Université Lille 1-MGM-Evolution Moléculaire-2005
Vincent CASTRIC, Xavier VEKEMANS

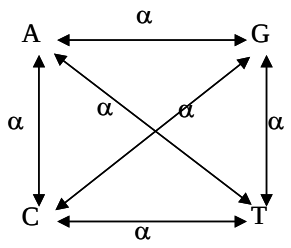
Maîtrise de Génétique et Microbiologie
TD 1
Corrigé

I) Evolution des séquences nucléotidiques

1) En supposant que se produisent 5.10^{-9} substitutions par site par année, calculer sous le modèle de Jukes & Cantor la proportion attendue de différences nucléotidiques entre deux séquences ayant divergé il y a 1, 10, 101 et 110 millions d'années. Quel phénomène mettez-vous ainsi en évidence ?

proportion attendue de différences nucléotidiques entre deux séquences = probabilité que deux séquences diffèrent à un site donné en X myrs. = 2 proba de changer d'ETAT le long d'une branche de X myrs.*
Hypothèse: même probabilité de substitution d'une base par n'importe laquelle des 3 autres

- *α = taux de substitution nucléotidique dans une direction = nombre attendu de substitutions d'1 nucléot. donné par 1 autre nucléot. donné par unité de temps*
- *3 α = taux de substitution nucléotidique total = nombre attendu de substitution d'1 nucléot. donné par n'importe quel autre nucléotide par unité de temps*



On aurait tendance à écrire :

*5.10^{-9} substitutions par site par année * 1, 10, 101 et 110 millions d'années * 2 ….*
Mais ca ne prend pas en compte les substitutions multiples.

Modèle de Jukes et Cantor : au fur et à mesure que le temps de divergence entre deux séquences augmente, la probabilité qu'une seconde substitution se produise au même site cesse d'etre négligeable, ce qui ralentit l'augmentation du nombre de différences réellement observé. Cette propriété est indésirable pour une mesure de distance.
Soit P(t)= proba d'etre une base donnée au temps t.

$$P(t+1)=P(t)-3\,\alpha\,P(t)+\alpha\,(1-P(t))$$

$$P(t+1)-P(t)=-3\,\alpha\,P(t)+\alpha\,(1-P(t))$$

Approx temps continu :

$$dP/dt=-3\,\alpha\,P(t)+\alpha\,(1-P(t))$$

$$dP/dt=-3\,\alpha\,P(t)+\alpha-\alpha\,P(t)$$

$$dP/dt=\alpha-4\,\alpha\,P(t)$$

La solution de cette équation différentielle est :

$$P(t)=x\exp(-4\,\alpha\,t)+1/4$$

Si on s'intéresse à probabilité que ca reste le même nucléotide, P(0)=1 et x=3/4

$$\mathbf{P(t)pas\ de\ changement = 3/4\ exp^{(-4\,\alpha\,t)} +1/4}$$

$$\mathbf{P(t)\ changement = 1-P(t)pas\ de\ changement =3/4.(1-exp^{(-4\,\alpha\,t)})}$$

Probabilité qu'il y ait eu un changement depuis les t générations écoulées le long des DEUX lignées de séquence :

$$\mathbf{P(2t)\ changement =3/4.(1-exp^{(-8\,\alpha\,t)})}$$

taux substitution	temps div (millions années)	Proportion de sites différents entre les deux séquences
5,00E-09	1	0,029
5,00E-09	10	0,247
5,00E-09	101	0,737
5,00E-09	118	0,741

On observe que la divergence atteint une limite, qui se situe à 0.75. Apres un grand nombre de substitutions, il n'y a en effet plus de corrélation entre l'etat de départ et l'etat final, la distribution des nucléotides tend donc vers sa fréquence à l'équilibre : 0.25 de chaque.

