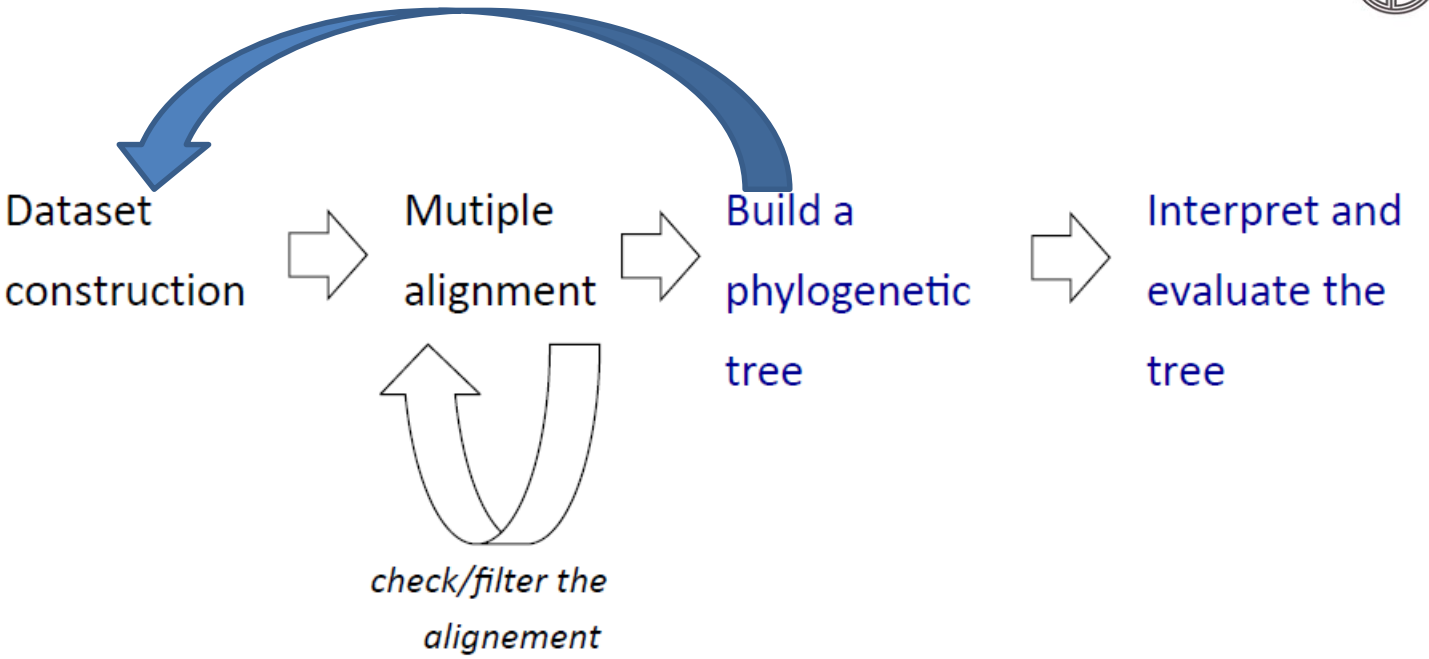
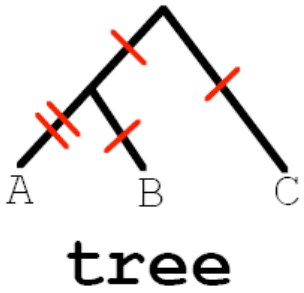
Inférence phylogénétique

Principales étapes d'une analyse phylogénétique



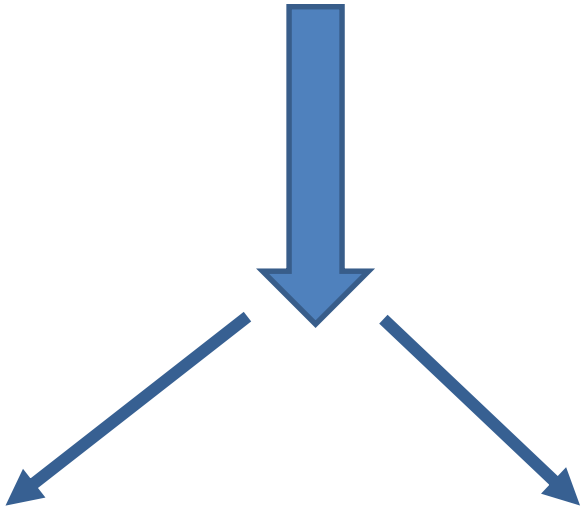

```
CAAACAGCGTT---GGCTCTCTA
AAAATAACACCaacATGCAAATG
AAAACAGCACCaacGTGCAAATG
AAAACAGCACCaacGTGCAAATG
```



Principales étapes d'une analyse phylogénétique

UNIVERSITÉ TOULOUSE III PAUL SABATIER

Dataset construction ⇨ Mutiple alignment ⇨ Build a phylogenetic tree ⇨ Interpret and evaluate the tree



Choix du modèle évolutif

Choix de la méthode de construction

Calcul d'une distance génétique (évolutive) entre deux séquences

UNIVERSITÉ TOULOUSE III PAUL SABATIER

Divergence observée ou *p*-distance : la plus simple

On compte le nombre *s* de substitutions observées entre deux séquences alignées que l'on rapporte au nombre de sites homologues *n* alignés, donc :

$$p = \frac{s}{n}$$

Proportional (*p*) Distance

DNA Site

Species	1	2	3	4	5	6	7	8	9	10
I	A	T	A	T	A	C	G	T	A	T
II	A	T	G	T	A	C	G	T	A	T
III	G	T	A	-	A	C	G	T	G	C
IV	G	C	G	T	A	T	G	C	A	C

$p =$ # differences / # sites

	I	II	III	IV
I	-	0.1	0.4	0.6
II		-	0.5	0.5
III			-	0.6
IV				-

facile à calculer mais quand les séquences ne sont pas proches (issues d'organismes distants dans l'évolution), elle sous-estime les distances évolutives.

Cause : l'existence de substitutions multiples. Phénomène plus critique pour les séquences d'acides nucléiques que pour les séquences protéiques.

Calcul d'une distance génétique (évolutive) entre deux séquences

Substitutions multiples

Séquence1 **GAAAAG**
Séquence2 **ATGAAG**

Type de substitution	Séquence 1	Séquence 2	Nombre de substitutions observé	Nombre de substitutions réel
Substitution unique (simple)	G	G ➤ A	1	1
Substitutions multiples	A	A ➤ C ➤ T	1	2
Substitutions coïncidentes au même site	T ➤ A	T ➤ G	1	2
Substitutions parallèles	T ➤ A	T ➤ A	0	2
Substitutions convergentes	C ➤ G ➤ A	C ➤ A	0	3
Substitution réverse (inverse)	G ➤T ➤ G	G	0	2

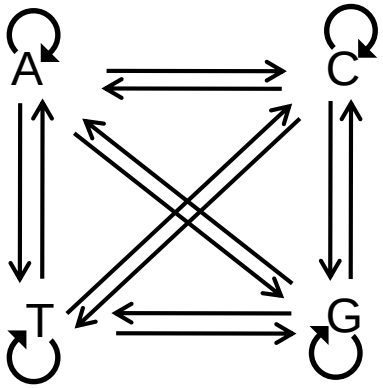
Calcul d'une distance génétique (évolutive) entre deux séquences

Pour tenter de corriger le biais du aux mutations multiples, des hypothèses sont faites sur la façon dont les bases se sont substituées à un locus donné

➤ Construction d'un modèle évolutif

➤ modéliser par un modèle de Markov en temps continu

Dans les modèles markoviens, l'information utile pour la prédiction du futur est contenue dans l'état présent du processus. Donc, l'état futur d'un site ne dépendra que de son état présent et pas des états passés.



Les substitutions à chaque site sont décrites par une chaîne de Markov dont les états correspondent aux quatre bases nucléotidiques et les probabilités de transitions sont données par les probabilités de passer d'un état à un autre ou de rester dans le même état.

L'évolution d'un site le long d'une branche d'un arbre phylogénétique est décrite par les probabilités de transition p_{ij} d'un état initial *i* au nœud ancêtre à un état *j* au nœud fils.

Calcul d'une distance génétique (évolutive) entre deux séquences

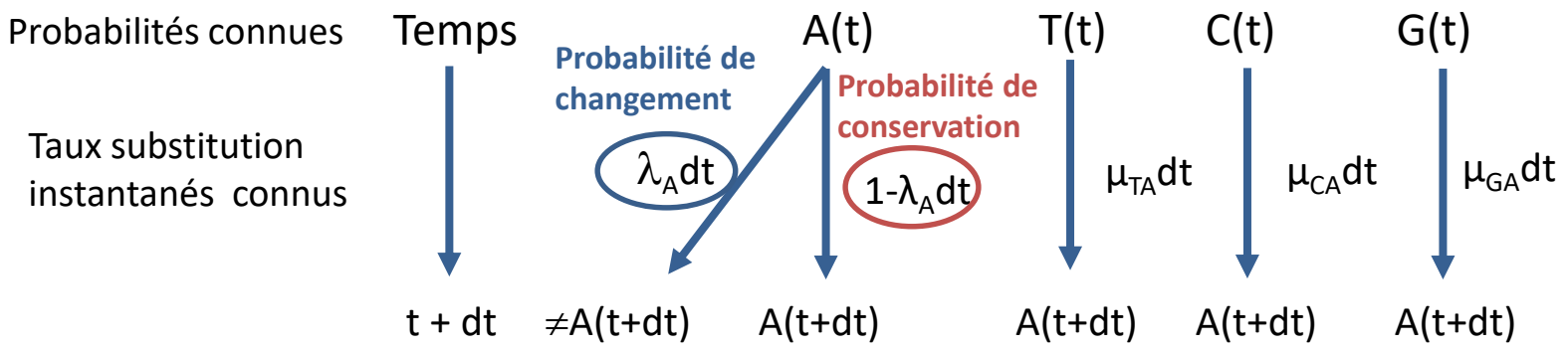
Hypothèses liées au modèle markovien

1. **Homogénéité du processus** : les probabilités de substitution ne changent pas au cours du temps. Donc même processus applicable le long de toutes les branches de l'arbre.

   On peut donc définir :

   - Le taux de substitution instantané d'une base d'un état *i* vers un état *j* μ_{ij} ($i \neq j$) c'est-à-dire le nombre attendu de substitutions du nucléotide *i* par le nucléotide *j* par unité de temps
   - Le taux de changement instantané d'un nucléotide dans l'état *i* vers un autre nucléotide λ_i, c'est-à-dire le nombre attendu de substitution du nucléotide *i* en n'importe quel autre nucléotide par unité de temps.

