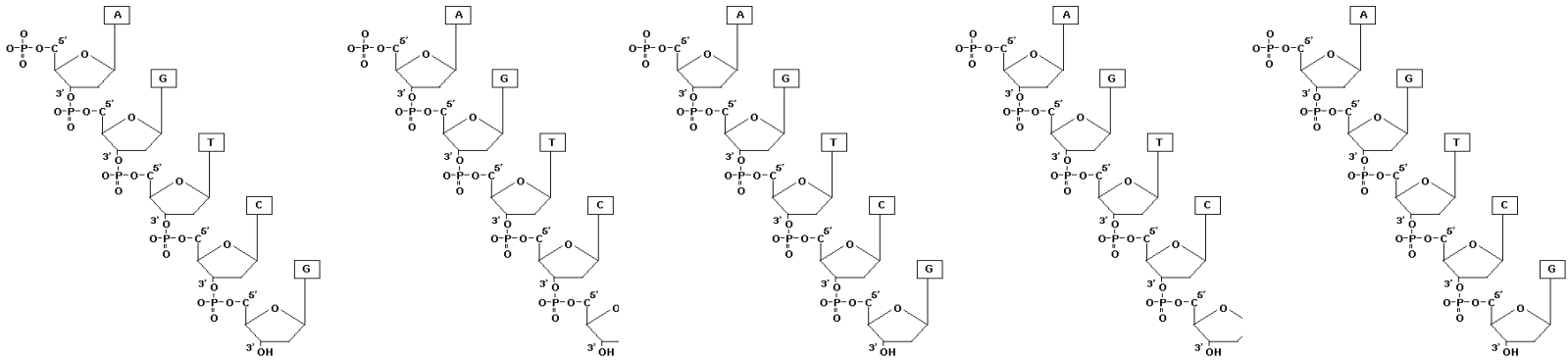


# Phylogénie et Evolution Moléculaire

10 – 14 septembre 2007



**Organisateurs:** Juan Montoya et Jan Pawlowski

**Assistants:** Fabien Burki et Loïc Pillet

**Horaires:** 9h15-12h15 / 13h30-17h30

**Fonctionnement:** Pour chaque sujet, une introduction théorique est suivie d'exercices pratiques. Les exercices se font sur PC individuels avec programmes et jeu de données installés.

**Attention:** Toutes configurations et données personnelles qui ne sont pas enregistrées sous **H**: sont perdues si vous éteignez l'ordinateur.  
- Ne pas éteindre l'ordinateur dans la journée.  
- Enregistrez vos données à la fin de chaque exercice.

**Conférences:** Quatre conférences.

**Evaluation:** Un rapport est demandé.  
La note est basée sur la participation et le rapport.

**Crédits:** 5 ECTS pour Masters et 3 ECTS pour Ecole doctorale E&E.

**Pour utiliser votre ordinateur:**

**User name:     *tpgenmol***

**Password:       *Az5uV9***

**Les données pré-installées se trouvent dans le serveur H:**

**H: \ Tpgenmol\PhylogénieTP \Données**

**Pour enregistrer vos résultats, vous devez créer un dossier à votre nom sous:**

**H: \ Tpgenmol\PhylogénieTP \Utilisateurs \**

# La phylogénie

## et l'étude de l'évolution moléculaire

**Relations évolutives entre organismes  
= topologie de l'arbre**

Premières méthodes de phylogénie  
(Maximum de Parcimonie, UPGMA)

- Améliore la recherche de la meilleure topologie
- Utilise la topologie pour extraire des informations concernant les processus d'évolution moléculaire

**=> Modèles d'évolution moléculaire**



# **La phylogénie moderne permet l'analyse de:**

- la relation évolutive entre organismes**

## **Mais aussi:**

- Duplication de gènes**
- Vitesse d'évolution**
- Pression de sélection**
- Variabilité génétique**
- Recombinaison**
- Temps de divergence entre lignées**
- Démographie**
- Etc...**

# **Théorie sur l'évolution moléculaire.**

# **Théorie sur l'évolution moléculaire.**

**- Les bases moléculaires de l'évolution.**

**- Evolution des séquences protéiques**

**- Evolution des séquences nucléotidiques**

# **Théorie sur l'évolution moléculaire.**

## **Les bases moléculaires de l'évolution.**

**1) Mécanismes de l'évolution**

**2) Structure des génomes**

**3) Structures et fonctions des gènes**

**4) Changements mutationnels dans l'ADN**

**5) Usage des codons dans les séquences codantes**

# 1) Mécanismes de l'Evolution

La première cause de l'évolution sont les changements mutationnels dans l'ADN, et en particulier dans les gènes ou les régions régulatrices.

Principaux changements mutationnels:

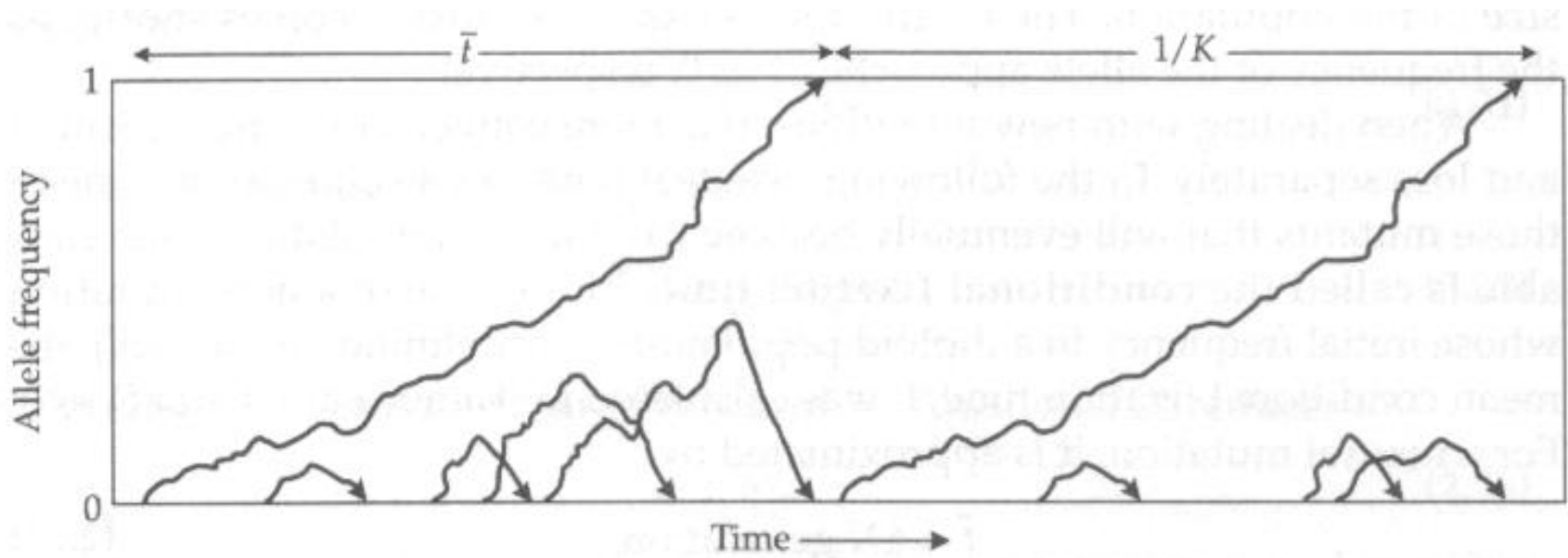
- substitution nucléotidique
- insertion/délétion de nucléotides (indels)
- recombinaison
- conversion génique
- duplications

Fixation d'une mutation dans une population/espèce:

Un gène muté peut se propager dans une population au cours des générations et peut éventuellement être fixé dans la population par dérive génétique et/ou par sélection naturelle.

# 1) Mécanismes de l'Evolution

Fixation d'une mutation par dérive génétique ou sélection positive



# 1) Mécanismes de l'Evolution

## Fixation des mutations

Une mutation est dite fixée lorsqu'elle a été incorporée dans le génome de tous les individus d'une espèce.

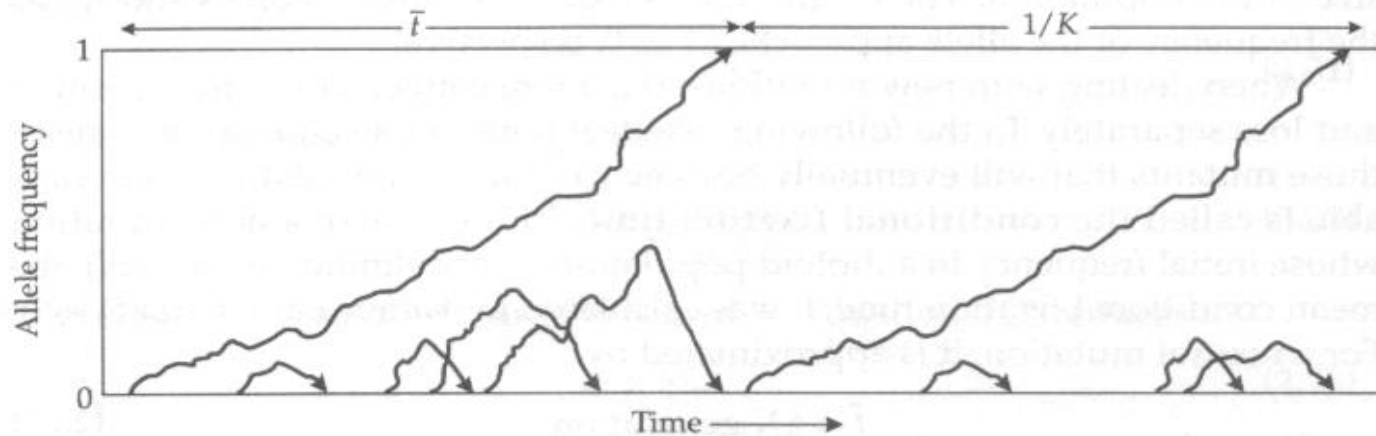
Le taux de fixation d'une mutation dépend

- de la taille de la population (N),
- de la fitness (s) de la mutation
- et du taux de mutation ( $\mu$ ).

$$r = 4Ns\mu$$

# 1) Mécanismes de l'Evolution

La fitness (s) d'une mutation influence la vitesse de fixation



temps de fixation

---

Mutation neutre:

$t_{neu}$

Mutation sélectionnée positivement:

$t_{pos} < t_{neu}$

Mutation sélectionnée négativement:

$t_{neg} > t_{neu}$ , OU  $t_{neg} = \infty$

---

# **Théorie sur l'évolution moléculaire.**

## **Les bases moléculaires de l'évolution.**

**1) Mécanismes de l'évolution**

**2) Structure des génomes**

**3) Structures et fonctions des gènes**

**4) Changements mutationnels dans l'ADN**

**5) Usage des codons dans les séquences codantes**

## 2) Structure du génome

Le génome contient des **régions non-codantes** et des **régions codantes** .

Régions non-codantes:

- ADN intergénique ou junk DNA
- régions régulatrices (promoteurs, enhanceurs, ... )
- éléments répétés non-codants (Alu, télomères,...)

Régions codantes:

- gènes codant pour des protéines avec ARN messagers (mRNA)
- gènes codant pour des ARNs structurels:  
ARN ribosomiques (rRNA),  
ARN de transfert (tRNA),  
Small nuclear RNAs (snRNA)

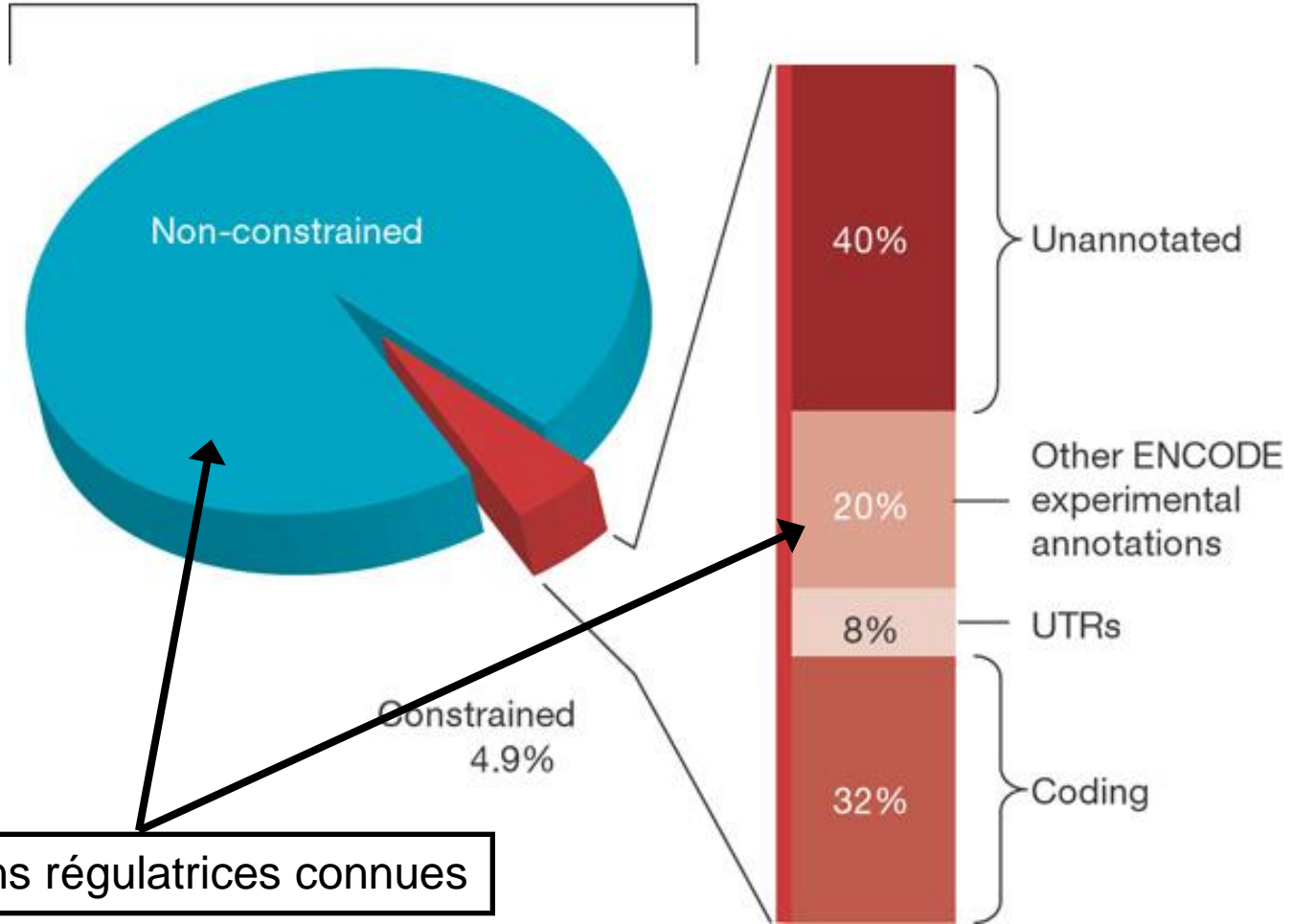
---

# Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project

The ENCODE Project Consortium\*

44 régions du génome à travers tous les chromosomes séquencées et comparées chez 23 mammifères.

All 44 ENCODE regions  
(29,998 kb)



Régions régulatrices connues

# **Théorie sur l'évolution moléculaire.**

## **Les bases moléculaires de l'évolution.**

**1) Mécanismes de l'évolution**

**2) Structure des génomes**

**3) Structures et fonctions des gènes**

**4) Changements mutationnels dans l'ADN**

**5) Usage des codons dans les séquences codantes**

## 2) Structure et fonction des gènes

### Structure des gènes codant pour des protéines

Région transcrite: succession d'exons et d'introns  
(exception: gènes mono-exoniques)

Région non-traduite 5' (= 5'UTR) dans le (ou les) premier(s) exon(s).

Codon start (AUG) en général dans le premier exon.

Codon stop en général dans le dernier exon.

Région non-traduite 3' (= 3'UTR) dans le (ou les) dernier(s) exon(s).

L'ARN pré-messager contient encore les introns.

Par l'**épissage** (splicing) les introns sont éliminés et l'on obtient l'ARN messager.

L'ARN messager contient toujours les 5' et 3' UTRs en plus de la région codante,  
= CDS (coding sequence), de l'AUG au codon STOP.

## 2) Structure et fonction des gènes

### **Le Code Génétique**

A partir du signal de départ de la traduction, chaque **codon** code pour un acide aminés de la protéine. C'est le **code génétique**.

**Le code génétique est universel.**

Il est le même pour les Prokaryotes, les Eukaryotes, les chloroplastes et presque identique chez les mitochondries.

**Le code génétique est dégénéré.**

Il y a 64 codons différents possibles ( $4^3$ ) mais seulement 20 acides aminés. Certains codons codes pour un même acide aminé, ils sont **synonymes**.

Le signal de départ de la traduction est en général le codon AUG qui code pour une méthionine.

Il y a généralement 3 codons de terminaison (STOP codons): UAA, UAG et UGA.

## 2) Structure et fonction des gènes

Tableau du code génétique universel

		2nd base in codon				
		U	C	A	G	
1st base in codon	U	Phe Phe Leu Leu	Ser Ser Ser Ser	Tyr Tyr STOP STOP	Cys Cys STOP Trp	U C A G
	C	Leu Leu Leu Leu	Pro Pro Pro Pro	His His Gln Gln	Arg Arg Arg Arg	U C A G
	A	Ile Ile Ile Met	Thr Thr Thr Thr	Asn Asn Lys Lys	Ser Ser Arg Arg	U C A G
	G	Val Val Val Val	Ala Ala Ala Ala	Asp Asp Glu Glu	Gly Gly Gly Gly	U C A G

## 2) Structure et fonction des gènes

Différents codes génétiques dans les mitochondries

Vertebrate mtDNA.

<b>Codon</b>	<b>New translation</b>
AGA	Stop
AGG	Stop
ATA	Met
TGA	Trp

Invertebrate mtDNA.

<b>Codon</b>	<b>New translation</b>
AGA	Ser
AGG	Ser
ATA	Met
TGA	Trp

Yeast mtDNA.

<b>Codon</b>	<b>New translation</b>
ATA	Met
CTA	Thr
CTC	Thr
CTG	Thr
CTT	Thr
TGA	Trp

## 2) Structure et fonction des gènes

Différents codes génétiques dans les mitochondries

Mold, Protozoan and Coelenterate mtDNA.