2) En supposant que se produisent 10.10^{-9} transitions et $2{,}5.10^{-9}$ transversions par site par année, quelles sont, sous le modèle de Kimura à deux paramètres, P(t) et Q(t), les proportions attendues de différences par transition et par transversion, respectivement, entre deux séquences qui ont divergé il y a 10, 100, 110 et 1100 millions d'années ?

$$I_{(t)} = \frac{1}{4} + \frac{1}{4}e^{-8\beta t} + \frac{1}{2}e^{-4(\alpha+\beta)t} \qquad P(t) = \frac{1}{4} + \frac{1}{4}e^{-8\beta t} - \frac{1}{2}e^{-4(\alpha+\beta)t} \qquad Q(t) = \frac{1}{2} - \frac{1}{2}e^{-8\beta t}$$

transition= purine→ purine ou pirimidine → pirimidine = 10.10^{-9} transitions
tranversion = inter =$2{,}5.10^{-9}$ transversions

	I(t),	*P(t),*	*Q(t)*
1 million d'années	*0,97066*	*0,01943*	*0,0099*
10 million d'années	*0,75795*	*0,15142*	*0,09063*
101 million d'années	*0,28637*	*0,27996*	*0,43367*
110 million d'années	*0,27974*	*0,27566*	*0,4446*

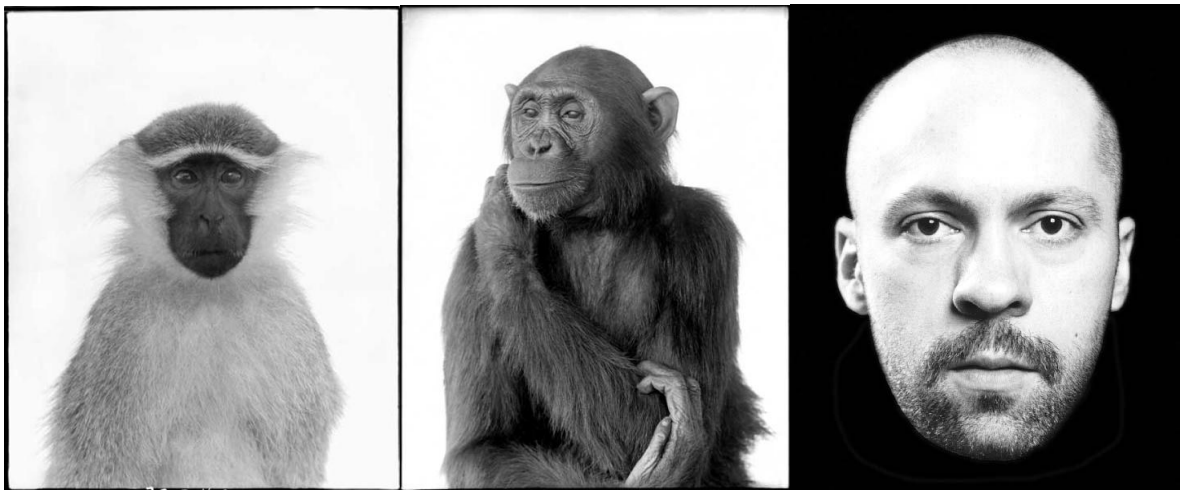
3) Montrer que, sous certaines conditions, le modèle de Kimura à deux paramètres se réduit au modèle de Jukes & Cantor dans le contexte de l'estimation des différences nucléotidiques entre deux séquences.

$$\begin{aligned}
&\frac{1}{4} + \frac{1}{4}e^{-8\beta t} + \frac{1}{2}e^{-4(\alpha+\beta)t} \\
&= \frac{1}{4} + \frac{1}{4}e^{-8\alpha t} + \frac{1}{2}e^{-4(\alpha+\alpha)t} \\
&= \frac{1}{4} + \frac{1}{4}e^{-8\alpha t} + \frac{1}{2}e^{-8\alpha t} \\
&= \frac{1}{4} + \frac{3}{4}e^{-8\alpha t}
\end{aligned}$$

Les deux modèles sont donc équivalents lorsque le taux de transition est égal à celui de transversion. Le modèle de J&C est donc d'autant plus faux que α est différent de β.

