Jérôme Fortin – 536 996 920
Nicolas Campeau – 536 781 691
Juliette Pourrain – 536 983 589
Daniel Ilboudo – 536 773 884
Nesrine Imloul- 111 290 248

Devoir #2 – Phylogénétique

Rapport présenté à
Nicolas Derome

Dans le cadre du cours
Statistiques génétiques : concepts et analyse
BIF-4002

4 avril 2025

Introduction

L'étude des relations phylogénétiques entre espèces constitue un pilier fondamental de la biologie évolutive. Grâce aux avancées des méthodes moléculaires, il est désormais possible d'élaborer des arbres phylogénétiques fiables à partir de séquences d'ADN ou de leurs traductions en acides aminés (*AA*), offrant ainsi une vision détaillée des liens de parenté entre organismes. Parmi les marqueurs génétiques les plus utilisés, le gène mitochondrial *cytochrome c oxydase I* (*COI*) occupe une place de choix dans les études de *DNA barcoding*, une méthode qui permet d'identifier les espèces à partir d'une courte séquence génétique standardisée, en raison de ses caractéristiques évolutives et fonctionnelles [1].

Le gène *COI*, qui code pour une sous-unité essentielle de la cytochrome c oxydase dans la chaîne respiratoire mitochondriale, présente plusieurs avantages pour l'analyse phylogénétique : il est strictement hérité de manière maternelle, ne subit pas de recombinaison et présente un taux d'évolution modéré permettant de distinguer aussi bien les relations récentes que plus anciennes [1]. En tant que gène codant, ses séquences peuvent être traduites en acides aminés, ce qui permet d'enrichir les analyses en contournant certains biais associés aux substitutions synonymes.

Notre étude porte sur les Cétacés, un groupe emblématique de mammifères marins, et leurs plus proches parents terrestres, les Hippopotamidés, avec lesquels ils forment un clade monophylétique bien documenté [2]. Malgré leur divergence écologique marquée, ces deux groupes partagent une ascendance commune au sein des Artiodactyles. Le gène *COI* est ici un marqueur particulièrement pertinent, car il permet d'explorer ces relations évolutives à différentes échelles.

La robustesse des analyses phylogénétiques dépend fortement du choix du modèle d'évolution moléculaire. Les gènes codants, comme *COI*, présentent une hétérogénéité du taux de substitution selon la position du codon, la troisième position étant généralement plus variable que les deux premières [3, 4]. Il est donc essentiel d'appliquer des modèles qui tiennent compte de cette structure et de recourir à des approches comme la correction gamma ou les modèles à sites invariants.

Un autre aspect déterminant dans l'inférence phylogénétique est le choix des groupes externes, qui permet de raciner correctement l'arbre et d'orienter l'interprétation des relations évolutives. Wiens et Tiu ont montré que l'ajout de groupes externes appropriés renforce la stabilité topologique des arbres [5]. Dans cette étude, nous avons comparé les Cétacés à divers

groupes externes, notamment les Pinnipèdes (phoques et otaries), les Siréniens (lamantin et dugong), les Éléphants (d'Afrique et d'Asie), les Hippopotames (commun et nain), ainsi que deux espèces de reptiles, la Tortue luth et la Cistude. L'ajout de ces groupes a permis d'évaluer l'impact du choix du groupe externe sur la robustesse des hypothèses phylogénétiques.

Enfin, cette analyse s'inscrit dans une démarche comparative entre alignements nucléotidiques et protéiques, dans différents cadres de lecture, afin d'évaluer l'impact de ces paramètres sur la topologie des arbres. Cette approche intégrative permettra de mieux comprendre l'évolution des Cétacés et d'évaluer les performances des méthodes utilisées à la lumière des connaissances actuelles en phylogénie moléculaire.

Matériel et méthodes

L'ensemble de données initial consiste en 32 séquences nucléotidiques codant pour le gène mitochondrial *COI* au format FASTA. Ce jeu de données peut être divisé en trois sous-groupes, soit les cétacés (25 séquences), l'hippopotame (groupe frère des cétacés, 1 séquence) et les pinnipèdes (groupe externe, 6 séquences). Chaque séquence a été étiquetée par son identifiant GenInfo (GI) afin de pouvoir l'identifier dans la base de données *NCBI*. Toutes les analyses ont été exécutées dans l'environnement de programmation R [6].

L'ensemble des séquences ont d'abord été importées dans l'environnement R et alignées à l'aide du paquetage *DECIPHER* [7]. Deux alignements ont été performés, soit un avec les pénalités offertes par défaut (–18 et –16 pour l'ouverture des *indels*, –2 et –1 pour leur prolongements) et un avec des pénalités nulles. Deux alignements d'acides aminés traduits ont ensuite été effectués à partir des séquences nucléotidiques, soit un avec le cadre de lecture par défaut et un avec le cadre de lecture 1, le tout en utilisant les pénalités par défaut.

Ensuite, quatre matrices de distances ont été produites à partir de l'alignement nucléotidique grâce au paquetage *ape* [8]. Ces matrices ont été construites selon les modèles évolutifs suivants, où α correspond au taux de transitions, β au taux de transversions et p à la distance observée.

Jukes et Cantor (*JC69*) [9] :

$$d_{xy} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p\right)$$

Kimura 2 paramètres 1980 (*K80*) [10] :

$$d_{xy} = -\frac{1}{2}\ln(1 - 2P - Q) + \frac{1}{4}\ln(1 - 2Q)$$

$$\text{où } Q = \left(\frac{1}{2}\right)\left(1 - e^{-8\beta t}\right) \text{ et } P = \frac{1}{4}\left(1 - 2e^{-4(\alpha+\beta)t} + e^{-8\beta t}\right)$$

Tamura et Nei 1993 (*TN93*) [11] :

$$d_{xy} = -B\ln\left(1 - \frac{p}{b}\right)$$

$$\text{où } B = 1 - \sum q_i^2 \text{ et } q_i = (\pi A^2 + \pi T^2 + \pi G^2 + \pi C^2)$$

Galtier et Gouy 1995 (*GG95*) [12] :

$$d_{xy} = d_1 rT + d_2\left(1 - e^{-\frac{(\alpha+1)rT}{2}}\right)$$

$$\text{où } d_1 = 1 + \alpha[\theta_1(1 - \theta_1) + \theta_2(1 - \theta_2)]$$

$$\text{et } d_2 = \left[\frac{\alpha}{(\alpha+1)}\right][(\theta_0 - \theta_1)(1 - 2\theta_1) + (\theta_0 - \theta_2)(1 - 2\theta_2)]$$

Une analyse *bootstrap* avec 1000 itérations a ensuite été exécutée pour chacun des modèles via la méthode *Neighbor-Joining* dans le but d'identifier le modèle maximisant la robustesse. Le paquetage *phangorn* [13, 14] a également été utilisé pour rechercher le meilleur modèle via la fonction *modelTest()* selon les critères d'information d'Akaike (AIC) et Bayésien (BIC). Enfin, l'arbre correspondant a été construit à partir des paramètres γ et *I*, qui ont été extraits en fonction des valeurs des variables *k*, *shape* et *inv*.

Trois matrices de distance ont ensuite été produites pour l'alignement d'acides aminés selon les modèles *LG*, *JTT* et *Blosum62* [15, 16, 17] avec un seuil de distance observée de 0,5. Ces matrices ont servi à construire de nouveaux arbres, toujours via la méthode *Neighbor-Joining* et avec 1000 itérations de *bootstrap*, qui ont été comparés aux arbres produits à partir des alignements nucléotidiques. Pour valider ces résultats, des alignements nucléotidiques ont été effectués une nouvelle fois pour chaque cadre de lecture possible. Un arbre phylogénétique (*Neighbor-Joining*, 1000 itérations de *bootstrap*) a finalement été produit pour le meilleur modèle de chacun des cadres de lecture tels qu'identifiés par la fonction *modelTest()* selon la liste de modèles suivante : *WAG, JTT, LG, Dayhoff, cpREV, mtmam, mtArt, MtZoa, mtREV24, VT, RtREV, HIVw, HIVb, FLU, Blosum62, Dayhoff_DCMut* et *JTT_DCMut*.

Enfin, d'autres groupes externes (lamantin + dugong, éléphant d'Afrique + d'Asie, hippopotame + hippopotame nain et tortue luth + cistude) (**Annexe 1 – Tableau 50**) ont été ajoutées au jeu de données existant pour le gène *COI* des cétacés dans le but de comparer les résultats obtenus. Les quatre paires ont été ajoutées séparément et ensemble et la topologie des arbres (*Neighbor-Joining*, 1000 itérations de *bootstrap*) a été comparée pour chacune des itérations.

Résultats

Figure 1 : Valeurs de distance entre les séquences par position nucléotidique pour l'alignement avec les paramètres par défaut.

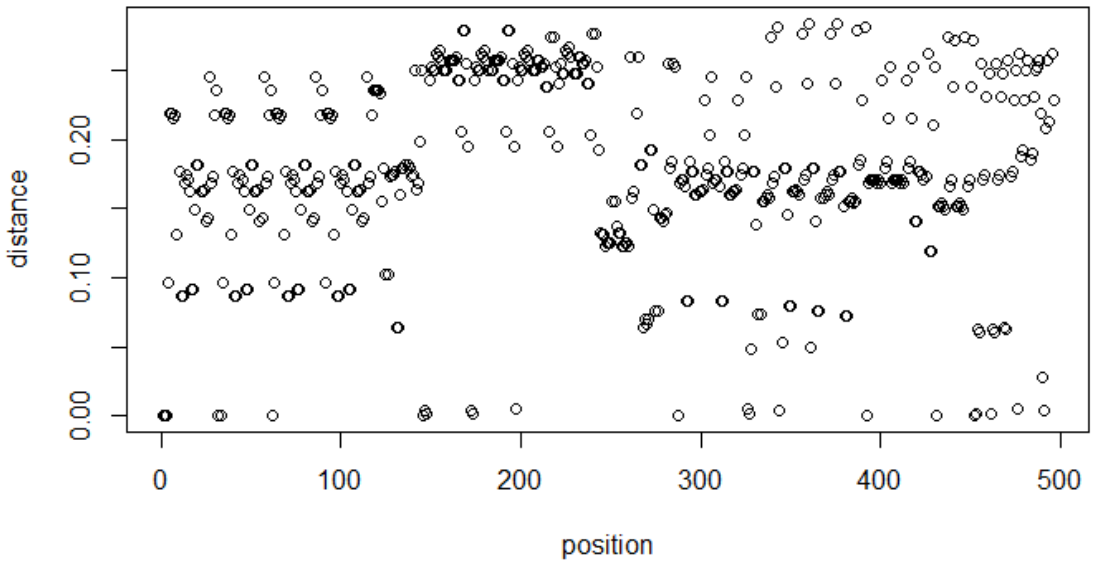

Figure 2 : Valeurs de distance entre les séquences par position nucléotidique pour l'alignement avec les paramètres à zéro.

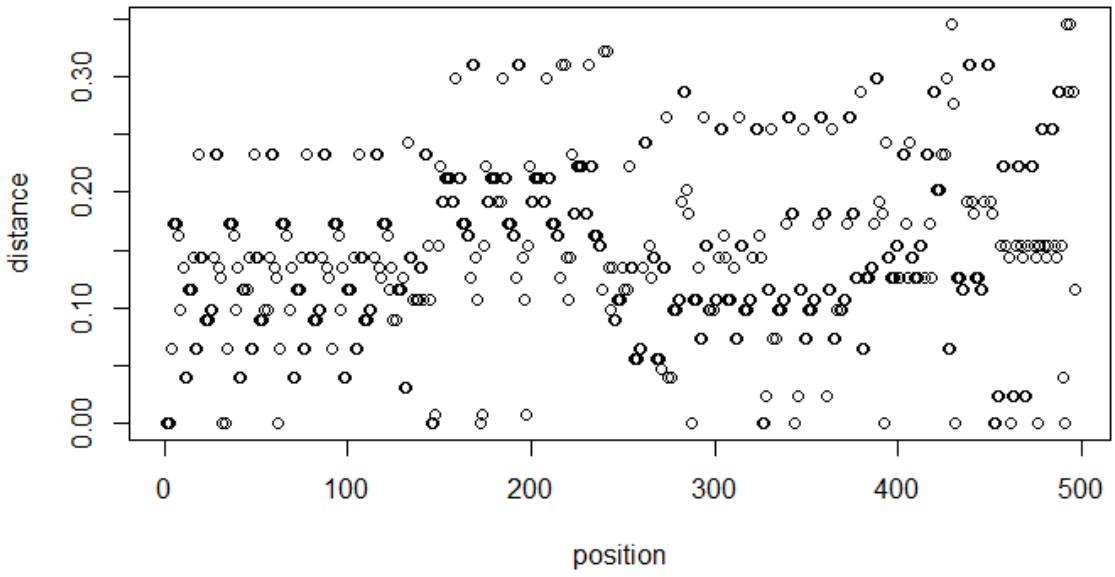

Figure 3 : Valeurs de distance entre les séquences par position nucléotidique pour l'alignement avec les paramètres par défaut selon la méthode de Jukes et Cantor.

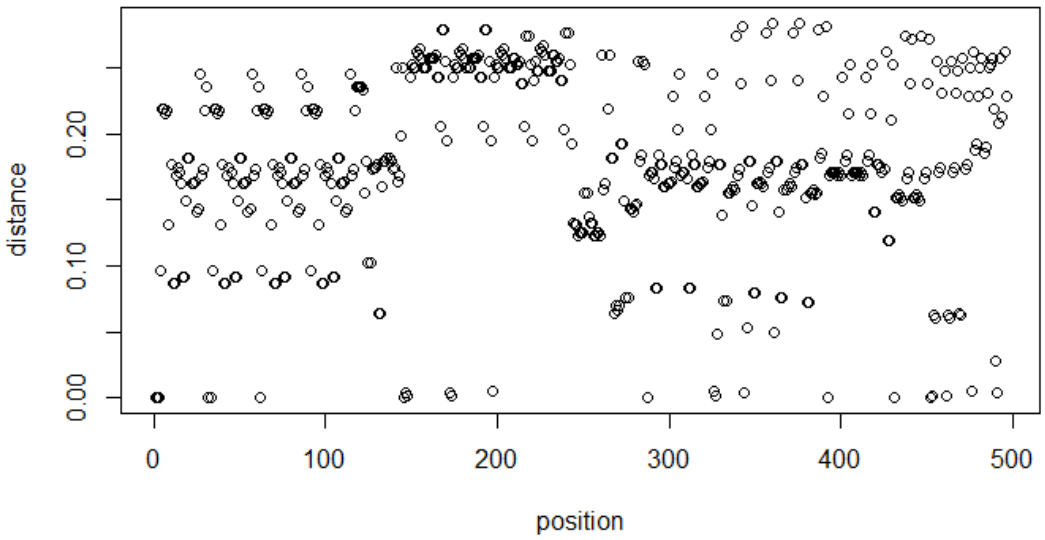

Figure 4 : Valeurs de distance entre les séquences par position nucléotidique pour l'alignement avec les paramètres par défaut selon la méthode de Kimura 2 paramètres.

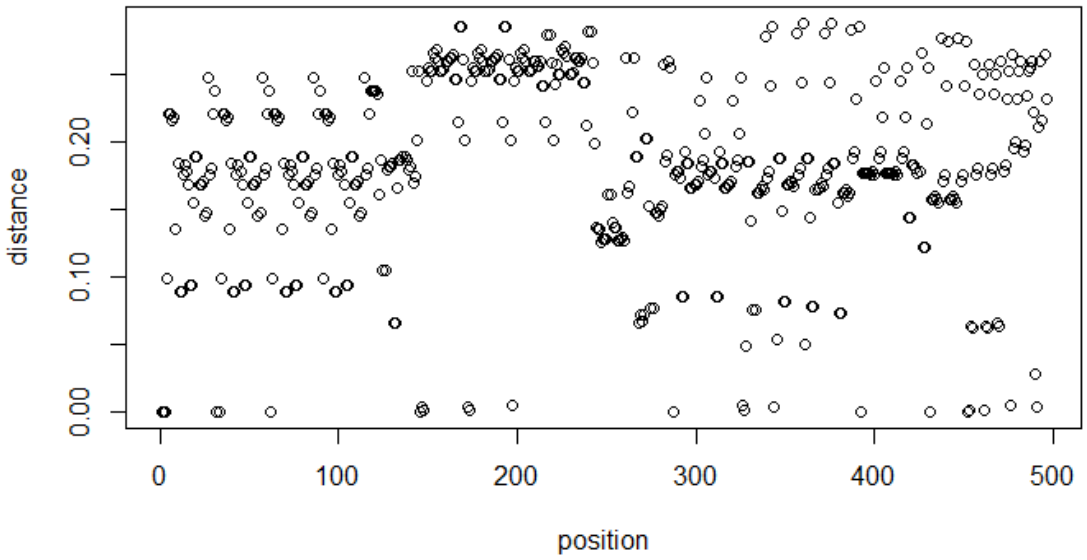

Figure 5 : Valeurs de distance entre les séquences par position nucléotidique pour l'alignement avec les paramètres par défaut selon la méthode de Tamura et Nei.

