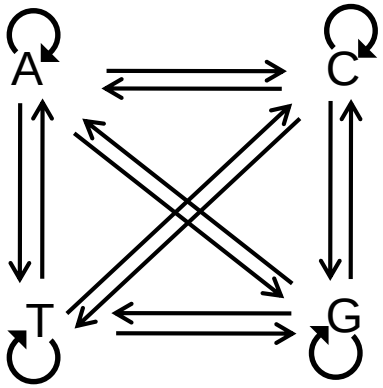
Calcul d'une distance génétique (évolutive) entre deux séquences

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

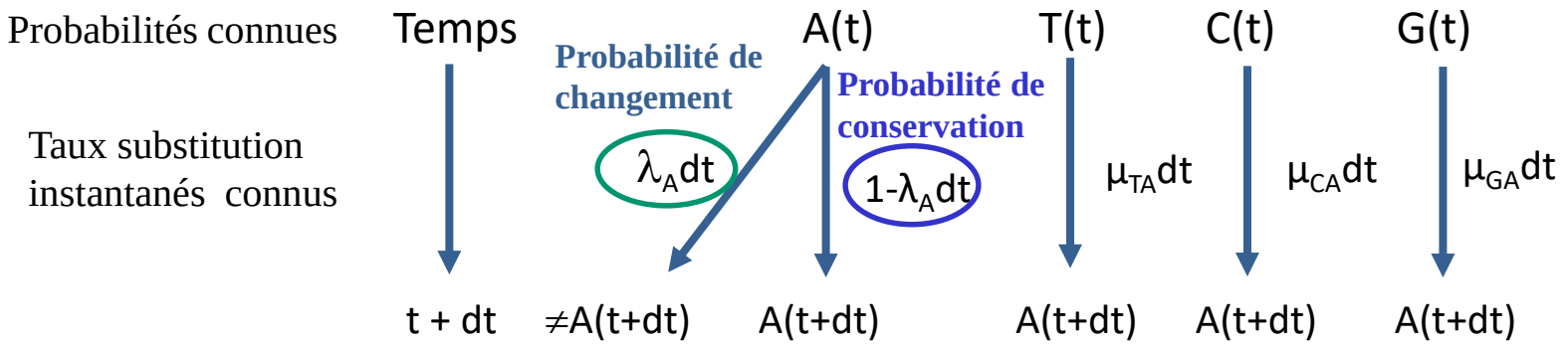
Hypothèses liées au modèle markovien

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:

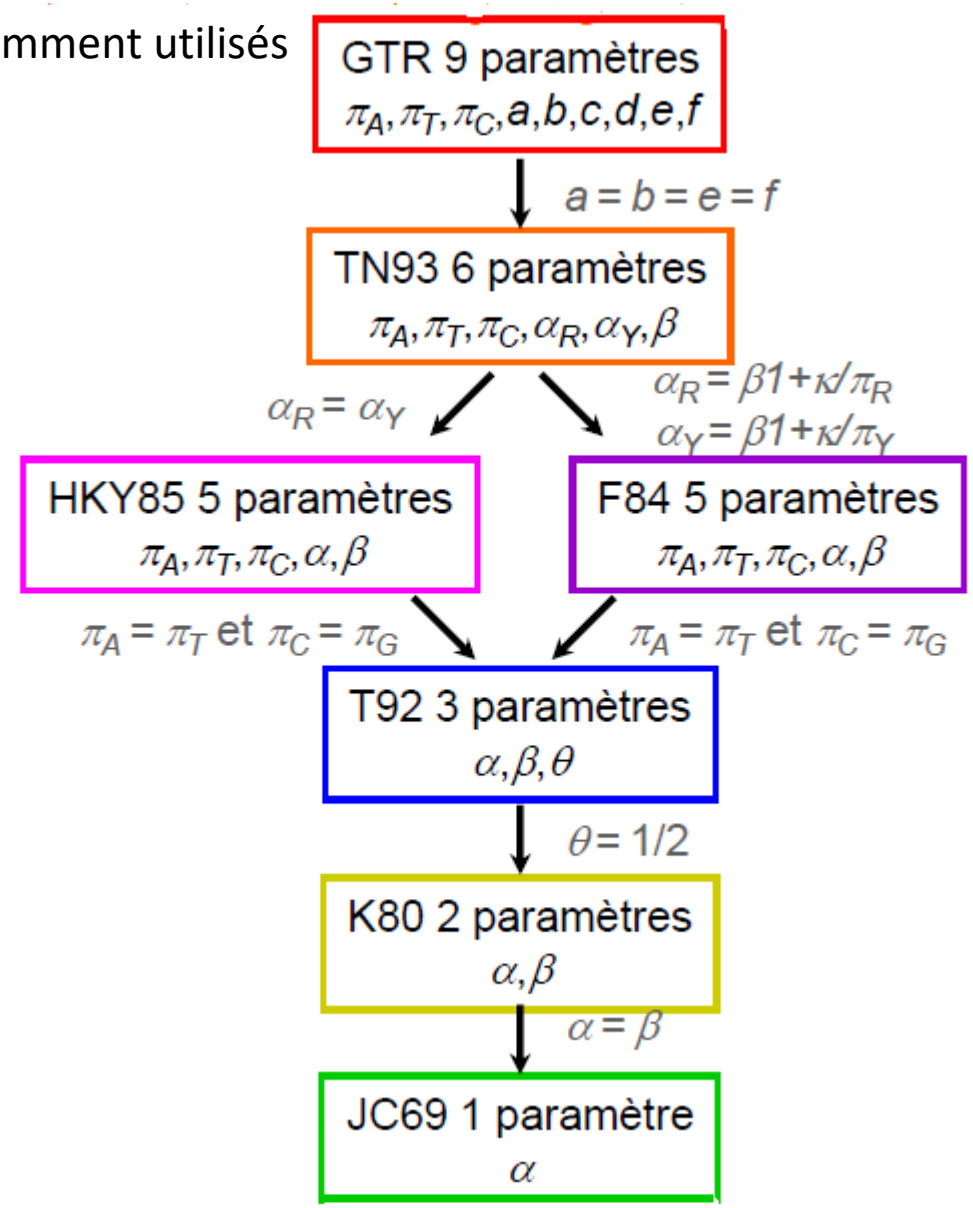
$$M = \begin{bmatrix} -\lambda_A & \mu_{AT} & \mu_{AC} & \mu_{AG} \\ \mu_{TA} & -\lambda_T & \mu_{TC} & \mu_{TG} \\ \mu_{CA} & \mu_{CT} & -\lambda_C & \mu_{CG} \\ \mu_{GA} & \mu_{GT} & \mu_{GC} & -\lambda_G \end{bmatrix}$$

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.