II) Estimation du nombre de substitutions nucléotidiques

1) a. En utilisant la Figure1, estimer le nombre de substitutions nucléotidiques par site (K) pour les séquences du gène de l'insuline chez l'humain et le singe vert en utilisant :



1. la méthode de Jukes et Cantor

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

comparaison	Nb de sites	Nb de différences	Proportion de sites différents (p)	K
homo-pan	177	2	0,01130	0,01138
homo-cerco	173	14	0,08092	0,08563
pan-cerco	175	13	0,07429	0,07822

2. la méthode de Kimura à 2 paramètres.

		tv	ti	Q	P	K
homo-pan	177	2	0	0,01130	0,00000	0,01140
homo-cerco	173	5	9	0,02890	0,05202	0,08621
pan-cerco	175	3	10	0,01714	0,05714	0,07917

$$K = \frac{1}{2}\ln\left(\frac{1}{1-2P-Q}\right)+\frac{1}{4}\ln\left(\frac{1}{1-2Q}\right)$$

1) b. Même question pour l'humain et le chimpanzé.

2) La date de divergence entre homme et chimpanzé est d'environ 5,5 millions d'années. Estimer le taux global de substitutions pour ces deux espèces.

Taux de substitution nucléotidique:
r = nombre de substitutions par site nucléotidique et par an (=3α)
= K/(2T) = 0,01140/ (2*5.5*10^6) = ***1,036.10^{-9}***

Estimer alors la date de divergence entre homme et singe vert. Quelle hypothèse faites-vous pour réaliser ce dernier calcul ?

r= K/(2T)
*→ T=K/(2*r) = 0,08621/(2*1,036.10$^{-9)}$* **=41,607 10^{6}**
Hypothèse nécessaire: les taux d'évolution sont constants le long de toutes les lignées.

3) Les données fossiles donnent une date de divergence de 20 millions d'années entre Hominidae et Cercopithecidae. Quelles hypothèses pouvez-vous avancer pour expliquer la différence avec vos résultats en 2) ?

On sur-estime la date de divergence. Plusieurs explications possibles :
Le modèle de substitutions utilisé n'est pas approprié, et notre ratio ti/tv n'est pas le bon.
Les taux d'évolution ne sont pas constants et se sont accéléré le long de la lignée qui mène au singe vert, ou se sont ralentis le long de la branche qui mène à l'homme. C'est une observation générale : ralentissement chez les hominoideae parmi les primates (Steiper et al. 2004 PNAS), d'un facteur 1.45.

- evolution des taux de mutations ?
- changement des pressions de sélection appliquées à cette protéine (nouvelle fonction acquise pour le gène de l'insuline ?)
- On pourrait aussi mettre en cause les données fossiles qui peuvent révéler quand une espèce est présente, mais ne peut pas dater jusqu'à quand une espèce n'est pas présente…

Figure 1. Alignement des séquences des premiers introns du gène de l'insuline chez l'humain, le singe vert (*Cercopithecus aethiops*), et le chimpanzé (*Pan troglodytes*) Notation : – indique une délétion.

```
Homo          GTCTGTTCCAAGGGCCTTTGCGTCAGGTGGGCTCAGGG------------------CCAGGGTGG
Pan           GTCTGTTCCAAGGGCCTTTGCGTCAGGTGGGCTCAGGGTT----------------CCAGGGTGG
Cercopithecus GTCTGTTCCAAGGGCCTTCGCGTCAGGTGGGCTCAGGGCTGC-CCACTTGGGGGTTCCAGGGTGG

Homo          CTGGACCCCAGGCCCCAGCTCTGCAGCAGGGAGGACGTGGCTGGGCTCGTGAAGCATGTGGGGGT
Pan           CTGGACCCCAGGCCCCAGCTCTGCAGCAGGGAGGACGTGGCTGGGCTCTTGAAGCATGTGGGGGT
Cercopithecus CTGGACCCCAGGCCCCAGCTCTGCAACAGGGAGGACATGGCTGGGCTCTTGAAGCGTTGAGGGT

Homo          GAGCCCAGGGGCCCCAAGGCAGGGCACCTGGCCTTCAGCCTGCCTCAGCCCTGCCTGTCTCCCAG
Pan           GAGCCCAGGGGCCCCAAGGCAGGGCACCTGGCCTTCAGCCGGCCTCAGCCCTGCCTGTCTCCCAG
Cercopithecus GAACCCAGGGGCCC-AGGGCAG-GCACCTGGCC-TCAGCTGGCCTCAGG-CTGCCTGTCTCTCAG
```

