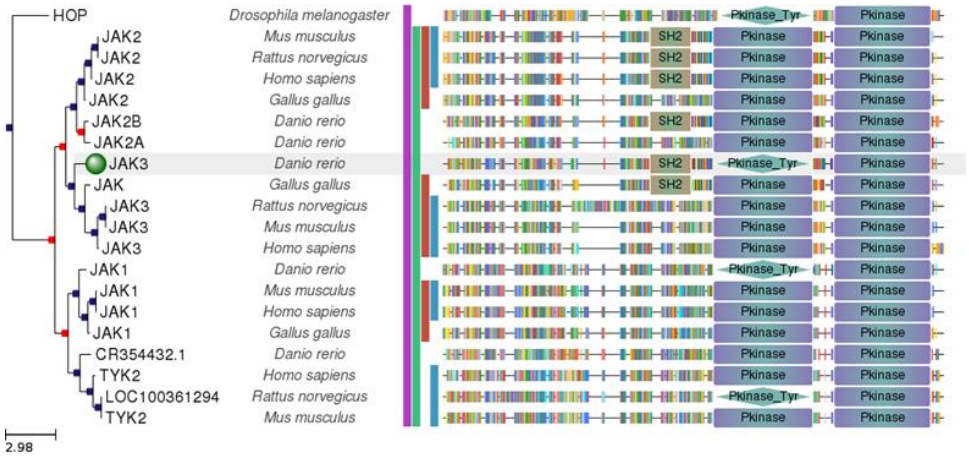
LE PASSAGE DU GÉNOME AU PHYLOME

La création et l'analyse de l'arbre phylogénétique



Travail de Recherche

Élève : Bruna Adell

Année : 2ème Batxibac, 2017-2018

« La vérité scientifique sera toujours plus belle
que les créations de notre imagination
et que les illusions de notre ignorance. »

Claude Bernard (1813 - 1878)

Biologiste, médecin, physiologiste et scientifique.

SOMMAIRE

1. INTRODUCTION

Depuis le début de l'année, je me demandais quel serait le thème de mon travail. Ensuite, j'ai eu du mal à choisir la thématique puisque, d'abord je désirais faire une étude sur le cerveau et la médecine. Puis, j'ai pensé aux chimères ou aux rôles de genre, mais à mon avis, c'était des sujets trop complexes pour cette sorte de travail.

Donc, j'ai finalement choisi de faire de la recherche sur la phylogénétique, car je me suis rendue compte que je suis vraiment intéressée par la génétique, l'évolution et les mutations. En outre, je désirais faire un travail innovateur et intéressant. J'ai pensé à montrer comment créer mon propre arbre phylogénétique à partir d'une protéine concrète et le comparer aussi avec un arbre phylogénétique général afin de montrer s'ils concordent ou pas. Ainsi, j'ai utilisé les sites PhylomeDB et UniProt (dont je vais parler dans mon travail), pour faire mes recherches. Les données sur ces sites sont générées automatiquement et analysées comme des études générales, mais la plupart des arbres phylogénétiques n'ont été jamais explorés en détail manuellement par personne. En plus, j'ai employé d'autres sites comme Phylogeny.fr.

C'est pour cela que j'ai décidé de me fixer, une fois choisi le thème du travail, un but. Alors, j'ai décidé de choisir quelques protéines différentes au hasard. Com par exemple Phy0036ZQQ ou Phy0007XBH. Puis, sur la base de différents arbres obtenus automatiquement dans le site Phylome DB, j'ai fait une analyse des modifications que ces protéines ont subi, par rapport à l'évolution; et finalement, j'ai comparé les résultats obtenus avec un arbre phylogénétique modèle de l'évolution. Ensuite, grâce à des programmes comme Pylogeny.fr, j'ai pu créer mon propre arbre phylogénétique, à partir de la protéine P08842, et ensuite je l'ai comparé avec l'arbre de l'évolution.

1.1. LA MÉTHODOLOGIE

La procédure que j'ai suivie pour commencer ce projet a été la suivante : tout d'abord, j'ai fait de la recherche sur Phylome DB et j'ai lu les différentes guides. Puis, j'ai choisi des protéines au hasard, le site a créé des arbres automatiques et j'ai analysé les protéines et les arbres phylogénétiques. Ensuite, j'ai comparé ces arbres phylogénétiques avec ceux déjà faits qui sont utilisés comme modèle pour comprendre l'évolution des espèces et je suis arrivée à des certaines conclusions. Tout cela, tandis que j'ai montré comment fonctionne le programme et comment on interprète les données. Après, j'ai cherché la protéine P08842 sur Phylome DB et, comment j'ai eu l'impression

qu’elle était suffisamment fiable, j’ai mis sa séquence protéique dans Uniprot et puis j’ai montré les résultats dans Phylogeny.fr. Dans la partie théorique du travail, j’ai d’abord expliqué certains concepts de base et génériques de la génétique, ainsi que le processus de synthèse des protéines, lesquelles j’ai analysées dans les arbres. Et, dans la dernière partie du projet, je me suis occupée de la phylogénétique, la création et l’analyse des arbres phylogénétiques.

1.2. LES HYPOTHÈSES

Les arbres d’espèces sont généraux et crées à partir de groupes de gènes. Donc, l’évolution d'un gène ne doit pas obligatoirement suivre l'évolution des espèces.

Dans le domaine des organismes modèles, qui sont des espèces qui ont été très étudiées, nous trouverons que les résultats attendus (selon les arbres phylogénétiques généraux) sont les résultats que j’ai obtenus dans mes arbres phylogénétiques, donc les hypothétiques diagrammes évolutifs généraux correspondent aux cas concrets.

2. INTRODUCTION À LA GÉNÉTIQUE

La comparaison des séquences d'ADN ou de protéines qui proviennent d'organismes différents peut reconstituer l'histoire de l'évolution de ces macromolécules. Cette histoire est normalement représentée comme un arbre phylogénétique. Si un arbre généalogique nous informe sur le lien de parenté entre les membres d'une famille, un arbre phylogénétique établit des relations entre molécules apparentées.

2.1. L'ADN (ACIDE NUCLEIQUE)

L'acide désoxyribonucléique est essentiel pour comprendre la génétique. L'ADN, présent dans toutes les cellules ainsi que chez de nombreux virus, est un polymère (assortiment) de nucléotides (monomère), reliés entre eux par des liaisons phosphodiester et qui forment une double hélice (sauf pour quelques virus). Donc, on dit que l'ADN est bicaténaire ou de double brin, et chaque chaine antiparallèle[1] de desoxinucléotides se synthétise dès l'extrémité 5'-P vers l'extrémité 3'-OH.

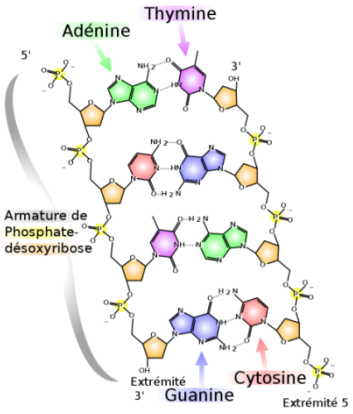


*Figure 1. Structure chimique de l'ADN illustrant les paires **AT** et **GC** formant la double hélice*

La source de toutes les photographies se trouve aux annexes du travail.

2.1.1. LES DÉSOXYRIBONUCLÉOTIDES

Au même temps, chaque nucléotide (desoxyribonucléotide) est constitué par : une base nucléique – *adénine* (A), *cytosine* (C), *guanine* (G) ou *thymine* (T) - un désoxyribose (un ribose, c'est-à-dire un monosaccaride avec cinq carbones, auquel lui manque un groupe hydroxyle (-OH) au troisième carbone), et un acide phosphorique (H_3PO_4).

Les bases nucléiques forment des liaisons hydrogène entre elles en respectant la complémentarieté : A s'accouple seulement avec T à l'aide de deux liaisons hydrogène, et C seulement avec G grâce à trois liaisons.

Concrètement, dans le cas de l'être humain, la macromolécule ADN est composée par $5{,}6 \cdot 10^9$ paires de nucléotides (toujours AT/TA, CG/GC).

1 Chaine antiparallèle : Les deux chaines son parallèles en directions inverses. L'une est orientée en sens 5'-3' et l'autre en sens inverse.

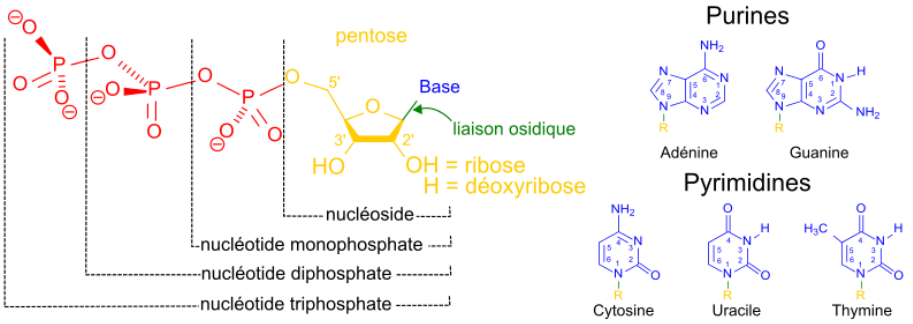
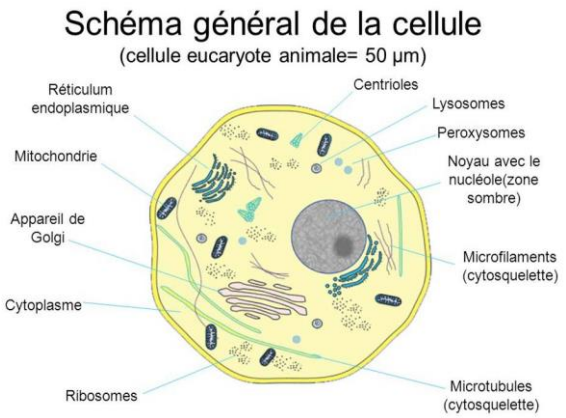
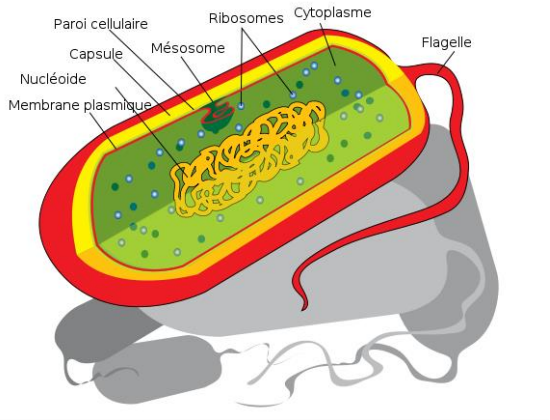


Figure 2. Structure des nucléotides. Les bases nucléiques peuvent se diviser entre Purines (A et G) et Pyrimidines (C, U et T)

D'un côté, dans les cellules eucaryotes, l'ADN se trouve principalement dans le noyau, mais aussi dans les mitochondries et les chloroplastes. L'ADN nucléaire est associé à des histones[2] et à des protéines non histoniques. D'autre part, l'ADN des procaryotes (sans noyau) est lié à des protéines égales aux histones, non histoniques et au ARN (formant le nucléoïde).





Figures 3 et 4. Modèles de cellule eucaryote (animale), puis procaryote.

On peut distinguer trois niveaux structurels de l'ADN :

- **La structure primaire de l'ADN :** c'est la séquence de nucléotides d'une seule chaine ou filament. Avec cette séquence, il est possible d'emmagasiner une information déterminée. C'est le « message biologique » ou « information génétique ».
- **La structure secondaire de l'ADN :** c'est la disposition dans l'espace des deux chaines de nucléotides en double hélice, avec les bases nucléiques mises face à

2 Histones : Des protéines localisées dans le noyau des cellules eucaryotes et dans les archéobactéries. Elles sont les principaux constituants protéiques des chromosomes. Elles permettent la compaction de l'ADN, puisque l'ADN s'y entoure autour.

face et liées grâce à des liaisons hydrogène.

- **La structure tertiaire de l'ADN :** la fibre d'ADN se trouve enroulée sur elle-même et elle forme une super hélice (surenroulement de l'ADN) qui réduit la longitude de l'ADN et facilite le processus de la duplication.

Puis, comme l'ADN n'est pas suffisamment emballé, il est nécessaire qu'il retourne sur les histones (sauf pour les spermatozoïdes) afin qu'il tient dans le noyau cellulaire et on puisse obtenir les chromosomes.

2.2. L'ARN

L'acide ribonucléique ou ARN est un acide nucléique, ainsi que l'ADN, et il est présent dans les virus, les cellules procaryotes et eucaryotes[3]. Il est constitué par nucléotides avec un acide phosphorique, un ribose et les bases nucléiques : adénine, guanine, cytosine et uracile. Ces nucléotides sont liés grâce à des liaisons phosphodiester dans le sens 5'-3'.

Dans les cellules eucaryotes, il y a cinq fois plus d'ARN que d'ADN. En plus, il est monocaténaire, sauf dans le vas des rétrovirus.

Il y a plusieurs types d'ARN, mais les plus importants sont :

- ARN messager (ARNm) : Il est formé par transcription de l'ADN. Son rôle consiste à transporter l'information génétique recueillie du noyau vers le cytoplasme où elle sera traduite en protéine ou polypeptide par les ribosomes du réticulum endoplasmique.

  ARN de transfert (ARNt) : Il sert à « traduire » les codons de l'ARNm en acides aminés. Ce sont des molécules qui se placent sur les sites du ribosome où va être lu l'ARN messager. Un ARNt est un brin court qui a un anticodon sur sa boucle, et un acide aminé attaché à l'autre extrémité et qui sera transféré à la protéine en formation.

3 Cellule procaryote : cellule dont l'acide désoxyribonucléique se trouve dans le cytoplasme et non à l'intérieur d'un noyau.

Cellule eucaryote : cellule dont l'acide désoxyribonucléique se trouve dans un noyau différencié, enveloppé par une membrane.

- ARN ribosomique : Associé à des protéines, il forme le ribosome qui constitue la tête de lecture de l'information génétique transcrite par l'ARN messager.

- RNA réglementaires : Ces ARN servent à réguler le processus d'expression des gènes et ils sont non codificateurs (ncRNA). Même s'ils ne codent pas de protéines, ils peuvent agir comme des riborégulateurs, et leur principale fonction est la régulation post-traductionnelle de l'expression des gènes. Ils se trouvent dans les cellules procaryotes et les eucaryotes et participent à la reconnaissance spécifique des cibles d'acide nucléique cellulaire grâce à un couplage de base complémentaire, contrôlant la croissance et la différenciation des cellules.