<b>Codon</b>	<b>New translation</b>
TGA	Trp

Ascidian mtDNA.

<b>Codon</b>	<b>New translation</b>
AGA	Gly
AGG	Gly
AGG	Met
TGA	Trp

Echinoderm mtDNA.

<b>Codon</b>	<b>New translation</b>
AAA	Asn
AGA	Ser
AGG	Ser
TGA	Trp

Flatworm mtDNA.

<b>Codon</b>	<b>New translation</b>
AAA	Asn
AGA	Ser
AGG	Ser
TAA	Tyr
TGA	Trp

## 2) Structure et fonction des gènes

Différents codes génétiques aussi dans le noyau

### Ciliate Nuclear Code.

<b>Codon</b>	<b>New translation</b>
TAA	Gln
TAG	Gln

### Alternative Yeast Nuclear.

<b>Codon</b>	<b>New translation</b>
CTG	Ser

# **Théorie sur l'évolution moléculaire.**

## **Les bases moléculaires de l'évolution.**

**1) Mécanismes de l'évolution**

**2) Structure des génomes**


**3) Structures et fonctions des gènes**

**4) Changements mutationnels dans l'ADN**

**5) Usage des codons dans les séquences codantes**

### 3) Changements mutationnels dans l'ADN

Tous les caractères morphologiques et physiologiques sont contrôlés par l'information génétique portée par l'ADN.

 Tout changement héritable dans ces caractères est causé par un changement dans l'ADN.

Il y a 4 changements de base à l'échelle nucléotidique:

1. Substitution d'un nucléotide par un autre
2. Insertion
3. Délétion
4. Inversion

Insertions, délétions et inversions peuvent impliquer plus d'un nucléotide.

### 3) Changements mutationnels dans l'ADN

#### 1. Substitution.

Thr Tyr Leu Leu  
ACC TAT TTG CTG  
↓  
ACC TCT TTG CTG  
Thr Tyr Leu Leu

#### 3. Insertion.

Thr Tyr Leu Leu  
ACC TAT TTG CTG  
↓  
ACC TAC TTT GCT G—  
Thr Tyr Phe Ala

#### 2. Deletion.

Thr Tyr Leu Leu  
ACC TAT TTG CTG  
↓  
ACC TAT TGC TG—  
Thr Tyr Cys

#### 4. Inversion.

Thr Tyr Leu Leu  
ACC TAT TTG CTG  
↓  
ACC TTT ATG CTG  
Thr Phe Met Leu

### 3) Changements mutationnels dans l'ADN

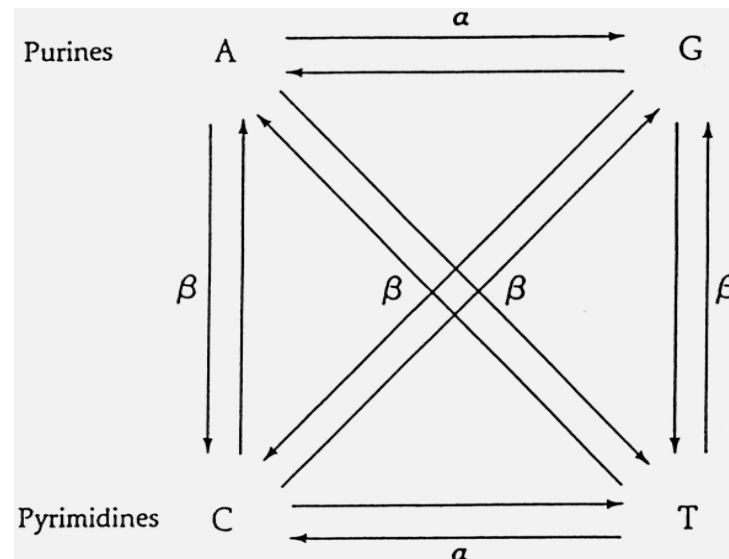
Deux types de substitutions: **transitions** et **transversions**

**Transition**: substitution d'un nucléotide par un autre nucléotide de même famille.

substitution d'une purine (A, G) par une autre purine (A, G) ou  
substitution d'une pyrimidine (C, T) par une autre pyrimidine (C, T).

**Transversions**: toutes les autres substitutions

Schéma des substitutions



### 3) Changements mutationnels dans l'ADN

Conséquences au niveau de la région codante:

Du fait d'un code génétique dégénéré, une substitution dans une région codante change ou ne change pas l'acide aminés codé.

Si la mutation ne change pas l'acide aminé, elle est dite **synonyme**.

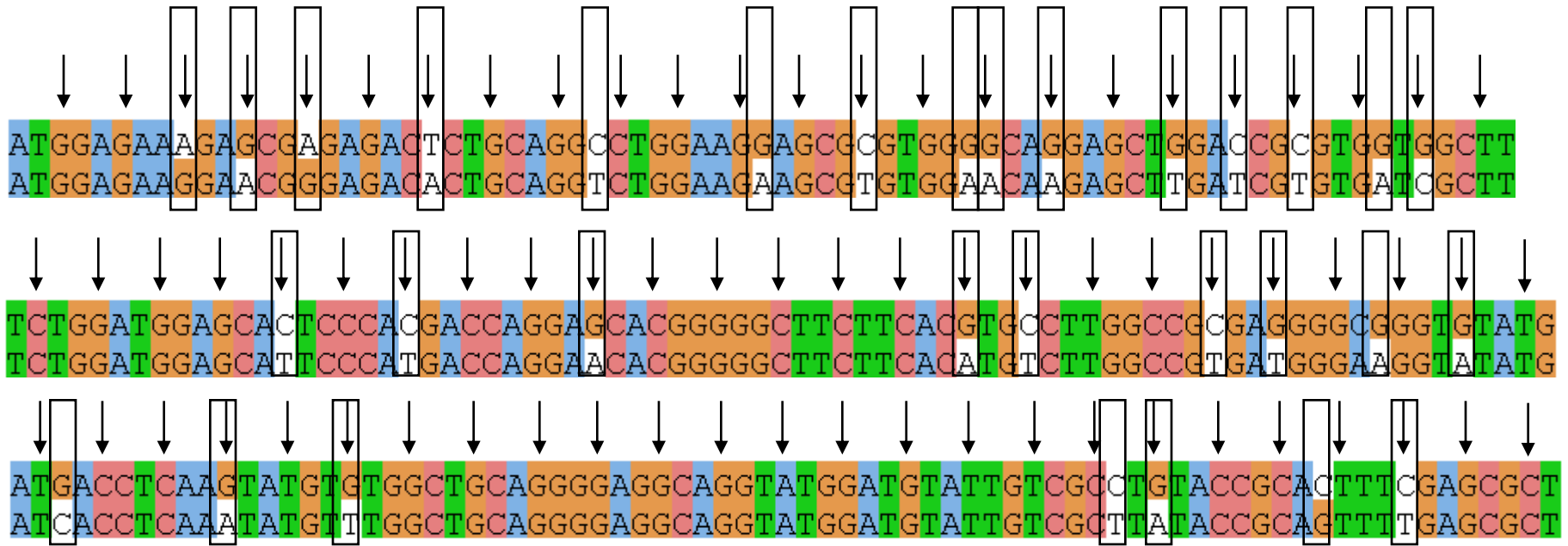
Si la mutation change l'acide aminé, elle est dite **non-synonyme**.

Si la mutation crée un codon STOP, elle est dite **nonsense**.

Du fait de la dégénérescence du code génétique, la plupart des mutations **synonymes** ont lieu en 3ème position du codon, quelques unes ont lieu en 1ère position.

Toute mutation en deuxième position est soit **non-synonyme** soit **nonsense**.

## Prépondérance des substitutions en 3ème position



Les 220 premiers nucléotides du gène de la renin binding protein de l'homme et de la souris

↓ Les 3ème positions sont indiquées

sur les 31 changements:

4 - 1ère position

4 - 2ème position

23 - 3ème position

### 3) Changements mutationnels dans l'ADN

Conséquences au niveau de la région codante:

Proportions théoriques des différents types de mutations.

Si l'on assume que tous les codons sont utilisés avec la même fréquence et que toutes les paires de nucléotides ont la même probabilité de substitution, alors les proportions des différents types de mutations sont:

synonymes:	25%
non-synonymes:	71%
non-sense:	4%

Dans la réalité, ces proportions sont toutes autres!  
(car les hypothèses de base ne sont pas réalistes).

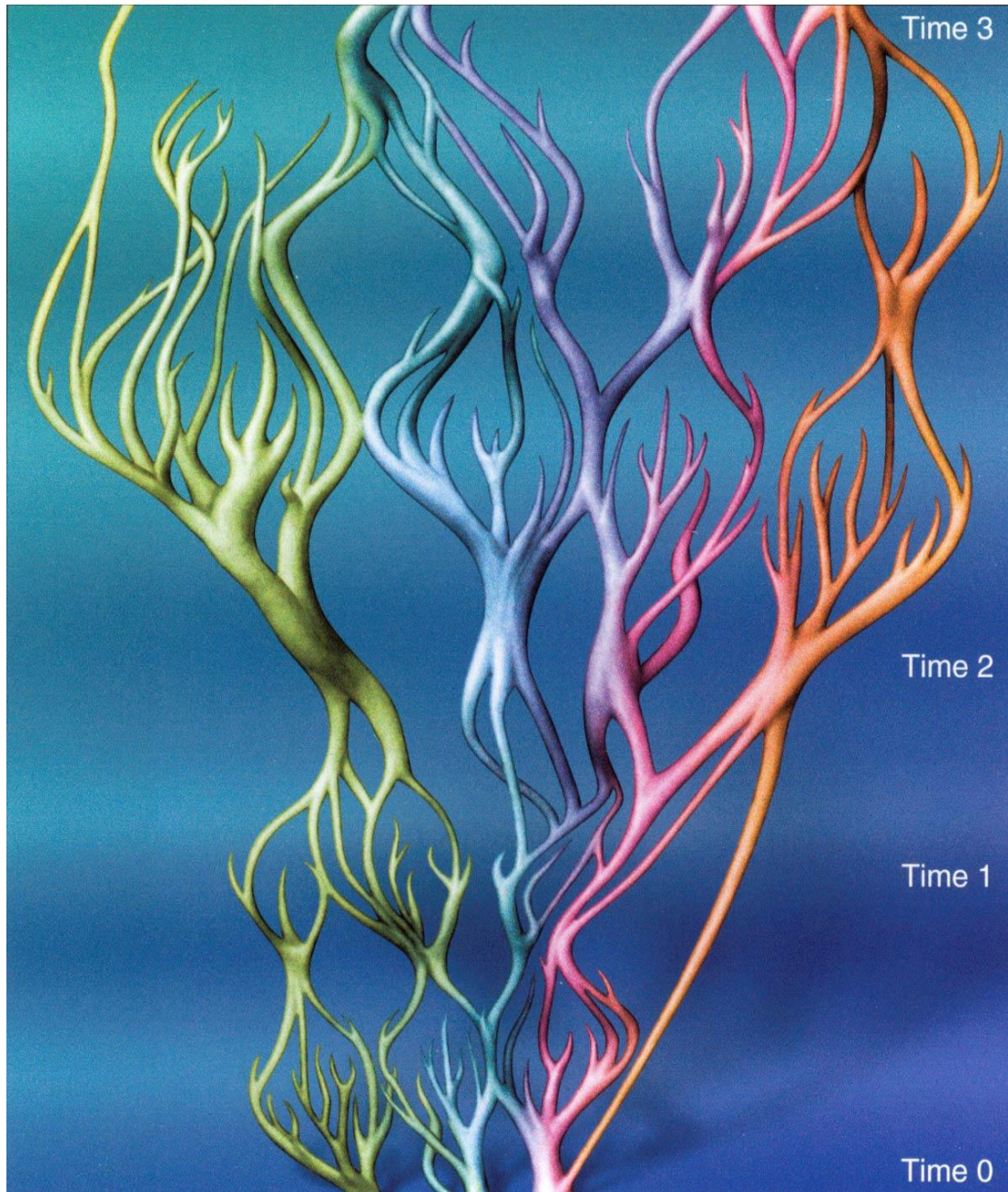
### 3) Changements mutationnels dans l'ADN

#### Les causes des insertions - délétions

Les insertions - délétions impliquant un petit nombre de nucléotides sont attribuables aux **erreurs de réplication** de l'ADN au cours de la méiose (comme pour la plupart des substitutions).

Les longues insertions - délétions semblent provenir de **crossing over inégaux** ou d'évènements de **transposition** via les transposons ou les éléments transposables (retrotransposons, LINEs, SINEs, Alu, ...).

Certaines insertions sont attribuables aux **transferts horizontaux** de gènes, évènements rares où une espèce reçoit un morceau d'ADN provenant d'une autre espèce (via des virus, plasmides, transposons...?).



L'évolution réticulée  
chez les coraux (Veron,  
2000)

# **Théorie sur l'évolution moléculaire.**

## **Les bases moléculaires de l'évolution.**

**1) Mécanismes de l'évolution**

**2) Structure des génomes**

**3) Structures et fonctions des gènes**

**4) Changements mutationnels dans l'ADN**

**5) Usage des codons dans les séquences codantes**

## 4) Usage des codons dans les séquences codantes

S'il n'y avait pas de pression de sélection ni de biais dans les fréquences de substitutions des nucléotides,

**alors** les codons qui codent pour un même acide aminé, ou **codons synonymes**, seraient utilisés avec la même fréquence à travers tout le génome.

Dans la réalité, ces conditions ne sont pas réunies et il existe un biais plus ou moins grand dans l'utilisation des codons, selon les organismes.

Exemple: gènes des RNA polymérases de *E.coli*. (entre parenthèses: *RSCU*)  
Relative Synonymous Codon Usage

Val	GUU: 55 (1.53)	Pro	CCU: 9 (0.48)
	GUC: 21 (0.58)		CCC: 0 (0.00)
	GUA: 34 (0.94)		CCA: 11 (0.59)
	GUG: 34 (0.94)		CCG: 55 (2.93)

## 4) Usage des codons dans les séquences codantes

Plusieurs causes expliquent le biais d'usage des codons: I

Chez les gènes à fort taux de transcription, le biais d'usage des codons est corrélé aux proportions relatives des tRNA "synonymes" disponibles dans les lieux de synthèse des protéines.

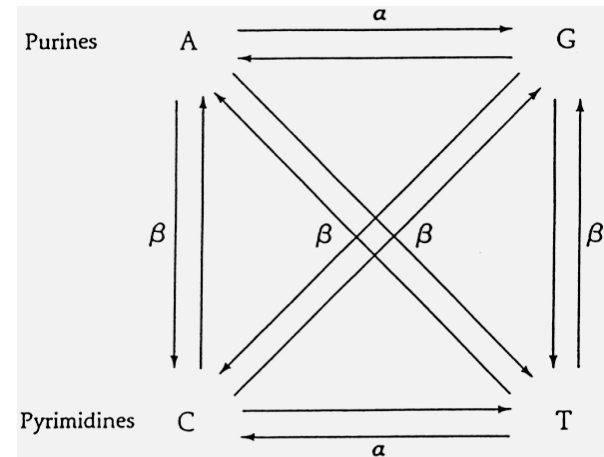
Les proportions relatives des tRNA "synonymes" semblent dépendre directement du nombre de gènes qui code pour chaque tRNA, et ceci est différent pour chaque espèce.

Ces observations ne s'appliquent que pour les gènes à fort taux de transcription. La **sélection purificatrice** élimine les mutations qui modifient un codon en un codon synonyme dont le tRNA serait moins abondant (moins de copie du gène), car le gène serait traduit plus lentement, et donc mauvais pour l'individu.

## 4) Usage des codons dans les séquences codantes

Plusieurs causes expliquent le biais d'usage des codons: II

Biais dans les taux de substitutions entre nucléotides.



Exemples:

Chez certaines bactéries (*Mycoplasma* sp.), le taux de substitution des G et C est si fortement biaisé en faveur des A ou T que les nucléotides en 3ème position sont presque exclusivement des A ou T (Muto et Osawa, 1987)

**Isochores** chez les mammifères: alternance de régions chromosomiques riches en G+C (moyenne 60%) et pauvres en G+C (moyenne 30%).

Le biais du taux de substitution peut donc être variable dans un même génome.

## 4) Usage des codons dans les séquences codantes

Mesure statistique du biais d'usage des codons.

L'indice **RSCU** (Relative Synonymous Codon Usage) mesure pour chaque codon son degré d'utilisation par rapport aux autres codons synonymes dans le même gène (ou set de gènes).

$$RSCU_i = \frac{Obs_i}{Exp_i}$$

où  $Obs_i$  est le nombre observé de codons  $i$  dans le gène et  $Exp_i$  est le nombre attendu de codons  $i$  dans le gène.

$$Exp_i = \frac{\sum aa_i}{\sum syn_i}$$

occurrence de l'**acide aminé** codé par le codon analysé  
-----  
nombre de **codons synonymes** pour cet acide aminé

## 4) Usage des codons dans les séquences codantes

Mesure statistique du biais d'usage des codons.

Un gène peut être caractérisé par l'indice **CAI** (Codon Adaptation Index)

[Sharp and Li \(1987\)](#)

Cet indice, qui varie de 0 à 1, utilise un set de gènes les plus fortement exprimés comme référence d'utilisation des codons synonymes.

Plus un gène possède un CAI élevé (entre 0.7 et 1), plus il se comporte comme un gène fortement exprimé.

Cet indice sert à définir le taux d'expression d'une gène ou à comparer les biais d'utilisation des codons entre organismes.