III) Patrons de substitution nucléotidique

1) Pour quel acide aminé code le codon ACT ?

1. ACT code pour Thr

2) Calculer la **proportion de changements synonymes** parmi l'ensemble des mutations possibles du codon ACT n'impliquant qu'un seul changement nucléotidique.

a) Sous l'hypothèse que la mutation est aléatoire (c.a.d. qu'il n'y a pas de différence de taux de mutation d'un nucléotide vers un autre)

b) Sous l'hypothèse que le patron de mutation se conforme à celui qui est observé dans les pseudogènes chez les Mammifères (Tableau 1).

De↓... vers →	A	T	C	G	Total
A	-	5.3	5.6	10.3	21.2
T	4.8	-	9.2	3.6	17.6
C	7.1	18.2	-	4.2	29.5
G	18.6	7.7	5.5	-	31.8
Total	30.5	31.2	20.3	18.1	

Tableau 1. Patron de substitution nucléotidique dans les pseudogènes (en pourcentages)

c) Sous l'hypothèse que le patron de mutation se conforme à celui qui est observé dans la région de contrôle de l'ADN mitochondrial chez l'humain (Tableau 2).

De↓... vers →	A	T	C	G	Total
A	**-**	**0.4**	**1.1**	**14.1**	**15.6**
T	**0.3**	**-**	**33.8**	**0.3**	**34.4**
C	**1.1**	**25.8**	**-**	**0.5**	**27.4**
G	**20.0**	**1.1**	**1.6**	**-**	**22.7**
Total	**21.4**	**27.3**	**36.5**	**14.9**	

Tableau 2. Patron de substitution nucléotidique dans l'ADN mitochondrial humain (en pourcentages)

3) Mêmes questions pour le codon GAT.

2. Suite à un changement nucléotidique, ACT peut muter en
CCT, TCT, GCT, AGT, ATT, AAT, ACA, ACC, ACG, dont seuls les trois derniers résultent en des changements synonymes. Donc 3/9=33% des mutations possibles du codon ACT sont synonymes.

A	C	T	Thr	Aléatoire	pseudogène		ADNmt	
C	C	T	Pro		5,6		1,1	
T	C	T	Ser		5,3		0,4	
G	C	T	Ala		10,3		14,1	
A	G	T	Ser		4,2		0,5	
A	T	T	Ile		18,2		25,8	
A	A	T	Asn		7,1		1,1	
A	C	A	Thr		4,8	0,07	0,3	0,00
A	C	C	Thr		9,2	0,13	33,8	0,44
A	C	G	Thr		3,6	0,05	0,3	0,00
					68,3	0,26	77,4	0,44

Ne pas oublier de pondérer les taux de mutation par la tendance générale du nucléotide à muter (0.07=4.8/68.3).
Dans ce cas, si la mutation affecte la troisième base, il s'agit d'un T, qui est le nucléotide qui a le moins tendance à muter dans les pseudogènes.
A l'inverse, ce même nucléotide est le plus mutable dans l'ADNmt : ceci explique la différence entre 26 % et 44%.
Les valeurs 26 % et 44% sont de part et d'autre de 33% (la valeur sous mutation aléatoire) parce que les changements synonymes (ici, troisième base) sont tributaire du changement d'un T en un autre nucléotide, qui se produit respectivement moins et plus souvent que dans 25% des cas.