Exemple : Calcul de la probabilité d'observer le nucléotide A à un site donné au temps $t + dt$



$$A(t+dt) = A(t)(1-\lambda_A dt) + T(t)\mu_{TA}dt + C(t)\mu_{CA}dt + G(t)\mu_{GA}dt$$

Calcul d'une distance génétique (évolutive) entre deux séquences

Si on fait le même raisonnement pour chacune des 4 bases on obtient le système de quatre équations différentielles linéaires :

$$A(t+dt) = A(t)(1-\lambda_A dt) + T(t)\mu_{TA}dt + C(t)\mu_{CA}dt + G(t)\mu_{GA}dt$$
$$T(t+dt) = T(t)(1-\lambda_T dt) + A(t)\mu_{AT}dt + C(t)\mu_{CT}dt + G(t)\mu_{GT}dt$$
$$G(t+dt) = G(t)(1-\lambda_G dt) + A(t)\mu_{AG}dt + T(t)\mu_{TG}dt + C(t)\mu_{CG}dt$$
$$C(t+dt) = C(t)(1-\lambda_C dt) + A(t)\mu_{AC}dt + T(t)\mu_{TC}dt + G(t)\mu_{GC}dt$$

On peut en déduire la matrice M des taux de substitution instantanés:

$$M = \begin{bmatrix} -\lambda_A & \mu_{AT} & \mu_{AC} & \mu_{AG} \\ \mu_{TA} & -\lambda_T & \mu_{TC} & \mu_{TG} \\ \mu_{CA} & \mu_{CT} & -\lambda_C & \mu_{CG} \\ \mu_{GA} & \mu_{GT} & \mu_{GC} & -\lambda_G \end{bmatrix}$$

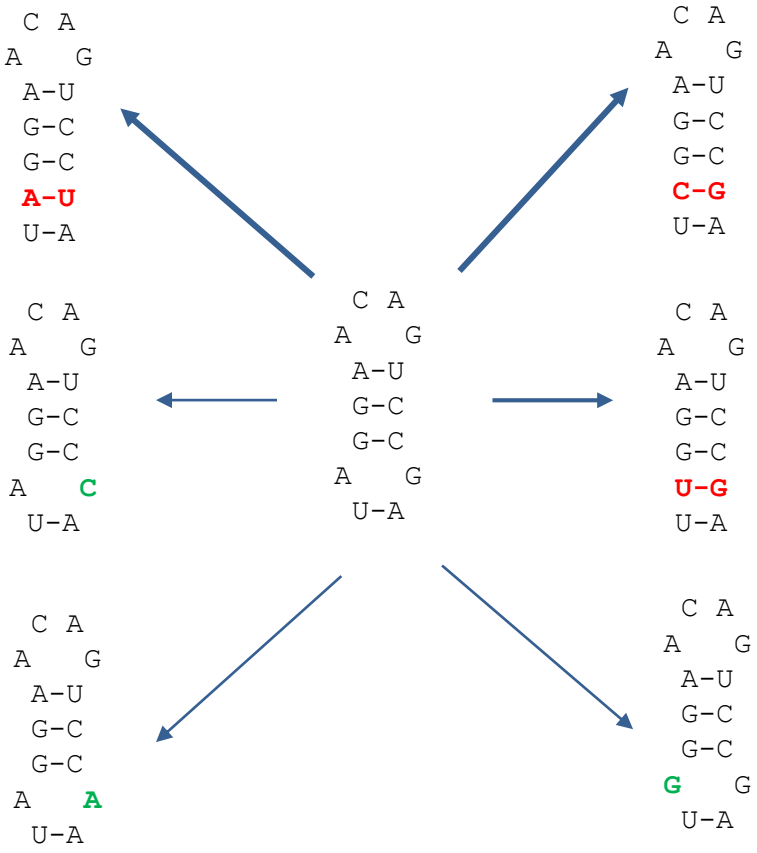
La différence entre les modèles d'évolution est liée à la définition des μ_{ij}

La matrice M décrit les fréquences relatives des différents types de substitutions, seuls les rapports entre les valeurs de μ_{ij} sont informatifs (par exemple le rapport transitions/tranversions, la fréquence des bases à l'équilibre, etc) et participent à la description du modèle.

Calcul d'une distance génétique (évolutive) entre deux séquences

Hypothèses liées au modèle markovien

2. **Indépendance des sites :** les sites évoluent indépendamment les uns des autres. Hypothèse pas vérifiée dans de nombreux cas, notamment pour les ARN structuraux où pour maintenir la structure secondaire, il y a apparition de mutations compensatrices (coévolution des sites).



Mutations qui rétablissent l'appariement
→ Plus forte probabilité de se fixer

Calcul d'une distance génétique (évolutive) entre deux séquences

Hypothèses liées au modèle markovien

3. Uniformité du processus : tous les sites d'une séquence suivent le même processus, c'est-à-dire que les probabilités et taux de substitution sont applicables à tous les sites. Conséquence, on suppose que les sites évoluent à la même vitesse. On sait que cette hypothèse est fausse mais elle utilisée dans la plupart des modèles d'évolution.

Des améliorations dans les modèles ont été proposées pour prendre en compte l'existence de vitesses d'évolution différentes. La plus courante est l'utilisation de la distribution Gamma.

4. Stationnarité : liée à l'hypothèse d'homogénéité, les taux d'évolution μ_{ij} sont constants au cours du temps pour l'ensemble des lignées. La matrice M s'applique donc en tout point de l'arbre. Si les μ_{ij} sont positifs, le processus markovien est dit stationnaire. Il existe alors, pour la chaîne de Markov associée, une distribution $\Pi = \{\pi_i\}$ qui est dite stationnaire.

Lorsque $t \rightarrow \infty$, les valeurs des fréquences des quatre bases tendent vers cette distribution. Ces valeurs correspondent aux compositions en bases à l'équilibre, soit les proportions de A, T, C et G attendues après un temps d'évolution infini.

Pas une contrainte absolue et certains modèles d'évolution non stationnaires ont été proposé.

Calcul d'une distance génétique (évolutive) entre deux séquences

5. Réversibilité : signifie que quand l'équilibre est atteint, la quantité de changement de l'état *i* vers l'état *j* est égale à la quantité de changement de l'état *j* vers l'état *i*. On a $\pi_i \mu_{ij} = \pi_j \mu_{ji}$.
Permet de simplifier les calculs, car les 12 paramètres non diagonaux de la matrice M peuvent être décrits par 9 paramètres, les 6 taux d'interchangeabilité μ_{ij}, et trois fréquences de bases à l'équilibre car $\sum_i \pi_i = 1$

$$M = \{\mu_{ij}\} = \begin{bmatrix} -\lambda_A & \mu_{AT} & \mu_{AC} & \mu_{AG} \\ \mu_{TA} & -\lambda_T & \mu_{TC} & \mu_{TG} \\ \mu_{CA} & \mu_{CT} & -\lambda_C & \mu_{CG} \\ \mu_{GA} & \mu_{GT} & \mu_{GC} & -\lambda_G \end{bmatrix}$$

$\mu_{AT} = \mu_{TA}$
$\mu_{AC} = \mu_{CA}$
$\mu_{AG} = \mu_{GA}$
$\mu_{CT} = \mu_{TC}$
$\mu_{CG} = \mu_{GC}$
$\mu_{GT} = \mu_{TG}$

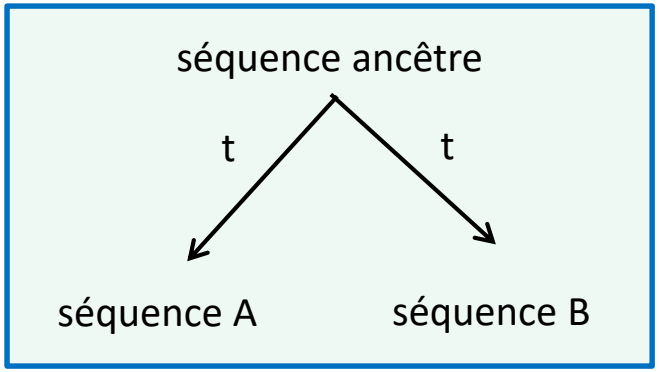
Calcul d'une distance génétique (évolutive) entre deux séquences

Distance évolutive entre deux séquences

Le nombre de substitution ayant eu lieu pendant un intervalle de temps *t* est donné par la multiplication du taux d'évolution global λ par *t*. Sous l'hypothèse de la stationnarité on a :

$$\lambda = \sum_i \pi_i \lambda_i \qquad \text{avec} \qquad \lambda_i = \sum_{i \neq j} \mu_{ij}$$

Quand on compare deux séquences homologues, le temps qui les sépare est non de *t* mais de 2*t*.



La distance évolutive séparant deux séquences est donc :

$$d = 2\sum_i \pi_i \lambda_i t$$

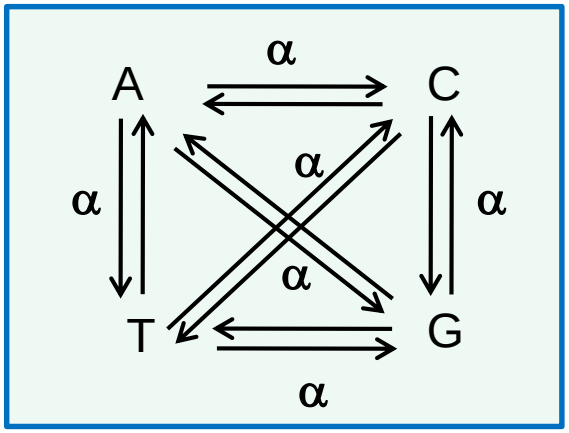
Modèles d'évolution ADN/ARN

Modèle de Jukes et Cantor (abrégé JC69)

- Modèle markovien le plus simple mais vision simplificatrice de l'évolution
- Toutes les positions changent avec la même probabilité
- toutes les substitutions sont équiprobables donc un seul taux de substitution instantané α pour chacun des changements possibles (tous les $\mu_{ij} = \alpha$) et un seul taux de conservation global instantané $1-3\alpha$.
- Egalité des fréquences nucléotidiques : $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$

$$M = \begin{bmatrix} -\lambda & \alpha & \alpha & \alpha \\ \alpha & -\lambda & \alpha & \alpha \\ \alpha & \alpha & -\lambda & \alpha \\ \alpha & \alpha & \alpha & -\lambda \end{bmatrix}$$

avec $\lambda = 3\alpha$



La matrice des probabilité de substitution P(t) peut donc s'écrire :

$$P(t) = \begin{bmatrix} q(t) & p(t) & p(t) & p(t) \\ p(t) & q(t) & p(t) & p(t) \\ p(t) & p(t) & q(t) & p(t) \\ p(t) & p(t) & p(t) & q(t) \end{bmatrix}$$

avec :
q(*t*) probabilité qu'après un temps *t* le nucléotide reste inchangé
p(*t*) probabilité qu'après un temps *t* il se soit substitué en un autre (passage de l'état *i* à l'état *j*).

Modèles d'évolution ADN/ARN

Modèle de Jukes et Cantor (abrégé JC69)

Pour calculer la distance entre deux séquences, il faut que l'on trouve une relation entre *d* et la probabilité d'observer une substitution à un site qui est donnée par la distance observée ou *p*-distance.

La démonstration à la fin du document pour les curieux….

la distance de Jukes et Cantor est donnée par :

$$d = -\frac{3}{4} Log\left(1 - \frac{4}{3} p_{dist}\right)$$

Un facteur correcteur est donc apporté à la p-distance p_{dist}

Quand p = ¾ d → ∞ (Log 0 pas défini). Donc ce modèle n'est pas utilisable pour des séquences dont la distance observée est supérieure à 75%.