2.3. LE GÈNE

De nos jours, un gène est défini comme une séquence d'acide nucléique (d'ADN, sauf chez certains virus dont l'acide nucléique est l'ARN) susceptible d'être transcrite en ARN. En plus, s'il est ensuite traduit en protéine, la séquence est nommée « codante ». La plupart des gènes commencent par une séquence de nucléotides appelée promoteur, dont le rôle est de permettre l'initiation mais surtout la régulation (pas tous les gènes sont exprimés dans toutes les cellules de notre corps) de la transcription de l'ADN en ARN, et se termine par une séquence terminatrice appelée terminateur, qui marque la fin de la transcription. Il y a environ 21.000 gènes chez l'humain.

2.4. LE GÉNOME

Un génome est l'ensemble complet d'ADN d'un organisme (sauf dans le cas des virus d'ARN) ; donc, il comprend tous ses gènes (les régions codantes), l'ADN non codant et le matériel génétique des mitochondries et des chloroplastes. Il est appelé aussi comme le « matériel génétique ». Chaque génome contient toute l'information nécessaire pour construire et maintenir cet organisme. Chez l'homme, une copie de l'ensemble du génome (plus de trois milliards de paires de bases d'ADN) est contenue dans toutes les cellules qui ont un noyau.

2.5. LA PROTÉINE

Les protéines sont des biomolécules composées par des chaînes linéaires d'acides aminés unis par des liens peptidiques. En outre, elles contiennent l'élément azote

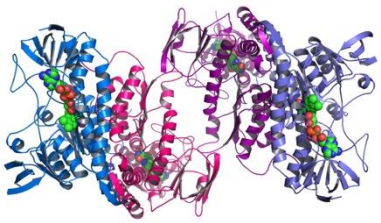


Figure 5. La protéine

ainsi que le carbone, l'hydrogène et l'oxygène. Les protéines exercent une énorme quantité de fonctions différentes:

- Structurelle : C'est la fonction la plus importante d'une protéine (par exemple, le Collagène)
- Contractile (actine et myosine)
- Enzymatique (sucrase et pepsine)
- Homéostatique: ils collaborent à la maintenance du pH
- Immunologique (anticorps)

La peau et les muscles sont composés de protéines; les anticorps et les enzymes sont des protéines; certaines hormones sont aussi des protéines.

2.5.1. LES ACIDES AMINÉS

Les protéines sont constituées par centaines d'unités plus petites appelées acides aminés qui sont attachés l'un à l'autre par des liaisons peptidiques, formant une longue chaîne (protéine ou polypeptide). Les protéines varient en longueur et en complexité en fonction du nombre et du type d'acides aminés qui composent la chaîne. Il existe environ vingt acides aminés différents, chacun ayant une structure chimique et des caractéristiques différentes. La structure protéique finale dépend des acides aminés qui la composent. En plus, la fonction de la protéine est directement liée à la structure de celle-ci. Si la structure tridimensionnelle de la protéine est altérée en raison d'une modification de la structure des acides aminés (une augmentation de la température, par exemple), la protéine devient dénaturée et elle perd sa fonction.

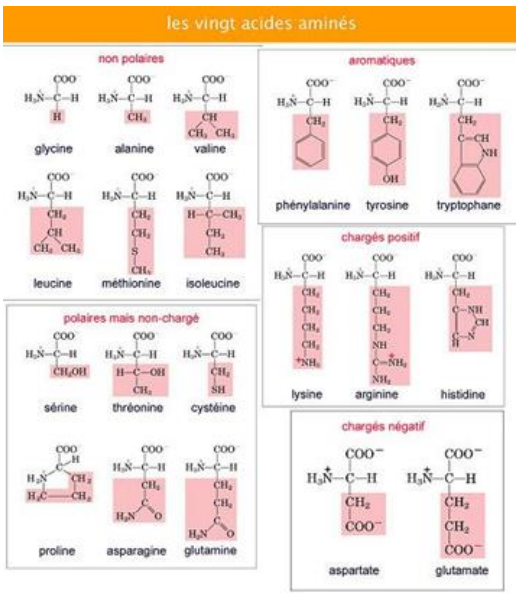


Figure 6. Les vingt acides aminés

Il existe seulement *huit* acides aminés essentiels à l'être humain : *l'isoleucine,* la *leucine,* la *lysine,* la *méthionine,* la *phentylalanine*, la *thréonine*, le *tryptophane* et la *valine*. On dit qu'ils sont essentiels, car l'organisme ne peut pas les fabriquer, et a donc besoin d'un apport externe (par l'alimentation). Un acide aminé manquant restreint la synthèse des

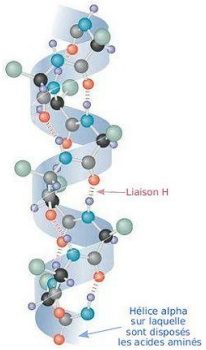
protéines et peut conduire à une déficience en protéines, c'est-à-dire, à un type grave de malnutrition.

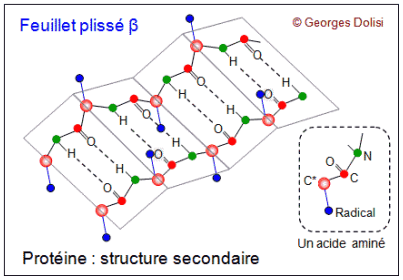
2.5.2. LA STRUCTURE DES PROTÉINES

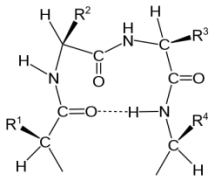
La géométrie tridimensionnelle d'une molécule de protéine est très importante pour sa fonction. Ainsi, quatre niveaux de structure sont utilisés pour déterminer les caractéristiques d'une protéine.

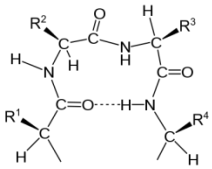
– Le premier niveau, ou la **structure primaire**, est la séquence linéaire des acides aminés qui crée la chaîne peptidique. La structure primaire d'une protéine est le fruit de la traduction de l'ARNm en séquence protéique par le ribosome.

– Dans la **structure secondaire**, la liaison hydrogène entre les groupes amide (-NH) et carbonyle (-CO) du squelette peptidique, crée une géométrie tridimensionnelle concrète. Il existe trois principales catégories de structures secondaires: les hélices, les feuillets et les coudes.









Figures 7, 8 et 9. Les hélices, les feuillets et les coudes.

– La **structure tertiaire** correspond au repliement de la chaîne polypeptidique dans l'espace. Donc, on parle de « structure tridimensionnelle ». Celle-ci est directement liée à sa fonction.

– Les **structures quaternaires** décrivent l'apparence des protéines lorsqu'une protéine est composée de deux ou plusieurs chaînes polypeptidiques unies par des liaisons non covalentes. *L'hémoglobine* est un exemple de structure quaternaire.

2.5.3. LA SYNTHÈSE DES PROTÉINES

La synthèse des protéines se divise en deux processus : transcription et traduction. Néanmoins, il y a un pas intermédiaire entre les deux qui s'appelle les modifications post-transcriptionnelles.

- **LA TRANSCRIPTION**

La synthèse des protéines commence avec la transcription d'un gène d'ADN en une molécule d'ARN messager (ARNm). Ce processus se déroule à l'intérieur du noyau des cellules d'eucaryotes et dans le cytosol des cellules de procaryotes.

Ce processus commence grâce à un promoteur, c'est-à-dire, à une région de l'ADN qui contrôle l'initiation de la transcription d'une partie concrète du même ADN. Ces séquences sont riches en paires adénine-thymine, et ceci facilite l'ouverture de la double hélice d'ADN par une hélicase (enzyme), libérant l'un des deux chaines pour être copié en ARN. Une ARN polymérase (ARN polymérase II pour les eucaryotes) lit ce segment d'ADN dans le sens 3' → 5' tout en synthétisant l'ARN messager dans le sens 5' → 3'. C'est-à-dire, la chaine d'ARN seulement peut se construire dans un sense concret.

Chez les procaryotes, le produit de la transcription d'un gène de protéine est directement ARN messager. Cependant, pour les eucaryotes on parle de transcrit primaire, qui doit encore subir une maturation de l'ARN messager avant de devenir fonctionnel. Les modifications post-transcriptionnelles provoquent cette maturation.

- **LES MODIFICATIONS POST-TRANSCRIPTIONNELLES**

Les principales modifications post-transcriptionnelles de l'ARN pré-messager sont l'ajout d'une coiffe de 7-méthylguanosine triphosphate à l'extrémité 5' et d'une queue poly(A) à l'extrémité 3'. Ensuite, l'épissage a lieu, consistant en l'élimination des introns (segments du gène qui ne codent pas un polypeptide) séparant les exons (qui sont codants). Donc, on coupe et on épisse des parties de l'ARN). Cet épissage peut être variable (épissage alternatif), et ainsi, on pourra obtenir des polypeptides (protéines) différents.

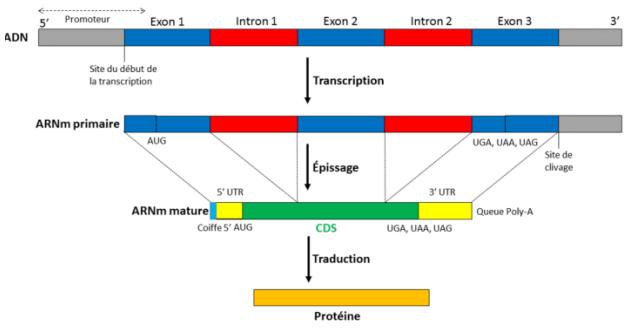


Figure 10. Processus de transcription et d'épissage d'un ARN messager.

– LA TRADUCTION

Elle a lieu au ribosome qui reçoit les ARNt (transporteurs des acides aminés) et l'ARNm (qui contient l'information nécessaire à la synthèse de la protéine). La traduction est le passage des codons ou triplets de nucléotides à acides aminés. Le **code génétique**, dont on parlera plus tard, est la clé qui permet de passer d'un langage à l'autre.

La traduction, qui se déroule dehors le noyau, se décompose en trois étapes :

- **L'initiation** : Dans le cytoplasme, l'assemblage des acides aminés a lieu aux ribosomes. La synthèse d'une protéine débute toujours par le même acide aminé : la méthionine (MET), codé par le même codon de l'ARNm, ou codon initiateur : AUG. Ce codon situé au niveau du ribosome s'associe avec l'ARNt.
- **L'élongation** : Le ribosome passe au codon suivant sur l'ARNm. Un nouvel ARNt se fixe sur le deuxième codon de l'ARNm. La première liaison peptidique se forme entre les deux acides aminés présents. Le ribosome se déplace à nouveau sur l'ARNm et ainsi de suite, assurant l'élongation de la chaîne peptidique.
- **La terminaison** : La terminaison se produit quand le ribosome rencontre un codon qui marque un « stop ». La synthèse protéique s'arrête alors. La protéine est libérée, la méthionine est éliminée, l'ARNt est séparé du reste de la molécule et le ribosome se libère de l'ARNm.

– LE CODE GÉNÉTIQUE

Comme nous avons déjà vu, l'information génétique est conservée par la cellule au niveau de son ADN. Cette information est transcrite en ARNm, puis traduite en protéines, processus dont nous venons de parler.

Le code génétique

Premier nucléotide	Deuxième nucléotide U		Deuxième nucléotide C		Deuxième nucléotide A		Deuxième nucléotide G		Troisième nucléotide
U	UUU	phényl-alanine	UCU	sérine	UAU	tyrosine	UGU	cystéine	U
	UUC		UCC		UAC		UGC		C
	UUA	leucine	UCA		UAA	STOP	UGA	STOP	A
	UUG		UCG		UAG		UGG	tryptophane	G
C	CUU	leucine	CCU	proline	CAU	histidine	CGU	arginine	U
	CUC		CCC		CAC		CGC		C
	CUA		CCA		CAA	glutamine	CGA		A
	CUG		CCG		CAG		CGG		G
A	AUU	isoleucine	ACU	thréonine	AAU	asparagine	AGU	sérine	U
	AUC		ACC		AAC		AGC		C
	AUA		ACA		AAA	lysine	AGA	arginine	A
	AUG	méthionine	ACG		AAG		AGG		G
G	GUU	valine	GCU	alanine	GAU	acide aspartique	GGU	glycine	U
	GUC		GCC		GAC		GGC		C
	GUA		GCA		GAA	acide glutamique	GGA		A
	GUG		GCG		GAG		GGG		G

Figure 11. Le code génétique.

Une étape essentielle dans l'expression de l'information génétique est donc la traduction de l'information sous forme nucléotidique (ARNm) en une forme protéique (acides aminés). Cette traduction est réalisée par triplets de nucléotides : trois nucléotides (« triplet » ou « codon ») codent pour un des vingt acides aminés. Cette correspondance est le code génétique.

3. LES MUTATIONS

Les mutations sont des altérations au hasard du matériel génétique (DNA chez les cellules et DNA ou RNA chez les virus). Généralement, elles sont récessives et elles ne se manifestent pas (parfois dû à l'environnement). Bien qu'elles soient normalement nocives pour l'individu, les mutations sont le moteur de la variabilité et de la diversité. En plus, elles peuvent avoir lieu aux cellules somatiques (mutations somatiques) ou aux cellules germinales (mutations germinales). Ces dernières, sont transcendantales, si l'on parle des individus pluricellulaires, puisque toutes les cellules du nouvel organisme auront la même information que le zygote (première cellule de l'organisme). Comme nous avons vu, les mutations affectent directement l'ADN et, par conséquent, la synthèse des protéines est modifiée.

Il existe de nombreuses façons différentes de modifier l'ADN, ce qui entraîne différents types de mutation :

3.1. LES MUTATIONS GÉNIQUES

Les mutations géniques sont des altérations dans la séquence de nucléotides d'un gène. Elles peuvent être :

- Substitutions : C'est une mutation qui échange une base nucléique (adénine, cytosine, guanine ou thymine) pour une autre. Ce type de mutation n'est pas nocive généralement, mais elle peut provoquer le changement d'un codon (triplet d'aminoacides). Pourtant, si un codon modifié indique "stop", cela peut provoquer la synthèse d'une protéine incomplète. Donc, les conséquents remplacements d'acides aminés, affectent le repliement, la stabilité et l'agrégation des protéines.