Codon GAT :

G	A	T	Asp	pseudogène		ADNmt	
A	A	T	asn	19		20	
C	A	T	his	5,5		1,6	
T	A	T	tyr	7,7		1,1	
G	C	T	ala	5,6		1,1	
G	G	T	gly	10		14,1	
G	T	T	val	5,3		0,4	
G	A	A	glu	4,8		0,3	
G	A	C	asp	9,2	0,13	33,8	0,5
G	A	G	glu	3,6		0,3	
				70,7	0,13	72,7	0,5

Dans ce cas, la mutation synonyme implique un changement de T vers C, qui est la plus probable dans l'ADNmt.
Si la valeur dans l'ADNmt reste très similaire, on remarque cette fois une forte différence avec dans les peudogènes.

Conclusion :
Les codons varient dans leur propension à donner des changements synonymes
Les régions de l'ADN varient dans leur taux de changements synonymes, meme pour un seul codon.
Il faut donc se méfier des modèles qui assument des t^x de mutations homogènes. On pourrait complexifier les modèles à l'infini pour prendre en compte ces variations.

Maîtrise de Génétique et Microbiologie
TD 2

IV) Polymorphisme nucléotidique au locus MHC

1) Chaque codon comporte trois sites nucléotidiques. Les mutations sur certains sites entraînent des changements synonymes, d'autres des changements non synonymes. Calculer, sous le modèle de Jukes et Cantor, le nombre de sites potentiellement synonymes au codon CGA.

CGA :

C	*G*	*A*	*Arg*
A	*G*	*A*	*Arg*
G	*G*	*A*	*Gly*
T	*G*	*A*	*STOP*
C	*A*	*A*	*Gln*
C	*C*	*A*	*Pro*
C	*T*	*A*	*Leu*
C	*G*	*C*	*Arg*
C	*G*	*G*	*Arg*
C	*G*	*T*	*Arg*

1,5

Le premier site du codon est synonyme avec une probabilité d'1/2, le second avec une proba de 0, le troisième avec une proba de 100%.
→ Au total, CGA comporte 1.5 sites synonymes sur les 3.
Pourquoi ne pas compter la mutation vers STOP comme une substitution possible ? Parce que les mutation vers STOP sont presque toujours éliminées (perte de fonction), donc ne peuvent pas donner lieu à des substitutions que l'on peut détecter.

2) Calculer, sous le modèle de Jukes et Cantor, le nombre de différences nucléotidiques synonymes et non-synonymes entre les codons TAC et GAC, puis entre les codons ACG et CGG.

TAC et GAC
Un seul chemin possible : T(Tyr)-> G.(Asp) : 1 substitution non synonyme, 0 synonyme.

ACG et CGG.
Deux chemins possibles :
ACG-> CCG->CGG (Thr, Pro, Arg)
Ou
ACG-> AGG->CGG (Thr, Arg, Arg)

Chemin 1 : 2 non syn + 0 syn
Chemin 2 : 1 non-syn + 1 syn
Plutot que de choisir entre ces deux chemins également parcimonieux, on leur attribue une proba identique, ½ .
Donc Nb de diff. Syn = 0 ½ +1* ½ =0.5*
Nb de diff. Non-Syn = 2 ½ +1* ½ =1.5*

3) La séquence suivante (figure 1) est extraite de la séquence du locus HLA-A humain. Calculer les proportions de différences nucléotidiques synonymes et non-synonymes qui séparent ces deux allèles.

