

4207	Ursusarctos
4241	vulpes velox
4242	vulpes vulpes

Tableau 5.1 : Liste des 30 espèces après la filtration des données avec leur ID.

5.2.1.2 Matrices de dissimilarité

a_lagopus	0	1	0.97	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
b_astutus	1	0	1	0.88	1	0.88	0.88	1	1	0.94	1	1	1	1	0.88	1	1	0.88	0.89	0.88	0.88	1	0.91	0.97	1	1	1	1	1	1
c_latrans	0.97	1	0	0.71	1	0.61	0.59	1	1	0.94	1	1	1	1	0.59	1	1	0.59	0.68	0.59	0.59	1	0.91	0.97	1	1	1	1	1	1
c_lupus	1	0.87	0.71	0	1	1	1	1	1	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
c_leuconotus	1	1	1	1	0	0.73	0.71	1	1	0.94	1	1	1	1	0.71	1	1	0.71	0.75	0.71	0.71	1	0.91	0.97	1	1	1	1	1	1
g_gulo	1	0.88	0.61	1	0.73	0	1	1	1	1.00	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
h_yaguarond	1	0.88	0.59	1	0.71	1	0	1	1	0.94	1	1	1	1	0.27	1	1	0.53	0.43	0.54	0.35	1	0.91	0.97	1	1	1	1	1	1
l_pardalis	1	1	1	1	1	1	1	0	1	0.94	1	1	1	1	0.25	1	1	0.51	0.41	0.52	0.33	1	0.91	0.97	1	1	1	1	1	1
lo_canaden	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ly_canaden	1	0.94	0.94	1	0.94	1	0.94	0.94	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
l_rufus	1	1	1	1	1	1	1	1	1	1	0	1	1	1	0.94	1	1	0.94	0.96	0.94	0.94	1	0.97	0.97	1	1	1	1	1	1
n_american	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
m_pennanti	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
m_mephitis	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
m_erminea	1	0.88	0.59	1	0.71	1	0.27	0.25	1	1	0.94	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
m_frenata	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	0.51	0.44	0.52	0.37	1	0.91	0.97	1	1	1	1	1	1
m_nivalis	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	1
m_vison	1	0.88	0.59	1	0.71	1	0.53	0.51	1	1	0.94	1	1	1	1	0.51	1	0	1	1	1	1	1	1	1	1	1	1	1	1
n_narica	1	0.89	0.68	1	0.75	1	0.43	0.41	1	1	0.96	1	1	1	1	0.44	1	1	0	0.52	0.51	1	0.91	0.97	1	1	1	1	1	1
p_onca	1	0.88	0.59	1	0.71	1	0.54	0.52	1	1	0.94	1	1	1	1	0.52	1	1	0.52	0	0.47	1	0.91	0.99	1	1	1	1	1	1
p_lotor	1	0.88	0.59	1	0.71	1	0.35	0.33	1	1	0.94	1	1	1	1	0.37	1	1	0.51	0.47	0	1	0.91	0.97	1	1	1	1	1	1
p_concolor	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1.00	1	1	1	1	1	0	0.91	0.97	1	1	1	1	1	1
s_putorius	1	0.91	0.91	1	0.91	1	0.91	0.91	1	1	0.97	1	1	1	1	0.91	1	1	0.91	0.91	0.91	0.91	0	1	1	1	1	1	1	1
s_pygmaea	1	0.97	0.97	1	0.97	1	0.97	0.97	1	1	0.97	1	1	1	1	0.97	1	1	0.97	0.99	0.97	0.97	1	0	1	1	1	1	1	1
t_taxus	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1
u_littoralis	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1
u_americanu	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	1
u_arctos	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1
v_velox	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1
v_vulpes	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0

Figure 5.1 : Matrice de dissimilarité entre les espèces habitantes de la zone géographique I de la précipitation moyenne.

Le calcul est fait selon la formule suivante : $1 - \frac{nb\,espece\,ensemble}{nbzone}$

a_lagopus	0	0.69	0.67	0.98	0.75	0.93	0.94	0.68	0.71	0.81	0.75	0.84	0.7	0.67	0.76	0.68	0.68	0.94	0.95	0.7	0.9	0.82	1	0.83	1	0.68	0.89	1	0.66	1
b_astutus	0.69	0	0.89	0.92	0.92	0.94	0.94	0.82	0.89	0.81	0.89	0.93	0.81	0.87	0.81	0.88	0.83	0.93	0.97	0.81	0.87	0.81	1	0.84	1	0.86	0.98	0.99	0.82	1
c_latrans	0.67	0.89	0	0.91	0.7	0.88	0.88	0.17	0.53	0.4	0.59	0.63	0.18	0.4	0.3	0.44	0.17	0.87	0.92	0.14	0.77	0.51	1	0.56	1	0.27	0.89	0.99	0.18	1
c_lupus	0.98	0.92	0.91	0	0.71	0.93	0.94	0.49	0.51	0.7	0.6	0.64	0.5	0.47	0.65	0.48	0.48	0.94	0.95	0.5	0.89	0.74	1	0.69	1	0.49	0.89	1	0.47	1
c_leuconotus	0.75	0.92	0.7	0.71	0	0.96	0.96	0.93	0.97	0.91	0.97	0.98	0.91	0.96	0.91	0.97	0.94	0.94	1	0.91	0.92	0.91	1	0.91	1	0.96	1	1	0.93	1
g_gulo	0.93	0.94	0.88	0.93	0.96	0	0.95	0.71	0.75	0.81	0.77	0.84	0.73	0.71	0.77	0.72	0.71	0.94	0.96	0.71	0.9	0.84	1	0.85	1	0.72	0.89	1	0.7	1
h_yaguarondi	0.94	0.94	0.88	0.94	0.96	0.95	0	0.91	0.94	0.91	0.95	0.98	0.91	0.92	0.82	0.93	0.91	0.83	0.86	0.82	0.81	0.93	1	0.91	1	0.93	0.98	0.99	0.9	1
l_pardalis	0.68	0.82	0.17	0.49	0.93	0.71	0.91	0	0.95	0.91	0.96	0.99	0.91	0.93	0.82	0.94	0.91	0.83	0.86	0.82	0.81	0.93	1	0.91	1	0.94	0.98	0.99	0.9	1
lo_canaden	0.71	0.89	0.53	0.51	0.97	0.75	0.94	0.95	0	0.43	0.59	0.64	0.19	0.4	0.34	0.43	0.16	0.9	0.93	0.18	0.8	0.53	1	0.59	1	0.26	0.89	0.99	0.18	1
ly_canaden	0.81	0.81	0.4	0.7	0.91	0.81	0.91	0.91	0.43	0	0.6	0.64	0.53	0.51	0.69	0.51	0.51	0.95	0.96	0.54	0.92	0.77	1	0.69	1	0.52	0.92	1	0.51	1
l_rufus	0.75	0.89	0.59	0.6	0.97	0.77	0.95	0.96	0.59	0.6	0	0.79	0.41	0.64	0.4	0.66	0.43	0.9	0.95	0.4	0.8	0.54	1	0.7	1	0.49	0.92	0.99	0.43	1
m_americana	0.84	0.93	0.63	0.64	0.98	0.84	0.98	0.99	0.64	0.64	0.79	0	0.59	0.59	0.72	0.6	0.59	0.96	0.97	0.59	0.93	0.77	1	0.74	1	0.59	0.93	1	0.59	1
m_pennanti	0.7	0.81	0.18	0.5	0.91	0.73	0.91	0.91	0.19	0.53	0.41	0.59	0	0.63	0.78	0.64	0.63	0.98	0.99	0.63	0.96	0.83	1	0.73	1	0.63	0.96	1	0.63	1
m_mephitis	0.67	0.87	0.4	0.47	0.96	0.71	0.92	0.93	0.4	0.51	0.64	0.59	0.63	0	0.33	0.44	0.19	0.9	0.95	0.18	0.8	0.51	1	0.56	1	0.27	0.9	0.99	0.2	1
m_erminea	0.76	0.81	0.3	0.65	0.91	0.77	0.82	0.82	0.34	0.69	0.4	0.72	0.78	0.33	0	0.43	0.39	0.93	0.94	0.41	0.87	0.7	1	0.63	1	0.44	0.89	1	0.38	1
m_frenata	0.68	0.88	0.44	0.48	0.97	0.72	0.93	0.94	0.43	0.51	0.66	0.6	0.64	0.44	0.43	0	0.34	0.81	0.87	0.24	0.71	0.51	1	0.66	1	0.42	0.91	0.99	0.34	1
m_nivalis	0.68	0.83	0.17	0.48	0.94	0.71	0.91	0.91	0.16	0.51	0.43	0.59	0.63	0.19	0.39	0.34	0	0.94	0.95	0.45	0.89	0.71	1	0.63	1	0.45	0.89	1	0.42	1
m_vison	0.94	0.93	0.87	0.94	0.94	0.94	0.83	0.83	0.9	0.95	0.9	0.96	0.98	0.9	0.93	0.81	0.94	0	0.93	0.18	0.81	0.54	1	0.59	1	0.25	0.89	0.99	0.18	1
n_narica	0.95	0.97	0.92	0.95	1	0.96	0.86	0.86	0.93	0.96	0.95	0.97	0.99	0.95	0.94	0.87	0.95	0.93	0	0.81	0.81	0.92	1	0.9	1	0.94	0.98	0.99	0.89	1
p_onca	0.7	0.81	0.14	0.5	0.91	0.71	0.82	0.82	0.18	0.54	0.4	0.59	0.63	0.18	0.41	0.24	0.45	0.18	0.81	0	0.86	0.97	1	0.95	1	0.95	0.98	0.99	0.92	1
p_lotor	0.9	0.87	0.77	0.89	0.92	0.9	0.81	0.81	0.8	0.92	0.8	0.93	0.96	0.8	0.87	0.71	0.89	0.81	0.81	0.86	0	0.51	1	0.56	1	0.27	0.9	0.99	0.19	1
p_concolor	0.82	0.81	0.51	0.74	0.91	0.84	0.93	0.93	0.53	0.77	0.54	0.77	0.83	0.51	0.7	0.51	0.71	0.54	0.92	0.97	0.51	0	1	0.85	1	0.85	0.95	0.99	0.8	1
s_putorius	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0.75	1	0.57	0.93	0.99	0.55	1
s_pygmaea	0.83	0.84	0.56	0.69	0.91	0.85	0.91	0.91	0.59	0.69	0.7	0.74	0.73	0.56	0.63	0.66	0.63	0.59	0.9	0.95	0.56	0.85	0.75	0	1	1	1	1	1	1
t_taxus	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0.64	0.95	0.99	0.58	1
u_littoralis	0.68	0.86	0.27	0.49	0.96	0.72	0.93	0.94	0.26	0.52	0.49	0.59	0.63	0.27	0.44	0.42	0.45	0.25	0.94	0.95	0.27	0.85	0.57	1	0.64	0	1	1	1	1
u_americanus	0.89	0.98	0.89	0.89	1	0.89	0.98	0.98	0.89	0.92	0.92	0.93	0.96	0.9	0.89	0.91	0.89	0.89	0.98	0.98	0.9	0.95	0.93	1	0.95	1	0	1	0.26	1
u_arctos	1	0.99	0.99	1	1	1	0.99	0.99	0.99	1	0.99	1	1	0.99	1	0.99	1	0.99	0.99	0.99	0.99	0.99	0.99	1	0.99	1	1	0	0.89	1
v_velox	0.66	0.82	0.18	0.47	0.93	0.7	0.9	0.9	0.18	0.51	0.43	0.59	0.63	0.2	0.38	0.34	0.42	0.18	0.89	0.92	0.19	0.8	0.55	1	0.58	1	0.26	0.89	0	1
v_vulpes	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0