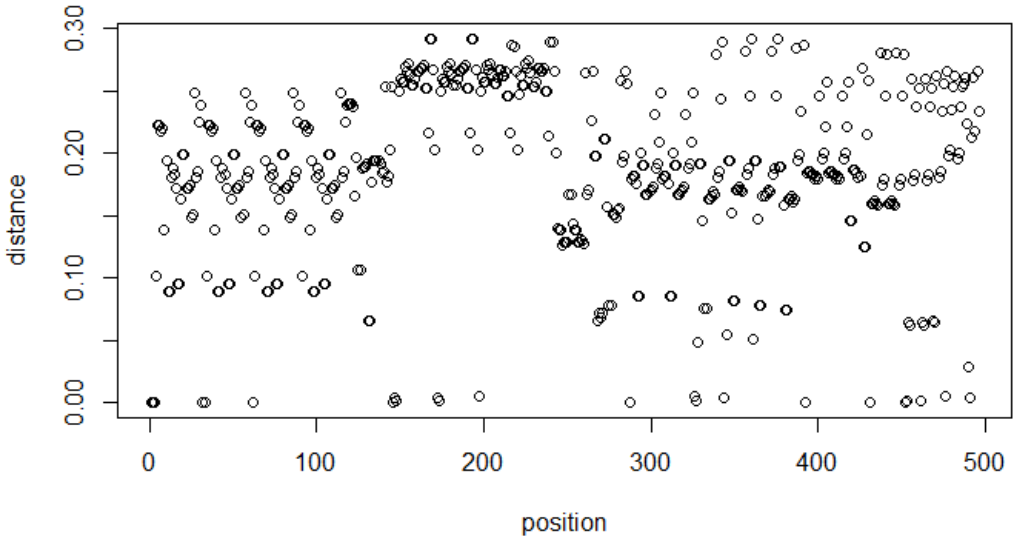

Figure 6 : Valeurs de distance entre les séquences par position nucléotidique pour l'alignement avec les paramètres par défaut selon la méthode de Galtier et Gouy.

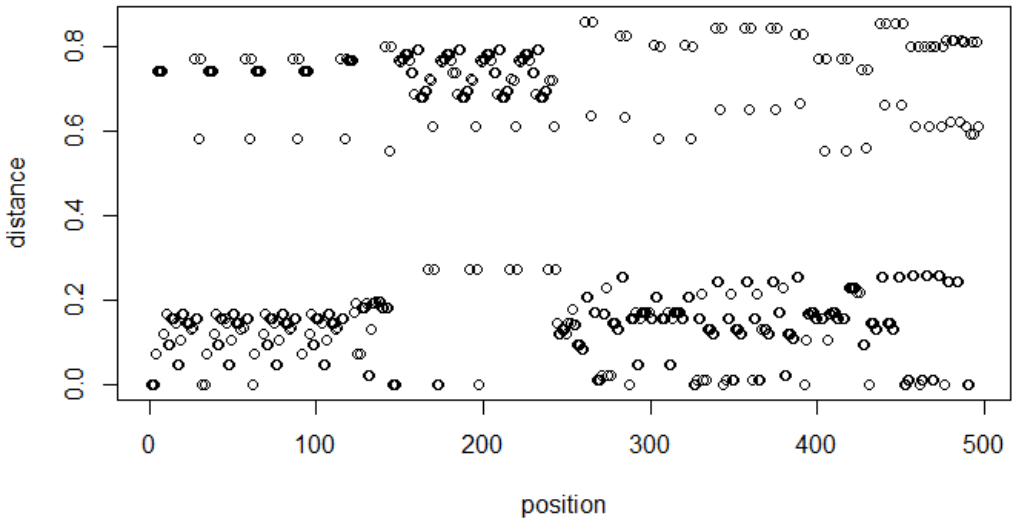

Figure 7 : Construction de l'arbre *Neighbor-Joining* selon la méthode de Jukes et Cantor.

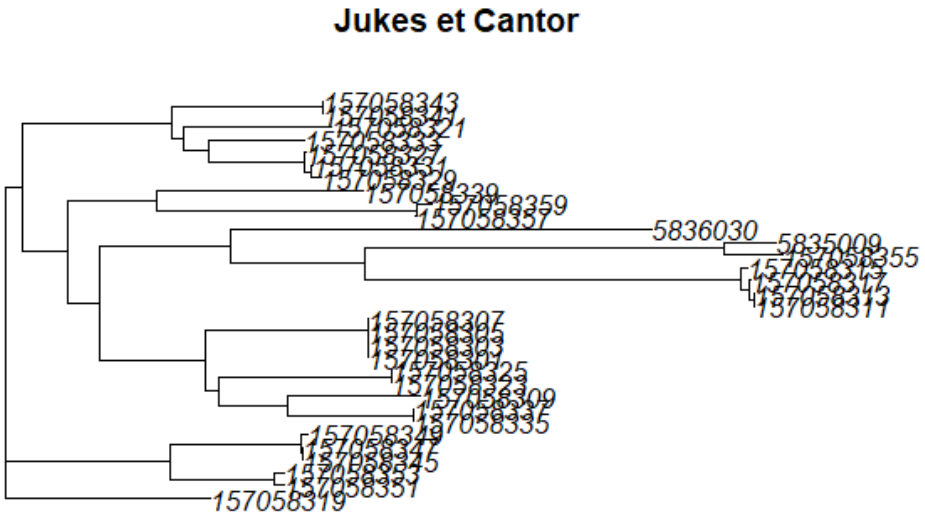

Figure 8 : Construction de l'arbre *Neighbor-Joining* selon la méthode de Kimura 2 paramètres 1980.

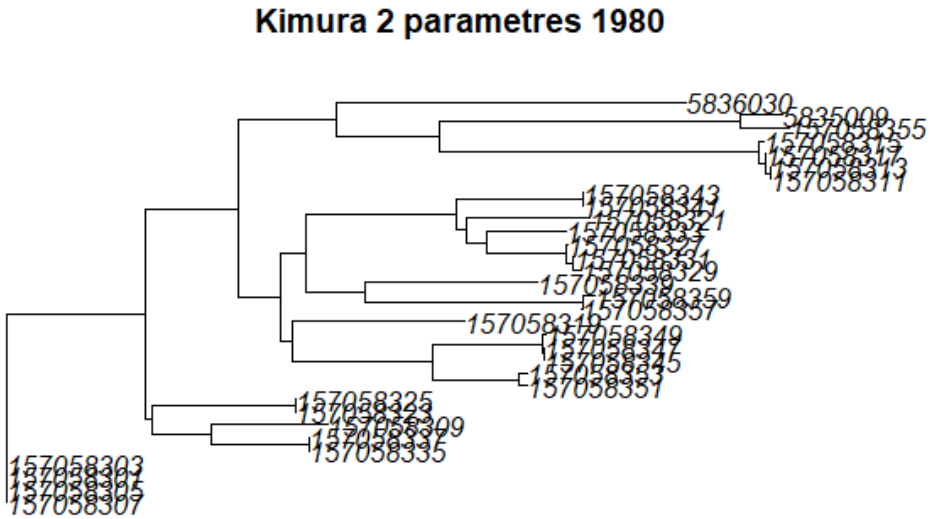

Figure 9 : Construction de l'arbre *Neighbor-Joining* selon la méthode de Tamura et Nei 1993.

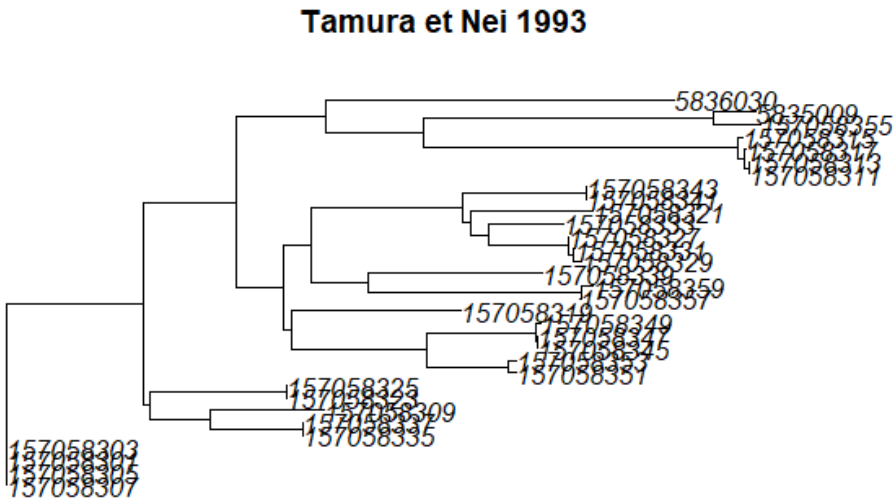

Figure 10 : Construction de l'arbre *Neighbor-Joining* selon la méthode de Galtier et Gouy 1995.

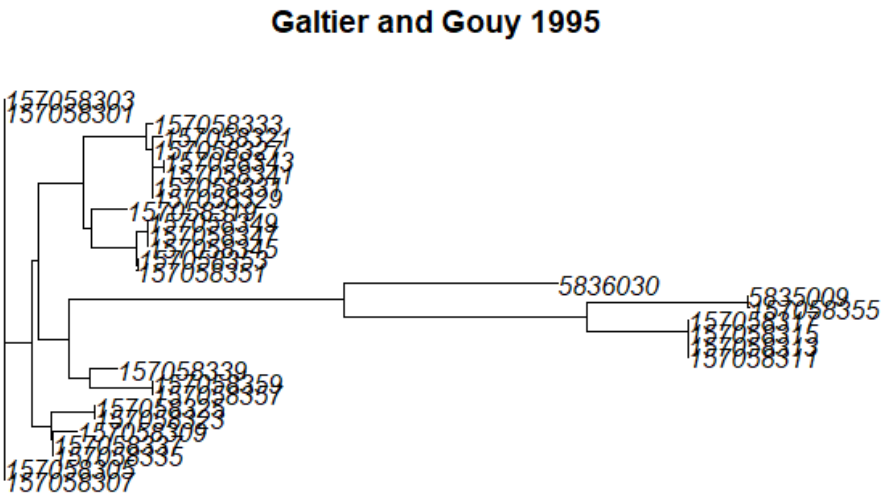

Tableau 1 : Comparaison de l'alignement des modèles entre eux par corrélation

Méthodes	Jukes et Cantor	Kimura 2 paramètres	Tamura et Nei	Galtier et Gouy
Jukes et Cantor	1,000	0,999	0,998	0,866
Kimura 2 paramètres	0,999	1,000	0,999	0,850
Tamura et Nei	0,998	0,999	1,000	0,840
Galtier et Gouy	0,866	0,850	0,840	1,000

Figure 11 : Construction de l'arbre *Neighbor-Joining* selon la méthode de Jukes et Cantor avec les valeurs de l'analyse *bootstrap*.

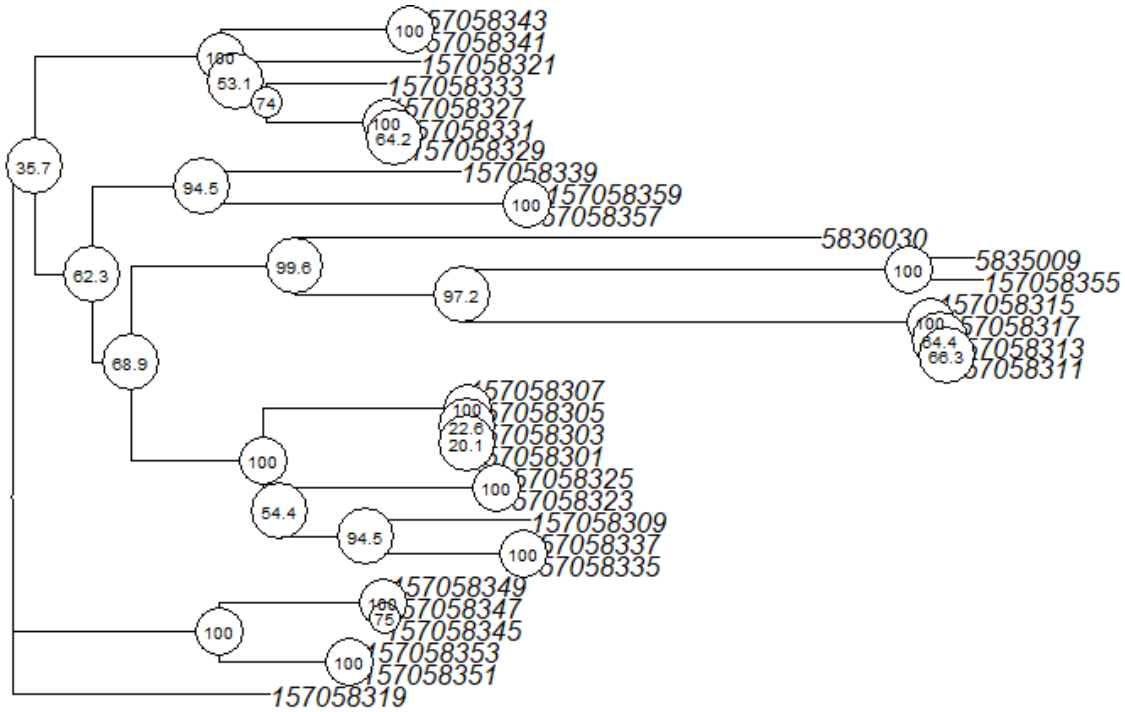

Figure 12 : Construction de l'arbre *Neighbor-Joining* selon la méthode de Tamura et Nei avec les valeurs de l'analyse *bootstrap*.

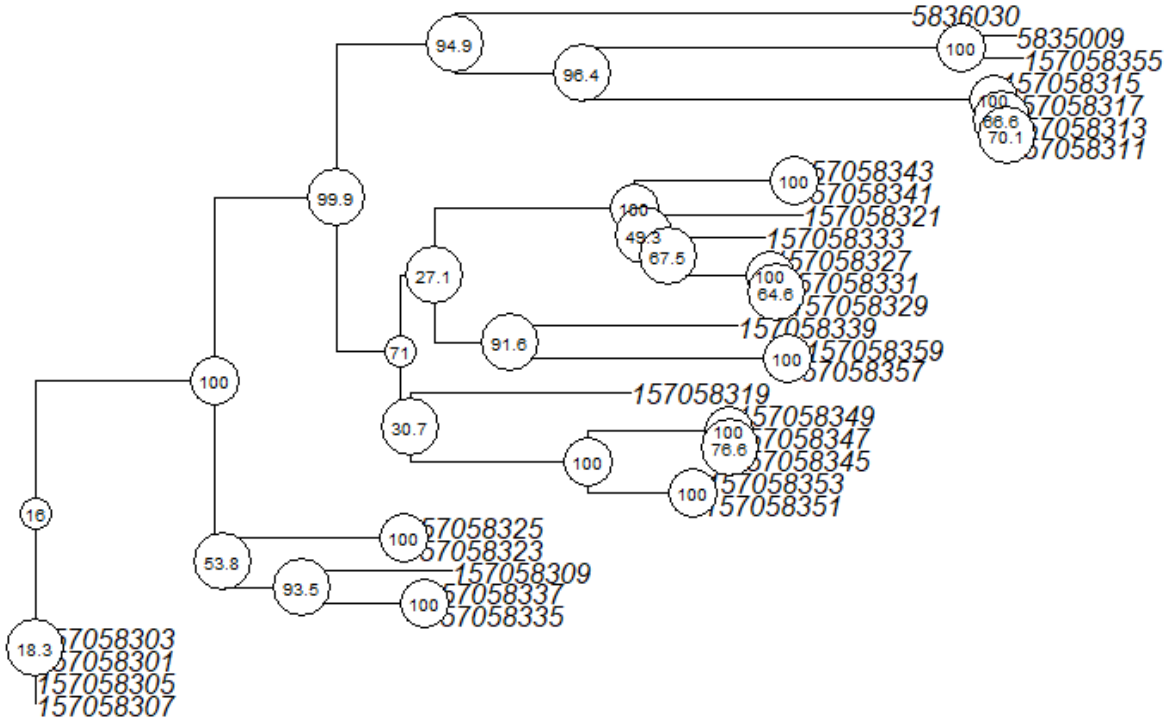

Figure 13 : Construction de l'arbre *Neighbor-Joining* selon la méthode de Kimura 2 parametres avec les valeurs de l'analyse *bootstrap.*

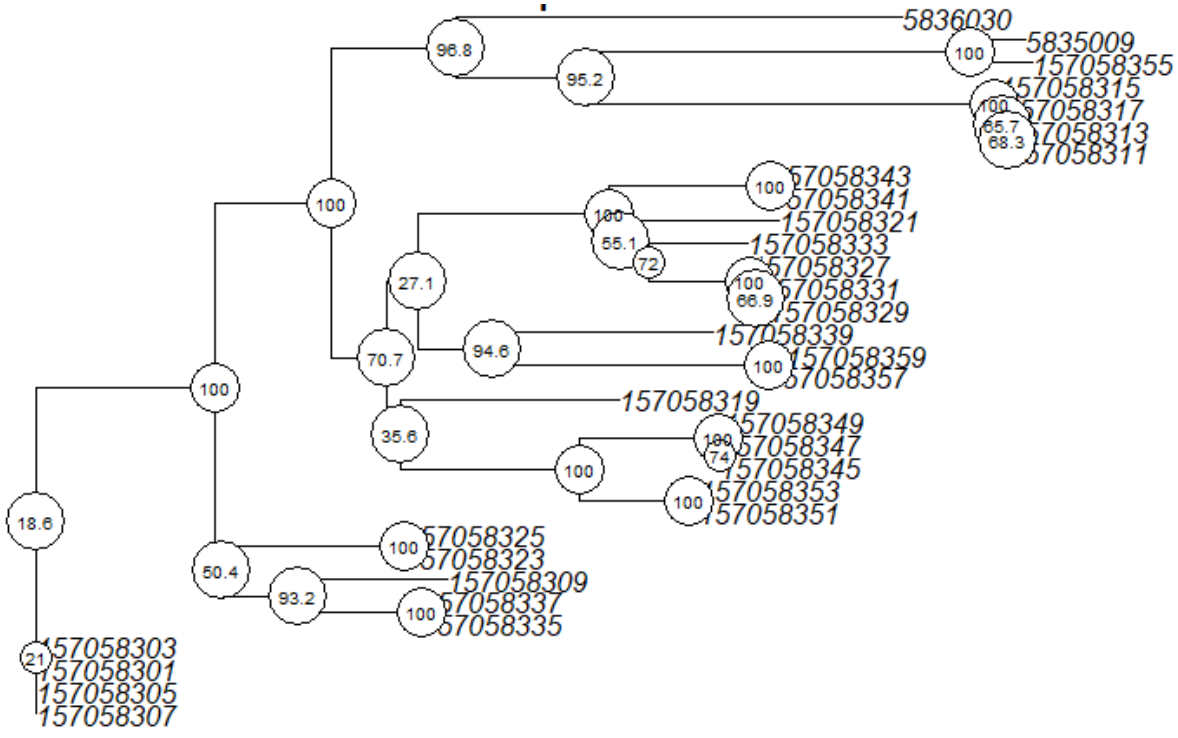

Figure 14 : Construction de l'arbre *Neighbor-Joining* selon la méthode de Galtier et Gouy avec les valeurs de l'analyse *bootstrap*.