Modèles d'évolution ADN/ARN

```
Seq1    TCAAGTCAGGTTCGA
           |   |  ||
Seq2    TCCAGTTAGACTCGA
          | | | |  | |
Seq3    TTCAATCAGGCCCGA
```

Distances observées

	Seq1	Seq2	Seq3
Seq2	0.266		
Seq3	0.333	0.333	

Distance observée

p-distance (seq1,seq2) = $\frac{4}{15}$ = 0.266

Distances évolutives Jukes et Cantor

	Seq1	Seq2	Seq3
Seq2	0.328		
Seq3	0.441	0.441	

Distance J&C

$$d = -\frac{3}{4} Log\left(1 - \frac{4}{3} p_{dist}\right)$$

$$d_{JC} = -\frac{3}{4} Log(1 - 0.266\frac{4}{3}) = 0.328$$

Modèles d'évolution ADN/ARN

Modèle de Kimura à deux paramètres (K80)

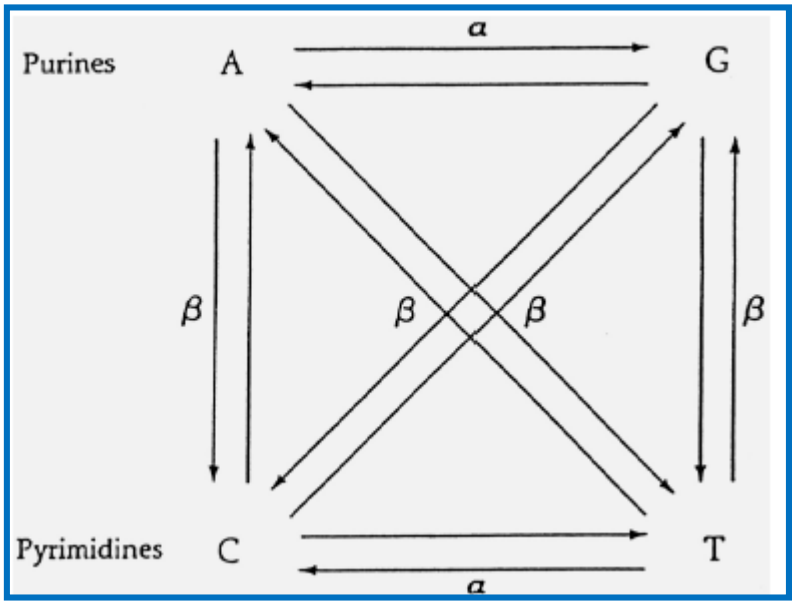
➢ Modèle moins simplificateur et biologiquement plus réaliste car il est observé que la fréquence des transitions est plus élevée que celle des transversions.

➢ les substitutions se produisent suivant deux taux distincts, l'un pour les transitions, l'autre pour les transversions, les transitions étant plus fréquentes (transition = A<->G ou T<->C).

➢ Egalité des fréquences nucléotidiques : $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$

$$M = \begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{bmatrix} -\lambda & \beta & \beta & \alpha \\ \beta & -\lambda & \alpha & \beta \\ \beta & \alpha & -\lambda & \beta \\ \alpha & \beta & \beta & -\lambda \end{bmatrix} \end{array}$$

avec α taux de transitions et β taux de tranversions
et λ le taux instantané de changement pour une base quelconque $\lambda = 2\beta + \alpha$



Modèles d'évolution ADN/ARN

Distance de Kimura à deux paramètres :

$$d = -\frac{1}{2} Log(1-2p-q) - \frac{1}{4} Log(1-2q)$$

Avec :

- *p* fréquence observée des transitions
- *q* fréquence observée des transversions

Modèles d'évolution ADN/ARN

Modèle de Tamura (T92)

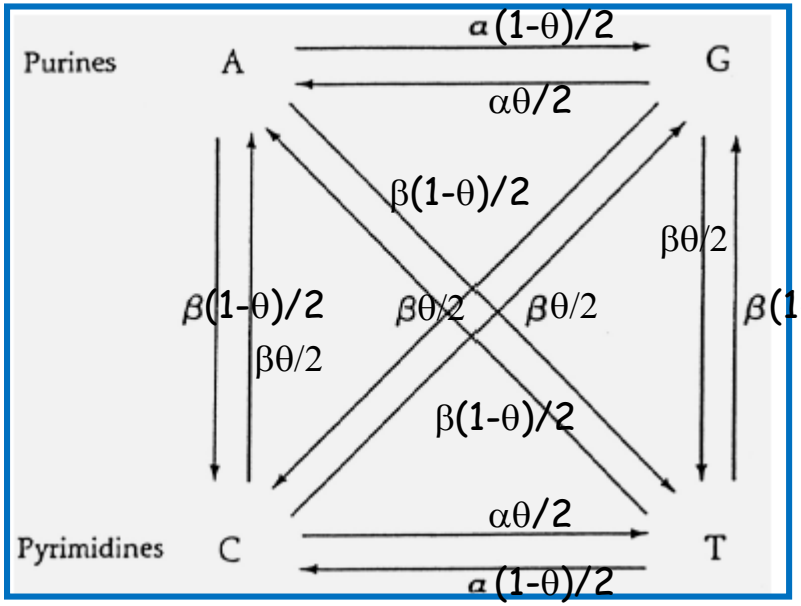
➢ Les modèles précédents JC69 et K80 imposent que les fréquences des bases à l'équilibre soient toutes égales à ¼, donc que le taux global de GC soit égal à ½. Or ceci est rarement vérifié sur les séquences réelles. Des modèles alternatifs ont été proposés pour rendre compte de cette réalité biologique.

➢ Le modèle de Tamura est une extension du modèle K80 en intégrant un paramètre supplémentaire θ représentant la fréquence de GC de la ou des séquence(s) considérée(s). Soit :

$$\pi_{GC} = \theta$$

Modèles d'évolution ADN/ARN

Modèle de Tamura (T92)



La matrice des taux de changements M devient :

$$M = \begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{bmatrix} -\lambda_A & \beta(1-\theta)/2 & \beta(1-\theta)/2 & \alpha(1-\theta)/2 \\ \beta(1-\theta)/2 & -\lambda_T & \alpha(1-\theta)/2 & \beta(1-\theta)/2 \\ \beta\theta/2 & \alpha\theta/2 & -\lambda_C & \beta\theta/2 \\ \alpha\theta/2 & \beta\theta/2 & \beta\theta/2 & -\lambda_G \end{bmatrix} \end{array}$$

avec α taux de transitions et β taux de tranversions
et λ le taux instantané de changement qui maintenant est fonction de la nature de la base (AT ou GC) :

$\lambda_A = \lambda_T = \beta\,(1 - \theta)/2 + \beta\theta/2 + \alpha\theta/2 = (\beta + \alpha\theta)/2$

$\lambda_C = \lambda_G = \beta\,(1 - \theta)/2 + \alpha\,(1 - \theta)/2 + \beta\theta/2 = (\beta + \alpha(1-\theta))/2$

$$\pi_{GC} = \theta$$

$$\pi_G = \pi_C = \theta/2$$

$$\pi_A = \pi_T = (1-\theta)/2$$

Modèles d'évolution ADN/ARN

Modèle de Tamura (T92)

La distance entre deux séquences est donnée par :

$$d = -hLog\left(1 - \frac{p}{h} - q\right) - \frac{1}{2}(1-h)Log(1-2q)$$

Avec :

- $h = 2\theta(1-\theta)$
- p fréquence observée des transitions
- q fréquence observée des transversions

Modèles d'évolution ADN/ARN

Modèle de Tamura et Nei (TN93)

➢ Ce modèle comprend 6 paramètres et est le modèle réversible le plus général dont l'expression analytique des probabilités est connue. Dans ce modèle, on distingue le taux de transition en fonction qu'il s'effectue entre deux purines (A<->G) , noté α_R, ou entre deux pyrimidines (T<->C) , noté α_Y. Le taux de tranversion reste β.

$$M = \begin{array}{c} \\ A \\ T \\ C \\ G \end{array} \begin{array}{c} \begin{array}{cccc} A & T & C & G \end{array} \\ \begin{bmatrix} -\lambda_A & \beta\pi_T & \beta\pi_C & \alpha_R\pi_G \\ \beta\pi_A & -\lambda_T & \alpha Y\pi_C & \beta\pi_G \\ \beta\pi_A & \alpha_Y\pi_T & -\lambda_C & \beta\pi_G \\ \alpha_R\pi_A & \beta\pi_T & \beta\pi_C & -\lambda_G \end{bmatrix} \end{array}$$

Ce modèle fait donc apparaître les fréquences π_i des bases à l'équilibre (calculée de façon empirique à partir de la moyenne des fréquences en bases des séquences étudiées)

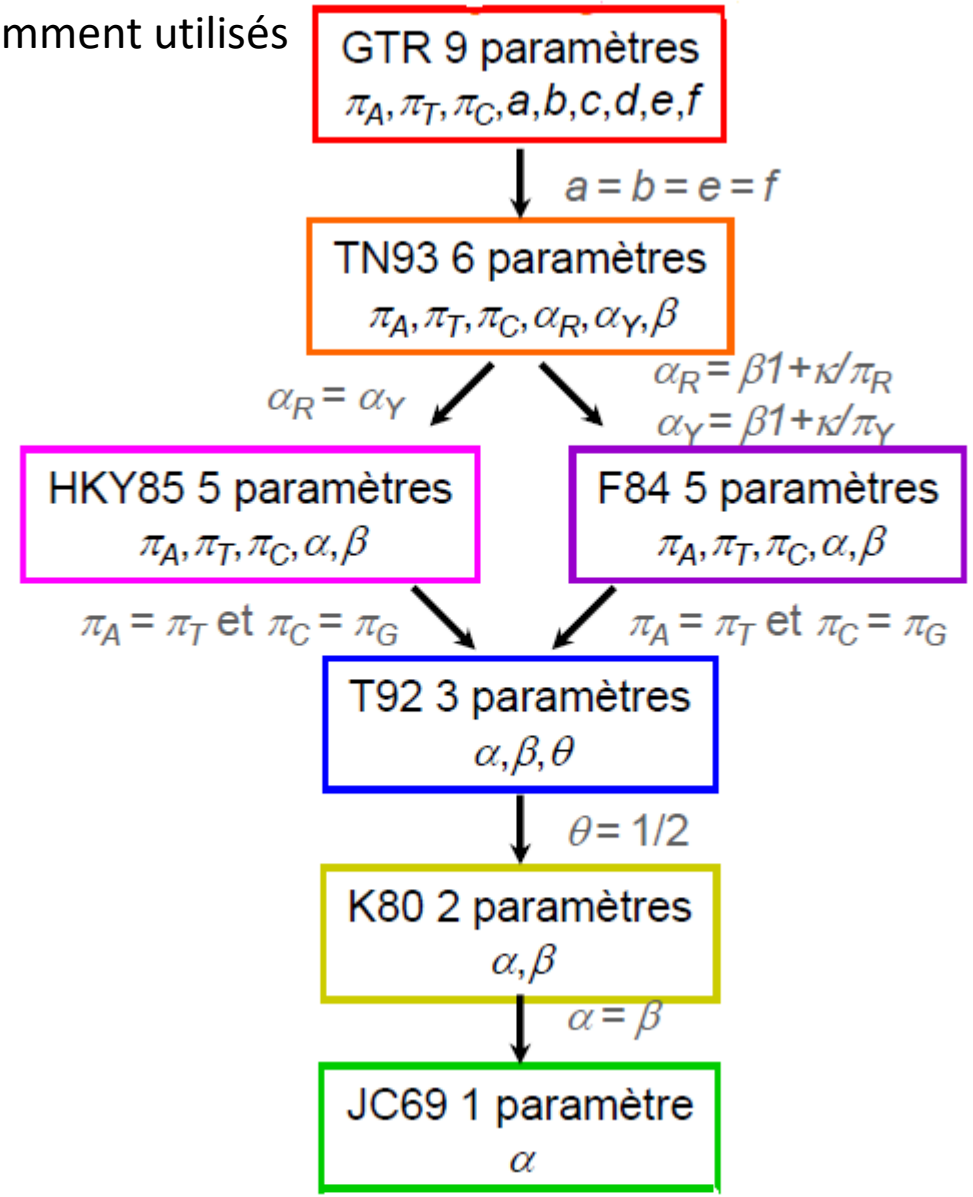
Pour les autres modèles voir :

Concepts et méthodes en phylogénie moléculaire, Guy Perriére et Céline Brochier-Armanet, collection IRIS, Springer-Verlag France

Inferring Phylogenies,J. Felsenstein, Sinauer Associates Inc. Publishers Sunderland, Massachutsetts