Figure 5.2 : Matrice de dissimilarité entre les espèces de la zone géographique II en fonction de la température moyenne. Le calcul est fait selon la formule suivante : $1 - \frac{nb\ especes\ enemble}{nbzone}$

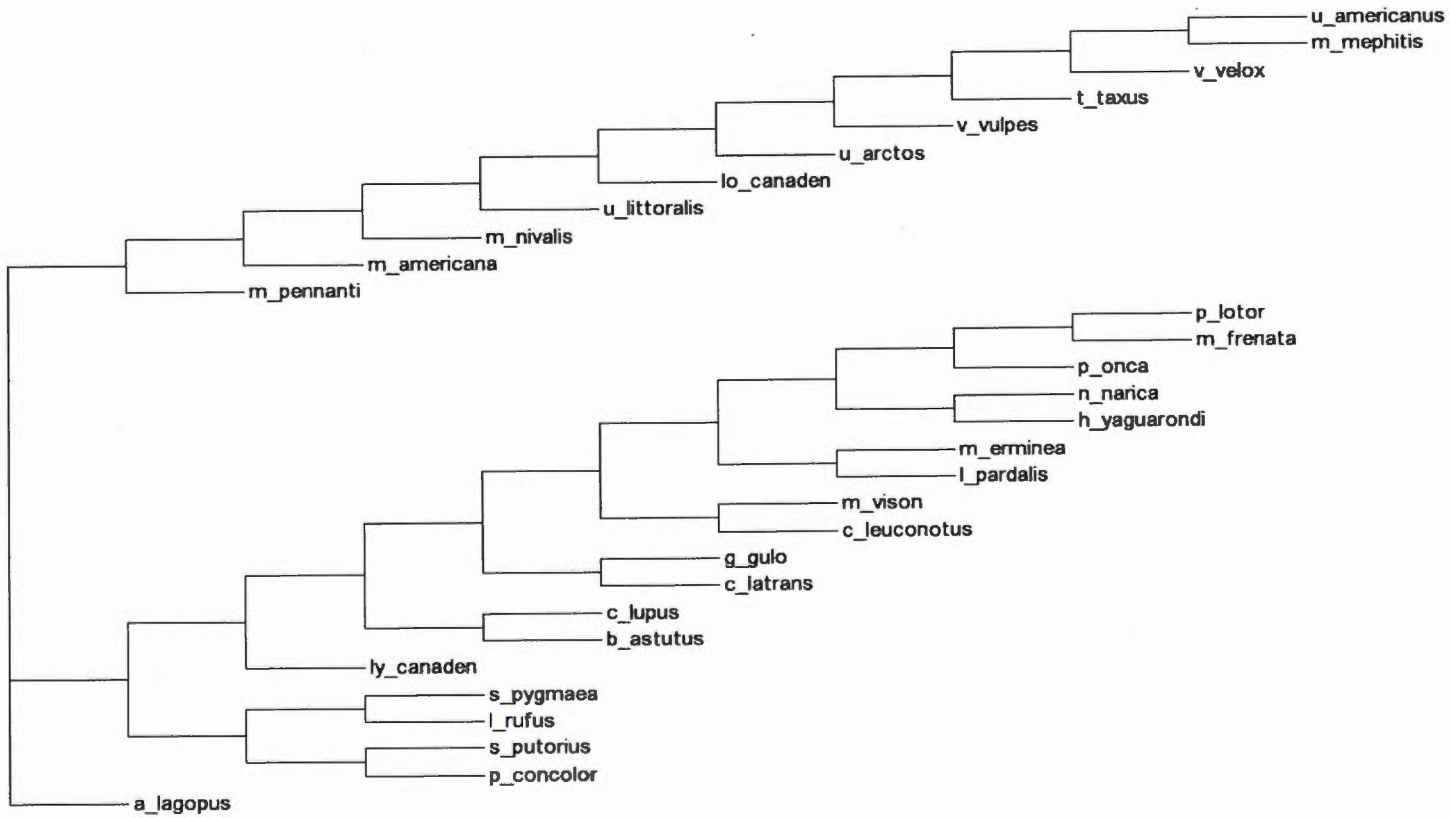


Figure 5.3 : Arbre phylogéographique de la distribution des espèces dans la zone I en fonction de la précipitation moyenne visualisé à l'aide de l'interface : http://trex.uqam.ca/.

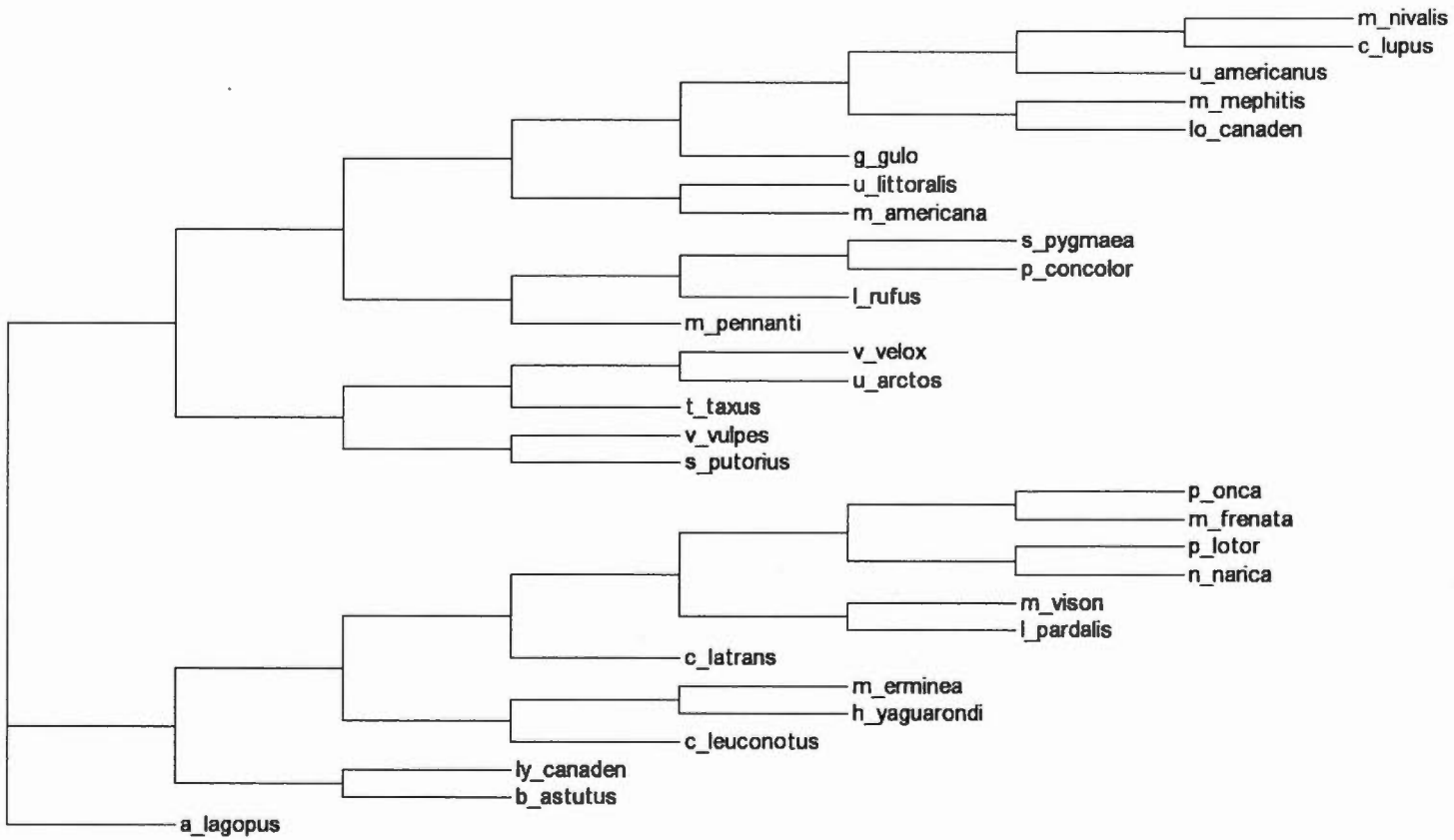


Figure 5.4 : Arbre phylogéographique de la distribution des espèces dans la zone II en fonction de la précipitation moyenne visualisé à l'aide de l'interface : **http://trex.uqam.ca/**

5. 2.1.3 Données phylogénétiques

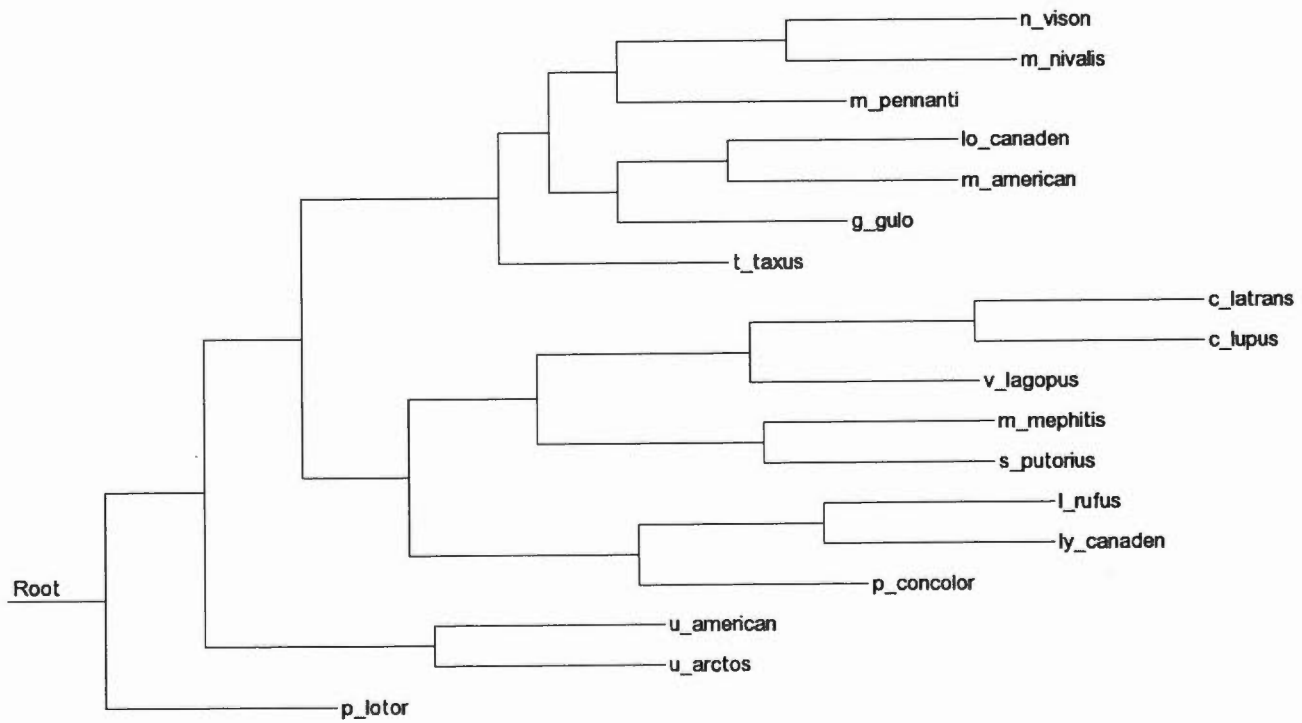


Figure 5.5 : Arbre phylogénétique des espèces. Le gène séquencé est le NADH_dehydrogenase_subunit_1 visualisé à l'aide de l'interface : http://trex.uqam.ca/ .

