PARTIE II

LES DISTANCES ÉVOLUTIVES
LES MODÈLES DE SUBSTITUTION

PARTIE II

LES DISTANCES ÉVOLUTIVES

LES MODÈLES DE SUBSTITUTION

Distances évolutives

➔ La notion de distance découle naturellement de celle de similitude :

La similitude Génétique

C x A
G x T
G =G
G x T
A= A
A= A
G= G
A= A
A= A
A= A

Sij= *(Sites identiques (i))/ Sites comparés (N)*

Distance génétique

C x A
G x T
G =G
G x T
A= A
A= A
G= G
A= A
A= A
A= A

Dij= *(Sites différents (k))/ Sites comparés (N)*

➔ Elles reflètent la quantité d'évolution survenue pendant un temps t,

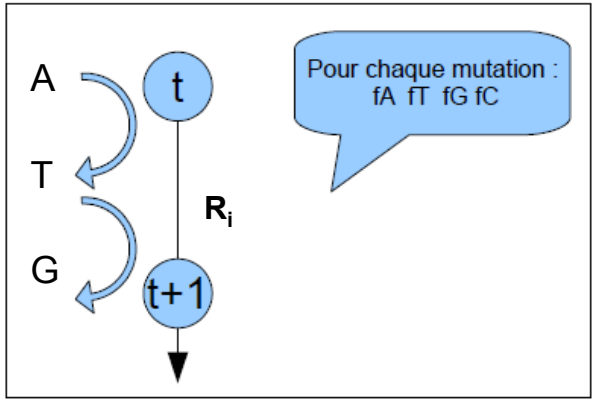
➔ Plus la similitude entre deux séquences S_{ij} est forte, plus la distance entre elles D_{ij} est faible

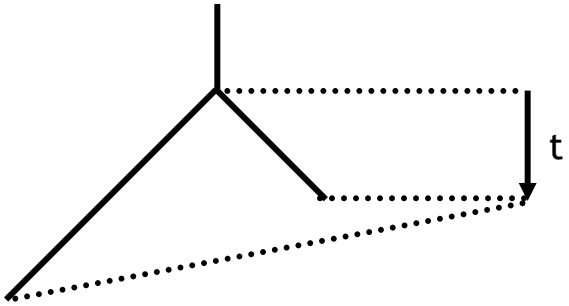
Distances évolutives

➔ Elles ne sont pas une simple fonction de temps,

ex: deux séquences qui sont issues d'un groupe frère dans une phylogénie sont séparées par le même intervalle de temps t, mais peuvent avoir subit une quantité d'évolution différente,

- r_i = taux d'évolution
- Longueur de la branche ⇔ quantité d'évolution = $r_i t_i$





Distances évolutives : Estimations

- Exemple

Évolution d'une séquence : 20 nucléotides ayant chacun la même probabilité de muter entre t et t+1

t=0	AAAAAAAAAAAAAAAAAAAA	t=10	AGAAAAAAAAGAAAACCACA
t=1	ATAAAAAAAAAAAAAAAAAA	t=11	CGAAAAAAAAGAAAACCACA
t=2	ATAAAAAAAAGAAAAAAAAA	t=12	CGAAAAAAAAGAAAACCACA
t=3	ATAAAAAAAAGAAAAAAAAG	t=13	CGAAATAAAAGAAAACCACA
t=4	ATAAAAAAAAGAAAACAAAG	t=14	CGAAATAAAAGAAAACCACA
t=5	ATAAAAAAAAGAAAACAAAC	t=15	CGAAATAAAAGAAAACCACA
t=6	ATAAAAAAAAGAAAACAAAC	t=16	CGATATAAAAGAAAACCACA
t=7	AGAAAAAAAAGAAAACAAAC	t=17	CGATATAAAAGAACACCACA
t=8	AGAAAAAAAAGAAAACAAAA	t=18	CGATATAAGAGAACACCACA
t=9	AGAAAAAAAAGAAAACAACA	t=19	CGATATAAGAGAACACCATA

Le nombre de différences observées entre la séquence à t=0 et la séquence à t=19 est de 10

$$D = k / n = 10 / 20 = 0.5$$

Or le nombre de mutations réelles entre la séquence à t=0 et la séquence à t=19 => 19

et le nombre de changements réels entre la séquence à t=0 et la séquence à t=19 => 16

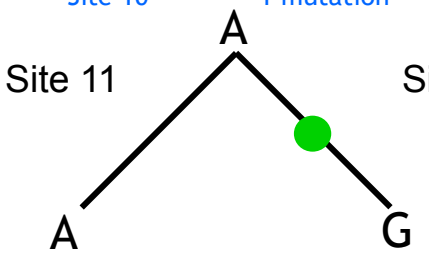
=> D réel $\geq k / n \Leftrightarrow D = k/n$ + changements cachés