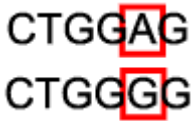
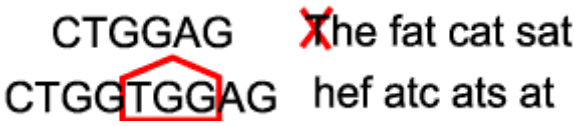


Figure 12. Une mutation par substitution

- Insertions et délétions : Les insertions sont des mutations dans lesquelles des paires de bases sont insérées dans l'ADN. Et, à cause des suppressions, une partie de l'ADN est perdue ou supprimée.

  Puisque l'ADN codant est divisé en codons de trois bases, les insertions et les délétions peuvent modifier un gène de sorte que l'analyse de ces codons sera incorrecte. Ces changements sont appelés des *frameshifts*.



Figures 13 et 14. Une insertion et un frameshift.

3.2. LES MUTATIONS CHROMOSOMIQUES

Les mutations chromosomiques provoquent des altérations dans la séquence des gènes des chromosomes. Il y en a de différents types :

- Délétions : C'est la perte d'un fragment du chromosome. Cela peut être létal.

Figure 15. Une déletion

- Duplications : C'est la répétition d'un fragment d'un chromosome. Les duplications permettent l'augmentation du matériel génétique et, grâce à des mutations postérieures, elles peuvent déterminer l'apparition de nouveaux gènes pendant le processus évolutif.

Figure 16. Une duplication

- Inversions : Les inversions sont des mutations dans lesquelles on trouve un changement de sens d'un fragment de chromosome.

Figure 17. Une inversion

- Translocations : C'est l'altération des positions d'un segment de chromosome.

Figure 18. Une translocation

3.3. LES MUTATIONS GÉNOMIQUES

Les mutations génomiques ce sont des altérations du nombre de chromosomes d'un individu.

4. LA SÉLECTION NATURELLE

Au long de la vie des organismes, leurs génomes (et, donc, leurs mutations) interagissent avec l'environnement. L'environnement d'un génome comprend la biologie moléculaire dans la cellule, d'autres cellules, d'autres individus, les populations, les espèces, ainsi que l'environnement abiotique[4]. Les individus présentant certaines variantes plus favorables dans leur entourage, peuvent survivre et se reproduire plus que les individus avec d'autres variantes moins adéquates. Par conséquent, la population évolue et les caractéristiques favorables (génétiques et donc héréditaires) deviennent plus fréquentes dans une population que dans une autre.
Au fil du temps, ce processus peut éventuellement entraîner une spéciation (création de nouvelles espèces, macroévolution). En d'autres termes, la sélection naturelle est un processus clé dans l'évolution d'une population.

5. LES SÉQUENCES OU DOMAINES CONSERVÉS

Dans la biologie évolutive, les séquences conservées sont des séquences semblables ou pareilles dans les acides nucléiques (ADN et ARN), les protéines ou les polysaccharides (par exemple la cellulose ou le glycogène, qui sont des glucides) à travers les espèces (séquences orthologues[5]) ou dans différentes molécules produites par le même organisme (séquences paralogues[6]).

La conservation à travers les espèces indique qu'une séquence a été maintenue par évolution malgré la spéciation. Une séquence hautement conservée est celle qui est conservée sans changements dans l'arrière-plan de l'arbre phylogénétique (au long du temps) et donc, cela signifie que la sélection naturelle a continuellement éliminé les formes avec des mutations dans cette séquence.

Les séquences sont susceptibles d'être fortement conservées pendant le temps géologique si elles sont nécessaires pour les fonctions cellulaires essentielles (comme

4 L'environnement abiotique comprend tous les facteurs et processus non vivants dans un écosystème. Par exemple, la lumière du soleil, le sol ou l'eau.

5 Les ORTHOLOGUES sont des gènes dans différentes espèces qui ont évolué à partir d'un gène ancestral commun par spéciation. Normalement, ils conservent la même fonction au cours de l'évolution.

6 Les PARALOGUES sont des gènes créés par duplication dans un génome. Ils développent de nouvelles fonctions, même si celles-ci sont liées à l'originelle.

le codage des enzymes vitales), la stabilité, le développement embryonnaire, la reproduction. La similarité des séquences est utilisée comme preuve de la conservation structurelle et fonctionnelle, et des relations évolutives entre les séquences. Par conséquent, les éléments fonctionnels sont fréquemment identifiés en recherchant des séquences conservées dans un génome.
La séquence du promoteur TATA (une région de l'ADN qui contrôle l'initiation de la transcription d'une partie concrète du même ADN) est un exemple d'une séquence d'ADN hautement conservée trouvée dans la plupart des eucaryotes.

6. L'ÉVOLUTION, LA PHYLOGÉNIE ET LA PHYLOGÉNÉTIQUE

Evolution, phylogénie et phylogénétique ont une relation importante et très étroite.

On pourrait définir l'évolution comme le changement de la composition génétique d'une population au cours de générations successives, qui peut être causé par la sélection naturelle, l'endogamie, l'hybridation ou la mutation.

A son tour, la phylogénie fait référence à l'histoire évolutive d'un groupe taxonomique[7] d'organismes. Elle est essentielle dans l'étude scientifique de l'identification, de la classification, de l'écologie et des histoires évolutives des organismes. La phylogénie montre les relations entre les groupes d'organismes (taxons), en particulier les différences et les similitudes entre eux.

La phylogénie est représentée par un diagramme d'arbre appelé arbre phylogénétique. Une branche de science étroitement liée qui utilise des diagrammes d'arbres phylogénétiques pour étudier les histoires évolutives et la relation entre différents groupes d'organismes est la phylogénétique. La relation entre taxons est habituellement démontrée par des données de séquençage moléculaire.
La phylogénie peut être représentée par un arbre phylogénétique enraciné ou pas. Un arbre phylogénétique enraciné implique un ancêtre commun où les taxons étroitement liés sont descendus. Un arbre phylogénétique non-racé, en revanche, ne montre pas un ancêtre commun, mais il émet l'hypothèse sur le degré de parenté évolutionnaire entre les taxons.

7 TAXONOMIE : La science de décrire, classer et nommer des organismes, y compris l'étude des relations entre les taxons et les principes qui soutiennent telle classification.

Par la suite, phylogénétique est l'étude scientifique de la phylogénie. Elle concerne principalement les relations d'un organisme avec d'autres organismes selon des similitudes et des différences évolutives. La phylogénétique est donc une partie de la systématique biologique.

Elle est aussi liée à la taxonomie, qui est une branche de la science qui s'occupe également de classer et nommer des organismes. La phylogénétique fournit de l'information à la taxonomie en matière de classification et d'identification des organismes.

En phylogénétique, les méthodes de séquençage de l'ADN sont utilisées pour analyser les traits héréditaires observables. On utilise également un arbre phylogénétique en diagramme pour montrer les histoires évolutives hypothétiques et les relations de groupes d'organismes basés sur les phylogénies de différentes espèces biologiques. L'arbre phylogénétique a été utilisé pour comprendre la biodiversité, la génétique, les évolutions et l'écologie des organismes.

7. L'ARBRE PHYLOGÉNÉTIQUE

Un arbre phylogénétique est un diagramme qui analyse et représente les relations évolutives entre les organismes ou espèces (taxa). Il montre seulement des hypothèses et pas de réponses définitives. L'hypothèse est basée sur des informations qui sont recueillies par rapport à l'ensemble d'espèces d'intérêt (qui se trouvent au bout des branches de l'arbre), comme par exemple, leurs caractéristiques physiques ou les séquences d'ADN de leurs gènes (ou protéines, dans le cas de mon étude). Le schéma de ramification dans un arbre phylogénétique reflète comment les espèces ou d'autres groupes ont évolué à partir des ancêtres communs. Dans ces arbres, deux espèces sont plus apparentées si elles ont un ancêtre commun plus récent. Par contre, elles sont moins liées si l'ancêtre commun est moins récent. Chaque point de branchement (appelé nœud interne) représente un événement de divergence, ou la séparation d'un seul groupe (qui serait l'ancêtre, le nœud interne) en deux groupes descendants. Un clade est un morceau d'une phylogénie qui comprend une lignée ancestrale et tous les descendants de cet ancêtre. Ce groupe d'organismes a la propriété de monophylie, de sorte qu'il peut également être appelé un groupe monophylétique. Un clade ou un groupe monophylétique est facile à identifier visuellement: c'est simplement un morceau d'un arbre plus grand qui peut être coupé de la racine avec une seule coupe. L'axe horizontal

de l'arbre ne représente pas le temps de façon directe. Donc, nous ne pouvons que comparer le chronométrage des événements de branchement qui se produisent sur la même lignée (même ligne directe depuis la racine de l'arbre).

À continuation, je rajoute l'arbre phylogénétique et taxonomique (général) avec lequel j'ai utilisé pour comparer mes résultats.

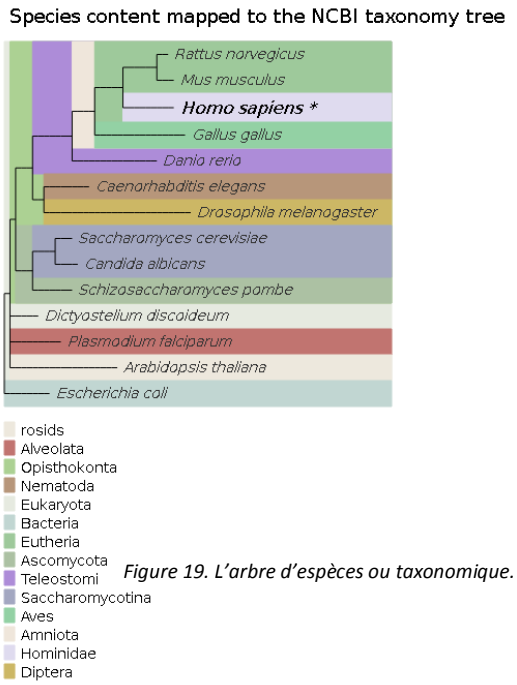


Figure 19. L'arbre d'espèces ou taxonomique.

8. COMMENT CONSTRUIRE ET ANALYSER UN ARBRE PHYLOGÉNÉTIQUE

Pour construire un arbre phylogénétique, les biologistes collectent des données sur les caractères de chaque organisme qui les intéresse. Les caractères sont des traits héréditaires qui peuvent être comparés entre les organismes, comme les caractéristiques physiques (morphologie), les séquences génétiques et les traits de comportement. Un phylome est défini comme la collection des phylogénies reconstruites pour chaque gène codé dans un génome. Ce n'est que récemment, grâce à de nouveaux algorithmes et ordinateurs plus rapides, que l'application de la phylogénétique à des génomes entiers est devenue possible. Donc, les arbres phylogénétiques peuvent être crées automatiquement avec certains sites web qui comparent les rapports évolutifs entre les espèces ou groupes, comme par exemple, UniProt ou Pylogénie.fr dont je me suis servie dans cette étude. Ainsi, les études phylogénétiques à grande échelle apportent des informations de grande valeur sur les relations évolutives entre les gènes de différentes espèces.

Dans mon travail, je me suis centrée dans les séquences des protéines et dans les altérations qu'elles ont subi au long du temps, car ce fait est une conséquence directe des mutations dans l'ADN.

Pour construire des arbres phylogénétiques, les outils qui ont cette fonction suivent une méthodologie peu variable :

- Ils cherchent la similarité entre les espèces ou molécules. L'option de recherche blast sur des collections de génomes nous permet de trouver des espèces qui ont des liens évolutifs entre elles.
- Ils font une reconstruction grâce aux alignements des séquences, qui ordonnent ces séquences afin d'analyser la similareté.
- Ils coupent les informations dans les alignements qui ne sont pas utiles pour la recherche. Par exemple, dans le cas de PhylomeDB, l'outil trimAl est ce qui réalise cette action.
- Les outils pour créer les arbres phylogénétiques fonts certains hypothèse de possibles organisation des espèces. Donc, ils font des estimations des models évolutifs les plus adéquats.
- Finalement, ils obtiennent un arbre phylogénetique fiable, avec une estimation de qualité des phylogénies entre les espèces. Des estimations de support des branches (la fiabilité des branches) sont faites aussi.

Pour lire un arbre phylogénétique, il faut suivre une série de pas :

- Déterminer l'ordre d'apparition des caractères. L'ordre d'apparition des caractères se lit de bas en haut. C'est-à-dire, du nœud les plus ancestral, jusqu'à les espèces de l'actualité. Dans mon cas il serait de gauche à droite. Il fait apparaitre les différentes étapes de l'évolution de manière chronologique. Cependant, cet ordre de description n'est pas obligatoire.
- Définir l'apparentement. Des animaux qui sont rangés dans un même groupe de la classification actuelle partagent des caractères communs.
- .Rechercher l'ancêtre commun. Partir à la recherche de l'ancêtre commun de deux êtres vivants revient à chercher le nœud, le niveau de réunion, des branches qui proviennent de deux êtres vivants. Les caractères de l'ancêtre commun sont ensuite déduits à partir du point de jonction des branches jusqu'au caractère le plus ancien.

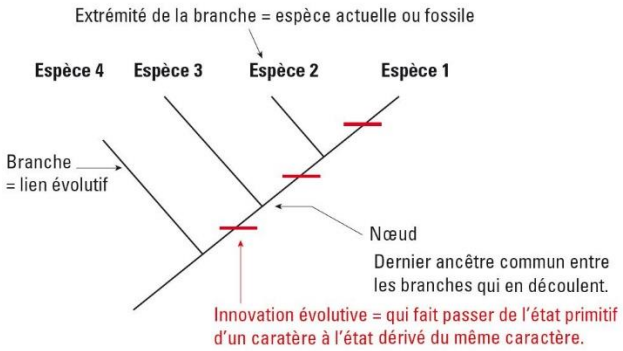


Figure 20. L'arbre phylogénétique et ses parties.

9. LES DIFFÉRENTS TYPES D'ARBRES

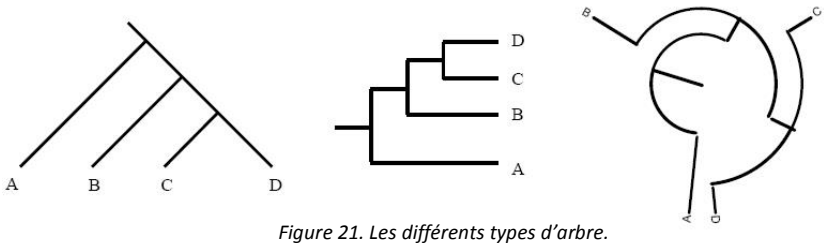


Figure 21. Les différents types d'arbre.