5.3 Déterminer les gènes significatifs

La première étape de l'algorithme consiste à déterminer les gènes qui sont influencés par les paramètres climatiques engendrant ainsi la distribution actuelle des espèces dans les différentes zones géographiques :

	prec 0,24 0,83	prec 0,83 0,9	prec 0,9 0,92	prec 0,92 0,93	prec 0,93 0,94	prec 0,94 0,95	prec 0,95 0,96	prec 0,96 0,97	prec 0,97 0,98	prec 0,98 0,99	prec 0,99 1,0	prec 1,0 1,1	prec 1,1 1,2	prec 1,2 1,3	prec 1,3 1,4	temp 1064 1345	temp 121 155	temp 12 120	temp 1345 1614	temp 155 201	temp 1614 1871	temp 1871 2278	temp 201 249	temp 2287 6065	temp 249 304	temp 304 372	temp 372 467	temp 467 588	temp 588 764	temp 764 1064	Minim
adenosine A3 receptor	0.86	0.96	1	1	0.93	1	0.82	1	0.96	1	0.86	0.93	0.89	0.89	0.89	0.89	1	1	1	1	0.96	0.89	0.96	1	0.96	1	0.93	1	0.96	0.96	0.82
NADH dehydrogenase subunit 1	0.77	0.91	0.91	1	1	1	0.91	1	0.95	1	0.91	0.91	0.91	0.91	0.91	1	0.91	1	1	1	0.95	0.91	0.95	1	0.91	0.82	1	1	1	1	0.77
Retinoid Binding Protein	0.81	0.94	0.94	1	1	1	0.94	1	0.97	1	0.81	0.94	0.91	0.94	0.94	0.94	1	0.91	0.94	1	0.94	0.91	0.97	1	0.91	0.94	1	1	0.91	0.97	0.81
adenosine A3 receptor1	0.86	0.89	0.93	1	1	1	0.89	1	0.96	1	1	1	0.89	0.89	0.96	0.89	1	0.93	0.93	1	0.96	0.89	0.96	1	0.96	1	1	1	0.89	0.96	0.86
NADH Dehydrogenase Subunit 2	0.83	0.93	0.87	1	1	1	0.9	1	0.97	1	0.8	0.87	0.77	0.77	0.77	0.9	1	1	0.93	1	0.97	0.9	0.97	1	0.9	0.87	0.93	1	0.9	0.97	0.77
rhodopsin	0.67	0.75	0.83	0.83	0.83	0.83	0.75	0.83	0.79	0.83	0.83	0.83	0.79	0.83	0.83	0.83	0.83	0.83	0.83	0.83	0.79	0.75	0.79	0.83	0.79	0.83	0.83	0.83	0.79	0.79	0.67
apolipoprotein B	0.87	0.97	0.93	1	1	1	0.9	1	0.97	1	0.8	0.93	0.83	0.83	0.83	0.9	1	1	0.93	1	0.97	0.9	0.97	1	0.97	1	0.93	1	0.9	0.97	0.8
NADH Dehydrogenase Subunit 4	0.79	0.92	0.92	1	1	1	0.92	1	0.96	1	0.92	0.92	0.92	0.92	0.92	1	0.92	1	1	1	0.96	0.88	0.96	1	0.88	0.83	1	1	1	0.96	0.79
Sex Determining Region Y Protein	0.92	0.92	0.83	1	0.83	0.83	0.83	0.83	0.92	1	0.67	0.83	0.67	0.67	0.67	0.83	0.67	1	0.67	0.67	1	0.83	1	1	0.75	0.67	0.83	1	0.67	0.75	0.67
ATP synthase F0 subunit 6	0.77	0.91	0.91	1	1	1	0.91	1	0.95	1	0.91	0.91	0.91	0.91	0.91	1	0.91	1	1	1	0.95	0.91	0.95	1	0.91	0.82	1	1	1	1	0.77
NADH Dehydrogenase Subunit 4L	0.77	0.82	0.91	1	1	1	0.91	1	0.95	1	1	1	1	1	1	1	0.82	0.91	1	1	0.95	0.91	0.95	0.91	1	0.91	1	1	1	1	0.77
ATP synthase F0 subunit 8	0.79	0.92	0.92	0.92	0.92	1	0.83	1	0.96	0.92	0.92	0.92	1	1	1	1	0.92	1	1	1	0.88	0.79	0.96	0.92	0.88	0.83	0.92	0.92	1	0.88	0.79
NADH Dehydrogenase Subunit 5	0.77	0.91	0.91	1	1	1	0.91	1	0.95	1	0.91	0.91	0.91	0.91	0.91	1	0.91	1	1	1	0.95	0.91	0.95	1	0.91	0.82	1	1	1	1	0.77
ATP synthase F0 subunit 8 1	0.79	0.92	0.92	0.92	0.92	1	0.83	1	0.96	0.92	0.92	0.92	1	1	1	1	0.92	1	1	1	0.88	0.79	0.96	0.92	0.88	0.83	0.92	0.92	1	0.88	0.79
NADH Dehydrogenase Subunit 6	0.9	0.8	0.6	0.8	1	1	0.8	1	1	1	1	1	1	1	1	1	0.8	1	1	1	1	1	1	1	0.8	1	1	1	1	1	0.6
brain derived neurotrophic factor	0.84	0.91	1	1	1	0.94	0.91	1	0.97	1	1	1	0.97	0.97	0.97	0.97	1	1	1	1	0.97	0.91	0.97	1	0.97	1	1	1	0.97	0.97	0.84
cancer susceptibility protein 1	0.82	0.95	0.91	1	1	1	0.95	1	0.95	1	0.82	1	0.91	0.91	0.91	0.91	1	1	0.91	1	1	0.91	0.95	1	0.95	1	0.91	1	0.91	0.86	0.82
polypeptide 1 precursor	0.69	0.77	0.77	0.85	0.85	0.85	0.77	0.85	0.81	0.85	0.85	0.85	0.81	0.85	0.85	0.85	0.77	0.77	0.85	0.77	0.81	0.77	0.81	0.85	0.81	0.85	0.85	0.85	0.81	0.81	0.69
Von Willebrand Factor	0.78	0.94	0.78	1	1	1	0.94	1	1	1	0.78	1	0.89	0.89	0.89	0.89	1	0.89	0.89	1	1	0.94	0.94	1	0.94	1	0.89	1	0.89	0.89	0.78
Cytochrome Oxidase Subunit I	0.82	0.82	0.91	1	1	1	0.86	1	0.95	1	1	1	1	0.95	0.95	0.95	1	1	1	1	0.95	0.86	1	0.91	1	0.91	1	1	0.95	1	0.82
prepronociceptin	0.85	0.96	0.92	1	1	1	0.88	1	0.96	1	0.92	0.92	0.88	0.88	0.88	0.96	1	0.92	1	1	0.96	0.88	0.96	1	0.96	1	0.92	1	0.96	0.96	0.85
growth hormone receptor	0.81	0.94	1	1	1	1	0.91	1	0.97	1	0.94	0.94	0.91	0.91	0.91	0.97	0.94	0.91	1	1	0.94	0.91	0.97	1	0.91	0.94	1	1	0.97	0.97	0.81
recombination activating protein 1	0.86	0.95	1	1	1	0.91	0.86	1	0.95	1	0.82	0.91	0.86	0.82	0.91	1	0.91	0.91	1	0.91	1	0.91	0.95	1	0.95	0.91	0.82	1	0.95	0.95	0.82
Minimum RF normalised	0.67	0.75	0.6	0.8	0.83	0.83	0.75	0.83	0.79	0.83	0.67	0.83	0.67	0.67	0.67	0.83	0.67	0.77	0.67	0.67	0.79	0.75	0.79	0.83	0.75	0.67	0.82	0.83	0.67	0.75	

Figure 5.6: Matrice de la distance *RF* normalisée entre les deux ensembles de données. Les deux arbres ayant une distance minimale sont mis en évidence. Les protéines pertinentes sont : rhodopsine, SRY et NADH.