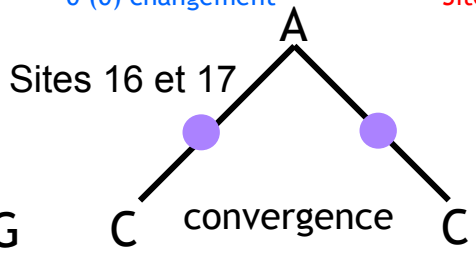
Distances évolutives : Estimations

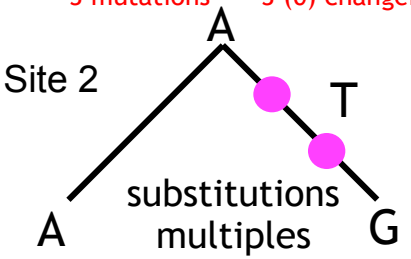
	1 3 5 7 9 11 13 15 17 19		1 3 5 7 9 11 13 15 17 19
t=0	AAAAAAAAAAAAAAAAAAAA	t=10	AGAAAAAAAAGAAAACCACA
t=1	ATAAAAAAAAAAAAAAAAAA	t=11	CGAAAAAAAAGAAAACCACA
t=2	ATAAAAAAAAGAAAAAAAAA	t=12	CGAAAAAAAAGAAAACCACA
t=3	ATAAAAAAAAGAAAAAAAAG	t=13	CGAAATAAAAGAAAACCACA
t=4	ATAAAAAAAAGAAAACAAAG	t=14	CGAAATAAAAGAAAACCACA
t=5	ATAAAAAAAAGAAAACAAAC	t=15	CGAAATAAAAGAAAACCACA
t=6	ATAAAAAAAAGAAAACAAAC	t=16	CGATATAAAAGAAAACCACA
t=7	AGAAAAAAAAGAAAACAAAC	t=17	CGATATAAAAGAACACCACA
t=8	AGAAAAAAAAGAAAACAAAA	t=18	CGATATAAGAGAACACCACA
t=9	AGAAAAAAAAGAAAACAACA	t=19	CGATATAAGAGAACACCATA

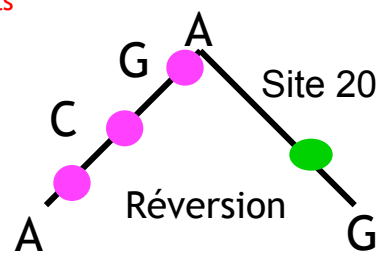
réels(visibles)

Site 1	1 mutation	1 (1) changement	Site 11	1 mutation	1 (1) changement
Site 2	2 mutations	2 (1) changements	Site 12	1 mutation	0 (0) changement
Site 3	0 mutation	0 (0) changement	Site 13	0 mutation	0 (0) changement
Site 4	1 mutation	1 (1) changement	Site 14	1 mutation	1 (1) changement
Site 5	0 mutation	0 (0) changement	Site 15	0 mutation	0 (0) changement
Site 6	2 mutations	1 (1) changements	Site 16	1 mutation	1 (1) changement
Site 7	0 mutation	0 (0) changement	Site 17	1 mutation	1 (1) changement
Site 8	0 mutation	0 (0) changement	Site 18	0 mutation	0 (0) changement
Site 9	1 mutation	1 (1) changement	Site 19	3 mutations	3 (1) changements
Site 10	1 mutation	0 (0) changement	Site 20	3 mutations	3 (0) changements







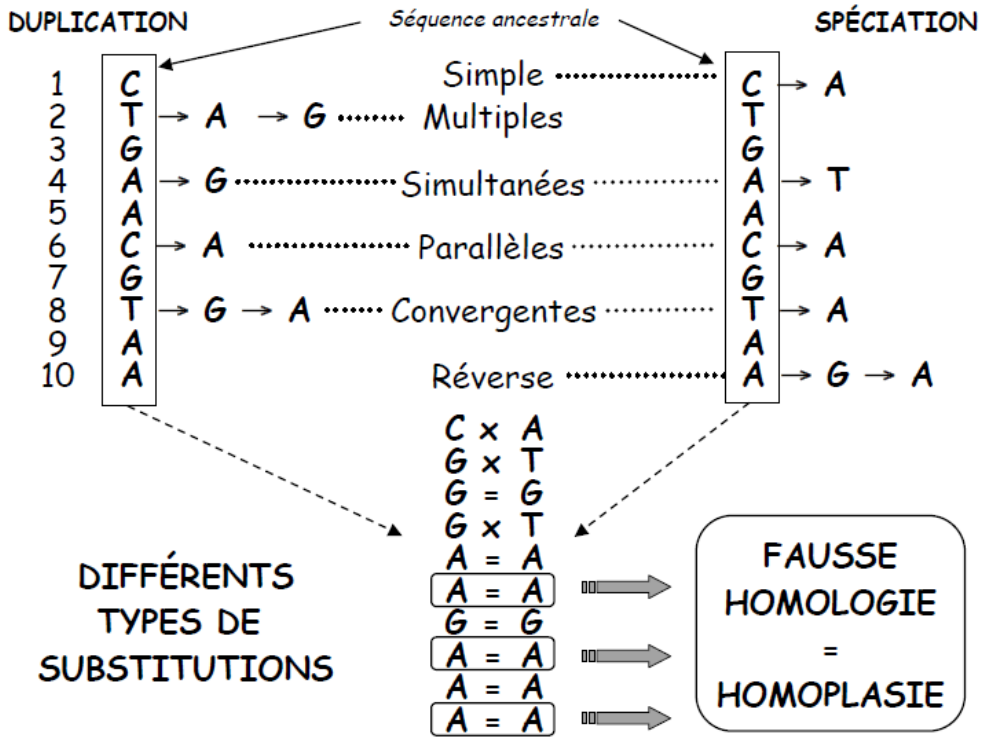


Distances évolutives : Estimations

De façon basique, les séquences d'ADN ne sont pas très informatives sur l'histoire évolutive :