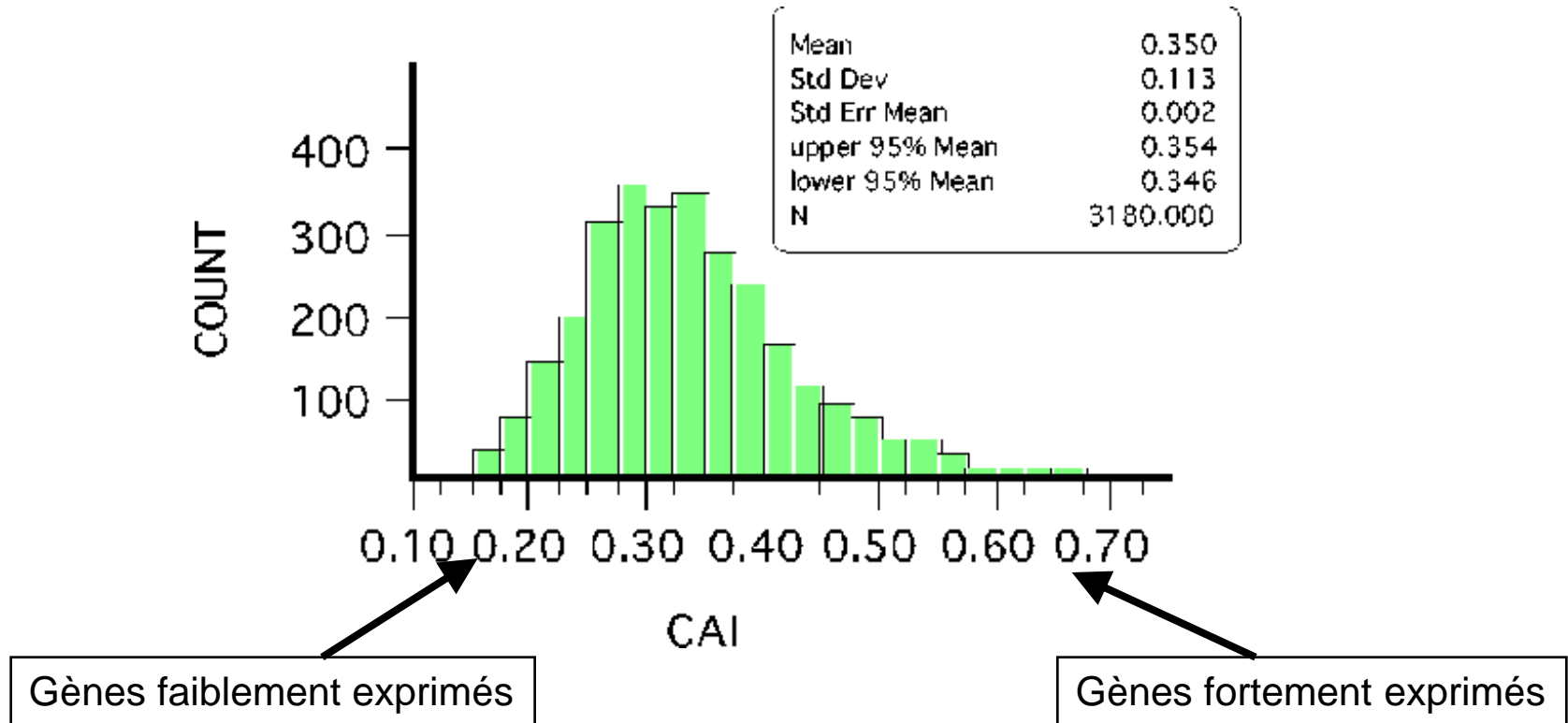
$$\ln(\text{CAI}) = \sum_{i=1}^{61} f_i \ln W_i$$

où  $f_i$  est la fréquence relative du codon  $i$  dans la séquence analysée

et  $W_i$  est le rapport de la fréquence du codon  $i$  sur la fréquence du codon synonyme le plus utilisés dans l'ensemble des gènes les plus fortement exprimés.

## 4) Usage des codons dans les séquences codantes

CAI (Codon Adaptation Index) [Sharp and Li \(1987\)](#)



**Figure:** Distribution of CAI for 3180 *E. coli* genes

# **Théorie sur l'évolution moléculaire.**

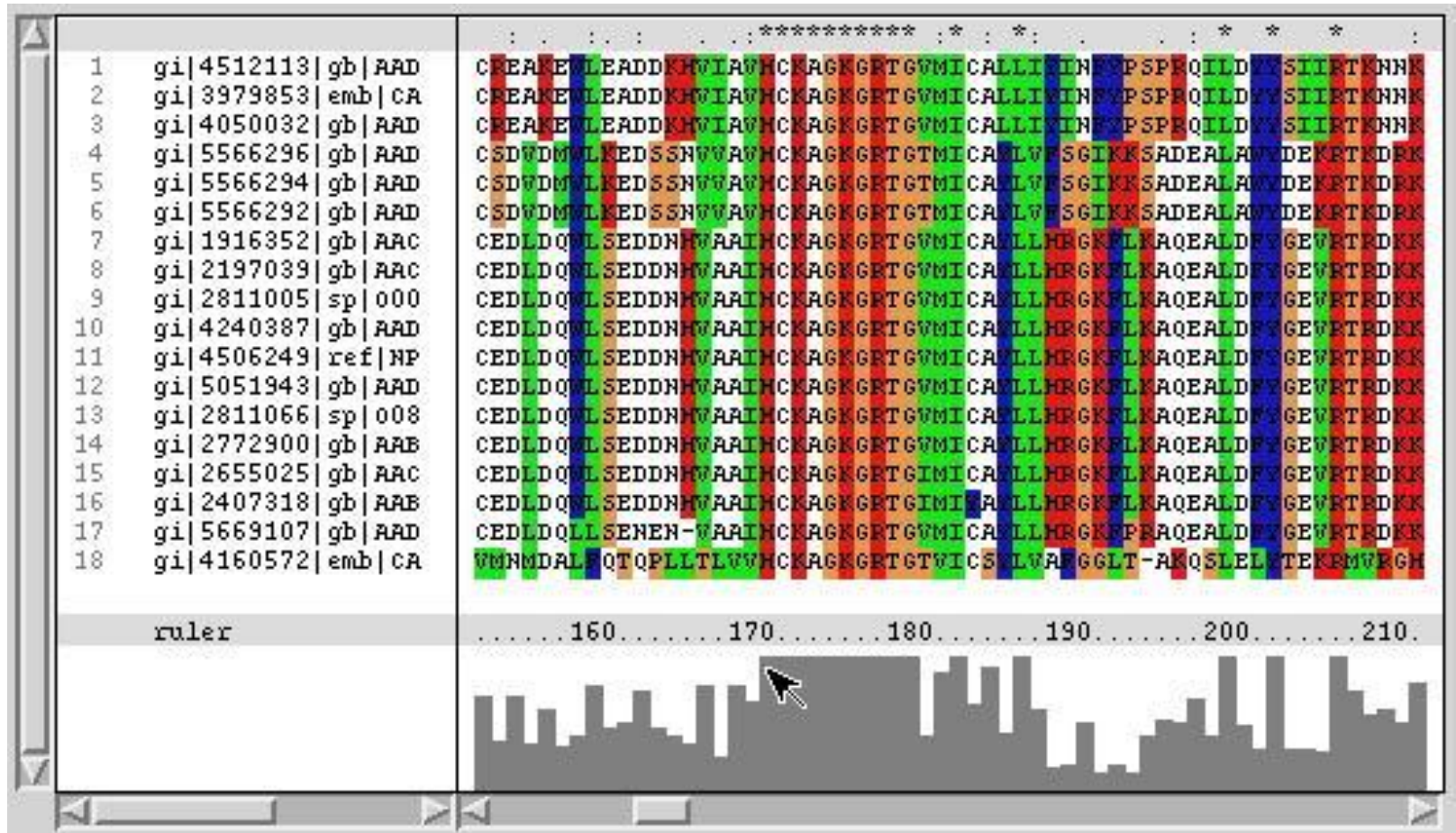
- **Les bases moléculaires de l'évolution.**

- **Evolution des séquences protéiques**

- **Evolution des séquences nucléotidiques**

# Evolution des séquences protéiques

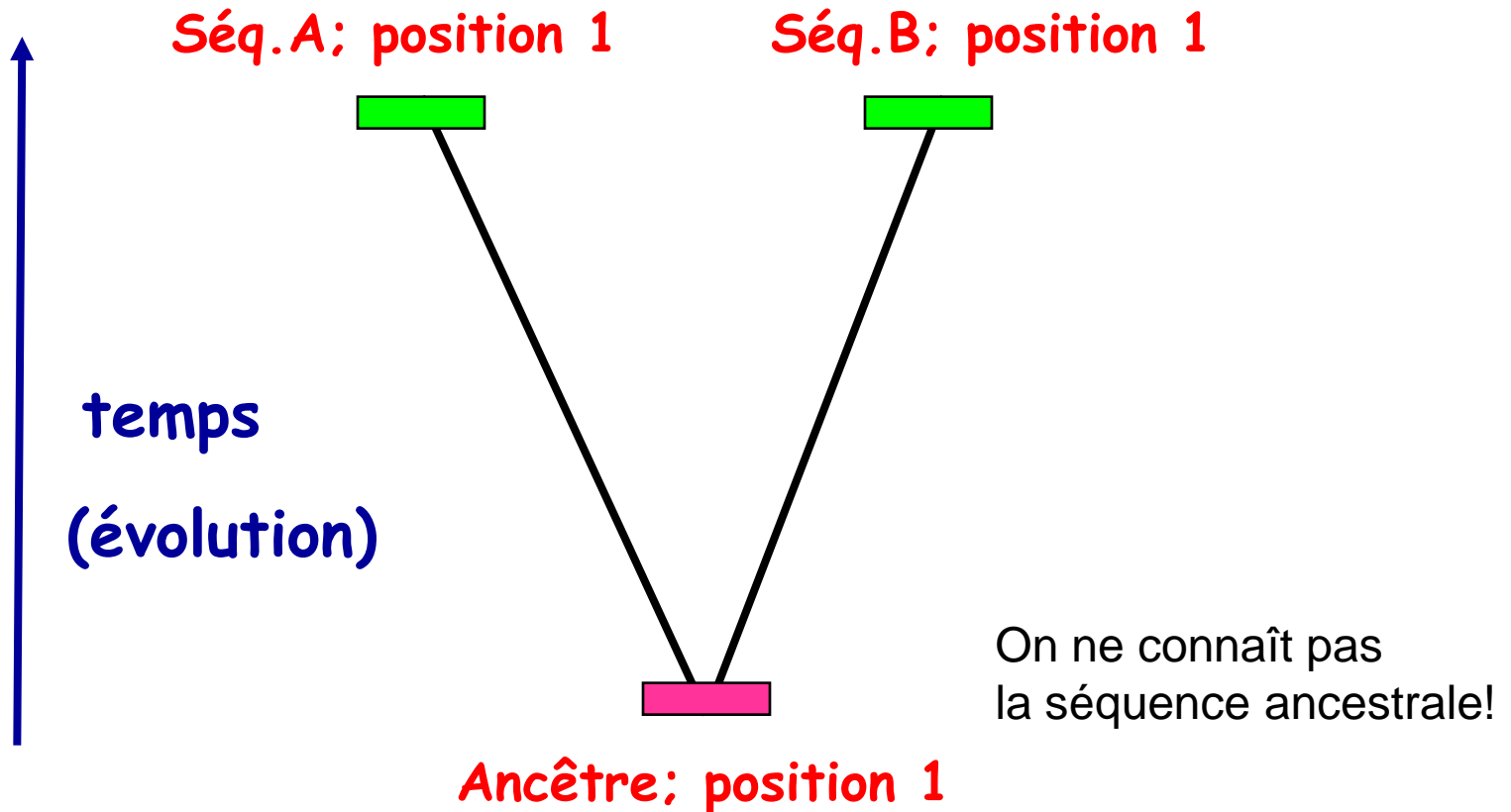
Comment mesurer les différences entre séquences ?



ClustalX window

# Evolution des séquences protéiques

Comment mesurer les différences entre séquences ?



# Evolution des séquences protéiques

Comment mesurer les différences entre séquences ?

Une mesure simple est de compter le nombre de aa différents entre paires de séquences. C'est la **distance observée**.

Cette distance peut être exprimée en proportion. C'est la **distance  $p$** .

**distance  $p$**  = distance observée / nombre de sites

$$p = n_d / n$$

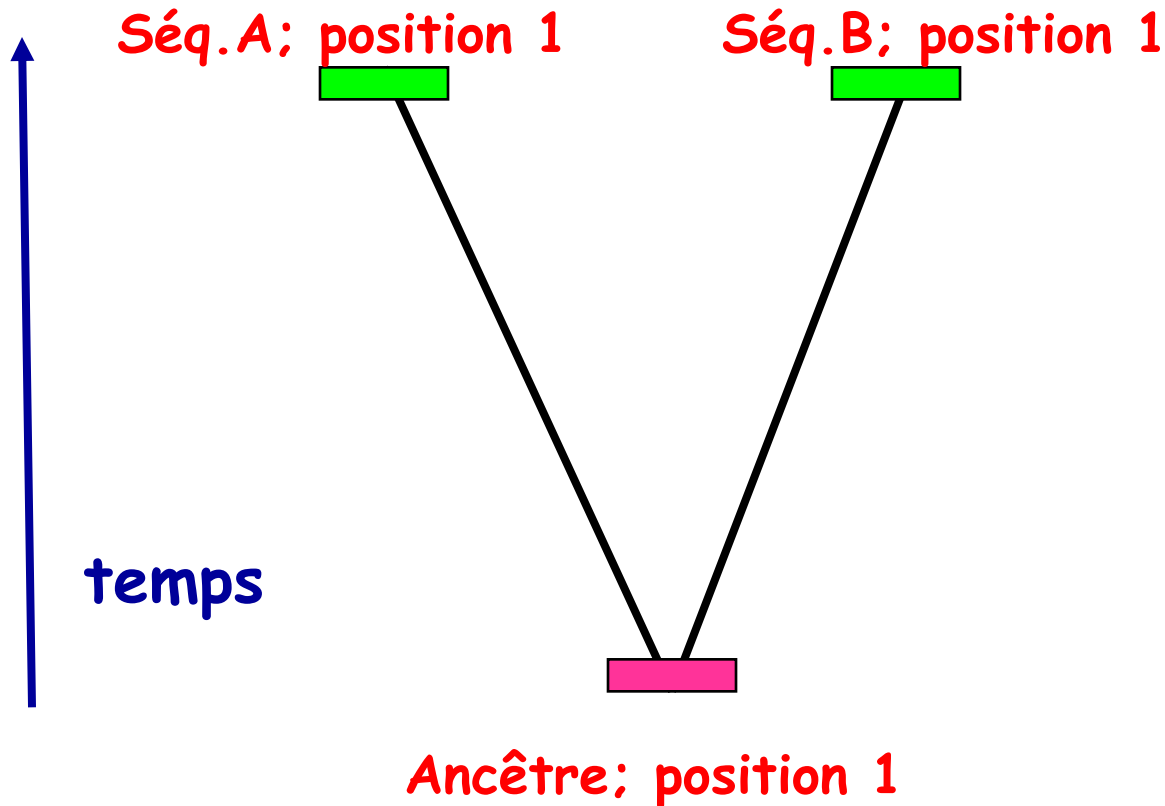
$n_d$  – number of amino acid differences between two sequences;  $n$  – number of aligned amino acids.

La distance  $p$  est une mesure exacte si tous les changements qui ont eu lieu au cours de l'histoire indépendante des ces séquences sont observables aujourd'hui.

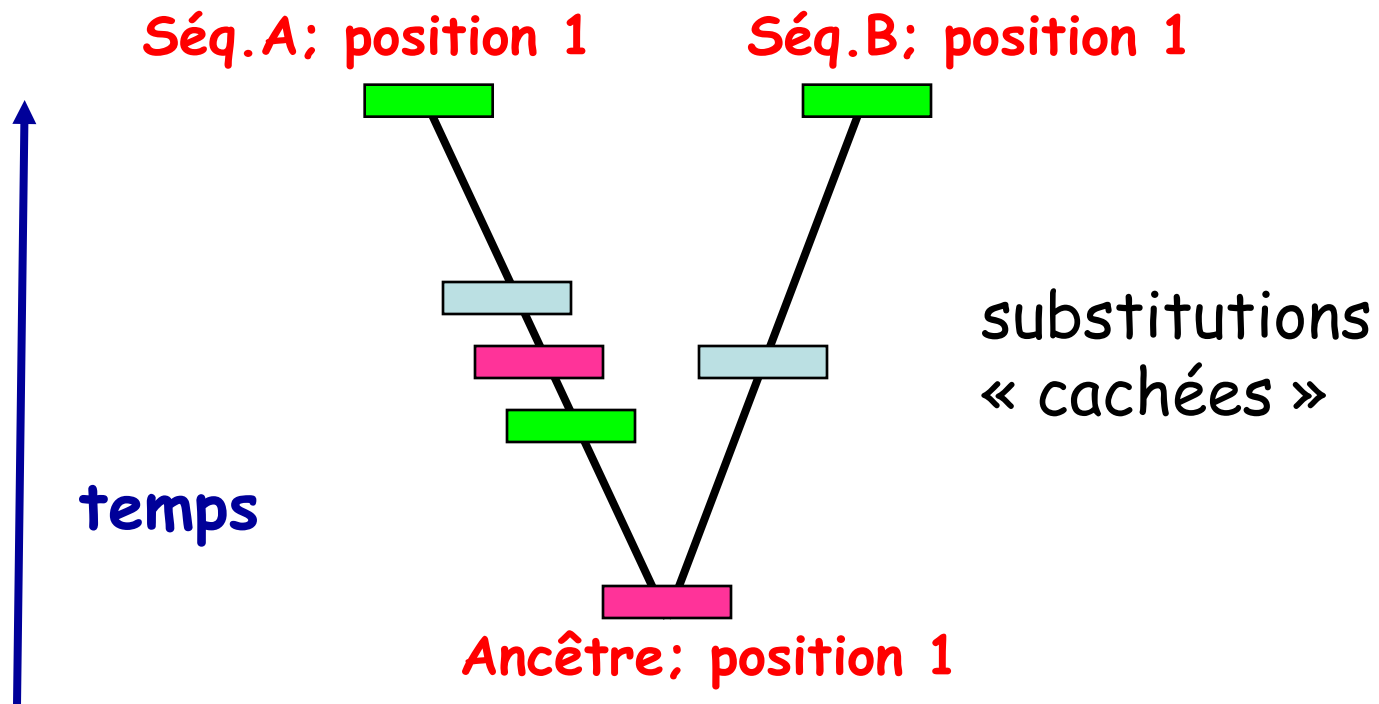
Ce n'est pas toujours le cas !

Est-ce que toutes les mutations sont visibles?

Distance observée vs distance évolutive



# Distance observée vs distance évolutive



Plus le temps est grand, plus le nombre de substitutions cachées est grand.

La distance  $p$  n'est donc pas proportionnelle au temps!

# Evolution des séquences protéiques

Parce que toutes les mutations ne sont pas forcément visibles, et que ceci augmente avec le temps, la **distance p** doit être corrigée pour être proportionnelle au temps.