5.4 Liste de protéines liées aux paramètres climatiques

1) prec_0.24_0.83 => [rhodopsin]

2) prec_0.83_0.9 => [rhodopsin]

3) prec_0.9_0.92 => [NADH_Dehydrogenase_Subunit_6]

4) prec_0.92_0.93 => [NADH_Dehydrogenase_Subunit_6]

5) prec_0.93_0.94 => [rhodopsin, Sex_Determining_Region_Y_Protein]

6) prec_0.94_0.95 => [rhodopsin, Sex_Determining_Region_Y_Protein]

7) prec_0.95_0.96 => [rhodopsin]

8) prec_0.96_0.97 => [rhodopsin, Sex_Determining_Region_Y_Protein]

9) prec_0.97_0.98 => [rhodopsin]

10) prec_0.98_0.99 => [rhodopsin]

11) prec_0.99_1.0 => [Sex_Determining_Region_Y_Protein]

12) prec_1.0_1.1 => [rhodopsin, Sex_Determining_Region_Y_Protein]

13) prec_1.1_1.2 => [Sex_Determining_Region_Y_Protein]

14) prec_1.2_1.3 => [Sex_Determining_Region_Y_Protein]

15) prec_1.3_1.4 => [Sex_Determining_Region_Y_Protein]

16) temp_1064_1345 => [rhodopsin, Sex_Determining_Region_Y_Protein]

17) temp_121_155 => [Sex_Determining_Region_Y_Protein]

18) temp_12_120 => [polypeptide_1_precursor]

19) temp_1345_1614 => [Sex_Determining_Region_Y_Protein]

20) temp_155_201 => [Sex_Determining_Region_Y_Protein]

21) temp_1614_1871 => [rhodopsin]

22) temp_1871_2278 => [rhodopsin]

23) temp_201_249 => [rhodopsin]

24) temp_2287_6065 => [rhodopsin]

25) temp_249_304 => [Sex_Determining_Region_Y_Protein]

26) temp_304_372 => [Sex_Determining_Region_Y_Protein]

27) temp_372_467 => [recombination_activating_protein_1]

28) temp_467_588 => [rhodopsin]

29) temp_588_764 => [Sex_Determining_Region_Y_Protein]

30) temp_764_1064 => [Sex_Determining_Region_Y_Protein]

5.5 Détection des gènes candidats

La dernière phase de l'algorithme consiste à trouver les positions de gènes liées à la distribution géographique des espèces selon un paramètre climatique. Après l'étape précédente qui nous a retourné la liste des gènes candidats (voir section 5.4), nous avons sélectionné les gènes ayant été retournés majoritairement. Nous procédons donc au traitement de la rhodopsine, la SRY et de la NADH. Ce traitement consiste à parcourir l'*ASM* en se basant sur le principe de la fenêtre coulissante, c.-à-d. parcourir l'*ASM* de chaque protéine avec une fenêtre coulissante de taille *TF* avec un avancement *P*. Pour chaque fenêtre nous calculons la distance *RF* entre l'arbre phylogénétique propre à cette fenêtre et l'arbre

phylogéographique correspondant. Le minimum de la distance *RF* indique une position significative.

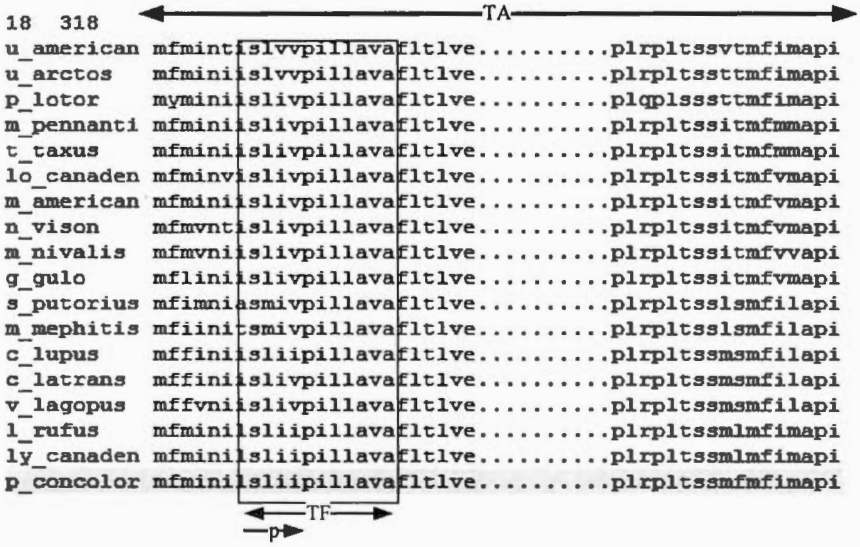


Figure 5.7 : Exemple de l'utilisation d'une fenêtre coulissante sur un *ASM* de 18 espèces étudiées.

5. 6 La distance *RF* minimale sur chaque protéine traitée.

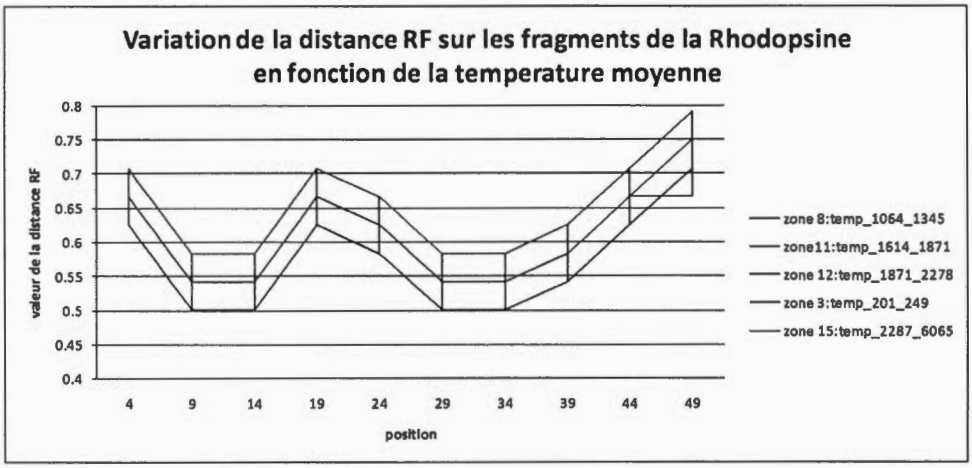


Figure 5.8 : Variations de la distance *RF* calculée à chaque fenêtre coulissante de la rhodopsine. La taille de la fenêtre est 7 et le pas d'avancement de la fenêtre est 5. L'axe horizontal présente la valeur de la position moyenne de chaque fenêtre.

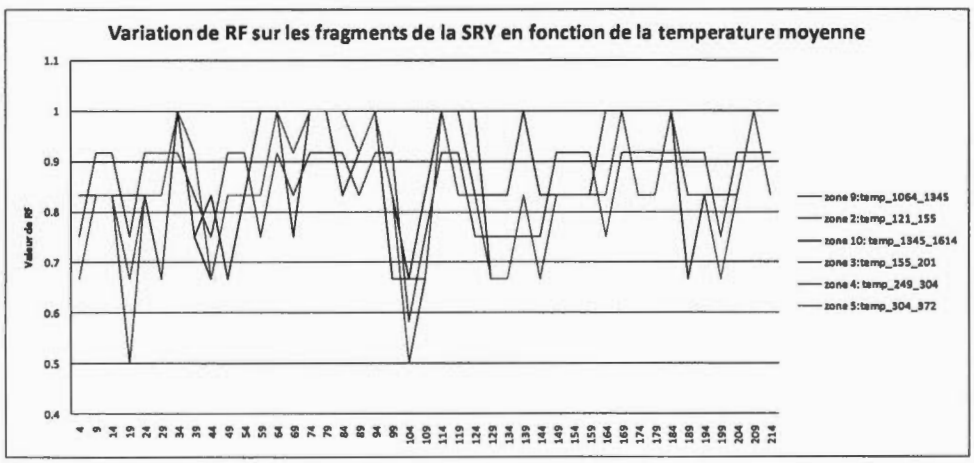


Figure 5.9 : Variations de la distance *RF* entre calculée dans chaque fenêtre coulissante sur la SRY. La taille de la fenêtre est 7 et le pas est 5. L'axe horizontal présente la valeur de la position moyenne de chaque fenêtre.

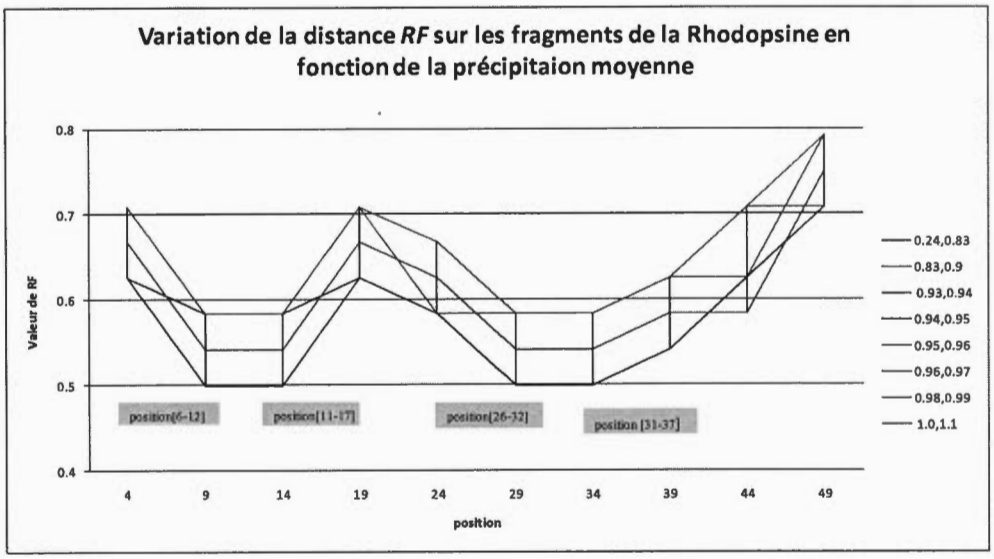


Figure 5.10 : Variations de la distance *RF* calculée dans chaque fenêtre coulissante de la Rhodopsine. La taille de la fenêtre est 7 et le pas d'avancement est 5. L'axe horizontal présente la valeur de la position moyenne de chaque fenêtre.

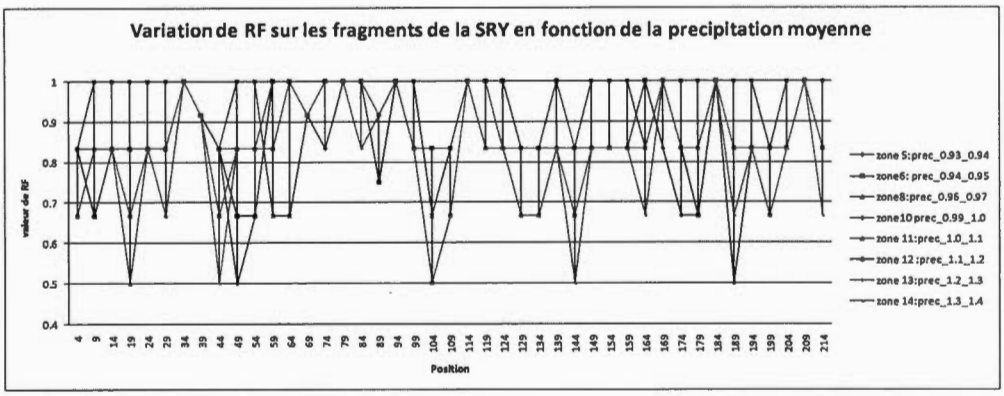


Figure 5.11 : Variations de la distance *RF* calculée dans chaque fenêtre coulissante sur la SRY. La taille de la fenêtre est 7 et le pas d'avancement de la fenêtre est 5. L'axe horizontal présente la valeur de la position moyenne de chaque fenêtre.