Les diagrammes en arbre peuvent représenter la même information tout en étant orientés de différentes manières. Les trois arbres de la figure 21, par exemple, ont la même topologie et ont donc les mêmes implications évolutives. Ils sont tous des Phylogrammes. Un phylogramme est un arbre phylogénétique dont les longueurs de branches sont proportionnelles au nombre de changements de caractères qui ont été inférés le long des branches. Néanmoins, les deux premiers s'agissent des cladogrammes (organisé par clades) et le dernière est un arbre circulaire.

Dans chaque cas, le premier événement de divergence sépare l'espèce du nœud qui a donné naissance à la branche A de la lignée qui a donné naissance aux pointes B, C et D. Cette dernière lignée s'est ensuite divisée en deux lignées dont l'une s'est transformée en pointe B. L'autre a donné lieu aux pointes C et D. Ce fait signifie que C et D partagent un ancêtre commun plus récent que l'un ou l'autre partage avec A ou B. Il peut sembler déroutant que des arbres si différents puissent contenir la même information. Il est essentiel de se rappeler que les lignes d'un arbre représentent des lignées évolutionnaires et que celles-ci n'ont pas de position ou de forme réelle.

10. LES APPLICATIONS DES ARBRES PHYLOGÉNÉTIQUES

L'utilisation la plus directe de l'arbre phylogénétique est de découvrir l'histoire évolutive des taxons (groupes d'espèces) et comment ils sont liés les uns aux autres. C'est la connaissance la plus basique, mais elle peut conduire à d'autres usages. L'alignement

des séquences multiples, la conservation de la structure et la détection des motifs sont quelques-unes des applications directes des arbres phylogénétiques. Ainsi la détection d'homologie, la divergence, la convergence, les attributs paralogues des séquences sont visualisables à travers les arbres.

Par exemple, en biologie de la conservation (et en écologie en général), nous pouvons mesurer la diversité phylogénétique (les différences évolutives entre les espèces) en utilisant des arbres phylogénétiques. Les branches sur l'arbre de l'évolution ne sont pas dessinées à des longueurs aléatoires, elles sont calculées à cette longueur. La diversité phylogénétique prend ces longueurs pour nous donner une idée de la diversité de la faune dans un écosystème. Par exemple, des branches qui sont longues indiquent que l'on a des taxons de base qui sont plus importants à conserver parce qu'ils sont apparus plus tôt et ont donc plus d'informations sur l'évolution à donner. D'un autre côté, des branches plus courtes nous indiquent que ces espèces sont très nouvelles, peut-être à cause d'un processus de divergence comme les mutations ou, dans mon cas, un événement de duplication ou spéciation d'un gène.

Une autre utilisation est dans la recherche de produits naturels. Supposons que l'on a une espèce d'éponge qui produit un produit chimique qui fonctionne très bien pour réduire l'inflammation. Mais les drogues produites à partir de cette espèce ont tendance à avoir des effets secondaires nocifs pour l'organisme. On fait une étude systématique des éponges et nous découvrons les espèces qui s'y ressemblent le plus, et on teste leurs produits chimiques (par ordre de parenté). Parfois, on peut n'obtenir un meilleur et ce fait implique que la recherche d'un produit qui ne soit pas nocif sera plus facile à réaliser.

En médecine, on peut utiliser des arbres phylogénétiques pour trouver des médicaments qui soient effectives dans certaines espèces qui se rassemblent. Si on trouve des espèces très proches a l'être humain, par exemple, on peut étudier leur génome pour trouver des drogues qui ont de l'effectivité sur nous.

En outre, on peut utiliser des arbres phylogénétiques pour guider notre recherche de nouvelles espèces. Si on le trace sur une carte, on découvrirait comment une espèce s'est étendue géographiquement dans son évolution. On peut ensuite aller à ces endroits pour chercher des populations qui ont été isolées et ont formé de nouvelles espèces. On peut aussi utiliser des arbres phylogénétiques pour nous dire quand les taxons sont apparus et où, qui nous aidera à trouver leurs correspondants fossiles.

11. LA RECHERCHE

11.1. LE PROCESSUS ET LE BUT

Comme nous avons déjà vu dans l'introduction du travail, tout d'abord, j'ai fait de la recherche sur Phylome DB et j'ai lu les différentes guides afin de connaître comment fonctionne cet outil. Dans un premier moment, je ne comprenais pas la démarche, et j'ai décidé de demander de l'aide à une créatrice du projet *Puja a l'arbre.* Dans projet en ligne, les étudiants qui veulent faire un travail de recherche sur la phylogénétique peuvent s'inscrire pour mieux connaître la page PhylomeDB.

Donc, après avoir parlé avec Marina Marcet-Houben (l'une des organisatrices), j'ai choisi cinq protéines au hasard. Puis, je les ai cherchées sur le site, j'ai choisi le phylome qui contenait la protéine, que je voulais comparer et le site a créé un arbre automatique sur ces protéines. Ensuite, je les ai analysées et j'ai comparé ces arbres phylogénétiques avec ceux déjà faits qui sont utilisés comme modèle pour comprendre l'évolution des espèces. Finalement, je suis arrivée à certaines conclusions.

Ensuite, j'ai cherché la protéine P08842 sur Phylome DB et j'ai mis sa séquence protéique dans Uniprot et puis j'ai créé mon arbre grâce à Phylogeny.fr.

11.2. LES OUTILS UTILISÉS

PHYLOME DB

PhylomeDB est une base de données publique des collections complètes de phylogénies génétiques (phylomes). Sur cet outil nous pouvons explorer de façon interactive l'histoire évolutive des gènes à travers d'arbres phylogénétiques et d'alignements de séquences. En bioinformatique, un alignement de séquence est un moyen d'arranger les séquences d'ADN, d'ARN ou de protéines pour identifier les régions de similarité qui peuvent être une conséquence des relations fonctionnelles, structurelles ou évolutives entre les séquences. Des espaces sont insérés de sorte que des caractères identiques ou similaires sont alignés dans des colonnes successives.

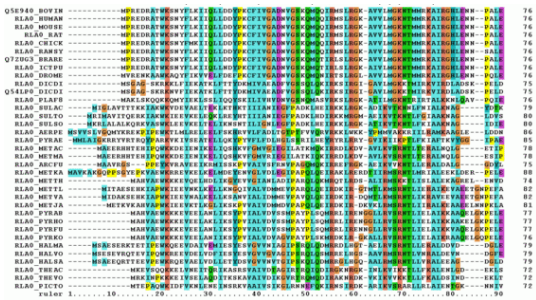


Figure 22. Un alignement de séquences.

Cet outil utilisé pour reconstruire des arbres sert à faire d'analyses phylogénétiques de différents génomes, d'alignements et à faire la création ou l'analyse du modèle évolutif. En outre, PhylomeDB inclut une section de téléchargement d'arbres, d'alignements et des prédictions d'orthologie. Pour faire ces prédictions, le site utilise *métaPhOrs* (*meta-Phylogeny based Orthologs*). A son tour, métaPhOrs combine les ressources de plusieurs bases de données (PhylomeDB, EnsemblCompara, EggNOG, OrthoMCL, COG, Fungal Orthogroups, et TreeFam) pour tester fiabilité des prédictions.

UNIPROT

Uniprot c'est la source universelle de protéines qui nous fournit avec leurs séquences et informations. Si nous cherchons une protéine qui nous intéresse, nous trouverons tout ce que nous savons sur celle-ci. Uniprot contient l'option de Blast (*Basic Local Alignment Search Tool*) dont je me suis servie pour créer mon arbre phylogénétique. Cette option nous permet de faire des recherches sur des séquences qui présentent des similarités.

Figure 23. Le site web UniProt.

PHYLOGENY.FR

Phylogeny.fr est un site web, relativement simple à utiliser, dédié à la reconstruction et à l'analyse des relations phylogénétiques entre séquences moléculaires. Il exécute et relie différents programmes bioinformatiques pour reconstruire un arbre phylogénétique robuste à partir d'un ensemble de séquences. Donc, c'est avec cette page que j'ai construit mon arbre phylogénétique.

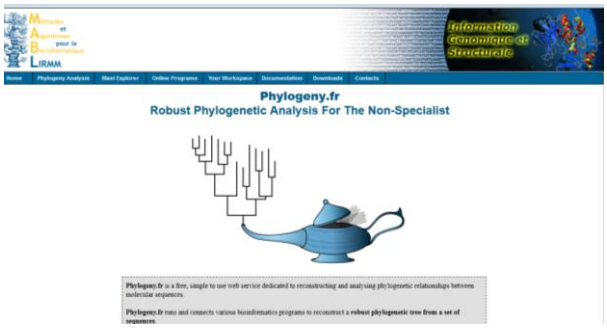


Figure 24. Le site web Phylogeny.fr.

ITOL

Interactive Tree Of Life est un outil en ligne pour l'affichage, l'annotation et la gestion des arbres phylogénétiques. Nous pouvons y explorer nos arbres directement dans le navigateur et les annoter avec différents types de données. En outre, il est possible de rajouter directement nos données dans l'arbre et nous pouvons modifier sa visualisation. Nous pouvons ajuster les couleurs des branches et des étiquettes, les styles et les fonts de manière interactive.

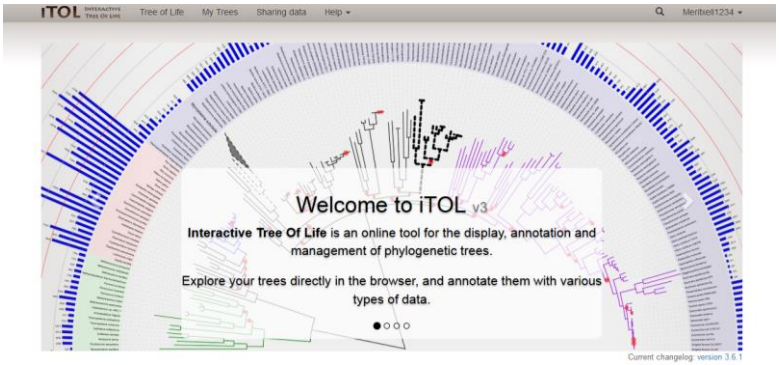


Figure 25. Le site web iTol.

11.3. L'ANALYSE

11.3.1. LA CONSTRUCTION DE MON ARBRE PHYLOGÉNÉTIQUE

L'un de mes objectifs pour la recherche était la construction d'un arbre phylogénétique à partir d'une protéine choisie. J'ai choisi la protéine P08842 ou Steryl-sulfatase (STS), qui se trouve dans la membrane du réticulum endoplasmique (un organite propre aux cellules eucaryotes), ainsi que dans le gène STS des humains (l'ADN qui la codifie se trouve dans ce gène). Les mutations de ce gène sont associées à l'ichtyose liée au chromosome[8] X (XLI). L'ichytose est une maladie de la peau causée par la déficience héréditaire de l'enzyme stéroïde sulfatase (STS) qui affecte de 1 sur 2000 à 1 en 6000 des hommes. Selon UniProt, d'où j'ai extrait sa séquence et son information, la protéine

8 Chromosome : L'un des plusieurs corps filiformes, constitués de chromatine, qui portent les gènes dans un ordre linéaire.

a été révisée manuellement par Swiss-Prot. Swiss-Prot est d'une base de données de séquences de protéines, qui unit des résultats expérimentaux, des caractéristiques calculées et des conclusions scientifiques. Ensuite, j'ai découvert que la protéine P08842 a la fonction hydrolase (enzyme qui catalyse la réaction d'hydrolyse[9]), plus concrètement, de conversion des précurseurs[10] de stéroïdes sulfatés en œstrogènes pendant la grossesse et que son activité catalytique (cette protéine augmente la rapidité d'une

réaction) est : 3-bêta-hydroxyandrost-5-en-17-one 3-sulfate + H2O = 3-bêta-hydroxyandrost-5-en-17-one + sulfate. En outre, il s'agit d'un enzyme dont le cofacteur[11] est le calcium, donc elle peut s'unir a cet élément. Elle participe dans le développement de l'épiderme et la grossesse féminine, entre d'autres.

Sa séquence d'aminoacides est : MPLRKMKIPFLLLFFLWEAESHAASRPNIILVMADD LGIGDPGCYGNKTIRTPNIDRLASGGVKLTQHLAASPLCTPSRAAFMTGRYPVRSGMA SWSRTGVFLFTASSGGLPTDEITFAKLLKDQGYSTALIGKWHLGMSCHSKTDFCHHPL HHGFNYFYGISLTNLRDCKPGEGSVFTTGFKRLVFLPLQIVGVTLLTLAALNCLGLLHVP LGVFFSLLFLAALILTLFLGFLHYFRPLNCFMMRNYEIIQQPMSYDNLTQRLTVEAAQFIQ RNTETPFLLVLSYLHVHTALFSSKDFAGKSQHGVYGDAVEEMDWSVGQILNLLDELRL ANDTLIYFTSDQGAHVEEVSSKGEIHGGSNGIYKGGKANNWEGGIRVPGILRWPRVIQ AGQKIDEPTSNMDIFPTVAKLAGAPLPEDRIIDGRDLMPLLEGKSQRSDHEFLFHYCNA YLNAVRWHPQNSTSIWKAFFFTPNFNPVGSNGCFATHVCFCFGSYVTHHDPPLLFDIS KDPRERNPLTPASEPRFYEILKVMQEAADRHTQTLPEVPDQFSWNNFLWKPWLQLCC PSTGLSCQCDREKQDKRLSR (583 aminoacides).

Ces lettres correspondent aux aminoacides suivants :

9 Hydrolyse : La rupture des liaisons entre molécules à cause de l'eau.

10 Précurseur : Substance nécessaire pour en produire une autre au moyen d'une réaction chimique. Ce sont des composés chimiques qui constituent une première étape dans un processus chimique et agissent comme substrat (molécule sur laquelle agit une enzyme) dans les étapes postérieures.

11 Cofacteur : Un cofacteur est toute substance non protéique requise pour qu'une protéine soit catalytiquement active.

Code d'une seule lettre	Code de trois lettres	Aminoacide
A	Ala	Alanine
R	Arg	Arginine
N	Asn	Asparagine
D	Asp	Acide aspartique
C	Cys	Cystéine
Q	Gln	Glutamine
E	Glu	Acide glutamique
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
L	Leu	Leucine
K	Lys	Lysine
M	Met	Méthionine
F	Phe	Phénylalanine
P	Pro	Proline
S	Ser	Serine
T	Thr	Thréonine
W	Trp	Tryptophane
Y	Tyr	Tyrosine
V	Val	Valine

Sa structure secondaire est :

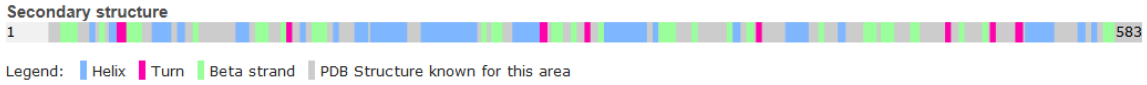


Figure 26. La structure secondaire de la protéine P08842.