Comment corriger la distance p?

De deux manières: - théorique: Poisson corrected distance (PC)

- empirique: Matrices de substitution (PAM, JTT,...)

# Evolution des séquences protéiques

- théorique: Poisson corrected distance (PC)

La distribution de Poisson est une distribution discrète de probabilité. Elle exprime la probabilité de réalisation d'un nombre d'évènements au cours d'un temps donné, si ces évènements se produisent avec un taux moyen connu.

La probabilité qu'il y aie exactement un nombre  $k$  d'évènements ( $K = 0, 1, 2, 3, \dots$ ) est donnée par:

$$f(k; \lambda) = \frac{e^{-\lambda} \lambda^k}{k!},$$

où

$e = 2.71828\dots$ ,

$k!$  est  $k$  factoriel.

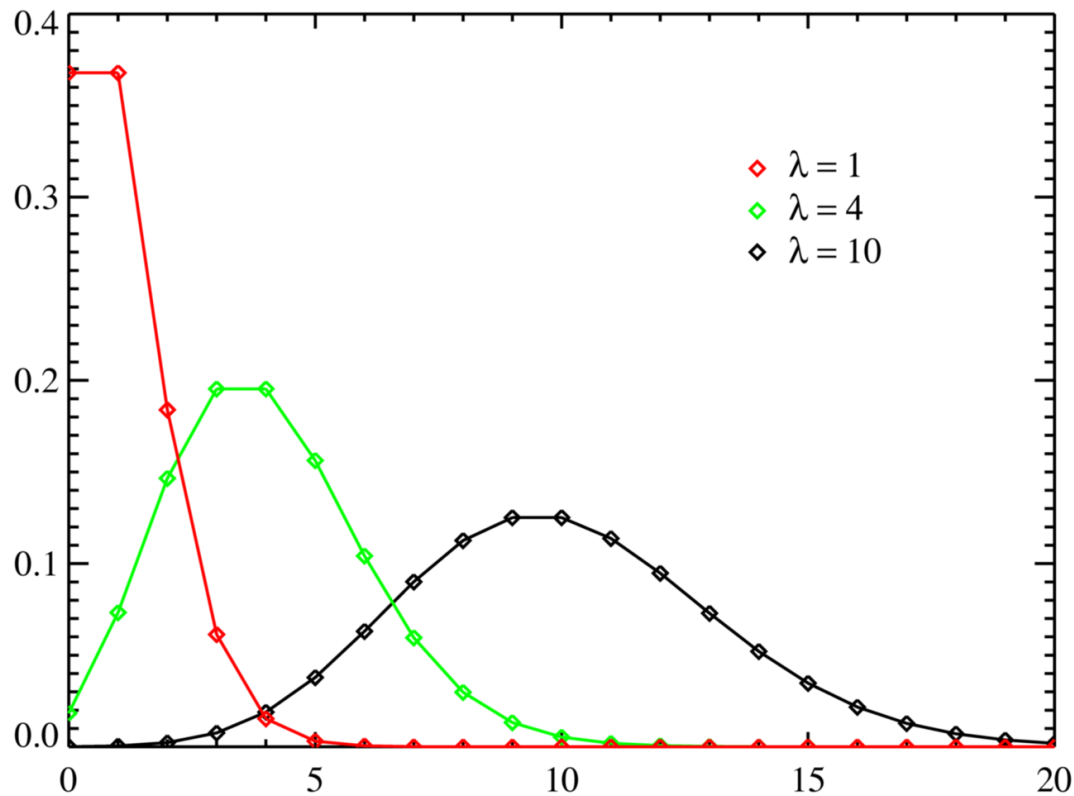
$\lambda$  est le nombre attendu d'occurrences au cours du temps donné.

$\lambda$  est la seule variable qui détermine la distribution.

# Evolution des séquences protéiques

Allure d'une distribution de Poisson en fonction de la valeur de  $\lambda$ .

( $\lambda$  = nombre attendu d'occurrences au cours du temps donné)



# Evolution des séquences protéiques

- théorique: Poisson corrected distance (PC)

$\lambda$  est le nombre attendu d'occurrences au cours du temps donné.  
 $\lambda$  est la seule variable qui détermine la distribution.

Dans notre cas quel est la valeur de  $\lambda$  ?

Supposons que  $r$  soit le taux de substitution d'un acide aminé par un autre au cours d'une année et qu'il soit le même pour tous les sites.

Le taux moyen de substitutions par site pendant une période de  $t$  années est donc:  $rt$

$\lambda$  est donc égale à  $rt$

$$f(k; \lambda) = \frac{e^{-\lambda} \lambda^k}{k!},$$

# Evolution des séquences protéiques

- théorique: Poisson corrected distance (PC)

$$f(k; \lambda) = \frac{e^{-\lambda} \lambda^k}{k!},$$

La probabilité qu'il n'y aie eu aucun changement ( $k = 0$ ) au cours du temps  $t$  à un site est:

$$f(0; t) = P(0; t) = e^{-rt}$$

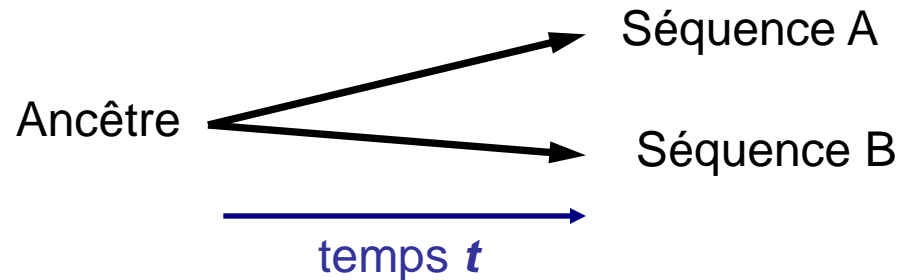
Cette équation n'est pas applicable car on ne connaît pas la séquence ancestrale à partir de laquelle on doit estimer le nombre de changements, c'est à dire l'état de la séquence au temps  $t = 0$ .

La proportion d'acides aminés inchangés au cours du temps  $t$  est obtenue empiriquement.

# Evolution des séquences protéiques

- théorique: Poisson corrected distance (PC)

Le nombre d'acides aminés inchangés est donc calculé à partir de la comparaison de deux séquences présentes aujourd'hui qui ont divergé il y a un temps  $t$ .



La probabilité ( $q$ ) qu'aucun des deux sites homologues dans les séquences A et B n'aie été substitué en un temps  $t$  est:

$$q = (e^{-rt}) \times (e^{-rt}) = e^{-2rt}$$

Cette probabilité  $q$  peut être estimée par  $q = 1 - p$

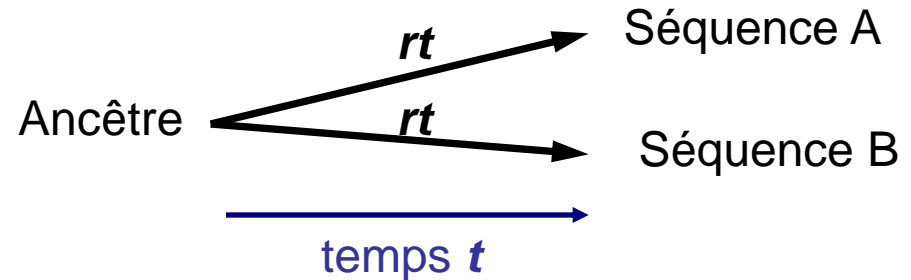
$p$  = proportion observée de sites différents entre deux séquences

# Evolution des séquences protéiques

- théorique: Poisson corrected distance (PC)

A partir des équations  $q = 1 - p$   
et  $q = e^{-2rt}$

on obtient  $1 - p = e^{-2rt}$  et  $2rt = -\ln(1-p)$



( $r = \text{taux de substitutions par sites}$ )

La distance "vrai" qui sépare les deux séquences A et B est:  $d = 2rt$

d'ou

$$d = -\ln(1-p)$$

C'est la *Poisson corrected* distance !

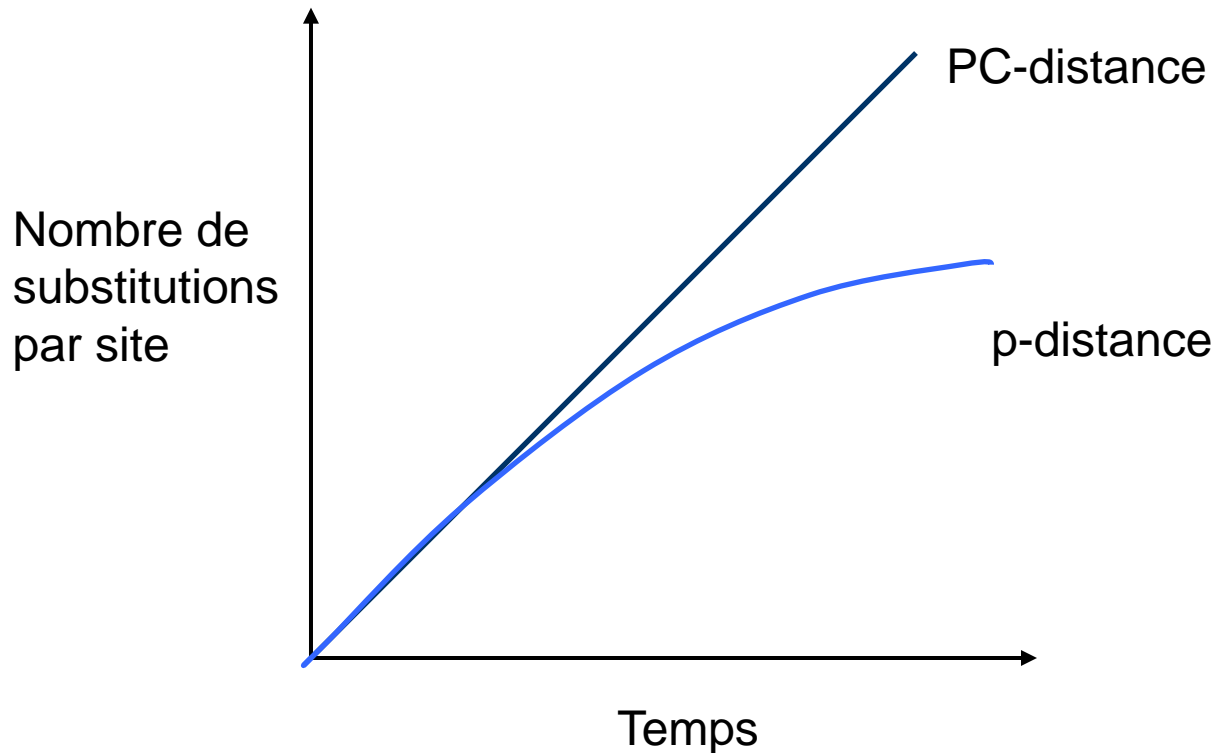
La variance est:

$$V(d) = p / [(1-p)n]$$

# Evolution des séquences protéiques

- théorique: Poisson corrected distance (PC)

Relation entre  $p$ -distance et *Poisson corrected*-distance (PC)



# Evolution des séquences protéiques

Dans le raisonnement de la correction de type Poisson, on assume que la probabilité de substitution est la même pour tous les sites.

Cette hypothèse ne tient généralement pas.

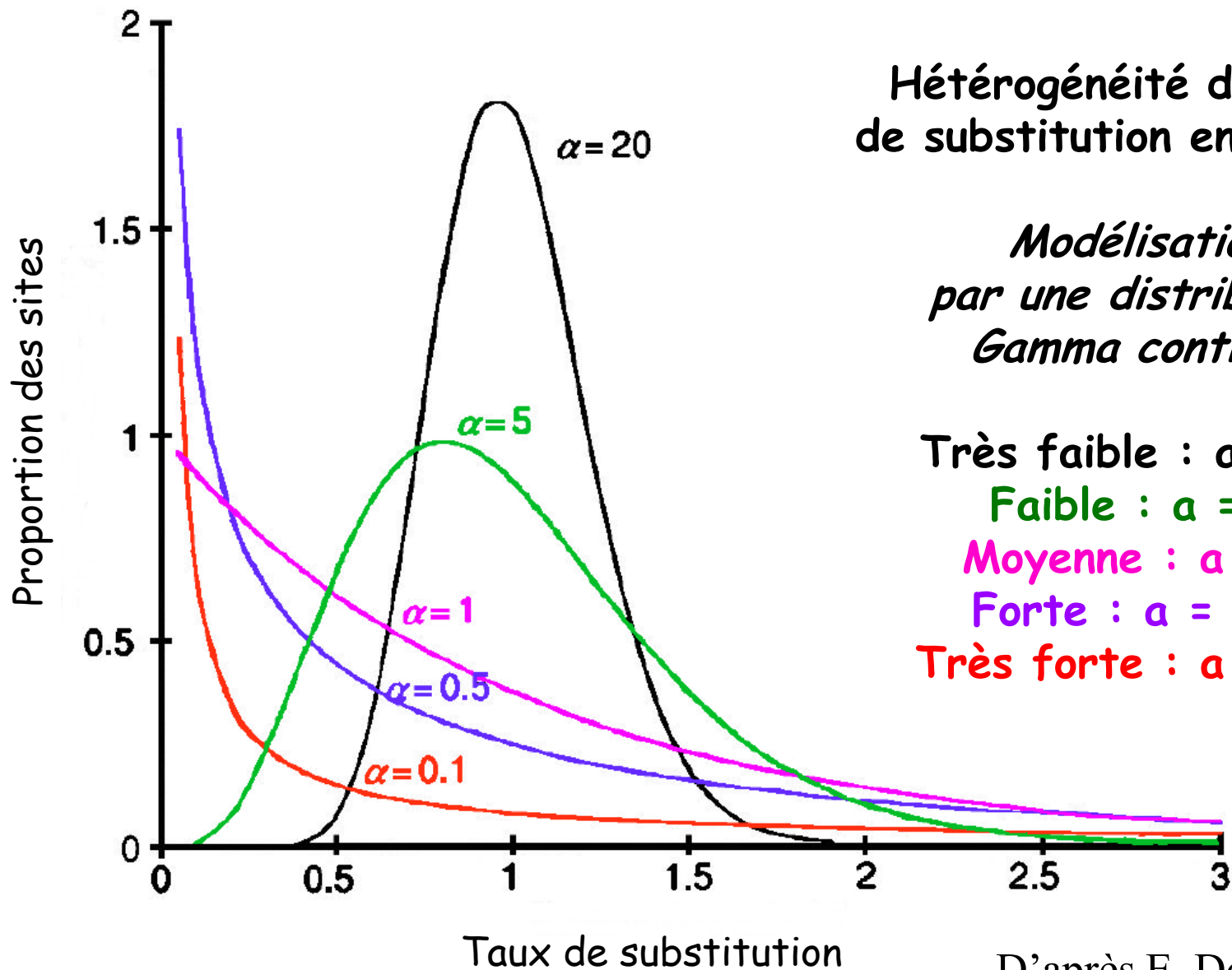
Uzzell et Corbin (1971) ont montré que la distribution du **nombre de substitution par site ( $k$ )** a une variance plus grande que celle donnée par la distribution de Poisson et que la distribution de  $k$  suit plutôt une distribution binomiale négative.

Pour que la distribution de  $k$  soit de type binomiale négative, il faut que le taux de substitution ( $r$ ) varie entre sites selon la **loi de distribution gamma**.

Un seul paramètre détermine l'allure de la distribution gamma, le paramètre  $\alpha$ .

La plupart des modèles d'évolution de séquences peuvent intégrer cette distribution.

# Distribution gamma



Hétérogénéité de taux  
de substitution entre sites

*Modélisation  
par une distribution  
Gamma continue*

Très faible :  $\alpha = 20$

Faible :  $\alpha = 5$

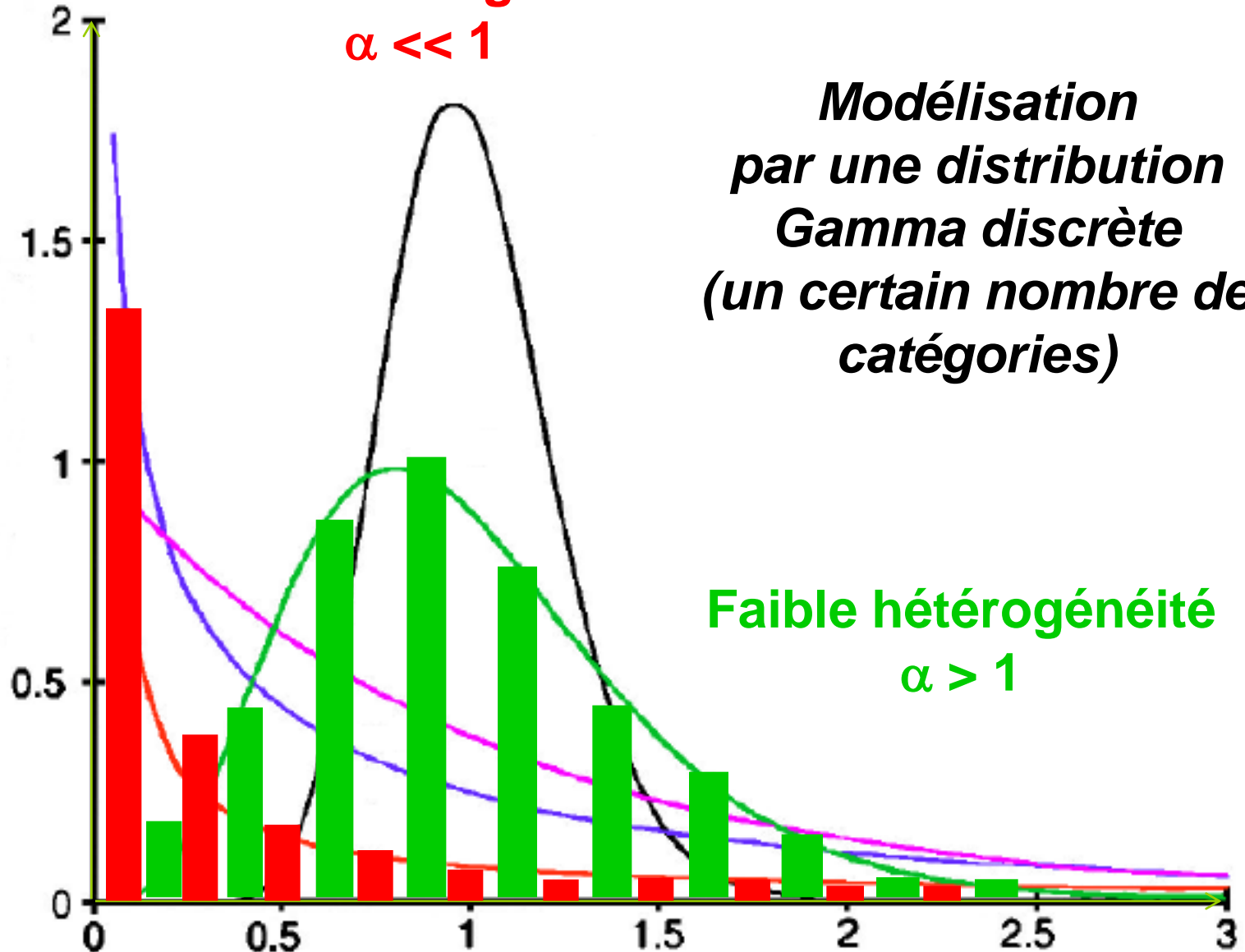
Moyenne :  $\alpha = 1$

Forte :  $\alpha = 0,5$

Très forte :  $\alpha = 0,1$

D'après E. Douzery

**Forte hétérogénéité**  
 $\alpha \ll 1$



D'après E. Douzery

# Evolution des séquences protéiques

Exemples de différence entre ces différentes distances

Estimation du taux de substitution par sites  
pour la chaîne alpha de l'hémoglobine