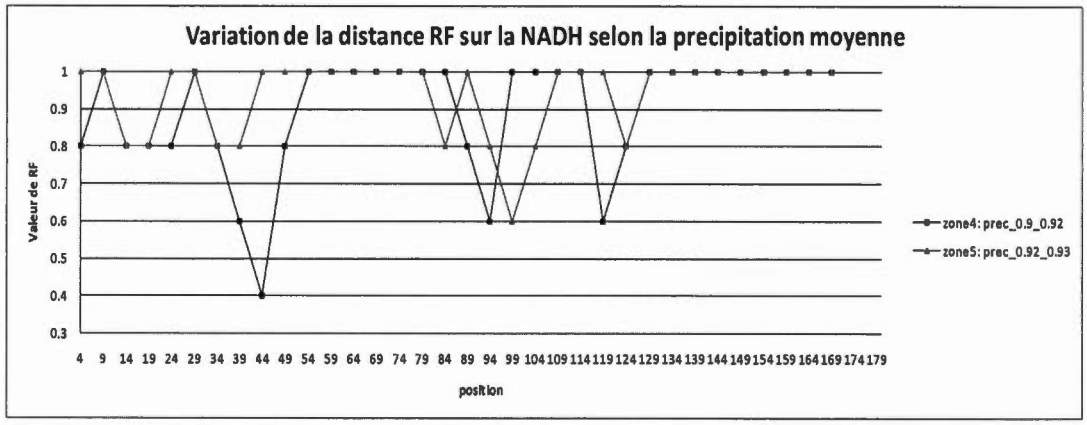


Figure 5.12 : Variations de la distance *RF* entre calculée dans chaque fenêtre coulissante sur la protéine NADH. La taille de la fenêtre est 7 et le pas d'avancement de la fenêtre est 5. L'axe horizontal présente la valeur de la position moyenne de chaque fenêtre.

5.7 Fragments de gènes liés aux paramètres climatiques

	[6-12]	[11-17]	[26-32]	[21-27]	[31-37]
prec_0.24_0.83	√	√	√	√	0
prec_0.93_0.94	√	√	√	√	√
prec_0.83_0.9	√	√	√	√	0
prec_0.96_0.97	√	√	√	√	0
prec_0.97_0.98	√	√	√	√	0
prec_0.98_0.99	√	√	√	√	0
prec_1.0_1.1	√	√	√	√	√
temp_1064_1345	√	√	√	√	0
temp_1614_1871	√	√	√	√	0
temp_1871_2278	√	√	√	√	0
temp_201_249	√	√	√	√	0
temp_2287_6065	√	√	√	√	0

Figure 5.13 : Matrice indiquant les positions significatives sur la rhodopsine dont la distance *RF* est minimale. La valeur « √ » indique une distance *RF* minimale.

	[1-7]	[16-22]	[41-47]	[101-107]	[106-112]
prec_0.9_0.92	0	0	0	0	0
prec_0.92_0.93	0	0	0	0	0
prec_0.93_0.94	0	√	0	0	0
prec_0.94_0.95	√	√	√	0	√
prec_0.96_0.97	0	0	0	√	0
prec_0.99_1.0	0	0	0	√	0
prec_1.0_1.1	0	0	0	√	0
prec_1.1_1.2	0	0	√	√	0
prec_1.2_1.3	0	0	0	√	√
prec_1.3_1.4	0	0	0	0	0
temp_1064_1345	0	0	√	√	√
temp_121_155	0	√	√	0	√
temp_1345_1614	0	0	√	0	√
temp_155_201	√	√	√	0	√
temp_249_304	√	√	√	0	√
temp_304_372	√	√	√	0	√

Figure 5.14 : Matrice indiquant les positions significatives sur la SRY dont la distance *RF* est minimale. La valeur « √ » indique une distance *RF* minimale.

5. 8 Illustration des protéines en 3D

- *NADH subunit6:*

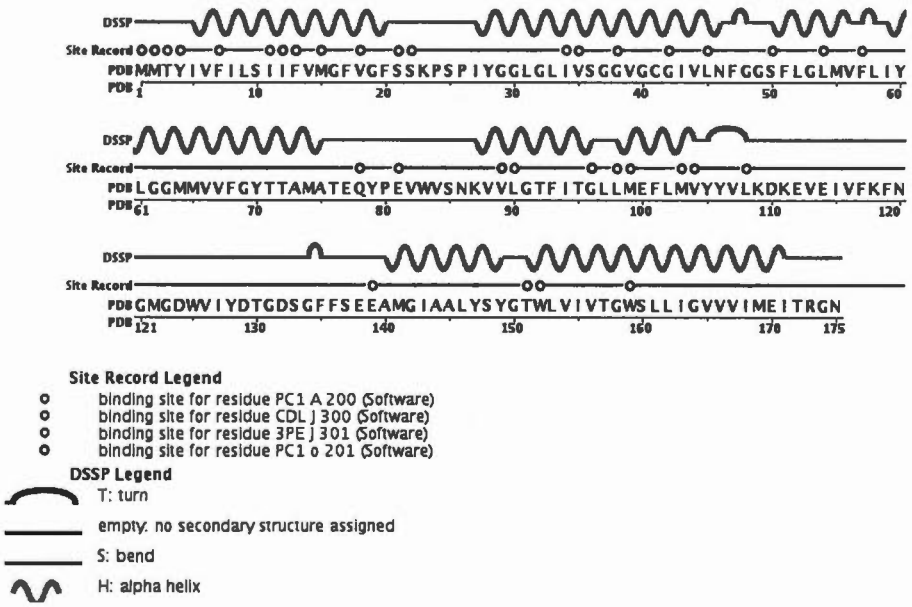


Figure 5.15 : Représentation de la structure complète de la protéine Source : www.rcsb.org.

Figure 5.16 : Illustration 3D de la position [40-47] de la NADH prouvée comme étant la partie qui explique la distribution géographique des espèces Source : UCSF Chimera.

- *La Rhodopsine*

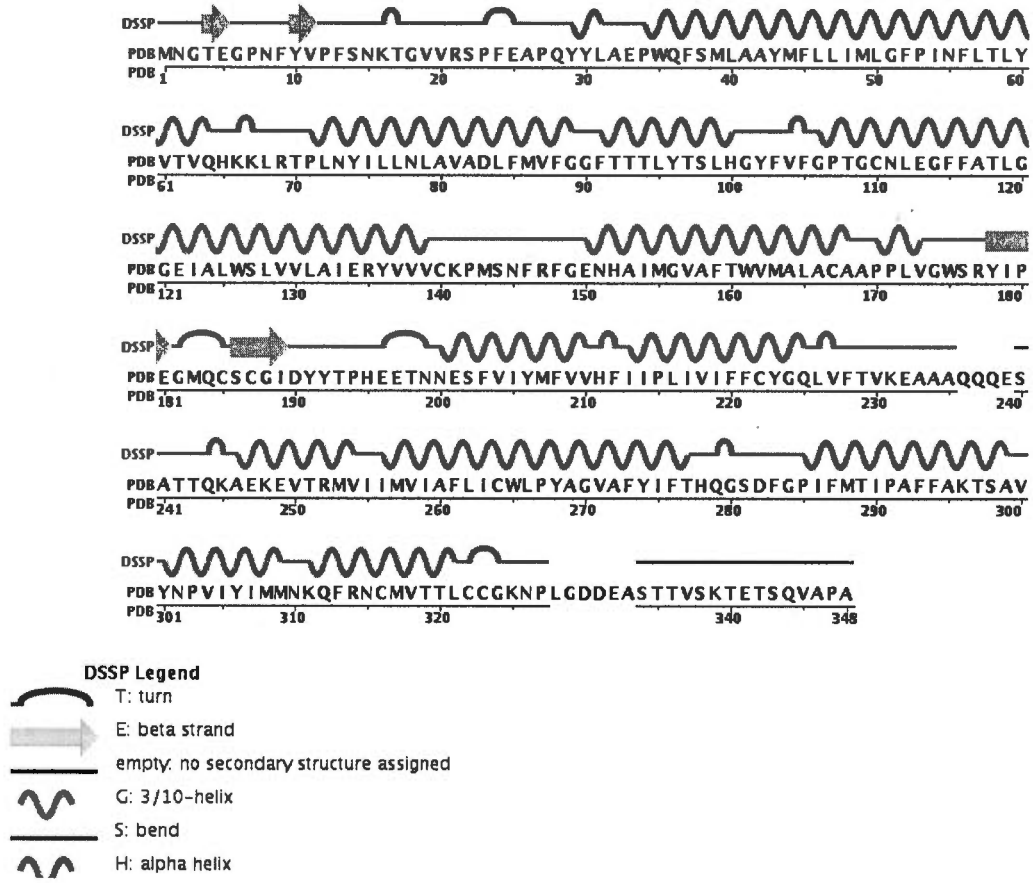


Figure 5.17 : La structure complète de la protéine rhodopsine. Source : www.rcsb.org.

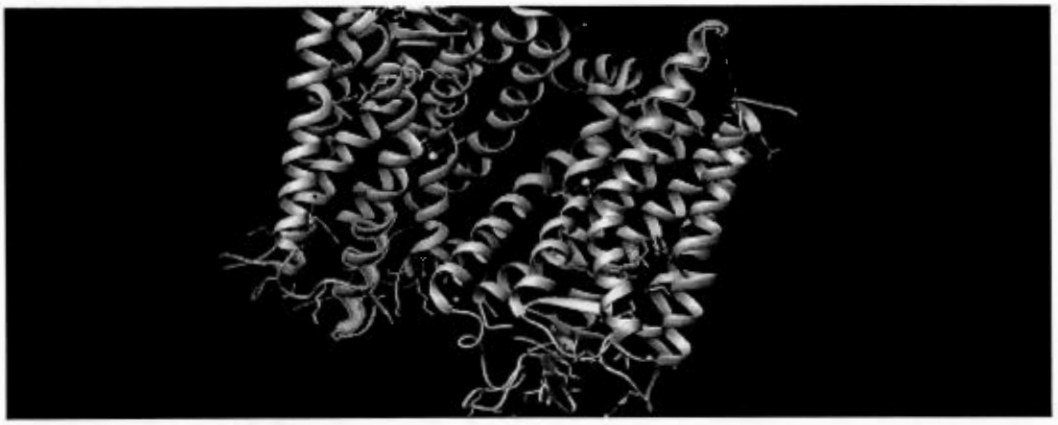


Figure 5.18 : Illustration en 3D des positions [21-27] et [26-32] sur la rhodopsine. Source : UCSF Chimera.