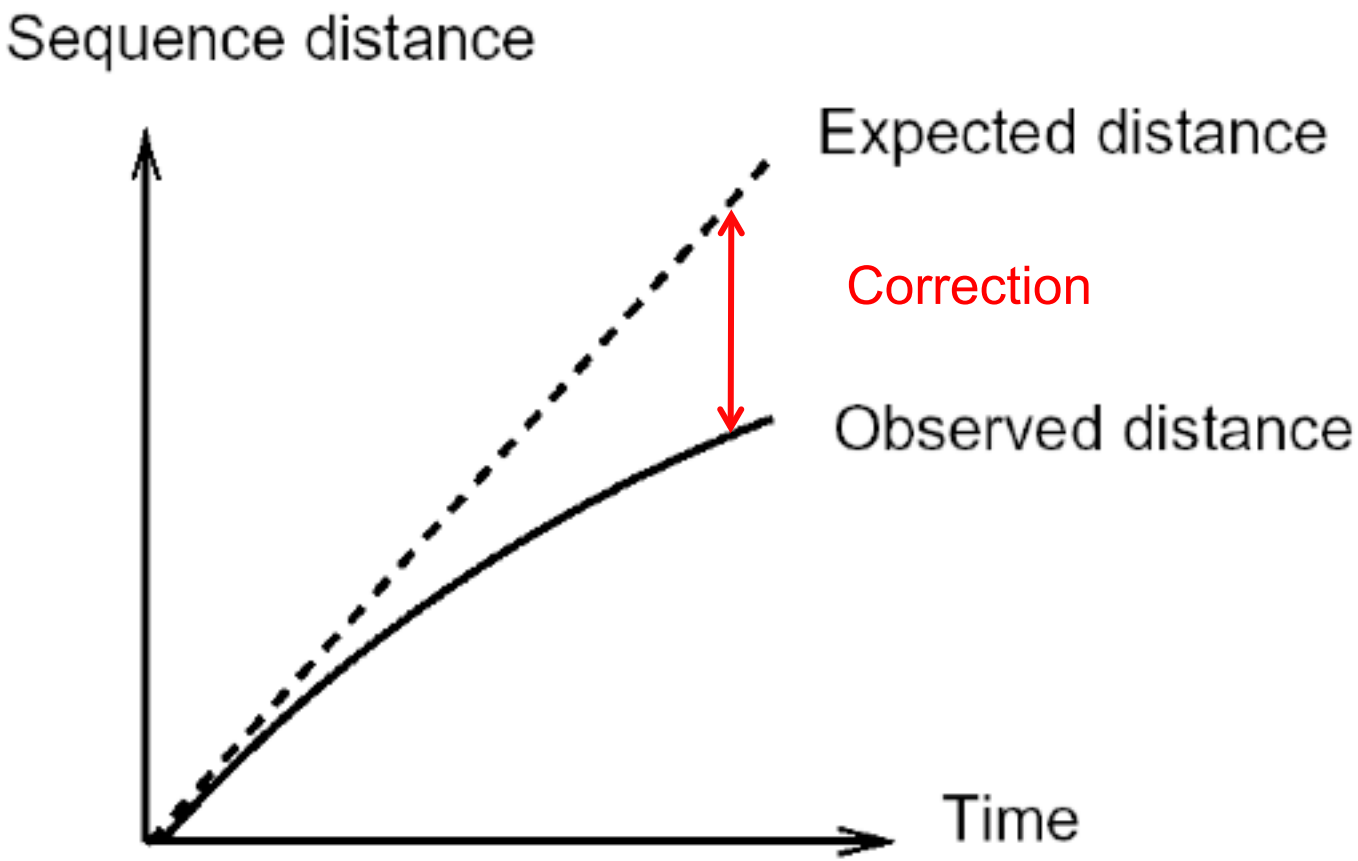
- Pour chaque site, le nombre maximal de différences est de 1.
- Il existe seulement 4 états. Donc, si il y a plus d'une substitution, nous perdons la substitution passée.
- Les substitutions multiples cachent l'histoire évolutive entre les séquences.
- Nous pouvons classifier les substitutions d'une autre manière.

Distances évolutives : Estimations



	Nombre de substitutions observées	Nombre de substitutions réelles
Simple	1	1
Multiples	1	2
Simultanées	1	2
Parallèles	0	2
Convergentes	0	3
Réverse	0	2

Distances évolutives : Estimations



Observed distance: $d_0(t \to \infty) = 3/4$

PARTIE II

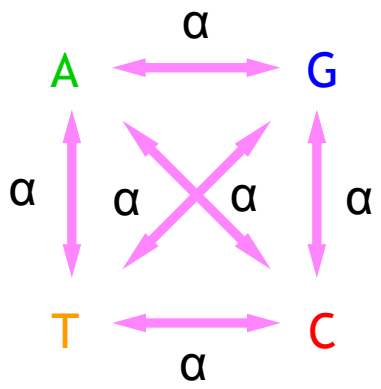
LES DISTANCES ÉVOLUTIVES

LES MODÈLES DE SUBSTITUTION

Modèles de Substitution : Jukes et Cantor

- C'est le modèle d'évolution le plus simple,
- Toutes les positions changent avec la même probabilité,
- Tous les changements sont équiprobables,
- Égalité des fréquences nucléotidiques :

$$\pi_A = \pi_G = \pi_C = \pi_T = 1/4$$



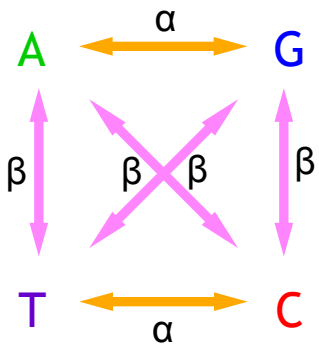
α = taux de changements par unité de temps