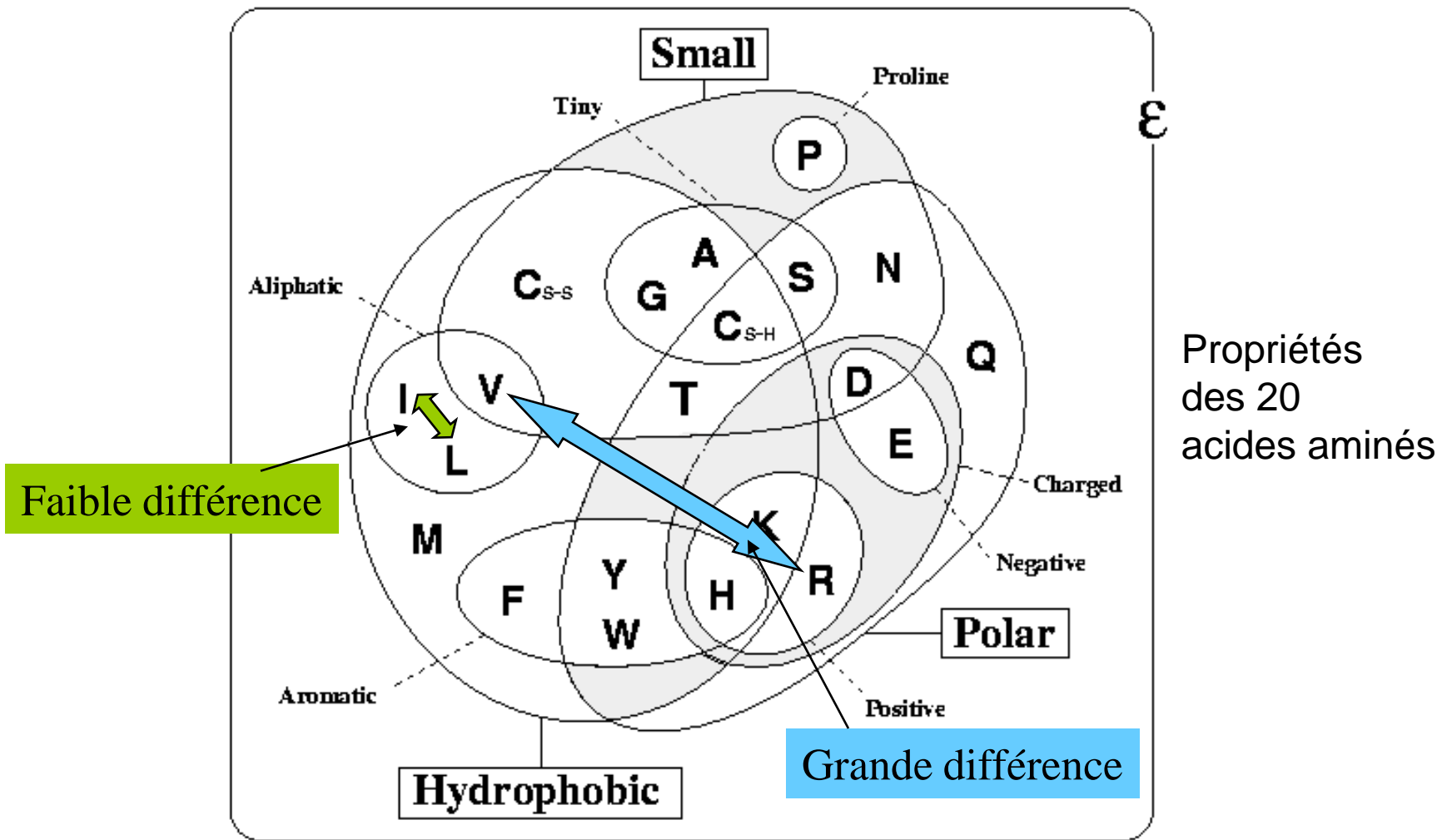
	P-distance	PC-distance	PC + Gamma-distance
Human/cow	0.121	0.129	0.134
Human/kangaroo	0.186	0.205	0.216
Human/carp	0.486	0.665	0.789

Le modèle PC + Gamma est cependant encore trop simpliste.

Les corrections empiriques reflètent mieux la complexité de l'évolution des protéines.

# Evolution des séquences protéiques

Correction empirique des distances: Matrices de substitutions  
(PAM, JTT....)





# **Théorie sur l'évolution moléculaire.**

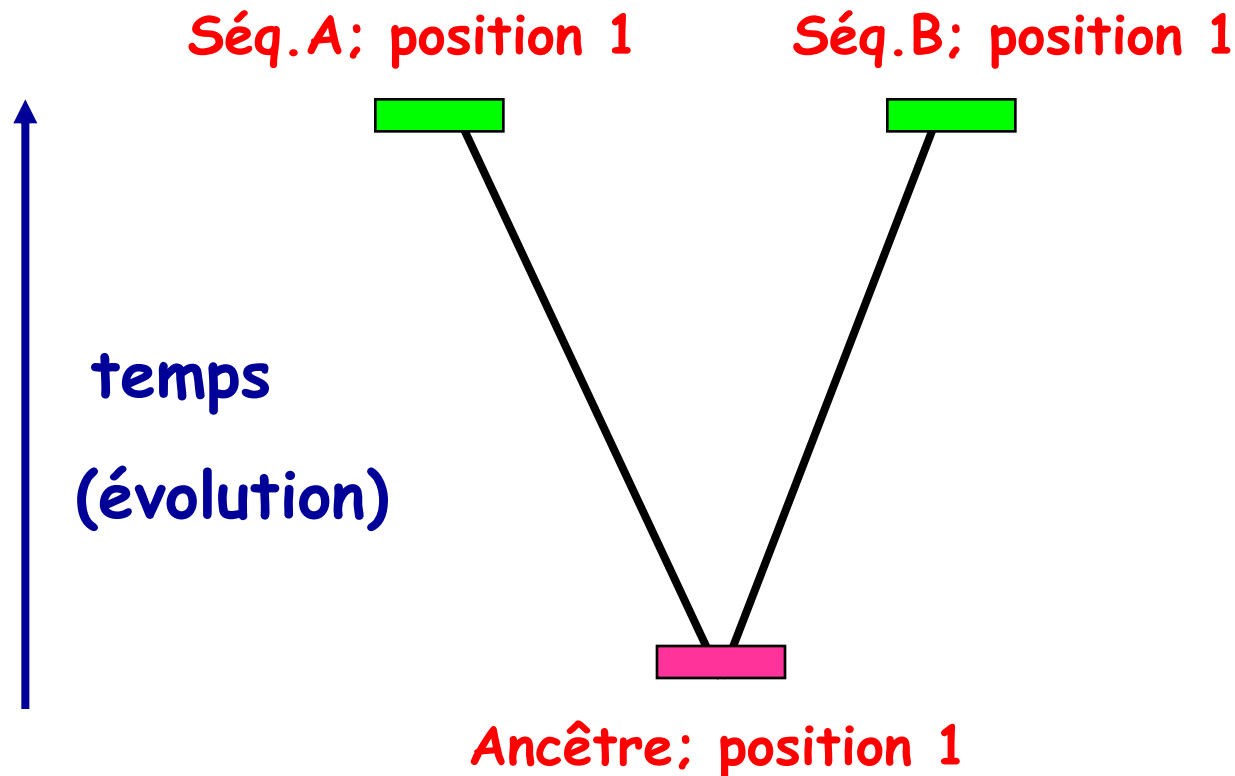
- Les bases moléculaires de l'évolution.**
- Evolution des séquences protéiques**

**- Evolution des séquences nucléotidiques**

# Evolution des séquences nucléotidiques

Distance observée ou distance- $p$ :

nombre de nucléotides différents / nombre total de nucléotides

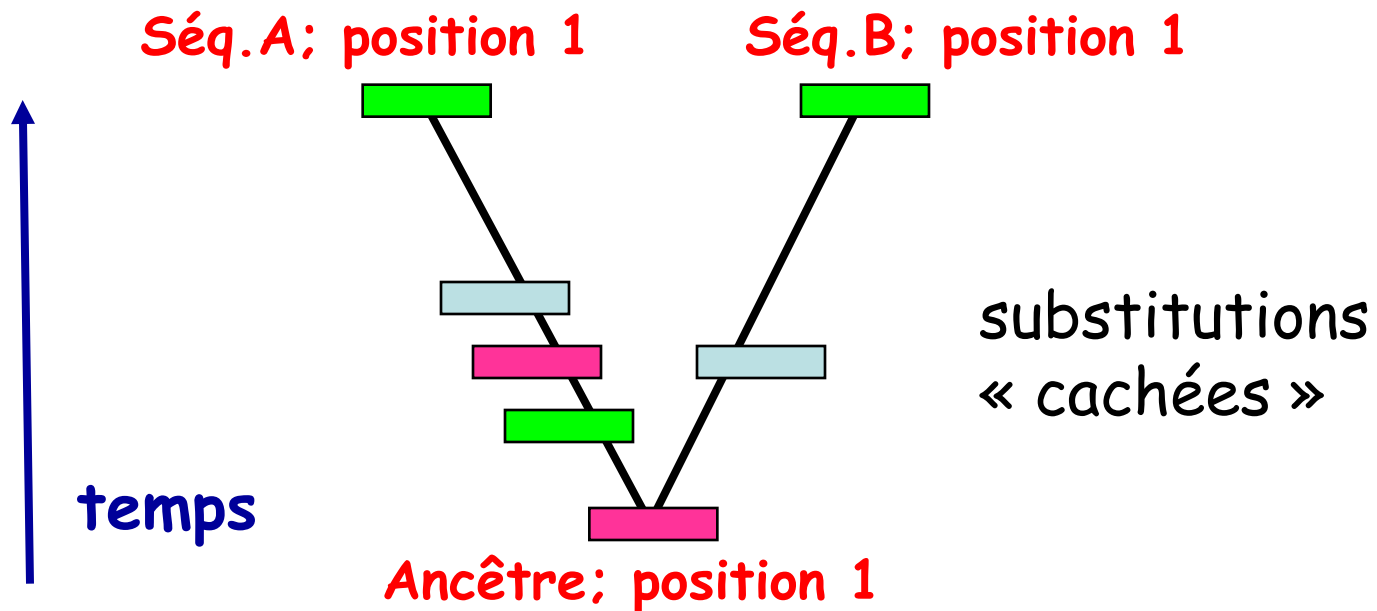


# Calcul de la distance-p

Séq X	C	G	A	T	G	A	T	A	A	C
Séq Y	C	G	A	A	A	A	C	A	G	C
				*	*		*		*	

$$D_{xy} = \frac{\text{Sites différents (k)}}{\text{Sites comparés (N)}} = 4 / 10 = 0.4$$

# Distance observée vs distance évolutive



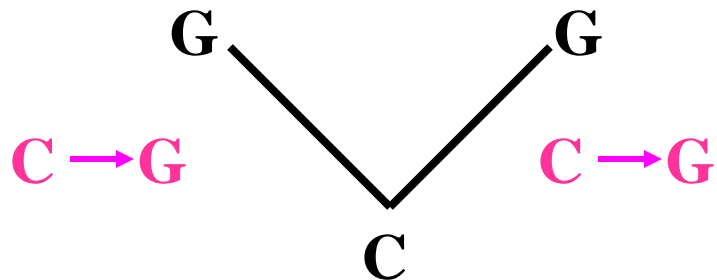
Plus le temps est grand, plus le nombre de substitutions cachées est grand.

Ce problème est plus important pour les séquences nucléotidiques (que 4 états par sites) que pour les séquences protéiques (20 états possibles par site).

# Substitutions "cachées"

## Substitution parallèle

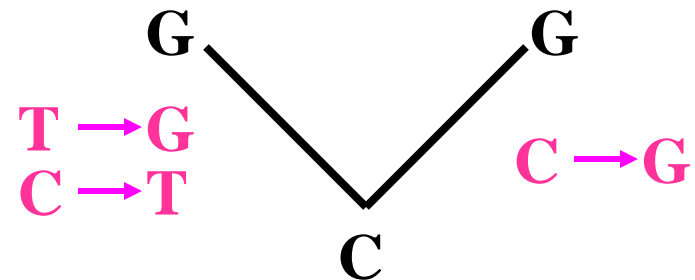
homme chimpanzé



0 sub. observée  
2 sub. réelles

## Substitution convergente

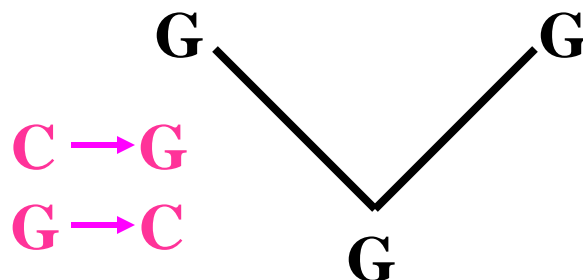
homme chimpanzé



0 sub. observée  
3 sub. réelles

## Substitution reverse

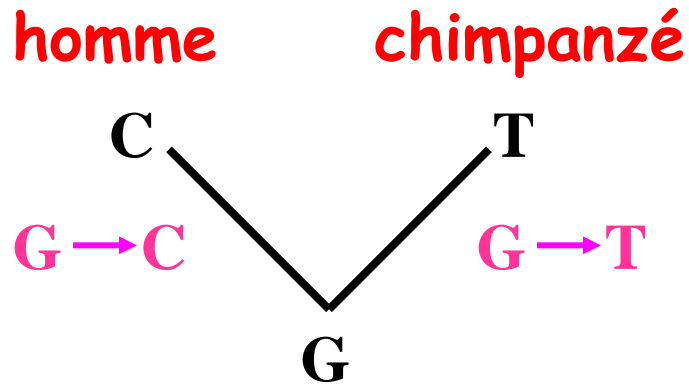
homme chimpanzé



0 sub. observée  
2 sub. réelles

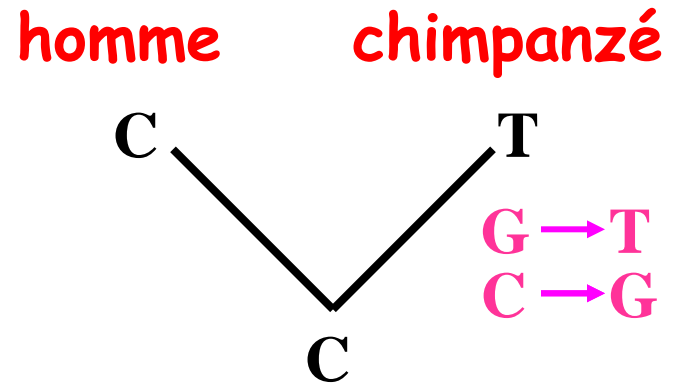
# Substitutions "cachées"

## Substitutions de coïncidence



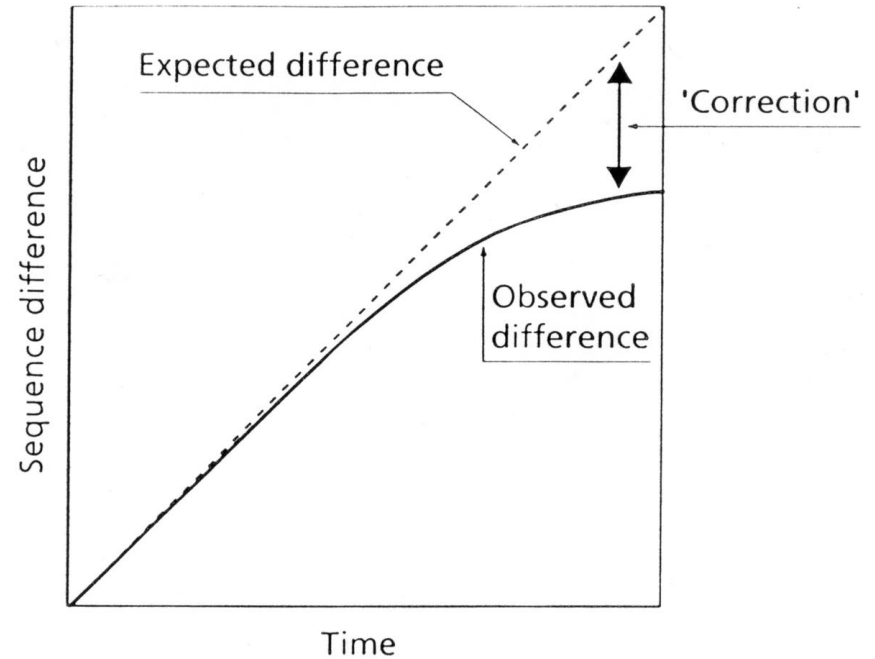
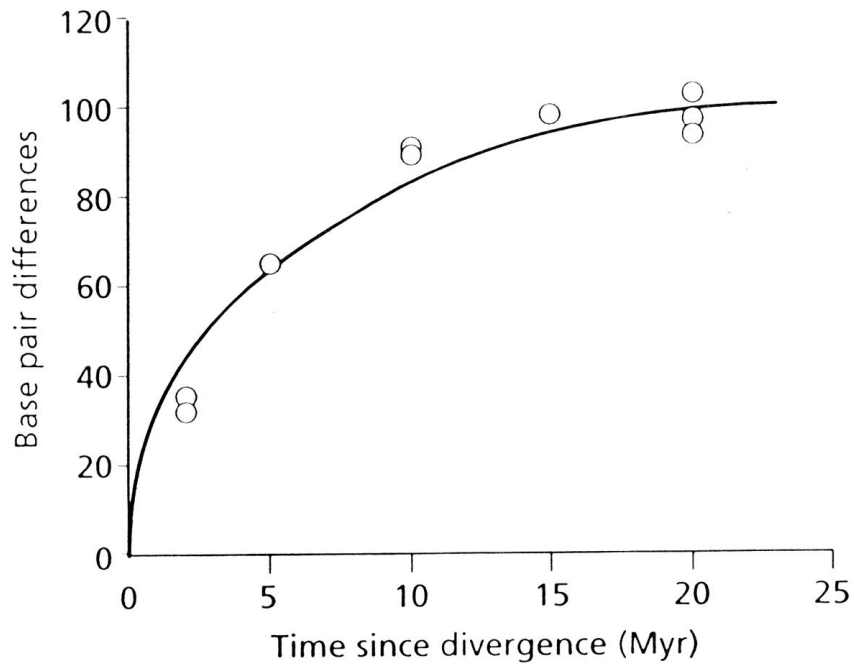
1 subst. observée  
2 subst. réelles

## Substitutions multiples



1 subst. observée  
2 subst. réelles

# Correction des distances observées



# Evolution des séquences nucléotidiques

Pour corriger la distance-p, il est nécessaire d'élaborer des modèles d'évolution des séquences nucléotidiques.