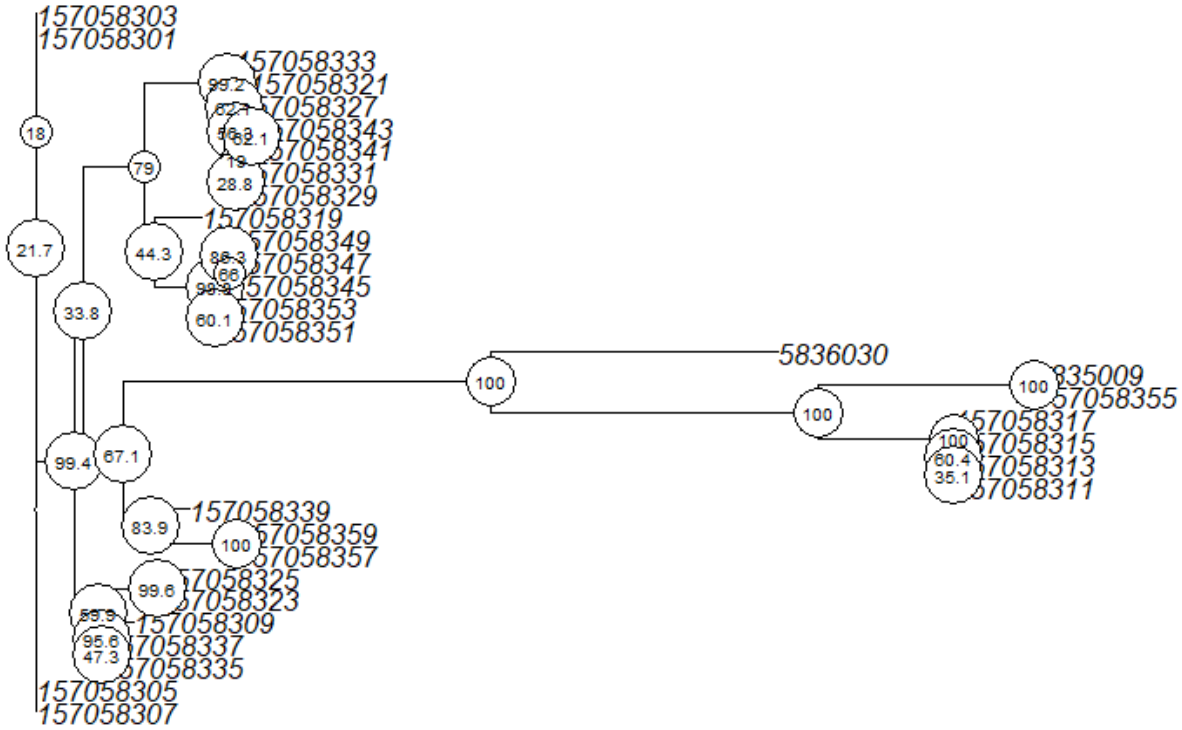

Tableau 2 : Comparaison des analyses *bootstrap* de chaque modèle entre eux par corrélation.

Méthodes	Jukes et Cantor	Kimura 2 paramètres	Tamura et Nei	Galtier et Gouy
Jukes et Cantor	1,000	0,693	0,653	0,257
Kimura 2 paramètres	0,693	1,000	0,826	0,452
Tamura et Nei	0,653	0,826	1,000	0,595
Galtier et Gouy	0,257	0,452	0,595	1,000

Tableau 3 : Les 5 meilleurs modèles selon le critère AIC pour l'alignement nucléotidique par défaut.

Modèles	AIC	BIC
GTR+G(4)+I	11745,01	12124,39
TIM2+G(4)+I	11748,61	12117,31
TVM+G(4)+I	11758,16	12132,20
TPM2u+G(4)+I	11760,88	12124,23
GTR+I	11763,43	12137,47

Tableau 4 : Les 5 meilleurs modèles selon le critère BIC pour l'alignement nucléotidique par défaut.

Modèles	BIC	AIC
TIM2+G(4)+I	12117,31	11748,61
TPM2u+G(4)+I	12124,23	11760,88
GTR+G(4)+I	12124,39	11745,01
TIM2+I	12131,47	11768,12
TVM+G(4)+I	12132,20	11758,16

Tableau 5 : Paramètres du meilleur modèle pour l'alignement nucléotidique par défaut.

Modèle	inv	k	shape
TIM2	0,58859	4	2,76728

Figure 15 : Arbre phylogénétique reconstruit avec le modèle TIM2 et analyse de *bootstrap* (1000 itérations)

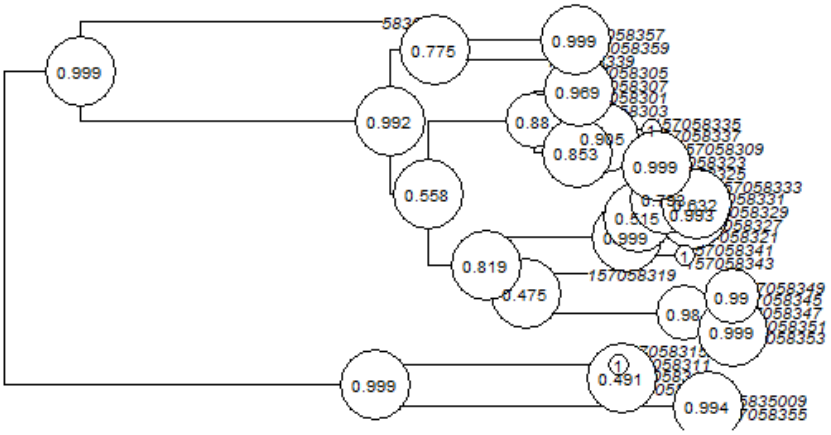

Figure 16 : Valeurs de distance entre les séquences d'acide aminés selon la méthode LG pour le cadre de lecture par défaut.

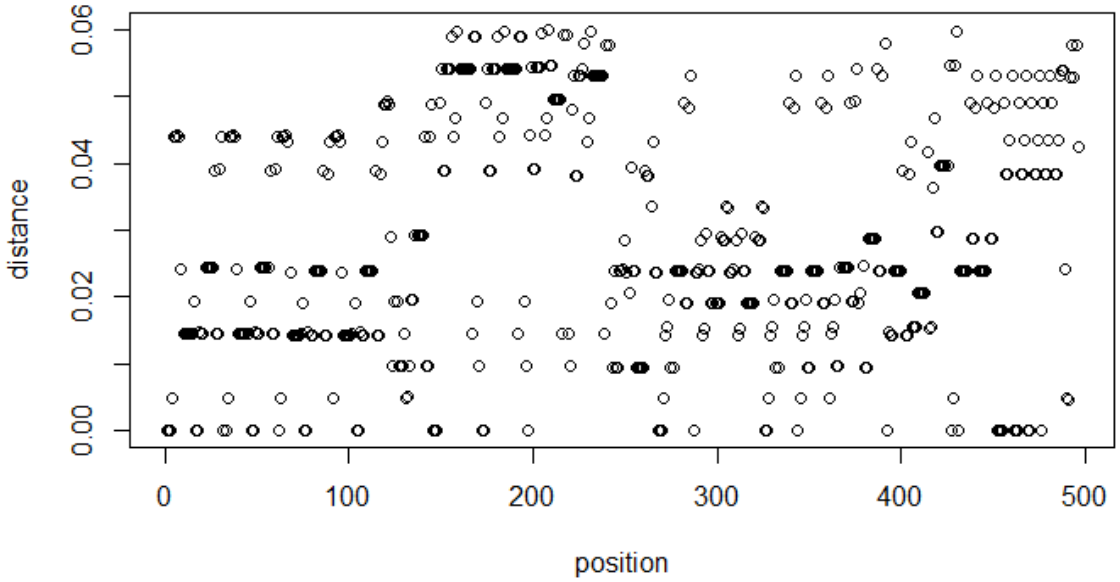

Figure 17 : Valeurs de distance entre les séquences d'acide aminés selon la méthode JTT pour le cadre de lecture par défaut.

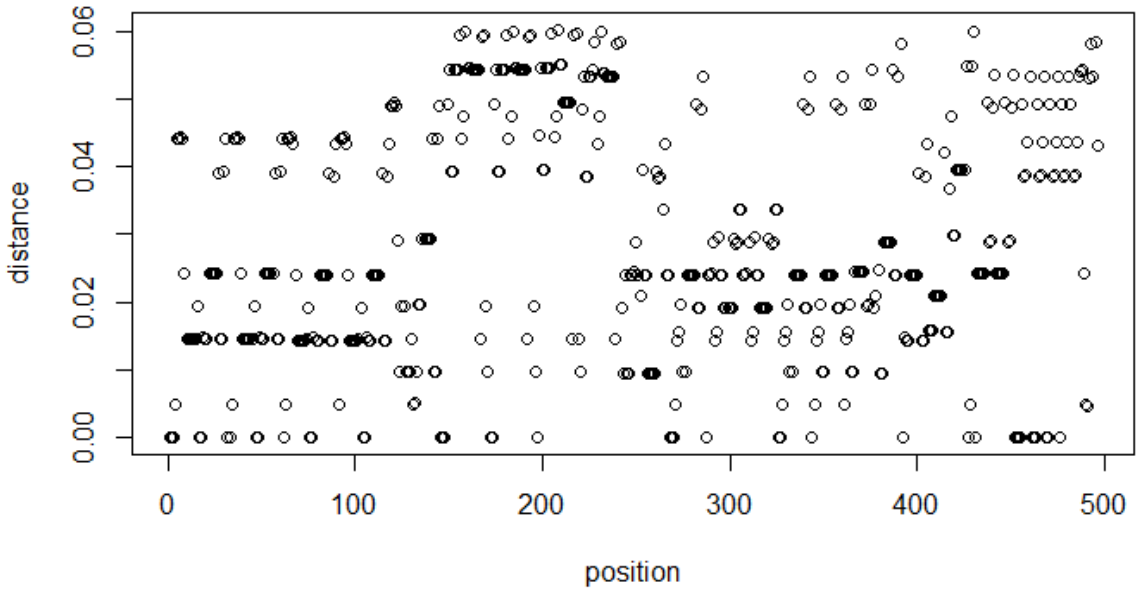

Figure 18 : Valeurs de distance entre les séquences d'acide aminés selon la méthode Blosum62 pour le cadre de lecture par défaut.

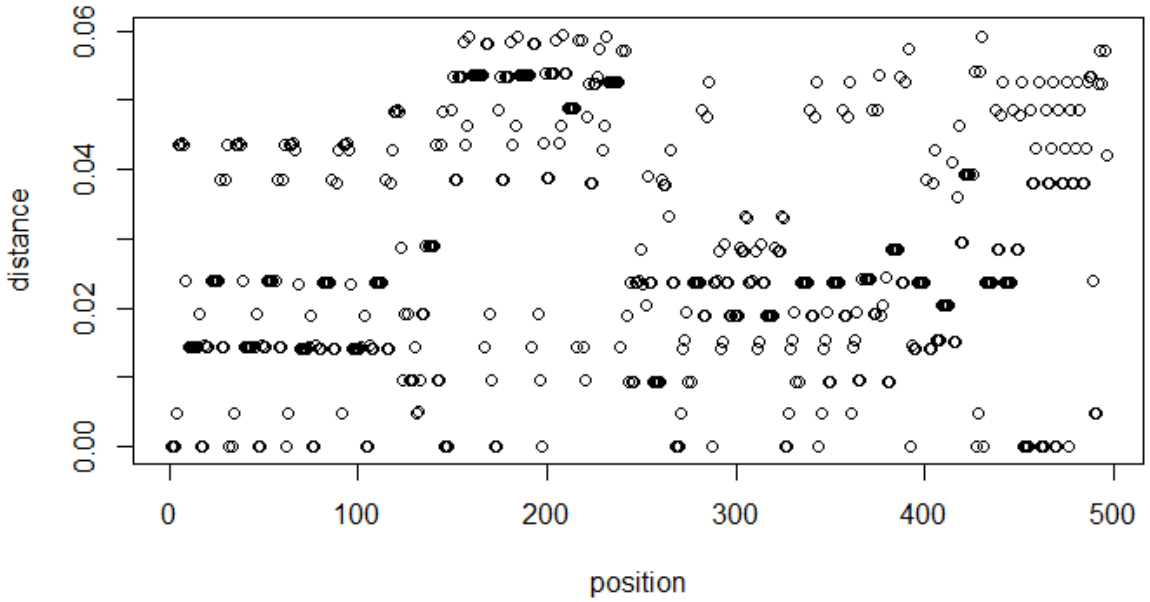

Figure 19 : Valeurs de distance entre les séquences d'acide aminés selon la méthode LG pour le cadre de lecture 1.

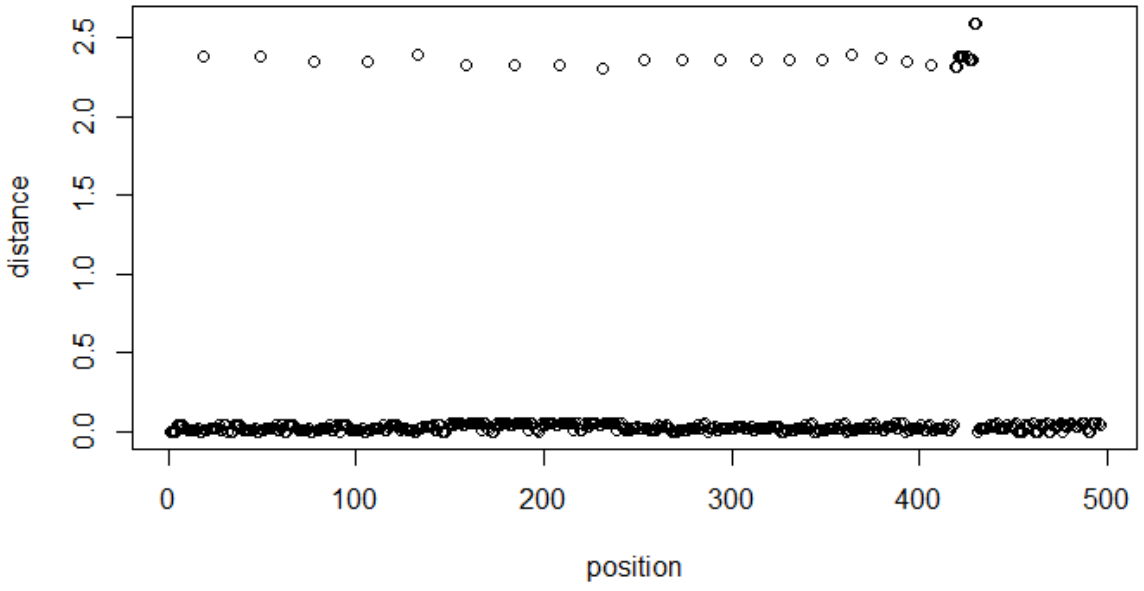

Figure 20 : Valeurs de distance entre les séquences d'acide aminés selon la méthode JTT pour le cadre de lecture 1.

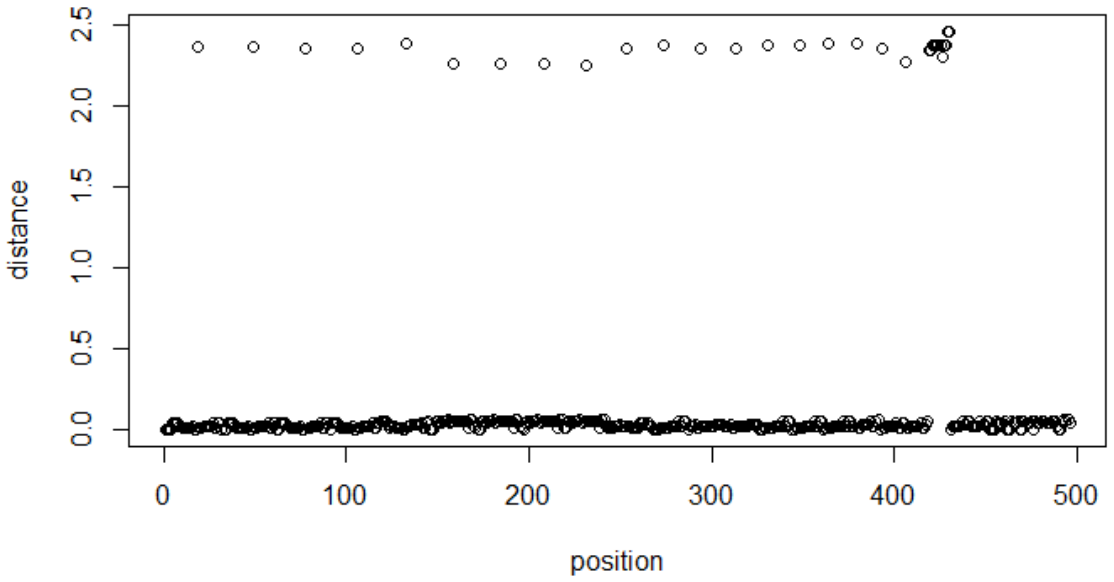

Figure 21 : Valeurs de distance entre les séquences d'acide aminés selon la méthode Blosum62 pour le cadre de lecture 1.