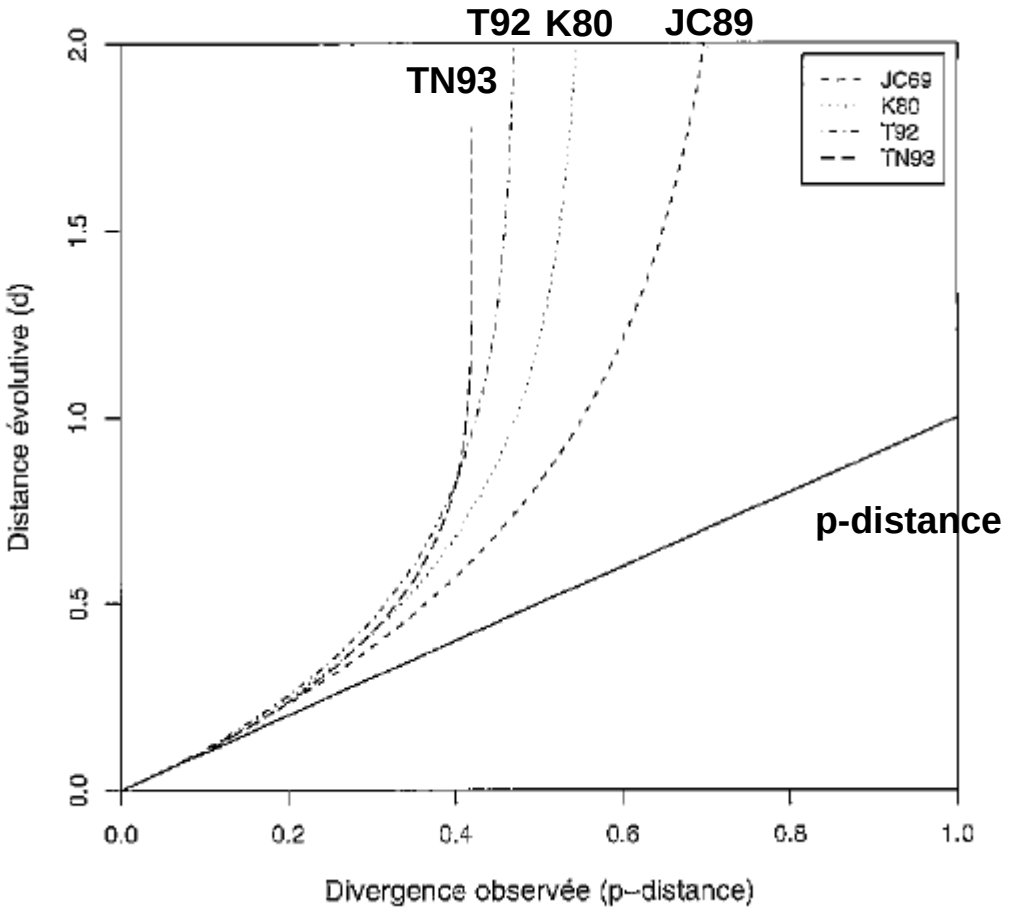
Modèles d'évolution ADN/ARN

Modèles couramment utilisés



Choix d'un modèle évolutif

Comparaison des différents modèles évolutifs de séquences d'ADN



Paramètres fixés :
taux GC (θ) = 0.3
π_A = 0.3, π_T = 0.4, π_C = 0.2 et π_G = 0.1
κ = transition/transversion = 4

- d ≤ 0.1 tous les modèles : même résultat (on peut utiliser modèle simple)
- 0.5 > d > 0.1 on peut utiliser JC89 ou K80, K80 préférable si séquence κ >5
- 1 > d > 0.5 utilisation de modèles avec nombre de paramètres plus important
- d > 1 pour beaucoup de paires de séquences, prudence sur la fiabilité de l'arbre. Eliminer les sites saturés manuellement ou avec méthodes appropriées.

Extrait de Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire.

Distances synonymes et non synonymes

- Hypothèses des modèles précédents:

  Tous les sites évoluent indépendamment selon le même processus.
- Problème: dans les gènes protéiques, il existe deux classes de sites avec des taux d'évolution très différents.
  - substitutions non synonymes (changent l'acide aminé): lent
  - substitutions synonymes (ne changent pas l'acide aminé): rapide
- Solution: calculer deux distances évolutives
  - **K_A ou d_N** = distance non-synonyme

    = nbr. substitutions non-synonymes / nbr. sites non-synonymes
  - **K_S ou d_S** = distance synonyme

    = nbr. substitutions synonymes / nbr. sites synonymes

Si les séquences sont soumises à une sélection purificatrice, on attend un déficit de substitutions non synonymes : $d_N/d_S < 1$

Si les séquences sont soumises à une sélection positive, on attend un excès de substitutions non synonymes : $d_N/d_S > 1$

Si les séquences évoluent de façon neutre on aura : $d_N \approx d_S$

Calcul des distances entre deux séquences protéiques

UNIVERSITÉ TOULOUSE III PAUL SABATIER Université de Toulouse

- Séquences protéiques fréquemment utilisées en phylogénie moléculaire car plus appropriées quand les analyses comportent des séquences issues de lignées séparées par de grandes distances évolutives ou quand les séquences évoluent rapidement (au niveau ADN perte du signal phylogénétique car les sites sont dits saturés, *i.e.*, ont subi de nombreuses substitutions multiples).

- Egalement plusieurs modèles pour estimer la distance entre deux séquences

Modèles d'évolution pour les séquences protéiques

Modèle de Poisson

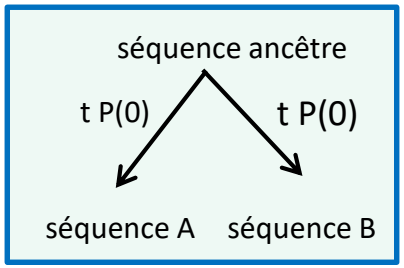
- Première estimation meilleure que la p-distance repose sur le concept de distribution de Poisson.
- Hypothèses :
  - tous les sites évoluent indépendamment et suivant un même processus
  - toutes les substitutions sont équiprobables
  - le taux de réversion est négligeable

Soit λ le taux de substitution à un site donné, alors λt correspond au nombre de substitutions s'étant produite au cours du temps *t* et la probabilité *P(k)* d'avoir *k* substitutions à un site (*k*={0,1,2,3,…} est donnée par la distribution de Poisson :

$$P(k) = e^{-\lambda t}\frac{(\lambda t)^k}{k!}$$

P(0) = $e^{-\lambda t}$ = proba aucun changement au temps *t*

soit *p* la probabilité d'observer une substitution entre deux séquences
et *q* = 1 – *p* la probabilité d'observer deux résidus identiques



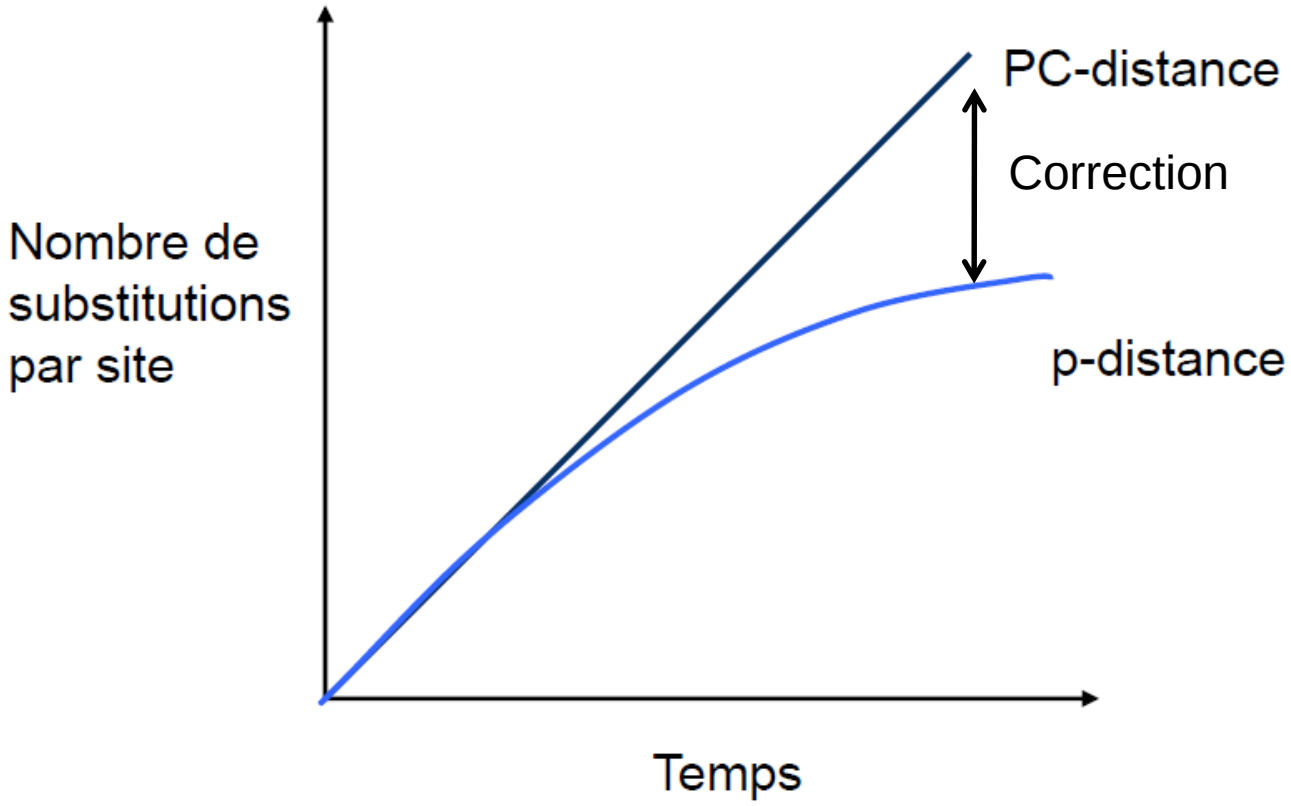
$q = P(0)^2 = e^{-2\lambda t}$ et $d = 2\lambda t$ (d distance entre les deux séquences)

On en déduit $q = 1-p = e^{-d}$

d'où ***la correction de Poisson*** : $d = -Log(1-p)$ Valeur de *p* estimée par la *p*-distance

Modèles d'évolution pour les séquences protéiques

Relation entre la *p*-distance et la distance corrigée de Poisson



Modèles d'évolution pour les séquences protéiques

Modèle de Poisson

- Cependant vision très simplificatrice car en particulier :
  - taux de substitutions plus ou moins élevé en fonction de l'importance fonctionnelle du site
  - présence aussi de substitution parallèle et de réversion donc on va sous-estimer la distance entre deux séquences
  - ne peut être utilisée que si séquences globalement peu divergentes