Elaboration d'un modèle mathématique simple d'évolution de séquences nucléotidiques:

le modèle de Jukes et Cantor (1969)

# Modèle Jukes-Cantor (JC)

$$P_t = \begin{vmatrix} \cdot & \alpha & \alpha & \alpha \\ \alpha & \cdot & \alpha & \alpha \\ \alpha & \alpha & \cdot & \alpha \\ \alpha & \alpha & \alpha & \cdot \end{vmatrix}$$

1 paramètre:  $\alpha$

Paramètres:

$$pA = pT = pG = pC$$

$$\alpha = \beta$$

Formule:

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

# Exemple

Pour deux séquences de la longueur de 500 nucléotides et qui diffèrent par 50 substitutions, la distance observée est égale à

$$D = 50 / 500 = 0.1$$

La distance évolutive calculée selon la formule de Jukes Cantor est égale à

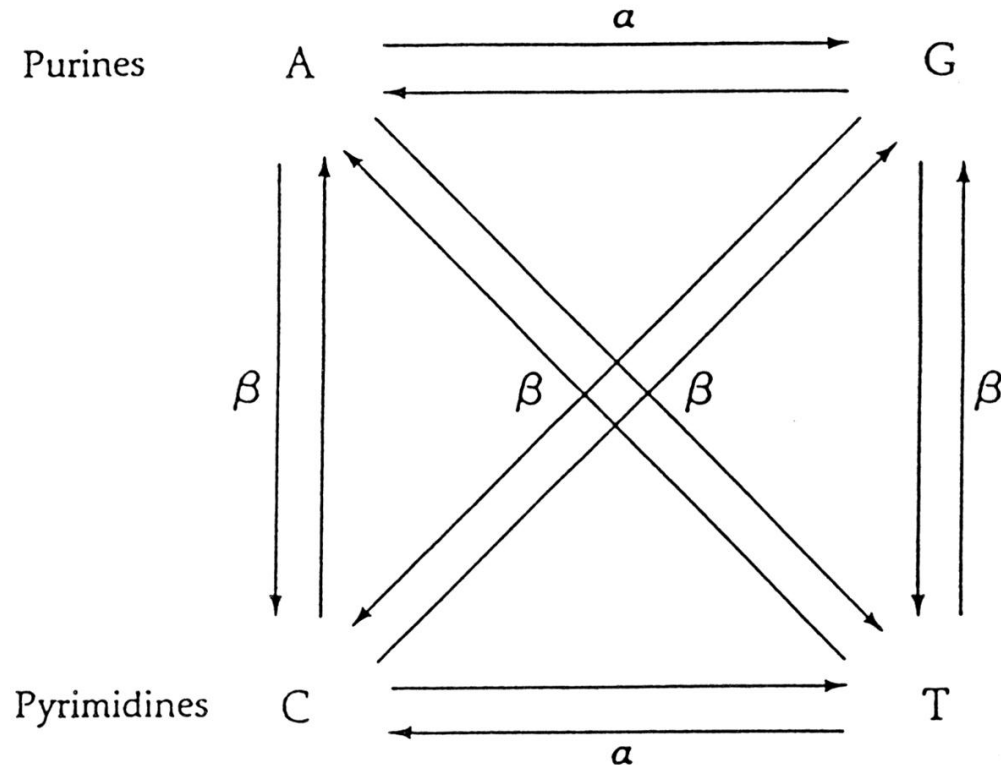
$$d_{xy} = 0.1073$$

Le nombre corrigé de substitutions est égal à 53.66, c'est-à-dire 3-4 substitutions n'ont pas été observées.

# Evolution des séquences nucléotidiques

Pour mieux coller à la réalité et avec l'augmentation du pouvoir de calcul des ordinateurs , les modèles d'évolution des séquences nucléotidiques se complexifient de plus en plus.

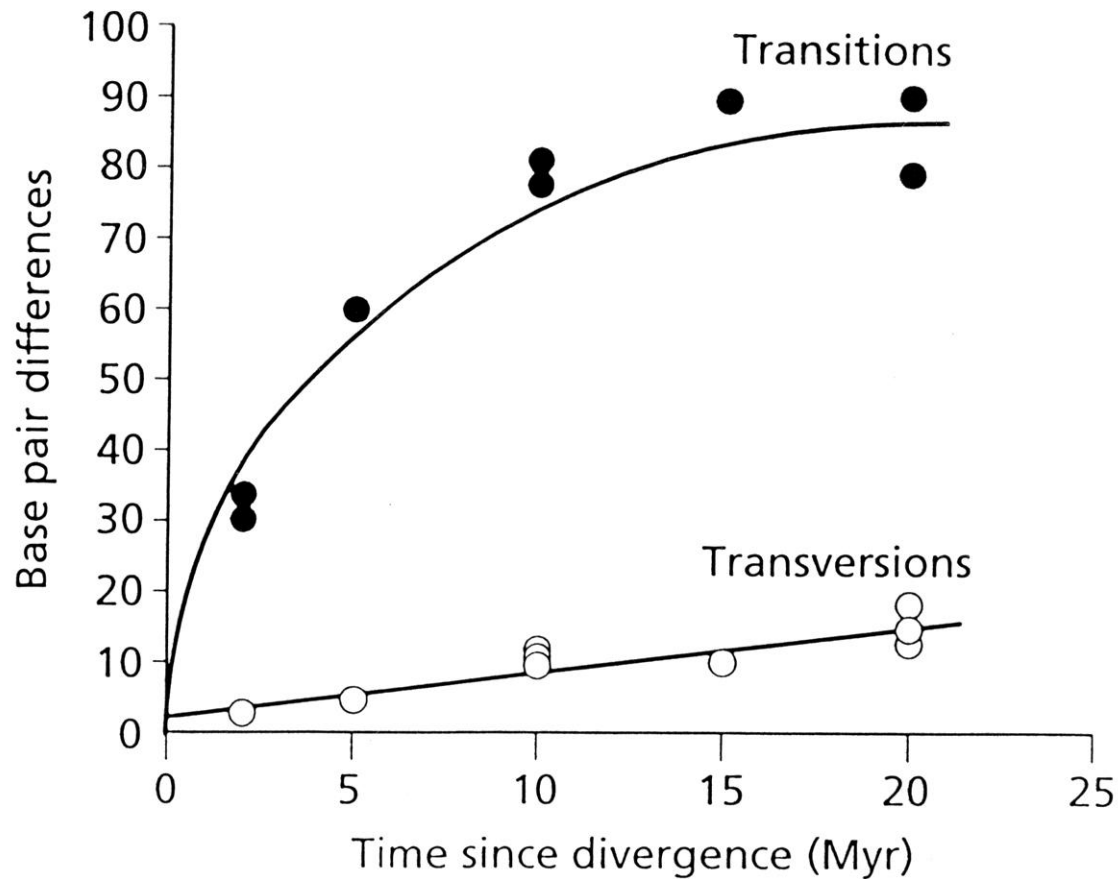
# Transitions vs transversions



Les transitions sont des substitutions de purines en purines ou de pyrimidines en pyrimidines **A↔G ; C↔T**

Les transversions sont des substitutions de pyrimidines en purines ou de purines en pyrimidines **A↔T ; A↔C ; C↔G ; G↔T**

# Les transitions sont souvent plus nombreuses que les transversions



	Ts/Tv ratio
mt DNA	9.0
12 S rRNA	1.75
$\alpha$ - et $\beta$ -globines	0.66

Le rapport entre les transitions et transversions dans les séquences du gène COII chez les bovidés.

## Modèle K2P (Kimura 2 paramètres)

$$P_t = \begin{pmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{pmatrix}$$

$$P_t = \begin{pmatrix} . & \beta & \alpha & \beta \\ \beta & . & \beta & \alpha \\ \alpha & \beta & . & \beta \\ \beta & \alpha & \beta & . \end{pmatrix}$$

2 paramètres:  $\alpha$  et  $\beta$

$\alpha$  - transitions (A-G, C-T)

$\beta$  - transversions

**Paramètres:**

$$p_A = p_T = p_G = p_C$$

$$\alpha \neq \beta$$

**Formule:**

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

## D 'autres modèles d'évolution:

### Modèle **F81** (Felsenstein, 1981)

$$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T \quad \alpha = \beta$$

4 paramètres

### Modèle **HKY85** (Hasegawa, Kishino, Yano, 1985)

$$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T \quad \alpha \neq \beta$$

6 paramètres

## Modèle GTR (General Time Reversible)

$$P_t = \begin{pmatrix} p_{AA} & p_{AC} & p_{AG} & p_{AT} \\ p_{CA} & p_{CC} & p_{CG} & p_{CT} \\ p_{GA} & p_{GC} & p_{GG} & p_{GT} \\ p_{TA} & p_{TC} & p_{TG} & p_{TT} \end{pmatrix}$$

$$P_t = \begin{pmatrix} . & p_{Aa} & p_{Ab} & p_{Ac} \\ p_{Ca} & . & p_{Cd} & p_{Ce} \\ p_{Gb} & p_{Gd} & . & p_{Gf} \\ p_{Tc} & p_{Te} & p_{Tf} & . \end{pmatrix}$$

### Paramètres:

$$\pi_A \neq \pi_C \neq \pi_G \neq \pi_T$$

$$a = A \leftrightarrow C \\ d = C \leftrightarrow G$$

$$b = A \leftrightarrow G \\ e = C \leftrightarrow T$$

$$c = A \leftrightarrow T \\ f = T \leftrightarrow G$$

Chaque substitution a une probabilité spécifique.

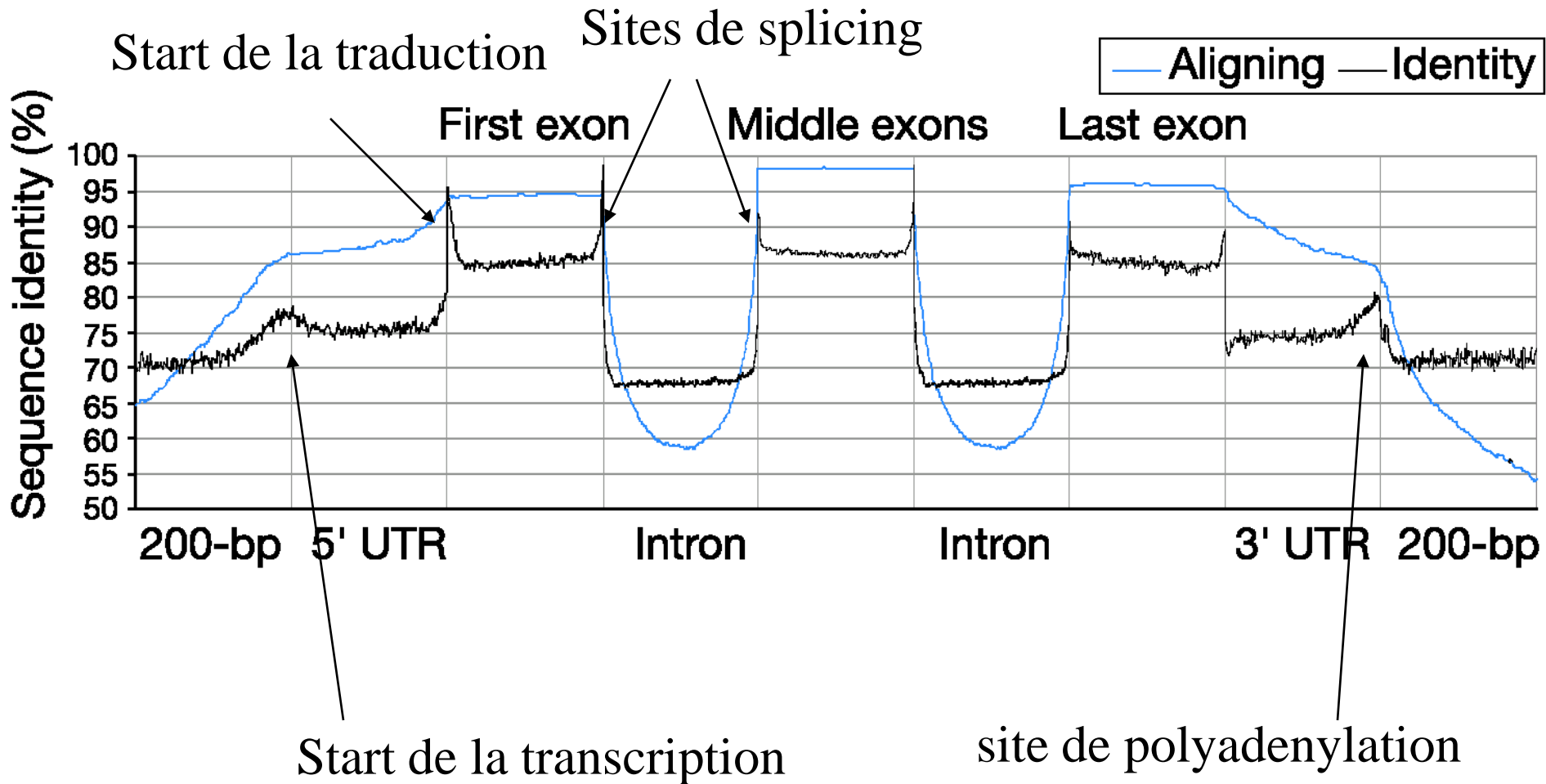
10 paramètres

Toutes les méthodes phylogénétiques courantes se basent sur les postulats suivants :

1. Les substitutions sont indépendantes les unes des autres.
2. Le taux de substitution est constant dans le temps et entre les lignées.
3. La fréquence des bases est identique entre lignées.
4. Les probabilités de substitutions sont les mêmes pour tous les sites et ne changent pas dans le temps.

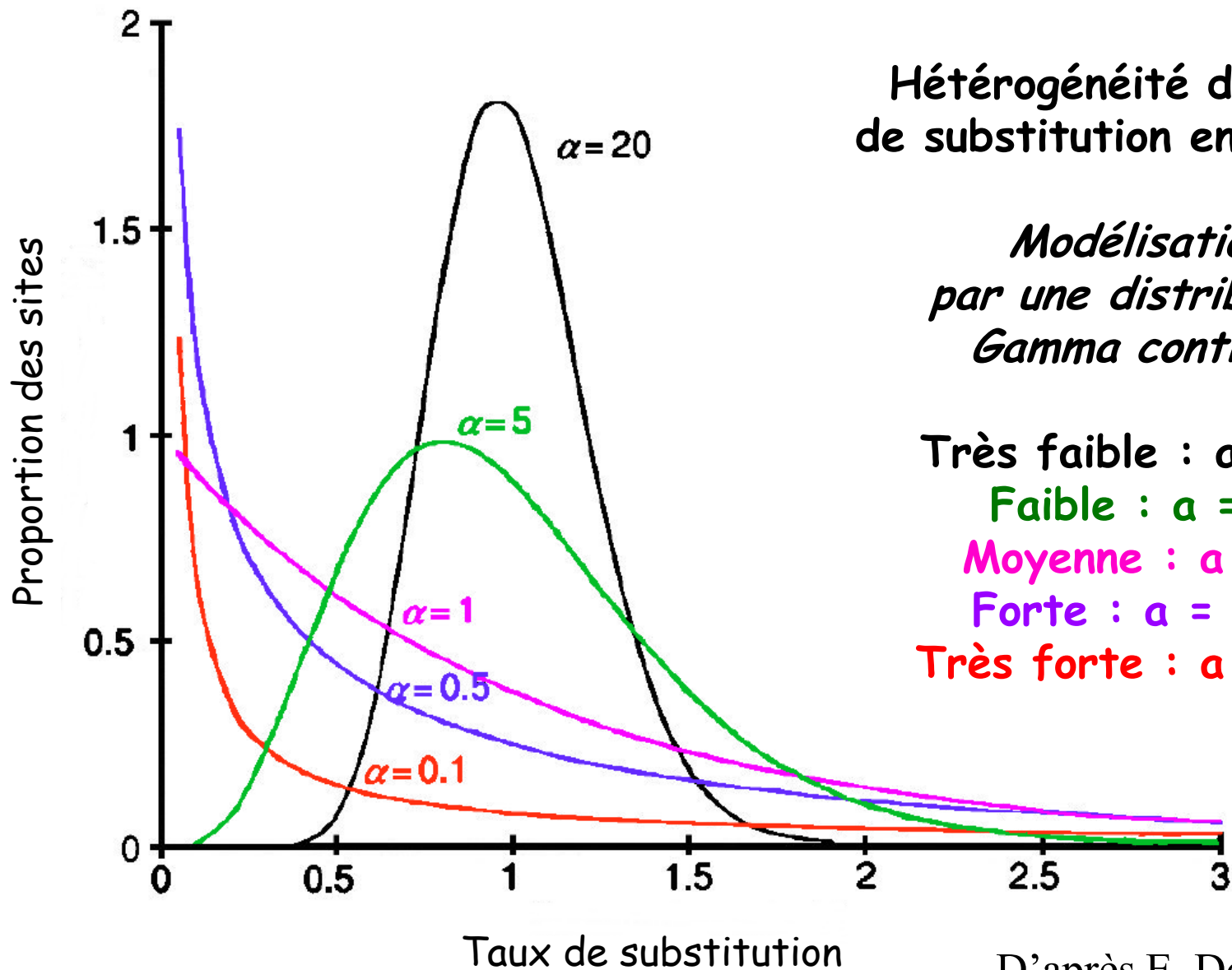
**Toutefois, ces postulats ne sont pas toujours respectés!**