Modèles de substitution : Kimura 2 paramètres

- Modèle plus sophistiqué que le modèle de Jukes et Cantor
- Les sites évoluent indépendamment et de la même façon (à taux constant au cours du temps)
- 2 taux de substitutions : transitions, transversions
- Egalité des fréquences nucléotidiques

$$\pi_A = \pi_G = \pi_C = \pi_T = 1/4$$

- Modèle symétrique dans les séquences on attend autant de purines que de pyrimidines



Transition A ⇔ G (purine ⇔ purine) ou T ⇔ C (pyrimidine vers pyrimidine)

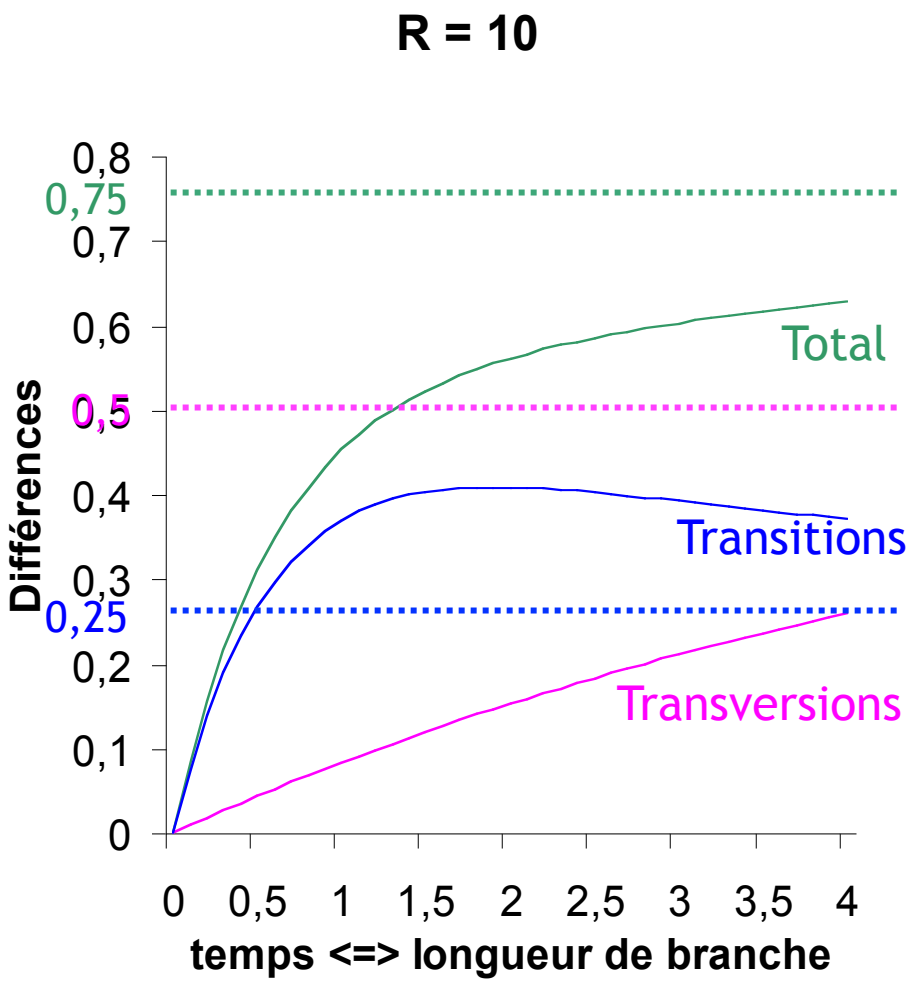
Transversion A ⇔ C ou A ⇔ T ou G ⇔ T ou G ⇔ C (purine ⇔ pyrimidine)