Modèles d'évolution ADN/ARN

Modèles couramment utilisés



Modèles d'évolution pour les séquences protéiques

Modèle de Poisson

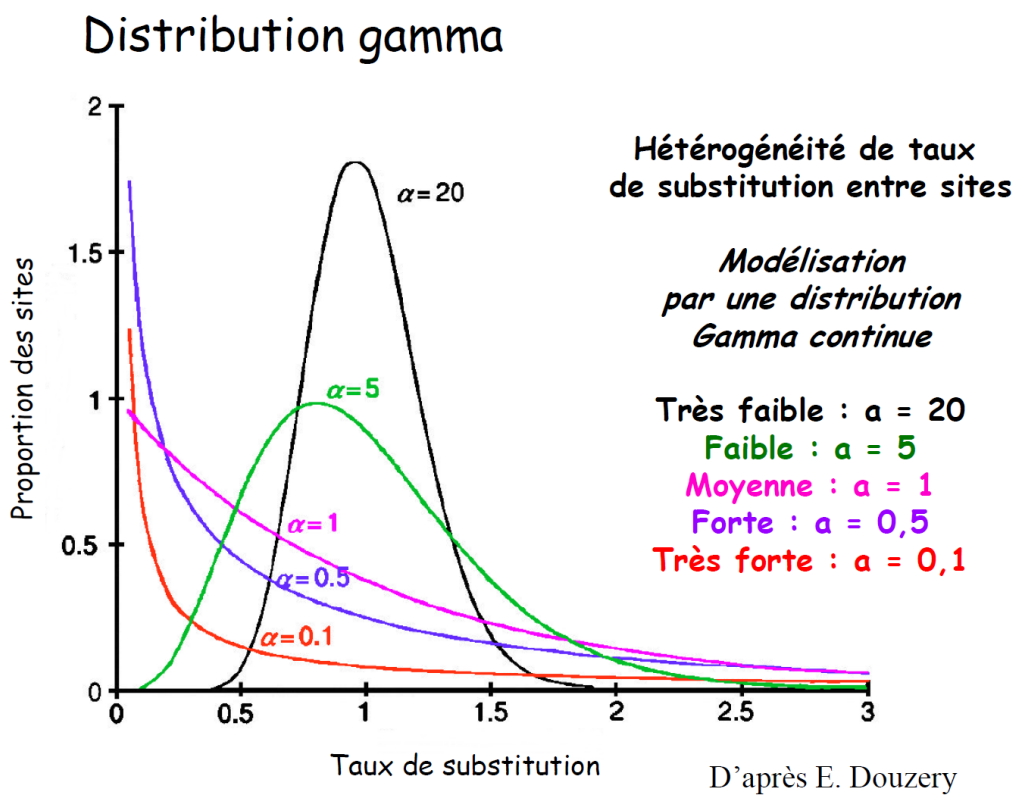
- Cependant vision très simplificatrice car en particulier :
  - taux de substitutions plus ou moins élevé en fonction de l'importance fonctionnelle du site
  - présence aussi de substitution parallèle et de réversion donc on va sous-estimer la distance entre deux séquences
  - ne peut être utilisée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

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.