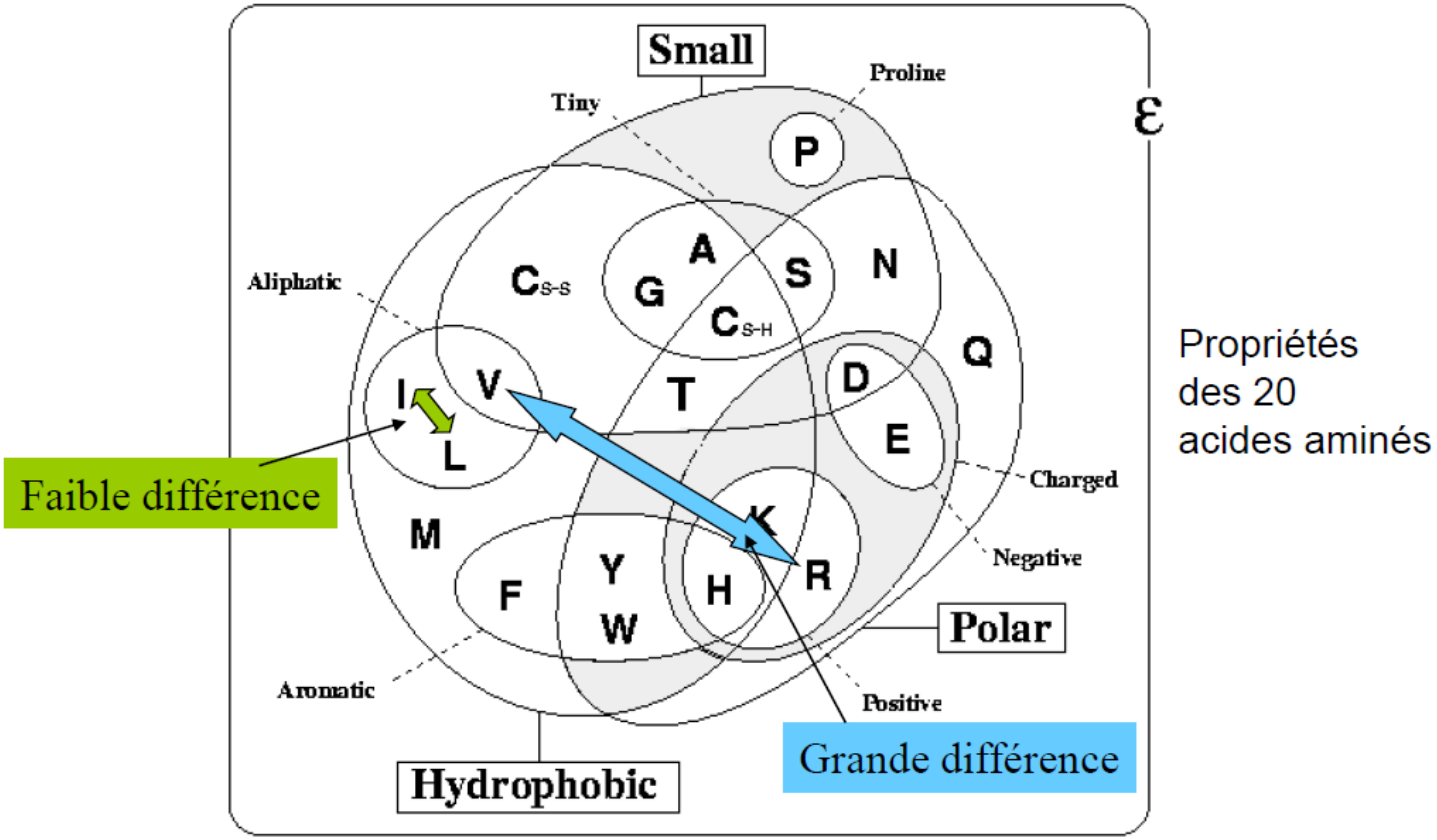
Donc autres modèles ont été développés.

Modèle	Référence
PAM	Dayhoff 1978
BLOSUM	Henikoff 1992
JTT (réactualisation de la PAM)	Jones 1992
WAG & LG	Whelan 2001, Le et Gascuel
Spécifiques (organelles etc..)	

Modèles d'évolution pour les séquences protéiques

Modèle PAM et assimilé

➢ Il est vite apparu que les substitutions entre acides aminés étaient d'autant plus fréquentes que ces acides aminés étaient proches en terme de propriétés physico-chimiques (polarité, hydrophobicité…)



Propriétés des 20 acides aminés

Modèles d'évolution pour les séquences protéiques

Modèle PAM et assimilé

Comment mesurer la similarité entre deux acides aminés ?

Méthode qui a été retenue : estimer en comparant des ensembles de séquences protéiques la fréquence de substitution d'un acide aminé en un autre → obtention d'une matrice 20x20 dont chacune des cases contient la valeur numérique attribuée à la paire d'acide aminé en question.

Première matrice construite par Dayhoff *et al.* En 1978 et appelée PAM pour Point Accepted Mutation : matrice empirique construite directement à partir de données issues de 71 familles de séquences homologues (environ 1300 séquences). Elle rend compte de deux processus :

- l'apparition de substitutions
- leur passage au travers du crible de la sélection.

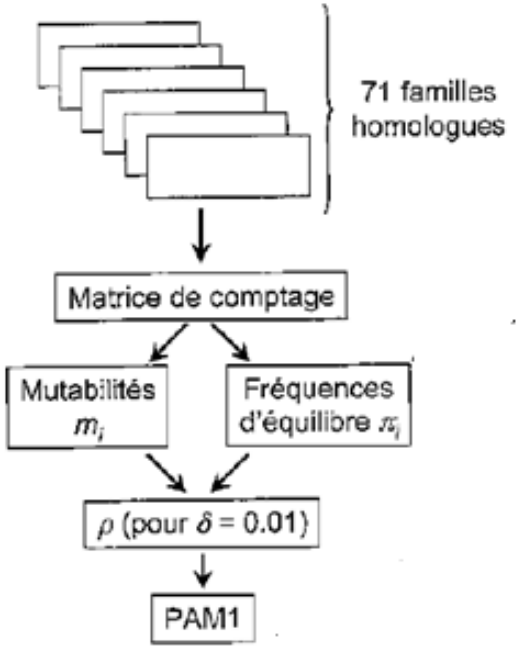
Hypothèse : processus d'évolution des protéines suit un modèle markovien de substitution d'ordre 0.

Choix des séquences : très proches (minimum 85% d'identité entre chaque paire de séquences de manière à éviter la présence de substitutions multiples).

Problème du modèle PAM : construit à partir d'un jeu de données limité.
Mise à jour par Jones *et al.* En utilisant un nombre de protéines beaucoup plus important (16300) mais en conservant une approche similaire.
Modèle et matrices correspondantes connus sous le nom de JTT (Jones, Taylor et Thronton).

Modèles d'évolution pour les séquences protéiques

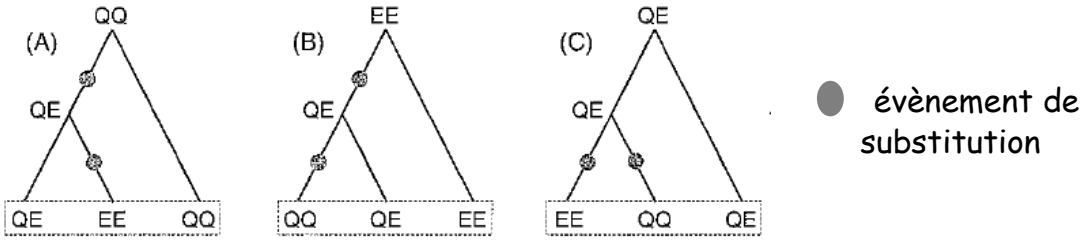
Principes généraux de la construction de la PAM



➢ Comptage du nombre de substitution pour chaque paire d'acides aminés :

Procédure :

- pour chaque famille alignement multiple des séquences et construction d'un arbre au moyen de la méthode du maximum de parcimonie qui permet de reconstruire les séquences ancêtres à chaque nœud de l'arbre.



Extrait de Perrière et Brochier-Armanet (2010) Concepts et méthodes en phylogénie moléculaire.

- Alignement de chaque séquence avec sa séquence ancêtre et comptage du nombre n_{ij} de substitution de l'acide aminé *i* vers l'acide aminé *j* et du nombre de conservations : obtention de la matrice 20x20 de résultats brutes qui est symétrique car on fait l'hypothèse de réversabilité $n_{ij} = n_{ji}$.

Modèles d'évolution pour les séquences protéiques

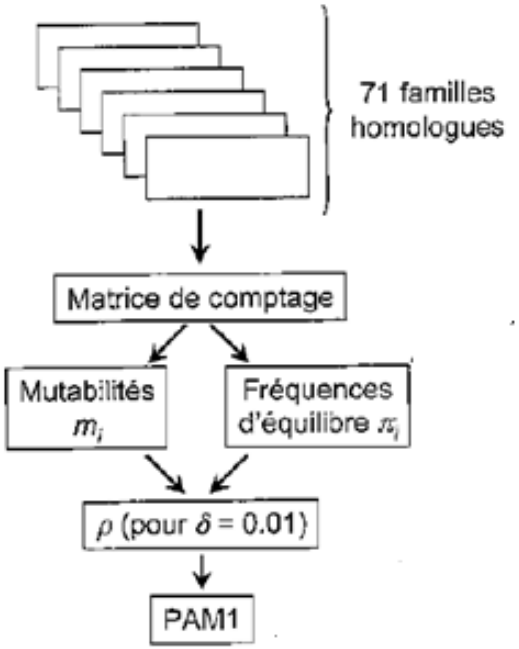
Exemple d'une matrice de cumul des mutations acceptées (x10)

	ala	arg	asn	asp	cys	gln	glu	gly	his	ile	leu	lys	met	phe	pro	ser	thr	trp	tyr	val
A																				
R	30																			
N	109	17																		
D	154	0	532																	
C	33	10	0	0																
Q	93	120	50	76	0															
E	266	0	94	831	0	422														
G	579	10	156	162	10	30	112													
H	21	103	226	43	10	243	23	10												
I	66	30	36	13	17	8	35	0	3											
L	95	17	37	0	0	75	15	17	40	253										
K	57	477	322	85	0	147	104	60	23	43	39									
M	29	17	0	0	0	20	7	7	0	57	207	90								
F	20	7	7	0	0	0	17	20	90	167	0	17								
P	345	67	27	10	10	93	40	49	50	7	43	43	4	7						
S	772	137	432	98	117	47	86	450	26	20	32	168	20	40	269					
T	590	20	169	57	10	37	31	50	14	129	52	200	28	10	73	696				
W	0	27	3	0	0	0	0	0	3	0	13	0	0	10	0	17	0			
Y	20	3	36	0	30	0	10	0	40	13	23	10	0	260	0	22	23	6		
V	365	20	13	17	33	27	37	97	30	661	303	17	77	10	50	43	186	0	17	

Nombre de fois où C (cys) est muté : 280
Nombre de fois où V (val) est muté : 2003

Pourquoi cette différence ?
V plus fréquent que C dans l'échantillon ?
V mute plus fréquemment que C ?

Modèles d'évolution pour les séquences protéiques



➢ Calcul de la *mutabilité* de chaque acide aminé *i*

La mutabilité est défini comme le rapport entre le nombre de substitutions affectant l'acide aminé *i* et le nombre d'acide aminé *i* observé dans les données :

$$m_i = \frac{\sum_{i \neq j} n_{ij}}{\sum_j n_{ij}}$$

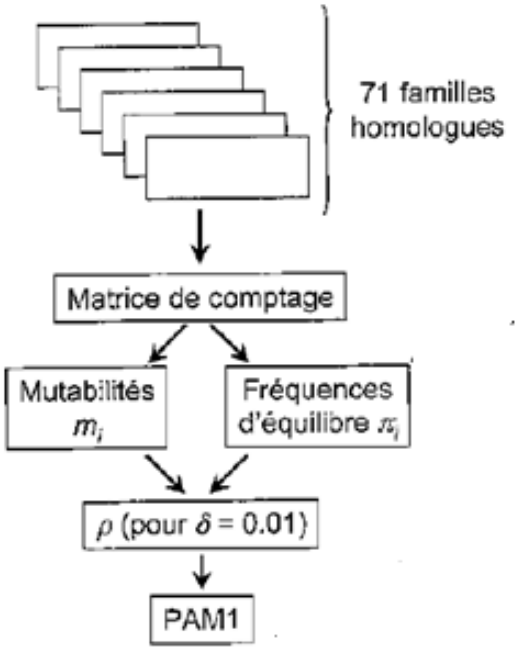
➢ Calcul de la probabilité de mutation de chaque paire d'acides aminés

Le calcul de la mutabilité nous indique si un acide aminé *i* est plus souvent (ou moins souvent) muté qu'un autre acide aminé *k.* Par contre, ceci ne nous indique pas quel est le pourcentage, par exemple, de *i* muté en *j.* Ce pourcentage correspond à la probabilité de mutation de la paire d'acides aminés *ij.* Cette valeur est donnée par :

$$p_{ij} = m_i \frac{n_{ij}}{\sum_{i \neq j} n_{ij}} = \text{mutabilité de i} \frac{\text{nombre d'acides aminés i muté en j}}{\text{nombre d'acides aminés i muté}}$$

Exemple : on à 100 acides aminés *R* et 10 sont mutés -> $m_R = 0,1$
5 *R* sont mutés en A -> $p_{RA} = 0,1 * 5/10 = 0,05$ (5%)
3 *R* sont mutés en V -> $p_{RV} = 0,1 * 3/10 = 0,03$ (3%)
2 *R* sont mutés en S -> $p_{RS} = 0,1 * 2/10 = 0,02$ (2%)

Modèles d'évolution pour les séquences protéiques



➢ <u>Prise en compte du temps : calcul de la matrice de probabilité PAM1</u>

LA matrice PAM1 estime les taux de substitution des acides aminées avec un taux de mutation attendu de 1 mutation fixée pour 100 sites, soit une distance évolutive de 0.01 (intervalle de temps évolutif pour permettre la fixation d'une mutation/100 sites).
Les valeurs de la matrice précédente ont donc été normalisées :

$$p_{ij_{i\neq j}} = \rho m_i \frac{n_{ij}}{\sum_{i\neq j} n_{ij}} \quad \text{et} \quad p_{ii} = 1 - \rho m_i$$

La valeur du facteur correctif ρ a été calculée pour que les fréquences de conservation des acides aminés (termes diagonaux de la matrice) représentent une conservation de 99%.