Rapport transition/transversion = $\alpha/2\beta$

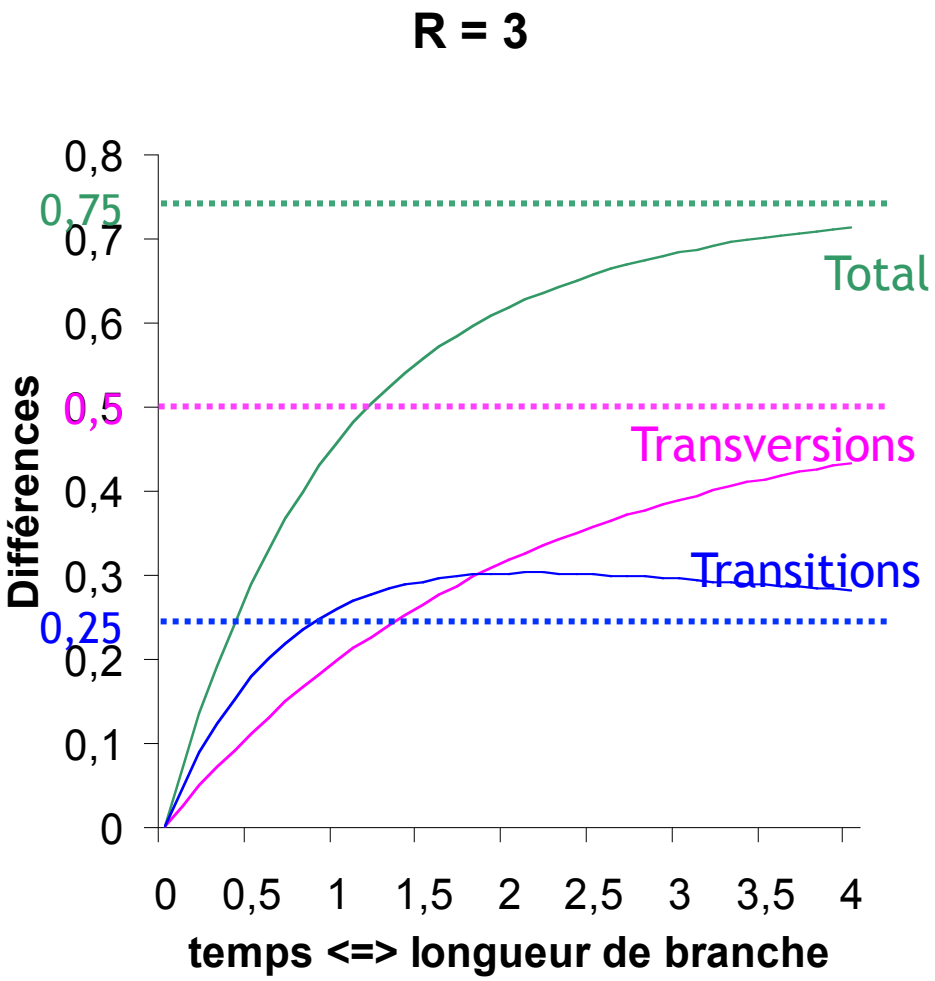
Taux de changement total = $\alpha + 2\beta$

Remarque Jukes et Cantor est un cas particulier du K2P quand $\beta = \alpha \Leftrightarrow R = 1/2$

Modèles de substitution : Kimura 2 paramètres



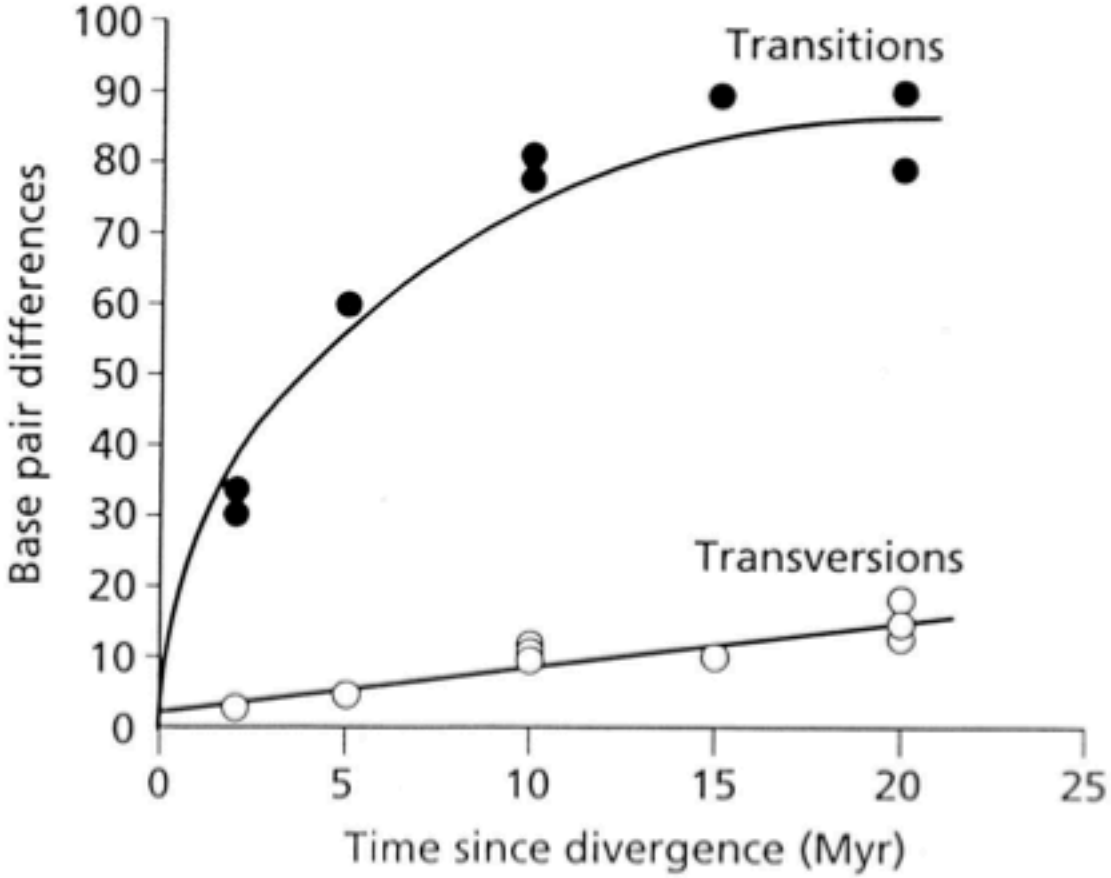
Typique de l' ADN mitochondrial



Typique de l' ADN nucléaire mammalien

Au bout d' un certain temps les transitions saturent et n'apportent plus d'informations.