# Conservation dans un gène 'modèle'



Sur la base de 3,165 comparaisons homme-souris

# Distribution gamma



Hétérogénéité de taux  
de substitution entre sites

*Modélisation  
par une distribution  
Gamma continue*

Très faible :  $\alpha = 20$

Faible :  $\alpha = 5$

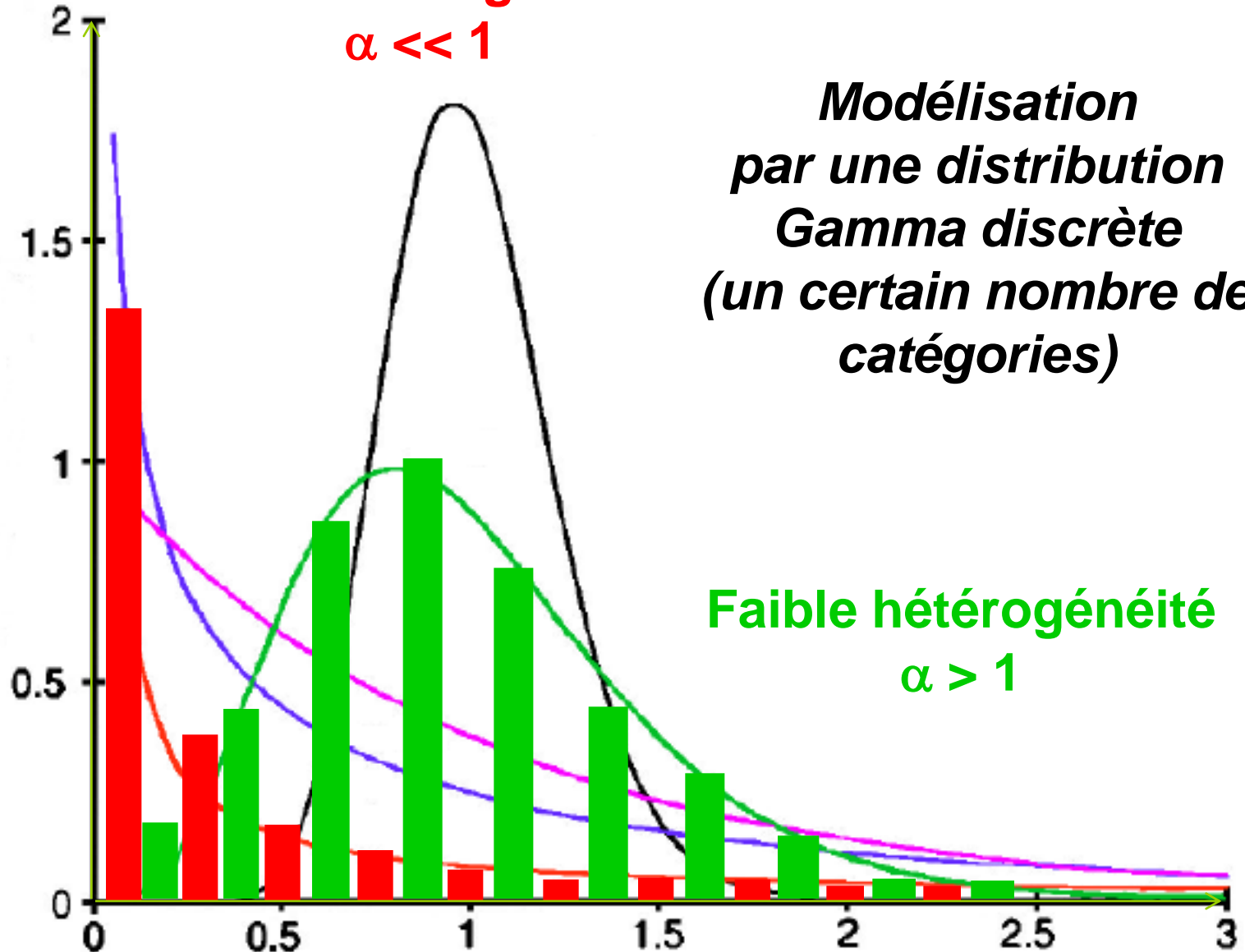
Moyenne :  $\alpha = 1$

Forte :  $\alpha = 0,5$

Très forte :  $\alpha = 0,1$

D'après E. Douzery

**Forte hétérogénéité**  
 $\alpha \ll 1$



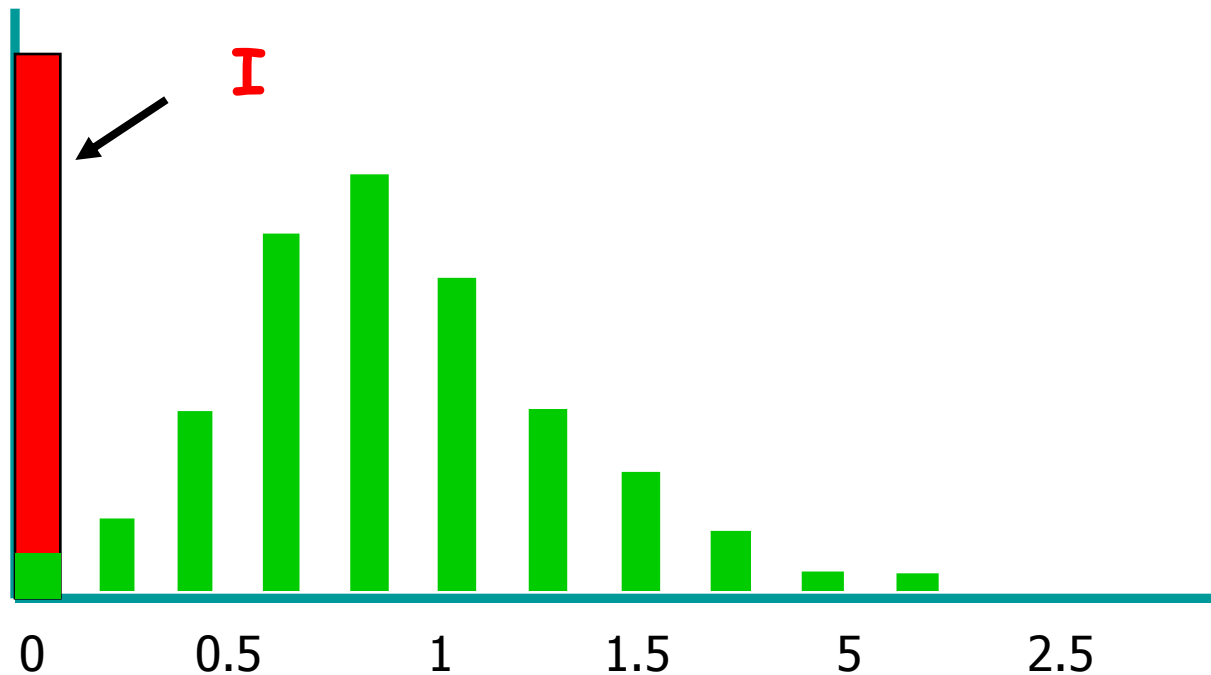
D'après E. Douzery

Estimations du paramètre  $\alpha$  (d'après Yang, 1996):

• albumine	1.05
• insulin	0.40
• prolactine	1.37
• 16S rRNA (stem)	0.29
• 16S rRNA (loop)	0.58
• 12S rRNA mt	0.16
• D-loop	0.17

# Sites invariants

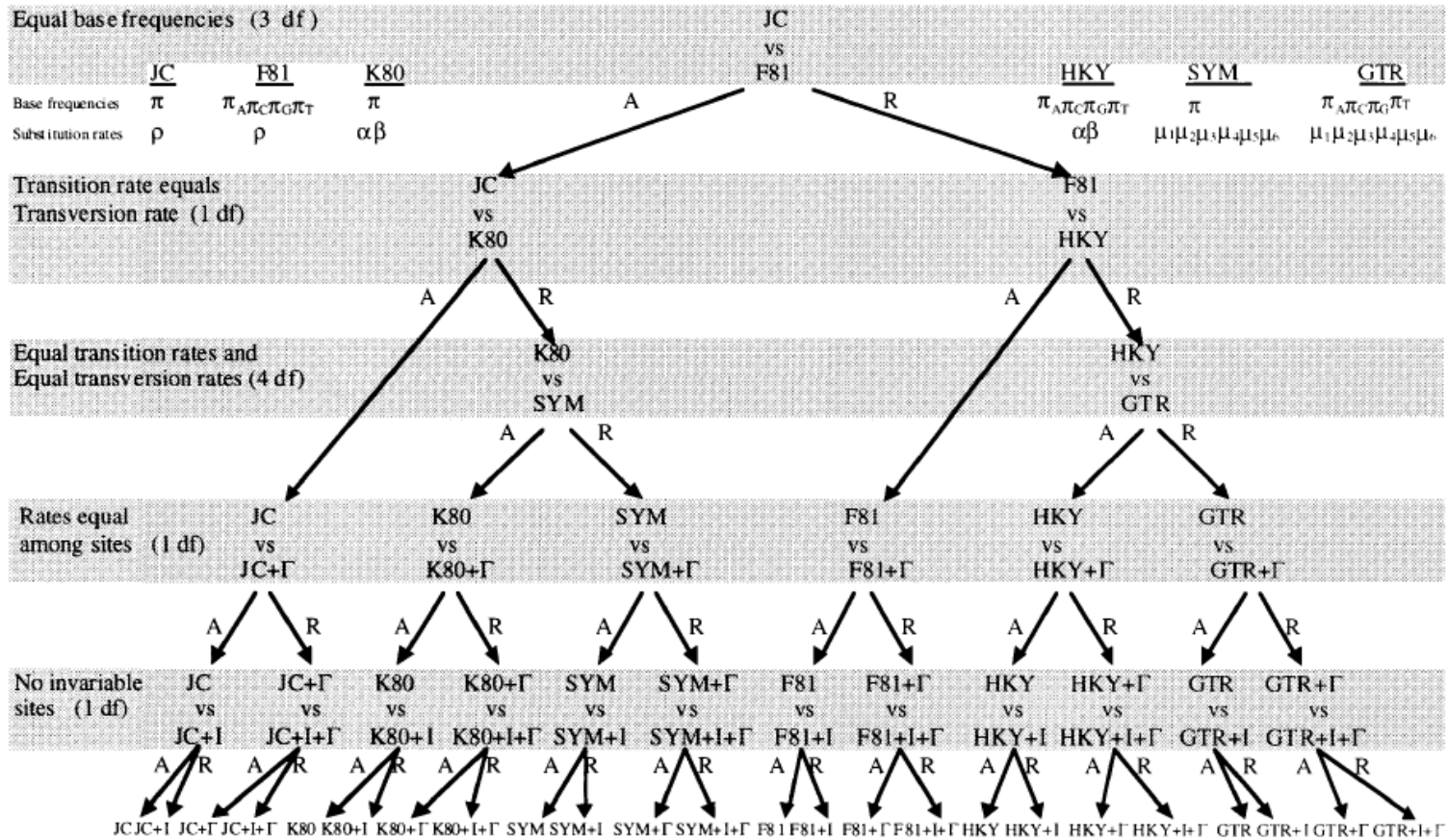
On peut traiter le nombre de sites invariants comme un paramètre supplémentaire indépendant:  $I$



La loi Gamma ne s'appliquera alors qu'aux sites libres de varier.

# Programme ModelTest: 56 modèles

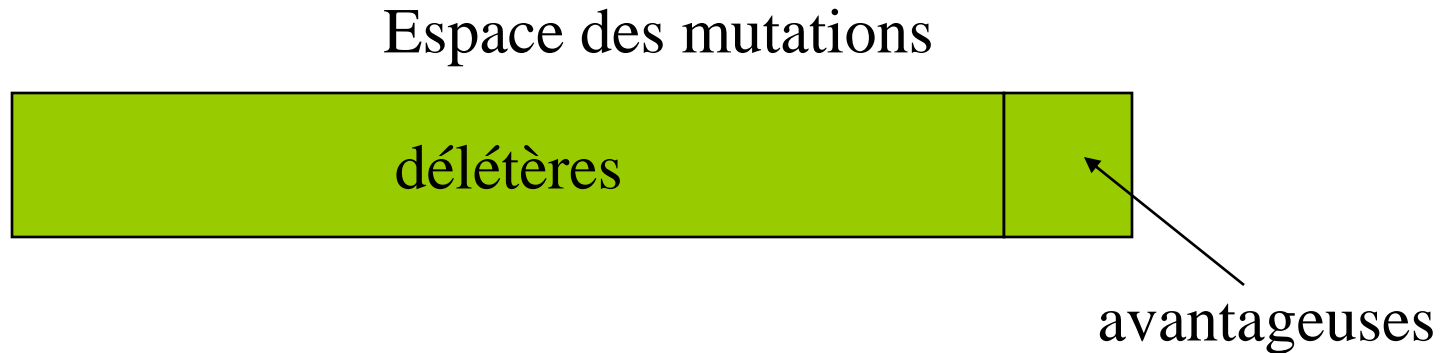
Posada D, Crandall KA, Bioinformatics. 1998;14(9):817-8



**Fig. 1.** Hierarchical hypothesis testing in MODELTEST. At each level the null hypothesis (upper model) is either accepted (A) or rejected (R). The models of DNA substitution are: JC (Jukes and Cantor, 1969), K80 (Kimura, 1980), SYM (Zharkikh, 1994), F81 (Felsenstein, 1981), HKY (Hasegawa *et al.*, 1985), and GTR (Rodríguez *et al.*, 1990).  $\Gamma$ : shape parameter of the gamma distribution; I: proportion of invariable sites. df: degrees of freedom. 1: equal base frequencies (0.25),  $\pi_A$ : frequency of adenine,  $\pi_C$ : frequency of cytosine,  $\pi_G$ : frequency of guanine,  $\pi_T$ : frequency of thymine.  $\rho$ : equal substitution rate,  $\alpha$ : transition rate,  $\beta$ : transversion rate;  $\mu_1$ : A $\Rightarrow$ C rate,  $\mu_2$ : A $\Rightarrow$ G rate,  $\mu_3$ : A $\Rightarrow$ T rate,  $\mu_4$ : C $\Rightarrow$ G rate,  $\mu_5$ : C $\Rightarrow$ T rate,  $\mu_6$ : G $\Rightarrow$ T rate.

# **Pression de sélection sur les séquences nucléotidiques**

Comment voyait-on l'évolution au niveau de la fonction d'un gène



- La plupart des mutations sont délétères, elles réduisent la fitness de l'organisme
  - elles sont éliminées de la population par **sélection purificatrice**.
- Quelques mutations sont avantageuses, elles augmentent la fitness de l'organisme
  - elles sont gardées par **sélection positive darwinienne (adaptative)**

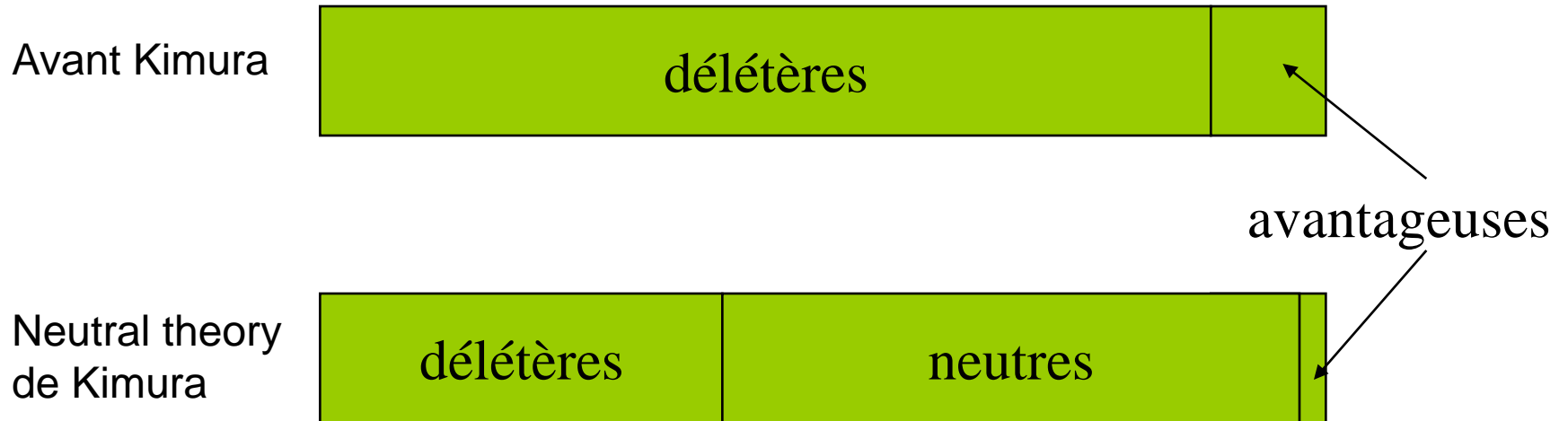
La fitness est fonction du taux de survie et de la fécondité

## Motoo Kimura



En 1968 Kimura analyse les changements dans l'hémoglobine, le cytochrome c et le triosephosphate dehydrogenase.

Kimura propose que la plupart des substitutions sont neutres du point de vue de la fitness et que très peu sont dues à la sélection positive darwinienne.