Organisation pratique

- une ligne de 20 codons pour chacun des x groupes de 2

- obtenir pour chaque ligne :
  - $N_{S1i,}$ N_{S2i} et N_{Si} nombres de sites potentiellement synonymes de chaque séquence et leur moyenne
  - N_{A1i}, N_{A2i} et N_{Ai} nombre de sites potentiellement non-synonymes de chaque séquence et leur moyenne
  - M_{Si}, nombre de différences nucléotidiques synonymes entre les deux séquences
    - M_{Ai}, nombre de différences nucléotidiques non-synonymes entre les deux séquences

ligne	Ns	Na	Ms	Ma
1	14	46	0	4
2	14,67	45,33	1	3
3	13,33	46,67	0	1
4	15,17	44,83	0,5	6,5
5	14,33	45,67	2	1
6	14,42	45,58	3	3
7	13,33	46,67	1	0
8	15,33	44,67	3	1
9	14,33	45,67	0	1
10	3,167	8,833	0	0

- Sur l'ensemble de la séquence,
  - Nombre moyen de sites potentiellement synonymes (N_S) et non synonymes (N_A).
    - *N_S =132.1 N_A =419.9*
  - Nombre total de différences nucléotidiques synonymes (M_S) et non-synonymes (M_A) entre les deux séquences.
    - *M_S =10.5 syn, M_A =20.5*
  - En déduire la proportion de sites nucléotidiques synonymes différents (p_S) et non-synonymes différents (p_A) entre les deux séquences.
    - *ps=0.0795, pa=0.0488*
  - Estimez les nombres de substitutions nucléotidiques par site potentiellement synonyme (K_S) et par site potentiellement synonyme (K_A).
    - *Ks=0.0840, Ka=0.0505*

4) En vous restreignant à la région ARS, que constatez-vous ? Comment interprétez-vous ce résultat ?

Conclusion :
Dans la région ARS, rA>> rS.
C.a.d : + de substitutions non-synonymes que synonymes.
Explication :
Une famille de gènes dont les produits contrôlent la capacité du système immunitaire à à reconnaître des protéines étrangères. Les molécules MHC se lient aux antigènes et les présentent à la surface des cellules. Un tel assemblage est reconnu par les lymphocytes T, qui lorsqu'ils sont activés initient la réaction immunitaire.
(historiquement, le terme histocompatibilité vient du domaine des greffes d'organes).

Le MHC humain s'appelle HLA (human leukocyte antigen), localisé sur le chomosome VI, dans une région qui contient un grand nombre des instruments de la machinerie cellulaire impliquée dans réaction immunitaire. Il existe un très grand polymorphisme de ces gènes.

Deux grands types d'explications, impliquant tous les deux sélection stabilisatrice :