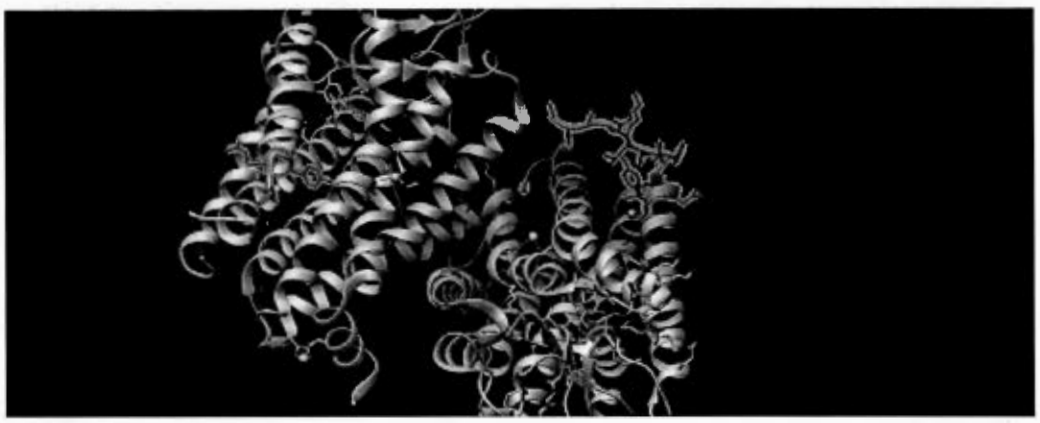


Figure 5.19 : Illustration en 3D des positions [6-12] et [11-17] sur la rhodopsine. Source : UCSF Chimera.

La SRY protéine liée au sexe

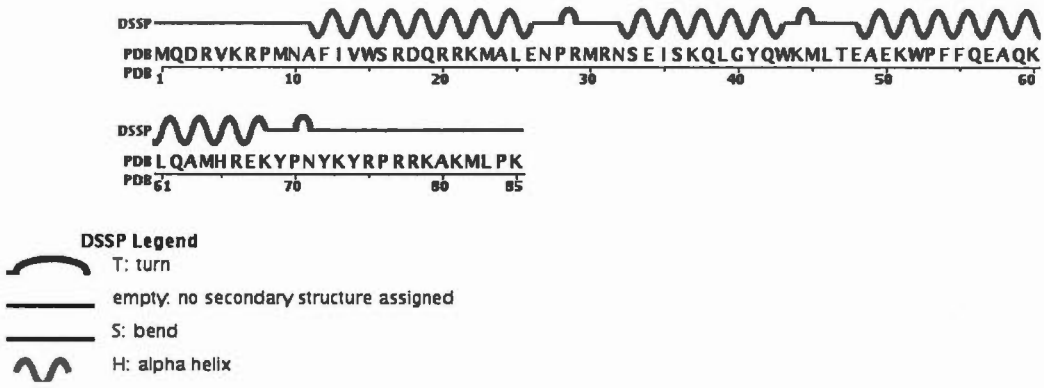


Figure 5.20 : La structure complète de la protéine SRY. Source : www.rcsb.org.

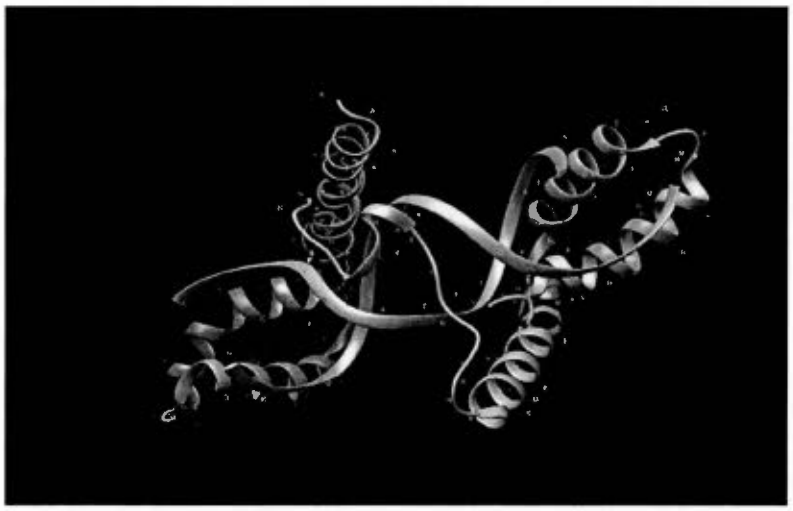


Figure 5.21 : Illustration 3D des positions [16-22] [41-47] [101,107] et [106-112] sur la SRY.

5. 9 Discussion et interprétation des résultats

Dans la partie ci-dessus nous avons présenté nos résultats les plus significatifs. Remarquons que dans le graphique 5.8 quatre fragments montrent une valeur minimale de *RF* sur les positions [6-12], [11-17], [26-32] et [31-37] de la **rhodopsine.** Ces zones correspondent à la zone dont la **température moyenne** varie entre 1871 et 2278°F. Sur le graphique 5.9 deux fragments montrent une valeur minimale de *RF* sur les positions [16-22] et [101-107] de la **SRY.** Ces zones correspondent à la zone ayant une **température moyenne** comprise dans l'intervalle suivant 1064-1345 °F. Le graphique 5.10 présente aussi les fragments ayant une valeur minimale de *RF* sur les positions [6-12], [11-17], [26-32] et [31-37] de la **rhodopsine.** Ces zones correspondent à la zone ayant une **précipitation moyenne** comprise dans l'intervalle suivante : [0.83-0.9] et [0.95-0.96]. Le graphique 5.11 montre un fragment de valeur *RF* minimale [101-107] sur la **SRY** : ce fragment correspond aux zones ayant une **précipitation moyenne** comprise dans l'intervalle suivante [0.99-1.0] [1.1-1.2] [1.2-1.3] [1.3-1.4]. Le graphique 5.12 montre aussi un fragment de valeurs *RF* minimales [41-47] sur la **NADH**. Ces zones correspondent à la zone ayant une **précipitation moyenne** comprise dans l'intervalle suivante [0.9-0.92].

L'analyse de ces données révèle que la distribution géographique des espèces pourrait être liée à des origines génétiques, ce qui affirme notre hypothèse de départ. En effet, l'algorithme a mis en évidence trois protéines pour lesquelles,

des positions significatives ont été retrouvées. Une illustration en 3D de ces positions montre que leurs formes présentent une structure secondaire en hélice α. L'hélice α est « une structure secondaire courante des protéines. Elle est formée par l'enroulement régulier sur elle-même d'une chaîne polypeptidique de forme hélicoïdale avec un pas de rotation droit »[10]. Le coût entropique élevé du repliement fait les interactions stabilisantes du polypeptide moins important. Les liaisons hydrogène dans les hélices α sont considérées moins stables que dans les autres structures, ce qui explique leur présence dans des milieux hydrophobes. Les hélices α ont 3 rôles fonctionnaires principaux :

- Elles jouent un rôle dans les interactions avec l'ADN, c'est le cas de la SRY (Murphy *et al*, 2001).

- Elles peuvent être une constitution principale des protéines transmembranaires, le cas de deux protéines NADH et rhodopsine (Letts et Sazanov, 2015; Ridge et Palcsewski 2007).

- Elles peuvent aussi jouer un rôle mécanique à cause de l'élasticité produite par leur forme en hélice.

[10] https://fr.wikipedia.org/wiki/H%C3%A9lice_alpha

De tels résultats montrent une forte corrélation entre la forme structurale des fragments des gènes qui peuvent expliquer la distribution géographique des espèces et les paramètres climatiques dans les régions considérées. De plus, cela suggère aussi que les données génétiques seules ne peuvent pas expliquer la macroévolution des espèces, ce qui peut être une réponse au débat présente dans la biologie évolutionnaire à savoir : est-ce que la microévolution d'une espèce peut être généralisée pour atteindre la macroévolution entre les taxons. Le couplage des données géographiques avec les données génétiques a permis d'aboutir à une forte inférence expliquant ainsi les résultats obtenus.

- *Rôle des protéines résultantes* :

- *Rhodopsine :* en analysant nos résultats (voir la section 5.4) nous observons que cette protéine a été mise en évidence 15 fois sur 30, pour les paramètres de précipitations moyennes et de 12 fois sur 30 pour le paramètre de la température moyenne. La rhodopsine appartient à la famille G-protéine dont on connaît plusieurs types : Gs (stimulateurs), Gi (inhibiteurs) et Go (inconnu). Elles sont des protéines transmembranaires, considérées comme étant des intermédiaires universels de l'extérieur vers l'intérieur. La rhodopsine est stimulée par la lumière engendrant ainsi une augmentation de GMPc (Guanylate monophosphate cyclique) qui finit par une phosphorylation et qui désensibilise le récepteur, découplant par conséquent la stimulation par

la lumière. Cette protéine se montre pertinente dans les résultats de sorte qu'elle est trop liée a l'extérieur et donc aux paramètres climatiques dominants une zone géographique donnée.

- *La protéine SRY :* En observant nos résultats, nous avons constaté l'influence des paramètres climatiques sur l'évolution de la SRY. En effet, ce gène est considéré comme étant l'inducteur essentiel de la détermination du sexe mal chez les mammifères. La Figure 4.21 représente une illustration en 3D de cette protéine, les fragments significatifs ont la forme structurale en hélice α.

- *La protéine NADH-6* : En analysant nos résultats, nous observons que cette protéine a été mise en évidence spécifiquement en lien avec les précipitations moyennes. Elle est une protéine transmembranaire, localisée dans la membrane interne de la mitochondrie cellulaire. Elle joue un rôle important dans la respiration cellulaire en augmentant ainsi l'énergie libre disponible de l'espèce par sa fonction dans l'échange extracellulaire vu sa localisation transmembranaire. Cela confirme sa présence dans la liste des gènes d'intérêts.

5. 10 Critique de la recherche

La collecte des données fut une étape cruciale et déterminante dans l'analyse de nos résultats. Cependant, nous avons constaté que cette étape n'était pas aussi

triviale que cela en avait l'air, en autre pour la récupération des distributions géographiques des espèces (c.-à-d. latitude et longitude). La recherche n'a pas pu être exploitée à son maximum puisque les données n'étaient pas toujours faciles à traiter. Le fait de ne pas considérer que deux paramètres climatiques dans notre étude la rend moins importante. Cependant, nous avons pu en tirer de bonnes conclusions. Les données brutes que nous avons considérées au début de la recherche ne couvraient pas le sujet de toutes les conditions climatiques. Cependant, l'algorithme a prouvé sa performance par les résultats interprétés et analysés ci-dessus. Plusieurs facteurs peuvent s'ajouter et jouer un rôle important dans la distribution géographique des espèces, par exemple : l'immigration saisonnière de certaines espèces. Pour ce dernier point, il faudrait introduire à l'algorithme un paramètre temporel.

5. 11 Conclusion

Dans ce chapitre nous nous sommes concentrés sur les résultats obtenus, leur analyse et leur interprétation. Nous avons évalué la performance de l'algorithme qui est capable de détecter une liaison entre la distribution géographique des espèces et leur aspect génétique. Nous avons identifié trois gènes pouvant décrire cette relation : la Rhodopsine, la SRY et la NADH d'où nous pouvons nous prononcer positivement concernant l'utilité de notre algorithme. L'illustration en 3D des fragments génétiques trouvés a montré que leur structure commune est en

hélice α. Dans le chapitre suivant, nous présenterons les prochaines étapes à réaliser aux deux niveaux : biologiques et informatiques.