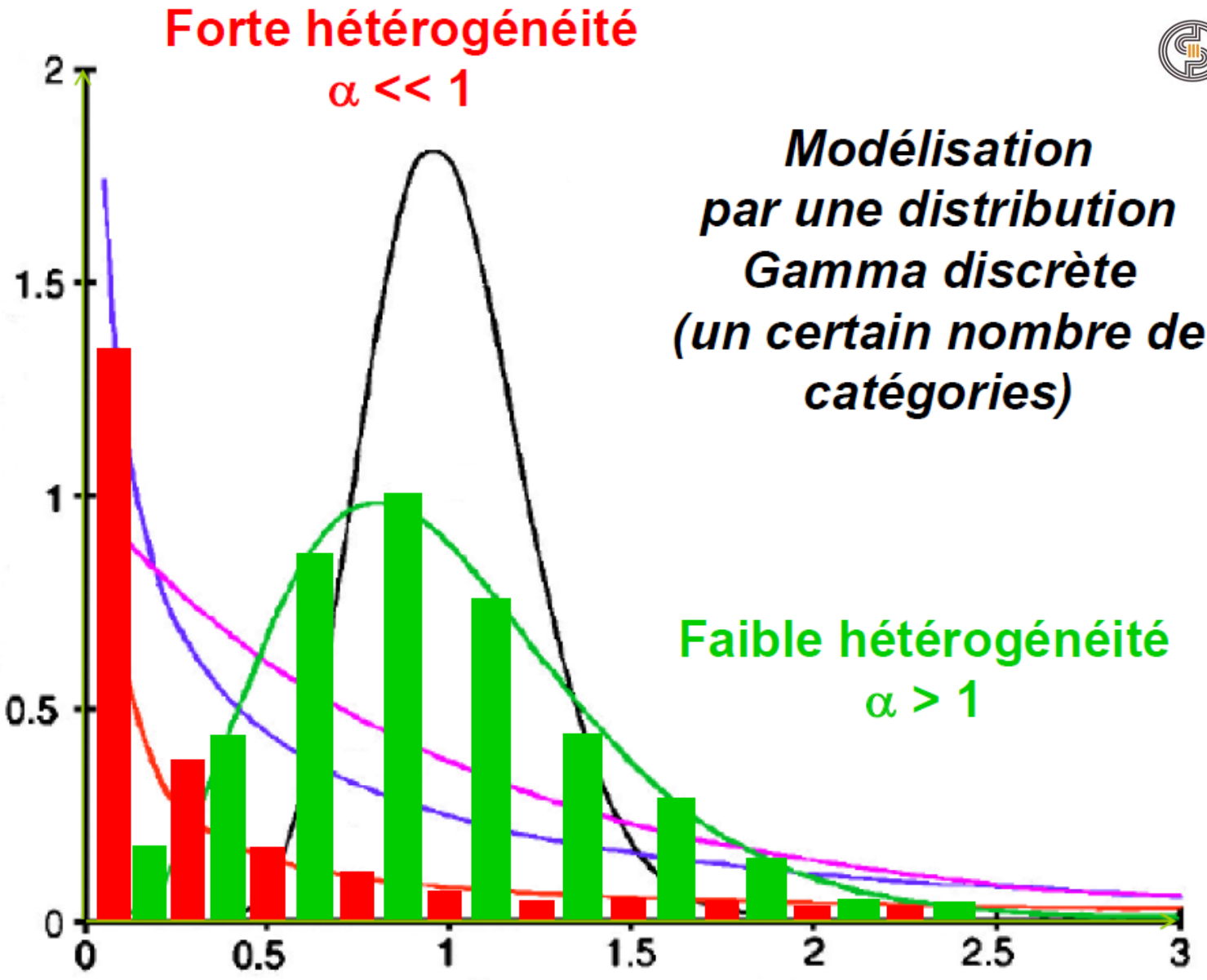
D'après E. Douzery

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

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 : $d_N \approx d_S$

An Empirical Codon Model for Protein Sequence Evolution

(Kosiol *et al.*, 2007, Mol. Biol. Evol. 24(7):1464-1479)

- ✓ Codon-level models are able to make distinctions between codons, which encode the same amino acid and those that do not.
- ✓ Allow to study if mutations maintaining the encoded amino acid (synonymous changes) is less, equally or ore frequently accepted by selection than nonsynonymous mutations.

→ Introduction of parameters describing the ratio of nonsynonymous to synonymous changes. Allows to measure the effect of natural selection on the sequence.

- ✓ Empirical models do not explicitly consider biological factors that shape protein evolution but simply attempt to summarize the substitution patterns observed in large quantities of data.
- ✓ These substitution patterns are described by parameters that aggregate all kinds of physicochemical properties of the amino acids and of their interaction with their local environment.
- ✓ These parameters are estimated once from a large data set and subsequently reused with the assumption that they are applicable to a wide range of sequence data sets.

An Empirical Codon Model for Protein Sequence Evolution

Previous codon models assumes that every mutation alters just one nucleotide.

If evolutionary change between two codons implies 2 or 3 nucleotides, it will result from a succession of independent single nucleotide mutation.