Donc, cette protéine dans sa structure secondaire on trouve des hélices (en bleu), des coudes (en rose) et des feuillets beta (en vert).

Par la suite, j'ai pris cette séquence et je l'ai copié-collé dans l'option de BLAST dans UniProt. BLAST (*Basic Local Alignment Search Tool*) trouve des régions de similarité locale entre les séquences, qui peuvent être utilisées pour déduire les relations fonctionnelles et évolutives entre les séquences (des autres espèces) ainsi que pour aider à identifier les membres des familles de gènes. Puis, j'ai eu comme résultat 250 protéines d'espèces différentes, qui ont des séquences très semblables à la protéine P08842. Donc, elles auront des fonctions très semblables aussi. J'en ai choisi douze (celles qui avaient un pourcentage d'identité entre le 99.5% et le 92.6%, donc elles étaient plus apparentées) qui appartiennent aux espèces : *Callithrix jacchus* (l'ouistiti), *Nomascus leucogenys* (le gibbon), *Pongo abelii* (l'orang-outan), *Pan troglodytes* (le chimpanzé), *Homo sapiens* (l'humain), *Papio anubis* (le babouin), *Chlorocebus sabaeus* (singe vert), *Macaca mulatta et Macaca fascicularis* (des macaques). L'identité ou similarité des protéines est exprimé en forme de pourcentage, auquel on lui attribue des

couleurs. Toutes mes protéines sont en couleur rouge. Donc elles sont toutes très semblables

et l'échelle d'identité.

Contrairement à la protéine P08842, celles-ci n'ont pas été révisées manuellement. Donc, elles ont été analysées moyennant TrEMBL qui fait un traitement automatique des données. J'ai cliqué sur l'option d'aligner et le programme a comparé les séquences de ces protéines grâce à un algorithme. Les résultats que nous obtenons, ce sont les alignements des séquences complètes de ces protéines que nous sommes en train de comparer. Nous pouvons déduire qu'elles sont vraiment très semblables grâce aux astérisques au-dessous de la dernière séquence, qui nous montrent que dans cet endroit les douze protéines ont le même aminoacide.

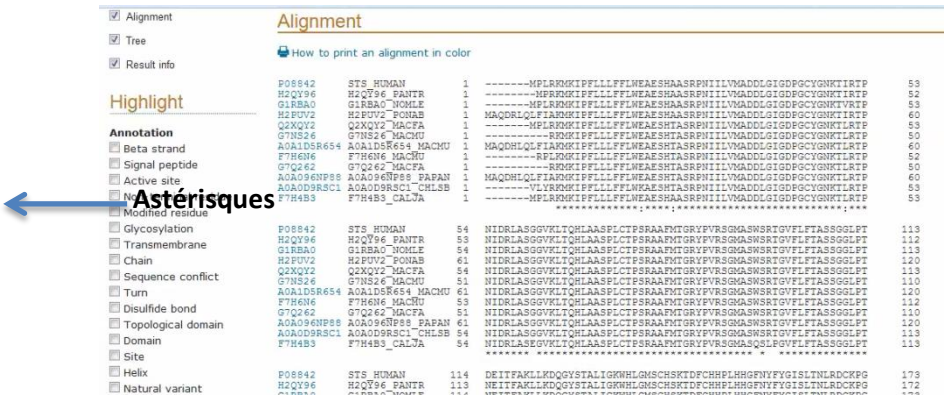


Figure 28. Les alignements des séquences des protéines que j'ai choisies.

Après les alignements, j'ai aussi obtenu un arbre :

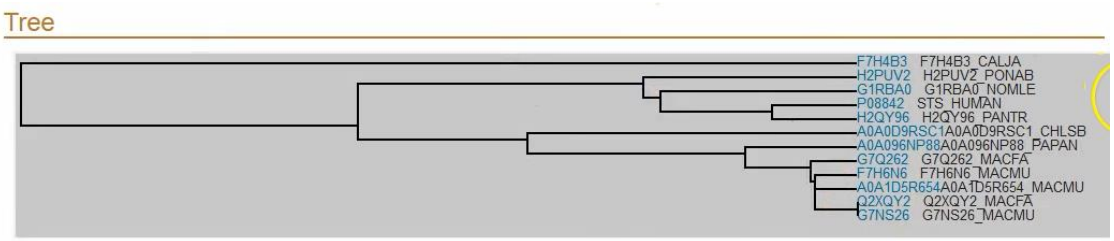


Figure 29. L'arbre phylogénétique de mes protéines, crée par UniProt.

Et à la fin, toutes les séquences des protéines choisies sont se montrées avec le nom et l'espèce devant. Je les ai toutes copiées et je les ai collées dans le site web Phylogeny.fr dans l'option One Click. J'ai effacé l'information qui n'était pas nécessaire et j'ai seulement laissé le nom de la protéine et l'espèce à laquelle elle appartenait. Puis, j'ai cliqué sur Submit et mon arbre phylogénétique a apparu. Néanmoins, celui-ci ne va pas être le définitif, car je désire de changer les couleurs des branches et des autres options de visualisation qu'on peut modifier grâce au site web iTol.

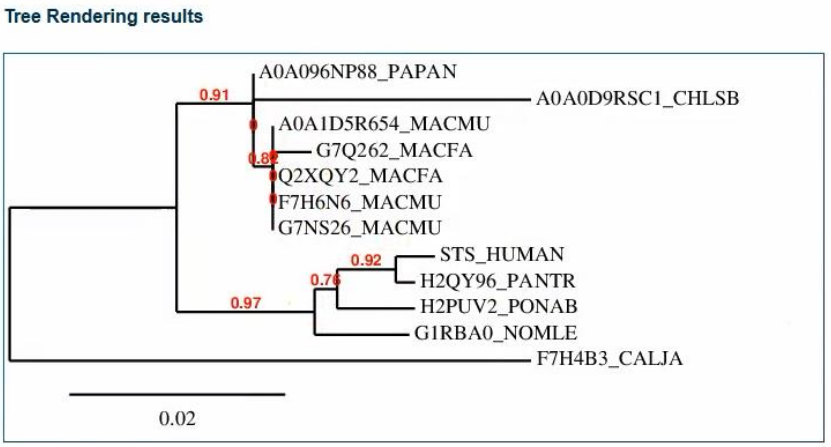


Figure 30. L'arbre phylogénétique crée par Phylogeny.fr.

Ensuite, au-dessous de cet arbre, nous avons l'option de le télécharger en différents formats. J'ai choisi le format Newick (comme la spécialiste de *Puja a l'arbre* m'avait dit). Cette option de télécharger en ce format, nous permettra d'obtenir l'arbre phylogénétique sur iTol. Les données obtenues ont été :

```
(F7H4B3_CALJA:0.04834,((G1RBA0_NOMLE:0.00877,(H2PUV2_PONAB:0.00722,
(H2QY96_PANTR:0.00179,STS_HUMAN:0.0036)0.92:0.00545)0.76:0.00208)0.97:0.01279,
((((G7NS26_MACMU:0,(F7H6N6_MACMU:0,(Q2XQY2_MACFA:0,G7Q262_MACFA:0.00359)0:0)0:0)
0:0,A0A1D5R654_MACMU:0)0.82:0.0018,A0A0D9RSC1_CHLSB:0.02574)0:0,
A0A096NP88_PAPAN:0)0.91:0.0071):0.01549);
```

Figure 30. L'arbre phylogénétique dans le format Newick.

Je les ai copiées et puis collées sur la page web iTol, qui nous servira pour créer notre arbre.

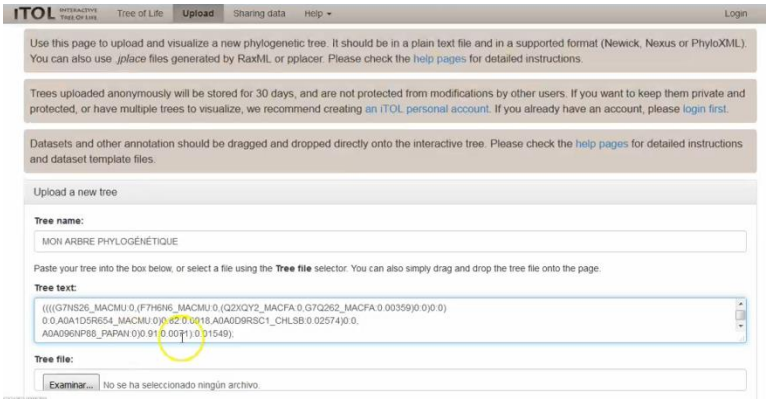


Figure 31. Le site web iTol, où j'ai collé mes données.

Ensuite, j'ai cliqué sur Upload et j'ai obtenu mon arbre, que j'ai modifié (la couleur des branches et la proximité des noms des espèces à celles-ci) avec les options de visualisation. A continuation, j'ai ajouté mon arbre phylogénétique et au-dessous, j'ai additionné des photos de chaque espèce qui y apparaît.

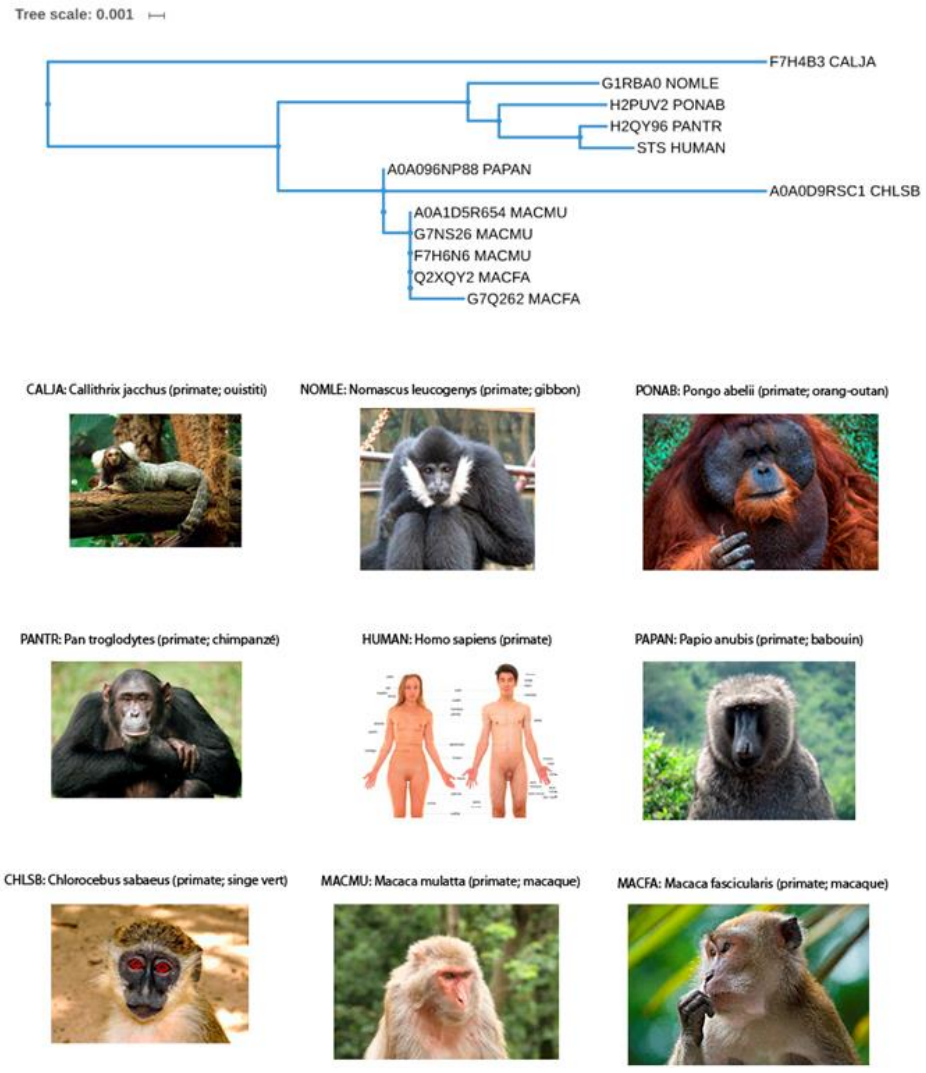


Figure 32. Mon arbre phylogénétique, avec des photographies des espèces.

Par la suite, pour faire l'analyse de cet arbre phylogénétique et pour déterminer sa fiabilité, j'ai pris comme modèle cet arbre taxonomique :

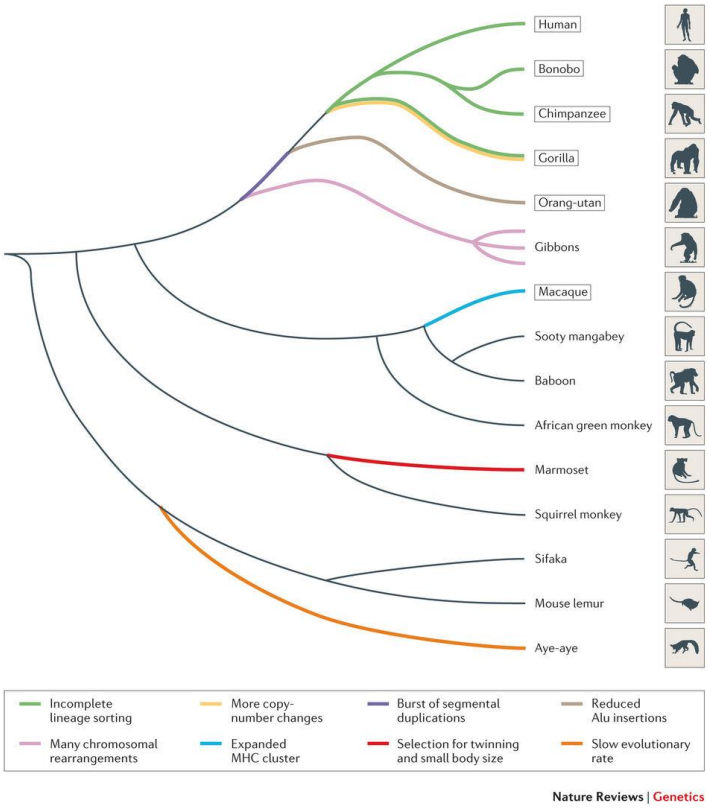


Figure 33. L'arbre phylogénétique d'espèces général.