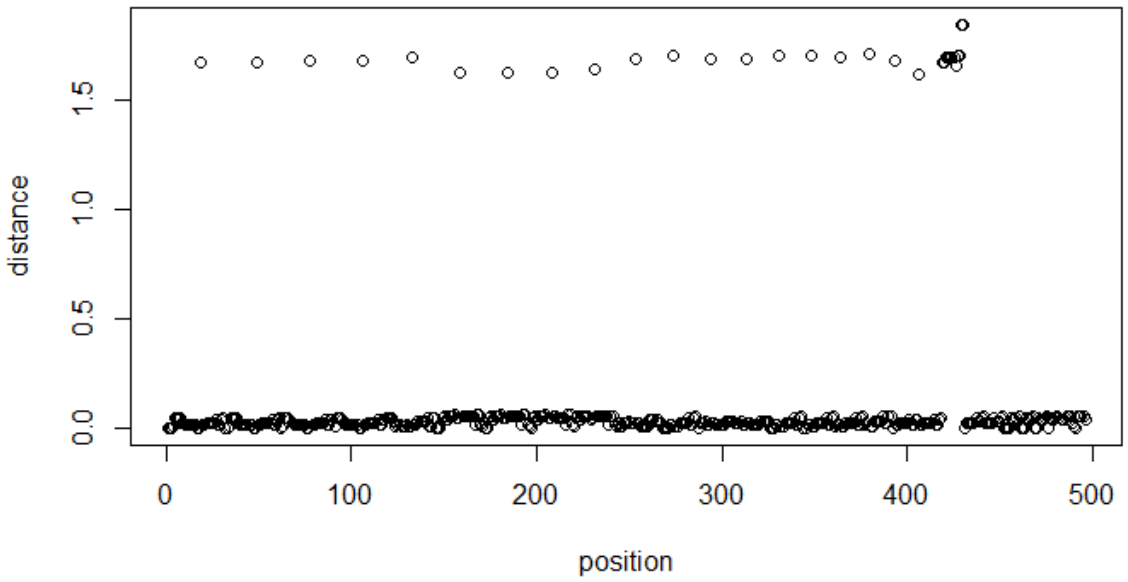

Tableau 6 : Les 5 meilleurs modèles selon le critère AIC pour l'alignement d'acide aminé pour la cadre de lecture par défaut.

Modèles	**AIC**	**BIC**
Blosum62+I	3674,411	3936,944
JTT+I	3674,647	3937,181
Blosum62+G(4)+I	3676,496	3943,263
JTT+G(4)+I	3676,731	3943,499
Blosum62+G(4)	3678,495	3941,028

Tableau 7 : Les 5 meilleurs modèles selon le critère BIC pour l'alignement d'acide aminé pour la cadre de lecture par défaut.

Modèles	**BIC**	**AIC**
Blosum62+I	3936,944	3674,411
JTT+I	3937,181	3674,647
Blosum62+G(4)	3941,028	3678,495
JTT+G(4)	3942,121	3679,587
Blosum62+G(4)+I	3943,263	3676,496

Tableau 8 : Les 5 meilleurs modèles selon le critère AIC pour l'alignement d'acide aminé pour la cadre de lecture 1.

Modèles	**AIC**	**BIC**
LG+G(4)	4623,775	4892,768
JTT+G(4)	4624,547	4893,540
Blosum62+G(4)	4624,596	4893,589
LG+G(4)+I	4625,713	4899,044
Blosum62	4526,160	4890,814

Tableau 9 : Les 5 meilleurs modèles selon le critère BIC pour l'alignement d'acide aminé pour la cadre de lecture 1.

Modèles	**BIC**	**AIC**
Blosum62	4890,814	4626,160
LG+G(4)	4892,768	4823,775
JTT+G(4)	4893,540	4624,547
Blosum62+G(4)	4893,589	4624,596
LG	4894,865	4630,211

Tableau 10 : Comparaison des analyses *bootstrap* de chaque modèle (cadre de lecture par défaut et cadre de lecture 1) entre eux par corrélation.

Méthodes	**LG**	**JTT**	**Blosum62**	**LG Cadre 1**	**JTT Cadre 1**	**Blosum62 Cadre 1**
LG	1,000	0,989	0,986	0,547	0,178	0,165
JTT	0,989	1,000	0,975	0,506	0,224	0,159
Blosum62	0,986	0,975	1,000	0,494	0,193	0,238
LG Cadre 1	0,547	0,506	0,494	1,000	0,010	-0,003
JTT Cadre 1	0,178	0,224	0,193	0,010	1,000	0,100
Blosum62 Cadre 1	0,165	0,159	0,238	-0,003	0,100	1,000

Figure 22 : Visualisation des valeurs de *bootstrap* selon la méthode LG pour le cadre de lecture par défaut.

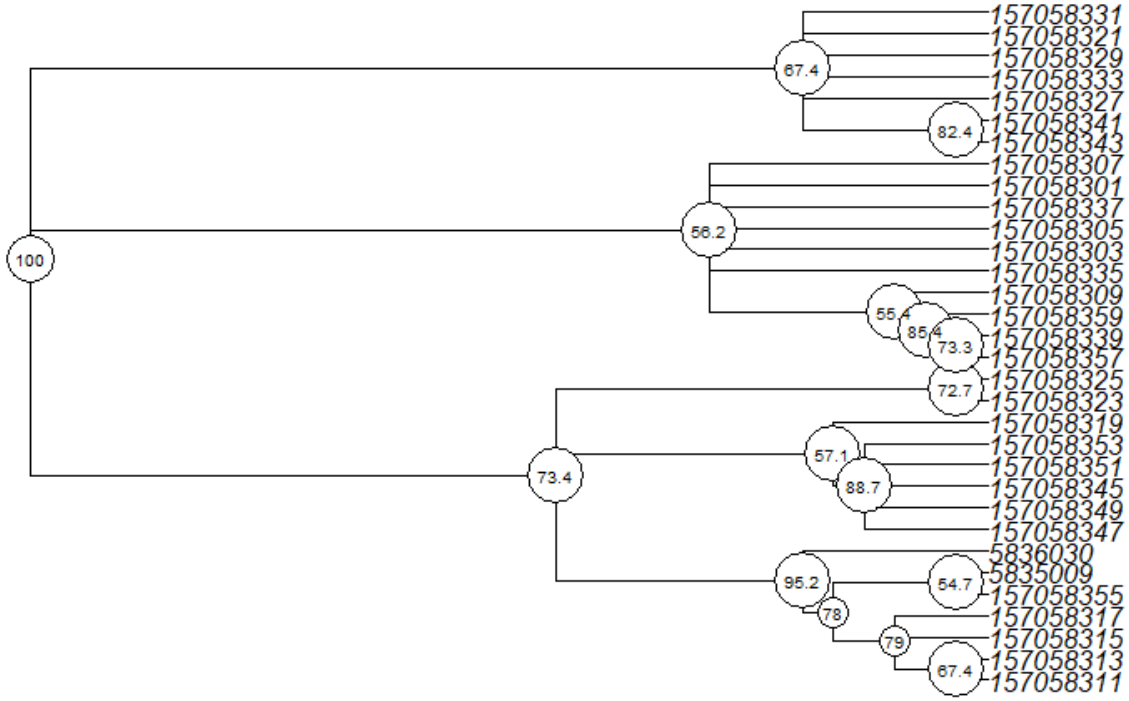

Figure 23 : Visualisation des valeurs de *bootstrap* selon la méthode JTT pour le cadre de lecture par défaut.

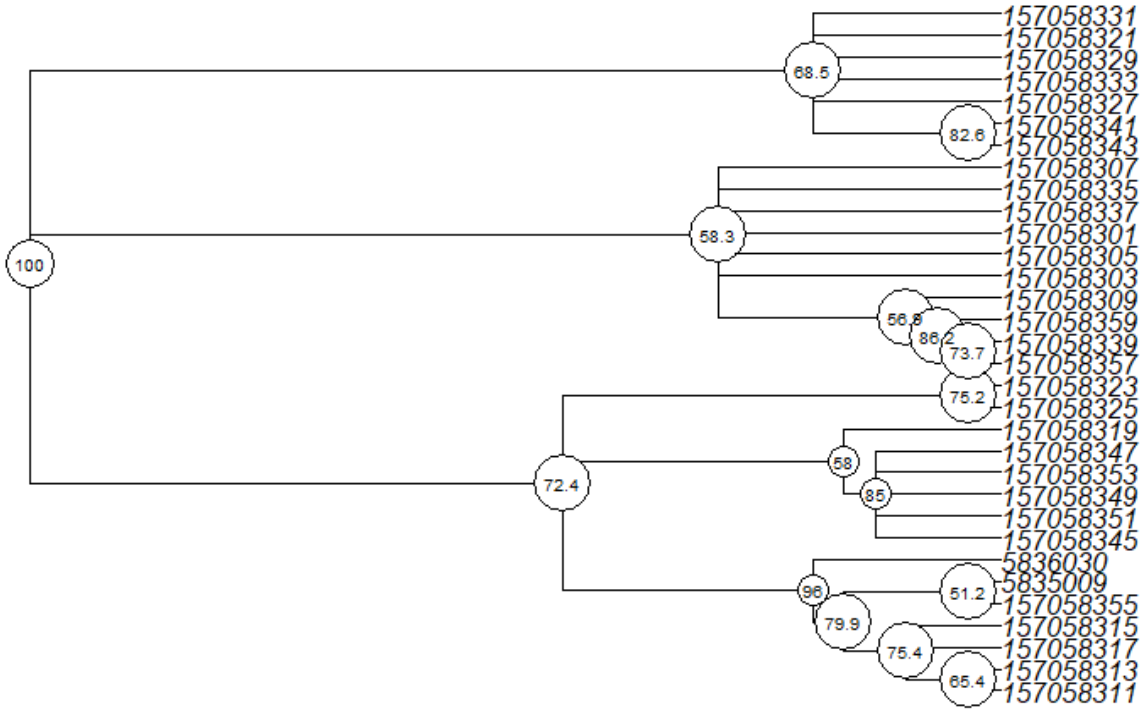

Figure 24 : Visualisation des valeurs de *bootstrap* selon la méthode Blosum62 pour le cadre de lecture par défaut.

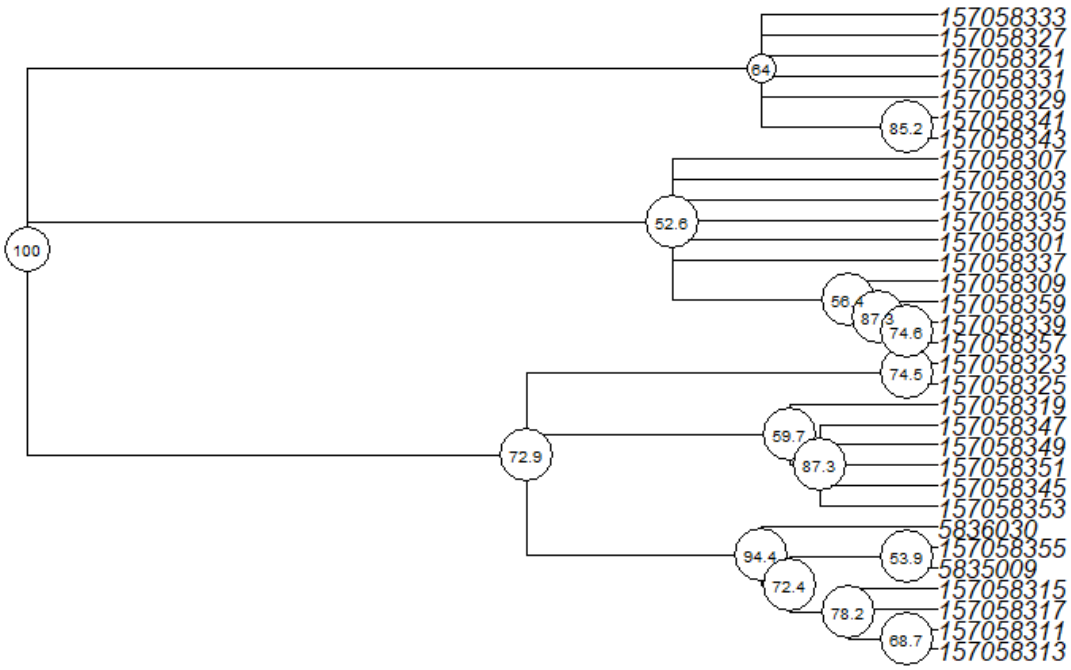

Figure 25 : Visualisation des valeurs de *bootstrap* selon la méthode LG pour le cadre de lecture 1.

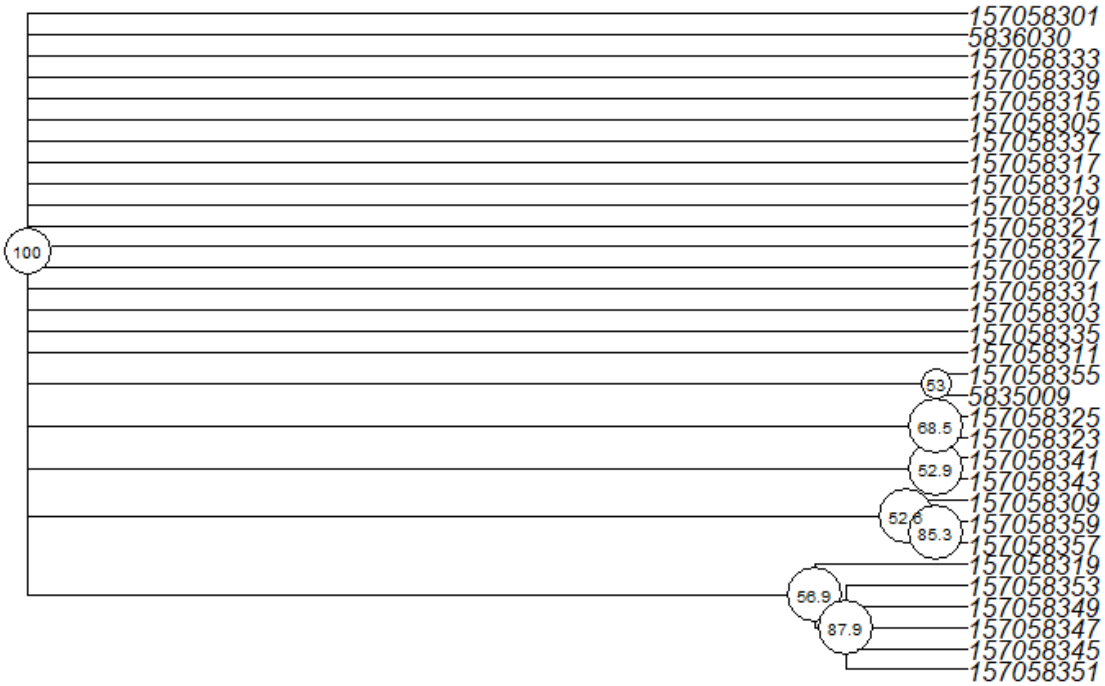

Figure 26 : Visualisation des valeurs de *bootstrap* selon la méthode JTT pour le cadre de lecture 1.

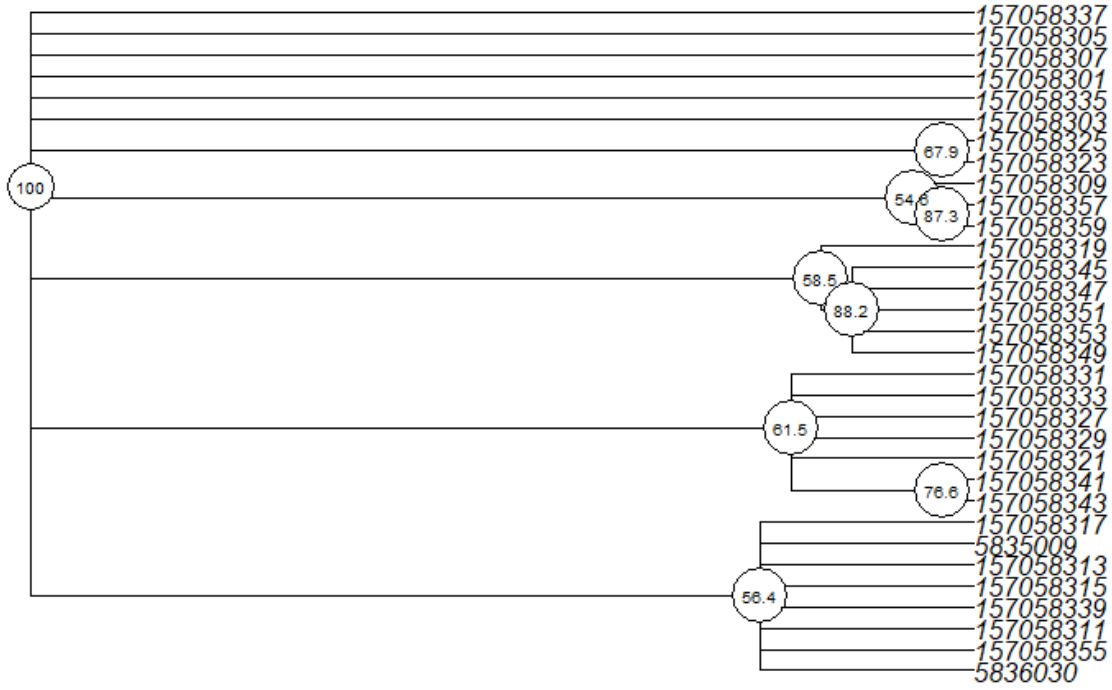

Figure 27 : Visualisation des valeurs de *bootstrap* selon la méthode Blosum62 pour le cadre de lecture 1.