1. *Sélection superdominante : les hétérozygotes peuvent faire face à un plus grand spectre de pathogènes. Un nouveau variant dans une population est nécessairement heterozygote au début car il est rare. A l'inverse, une mutation ne changeant pas la séquence d'a.a. ne bénéficie pas de cet avantage, et a donc une probabilité supérieure d'être perdue. Il résulte le maintien préférentiel des mutations non-synonymes. Au final, on observe plus de différences non-syn que synonymes.*
2. *Course évolutive : Les pathogènes s'adaptent aux allèles MHC présents, d'autant plus facilement qu'ils sont fréquents. Donc un individu portant un nouvel allèle a un avantage parce que les parasites auxquels cet allèle permet de résister n'ont pas eu le temps de s'y adapter.*

```
#A-2301   |GGT TCT CAC ACC CTC CAG ATG ATG TTT GGC TGC GAC GTG GGG TCG GAC GGG CGC TTC CTC      20
#A-2501   |GGT TCT CAC ACC ATC CAG AGG ATG TAT GGC TGC GAC GTG GGG CCG GAC GGG CGC TTC CTC

#A-2301    CGC GGG TAC CAC CAG TAC GCC TAC GAC GGC AAG GAT TAC ATC GCC CTG AAA GAG GAC CTG      40
#A-2501    CGC GGG TAC CAG CAG GAC GCT TAC GAC GGC AAG GAT TAC ATC GCC CTG AAC GAG GAC CTG

#A-2301    CGC TCT TGG ACC GCG GCG GAC ATG GCG GCT CAG ATC ACC CAG CGC AAG TGG GAG GCG GCC      60
#A-2501    CGC TCT TGG ACC GCG GCG GAC ATG GCG GCT CAG ATC ACC CAG CGC AAG TGG GAG ACG GCC

#A-2301    CGT GTG GCG GAG CAG TTG AGA GCC TAC CTG GAG GGC ACG TGC GTG GAC GGG CTC CGC AGA      80
#A-2501    CAT GAG GCG GAG CAG TGG AGA GCC TAC CTG GAG GGC CGG TGC GTG GAG TGG CTC CGC AGA

#A-2301    TAC CTG GAG AAC GGG AAG GAG ACG CTG CAG CGC ACG|GAC CCC CCC AAG ACA CAT ATG ACC     100
#A-2501    TAC CTG GAG AAC GGG AAG GAG ACG CTG CAG CGC ACG|GAC GCC CCC AAG ACG CAT ATG ACT

#A-2301    CAC CAC CCC ATC TCT GAC CAT GAG GCC ACT CTG AGA TGC TGG GCC CTG GGC TTC TAC CCT     120
#A-2501    CAC CAC GCT GTC TCT GAC CAT GAG GCC ACC CTG AGG TGC TGG GCC CTG AGC TTC TAC CCT

#A-2301    GCG GAG ATC ACA CTG ACC TGG CAG CGG GAT GGG GAG GAC CAG ACC CAG GAC ACG GAG CTT     140
#A-2501    GCG GAG ATC ACA CTG ACC TGG CAG CGG GAT GGG GAG GAC CAG ACC CAG GAC ACG GAG CTC

#A-2301    GTG GAG ACC AGG CCT GCA GGG GAT GGA ACC TTC CAG AAG TGG GCA GCT GTG GTG GTA CCT     160
#A-2501    GTG GAG ACC AGG CCT GCA GGG GAT GGG ACC TTC CAG AAG TGG GCG TCT GTG GTG GTG CCT

#A-2301    TCT GGA GAG GAG CAG AGA TAC ACC TGC CAT GTG CAG CAT GAG GGT CTG CCC AAG CCC CTC     180
#A-2501    TCT GGA CAG GAG CAG AGA TAC ACC TGC CAT GTG CAG CAT GAG GGT CTG CCC AAG CCC CTC

#A-2301    ACC CTG AGA TGG                                                                   184
#A-2501    ACC CTG AGA TGG
```

Figure 1. Séquences nucléotidiques de deux des trois domaines extracellulaires, α2, α3 de deux allèles du locus HLA-A humain. Les barres verticales représentent les limites des exons. On ne montre ici que les exons 2 et 3. Les nucléotides du site de reconnaissance de l'antigène (ARS) sont figurés en gras.

Maîtrise de Génétique et Microbiologie
TD 3

Méthodes de reconstruction phylogénétique

I) A. Par maximum de parcimonie

50

```
(1) GTCTGTTCCAAGGGCCTTTGCGTCAGGCTGGGCCTCAGGGTTGCCCCACT
(2) GTCTGTTACAAGGGCCTTCGCGCCAGGCTAGGCCTCAGGGTTGCCTCACT
(3) GTGTGTTTCAAGGGCCTTTGCGCCAGTCTGGGCCCCAGGGCTGCCCCACT
(4) GTGCGTTACATGGGCCTTCGCGTCAGTCTGGGCCCCAGGGCTGCCTCACT

(1) CGGGGTTCCAGGGCAGCTGGACCCCAGGCCCCAGCTCTGCATCAGGGAGG
(2) CGAGGTGCCAGGGCGGCTGGTCCCCAGGCCTCAGCTCTGCAGCAAGGAGG
(3) CGTGGTACCAGAGCAGTTGGACCCCAGGTCCCAGCTCTACAGCATTGAGG
(4) TGCGGTTCCAGAGCGGTTGGACCCTAGGTCTCAACTCTACAGCAGGGAGG

(1) ACGTGGCTGGGCCCGTGAAACATGTGTGGGTGAGTCCAGGC
(2) GCGTGGTTGGGCTCGTGAAGCATGTGGGGGTGAGTCCGGGG
(3) ATGTGGCTGGACTCGTGAAGCATTTGTGGGTGAGCCCGGGG
(4) GTGTGGTTGGGCCCGTGAAGCATTTGGGGGTAAGCCCAGGG
```