Finalement, dans mon arbre on trouve un nœud initial qui se divise en deux branches. L'une contient *Callithrix jacchus* (l'ouistiti ; *marmoset* en anglais) et l'autre a souffert une spéciation, qui donne comme résultat deux branches avec de différentes espèces. D'un côté, nous avons *Pan troglodytes* (le chimpanzé) avec *Homo sapiens* (l'humain), et puis, *Pongo abelii* (orang-outan). Ensuite, nous avons *Nomascus leucogenys* (le gibbon). Jusqu'à ici, tous coïncident avec la distribution de l'arbre taxonomique pris de PhylomeDB. De l'autre côté, nous trouvons *Papio anubis* (le babouin), *Chlorocebus sabaeus* (le singe vert) et une branche qui contient, à son tour, *Macaca mulatta* (le macaque) et *Macaca fascicularis* (le macaque) plusieurs fois. Cela peut être dû à une duplication d'une protéine ancestrale de leur génome (le gène qui codifiait cette protéine s'est répliqué) qui a souffert des modifications au fil du temps et a produit des protéines différentes à cause de l'accumulation des mutations.

Pour conclure, tous les résultats obtenus coïncident avec l'arbre pris comme référence du site web Nature (l'une des revues scientifiques les plus prestigieuses au monde). Donc, on peut affirmer que l'arbre phylogénétique est fiable et il est bien fait.

11.3.2. L'ANALYSE DES ARBRES PHYLOGÉNÉTIQUES

Le deuxième but que je m'étais proposé c'était l'analyse de certains arbres phylogénétiques « automatiques » crées par le site web PhylomeDB. Donc, la méthodologie que j'ai suivie a été la suivante : tout d'abord, sur PhylomeDB j'ai fait le choix des protéines au hasard moyennant l'option de « random search » de l'outil de recherche. Puis, selon la taille (l'arbre n'est ni trop large ni trop simple), les espèces avec lesquelles cette protéine a des liens évolutifs et la fiabilité entre celles-ci, j'ai choisi mes arbres. J'ai choisi le phylome que j'ai désiré, qui a été en fait les organismes modèles et ensuite, j'ai obtenu mes arbres.

LES ORGANISMES MODÈLES

Les organismes modèles sont des espèces qui ont été largement étudiées parce qu'elles sont faciles à entretenir, à nourrir et à reproduire dans un laboratoire. Ils sont utilisés dans le laboratoire pour aider les scientifiques à comprendre les processus biologiques. Ils sont utiles dans la recherche en génétique parce qu'ils peuvent se reproduire en grand nombre, ils ont un temps de génération très court (le temps entre la naissance et la reproduction). Les mutants (des organismes modèles qui ont subi un changement ou une mutation dans leur ADN) permettent aux scientifiques d'étudier certaines caractéristiques ou maladies. En plus, certains organismes modèles ont des gènes semblables ou des génomes de taille similaire aux humains.
Les organismes modèles peuvent être utilisés pour créer des cartes génétiques très détaillées. Les cartes génétiques sont une représentation visuelle de la localisation des différents gènes d'un chromosome et les zones d'intérêt dans le génome. Certains exemples d'organismes modèles sont :

- *Mus musculus* (la suris domestique)
- *Rattus norvegicus* (rat brun)
- *Gallus gallus* (la poule)
- *Danio rerio* (le poisson-zèbre)

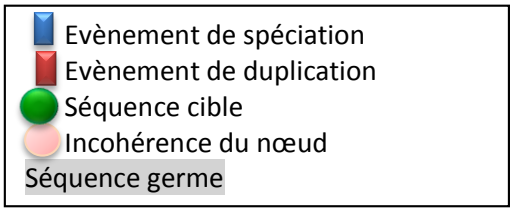
Dans mon expérience, on trouve:

- *Caenorhabditis elegans (un petit ver)*
- *Drosophila melanogaster* (la mouche du vinaigre)
- *Saccharomyces cerevisiae* (une levure)
- *Candida albicans* (une levure)

- *Schizosaccharomyces pombe* (une levure)
- *Plasmodium falciparum* (un plasmodium)
- *Arabidopsis thaliana* (une plante)
- *Escherichia coli* (une bactérie intestinale)

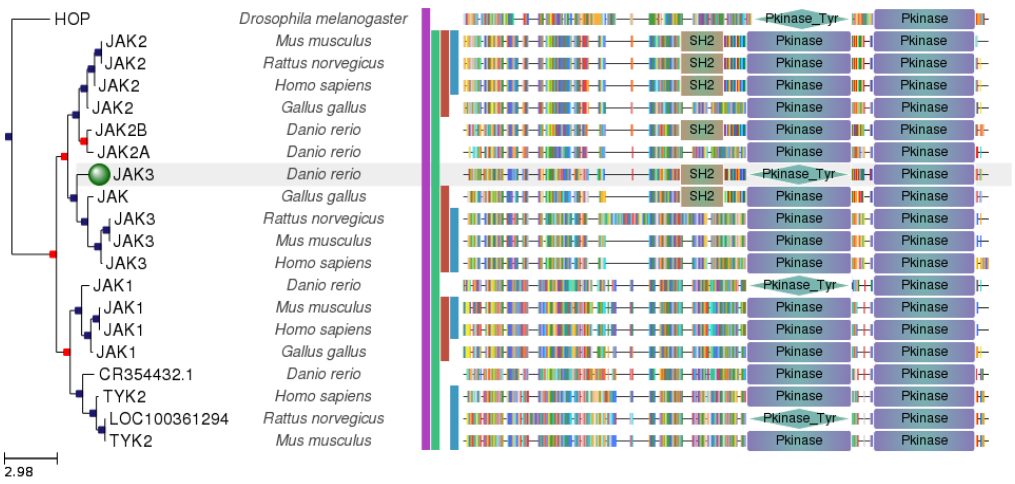
Dans certains arbres, j'ai changé la racine ou ancêtre commun afin d'obtenir des résultats logiques et adéquats, on le voit dans les analyses que j'ai réalisées grâce aux informations que le site PhylomeDB et Pfam (pour les domaines protéiques qui se trouvent à droite des arbres). Pfam est une base de données avec une grande collection de familles de protéines. On y trouve des informations sur des protéines individuelles ou elle génère des groupes d'entrées (protéines) liées, appelées clans. Un clan est une collection d'entrées Pfam qui sont liées par similarité de séquence. Les protéines sont généralement composées d'une ou plusieurs régions fonctionnelles, communément appelées domaines. Différentes combinaisons de domaines donnent comme résultat la grande variété de protéines trouvées dans la nature. L'identification des domaines qui se produisent dans les protéines peut donc donner un aperçu de leur fonction.

LÉGENDE



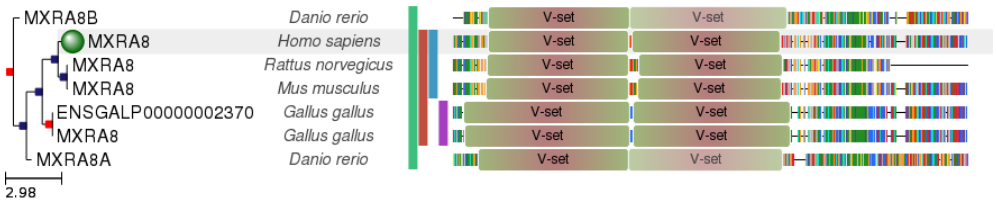
L'arbre se trouve enraciné dans la feuille (en ce cas, une protéine) la plus ancienne.

ARBRE 1



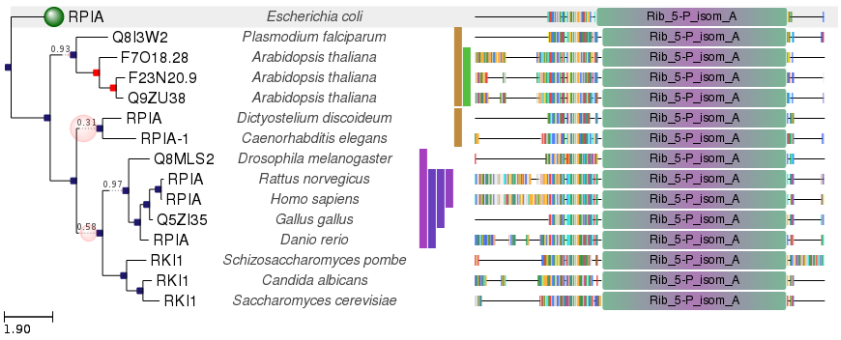
Les protéines de cet arbre sont très bien conservées parmi les espèces au fil du temps. On peut alors en déduire qu'elles ont une fonction identique ou très semblable entre elles (si nous cliquons sur chaque nœud nous trouvons que la distance entre branches est très petite, car elles sont très similaires et elles ont accumulé peu de mutations) et que les protéines jouent un rôle très important pour la vie des organismes. Sinon elles auraient accumulé des mutations. Il y a eu un processus de duplication après la spéciation dans la protéine ancestrale (racine), c'est-à-dire, le gène codant pour la protéine dans le génome est devenu doublé et a provoqué deux copies résultantes qui ont pris différents chemins d'évolution et qui ont conservé des fonctions similaires. La spéciation est, en biologie, le processus évolutif par lequel de nouvelles espèces vivantes apparaissent. Dans ce cas, on fait référence à la création des nouvelles protéines. En comparant ces arbres phylogénétiques avec ceux généraux des organismes modèles (en nous centrant sur la protéine cible) nous trouvons que *Mus musculus* et *Rattus norvegicus* vont correctement de pair, ainsi comme dans le cas d'*Homo sapiens*. Dans les autres branches nous voyons le même: *Mus musculus* et *Rattus norvegicus* sont vont de pair, au-dessus se situe l'Humain, ensuite il y a *Gallus gallus* et, en haut, on trouve *Danio rerio* qui contient la protéine cible que j'ai cherché (JAK3). La protéine JAK3 appartient à la famille des tyrosines kinases qui participent à la croissance cellulaire et font partie du système immunitaire. À son tour, *Drosophila melanogaster* est dans une position évolutive antérieure que tous ceux-ci, aussi comme prévu. Dans l'espèce *Danio rerio* nous trouvons qu'une protéine paralogue à que nous avions choisi s'est doublée. De la racine de l'arbre, on aperçoit une spéciation, où le chemin de la *Drosophila melanogaster* est bifurqué du reste. Ensuite, il y a une duplication relativement ancienne qui, à son tour, est suivie d'une duplication pour chaque branche aussi. Cela donne comme résultat des différentes protéines homologues qui prennent différents chemins évolutifs basées sur la même fonction (qui deviendra une autre). En outre, si nous cliquons sur chaque nœud des branches, nous trouvons que le support est 1. C'est donc un arbre hypothétique très fiable. Dans le panneau des domaines (selon la page Pfam) et des séquences nous perçons que certaines protéines ont perdu le domaine SH2, qui sert à fixer les protéines aux résidus de tyrosine phosphorylés des autres protéines. Certaines possèdent également un domaine de tyrosine kinase (Pkinase_Tyr), qui est une sous-classe de Pkinase (protéine kinase) que contiennent toutes les séquences de l'arbre. La tyrosine kinase agit en transférant un groupe phosphate de l'ATP (cofacteur adénosine triphosphate) à une protéine dans une cellule. Elle fonctionne comme un interrupteur "on" ou "off" dans de nombreuses fonctions cellulaires.

ARBRE 2



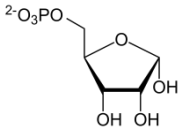
Dans cet arbre, comme dans le précédent, nous pouvons observer que les protéines de différentes espèces sont également très bien conservées. Encore une fois, les branches de *Ratus norvegicus* et de *Mus musculus* quittent le même nœud et au-dessus ils ont *Homo sapiens (qui contient la protéine MXRA8, laquelle j'ai cherché)*, comme prévu. Ce dernier, à son tour, sort d'un nœud de spéciation avec *Gallus gallus*, qui a subi une duplication d'une protéine, et une qui devrait être au-dessus de lui*. Danio rerio* se trouve au-dessus de toutes les autres espèces. En ce qui concerne le domaine des séquences, ils ont tous deux domaines V-set qui sont des domaines similaires aux immunoglobulines qui ressemblent au domaine variable de l'anticorps. Il y a une espèce, *Rattus norvegicus*, qui a perdu une partie de la séquence puisqu'il y a un trou (ligne droite à la fin). Les branches ont une longueur comprise entre 0,00 et 0,4, par conséquent, les protéines sont assez liées les unes aux autres. Et le support est de 1.00, ce qui signifie que l'arbre est très fiable. On croit que la protéine MXRA8 participe dans la maturation et le maintien de la barrière hémato-encéphalique (une barrière physiologique présente dans le cerveau chez tous les vertébrés terrestres, entre la circulation sanguine et le système nerveux central).

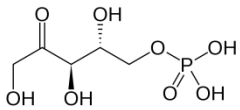
ARBRE 3



Dans cet arbre nous pouvons observer qu'Homo sapiens est mis avec *Rattus norvegicus* qui, à son tour, étaient sortis du même nœud que *Gallus gallus*. Au-dessus nous avons *Danio rerio*. Donc, jusqu'à ici, tout est comme prévu. Si on continue à rétrocéder dans l'évolution, nous trouvons que dans le nœud il y a un cercle rouge. Cela veut dire que

cette branche-ci n'est pas trop fiable. Sous le nœud que je viens de décrire, nous avons *Candida albicans* avec *Saccharomyces cerevisiae*, et elles sortent d'une spéciation conjointement avec *Schizosaccharomyces pombe.* Ceci est correct, néanmoins la position ne l'est pas. Ces espèces devraient être au-dessus (évolutivement) de la branche qui contient *Drosophila melanogaster*, selon l'arbre taxonomique de NCBI (dans le point 7 du sommaire). Nous pourrons déduire que ce fait est dû à la basse fiabilité du nœud d'où ces espèces partent. Puis, en haut, nous avons les espèces *Plasmodium falciparum* et *Arabidopsis thaliana* correctement situées. Cette dernière espèce est la seule qui a souffert une duplication du gène qui codifie la protéine, qui donne comme résultat trois protéines vraiment semblables entre elles. Et finalement, *Escherichia coli* est au-dessus, car je l'ai mise comme origine ou racine de cet arbre phylogénétique afin que les résultats soient plus exacts. Cette espèce contient la protéine cible RPIA qui catalyse la conversion réversible du ribose-5-phosphate en ribulose 5-phosphate.