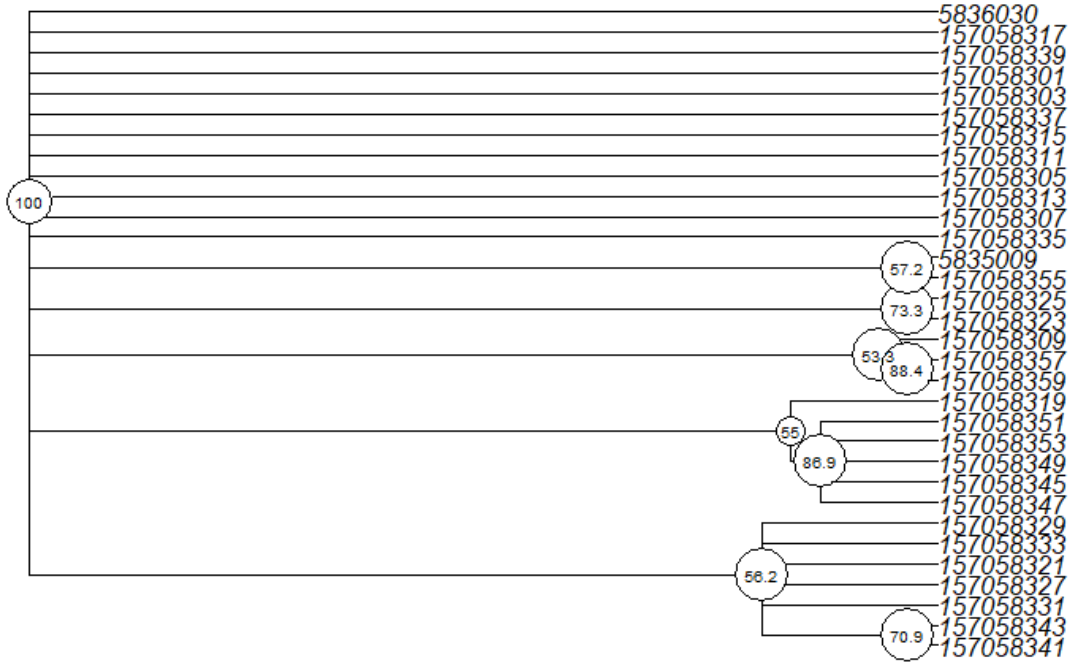

Tableau 11 : Les 5 meilleurs modèles selon le critère AIC pour l'alignement d'acide aminés pour la cadre de lecture 1.

Modèles	**AIC**	**BIC**
MtZoa+G(4)	4478,002	4746,994
MtZoa+G(4)+I	4479,944	4753,275
MtZoa+I	4481,854	4750,847
MtZoa	4482,641	4747,295
mtArt+G(4)	4502,102	4771,095

Tableau 12 : Les 5 meilleurs modèles selon le critère BIC pour l'alignement d'acide aminés pour la cadre de lecture 1.

Modèles	**BIC**	**AIC**
MtZoa+G(4)	4746,994	4478,002
MtZoa	4747,295	4482,641
MtZoa+I	4750,847	4481,854
MtZoa+G(4)+I	4753,275	4479,944
mtArt+G(4)	4771,095	4502,102

Tableau 13 : Paramètres du meilleur modèle pour l'alignement d'acide aminés pour la cadre de lecture 1.

Modèle	**k**	**shape**
MtZoa	4	0,83135

Tableau 14 : Les 5 meilleurs modèles selon le critère AIC pour l'alignement d'acide aminés pour la cadre de lecture 2.

Modèles	**AIC**	**BIC**
HIVw+I	10327,97	10596,19
HIVw+G(4)+I	10327,98	10600,52
HIVw+G(4)	10327,71	10596,93
HIVw	10327,23	10597,12
HIVb+G(4)	10587,82	10656,04

Tableau 15 : Les 5 meilleurs modèles selon le critère BIC pour l'alignement d'acide aminés pour la cadre de lecture 2.

Modèles	**BIC**	**AIC**
HIVw+I	10596,19	10327,97
HIVw+G(4)	10596,93	10327,71
HIVw	10597,12	10327,23
HIVw+G(4)+I	10600,52	10327,98
HIVb+G(4)	10656,04	10587,82

Tableau 16 : Paramètres du meilleur modèle pour l'alignement d'acide aminés pour la cadre de lecture 2.

Modèle	**inv**
HIVw	0,03177

Tableau 17 : Les 5 meilleurs modèles selon le critère AIC pour l'alignement d'acide aminés pour la cadre de lecture 3.

Modèles	**AIC**	**BIC**
HIVw+G(4)	10665,34	10930,61
HIVw+G(4)+I	10666,82	10936,37
HIVw+I	10696,16	10961,43
HIVw	10707,87	10968,86
mtmam+G(4)	10814,11	11079,38

Tableau 18 : Les 5 meilleurs modèles selon le critère BIC pour l'alignement d'acide aminés pour la cadre de lecture 3.

Modèles	**BIC**	**AIC**
HIVw+G(4)	10930,61	10665,34
HIVw+G(4)+I	10936,37	10666,82
HIVw+I	10961,43	10696,16
HIVw	10968,86	10707,87
mtmam+G(4)	11079,38	10814,11

Tableau 19 : Paramètres du meilleur modèle pour l'alignement d'acide aminés pour la cadre de lecture 3.

Modèle	**k**	**shape**
HIVw	4	3,11593

Figure 28 : Visualisation des valeurs de *bootstrap* selon la méthode MtZoa pour le cadre de lecture 1 sans correction.

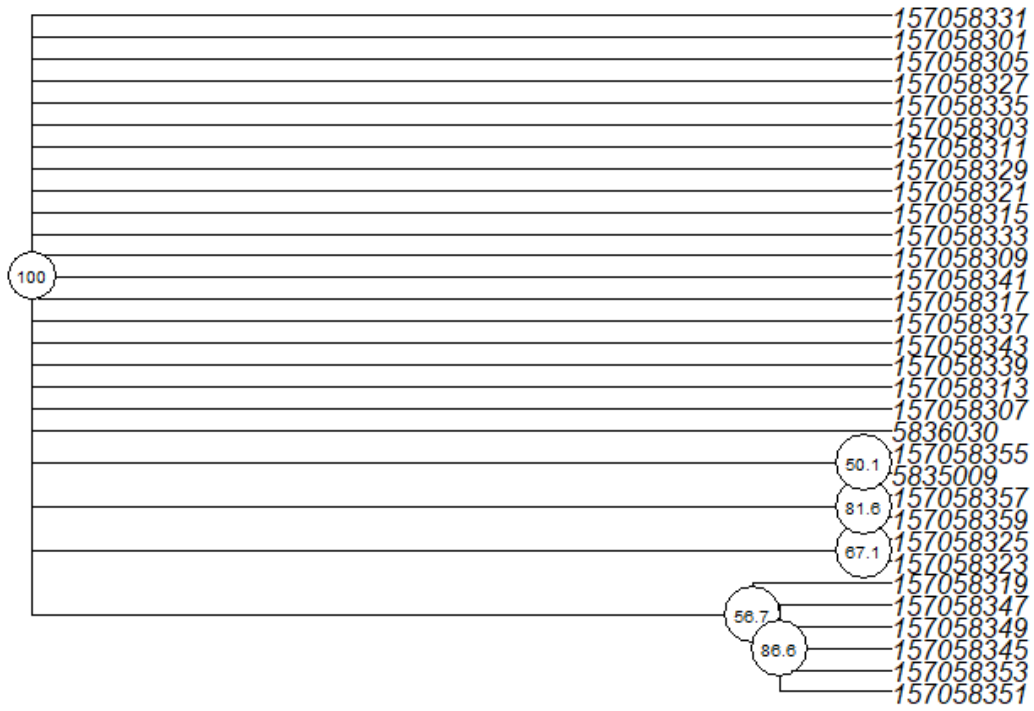

Figure 29 : Visualisation des valeurs de *bootstrap* selon la méthode HIVw pour le cadre de lecture 2 sans correction.

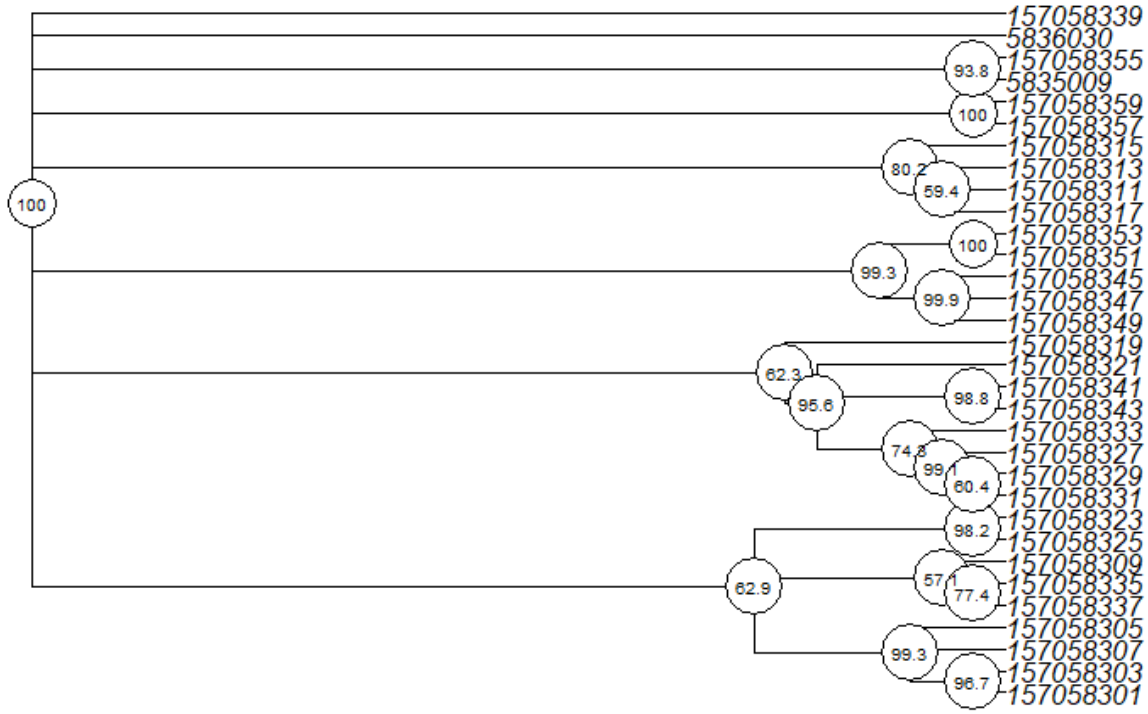

Figure 30 : Visualisation des valeurs de *bootstrap* selon la méthode HIVw pour le cadre de lecture 3 sans correction.

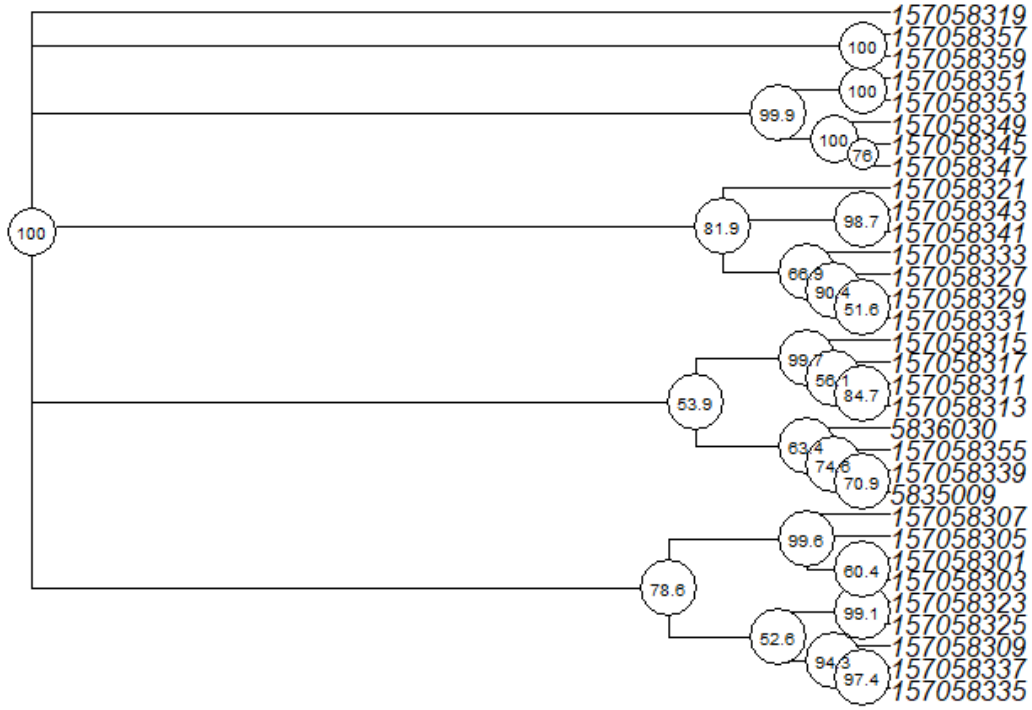

Figure 31 : Visualisation des valeurs de *bootstrap* selon la méthode MtZoa+G(4) pour le cadre de lecture 1 avec correction.

Figure 32 : Visualisation des valeurs de *bootstrap* selon la méthode HIVw+I pour le cadre de lecture 2 avec correction.

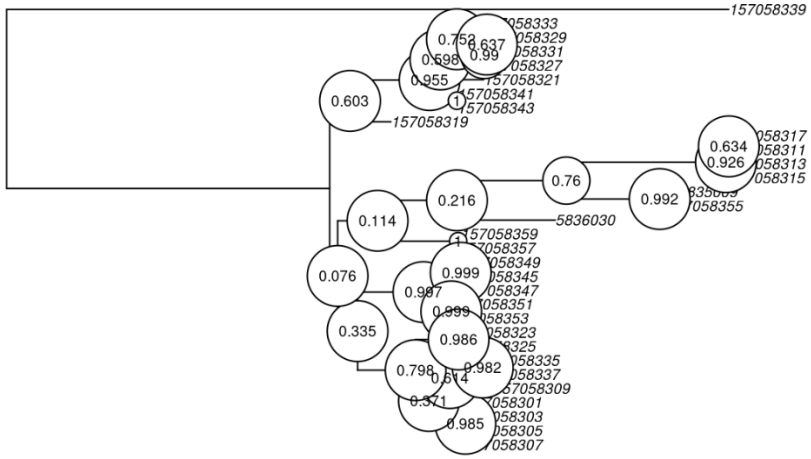

Figure 33 : Visualisation des valeurs de *bootstrap* selon la méthode HIVw+G(4) pour le cadre de lecture 3 avec correction.

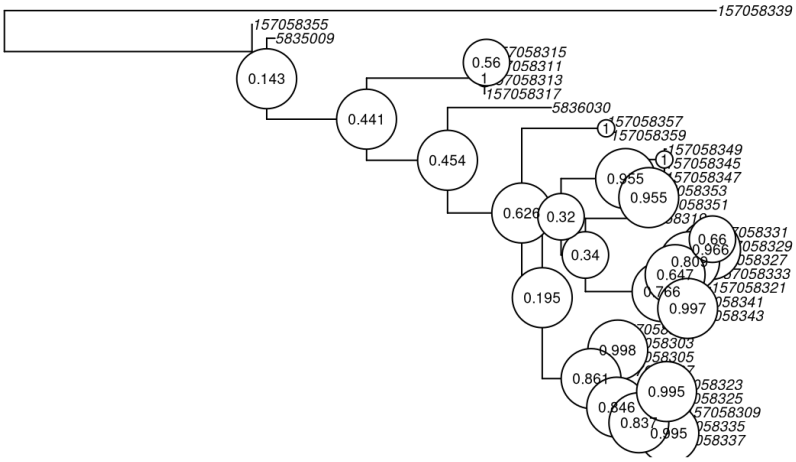

Tableau 20 : Les 5 meilleurs modèles selon le critère AIC pour l'alignement nucléotique des cétacés et du groupe externe composé du lamantin et du dugong (Siréniens).

Modèles	AIC	BIC
GTR+G(4)+I	7297,943	7571,785
TVM+G(4)+I	7301,169	7570,521
TIM3+G(4)+I	7301,604	7566,467
GTR+I	7301,778	7571,130
TVM+I	7304,062	7568,925

Tableau 21 : Les 5 meilleurs modèles selon le critère BIC pour l'alignement nucléotique des cétacés et du groupe externe composé du lamantin et du dugong (Siréniens).

Modèles	BIC	AIC
TPM3u+I	7563,715	7307,831
TPM3u+G(4)+I	7564,783	7304,409
TIM3+I	7566,371	7305,997
TIM3+G(4)+I	7566,467	7301,604
HKY+I	7568,394	7316,999

Tableau 22 : Paramètres du meilleur modèle pour l'alignement nucléotique des cétacés et du groupe externe composé du lamantin et du dugong (Siréniens).

Modèle	inv	k	shape
TIM3	0,60011	4	5,33818

Figure 34 : Construction de l'arbre *Neighbor-Joining* selon la méthode de TIM3+G(4)+I avec l'analyse *bootstrap* avec correction pour l'alignement nucléotique des cétacés et du groupe externe composé du lamantin et du dugong (Siréniens).