Modèles d'évolution pour les séquences protéiques

Principes généraux de la construction de la PAM

➢ famille de matrices : construction des PAMk

Comme on a fait l'hypothèse que la probabilité de mutation d'un acide aminé est indépendante de ce qui s'est produit à cette position dans le passé, on va pouvoir obtenir les probabilités de mutation pour des intervalles d'évolution plus grands par la multiplication de la PAM1 avec elle-même. Une PAMk sera obtenue en multipliant la PAM1 k fois par elle-même (k mutations acceptées pour 100 sites)

PAM2 = PAM1 x PAM1 = $PAM1^2$	intervalle d'évolution : 2 mutations acceptées pour chaque 100 résidus : d = 0.02
$PAM40 = PAM1^{40}$	intervalle d'évolution : 40 mutations acceptées pour chaque 100 résidus : d = 0.4
$PAM120 = PAM1^{120}$	intervalle d'évolution : 120 mutations acceptées pour chaque 100 résidus : d = 1.2
$PAM250 = PAM1^{250}$	intervalle d'évolution : 250 mutations acceptées pour chaque 100 résidus : d = 2.5

Modèles d'évolution pour les séquences protéiques

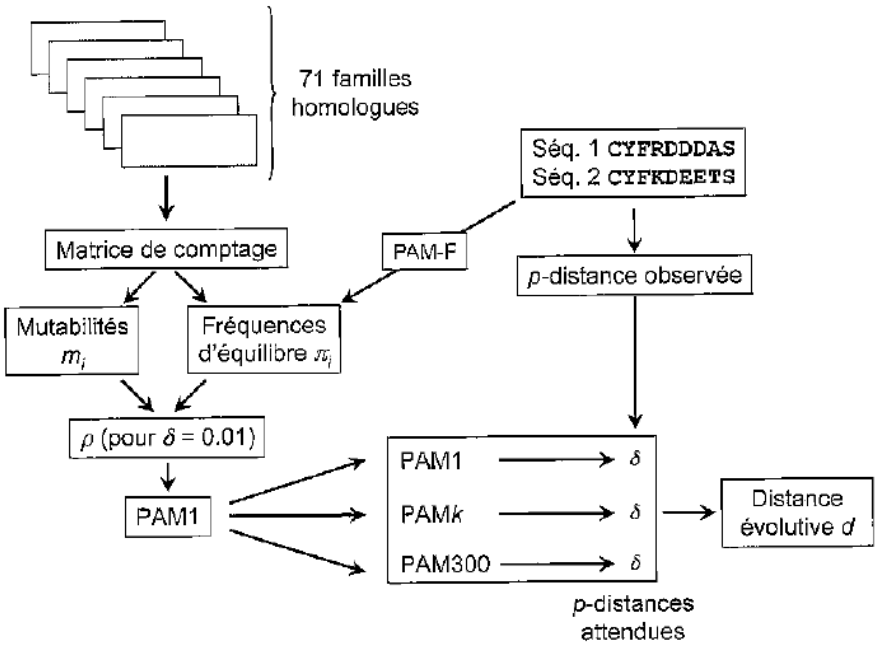
Modèle PAM : Calcul d'une distance PAM

Pour une matrice PAM donnée, on peut donc calculer δ la proportion attendue de sites qui ont été substitués entre deux séquences grâce à la relation suivante :

$$\delta = \rho \sum_i \pi_i m_i$$

avec : ρ le facteur d'échelle de la PAM, π_i la fréquence à l'équilibre de l'acide aminé *i* et m_i sa mutabilité

δ ne dépend que de ρ car les π_i et les m_i sont estimés à partir des données. δ correspond à la p-distance attendue.



On va choisir la PAM qui fournira une valeur de δ égale à la p-distance observée.

Connaissant alors la PAM, on pourra calculer la distance évolutive entre nos deux séquences.

Modèles d'évolution pour les séquences protéiques

Modèle WAG

Problème avec le modèle PAM ou JTT : utilisation de séquences très similaires pour estimer les taux de substitutions. Pour séquences distantes inférés.

➡ Utilisation du maximum de vraisemblance pour estimer les taux de substitutions des acides aminés.

Premières tentatives, modèles adaptés :

- aux séquences mitochondriales de vertébrés (mtREV).
- aux séquences mitochondriales de mammifères (mtMAM).
- aux séquences chloroplastiques (cpREV).

Modèle plus général proposé par Whelan et Goldman : modèle WAG

- utilise 182 familles de protéines homologues (3905 séquences).
- utilisation du modèle WAG permet d'obtenir des arbres dont la vraisemblance est significativement supérieure à ceux obtenus avec les modèles PAM ou JTT
- faiblesse du modèle : hypothèse d'uniformité (même vitesse d'évolution pour tous les sites)

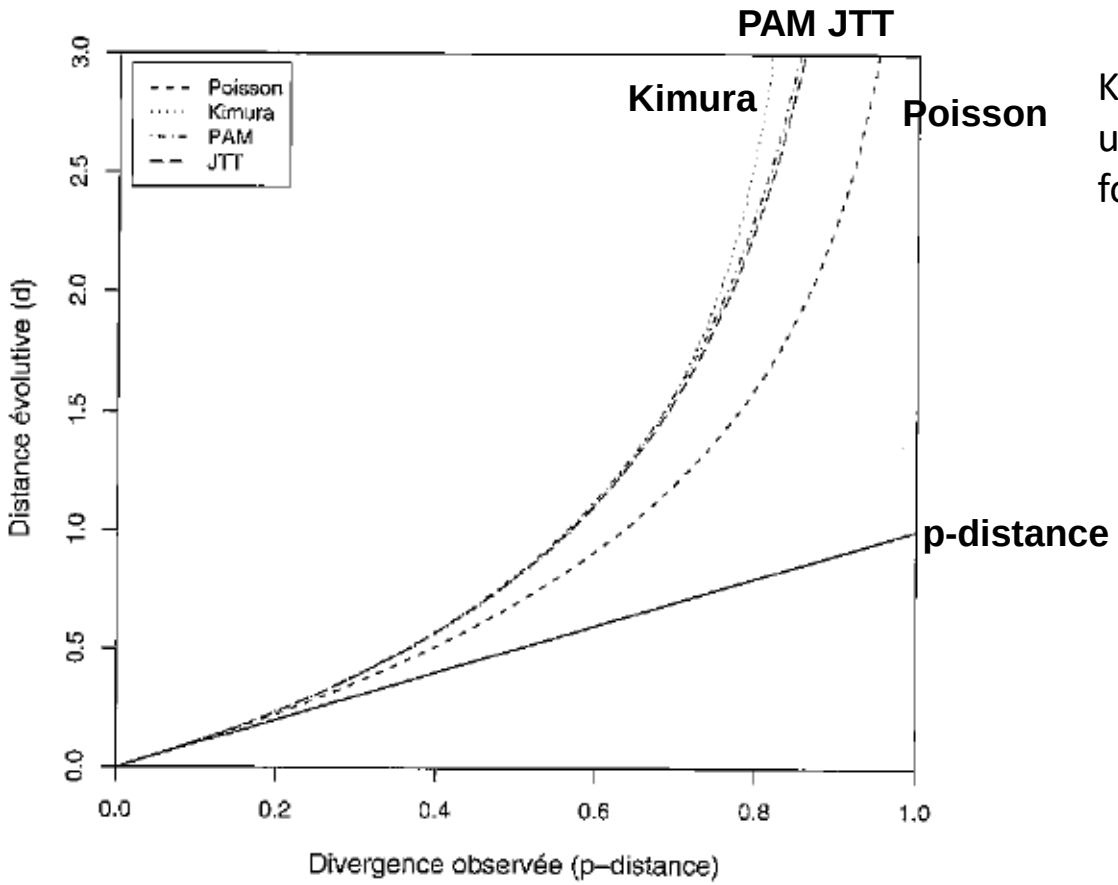
Modèle proposé par Le et Gascuel : modèle LG extension du modèle WAG

- prise en compte de différentes vitesses d'évolution pour les sites
- construit à partir de 3912 familles comprenant au total 49637 séquences

Choix d'un modèle évolutif

➢ Utilisation de séquences protéiques

- modèles les plus performants étant ceux bâtis sur le plus grand nombre de données (car estimation des taux de substitutions pour tous les modèles).
- modèles WAG et LG supérieurs aux modèles PAM et JTT.
- si distances évolutives faibles, on peut utiliser le modèle de Poisson ou de Kimura car aussi bons résultats. Donc si même résultat avec deux modèles, utiliser le plus simple car plus rapide.

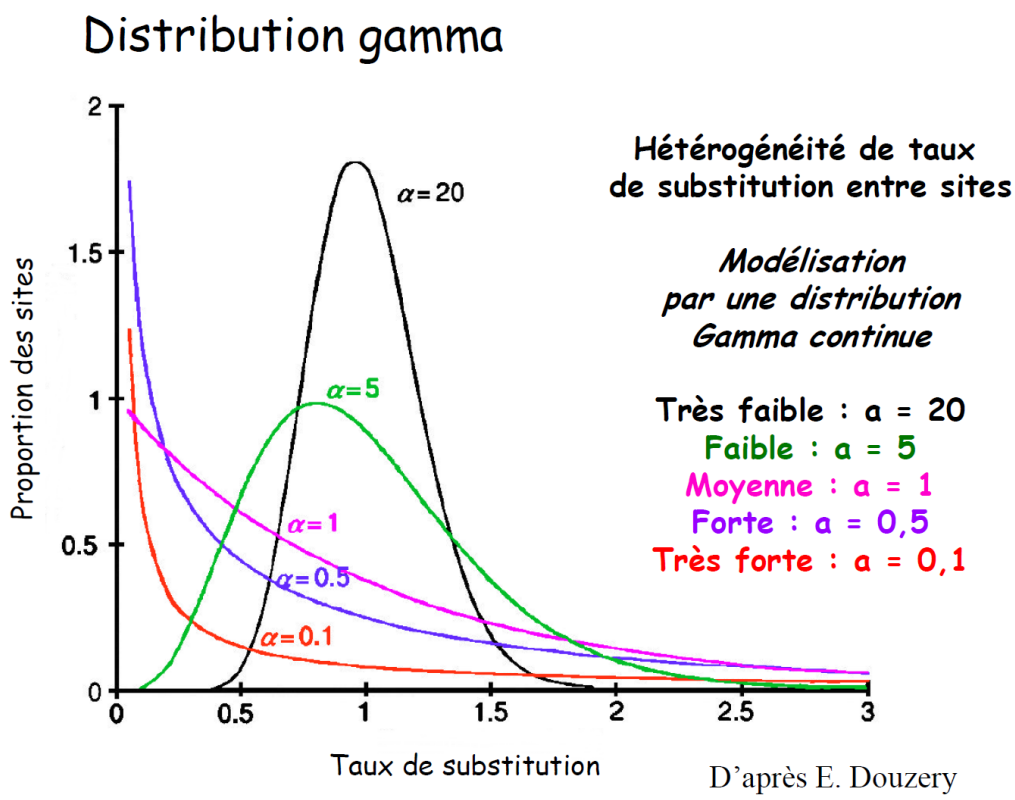


Kimura : modèle simplifié donnant une estimation de la distance PAM en fonction de la p-distance p :

$$d = -Log\left(1 - p - 0.2p^2\right)$$

Correction des distances pour différentes vitesses d'évolution

Hypothèse des différents modèles évolutifs présentés :
tous les sites évoluent à la même vitesse, or les contraintes fonctionnelles engendrent des taux d'évolution (*r*) différents selon les sites. Il a été démontré que ce taux *r* est modélisable par une loi Gamma (séquences nucléiques ou protéiques). Choix de la distribution Gamma : pas de justification biologique mais commodité mathématique car la forme de la distribution ne dépend que d'un seul paramètre a.