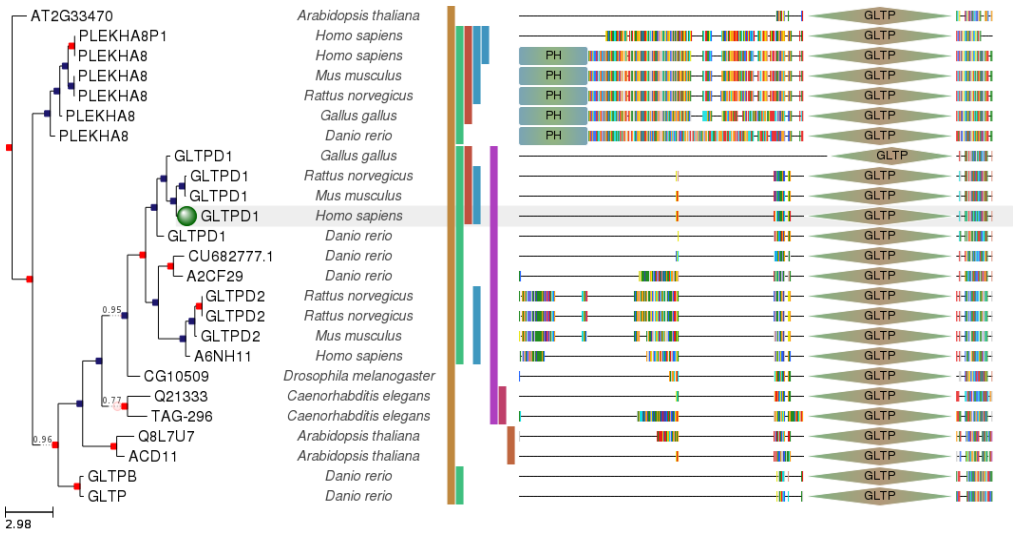


Figures 34 et 35. Une ribose-5-phosphate et une ribulose 5-phosphate

Dans le panneau des domaines et séquences, nous trouvons que dans la première partie des séquences de toutes les protéines, la conservation des aminoacides est variable. Puis, presque toutes ont gardé la séquence et enfin, toutes ont le domaine « ribose 5-phosphate isomerase A » qui est donc, bien conservé. Il joue un rôle essentiel dans le métabolisme des glucides, car il catalyse la conversion entre le ribose-5-phosphate et le ribulose-5-phosphate dans la route métabolique[11] du pentose-phosphate.

11 Route métabolique : une séquence de réactions chimiques où un substrat initial est transformé et donne naissance à des produits finaux.

ARBRE 4

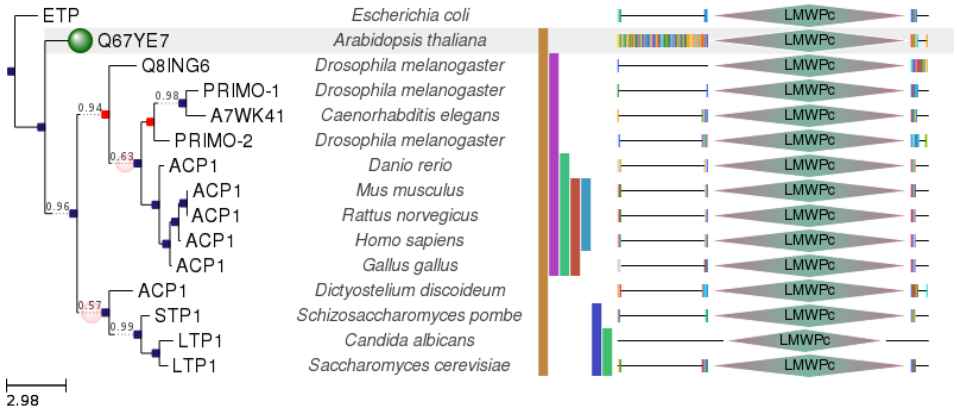


Nous pouvons voir dans cet arbre phylogénétique que la protéine choisie appartient à l'espèce humaine, qui est au même niveau que *Mus musculus*. Puis, nous trouvons *Rattus norvegicus, Gallus gallus* et *Danio rerio*. Donc, jusqu'ici tout est comme prévu. Suivant le nœud qui vient à continuation, qui montre une duplication de la protéine ancestrale, nous voyons que *Rattus norvegicus* a la protéine GLTPD2 doublée dans son génome. Puis, il se rejoint avec *Mus musculus*, et Homo sapiens, qui est adéquat à nouveau. Plus en arrière dans le temps, nous localisons *Drosphila melanogaster* et ensuite, *Caenorhabditis elegans*, dont les protéines Q21333 et TAG-296 proviennent d'une duplication d'un gène dans son génome. Puis, nous trouvons *Arabidopsis thaliana,* qui est bien placé. Cependant, au même niveau on trouve Danio rerio qui a eu une duplication, qui a eu comme résultat les protéines GLTPB et GLTP. Cette espèce devrait être égale à *Gallus gallus* ou *Homo sapiens.* Donc, cette localisation est très bizarre. Ce fait peut avoir de la rélation avec la longueur de cette branche, qui est de 1.011 (trop haute et très peu fiable). De l'autre côté de l'arbre (en haut), nous trouvons une duplication dans le génome *d'Homo sapiens* qui génère les protéines PLEKHA8P1 et PLEKHA8. *Mus musculus* et *Rattus norvegicus* sont ensemble, et puis on a *Gallus gallus* et *Danio rerio*, qui sont tous corrects. Dans une branche à l'écart, nous avons *Arabidopsis thaliana,* qui est plus ancienne que toutes les autres espèces. Par conséquent, la distribution de cet arbre est appropriée, sauf pour la localisation de *Danio rerio*.

Dans les branches en haut, nous voyons que les séquences d'aminoacides son très conservées. Donc, ces protéines auront un rôle essentiel pour ces espèces. Par ailleurs, toutes elles contiennent le domaine PH qui sont liés avec le recrutement de protéines à différentes membranes, les ciblant ainsi à des compartiments cellulaires appropriés ou leur permettant d'interagir avec d'autres composants des voies de transduction du signal,

selon Pfam. Finalement, toutes les espèces de l'arbre ont le domaine GLTP qui est chargé du transport des différents glycosphingolipides et glycéroglycolipides entre les membranes intracellulaires. La protéine GLTPD1 intervient dans le transfert intracellulaire du céramide-1-phosphate entre les membranes des organites et la membrane cellulaire.

ARBRE 5



La protéine cible appartient à l'espèce *Arabidopsis thaliana*, qui est la seule qui conserve la séquence initiale d'aminoacides. De son même nœud, nous trouvons une autre branche qui a souffert une spéciation. Donc, celle-ci se bifurque en deux branches. En bas, nous avons *Candida albicans* avec *Saccharomyces cerevisiae*, puis *Schizosaccharomyces pombe* et ensuite, *Dictyostelium discoideum*. Ces résultats, selon l'arbre taxonomique, sont corrects. Néanmoins, leur position dans l'arbre n'est pas très commune, puisqu'ils devraient se localiser dans des branches antérieures aux espèces *Mus musculus, Danio rerio*, etc.

Dans la branche en haut, qui sort de la troisième spéciation, nous trouvons d'un côté *Mus musculus* avec *Rattus norvegicus*, puis Homo sapiens et enfin *Gallus gallus*. Ces résultats sont adéquats aussi et, en plus, leurs protéines sont identiques (ACP1). Ensuite, dans la branche qui sort du nœud par spéciation au-dessus de celle que nous venons de réviser, nous voyons *Drosophila melanogaster* pair avec *Caenorhabditis elegans*, et puis nous trouvons une autre fois *Drosophila melanogaster,* qui est un peu bizarre. Cependant, ses protéines sont orthologues : PRIMO-1 et PRIMO-2.

Quant aux domaines, toutes les protéines ont conservé LMWPc (Protein tyrosine phosphatase), qui a comme fonction éliminer les groupes phosphate des résidus tyrosine phosphorylés sur les protéines.

11.4. LES CONCLUSIONS

Pour conclure, grâce à cette étude j'ai pu créer mon propre arbre phylogénétique à partir de la protéine P08842 qui se trouve dans différentes espèces de primates. J'ai aussi analysé certains arbres qui ont été générés par le site PhylomeDB et qui contenaient des organismes modèles. Enfin, j'ai obtenu des résultats qui se correspondent généralement à ceux que j'attendais.

En premier lieu, en ce qui concerne ma première hypothèse, dans le cas de l'arbre que j'ai créé personnellement, toutes les relations évolutives étaient bien faites. Conséquemment, on pourrait réfuter la première hypothèse. Néanmoins, le fait de refuser cette déduction a très peu de crédibilité, puisque pour arriver à une conclusion déterminante et définitive il faudrait réaliser un large nombre de répliques afin de montrer que ces résultats ne sont pas à cause du hasard. D'ailleurs, mon but était de montrer comment créer un arbre phylogénétique et pas de confirmer ou de réfuter l'hypothétique arbre phylogénétique général des primates. Cette correspondance entre l'arbre phylogénétique concret et le général correspond aux primates et nous montre qu'il est bien fait.

Dans la deuxième partie de mon expérience, les résultats que j'ai obtenus ont aussi été en correspondance avec les arbres généraux. J'ai fait cinq répliques et elles pourraient confirmer la deuxième hypothèse du travail. On pourrait affirmer que finalement les résultats obtenus dans mes arbres phylogénétiques sont très semblables aux arbres d'espèces généraux et que l'évolution et la phylogénétique moléculaire sont très étroitement liées. Ainsi, les hypothétiques diagrammes évolutifs généraux correspondent aux cas concrets et donc, dans le domaine des organismes modèles les résultats attendus (selon les arbres phylogénétiques généraux) sont les résultats que j'ai obtenus dans mes arbres. Néanmoins, on peut réfuter la première hypothèse puisque dans mes arbres, les gènes, qui codifient les protéines, suivent l'évolution des espèces. Donc, les arbres généraux correspondent à mes arbres particuliers de protéines.

12. LA BIBLIOGRAPHIE

Prensa Científica. Del genoma al filoma [en línia]

Muntaner 339, pral 1a. - 08021 Barcelona – España

[Consultat 30 de juny 2017]

< http://www.investigacionyciencia.es/revistas/investigacion-y-ciencia/la-verdad-del-desnudo-502/del-genoma-al-filoma-1429>

The Chemistry of Biology: Proteins [en línia]

Sandbox Networks, Inc. 2000–2017

[Consultat 5 de juliol 2017]

<https://www.infoplease.com/science/biology/chemistry-biology-proteins>

Wikipedia. Nucleotide [en línia]

Wikipedia®

[Consultat 7 juliol 2017]

<https://en.wikipedia.org/wiki/Nucleotide>

Wikipedia. Natural selection [en línia]

Wikipedia®

[Consultat 10 juliol 2017]

<https://en.wikipedia.org/wiki/Natural_selection>

Your Genome (2017). Copyright information. [en línia]

[Consultat 19 juliol 2017].

< https://www.yourgenome.org/facts/what-types-of-mutation-are-there >

BIO-WEB 2.0. Synthèse des protéines [en línia]

[Consultat 15 juliol 2017]

<http://www.jpboseret.eu/biologie/index.php/cellules/33-synthese-des-proteines>

PhylomeDB. [en línia]

Comparative Genomics Group at CRG (Barcelona, Spain)

[Consultat 30 juny-30 ocutbre 2017]

<http://www.phylomedb.org/>

PhylomeDB: A database for genome-wide collections of gene phylogenies. [en línia]

2008-2017 ResearchGate GmbH

[Consultat 20 juliol 2017]

<https://www.researchgate.net/publication/5884178_PhylomeDB_A_database_for_genome-wide_collections_of_gene_phylogenies>

Khan Academy. Building a phylogenetic tree [en línia]

[Consultat 20 juliol 2017]

<https://www.khanacademy.org/science/biology/her/tree-of-life/a/building-an-evolutionary-tree>

Understanding evolution. Reading trees: A quick review. [en línia]

[Consultat 7 octubre 2017]

<https://evolution.berkeley.edu/evolibrary/home.php>

13. LES ANNEXES

Indépendamment du travail écrit, j'ai réalisé une vidéo montrant le processus de création de mon arbre phylogénétique laquelle je vais montrer à l'exposition orale.

13.1. LES REMERCIEMENTS

Tout d'abord, je voudrais remercier Marina Marcet-Houben, une des chercheuses du projet Puja a l'Arbre, qui travaille dans le PRBB à Barcelona et qui m'a aidé vraiment à comprendre le fonctionnement de PhylomeDB et les outres outils que j'ai utilisé pour faire ce travail. Sans elle, je n'aurais pas pu finir mon étude.

En outre, je voudrais remercier ****, ma professeure du Travail de Recherche. Elle m'a aidé beaucoup aussi et elle m'a guidé dans le processus de création du travail. Même si elle ne comprenait pas les thèmes, elle a essayé de le faire afin de me conseiller. Il faut remarquer qu'elle a été très stricte et rigoureuse, afin que je fisse le travail le mieux que possible.

Finalement, je voudrais remercier mes parents, ma famille et mes amis qui ont été un support très important dans la période d'obtention de mon travail.

13.2. LES DIFFICULTÉS RETROUVÉES

Ainsi, je veux exprimer les difficultés que j'ai eues au début à cause du niveau de technicité du sujet et de toute la théorie à apprendre avant de commencer à travailler. Par ailleurs, j'ai eu du mal à utiliser les sites web de la partie pratique. Avant de faire aucune analyse, il fallait savoir clairement la méthodologie de travail. Par conséquent, j'ai perdu beaucoup de temps à lire des guides pour chaque outil, avant de contacter Marina.

En outre, faire des hypothèses adéquates a été un peu compliqué et c'est pour cela que j'ai dû les refaire deux fois. Néanmoins, je suis satisfaite avec les définitives.

13.3. L'INTERVIEW

Comme Marina Marcet-Houben m'a aidé beaucoup et elle travaile la bioinformatique et les arbres phylogéniques, j'ai décidé de lui faire une petite interview.

Salut Marina. Tout d'abord, je vous remercie pour toute l'aide que vous m'avez apportée dans le processus de création de mon travail de recherche. Puisqu'il s'agit de phylogénétique et que vous êtes une personne bien comprise dans ce sujet, j'ai décidé de vous faire une petite interview pour savoir davantage sur votre métier et sur la phylogénétique en question.

- Quel est votre métier?