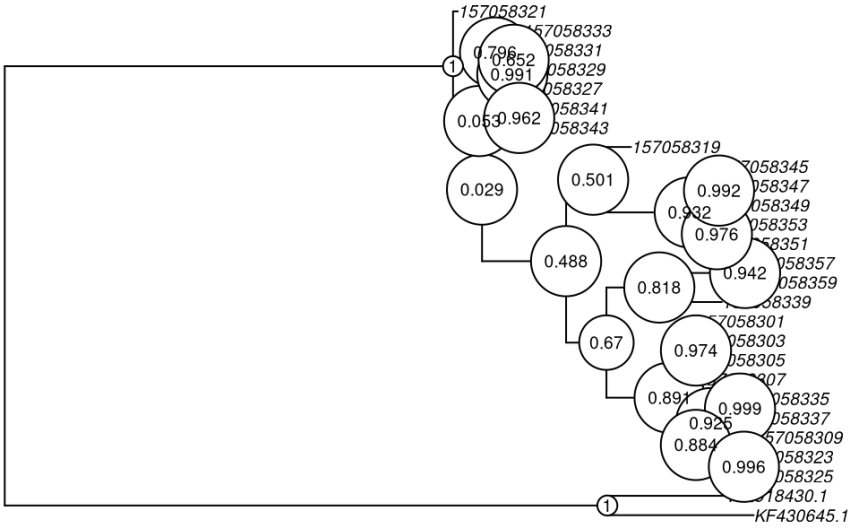

Tableau 23 : Les 5 meilleurs modèles selon le critère AIC pour l'alignement d'acide aminés des cétacés et du groupe externe composé du lamantin et du dugong (Siréniens).

Modèles	AIC	BIC
MtZoa+I	1588,189	1763,462
MtZoa+G(4)	1588,432	1763,705
MtZoa+G(4)+I	1590,213	1768,856
mtmam+I	1596,645	1771,918
mtmam+G(4)	1596,839	1772,112

Tableau 24 : Les 5 meilleurs modèles selon le critère BIC pour l'alignement acide aminés des cétacés et du groupe externe composé du lamantin et du dugong (Siréniens).

Modèles	BIC	AIC
MtZoa+I	1763,462	1588,189
MtZoa+G(4)	1763,705	1588,432
MtZoa+G(4)+I	1768,856	1590,213
MtZoa	1769,165	1597,262
mtmam+I	1771,918	1596,645

Tableau 25 : Paramètres du meilleur modèle pour l'alignement acide aminés des cétacés et du groupe externe composé du lamantin et du dugong (Siréniens).

Modèle	inv
MtZoa	0,76984

Figure 35 : Construction de l'arbre *Neighbor-Joining* selon la méthode de MtZoa+I avec l'analyse *bootstrap* avec correction pour l'alignement acide aminés des cétacés et du groupe externe composé du lamantin et du dugong (Siréniens).

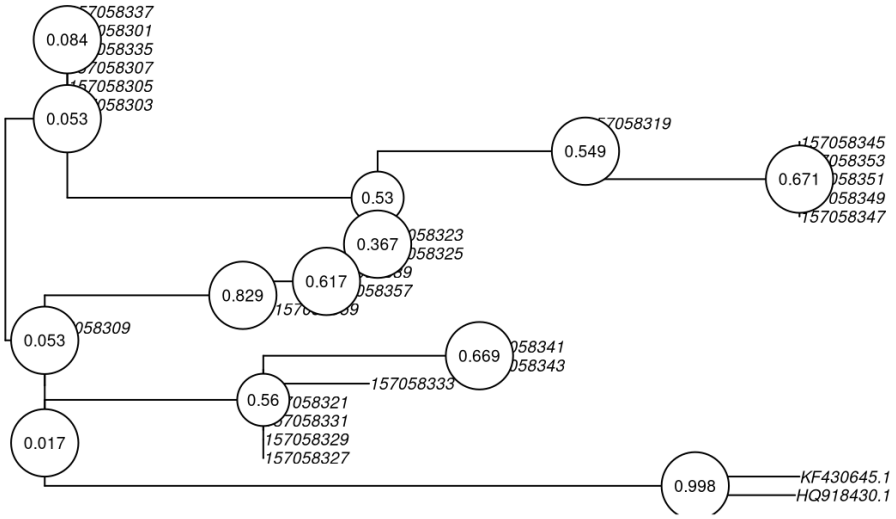

Tableau 26 : Les 5 meilleurs modèles selon le critère AIC pour l'alignement nucléotique des cétacés et du groupe externe composé de l'éléphant d'Afrique et d'Asie.

Modèles	AIC	BIC
GTR+G(4)+I	7218,794	7497,361
TIM3+G(4)+I	7224,217	7493,651
TVM+G(4)+I	7225,478	7499,479
TIM2+G(4)+I	7226,113	7495,546
GTR+I	7228,271	7502,272

Tableau 27 : Les 5 meilleurs modèles selon le critère BIC pour l'alignement nucléotique des cétacés et du groupe externe composé de l'éléphant d'Afrique et d'Asie.

Modèles	BIC	AIC
TIM3+G(4)+I	7493,651	7224,217
TPM3u+G(4)+I	7494,955	7230,088
TIM2+G(4)+I	7495,546	7226,113
TrN+G(4)+I	7496,987	7232,120
GTR+G(4)+I	7497,361	7218,794

Tableau 28 : Paramètres du meilleur modèle pour l'alignement nucléotique des cétacés et du groupe externe composé de l'éléphant d'Afrique et d'Asie.

Modèle	inv	k	shape
TIM3	0,57880	4	3,51927

Figure 36 : Construction de l'arbre *Neighbor-Joining* selon la méthode de TIM3+G(4)+I avec l'analyse *bootstrap* avec correction pour l'alignement nucléotique des cétacés et du groupe externe composé de l'éléphant d'Afrique et d'Asie.

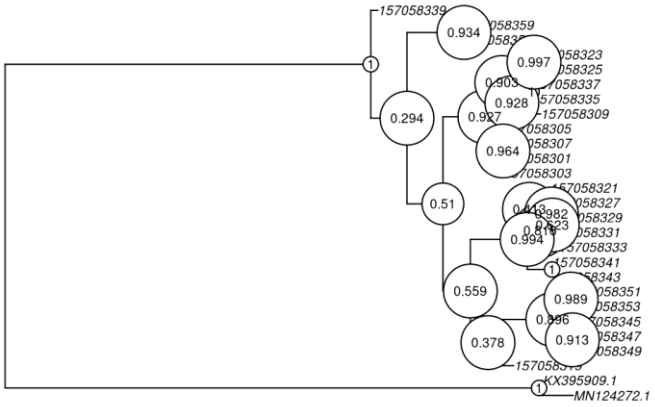

Tableau 29 : Les 5 meilleurs modèles selon le critère AIC pour l'alignement d'acide aminés des cétacés et du groupe externe composé de l'éléphant d'Afrique et d'Asie.

Modèles	AIC	BIC
MtZoa+I	1769,953	1949,183
MtZoa+G(4)	1770,108	1949,338
MtZoa+G(4)+I	1772,039	1954,716
mtmam+I	1777,097	1956,328
mtmam+G(4)	1777,097	1956,493

Tableau 30 : Les 5 meilleurs modèles selon le critère BIC pour l'alignement acide aminés des cétacés et du groupe externe composé de l'éléphant d'Afrique et d'Asie.

Modèles	BIC	AIC
MtZoa+I	1949,183	1769,953
MtZoa+G(4)	1949,338	1770,108
MtZoa+G(4)+I	1954,716	1772,039
mtmam+I	1956,328	1777,097
mtmam+G(4)	1956,493	1777,097

Tableau 31 : Paramètres du meilleur modèle pour l'alignement acide aminés des cétacés et du groupe externe composé de l'éléphant d'Afrique et d'Asie.

Modèle	inv
MtZoa	0,73435

Figure 37 : Construction de l'arbre *Neighbor-Joining* selon la méthode de MtZoa+I avec l'analyse *bootstrap* avec correction pour l'alignement acide aminés des cétacés et du groupe externe composé de l'éléphant d'Afrique et d'Asie.

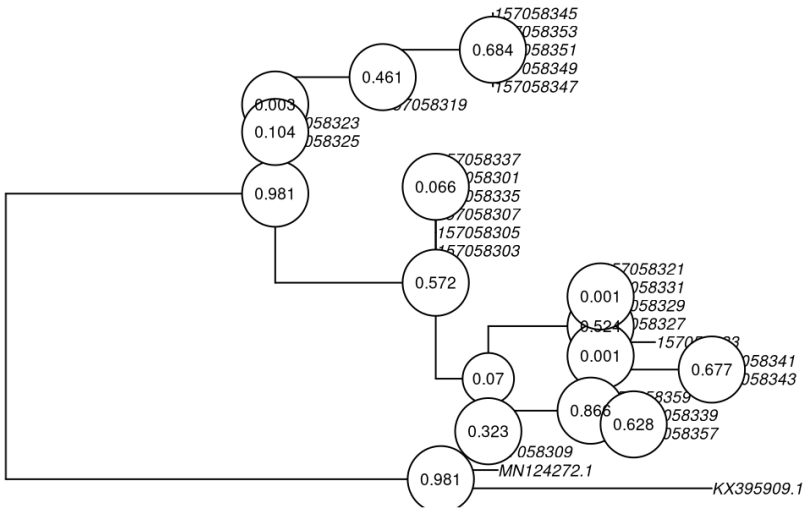

Tableau 32 : Les 5 meilleurs modèles selon le critère AIC pour l'alignement nucléotique des cétacés et du groupe externe composé de l'hippopotame et l'hippopotame nain.

Modèles	AIC	BIC
TPM3u+G(4)+I	57083,48	57530,41
TPM2u+G(4)+I	57084,21	57531,14
HKY+G(4)+I	57084,42	57523,64
TVM+G(4)+I	57084,71	57547,05
TIM3+G(4)+I	57085,45	57540,08

Tableau 33 : Les 5 meilleurs modèles selon le critère BIC pour l'alignement nucléotique des cétacés et du groupe externe composé de l'hippopotame et l'hippopotame nain.

Modèles	BIC	AIC
HKY+I	57519,72	57088,20
HKY+G(4)+I	57523,64	57084,42
TPM3u+I	57526,67	57084,42
TPM2u+I	57527,01	57087,79
TrN+I	57529,28	57090,06

Tableau 34 : Paramètres du meilleur modèle pour l'alignement nucléotique des cétacés et du groupe externe composé de l'hippopotame et l'hippopotame nain.

Modèle	inv	k	shape
HKY	0,61644	4	5,42279

Figure 38 : Construction de l'arbre *Neighbor-Joining* selon la méthode de TIM3+G(4)+I avec l'analyse *bootstrap* avec correction pour l'alignement nucléotique des cétacés et du groupe externe composé de l'hippopotame et l'hippopotame nain.

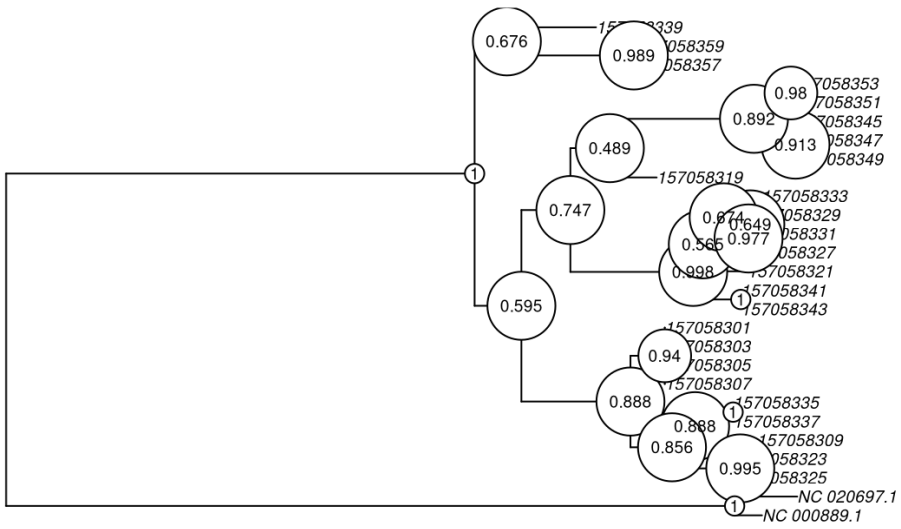

Tableau 35 : Les 5 meilleurs modèles selon le critère AIC pour l'alignement d'acide aminés des cétacés et du groupe externe composé de l'hippopotame et l'hippopotame nain.

Modèles	AIC	BIC
mtREV24+G(4)	58359,72	58721,36
mtREV24+G(4)+I	58363,04	58731,64
mtREV24+I	58417,81	58779,45
mtREV24	58616,57	58971,26
FLU+G(4)+I	58745,90	59112,22

Tableau 36 : Les 5 meilleurs modèles selon le critère BIC pour l'alignement acide aminés des cétacés et du groupe externe composé de l'hippopotame et l'hippopotame nain.

Modèles	BIC	AIC
mtREV24+G(4)	58721,36	58359,72
mtREV24+G(4)+I	58731,64	58363,04
mtREV24+I	58779,45	58417,81
mtREV24	58971,26	58616,57
FLU+G(4)+I	59112,22	58745,90

Tableau 37 : Paramètres du meilleur modèle pour l'alignement acide aminés des cétacés et du groupe externe composé de l'hippopotame et l'hippopotame nain.

Modèle	k	shape
mtREV24	4	0,56739

Figure 39 : Construction de l'arbre *Neighbor-Joining* selon la méthode de MtZoa+I avec l'analyse *bootstrap* avec correction pour l'alignement acide aminés des cétacés et du groupe externe composé de l'hippopotame et l'hippopotame nain.

Tableau 38 : Les 5 meilleurs modèles selon le critère AIC pour l'alignement nucléotique des cétacés et du groupe externe composé de la tortue luth et cistude.

Modèles	AIC	BIC
GTR+G(4)+I	7754,860	8041,754
TIM2+G(4)+I	7760,169	8037,657
TIM3+G(4)+I	7761,520	8039,008
TVM+G(4)+I	7766,360	8048,551
TrN+G(4)+I	7766,934	8039,719

Tableau 39 : Les 5 meilleurs modèles selon le critère BIC pour l'alignement nucléotique des cétacés et du groupe externe composé de la tortue luth et cistude.

Modèles	BIC	AIC
TIM2+G(4)+I	8037,657	7760,169
TIM3+G(4)+I	8039,008	7761,520
TrN+G(4)+I	8039,719	7766,934
GTR+G(4)+I	8041,754	7754,860
TPM3u+G(4)+I	8043,872	7771,087

Tableau 40 : Paramètres du meilleur modèle pour l'alignement nucléotique des cétacés et du groupe externe composé de la tortue luth et cistude.

Modèle	inv	k	shape
TIM2	0,57267	4	2,48824

Figure 40 : Construction de l'arbre *Neighbor-Joining* selon la méthode de TIM3+G(4)+I avec l'analyse *bootstrap* avec correction pour l'alignement nucléotique des cétacés et du groupe externe composé de la tortue luth et cistude.

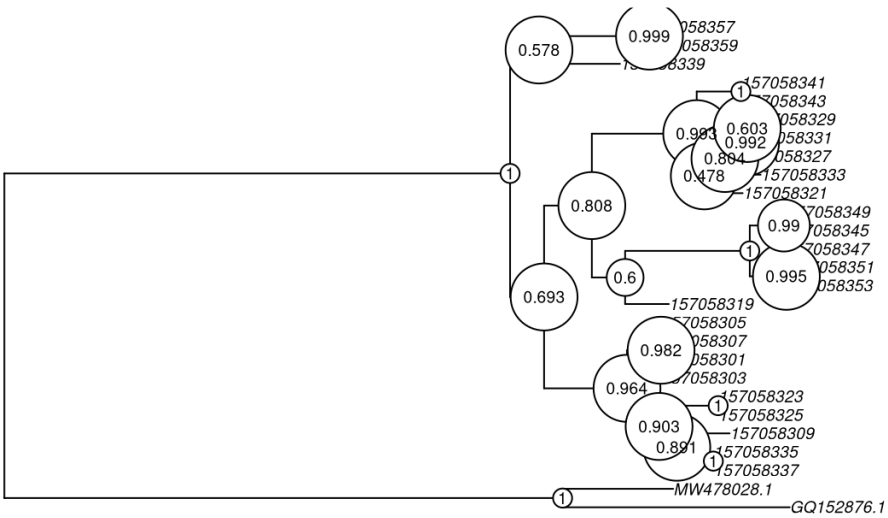

Tableau 41 : Les 5 meilleurs modèles selon le critère AIC pour l'alignement d'acide aminés des cétacés et du groupe externe composé de la tortue luth et cistude.

Modèles	AIC	BIC
MtZoa+I	1970,461	2156,998
MtZoa+G(4)	1970,548	2157,085
MtZoa+G(4)+I	1972,658	2162,782
mtArt+I	1980,827	2167,364
mtArt+G(4)	1980,937	2167,474

Tableau 42 : Les 5 meilleurs modèles selon le critère BIC pour l'alignement acide aminés des cétacés et du groupe externe composé de la tortue luth et cistude.

Modèles	BIC	AIC
MtZoa+I	2156,998	1970,461
MtZoa+G(4)	2157,085	1970,548
MtZoa+G(4)+I	2162,782	1972,658
mtArt+I	2167,364	1980,827
mtArt+G(4)	2167,474	1980,937

Tableau 43 : Paramètres du meilleur modèle pour l'alignement acide aminés des cétacés et du groupe externe composé de la tortue luth et cistude.