Modèles de substitution : Kimura 2 paramètres



	Ts/Tv ratio
mt DNA	9.0
12 S rRNA	1.75
α- et β-globines	0.66

Modèles de substitution : autres

- Les modèles HKY et F84 acceptent des fréquences nucléotidiques arbitraires
  - F81 - F84 (Felsenstein , 1981; Kishino et Hasegawa – 1989)
    - $\pi A \neq \pi C \neq \pi G \neq \pi T$ et $\alpha = \beta$ (4 paramètres)

M =

↗	A	T	C	G
A	–	π_T	π_C	π_G
T	π_A	–	π_C	π_G
C	π_A	π_T	–	π_G
G	π_A	π_T	π_C	–

  - HKY (Hasegawa, Kishino et Yano – 1985)
    - $\pi A \neq \pi C \neq \pi G \neq \pi T$ et $\alpha \neq \beta$ (6 paramètres)

M =

↗	A	T	C	G
A	–	$\beta\ \pi_T$	$\beta\ \pi_C$	$\alpha\ \pi_G$
T	$\beta\ \pi_A$	–	$\alpha\ \pi_C$	$\beta\ \pi_G$
C	$\beta\ \pi_A$	$\alpha\ \pi_T$	–	$\beta\ \pi_G$
G	$\alpha\ \pi_A$	$\beta\ \pi_T$	$\beta\ \pi_C$	–

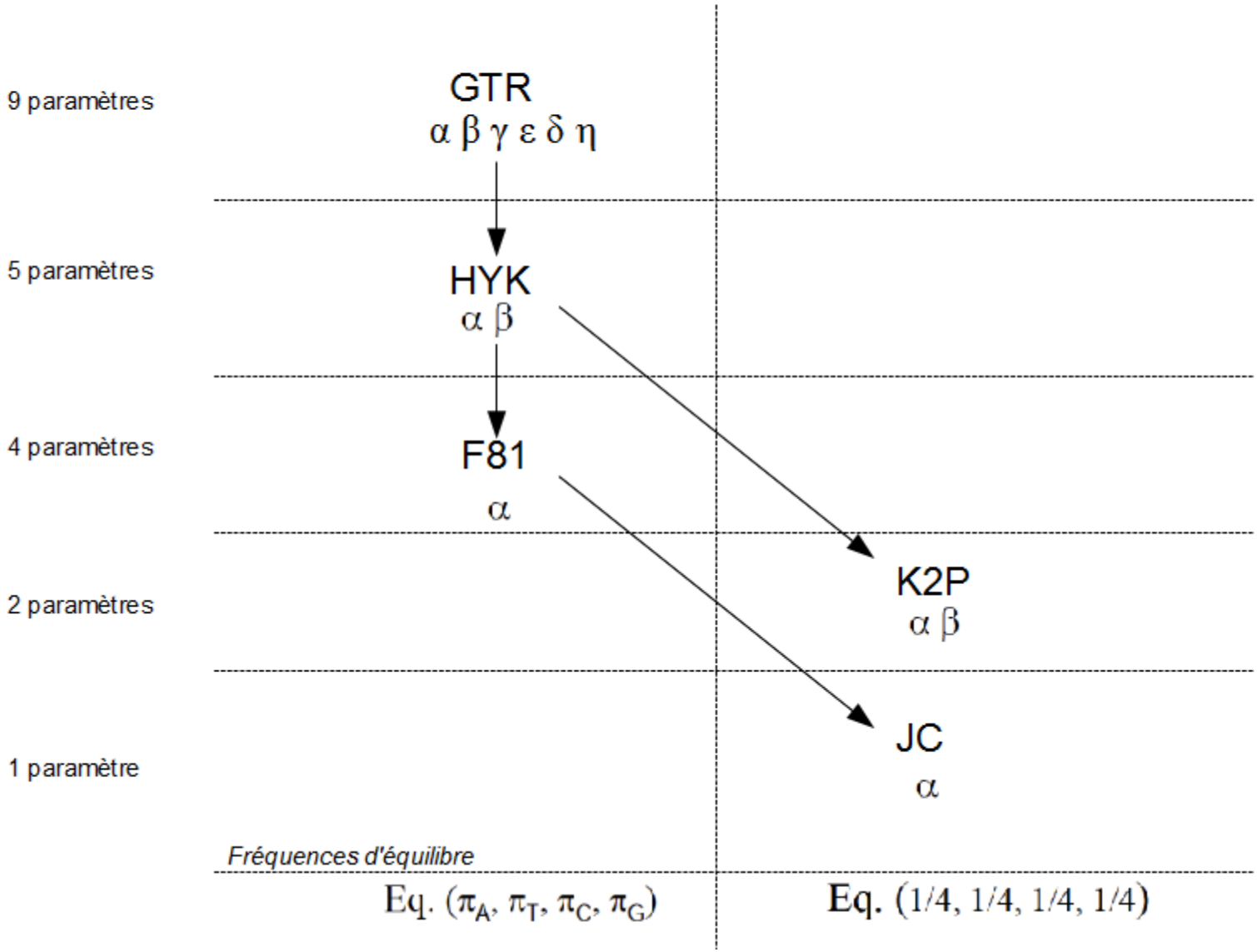
- Le modèle GTR (General Time Reversible)
  - $\pi A \neq \pi C \neq \pi G \neq \pi T$
  - a=A<->C; b=A<->G; c=A<->T; d=C<->G; e=C<->T; f=T<->G

(10 paramètres)

M =

↗	A	T	C	G
A	–	$\rho_{AT}\pi_T$	$\rho_{AC}\pi_C$	$\rho_{AG}\pi_G$
T	$\rho_{AT}\pi_A$	–	$\rho_{CT}\pi_C$	$\rho_{GT}\pi_G$
C	$\rho_{AC}\pi_A$	$\rho_{CT}\pi_T$	–	$\rho_{CG}\pi_G$
G	$\rho_{AG}\pi_A$	$\rho_{GT}\pi_T$	$\rho_{CG}\pi_C$	–

Modèles de substitution : Résumé



Modèles de substitution

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.