Attention pour l'an prochain : je ne retrouve pas les sequences sur Genebank : 3et4 = la meme espece = singe de nuit. (a verifier !)

1) Combien existe-t-il de topologies non enracinées possibles reliant quatre séquences ?

Il existe trois topologies possibles :
(1,2),(3,4)
(1,3),(2,4)
(1,4),(2,3)

2) Sous quelle(s) conditions un site nucléotidique peut-il vous aider à reconstituer l'histoire évolutive évolutive de ces taxons ? Identifier ces sites sur les quatre séquences ci-dessus (sites phylogénétiquement informatifs).

Il faut que l'une des topologies implique moins de substitutions nucléotidiques que les deux autres.
En pratique, il faut que le site comporte deux états qui soient partagés par au moins deux taxons.
Il y en a 21.

3) Quel arbre non enraciné retenez-vous pour représenter l'histoire évolutive évolutive de ces taxons ?

(1,2),(3,4) 11 sites
(1,3),(2,4) 7 sites
(1,4),(2,3) 3 sites

La topologie (1,2), (3,4) est plébiscitée.
S'il s'agit de la véritable histoire, alors il s'est produit 11+7*2+3*2=**31** substitutions au minimum
Si par contre la véritable histoire était (1,3),(2,4), alors faut 11*2+7+3*2=**35** substitutions
Enfin, si (1,4),(2,3), il faut 11*2+7*2+3=**39** substitutions

3) Cette conclusion vous paraît-elle robuste ?

Approche intéressante car se base sur le vrai détail du processus de substitunios (meme si seulement inféré).
Mais tres faible nombre de sites analysés.

De +, aucune estimation de la confiance statistique qu'il faut accorder aux résultats : bootstrap serait nécessaire.

B. Par UPGMA

1) Dénombrer le nombre de sites nucléotidiques différents entre chaque paire de séquence. Sous le modèle de Jukes et Cantor, estimez la proportion de sites ayant été affecté par une substitution.
2) Regrouper les deux taxons ayant la plus faible divergence en un taxon composite et estimez la distance de ce groupe à tous les autres taxons (moyenne des distances).
3) Continuer ainsi jusqu'à l'ancetre commun de tous les taxons.
4) Comparer ces resultats à ceux obtenus par la méthode de maximum de parcimonie.

	1	2	3	4
1				
2	**0,1348**			
3	0,1631	0,1844		
4	0,2057	0,1773	0,1560	

	(1,2)	3	4
(1,2)			
3	0,174		
4	0,191	**0,1560**	

	(1,2)	(3,4)
(1,2)		
(3,4)	**0,183**	

On retrouve bien le groupement (1,2),(3,4).
Une approche plus globale, car caractérise la distance moyenne entre tous les taxons, sans s'intéresser à ce qui se passe site par site.
Avec une horloge moléculaire, peut dater directement les MRCA.

Meme limite que précédemment, mais tres faible nombre de sites analysés.
De +, aucune estimation de la confiance statistique qu'il faut accorder aux résultats : bootstrap serait nécessaire.

Quelle hypothèse forte avez-vous posé pour construire cet arbre ?

Hypothèse d'égalité des taux d'évolution, dont on a vu qu'elle était parfois fausse. Cette méthode peut donner lieu à des topologies fausses si les taux d'évolution ne sont pas constants.