Une mutation neutre (ou allèle) ne change pas la fitness du wildtype

# Les changements synonymes peuvent être des mutations neutres

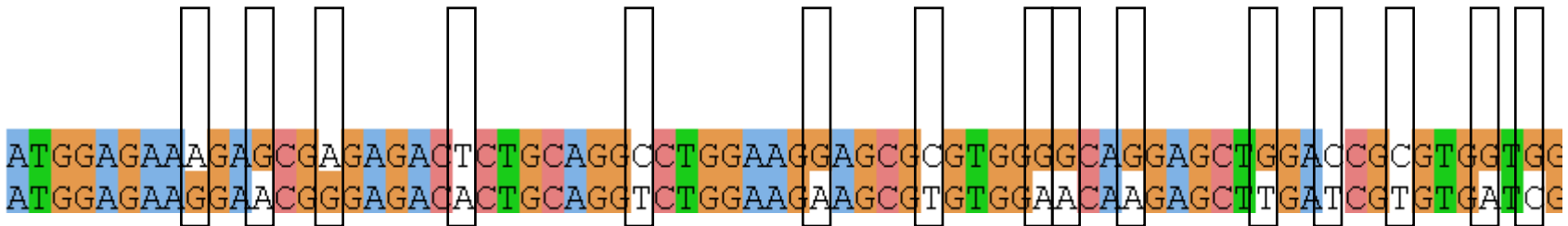
Prédictions:

- Si la plupart des changements dans l'ADN étaient dus à la sélection positive alors on déduirait que la plupart des changements auraient lieu en 1ère ou 2ème position des codons.
- Si les changements dans l'ADN incluent des mutations neutres, alors les 3ème positions devraient changer plus rapidement parce que les mutations synonymes sont plus vraisemblablement neutres.

King, J. L., and Jukes, T. H. 1969. Non-Darwinian evolution, Science 164, 788-798.

# Taux de substitution synonyme et taux de substitution non-synonyme.

- $K_S$  = nombre de substitutions **S**ynonymes par site synonyme
- $K_A$  = nombre de substitutions non-synonymes (**A**ltering) par site non-synonyme



On pourrait estimer le  $K_S$  et le  $K_A$  de manière empirique, en examinant les données mais des méthodes ont été développées pour le faire de manière systématique.

Les sites nucléotidiques peuvent être classifiés en 3 types en fonction de leur dégénérescence.

AAA	AAC	ACA	ACC	CAA	CAC	CCA	CCC
AAG	AAU	ACG	ACU	CAG	CAU	CCG	CCU
AGA	AGC	AUA	AUC	CGA	CGC	CUA	CUC
AGG	AGU	AUG	AUU	CGG	CGU	CUG	CUU
GAA	GAC	GCA	GCC	UAA	UAC	UCA	UCC
GAG	GAU	GCG	GCU	UAG	UAU	UCG	UCU
GGA	GGC	GUA	GUC	UGA	UGC	UUA	UUC
GGG	GGU	GUG	GUU	UGG	UGU	UUG	UUU

**2-fold Degenerate**  
deux possibilités synonymes

**4-fold degenerate**  
4 possibilités synonymes

**0-fold degenerate** - aucune possibilité synonyme  
tout changement modifie le aa

**Synonyme - Altère**

**Il y a 32 situations 4-fold degenerated en 3ème position sur les 61 codons**

AAA	AAC	ACA	ACC	CAA	CAC	CCA	CCC
AAG	AAU	ACG	ACU	CAG	CAU	CCG	CCU
AGA	AGC	AUA	AUC	CGA	CGC	CUA	CUC
AGG	AGU	AUG	AUU	CGG	CGU	CUG	CUU
GAA	GAC	GGA	GGC	UAA	UAC	UCA	UCC
GAG	GAU	GGG	GGU	UAG	UAU	UCG	UCU
GGA	GGC	GUA	GUC	UGA	UGC	UUA	UUC
GGG	GGU	GUG	GUU	UGG	UGU	UUG	UUU

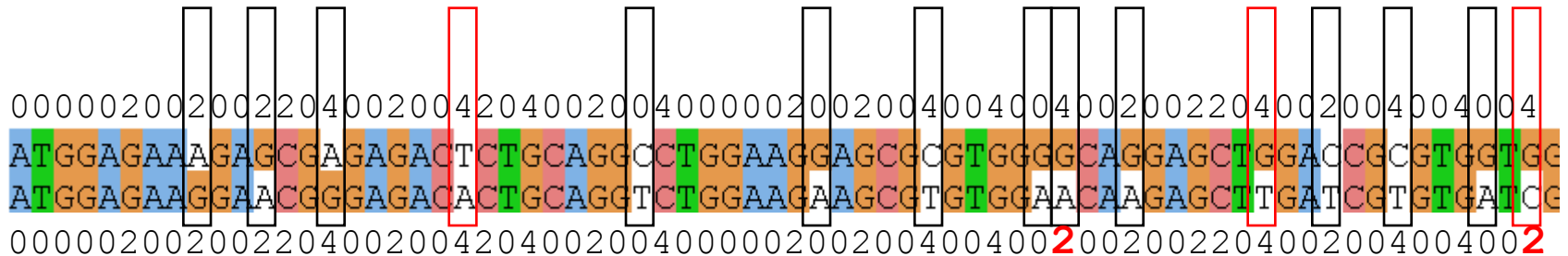
**Il y a 25 situations 2-fold degenerated en 3ème position  
et 8 situations en 1ère position des codons.**

AAA	AAG	ACA	ACC	CAA	CAC	CCA	CCC
AAG	AUU	ACG	ACU	CAG	CAU	CCG	CCU
AAA	AGC	AUA	ALC	CGA	GGC	CUA	CUC
AGG	AGU	AUG	AUU	CGG	CGU	CUG	CUU
GAA	GAC	GCA	GCC	UAA	UAC	UCA	UCC
GAG	GAU	GCG	GCU	UAG	UAU	UCG	UCU
GGA	GGC	GUA	GUC	UGA	UGC	UUA	UUC
GGG	GGU	GUG	GUU	UGG	UGU	UUG	UUU

**Les situations 0-fold degenerated sont les 2ème positions (61)  
et 53 des 1ère positions.**

AAA	AAC	ACA	ACC	GAA	GAC	GCA	GCC
AAG	AAU	AGG	AUU	GAG	GAU	GCG	GUU
AGA	AGC	AAA	AUC	GGA	GGC	GUA	GUC
AGG	AGU	AUG	AUU	GGG	GGU	GUG	GUU
GAA	GAC	GCA	GCC	UAA	UAC	LCA	LCC
GAG	GAU	GCG	GUU	UAG	UAU	LCG	LCU
GGA	GGC	GUA	GUC	UGA	LGC	LUA	LUC
GGG	GGU	GUG	GUU	LGG	LGU	LUG	LUU

En partant d'une paire de séquences alignées, classifier chaque site de la séquence en fonction de sa dégénérescence.



Compter le nombre de sites 4-,2-,0-fold degenerated.  
Si besoin, faire la moyenne entre les deux séquences.

Exemple:

$$L_0 = (45+45)/2 = 45$$

$$L_2 = (13+15)/2 = 14$$

$$L_4 = (10+8)/2 = 9$$

Classifier les mutations en fonction de leur dégénérescence et du type de substitution.

a. transition (S) or transversion (V)

b. dégénérescence (0,2,4)

	0-fold	2-fold	4-fold
transition	S0	S2	S4
transversion	V0	V2	V4

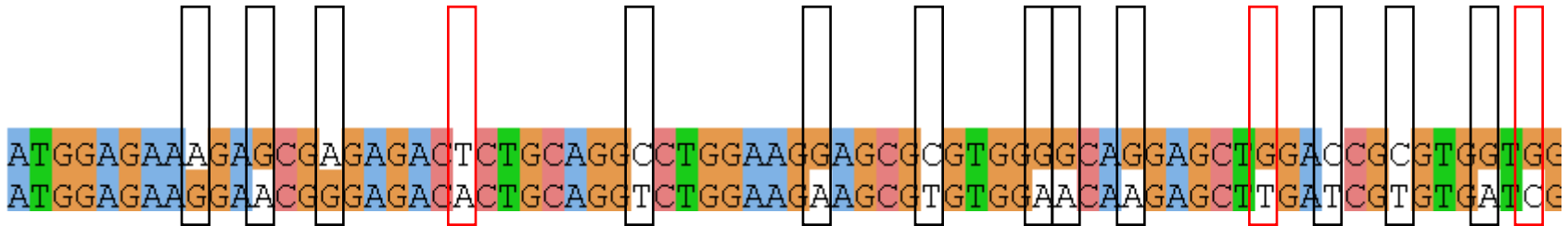
La simplification clé de cette méthode est la relation intéressante entre les transitions - transversions et la dégénérescence:

## Mutations synonymes

	0-fold	2-fold	4-fold
transitions	$S_0$	$S_2$	$S_4$
transversions	$V_0$	$V_2$	$V_4$

## Mutations non-synonymes

(Exceptions: 1ère position de l'arginine (CGA, CGG, AGA, AGG), dernière position de l'isoleucine (AUU, AUC, AUA)).



Calculer la proportion de substitutions pour les transitions et les transversions séparément pour pouvoir ensuite utiliser le modèle de Kimura (K2P) pour corriger les substitutions multiples:

- proportion de transitions de type  $i$ :  $P_i = S_i/L_i$   $i=0\text{-fold}, 2\text{-fold}, 4\text{-fold}$
- proportion de transversions de type  $i$ :  $Q_i = V_i/L_i$

Le modèle de K2P est utilisé pour corriger les substitutions multiples:

$$A_i = (1/2) \ln (1/(1- 2P_i - Q_i)) - (1/4) \ln (1/(1- 2Q_i))$$

$$B_i = (1/2) \ln (1/(1- 2Q_i))$$

Ici  $A_i$  = taux corrigé de transitions de type  $i$

## Calcul du $K_S$ et du $K_A$

$K_S$  = substitutions synonymes par site synonyme

nb.sites x taux  $S_2$   $S_4$   $V_4$  exprimés en proportion + corrigé avec K2P

$$K_S = \frac{L_2 A_2 + L_4 A_4 + L_4 B_4}{L_2/3 + L_4}$$

Par convention, 1/3 des sites 2-fold degenerated sont considérés synonymes

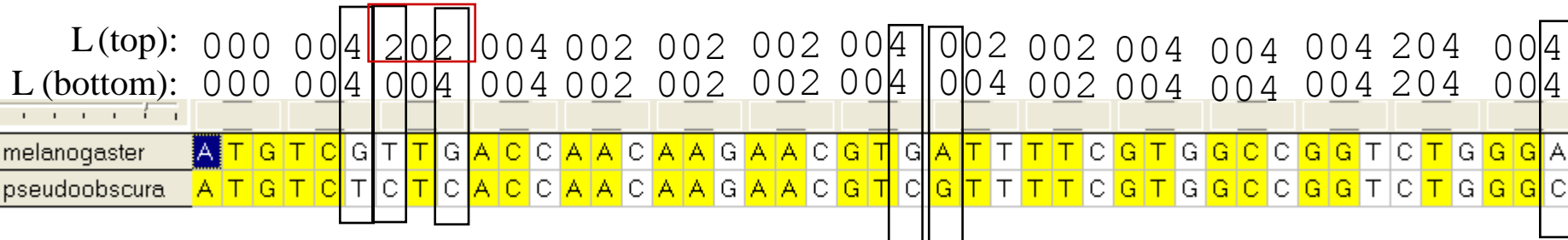
$K_A$  = substitutions non-synonymes par site non-synonyme

$V_0$   $V_2$   $S_0$  en proportion + K2P

$$K_A = \frac{L_0 B_0 + L_2 B_2 + L_0 A_0}{(2/3)L_2 + L_0}$$

Ici, 2/3 des sites 2-fold degenerated sont considérés non-synonymous

## un exemple



$$L_0 = (29+30)/2 = 29.5$$

$$S_0 = (2+1)/2 = 1.5$$

$$V_0 = 0$$

$$L_2 = (7+5)/2 = 6$$

$$S_2 = (1+0)/2 = 0.5$$

$$V_2 = 0$$

$$L_4 = (9+10)/2 = 9.5$$

$$S_4 = 0$$

$$V_4 = (4+4)/2 = 4$$

Proportions (P = transitions, Q = transversions)

$$P_0 = 1.5/29.5 \quad Q_0 = 0$$

$$P_2 = 0.5/6 \quad Q_2 = 0$$

$$P_4 = 0 \quad Q_4 = 4/9.5$$

L (top):	000	004	204	004	004	002	002	002	004	002	002	004	004	004	204	004																													
L (bottom):	000	004	004	004	004	002	002	002	004	004	002	004	004	004	204	004																													
<input checked="" type="checkbox"/> D. melanogaster	A	T	G	T	C	G	T	T	G	A	C	C	A	A	C	A	A	G	A	A	C	G	T	G	A	T	T	T	T	C	G	T	G	G	C	C	G	G	T	C	T	G	G	G	A
<input checked="" type="checkbox"/> D. pseudoobscura	A	T	G	T	C	T	C	T	C	A	C	C	A	A	C	A	A	G	A	A	C	G	T	C	G	T	T	T	T	C	G	T	G	G	C	C	G	G	T	C	T	G	G	G	C

## K2P corrections

$$A_i = \frac{1}{2}(\ln(1/(1-2P_i-Q_i))) - \frac{1}{4}(\ln(1/1-2Q_i))$$

$$B_i = \frac{1}{2}(\ln(1/(1-2Q_i)))$$

$$A_0 = 0.054 \quad B_0 = 0$$

$$A_2 = 0.091 \quad B_2 = 0$$

$$A_4 = -0.188 \quad B_4 = 0.923$$

les A et B sont les fréquences des transitions et transversions corrigées avec le modèle K2P.

$$K_S = B_4 + (A_2L_2 + A_4L_4)/(L_2 + L_4) = 0.843$$

$$K_A = A_0 + (L_0B_0 + L_2B_2)/(L_0 + L_2) = 0.054$$

Les taux de substitutions synonymes sont généralement plus grands que les taux de substitutions non-synonymes.

**TABLE 8.1** Numbers of nucleotide substitutions per 100 sites between species<sup>a</sup>

<i>Species pair</i>	<i>Synonymous sites</i>		<i>Nonsynonymous sites</i>	
	$K_S$	$L^b$	$K_A$	$L^b$
Mouse–rat	18.0 ± 0.7	4,229	1.8 ± 0.1	15,217
Mouse–hamster	30.3 ± 1.0	4,229	2.9 ± 0.1	15,217
Rat–hamster	31.3 ± 1.0	4,229	2.7 ± 0.1	15,217
Mouse–human	53.4 ± 1.5	4,229	5.2 ± 0.2	15,217
Rat–human	51.6 ± 1.5	4,229	5.0 ± 0.2	15,217
Hamster–human	52.3 ± 1.5	4,229	5.1 ± 0.1	15,217

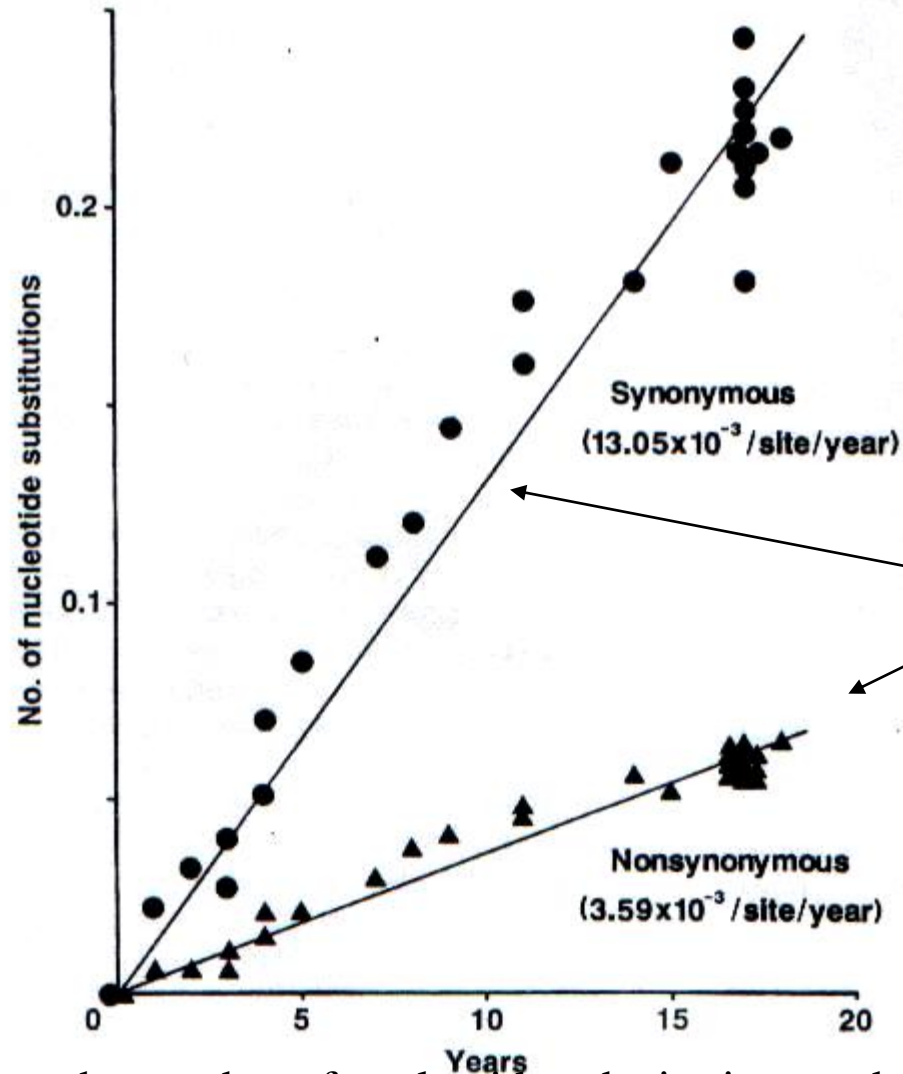
From O’Uigin and Li (1992).

<sup>a</sup>Computed by Li et al.’s (1985b) method.

<sup>b</sup>Number of sites compared.

From Li, W-H Molecular Evolution who took it from:  
Oh’Uigin and Li. JME 1992 35: 377-384

# The Molecular Clock of Viral Evolution



Relationship between the number of nucleotide substitutions and the difference in the year of isolation for the H3 hemagglutinin gene of human influenza A viruses. All sequence comparisons were made with the strain isolated in 1968.

**Le taux  $K_s/K_a$  est utilisé pour évaluer la pression de sélection sur les régions codantes.**

$K_A/K_s \approx 1$  -> les séquences évoluent de manière neutre, sans sélection.

$K_A/K_s > 1$  -> La région codante est sous sélection positive.

$K_A/K_s \ll 1$  -> La région codante est sous sélection purificatrice.