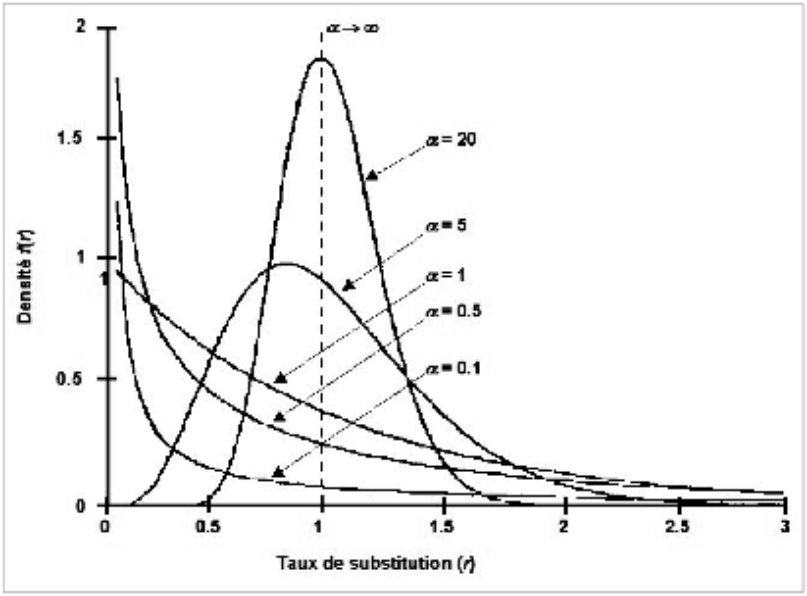
⇒ Toutefois, ces postulats ne sont pas toujours respectés!

⇒ **NON REALISTES D' UN POINT DE VUE BIOLOGIQUE**

⇒ **Développement d'autres modèles d'évolution**

Modèles de substitution : loi gamma

Modèles de distribution gamma des sites ou log normal



Un seul paramètre (alpha) définit la variation des sites :

$\alpha \rightarrow \infty$: tous les sites changent au même taux
$\alpha \rightarrow 0$: sites invariants ou évoluant lentement
$\alpha = 2$: loi de poisson
$\alpha > 10$: loi de distribution normale

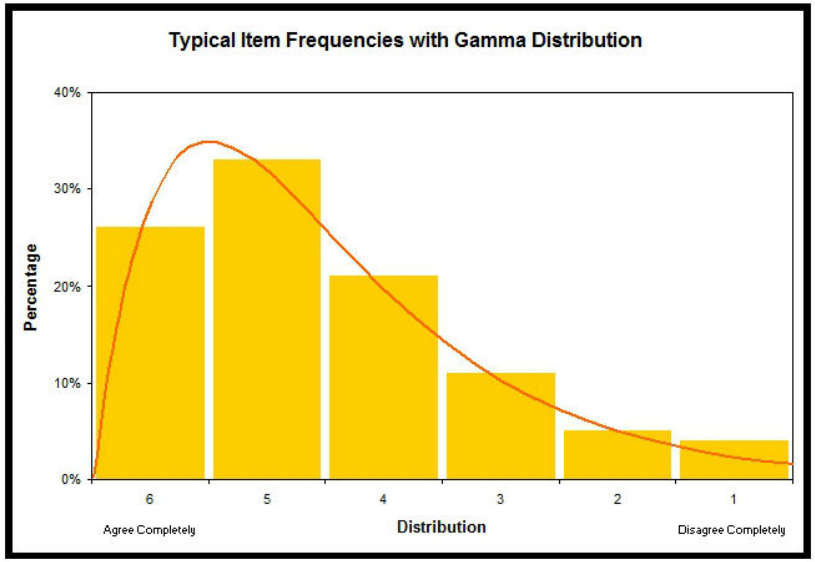
<u>Estimations du paramètre alpha (d'après Yang, 1996):</u>

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

<u>Alpha =0,5 pour ADN et 0,7 pour proteines</u>

Modèles de substitution : Loi gamma

- Simulation de la vitesse d'évolution des sites



La vitesse d'évolution des sites suit une loi gamma.
La loi gamma est une distribution continue=>Informatiquement impossible de calculer (trop long)

⟹ Transformation en loi discrète en la découpant en catégories (N)
Habituellement, N=4 pour ADN et N=6 pour protéines

Modèles de substitutions : évolution protéique

Modèle de dépendance des sites (Tillier) et les modèles d'évolution protéique

Modèle de Dayhoff et Eck (1968)