Modèle	inv
MtZoa	0,75357

Figure 41 : Construction de l'arbre *Neighbor-Joining* selon la méthode de MtZoa+I avec l'analyse *bootstrap* avec correction pour l'alignement acide aminés des cétacés et du groupe externe composé de la tortue luth et cistude.

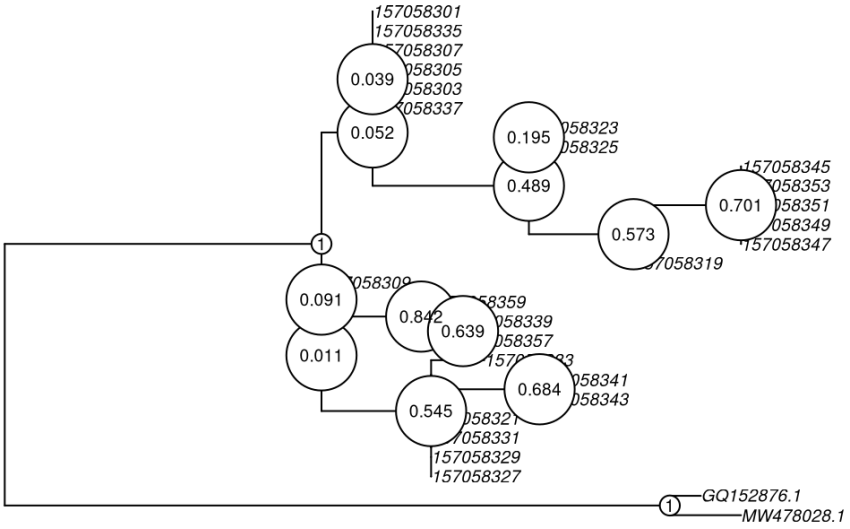

Tableau 44 : Les 5 meilleurs modèles selon le critère AIC pour l'alignement nucléotique des cétacés et des 4 groupes externes.

Modèles	AIC	BIC
TVM+G(4)+I	60602,44	61157,25
TPM3u+G(4)+I	60602,74	61142,14
GTR+G(4)+I	60603,76	61166,27
TPM2u+G(4)+I	60603,96	61143,36
TIM3+G(4)+I	60604,29	61151,39

Tableau 45 : Les 5 meilleurs modèles selon le critère BIC pour l'alignement nucléotique des cétacés et des 4 groupes externes.

Modèles	BIC	AIC
HKY+G(4)+I	61137,58	60605,89
HKY+G(4)	61140,50	60616,52
TPM3u+G(4)+I	61142,14	60602,74
TPM2u+G(4)+I	61143,36	60602,74
TPM3u+G(4)	61145,77	60614,08

Tableau 46 : Paramètres du meilleur modèle pour l'alignement nucléotique des cétacés et des 4 groupes externes.

Modèle	inv	k	shape
TPM3u	0,55275	4	1,18214

Figure 42 : Construction de l'arbre *Neighbor-Joining* selon la méthode de TIM3+G(4)+I avec l'analyse *bootstrap* avec correction pour l'alignement nucléotique des cétacés et des 4 groupes externes.

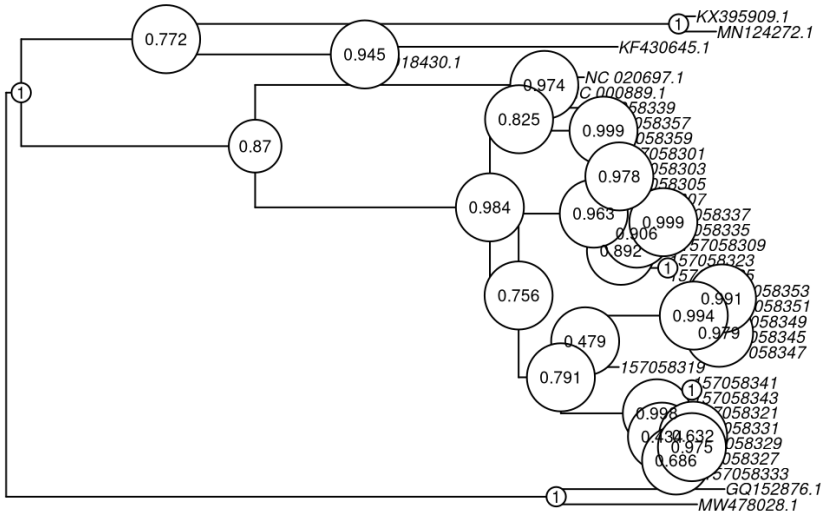

Tableau 47 : Les 5 meilleurs modèles selon le critère AIC pour l'alignement d'acide aminés des cétacés et des 4 groupes externes.

Modèles	AIC	BIC
mtREV24+G(4)	58678,12	59123,19
mtREV24+G(4)+I	58681,52	59133,54
mtREV24+I	58746,43	59191,50
mtREV24	58951,70	59389,81
FLU+G(4)+I	59058,09	59510,11

Tableau 48 : Les 5 meilleurs modèles selon le critère BIC pour l'alignement acide aminés des cétacés et des 4 groupes externes.

Modèles	BIC	AIC
mtREV24+G(4)	59123,19	58678,12
mtREV24+G(4)+I	59133,54	58681,52
mtREV24+I	59191,50	58746,43
mtREV24	59389,81	58951,70
FLU+G(4)+I	59510,11	59058,09

Tableau 49 : Paramètres du meilleur modèle pour l'alignement acide aminés des cétacés et des 4 groupes externes.

Modèle	k	shape
mtREV24	4	0,55330

Figure 43 : Construction de l'arbre *Neighbor-Joining* selon la méthode de MtZoa+I avec l'analyse *bootstrap* avec correction pour l'alignement acide aminés des cétacés et des 4 groupes externes.

Discussion

Section A

Pour l'alignement par défaut des séquences, l'algorithme applique une pénalité standard pour l'insertion de *gaps*, ce qui réduit la fragmentation de l'alignement et favorise les substitutions plutôt que des *indels* excessifs. On observe un alignement plus conservatif, avec une région bien alignée au début des séquences (**Figure 1**). Les *gaps* introduits sont regroupés en fin de séquences, là où leur présence est biologiquement plus probable.

En revanche, lorsque les pénalités pour l'ouverture et l'extension des *gaps* sont nulles, l'algorithme insère des *gaps* beaucoup plus librement. On observe alors des *gaps* dispersés un peu partout, ce qui conduit à un alignement plus fragmenté, avec de nombreux *gaps* placés au moindre désalignement (**Figure 2**). Cela peut réduire la fiabilité des analyses phylogénétiques. L'absence de pénalités pousse l'algorithme à aligner les séquences uniquement en fonction de leur similarité brute, ce qui peut masquer certaines relations évolutives.

Les pénalités de *gap* influencent donc fortement la structure de l'alignement. Un bon alignement repose sur un compromis entre la conservation des blocs homologues et l'introduction de gaps biologiquement pertinents. En phylogénie, il est crucial de bien choisir ces paramètres pour éviter un excès d'alignements artificiels dus à un relâchement trop important des contraintes sur les *gaps*.

Section B

Dans le cadre de l'analyse du gène COI, le choix du paramètre *GeneticCode* doit correspondre au type de matériel génétique étudié. L'option retenue est l'option 2, soit le code génétique de l'ADN mitochondrial des vertébrés. En effet, le gène COI (*cytochrome c oxydase I*) est un marqueur mitochondrial couramment utilisé pour l'identification des animaux. De plus, les différents jeux de données utilisés — incluant les cétacés, les hippopotames et les autres groupes externes mentionnés — concernent tous des vertébrés, c'est-à-dire des organismes possédant une colonne vertébrale.

Concernant la question de savoir si le gène COI est codant, la réponse est oui : il code pour une sous-unité de la protéine cytochrome c oxydase (COX), une enzyme clé de la chaîne de transport des électrons mitochondriale [19]. Cette enzyme joue un rôle essentiel dans la

production d'ATP, en réduisant l'oxygène et en pompant des protons à travers la membrane mitochondriale interne [19].

Section C

Le modèle d'évolution nucléotidique qui maximise la robustesse de la topologie, mesurée par les valeurs de *bootstrap* aux nœuds, semble être *TN93* ou *K80*. En effet, ces deux modèles produisent des arbres phylogénétiques très similaires (**Figures 12 et 13**), dans lesquels la majorité des nœuds présentent des scores de *bootstrap* supérieurs à 70 %, ce qui indique un bon niveau de confiance dans les regroupements observés. De plus, une forte corrélation a été observée entre les valeurs de *bootstrap* générées par *K80* et *TN93* (**Tableaux 2**), suggérant que ces deux modèles aboutissent à des topologies similaires et stables. Après avoir calculé la moyenne des scores de *bootstrap*, il apparaît que *TN93* est le meilleur modèle, bien que *K80* soit vraiment proche en termes de performance.

Le meilleur modèle, tel qu'identifié par la fonction *modelTest()* (**Tableaux 3 et 4**), est *TIM2+G(4)+I*, avec un paramètre *inv* d'environ 0,589, un paramètre *k* de 4, et un paramètre *shape* d'environ 2,767 (**Tableau 5**). Le modèle choisi n'est donc pas le même que celui qui maximise la robustesse de la topologie.

En comparant les deux arbres, on observe que, même si la structure globale est similaire, il existe plusieurs différences dans l'ordre des bifurcations. Les valeurs de *bootstrap* sont légèrement plus élevées pour le modèle *TN93*, qui maximise la robustesse, ce qui suggère une meilleure stabilité des relations évolutives.

Concernant la longueur des branches, l'arbre généré par le modèle sélectionné via *modelTest()* présente des branches plus longues et plus variées, reflétant une plus grande hétérogénéité des taux d'évolution. En revanche, l'arbre obtenu avec *TN93* (**Figure 12**) montre des branches plus courtes et plus homogènes, ce qui est souvent associé à une topologie plus robuste.

Pour résumer, bien que l'arbre proposé par *modelTest()* soit, en théorie, le plus adapté selon des critères statistiques d'ajustement, l'arbre généré avec le modèle *TN93* semble, en pratique, plus fiable. Avec ses valeurs de *bootstrap* globalement plus élevées et sa structure plus

homogène, il offre une meilleure robustesse, ce qui renforce la confiance dans sa topologie. Dans ce cas précis, il pourrait donc être considéré comme le meilleur arbre nucléotidique.

À partir du meilleur arbre nucléotidique retenu :

Dans l'Atelier 2, les modèles *K80* et *TN93* présentent des résultats similaires, tant au niveau des graphiques que des valeurs de *bootstrap*. Bien que ces modèles affichent un nombre plus important de nœuds, leurs valeurs de *bootstrap* sont globalement inférieures à celles du modèle sélectionné par *modelTest()*. Ce dernier, basé sur *GTR+G*, offre une topologie plus homogène et stable (**Figure 50**). Cette homogénéité renforce la fiabilité des relations évolutives inférées, suggérant que *GTR+G* serait le modèle le plus approprié pour cette analyse.

Pour comparer les deux jeux de données, le premier critère à examiner est le temps de divergence des taxons, estimé en comparant les longueurs de branche dans l'arbre phylogénétique. Les Cétacés (**Figures 12**) montrent des branches plus courtes, indiquant un temps de divergence plus récent ou une évolution plus lente. À l'inverse, les Bactéries (**Annexe 2 – Figure 44**) présentent des branches plus longues, suggérant une divergence plus rapide et une évolution plus dynamique.

Concernant la vitesse d'évolution, le gène *COI* des Cétacés évolue plus lentement en raison de la sélection purificatrice, qui réduit la variation génétique. Ce processus de sélection favorise la stabilité de la séquence et conduit à des branches plus courtes. En revanche, le gène *SQR* des Bactéries montre une évolution plus rapide, ce qui entraîne des branches plus longues et une plus grande variabilité génétique.

Les différences dans les modèles d'évolution choisis reflètent les caractéristiques de chaque jeu de données. Les modèles *TN93* pour les Cétacés sont adaptés à un gène moins variable et soumis à une sélection purificatrice, tandis que le modèle *GTR* pour les Bactéries prend en compte la plus grande variabilité et les taux de substitution plus rapides. Les résultats différents de *modelTest()* durant les deux travaux montrent que les meilleurs modèles peuvent différer significativement entre les deux types de données, soulignant ainsi l'importance de choisir un modèle d'évolution approprié à chaque situation.

Section D

L'analyse comparative des arbres phylogénétiques obtenus à partir des alignements traduits avec le cadre de lecture par défaut (**Figures 22, 23 et 24**) et un cadre de lecture modifié (cadre 1) (**Figures 25, 26 et 27**) révèle un impact significatif sur la topologie des arbres.

Avec le cadre de lecture par défaut, les arbres obtenus à partir des modèles *LG*, *JTT* et *Blosum62* montrent une topologie bien structurée, avec des regroupements évolutifs cohérents et des valeurs de *bootstrap* majoritairement élevées (souvent supérieures à 70-80 %). En revanche, l'utilisation du cadre 1 entraîne une dégradation importante de la topologie. En effet, plusieurs regroupements disparaissent, les séquences sont moins bien classées, et les valeurs de *bootstrap* diminuent fortement, traduisant une perte de robustesse.

Ces résultats confirment que le choix du cadre de lecture est crucial pour la traduction des séquences nucléotidiques en acides aminés. Un cadre incorrect fausse la séquence protéique, introduit du bruit dans l'alignement multiple et compromet la fiabilité de l'inférence phylogénétique.

Section E

Les vitesses d'évolution entre les différents cadres de lecture sont très différentes. Cela est particulièrement visible sur les graphiques réalisés avec correction (**Figures 31, 32 et 33**). Par exemple, pour le cadre 1 (**Figure 31**), on constate que le temps d'évolution du premier nœud est tellement élevé que les temps d'évolution des sous-nœuds deviennent négligeables et n'apparaissent pratiquement pas. Cette figure n'est pas une erreur, comme on peut le confirmer en consultant la version sans correction (**Figure 28**) où toutes les espèces sont considérées comme étant très éloignées les unes des autres. Seuls les échantillons provenant de la même espèce sont alors liés par un nœud.

Ainsi, le cadre de lecture 1 ne semble manifestement pas être le bon, puisque les figures produites ne correspondent pas du tout à la réalité. Il est donc inutile d'aller plus loin dans l'analyse avec ce cadre. En revanche, les deux autres cadres de lecture donnent des résultats plus similaires entre eux, tout en présentant suffisamment de différences pour rendre leur comparaison intéressante.

La première chose notable dans ces deux cadres est la présence du taxon *157058339*, isolé des autres taxons par un nœud affichant un très grand temps d'évolution dans les graphiques corrigés (**Figures 32 et 33)**. Ce taxon correspond à la baleine à bec de Stejneger, un cétacé de la famille des odontocètes. [23] On s'attendrait donc à ce qu'il soit proche d'autres espèces de cette famille. Pourtant, dans le cadre de lecture 3 (**Figure 33**), ses plus proches voisins sont le phoque gris (*5835009*) et le phoque annelé (*157058355*), deux pinnipèdes qui ne font pas partie des cétacés.

Même si les pinnipèdes ne sont pas très éloignés des cétacés sur le plan taxonomique, puisqu'ils font également partie des mammifères placentaires revenus à la vie aquatique, cette proximité reste surprenante. [22] On peut confirmer leur parenté en observant les autres pinnipèdes présents dans le jeu de données : ils sont tous regroupés sous les identifiants retenus du *Fasta, 157058311, 157058313, 157058315* et *157058317*. Cette observation révèle aussi une autre différence entre les cadres de lecture 2 et 3 : dans le cadre 2 (**Figure 32**), les pinnipèdes apparaissent tout en bas de l'arbre phylogénétique, associés au temps d'évolution le plus élevé, tandis que dans le cadre 3 (**Figure 33**), ils se retrouvent au début de l'arbre, avec peu de divergence depuis le nœud de séparation avec les cétacés.

On retrouve ensuite les deux sous-ordres des cétacés, les mysticètes et les odontocètes, qui se distinguent notamment par la présence de fanons chez les premiers et de dents chez les seconds. [21] Bien qu'il existe d'autres différences plus subtiles, celle-ci est la principale. Dans notre jeu de données, plusieurs espèces des deux sous-ordres sont présentes : principalement des baleines chez les mysticètes et des dauphins chez les odontocètes, avec quelques autres espèces. (**Annexe 3 – Figure 45**).

Dans les deux figures comparées (**Figures 32 et 33**), les deux sous-ordres sont généralement bien séparés. Cette séparation se retrouve notamment dans la littérature. En effet, Nikaido et al. 2001 ont également remarqué une séparation entre ces deux groupes lors de la construction de l'arbre phylogénétique. De plus, cette séparation est particulièrement visible dans le cadre 2 (**Figure 32**). Toutefois, elle n'est pas parfaite : dans le nœud supérieur, on retrouve uniquement des odontocètes (principalement des dauphins), tandis que dans le nœud inférieur, on observe la branche des pinnipèdes ainsi qu'un autre nœud qui se divise en un groupe de mysticètes vers le bas et d'odontocètes vers le haut, ceux-ci étant plus proches des phoques et des orques. Pour le cadre 3 (**Figure 33**), cette séparation est également visible au niveau du nœud de confiance *0,195*, où toutes les espèces de la branche inférieure sont des

mysticètes. Toutefois, en termes de vitesse d'évolution, elles apparaissent toutes au même niveau, et cela dans les deux figures.

Il est donc difficile de tirer des conclusions définitives à partir de la vitesse d'évolution et de la séparation des espèces, car les deux graphiques sont très similaires, avec des différences qui nécessiteraient une analyse plus précise, espèce par espèce.