CHAPITRE VI

CONCLUSION ET PERSPECTIVES

Dans le cadre de ma maîtrise en informatique, j'ai développé un nouvel algorithme pour retrouver les relations entre la génétique des espèces et leur distribution géographique. Les résultats que j'ai obtenus sont satisfaisants et répondent bien à mes objectifs de départ, à savoir :

- Déterminer les gènes expliquant la distribution géographique des espèces.
- Localiser les fragments significatifs sur ces gènes.
- Schématiser ces fragments de gènes en 3D et déterminer leur fonctionnalité.

Notre algorithme a été écrit en langage Java, tandis que la partie du prétraitement des données a été programmée en langage Perl. Le programme a été validé sur un jeu de données réel (voir chapitre III) utilisé pour tester nos algorithmes. Nous avons commencé par la reconstruction des arbres phylogéographiques, puis les

arbres phylogénétiques. Nous avons sélectionné les espèces appartenant au groupe des carnivores se trouvant en Amérique du Nord. Nous avons examiné les protéines des 52 espèces étudiées. Le choix des protéines a été fait en considérant que la protéine sélectionnée devrait être séquencée pour au moins la moitié des espèces (c.-à-d. 26 espèces). Nous avons téléchargé les alignements pour 23 protéines de la base de données GenBank.

Puis, nous avons inféré 23 arbres phylogénétiques. En ce qui concerne les données phylogéographiques, nous avons subdivisé la zone choisie en sous-zones, chaque sous-zone se caractérisant par des paramètres climatiques et par la présence de certaines espèces. Nous avons étudié la distribution des espèces dans ces sous-zones selon les deux paramètres climatiques suivantes : la température moyenne et les précipitations moyennes. Nos résultats sont prometteurs et encourageants : nous avons mis en évidence trois protéines : 1) la Rhodopsine, qui joue un rôle dans la vision, 2) la NADH, qui a un rôle important dans la respiration et 3) la SRY, qui détermine le sexe de l'espèce et est localisée sur le chromosome Y. En plus, nous avons déterminé les positions cruciales sur chaque gène résultant. L'illustration de ces régions a été en amont d'une analyse faite sur les aboutissements (voir le chapitre V). Ces résultats prometteurs nous ont amené à présenter nos travaux dans des conférences nationales, telles que le symposium de Biologie et Écologie à l'Université de Montréal en 2016 et l'ACFAS à McGill en 2017.

Finalement, au niveau des perspectives nous signalons la continuité et l'amélioration de cette étude qui a été liée aux plusieurs domaines scientifiques tels que : la génétique, l'informatique et la bioinformatique, qui représentent actuellement des défis majeurs.

6. 1 Du point de vue génétique

Nous proposons le passage vers les réseaux réticulés. Le fait de considérer que toutes les espèces ont évolué depuis un seul ancêtre commun semble être non réaliste. Plusieurs problèmes phylogénétiques ne peuvent pas être expliqués par un simple arbre binaire, citons les plus importants : l'hybridation entre les espèces et le transfert latéral des gènes (surtout dans les communautés microbiennes) (Makarenkov et Legendre, 2004). Il serait donc intéressant d'utiliser les réseaux phylogénétiques à la place des arbres dans notre étude.

Dans l'étude de Makarenkov et Legendre de 2004, les auteurs ont expliqué le passage d'un arbre phylogénétique vers un réseau, en proposant un algorithme permettant d'ajouter à un arbre phylogénétique une branche de longueur minimale telle que cet ajout n'affecte pas les longueurs respectives des autres branches préexistantes.

6. 2 Du point de vue informatique

Concernant la partie informatique, nous pourrions proposer l' amélioration suivante :

- Paralléliser l'algorithme

Vu la quantité énorme des données génétiques, la version séquentielle du programme ne sera pas toujours capable de répondre à notre objectif, d'où nous proposons un passage envers la version parallélisée. Plusieurs environnements en « big data » peuvent être utiles dans ce sens, par exemple : Spark Apache ou Hadoop Mapreduce (Taylor, 2010).

Notons aussi que d'un point de vue écologique, il serait important d'avoir un jeu de données complet contenant plus de paramètres climatiques et plus d'espèces. Finalement, l'algorithme de cette étude a permis de lier la phylogéographie à la génétique, en mettant en évidence la corrélation entre les protéines suivantes : la rhodopsine, la SRY et la NADH et les paramètres climatiques. L'utilisation des fenêtres coulissantes a permis de mieux cibler les fragments d'intérêt des protéines pour ensuite les visualiser en 3D. Suivant cette idée il serait aussi intéressant d'incorporer à notre étude l'outil de visualisation GenGIS (Parks *et al*, 2009) qui permettrait de présenter géographiquement cette corrélation, ainsi que de la valider statistiquement (voir la Figure 6.1).

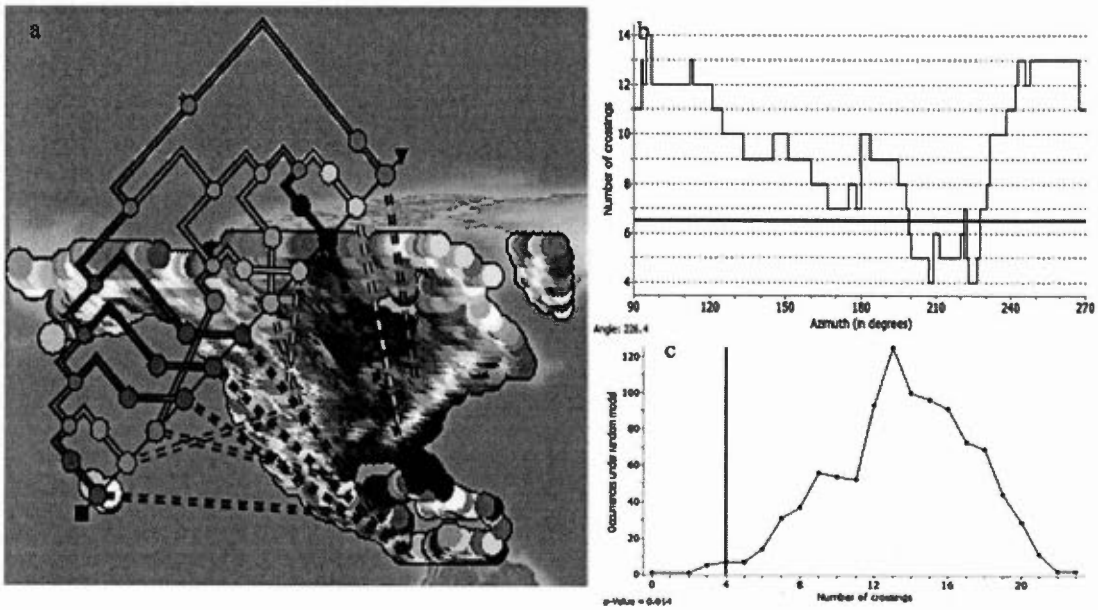


Figure 6.1 : (a) arbre de SRY corrélé au gradient de température ;(b) Analyse des axes géographiques de GenGIS (la ligne rouge représente un p-valeur de 0.05 pour les permutations de Monte-Carlo) ; (c) Permutation de Monte-Carlo pour l'arbre a (la ligne rouge représente le nombre de croisements du modèle testé).

BIBLIOGRAPHIE

Avise, J. C., Arnold, J. ,Ball, R. M., Bermingham, E., Lamb, T., Neigel, J. E ,Reeb, C. A., et Saunders, N. C. (1987). « Intraspecific phylogeography: the mitochondrial DNA bridge between population genetics and systematics.» *Annual review of ecology and systematics* 18(1): 489-522.

Avise, J. C. et Nelson, W. S. (1989). «Molecular Genetic Relationships of the Extinct Dusky Seaside Sparrow.» *Science* 243(4891) : 646-648.

Barrett, M., Donoghue, M. J et Sober, E.(1991). « Against consensus. » *Systematic Zoology* 40(4) : 486-493.

Barthélemy, J. P., et Guénoche, A. (1991). «Trees and proximity representations.» *John Wiley & Sons.*

Barthélemy, J. P., et Luong, N. X. (1987). «Sur la topologie d'un arbre phylogénétique: aspects théoriques, algorithmes et applications à l'analyse de données textuelles». *Mathématiques et Sciences humaines, 100*, 57-80.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J.,Wheeler, D. L., (2008). « GenBank. » *Nucleic Acids Research* 36(suppl 1) : D25-D30.

Bermingham, E., Moritz, C. (1998). « Comparative phylogeography: concepts and applications. » *Mol Ecol* 7(4) : 367-369.

Besson, S. (2012). « Les espèces d'invertébrés, ingénieurs de la biodiversité, sont menacées. » *Actualités News -Environnement.*

Bekinschtein, P., Cammarota, M., Katche, C., Slipczuk, L., Rossato, J. I., Goldin, A., et Medina, J. H. (2008). « BDNF is essential to promote persistence of long-term memory storage ». *Proceedings of the National Academy of Sciences, 105*(7), 2711-2716.

Bluis, J.et Shin, D.-G.(2003). « Nodal distance algorithm: Calculating a phylogenetic tree comparison metric. Bioinformatics and Bioengineering ».2003. *Proceedings. Third IEEE Symposium on, IEEE.*

Boc, A. et Makarenkov, V.(2012). « T-REX : a web server for inferring, validating and visualizing phylogenetic trees and networks. » *Nucleic Acids Research* .40(W1) : W573-W579.

Boc, A., Philippe, H. et Makarenkov, V. (2010). « Inferring and validating horizontal gene transfer events using bipartition dissimilarity. » *Systematic biology* : syp103.

Bourguignon, P. Y., et Robelin, D. (2004). « Modèles de Markov parcimonieux : sélection de modèles et estimation ». *Proceedings of JOBIM. Montréal.*

Buneman, O. P. (1971). « The recovery of trees from measures of dissimilarity ». *Mathematics in the archaeological and historical sciences.*

Cracraft, J. (1983). «Species concepts and speciation analysis ». *Current ornithology*, Springer : 159-187.

Danchin, A. (2011). « History of biology 1800-1849 ».

Darlu, P. et Tassy, P. (1993). « La reconstruction phylogénétique». *Concepts et méthodes. Paris, France.*

Darwin, C. (1859). « De l'origine des espèces par voie de sélection naturelle. » *On the Origin of Species by Means of Natural Selection.*

Donald, D. B. et Alger, D. J. (1993). « Geographic distribution, species displacement, and niche overlap for lake trout and bull trout in mountain lakes. » *Canadian Journal of Zoology* 71(2) : 238-247.