It specifies the relative instantaneous substitution rate from codon *i* to codon *j* as *(*i ≠ *j*):

$$q_{ij} = \begin{cases} 0 & \text{if } i \text{ or } j \text{ is a stop codon or} \\ & i \to j \text{ requires} > 1 \text{ nt substitution,} \\ \pi_j & \text{if } i \to j \text{ is a synonymous transversion,} \\ \pi_j \kappa & \text{if } i \to j \text{ is a synonymous transition,} \\ \pi_j \omega_M & \text{if } i \to j \text{ is a nonsynonymous transversion,} \\ \pi_j \kappa \omega_M & \text{if } i \to j \text{ is a nonsynonymous transition.} \end{cases}$$

where : κ is the transition-transversion rate ratio

ω_M represents the nonsynonymous/synonymous ratio

π_j is the equilibrium frequency of codon *j* with $\sum_{j=1}^{61} \pi_j$

An Empirical Codon Model for Protein Sequence Evolution

New model : taking into account mutations of 2 or 3 nucleotides at a time.
As previously, the definition of the relative instantaneous substitution rate from codon *i* to codon *j* as *(*i ≠ *j*) in the ECM (Empirical Codon Model) is calculated as :

$$q_{ij} = \begin{cases} 0 & \text{if } i \text{ or } j \text{ is a stop codon} \\ s_{ij}^{*}\pi_j\kappa(i,j) & \text{if } i \rightarrow j \text{ is a synonymous change} \\ s_{ij}^{*}\pi_j\kappa(i,j)\omega & \text{if } i \rightarrow j \text{ is a nonsynonymous change.} \end{cases}$$

where : $\kappa(i,j)$ is the transition-transversion bias between codon *i* and *j*
ω represents the nonsynonymous/synonymous ratio
π_j is the equilibrium frequency of codon *j* estimated from each particular data set analyzed
S_{ij}^{*} is the substitution rate estimated from the Pandit database (data set analyzed)

The Pandit database: each protein family includes an alignment of protein sequences, the corresponding alignment of DNA sequences. Each alignment have an estimated associated phylogenetic tree. 7332 protein families were used to trained the models

An Empirical Codon Model for Protein Sequence Evolution

$\kappa(i,j)$ in the previous equation represents a measure of the relative strength of the transition–transversion bias with respect to the average level implicit in the Pandit database.

Taking into account double and triple nucleotide changes leads to new scenarios in addition to the single transitions or single transversions inherent in single nucleotide changes.

- 9 possible ways to combine transitions (ts) and transversions (tv) in multiple nucleotide changes within 1 codon :
  - 1 nucleotide change : (1ts, 0tv); (0ts, 1tv)
  - 2 nucleotide changes : (2ts, 0tv); (1ts, 1tv); (0ts; 2tv)
  - 3 nucleotide changes : (3ts, 0tv); (2ts, 1tv); (1ts, 2tv); (0ts, 3tv)

Thus the transition–transversion bias may now be modeled as a function $\kappa(i, j)$ that depends on the numbers of transitions (nts) and transversions (ntv) of the change from codon *i* to codon *j*.

6 formulations for $\kappa(i, j)$ have been proposed among which the MG2K model of IQ-TREE .

An Empirical Codon Model for Protein Sequence Evolution

the MG2K model of IQ-TREE : ECM + F + ω +2κ

In this model, transitions and transversions are modeled with individual parameters (κ_1 for transitions and κ_2 for transversions) and the effect is seen as multiplicative in terms of the relative rates:

$$\kappa(i,j) = \kappa_1^{n_{ts}} \kappa_2^{n_{tv}}$$

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.

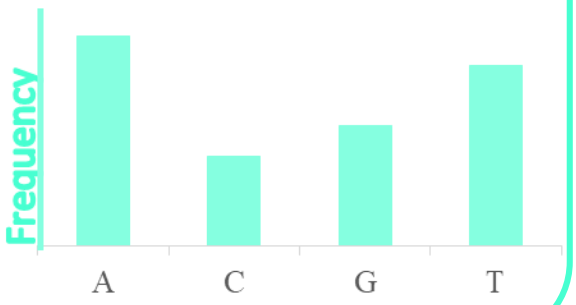
Les tests de vraisemblance sont des méthodes bien adaptées qui permettent non seulement de déterminer les hypothèses qui expliquent le mieux le jeu de données mais aussi de comparer des hypothèses.