Je suis chercheuse postdoctoral au Centre de réglementation génomique (CRG) à Barcelone et je suis spécialisée en génomique et en bioinformatique. Ce domaine de recherche consiste à utiliser des outils informatiques pour essayer d'étudier les aspects biologiques à travers des génomes complètement séquencés. Dans ce champ, j'ai travaillé sur le séquençage et l'assemblage des espèces Penicillium, qui sont des espèces de champignons d'un grand intérêt, non pas seulement du point de vue industriel (ils sont utilisés pour produire des antibiotiques tels que la pénicilline) mais aussi économique, puisque certains de ces organismes sont pathogènes des plantes et causent des pertes importantes dans le secteur agricole.
En outre, j'utilise des génomes entièrement séquencés pour étudier l'évolution des espèces. Je m'intéresse particulièrement à comment les espèces hybrides évoluent puisqu'elles ne sont pas des espèces fertiles, mais pour survivre, elles doivent récupérer leur capacité de reproduction. Nous avons suffisamment des preuves que les hybrides anciens ont survécu au fil du temps, donc c'est possible.

- Qu'est-ce que c'est exactement le CRG et qu'est-ce que l'on y fait?

Le CRG est un centre de recherche où la recherche de base est réalisée. La recherche de base est celle qui fournit des connaissances mais qui ne s'avère pas souvent pas sur des applications pratiques immédiates. Plusieurs fois, des collaborations avec d'autres centres de recherche ou entreprises sont nécessaires pour parvenir à une application utile pour la société. Cependant, cette recherche de base est essentielle pour pouvoir trouver des applications qui améliorent notre qualité de vie. Le CRG est principalement axé sur la biomédecine. Les différents groupes qui dont y font partie étudient de nombreux aspects différents tels que le cancer, les maladies mentales, le syndrome de Down et j'en passe et des meilleures. Dans notre groupe en particulier, nous travaillons avec des champignons pathogènes, ceux qui vivent à l'intérieur de nous et que lorsque notre système immunitaire est détérioré ils peuvent causer des maladies. Même la mort.

- Avez-vous toujours su que vous aimeriez ce métier et travailler dans le CRG? Qu'est-ce qui vous a amené à opter pour ce métier?

Non, mon cas a été un peu dû à la casualité. J'ai étudié un master en biochimie parce que l'on y étudiait la génétique et j'avais une grande curiosité pour savoir ce qui nous rendait différents les uns des autres. Pendant les études, je me suis rendue compte de que je n'aimais pas le travail de laboratoire, ce qui était un problème puisque j'aimais la partie théorique. J'ai eu de la chance car tandis que j'étais en train d'étudier le master, on a commencé à investir dans le domaine de la bioinformatique. Après avoir eu quelques matières sur la bioinformatique, j'ai décidé de faire ma thèse dans un groupe bioinformatique qui étudiait l'évolution des bactéries. Principalement c'était pour vérifier si la bioinformatique me plaisait. Le résultat est que des années plus tard, je continue à travailler dans le même domaine, bien que j'aie changé de bactéries par des champignons.

- Avez-vous des travails de recherche publiés?

Oui, pendant ma carrière j'ai publié 46 travails. Certains d'entre eux sont des collaborations avec d'autres groupes de recherche nationaux et internationaux. Par exemple, j'ai participé au séquençage du lynx ibérique et de l'olivier. D'autres sont des travaux réalisés principalement pour moi et pour mon superviseur.

- Vous êtes l'un des organisateurs du projet Puja a l'Arbre. Qu'est-ce que l'on fait dans ce projet?

Le projet Puja a l'Arbre était un projet pilote visant à amener l'évolution aux lycées. Bien que l'évolution de l'espèce soit majoritairement acceptée, sa présence dans les salles de classe peut encore être améliorée. Dans ce domaine, notre groupe a proposé d'offrir de l'aide pour créer des travaux de recherche en utilisant notre site web: PhylomeDB. PhylomeDB recueille près de 6.5 millions d'arbres phylogénétiques qui montrent

l'évolution de différents gènes dans différents groupes d'espèces. L'idée était que les étudiants, aidés par leurs professeurs et par un membre de notre groupe, essaieraient d'incorporer l'étude de l'évolution de certains gènes dans leur travail de recherche.

- Comment définiriez-vous un arbre phylogénétique et quelles applications a-t-il? Pensez-vous qu'ils sont vraiment utiles?

Un arbre phylogénétique est un outil que nous utilisons pour représenter l'évolution d'un groupe d'espèces ou d'une famille de gènes. Comme tout outil, il a ses avantages et des inconvénients. Un des avantages est qu'il nous permet de visualiser quelles séquences ou espèces sont plus proches et donc moins séparées d'un ancêtre commun. La partie la plus informative dans notre cas est quand un arbre construit pour représenter l'évolution d'un gène ne suit pas le même modèle que celui observé dans un arbre d'espèces. Ce fait-ci indique que quelque chose s'est passé et il nous permet d'analyser quels événements évolutionnaires ont pu causer ce changement dans l'arbre. Les arbres phylogénétiques, au-delà de la représentation des événements évolutifs, peuvent être utilisés pour quelque chose de très important en biologie: l'annotation fonctionnelle des protéines. La façon de trouver ce qu'une protéine fait dans un organisme est de faire des expériences. Mais cela est très coûteux et lent, et actuellement, c'est seulement fait systématiquement dans un groupe d'espèces modèles comme la levure ou la souris. Mais nous avons beaucoup plus de génomes et nous voulons savoir ce qu'ils font. Une des façons de prédire ce qu'un gène fait est de réaliser un arbre phylogénétique et de voir si l'un des autres gènes qui apparaissent dans l'arbre a une fonction connue et ensuite transférer cette fonction dans notre gène cible. Ceci, bien que n'étant pas aussi fiable qu'une expérience, nous permet d'économiser beaucoup d'argent et nous donne une base de travail.

- Comment a été votre expérience en aidant les étudiants?

Aider les étudiants est généralement très gratifiant, ils me font penser à des projets qui sont à sa portée et beaucoup des questions qu'ils me posent sont parfois des choses que je n'avais jamais considérées. En outre, des bioinformatiques sont nécessaires, maintenant que la bioinformatique existe, je pense qu'il est favorable de faire connaître certaines choses qui sont possibles de faire et de motiver les intéressés intéressées à se dédier à ce métier.

- Trouvez-vous facile d'utiliser l'outil PhylomeDB? Les étudiants qui participent au projet Puja a l'Arbre ont généralement des difficultés avec cette site web?

Le principal problème avec phylomeDB est que nous ne l'utilisons pas. Principalement le web a été créé comme un lieu pour laisser toutes les données que nous avons construit, afin que les autres puissent en avoir accès. Cependant, nous travaillons directement avec les données qui constituent la base de données, surtout parce que nous travaillons habituellement avec des milliers d'arbres et nous ne les pouvons pas regarder

individuellement. Nous savons que phylomeDB a un problème principal, qui est la recherche de séquences d'intérêt. En général, tout le domaine de la biologie pose un gros problème lorsqu'il s'agit de nommer des choses. Il n'y a pas de consensus commun. Le même gène peut être appelé de différentes façons, ce qui complique la recherche. Dans PhylomeDB nous essayons d'inclure autant d'identifiants que possible mais c'est toujours difficile. C'est à cause de cela que nous avons l'alternative de chercher per blast (séquence), qui est pour l'instant la meilleure alternative disponible. Le site web n'est pas si compliqué à naviguer une fois que quelqu'un nous raconte comment il fonctionne, je ne sais pas si quelqu'un a lu la guide qui est sur le site, donc je ne sais pas si c'est suffisant ou pas. A mon avis, les étudiants ont eu plus de problèmes dans l'interprétation des arbres, ce qui n'est pas strictement un problème de PhylomeDB. C'est dû à un manque de connaissances générales sur la phylogénie. Je crois que avec notre aide, ils ont suffisamment appris pour interpréter correctement ce qu'ils voyaient et en tirer des conclusions correctes.

- Le secteur phylogénétique (en Espagne et dans le monde entier) est-il suffisamment développé? Y a-t-il eu des découvertes importantes ces dernières années?

Comme dans beaucoup de domaines de la bioinformatique, la phylogénétique est en constante évolution. Les arbres phylogénétiques sont des prédictions, ils prennent des alignements ou d'autres données pour pouvoir établir les relations entre les séquences, mais ils peuvent avoir des erreurs ou l'information peut être insuffisante. Par ailleurs, certains arbres prennent beaucoup de temps pour être reconstruits, et l'un des objectifs actuels est d'essayer de créer des programmes plus rapides sans perdre la qualité. En ce qui concerne les découvertes, il y a toute une série de recherches dédiés à l'étude de comment les espèces ont évolué entre elles. Il y a encore beaucoup de questions et de nombreux endroits de l'arbre de l'espèce qui ne sont pas résolus. Il y a beaucoup de chercheurs qui consacrent des efforts pour construire des arbres de toutes les espèces connues parce que nous sommes intéressés à savoir comment ils ont évolué. Je pense que dans cette ligne, les travaux les plus intéressants se font chez les protistes (des êtres vivants qui sont constitués par une seule cellule eucaryote, comme par exemple les protozoaires), puisque nous en savons très peu sur ces organismes ancestraux.

- Bon, merci Marina ! J'ai aimé faire ce travail de recherche sur la phylogénétique. Sans votre aide, il m'aurait pris beaucoup plus de temps et surement je n'aurais pas compris certaines concepts ou comment analyser les arbres. J'espère que vous avez eu une bonne expérience avec moi.

De rien Meritxell, j'aime beaucoup ce métier et t'aider a été un processus très agréable. Je désire que tu aies une très bonne note pour ton travail !

13.3. LA SOURCE DES PHOTOS ET DES ARBRES ANALYSÉS

Photographie de la couverture : Arbre numéro 1 :
http://phylomedb.org/?q=search_tree&seqid=Phy0036ZQQ&phyid=507

Figure 1. https://en.wikipedia.org/wiki/DNA#/media/File:DNA_chemical_structure.svg

Figure 2. https://en.wikipedia.org/wiki/Nucleotide#/media/File:Nucleotides_1.svg

Figure 3. https://4.bp.blogspot.com/-bfSL8BBvm6k/Vy4Q-1TLoII/AAAAAAAAAX0/yNokGtt_4N8nn7MLGiL03htIykF_bnptwCKgB/s1600/cellule%2Beucaryote.jpg

Figure 4. https://jeretiens.net/wp-content/uploads/2017/03/cellule_procaryote_apprendre_astuce.jpg

Figure 5. https://upload.wikimedia.org/wikipedia/commons/thumb/8/86/Argonne%27s_Midwest_Center_for_Structural_Genomics_deposits_1%2C000th_protein_structure.jpg/320px-Argonne%27s_Midwest_Center_for_Structural_Genomics_deposits_1%2C000th_protein_structure.jpg

Figure 6. https://i.pinimg.com/736x/66/ef/ee/66efee5ce8d11d8abcf4963192545188--cours-svt-la-nutrition.jpg

Figure 7. https://upload.wikimedia.org/wikipedia/commons/thumb/9/96/AlphaHelixProtein_fr.jpg/250px-AlphaHelixProtein_fr.jpg

Figure 8. http://www.bio-top.net/Schemas/Proteine_feuillet_beta.gif

Figure 9. https://upload.wikimedia.org/wikipedia/commons/thumb/9/9c/Beta_turn.svg/916px-Beta_turn.svg.png

Figure 10. https://upload.wikimedia.org/wikipedia/commons/thumb/5/56/Sch%C3%A9ma_biologie_mol%C3%A9culaire.png/1200px-Sch%C3%A9ma_biologie_mol%C3%A9culaire.png

Figure 11. http://raymond.rodriguez1.free.fr/Documents/Cellule-genome/codeG.jpg

Figure 12. http://evolution.berkeley.edu/evolibrary/article/mutations_03

Figure 13. http://evolution.berkeley.edu/evolibrary/article/mutations_03

Figure 14. http://evolution.berkeley.edu/evolibrary/article/mutations_03

Figure 15. https://socratic.org/questions/what-are-four-types-of-chromosomal-mutations

Figure 16. https://socratic.org/questions/what-are-four-types-of-chromosomal-mutations

Figure 17. https://socratic.org/questions/what-are-four-types-of-chromosomal-mutations

Figure 18. https://socratic.org/questions/what-are-four-types-of-chromosomal-mutations

Figure 19. http://www.phylomedb.org/phylome_500

Figure 20. http://sites.crdp-aquitaine.fr/stl/files/2013/12/Arbre-phylogenique.jpg

Figure 21. https://upload.wikimedia.org/wikipedia/commons/thumb/7/79/RPLP0_90_ClustalW_aln.gif/575px-RPLP0_90_ClustalW_aln.gif

Figure 22. https://upload.wikimedia.org/wikipedia/commons/thumb/7/79/RPLP0_90_ClustalW_aln.gif/575px-RPLP0_90_ClustalW_aln.gif

Figure 23. http://www.uniprot.org/

Figure 24. http://www.phylogeny.fr/

Figure 25. https://itol.embl.de/

Figure 26. http://www.uniprot.org/uniprot/P08842

Figure 27. http://www.uniprot.org/blast/uniprot/B2017110448CF0A2DF181CEB7EC2BC48F8F5F3B0598642F7

Figure 28. http://www.uniprot.org/blast/uniprot/B20171104A7434721E10EE6586998A056CCD0537E2171F86

Figure 29. http://www.uniprot.org/blast/uniprot/B20171104A7434721E10EE6586998A056CCD0537E2171F86

Figure 30. http://www.uniprot.org/blast/uniprot/B20171104A7434721E10EE6586998A056CCD0537E2171F86

Figure 31. http://www.phylogeny.fr/simple_phylogeny.cgi

Figure 32. https://itol.embl.de/tree/80102124531426015092724 42

Figure 33. http://www.nature.com/nrg/journal/v15/n5/fig_tab/nrg3707_F1.html

Figure 34. https://upload.wikimedia.org/wikipedia/commons/thumb/b/bc/Alpha-D-Ribose-5-phosphat.svg/1200px-Alpha-D-Ribose-5-phosphat.svg.png

Figure 35. https://upload.wikimedia.org/wikipedia/commons/thumb/9/94/Ribulose_5-phosphate.svg/1200px-Ribulose_5-phosphate.svg.png

Arbre 1. http://phylomedb.org/?q=search_tree&seqid=Phy0036ZQQ&phyid=507

Arbre 2. http://phylomedb.org/?q=search_tree&seqid=Phy0007XBH_HUMAN&phyid=500

Arbre 3. http://phylomedb.org/?q=search_tree&seqid=Phy0035MPA&seedid=0035MPA&phyid=505&method=LG&tree_features=lineage,best_name,support,motifs,spname&snodes

Arbre 4. http://phylomedb.org/?q=search_tree&seqid=Phy0007XBD&phyid=500

Arbre 5. http://phylomedb.org/?q=search_tree&seqid=Phy00016YX&phyid=502