Eddy, S. R. (1996). « Hidden markov models. » *Current opinion in structural biology* 6(3) : 361-365.

Edgar, R. C. (2004). « MUSCLE : multiple sequence alignment with high accuracy and high throughput. » *Nucleic Acids Research* 32(5) : 1792-1797.

Etien, A. (2006) . «Ingénierie de l'alignement : concepts, modèles et processus : la méthode ACEM pour l'alignement d'un système d'information aux processus d'entreprise», *Paris 1*

Felsenstein, J. (1981). « Evolutionary trees from DNA sequences: a maximum likelihood approach. ». *Journal of molecular evolution* 17(6) : 368-376.

Felsenstein, J. (2004). « Inferring phylogenies », *Sinauer Associates Sunderland.*

Fitch, W. M. (1971). «Toward defining the course of evolution: minimum change for a specific tree topology ». *Systematic biology* 20(4) : 406-416.

Foulds, L. et Robinson, R. (1981). « Enumeration of binary phylogenetic trees ». *Combinatorial Mathematics VIII*, Springer : 187-202.

Delsuc, F., et Douzery, E. J. (2004). « Les méthodes probabilistes en phylogénie moléculaire :(1) Les modèles d'évolution des séquences et le maximum de vraisemblance ». *Biosystema*, *22*, 59-74.

Fukunaga, K. et Narendra, P. M. (1975). « A branch and bound algorithm for computing k-nearest neighbors ». *Computers, IEEE Transactions on* 100(7) : 750-753.

Gomez-Zurita, J, Petitpierre, E. , et Juan, C.(2000). « Nested cladistic analysis, phylogeography and speciation in the Timarcha goettingensis complex (Coleoptera, chrysomelidae)». *Mol Ecol* 9(5) : 557-570.

Guindon, S. Dufayard ,J.-F., Lefort, V., Anisimova, M. , Hordijk, W., et Gascuel, O.(2010). «New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of *PhyML* 3.0». *Systematic biology* 59(3) : 307-321.

Haeckel, E. (1866). « Generelle Morphologie der Organismen». (Vol. 2). *Éditeur de Georg Reimer.*

Hanada, H., Nakamura, A., et Kudo, M. (2011). « A practical comparison of edit distance approximation algorithms ». *Granular Computing (GrC), 2011 IEEE International Conference on, IEEE.*

Ihaka, R., et Gentleman, R. (1996). «R: a language for data analysis and graphics ». *Journal of computational and graphical statistics*, *5*(3), 299-314.

Iliopoulos, D., Volakakis, N., Tsiga, A., Rousso, I., et Voyiatzis, N. (2004) « Description and molecular analysis of SRY and AR genes in a patient with 46, XY pure gonadal dysgenesis (Swyer syndrome) ». In *Annales de génétique* (Vol. 47, No. 2, pp. 185-190). Elsevier Masson.

Ingelsson, E., Schaefer, E. J., Contois, J. H., McNamara, J. R., Sullivan, L., Keyes, M. J., et Vasan, R. S. (2007). « Clinical utility of different lipid measures for prediction of coronary heart disease in men and women ». *Jama*, *298*(7), 776-785.

Jukes, T. H. et Cantor, C. R. (1969). « Evolution of protein molecules ». *Mammalian protein metabolism* 3(21) : 132.

Kays, R. W. et Wilson, D. E.(2009) « Mammals of North America ». *Princeton University Press.*

Kimura, M. (1980). « A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences». *Journal of molecular evolution* 16(2) : 111-120.

Kimura, M. et Ohta, T. (1972). «On the stochastic model for estimation of mutational distance between homologous proteins ». *Journal of molecular evolution*, *2*(1), 87-90.

Katoh, K., Kuma, K. I., Toh, H., et Miyata, T. (2005). «MAFFT version 5: improvement in accuracy of multiple sequence alignment». *Nucleic acids research*, *33*(2), 511-518.

Kosiol, C. et Goldman, N. , (2005). « Different versions of the Dayhoff rate matrix ». *Molecular biology and evolution* 22(2) : 193-199.

Lamarck, J.-B.-P. (1809). « Philosophie zoologique».

Letts, J. A., et Sazanov, L. A. (2015). « Gaining mass: the structure of respiratory complex I,from bacterial towards mitochondrial versions». *Current opinion in structural biology*, *33*, 135-145.

Li, K.-B. (2003). «ClustalW-MPI: ClustalW analysis using distributed and parallel computing». *Bioinformatics* 19(12) : 1585-1586.

Liu, Y., Fiskum, G., et Schubert, D.(2002). «Generation of reactive oxygen species by the mitochondrial electron transport chain». *Journal of neurochemistry* 80(5) : 780-787.

Mailund, T., et Pedersen, C. N. (2004). «QDist—Quartet distance between evolutionary trees». *Bioinformatics*, *20*(10), 1636-1637.

Makarenkov, V. (2001). «T-REX : reconstructing and visualizing phylogenetic trees and reticulation networks ». *Bioinformatics* 17(7) : 664-668.

Makarenkov, V., et Lapointe, F. J. (2004). « A weighted least-squares approach for inferring phylogenies from incomplete distance matrices ». *Bioinformatics*, *20*(13), 2113-2121.

Makarenkov, V., et Legendre, P. (2004). « From a phylogenetic tree to a reticulated network». *Journal of Computational Biology*, *11*(1), 195-212.

Marx, V. (2013). «Biology: The big challenges of big data». *Nature* 498(7453) : 255-260.

Mayr, E. (1942). «Systematics and the origin of species, from the viewpoint of a zoologist ». *Harvard University Press*.

Mullahy, J. (1986). «Specification and testing of some modified count data models. » *Journal of econometrics* 33(3) : 341-365.

Neyman, J. (1971). «Molecular studies of evolution: a source of novel statistical problems ».

Park, D. H., Kim, S. K., Shin, I. H., et Jeong, Y. J. (2000). « Electricity production in biofuel cell using modified graphite electrode with neutral red ». *Biotechnology Letters*, *22*(16), 1301-1304.

Parks, D. H., Porter, M., Churcher, S., Wang, S., Blouin, C., Whalley, J.et Beiko, R. G. (2009). «GenGIS: A geospatial information system for genomic data». *Genome Research*, *19*(10), 1896-1904.

Pattengale, N. D., Gottlieb, E. J., et Moret, B. M. (2007). « Efficiently computing the Robinson-Foulds metric ». *Journal of Computational Biology*, *14*(6), 724-735.

Peichl, L., Behrmann, G., et Kröger, R. H. (2001). «For whales and seals the ocean is not blue: a visual pigment loss in marine mammals». *European Journal of Neuroscience*, *13*(8), 1520-1528.

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., et Ferrin, T. E. (2004). «UCSF Chimera—a visualization system for exploratory research and analysis ». *Journal of computational chemistry*, *25*(13), 1605-1612.

Pischon, T., Girman, C. J., Sacks, F. M., Rifai, N., Stampfer, M. J., et Rimm, E. B. (2005). « Non–high-density lipoprotein cholesterol and apolipoprotein B in the prediction of coronary heart disease in men ». *Circulation*, *112*(22), 3375-3383.

Plotree, D. et Plotgram, D., (1989). « PHYLIP-phylogeny inference package (version 3.2) ». *cladistics* 5 : 163-166.

Posada, D. et Crandall, K. A., (2001). « Selecting the best-fit model of nucleotide substitution». *Systematic biology* 50(4) : 580-601.

Ridge, K. D., et Palczewski, K. (2007). «Visual rhodopsin sees the light: structure and mechanism of G protein signaling». *Journal of Biological Chemistry*, *282*(13), 9297-9301.

Sadler, J. E. (1998). « Biochemistry and genetics of von Willebrand factor ». *Annual review of biochemistry*, *67*(1), 395-424.bb.

Sadler, D. R. (1989). « Formative assessment and the design of instructional systems ». *Instructional science* 18(2) : 119-144.

Saitou, N. et Nei, M. (1987). «The neighbor-joining method: a new method for reconstructing phylogenetic trees». *Molecular biology and evolution* 4(4) : 406-425.

Sakai, A. K., Allendorf, F. W., Holt, J. S., Lodge, D. M., Molofsky, J., With, K. A. et McCauley, D. E. (2001).«The population biology of invasive species ». *Annual review of ecology and systematics*, *32*(1), 305-332

Slatkin, M. (1987). « Gene flow and the geographic structure of natural populations». *Science* 236 (4803) : 787-792.

Slatkin, M. et Hudson, R. R., (1991). « Pairwise comparisons of mitochondrial DNA sequences in stable and exponentially growing populations ». *Genetics* 129(2) : 555-562.

Sneath, P. H et Sokal, R. R. (1973). « The principles and practice of numerical classification ». *Numerical taxonomy.*

Stamatakis, A. (2006). « RAxML-VI-HPC : maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models ». *Bioinformatics* 22 (21) : 2688-2690.

Swofford, D. L. et. Begle, D. P (1993). « *PAUP :* Phylogenetic Analysis Using Parsimony». *Version 3.1, March 1993, Center for Biodiversity, Illinois Natural History Survey.*

Taberlet, P., Fumagalli, L., Wust-Saucy, A.-G., Cosson, J.-F.,(1998). « Comparative phylogeography and postglacial colonization routes in Europe ». *Mol Ecol* 7(4) : 453-464.

Tahiri, N. et Makarenkon,V. (2012). « Un nouvel algorithme pour retrouver les relations phylogénétiques entre la distribution géographique des espèces et leurs compositions génétiques». *rapport de recherche.*

Tavaré, S. (1986). « Some probabilistic and statistical problems in the analysis of DNA sequences». *Lectures on mathematics in the life sciences* 17 : 57-86.

Taylor, R. C. (2010). « An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics ». *BMC bioinformatics*, *11*(12), S1

Wheeler, T. J. (2009). « Large-scale neighbor-joining with NINJA ». *International Workshop on Algorithms in Bioinformatics, Springer.*

West-Eberhard, M. J. (2005). « Developmental plasticity and the origin of species differences ». *Proceedings of the National Academy of Sciences* 102(suppl 1): 6543-6549

Whelan, S., et Goldman, N. (2001). « A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach ». *Molecular biology and evolution*, *18*(5), 691-699.

Yap, I. et Nelson, R., (1996). « WinBoot: a program for performing bootstrap analysis of binary data to determine the confidence limits of UPGMA-based dendrograms ». *International Rice Research Institute, Manila* : 1-22.

Yamada, K., et Nabeshima, T. (2003). « Brain-derived neurotrophic factor/TrkB signaling in memory processes ». *Journal of pharmacological sciences*, *91*(4), 267-270.

Yusnita, Y., Norsiah, M. D., et Rahman, A. J. (2010). « Mutations in mitochondrial NADH dehydrogenase subunit 1 (mtND1) gene in colorectal carcinoma ». *The Malaysian journal of pathology*, *32*(2), 103-110.