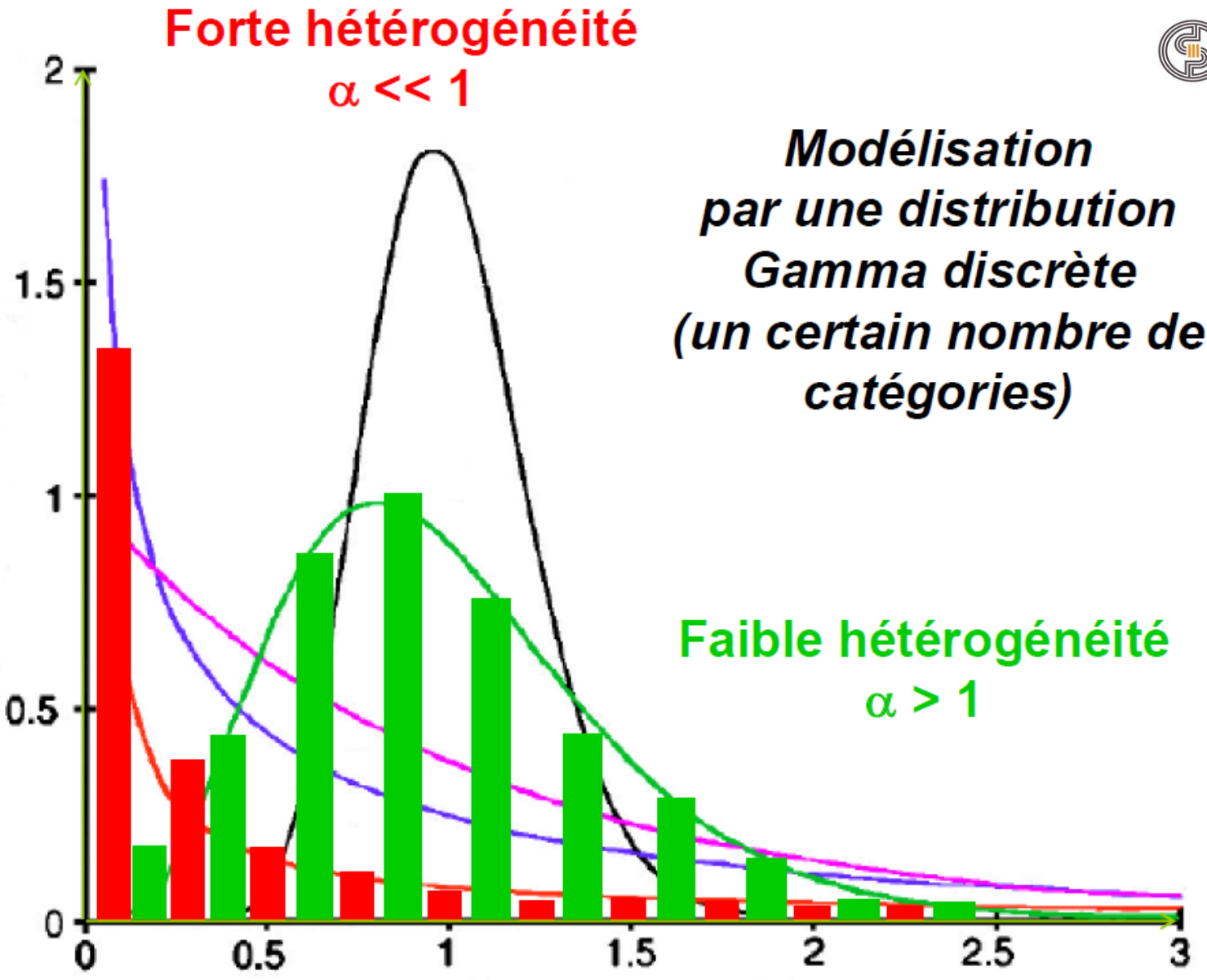
D'après E. Douzery

Si $\alpha > 1 \rightarrow$ forme de cloche.
Plus α est grand, plus la variance de *r* diminue traduisant une faible hétérogénéité des taux de substitutions par rapport à la moyenne.

Si $\alpha \leq 1 \rightarrow$ forme de L.
Nombre important de sites avec un r proche de 0 (sites quasiment invariants).
Donc forte hétérogénéité dans les taux d'évolution.

α est estimé à partir des données.
Distribution Gamma est discrétisée (nombre de catégories pour *r* variant de 4 à 8).

Correction des distances pour différentes vitesses d'évolution



Correction des distances pour différentes vitesses d'évolution

La plupart des modèles vus précédemment peuvent intégrer dans leur calcul de la distance une correction par la loi Gamma.

Exemple séquences nucléiques : le modèle de Jukes et Cantor (JC89) qui s'identifie par JC89+Γ

Modèle JC89

$$d = -\frac{3}{4} Log\left(1 - \frac{4}{3} p_{dist}\right)$$

Modèle JC89+Γ

$$d = \frac{3}{4}\alpha\left[\left(1 - \frac{4}{3} p_{dist}\right)^{-1/\alpha} - 1\right]$$

Exemple séquences protéiques : le modèle de Poisson (Poisson+Γ)

Modèle Poisson

$$d = -Log(1 - p)$$

Modèle Poisson+Γ

$$d = \alpha\left[(1 - p)^{-1/\alpha} - 1\right]$$

α = 2.25 distance peu différentes de celles obtenues avec le modèle PAM
α = 2.4 distance peu différentes de celles obtenues avec le modèle JTT

Correction des distances pour différentes vitesses d'évolution

UNIVERSITÉ TOULOUSE III PAUL SABATIER

Exemple de l'estimation du taux de substitution par site : chaîne α hémoglobine

	P-distance	PC-distance	PC + Gamma-distance
Human/cow	0.121	0.129	0.134
Human/kangaroo	0.186	0.205	0.216
Human/carp	0.486	0.665	0.789

PC = correction de poisson

Choix des modèles évolutifs

Choix d'un modèle évolutif

Des méthodes permettant de tester l'adéquation du modèle aux données existent mais souvent le choix du modèle est du fait de l'utilisateur et de ses connaissances.

Quelques règles simples :

- construction d'une phylogénie à partir de gènes protéiques :
  - séquences très distantes dans l'évolution : utilisation des séquences protéiques.
  - séquences proches dans l'évolution : utilisation des séquences acides nucléiques voir travailler uniquement sur les positions synonymes.

- Utilisation de séquences nucléiques : grand nombre de modèles
  - critère important : le degré de divergence entre les séquences.
  - pas toujours pertinent d'utiliser les modèles avec beaucoup de paramètres :
    - si les séquences sont courtes ou trop similaires les estimations des paramètres sont mauvaises.
    - modèle arrivant à saturation plus rapidement (cf. figure suivante) donc si séquences très divergentes, fréquemment impossible de calculer les distances.
  - donc si même résultat avec deux modèles, utiliser le plus simple car la variance de la distance augmente avec le nombre de paramètres.
  - application de la correction Gamma que si nombre de sites utilisés important car nécessite d'estimer un paramètre supplémentaire (la forme a de la distribution).

Choix des modèles évolutifs

Grand nombre de modèles d'évolution dont certains très complexes et intégrant un grand nombre de paramètres.

Problème : la précision de l'estimation des paramètres peut être mauvaise notamment quand peu de données (nombre de séquences et/ou de sites).

- Primordial de choisir le modèle qui est le plus en adéquation avec les données.

- Etape indispensable à toute analyse phylogénétique rigoureuse.

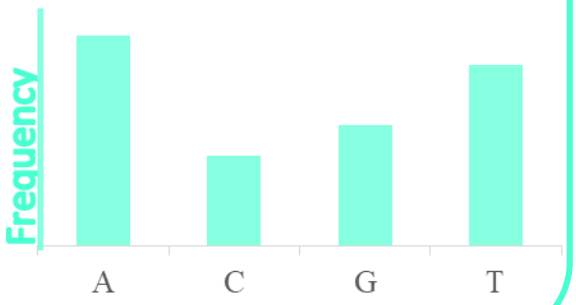
Les tests de vraisemblance sont des méthodes bien adaptées qui permettent non seulement de déterminer les hypothèses qui expliquent le mieux le jeu de données mais aussi de comparer des hypothèses.

- Test du rapport de vraisemblance appelé LRT pour Likelihood Ratio Test
- Akaike Information Criterion (AIC).

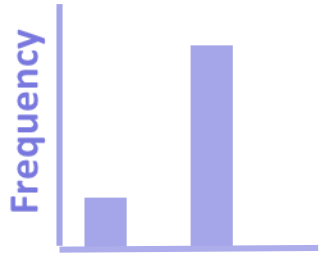
Choix des modèles évolutifs

UNIVERSITÉ TOULOUSE III PAUL SABATIER

Fréquences en bases (π)



Sites invariables (I)



CHOIX

Modèle d'évolution

Taux évolution des sites (Γ)

Likelihood Ratio Test

Nécessite que les modèles que l'on veut tester soit imbriqués (du plus simple au plus complexe)