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

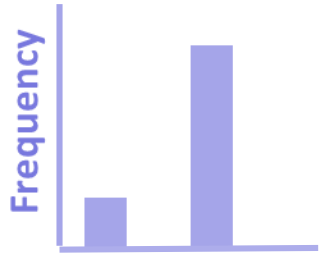
Choix des modèles évolutifs

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)

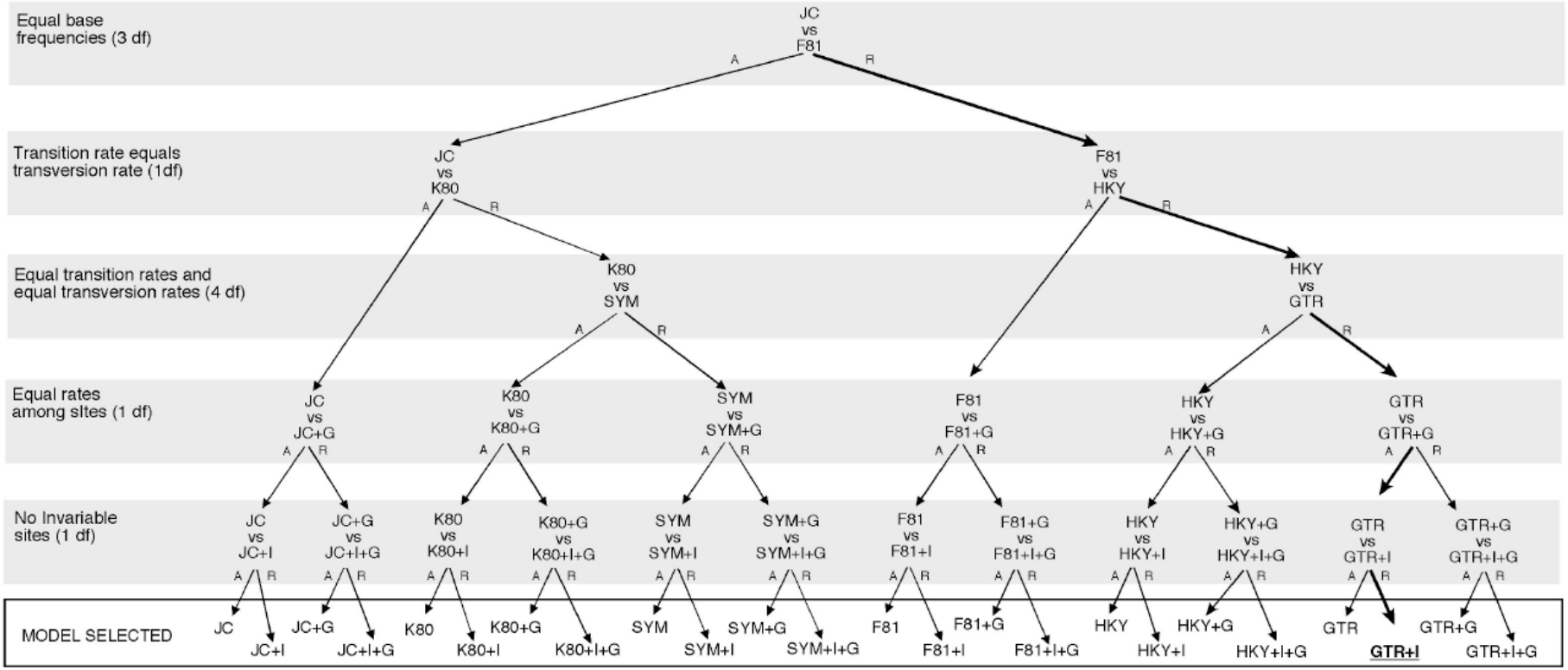


Figure 17. Example of a particular forward hierarchy of likelihood ratio tests for 24 models. At any level the null hypothesis (model on top) is either accepted (A) or rejected (R). In this example the model selected is GTR+I.

(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.

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