On peut néanmoins conclure que la correction gamma améliore, dans certains aspects, la précision et la robustesse des graphiques, sans pour autant être systématiquement meilleure. Par exemple, elle permet de visualiser le temps d'évolution, ce qui est très utile. Toutefois, avec ou sans correction, les groupes supposés rester ensemble restent effectivement groupés. De plus, la confiance des nœuds est en moyenne plus élevée sans la correction gamma. Une approche pertinente pourrait donc consister à identifier les bons groupements à partir des figures sans correction, puis à analyser les temps d'évolution et les distances réelles entre espèces à l'aide des figures corrigées.

En conclusion, l'arbre obtenu avec le cadre de lecture 2 semble le plus proche du véritable patron de divergence entre les taxons. Le cadre de lecture 3 donne également de bons résultats, suffisamment convaincants pour ne pas être ignorés. En revanche, le cadre de lecture 1 s'éloigne trop de la réalité pour être exploité utilement. Pour ce qui est de la littérature récente entourant notre sujet, nous n'avons pas trouver de données assez bonnes pour les comparer avec nos résultats.

Section F

L'utilisation de groupes externes est une étape essentielle en phylogénie, car elle permet de raciner l'arbre et de donner une direction à l'évolution des lignées étudiées. Dans cette analyse, nous avons évalué l'impact de l'ajout de quatre paires de groupes externes (**Annexe 1 – Tableau 50**) au jeu de données du gène mitochondrial *COI*, en les testant d'abord séparément, puis tous ensemble : les Siréniens (*Lamantin Trichechus manatus* et *Dugong Dugong dugon*), les Éléphants (d'Afrique *Loxodonta africana* et d'Asie *Elephas maximus*), les Hippopotames (commun *Hippopotamus amphibius* et nain *Hexaprotodon liberiensis*), ainsi que les Reptiles (*Tortue luth Dermochelys coriacea* et *Cistude Emys orbicularis*).

L'ajout des Siréniens a donné une topologie stable et cohérente (**Figure 34**), avec comme modèle optimal pour l'alignement nucléotidique *TIM3+G(4)+I* (**Tableau 22**),

affichant une valeur *AIC* de 7301,60 (**Tableau 20**) et une valeur *BIC* de 7566,47 (**Tableau 21**). Pour les acides aminés, le modèle *MtZoa+I* a été retenu avec *AIC* = 1588,19 et *BIC* = 1763,46 (**Tableaux 23-24**). L'arbre obtenu **(Figure 35**) est bien raciné, les branches sont équilibrées et la monophylie des Cétacés reste intacte.

L'inclusion des Éléphants a également permis un racinage correct, malgré une augmentation de la longueur des branches externes due à leur éloignement évolutif (**Figures 36-37**). Le meilleur modèle en ADN est *TIM3+G(4)+I* (**Tableau 28**), avec *AIC* = 7224,22 (**Tableau 26**) et *BIC* = 7493,65 (**Tableau 27**), et pour les AA, *MtZoa+I* (**Tableau 31**) (*AIC* = 1769,95, *BIC* = 1949,18) (**Tableaux 29-30**).

En revanche, les Hippopotames ne peuvent pas remplir le rôle de groupe externe, car ils forment un groupe frère direct des Cétacés. Leur inclusion empêche un bon enracinement de l'arbre (**Figure 38**). Le modèle optimal pour l'ADN dans ce cas est *TIM3+G(4)+I* (**Tableau 34**), avec *AIC* = 57085,45 (**Tableau 32**) et *BIC* = 57540,08 (**Tableau 33**), tandis que pour les AA, le modèle *mtREV24+G(4)* (**Tableau 37**) s'est avéré le plus approprié (*AIC* = 58359,72, *BIC* = 58721,36) (**Tableaux 35-36**), bien que l'arbre soit mal raciné et peu interprétable (**Figure 39**).

L'ajout des Reptiles, trop éloignés des Cétacés, provoque un phénomène de « longue branche » qui déséquilibre l'arbre (**Figures 40-41**). Le modèle ADN *TIM2+G(4)+I* (**Tableau 40**) affiche un *AIC* de 7760,17 (**Tableau 38**) et un *BIC* de 8037,66 (**Tableau 39**), alors que pour les AA, *MtZoa+I* (**Tableau 43**) reste le modèle optimal avec *AIC* = 1970,46 et *BIC* = 2156,99 (**Tableaux 41-42**).

Enfin, l'ajout simultané des quatre paires de groupes externes permet un enracinement multiple, mais augmente la complexité et la longueur des branches, notamment celles des Reptiles (**Figures 43-44**). Le modèle optimal en ADN devient *TPM3u+G(4)+I* (**Tableau 46**) avec *AIC* = 60602,74 (**Tableau 44**) et *BIC* = 61142,14 (**Tableau 45**), tandis que pour les acides aminés, *mtREV24+G(4)* (**Tableau 49**) est retenu (*AIC* = 58678,12, *BIC* = 59123,19) (**Tableaux 47-48**).

Au terme de cette analyse, il apparaît clairement que le choix du groupe externe a une influence déterminante sur la topologie et la fiabilité des arbres phylogénétiques. Un bon groupe externe doit être extérieur au clade étudié, présenter une distance évolutive suffisante mais pas excessive, éviter la création de longues branches susceptibles de fausser les relations internes, et permettre un enracinement clair de l'arbre. Parmi les groupes testés, les Siréniens

se sont révélés être les plus appropriés, car ils sont suffisamment proches des Cétacés pour fournir une racine informative, tout en étant clairement distincts. Les Éléphants peuvent également être utilisés avec prudence, car leur inclusion ne perturbe pas significativement la structure de l'arbre. Les Hippopotames, en revanche, ne remplissent pas les conditions d'un vrai groupe externe et nuisent à l'interprétation des relations évolutives, tandis que les Reptiles, par leur éloignement extrême, engendrent des artefacts topologiques comme l'allongement excessif des branches. Ces observations soulignent l'importance de sélectionner avec rigueur les groupes externes, car leur influence peut affecter de manière significative les inférences phylogénétiques et la compréhension des liens évolutifs entre les espèces.

Section G

Tout d'abord, le *DNA barcoding* est une méthode utilisée pour identifier une espèce de manière rapide et efficace à partir de son ADN [18]. Cette technique repose sur l'analyse de petites séquences standardisées, appelées *barcodes*, qui sont présentes chez la plupart des espèces [18]. Pour qu'une région génétique soit considérée comme un code-barre, elle doit répondre à certains critères [18]. En effet, elle doit présenter une bonne variabilité génétique entre les espèces, tout en étant conservée aux extrémités pour permettre la création d'amorces PCR universelles [18]. De plus, la séquence doit être courte afin de faciliter l'extraction et l'amplification de l'ADN, même à partir d'échantillons dégradés [18].

En ce qui concerne le gène *COI (cytochrome c oxydase I*), c'est un code-barre utilisé principalement chez les animaux [19]. Il se situe dans l'ADN mitochondrial, ce qui offre plusieurs avantages : il y a plusieurs copies par cellule, ce qui facilite son extraction, et il évolue d'une manière particulière [19]. Bien que ce segment code pour la protéine *COX*, l'enzyme clé de la chaîne de transport des électrons mitochondriaux, l'ADN mitochondrial est soumis à une sélection purificatrice (c'est-à-dire qu'il conserve surtout les mutations bénéfiques) [19]. Le gène *COI* présente donc des variations suffisantes entre les espèces pour les distinguer, tout en étant suffisamment stable pour ne pas perturber la structure et la fonction de la protéine [19]. Cela en fait un marqueur idéal pour le *DNA barcoding*, car il permet de différencier efficacement les espèces tout en restant stable au sein d'une même espèce [19]. Enfin, le gène *COI* est largement représenté dans les bases de données comme *BOLD* [19] et *NCBI*, ce qui permet une comparaison efficace entre espèces.

Conclusion

Ce travail avait pour objectif de construire un arbre phylogénétique fiable des cétacés, suffisamment robuste pour être comparé à d'autres groupes externes et avoir du sens. Pour cela, de nombreuses sous-étapes étaient nécessaires afin de s'assurer que chacun des choix méthodologiques était pertinent. Par exemple, l'alignement des séquences est une étape cruciale, où le choix du modèle et des paramètres est très important. Dans le cas présent, il s'agissait d'un gène codant issu de l'ADN mitochondrial de vertébrés.

Pour déterminer le meilleur modèle, plusieurs techniques ont été utilisées. La fonction *modelTest()* s'est révélée efficace pour identifier le modèle le plus adapté. Les matrices de distance générées à l'aide des bibliothèques *ape* et *phangorn* ont également été exploitées. Des analyses de *bootstrap* ont été menées, car elles sont très utiles pour évaluer rapidement la robustesse de plusieurs modèles. Il n'existe donc pas une seule bonne réponse pour déterminer quel modèle est le meilleur : cela dépend du contexte. Cependant, selon les résultats obtenus ici, le modèle *TN93* a été un bon choix pour les analyses simples, tandis que *TIM2* et *TIM3* se sont démarqués lors des analyses plus complexes, notamment avec l'ajout de corrections gamma et d'un site invariant.

Le cadre de lecture est également un aspect important. Lors de la comparaison des trois cadres de lecture possibles, nous avons réalisé que le premier était inutilisable, tandis que les deux autres donnaient des résultats différents. Cependant, il est difficile de conclure quel est le meilleur choix. Enfin, le choix du groupe externe est une étape importante. Les résultats montrent que les siréniens et les éléphants étaient les meilleurs choix.

Bibliographie

[1] Hebert, P. D. N., Ratnasingham, S., & deWaard, J. R. (2003). *Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species*. **Proceedings of the Royal Society of London. Series B: Biological Sciences, 270**(Suppl 1), S96–S99. https://doi.org/10.1098/rsbl.2003.0025

[2] Gatesy, J., Geisler, J. H., Chang, J., Buell, C., Berta, A., Meredith, R. W., Springer, M. S., & McGowen, M. R. (2013). *A phylogenetic blueprint for a modern whale*. **Molecular Phylogenetics and Evolution, 66**(2), 479–506. https://doi.org/10.1016/j.ympev.2012.10.012

[3] Tamura, K., Nei, M., & Kumar, S. (2004). *Prospects for inferring very large phylogenies by using the neighbor-joining method*. **PNAS, 101**(30), 11030–11035. https://doi.org/10.1073/pnas.0404206101

[4] Posada, D., & Buckley, T. R. (2004). *Model selection and model averaging in phylogenetics: advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests*. **Systematic Biology, 53**(5), 793–808. https://doi.org/10.1080/10635150490522304

[5] Wiens, J. J., & Tiu, J. (2012). *Highly incomplete taxa can rescue phylogenetic analyses from the negative impacts of limited taxon sampling*. **PLoS ONE, 7**(8), e42925. https://doi.org/10.1371/journal.pone.0042925

[6] R Core Team (2025). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/

[7] Wright, E. S. (2016). *Using DECIPHER v2.0 to Analyze Big Biological Sequence Data in R*. **The R Journal, 8**(1), 352-359.

[8] Paradis, E., & Schliep, K. (2019). *ape 5. 0: An environment for modern phylogenetics and evolutionary analyses in R*. **Bioinformatics, 35**(3), 526–528. https://doi.org/10.1093/bioinformatics/bty633

[9] Jukes TH, Cantor CR (1969). *Evolution of Protein Molecules.* New York: Academic Press. pp. 21–132.

[10] Kimura, M. (1980). *A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences*. **Journal of Molecular Evolution, 16**(2), 111–120. https://doi.org/10.1007/BF01731581

[11] Tamura, K., & Nei, M. (1993). *Estimation of the number of nucleotide substitutions in the control region of mitochondrial DNA in humans and chimpanzees*. **Molecular Biology and Evolution, 10**(3), 512–526. https://doi.org/10.1093/oxfordjournals.molbev.a040023

[12] Galtier, N., & Gouy, M. (1995). *Inferring phylogenies from DNA sequences of unequal base compositions*. **Proceedings of the National Academy of Sciences of the United States of America, 92**(24), 11317–11321. https://doi.org/10.1073/pnas.92.24.11317

[13] Schliep, K. P. (2011). *phangorn: phylogenetic analysis in R*. **Bioinformatics, 27**(4), 592-593. https://doi.org/10.1093/bioinformatics/btq706

[14] Schliep, K., Potts, A. J., Morrison, D. A., & Grimm, G. W. (2017). *Intertwining phylogenetic trees and networks*. **Methods in Ecology and Evolution, 8**(10), 1212–1220. https://doi.org/10.1111/2041-210X.12760

[15] Le, S. Q., & Gascuel, O. (2008). *An improved general amino acid replacement matrix*. **Molecular Biology and Evolution, 25**(7), 1307–1320. https://doi.org/10.1093/molbev/msn067

[16] Jones, D. T., Taylor, W. R., & Thornton, J. M. (1992). *The rapid generation of mutation data matrices from protein sequences*. **Bioinformatics, 8**(3), 275–282. https://doi.org/10.1093/bioinformatics/8.3.275

[17] Henikoff, S., & Henikoff, J. G. (1992). *Amino acid substitution matrices from protein blocks*. **Proceedings of the National Academy of Sciences, 89**(22), 10915–10919. https://doi.org/10.1073/pnas.89.22.10915

[18] Kress, W. J., & Erickson, D. L. (2008). DNA barcodes : Genes, genomics, and bioinformatics. *Proceedings of the National Academy of Sciences*, *105*(8), 2761-2762. https://doi.org/10.1073/pnas.0800476105

[19] Pentinsaari, M., Salmela, H., Mutanen, M., & Roslin, T. (2016). Molecular evolution of a widely-adopted taxonomic marker (COI) across the animal tree of life. *Scientific Reports*, *6*(1), 35275. https://doi.org/10.1038/srep35275

[20] Baleine en direct; (2021). CÉTACÉ, PINNIPÈDE, RORQUAL, MAMMIFÈRE MARIN… C'EST LA MÊME CHOSE? https://baleinesendirect.org/cetace-baleine-rorqual-cest-la-meme-chose

[21] Wikipédia; (2025). Cétacé. https://fr.wikipedia.org/wiki/Cetacea

[22] Wikipédia; (2025). Pinniped. https://en.wikipedia.org/wiki/Pinniped

[23] Wikipédia; (2023). Baleine à bec de Stejneger https://fr.wikipedia.org/wiki/Baleine_%C3%A0_bec_de_Stejneger

[24] Nikaido, M., Matsuno, F., Hamilton, H., Brownell, R. L., Cao, Y., Ding, W., Zuoyan, Z., Shedlock, A. M., Fordyce, R. E., Hasegawa, M., & Okada, N. (2001). Retroposon analysis of major cetacean lineages : The monophyly of toothed whales and the paraphyly of river

dolphins. Proceedings of the National Academy of Sciences, 98(13), 7384-7389. https://doi.org/10.1073/pnas.121139198

Annexe 1 : Données utilisés pour les analyses

Tableau 50 : Listes des animaux utilisés pour les différents groupes externes pour répondre a la question f).

ID NCBI	Nom scientifique	Nom courant
HQ918430.1	Trichechus manatus	Lamantin des Caraïbes
KF430645.1	Dugong dugon	Dugong
MN124272.1	Loxodonta africana	Éléphant d'Afrique
KX395909.1	Elephas maximus	Éléphant d'Asie
NC_000889.1	Hippopotamus amphibius	Hippopotame
NC_020697.1	Hexaprotodon liberiensis	Hippopotame nain
GQ152876.1	Dermochelys coriacea	Tortue luth
MW478028.1	Emys orbicularis	Cistude

Annexe 2 : Résultats de l'atelier 2

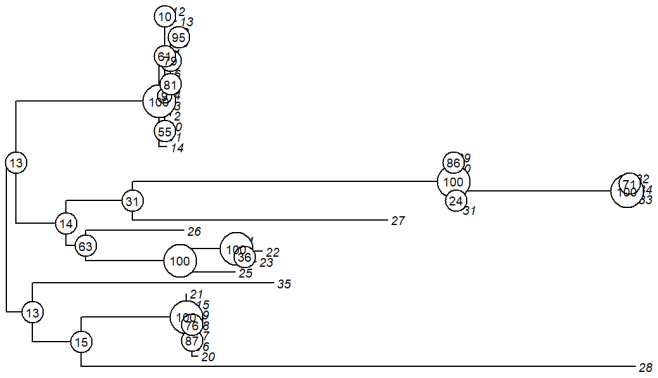

Figure 44 – Meilleur arbre nucléotidique de l'atelier 2 selon le modèle GTR avec correction.

Annexe 3 : Arbre phylogénique modèle des cétacés

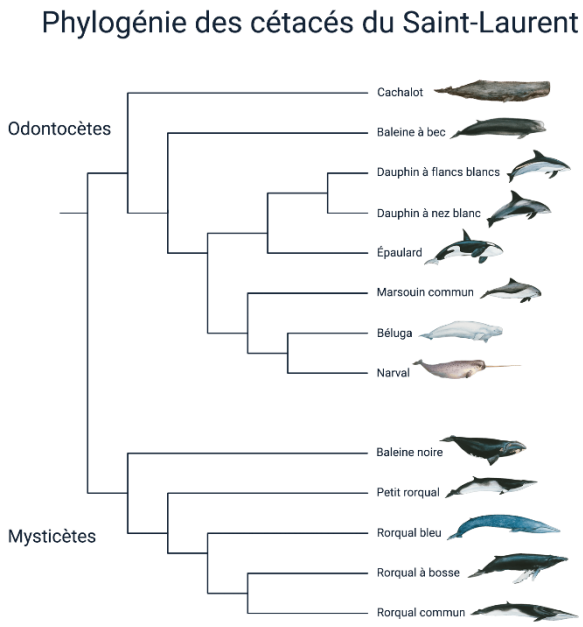

Figure 45 – Arbre phylogénique modèle des cétacés du Saint-Laurent [20]