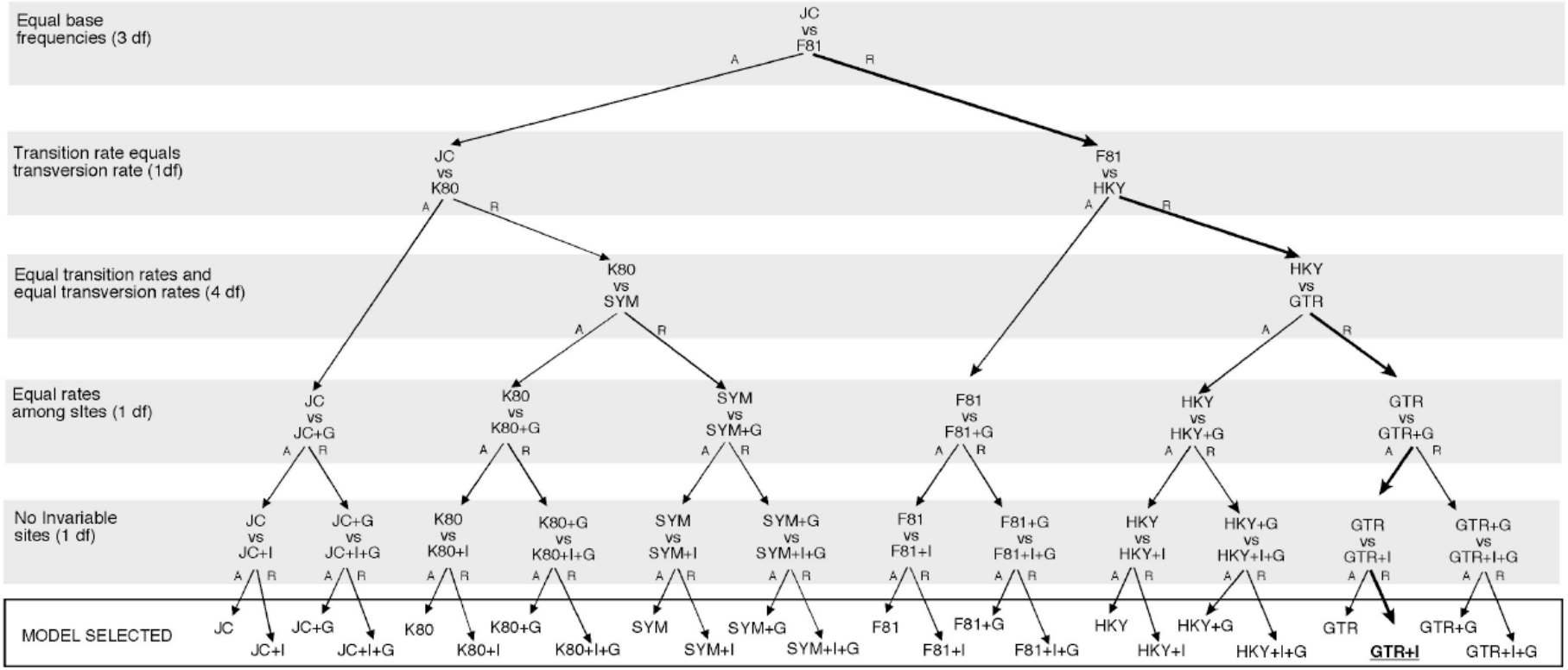


Figure 17. Example of a particular forward hierarchy of likelihood ratio tests for 24 models. At any level the null hypothesis (model on top) is either accepted (A) or rejected (R). In this example the model selected is GTR+I.

(Extrait du manuel de JModelTest)

Likelihood Ratio Test

Ce test est utilisé quand l'on désire comparer deux arbres qui ont la même topologie mais qui ont été obtenus avec des modèles d'évolution différents. On compare deux modèles :

• Le modèle M_0 (le plus simple, i.e. qui a le plus petit nombre de paramètres k_0) qui correspondra à l'hypothèse nulle H0.

• Le modèle M_1 (le plus complexe, k_1 paramètres, $k_1 > k_0$) qui correspondra à l'hypothèse alternative H1.

Le rapport de vraisemblance est donné par : $$\Delta = 2\ln\left[\frac{L(\Theta_1)}{L(\Theta_0)}\right] = 2\left[\ln L(\Theta_1) - \ln L(\Theta_0)\right]$$

Δ suit une loi du χ^2 à $k_1 - k_0$ d.d.l., soit le nombre de paramètre du modèle M_1 à contraindre pour se ramener au modèle M_0. Le modèle nul sera rejeté si Δ est supérieur au niveau de confiance fixé par l'utilisateur

Critiques majeures de ce test :

• la sélection des modèles testés dépend du parcours de l'arbre hiérarchique. Par exemple, si le modèle le plus adapté est le F81+I+G, il ne pourra pas être testé si à l'étape précédente on a rejeté le modèle F81 au profit du modèle HKY. Pour palier à ce problème, on peut faire des tests dynamiques.

• le choix du modèle se fait sur la base d'un arbre dont la topologie est fixée. Si celle-ci n'estime pas bien l'histoire évolutive des données, les vraisemblances obtenues peuvent être irréalistes. Or le choix du modèle précède cette reconstruction.

Akaike Information Criterion (AIC)

C'est un estimateur qui correspond à la minimisation de la distance attendue entre un modèle vrai et son estimation. Les modèles correspondant aux valeurs minimales de l'AIC sont considérés comme les plus appropriés pour la reconstruction. Une même topologie de référence doit être utilisée pour tester les différents modèles. L'AIC permet de tester des modèles sans que ceux-ci soient imbriqués.

$$AIC = -2\ln L(\Theta) + 2k$$

k = nombre de paramètres libres du modèle

L'*AIC* apparaît biaisé pour les modèles riches en paramètres comparativement au LRT.

Si la taille *n* du jeu de données est petite comparée au nombre de paramètres *k* du modèle ($n/k < 40$) l'utilisation de l'*AIC* corrigé AIC_c est recommandée.

$$AIC_c = AIC + \frac{2k(k+1)}{n-k-1}$$

Deux logiciels : **JModelTest** pour les séquences d'acides nucléiques et **ProtTest**

Aujourd'hui, nouvelle implémentation de ces deux logiciels dans ModelTest-NG (Darriba *et al.*, 2020, MBE, 37:291-94)

Démonstration du calcul de la distance de Jukes et Cantor

Calcul des distances entre deux séquences nucléiques

Modèle de Jukes et Cantor (abrégé JC69)

Pour calculer la distance entre deux séquences, il faut que l'on trouve une relation entre *d* et la probabilité d'observer une substitution à un site qui est donnée par la distance observée ou *p*-distance.

On a : $q(t) + 3p(t) = 1$
Il est possible de résoudre ce système en diagonalisant la matrice M afin d'obtenir une solution analytique pour les valeurs de p(t). Mais comme le modèle est simple, autre solution.

Soit $q(t+\Delta(t))$ la probabilité qu'un nucléotide quelconque présente le même état *i* après un intervalle de temps $t+\Delta(t)$. Deux possibilités :

- soit au temps *t* la base considérée était dans l'état *i* (probabilité q(t)), elle restera dans cet état avec un taux de conservation $1-\lambda$ (or $\lambda = 3\alpha$) donc avec un taux égal à $1-3\alpha$. Donc la probabilité de ce scénario est $q(t)(1-3\alpha)$.

- soit au temps *t* la base est dans un autre des trois états possibles (probabilité p(t)), elle pourra être substituée dans l'état *i* avec un taux de 3α. La probabilité associée étant $p(t)(3\alpha)$.

Calcul des distances entre deux séquences nucléiques

UNIVERSITÉ TOULOUSE III PAUL SABATIER

Modèle de Jukes et Cantor (abrégé JC69)

$q(t+\Delta(t)) = (1-3\alpha)q(t) + 3\alpha p(t) = (1-3\alpha)q(t) + \alpha\ [1-q(t)]$ car $3p(t) = 1-q(t)$

A partir de cette équation, il est possible d'exprimer la probabilité $q(\Delta t)$ que la base considérée soit dans l'état i après un intervalle de temps Δt :

$$\Delta q(t) = q(t+\Delta(t)) - q(t) = q(t) - 3\alpha q(t) + \alpha - \alpha q(t) - q(t)$$

$$\Delta q(t) = -4\alpha q(t) + \alpha$$

Si $\Delta(t) \rightarrow 0$, on peut en déduire l'équation différentielle :

$$\frac{dq(t)}{dt} = -4\alpha q(t) + \alpha$$

Équation différentielle linéaire du premier ordre

$$q'(t) + 4\alpha\ q(t) = \alpha$$

Calcul des distances entre deux séquences nucléiques

UNIVERSITÉ TOULOUSE III PAUL SABATIER

Modèle de Jukes et Cantor (abrégé JC69)

$q'(t) + 4\alpha\ q(t) = \alpha$

On résout d'abord l'équation sans second membre : $$\frac{dq(t)}{dt} + 4\alpha q(t) = 0$$

On se ramène à une équation à variables séparables : $$\frac{dq(t)}{q(t)} = -4\alpha dt$$

On intègre : $$\int \frac{dq(t)}{dt} = \int -4\alpha dt \Rightarrow Log\left|\frac{q(t)}{C}\right| = -4\alpha t$$

On peut écrire $$e^{Log\left|\frac{q(t)}{C}\right|} = \frac{q(t)}{C} = e^{-4\alpha t}$$

soit $$\boxed{q(t) = Ce^{-4\alpha t}}$$

Calcul des distances entre deux séquences nucléiques

Modèle de Jukes et Cantor (abrégé JC69)

Recherche d'une solution particulière :

On pose q(t) = a car le second membre est une constante (α) et donc q'(t) (dérivée de q(t)) est égal à 0.

On a donc : $4\alpha a = \alpha$ soit $a = \frac{1}{4}$

Solution générale de l'équation différentielle (solution sans second membre + solution particulière) :

$$q(t) = Ce^{-4\alpha t} + \frac{1}{4}$$

A t=0 on a q(t) =1 (aucune substitution ne s'est encore produite) et l'équation devient :

$$1 = C + \frac{1}{4} \qquad \text{soit} \quad C = \frac{3}{4}$$

On remplace C dans la solution générale :

$$q(t) = \frac{3}{4}e^{-4\alpha t} + \frac{1}{4}$$

Calcul des distances entre deux séquences nucléiques

Modèle de Jukes et Cantor (abrégé JC69)

$$q(t) = \frac{3}{4} e^{-4\alpha t} + \frac{1}{4}$$

On peut donc en déduire p(t) car q(t) + 3p(t) =1

$$p(t) = \frac{1}{3}\left[1 - \frac{3}{4} e^{-4\alpha t} - \frac{1}{4}\right] \quad \text{soit} \quad p(t) = \frac{1}{4}\left[1 - e^{-4\alpha t}\right]$$

Considérons la probabilité d'observer une substitution à un site entre deux séquences, l'estimation de cette probabilité est donnée par la distance observée, c'est-à-dire la p-distance. Le temps de divergence séparant deux séquences homologues est de 2t.
On a donc:

p_{dist} = 3p(2t) (3 possibilités de substitution avec la même probabilité de substitution)

En remplaçant p(2t) par la valeur obtenue ci-dessus on obtient : $p_{dist} = \frac{3}{4} - \frac{3}{4} e^{-8\alpha t} = \frac{3}{4}(1 - e^{-8\alpha t})$

soit $e^{-8\alpha t} = \left(1 - \frac{4}{3} p_{dist}\right)$ et $8\alpha t = -Log\left(1 - \frac{4}{3} p_{dist}\right)$

d'où $\alpha t = -\frac{1}{8} Log\left(1 - \frac{4}{3} p_{dist}\right)$

On a donc obtenu une relation entre le taux de substitution instantané et la p-distance

Calcul des distances entre deux séquences nucléiques

Modèle de Jukes et Cantor (abrégé JC69)

Calcul de la distance évolutive entre deux séquences dans le cadre de ce modèle :

On a vu précédemment que : $$d = 2\sum_i \pi_i \lambda_i t$$

Quand t → ∞ q(t) → ¼ et p(t) → ¼ (car $e^{-4\alpha t}$ vaut 0). A l'équilibre les fréquences des quatre bases sont donc toutes égales à ¼.
De plus, quelque soit *i* λ_i est égal à λ donc la formule de la distance devient :

$$d = 2 \times 4\left(\frac{1}{4}\lambda t\right) = 2\lambda t$$ or $\lambda = 3\alpha$ donc en remplaçant λ dans l'équation on a $d = 6\alpha t$

Précédemment on a vu que $$\alpha t = -\frac{1}{8} Log\left(1 - \frac{4}{3} p_{dist}\right)$$

En remplaçant dans *d* on obtient la distance de Jukes et Cantor :

$$d = -\frac{3}{4} Log\left(1 - \frac{4}{3} p_{dist}\right)$$

Un facteur correcteur est donc apporter à la p-distance p_{dist}