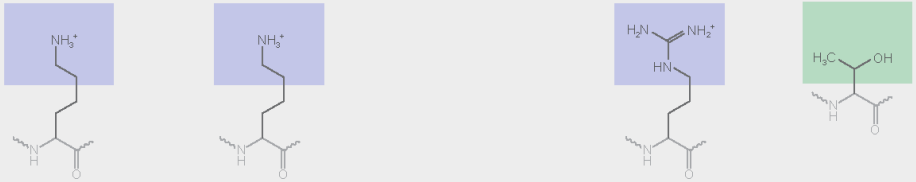
Utilisation d'un modèle à 20 états de caractères

- Les probabilités de passage d'un acide aminé à l'autre sont calculées à partir de comparaisons de séquences réelles
- Le code génétique n'intervient pas directement pour le calcul des fréquences de changements
- Les fréquences empiriques de changements reflètent les contraintes dues au code

Modèles de substitutions : évolution protéique

- Analyse de 1572 changements inférés par parcimonie à partir de 71 jeux de données de protéines proches
- Établissement d'une matrice de probabilité de mutation qui montre les probabilités de changement d'un acide aminé en un autre pour différents temps de divergence => **Matrices PAM = Point of accepted mutation**

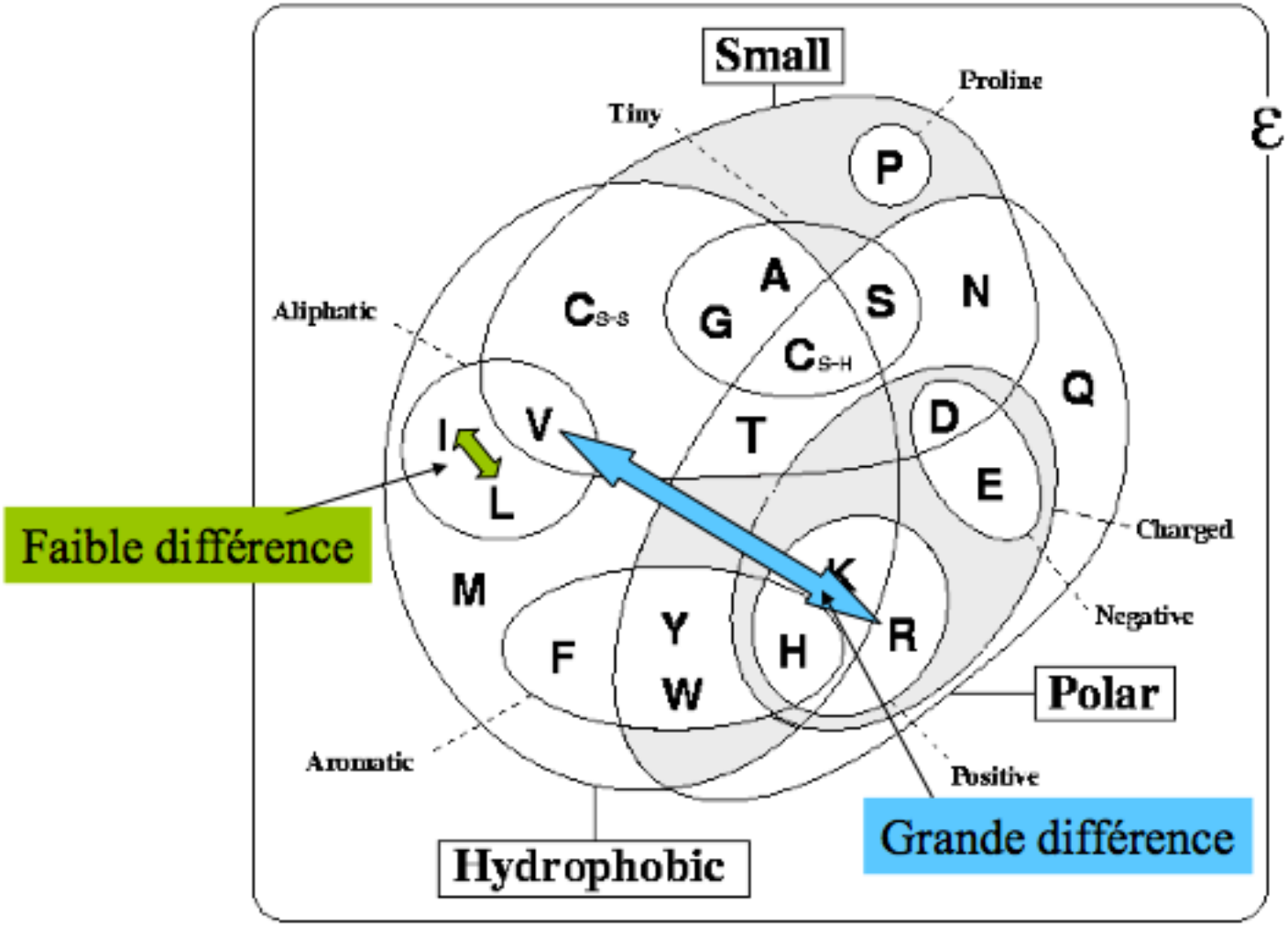
	No mutation	Point mutations			
		Silent	Nonsense	Missense	
				conservative	non-conservative
DNA level	TTC	TTT	ATC	TCC	TGC
mRNA level	AAG	AAA	UAG	AGG	ACG
protein level	**Lys**	**Lys**	**STOP**	**Arg**	**Thr**



PAM 001 :

- Matrice de probabilité de changement d'un acide aminé en un autre le long d'une branche le long de laquelle 1% des acides aminés ont changé

Modèles de substitutions : évolution protéique



Propriétés des 20 acides aminés

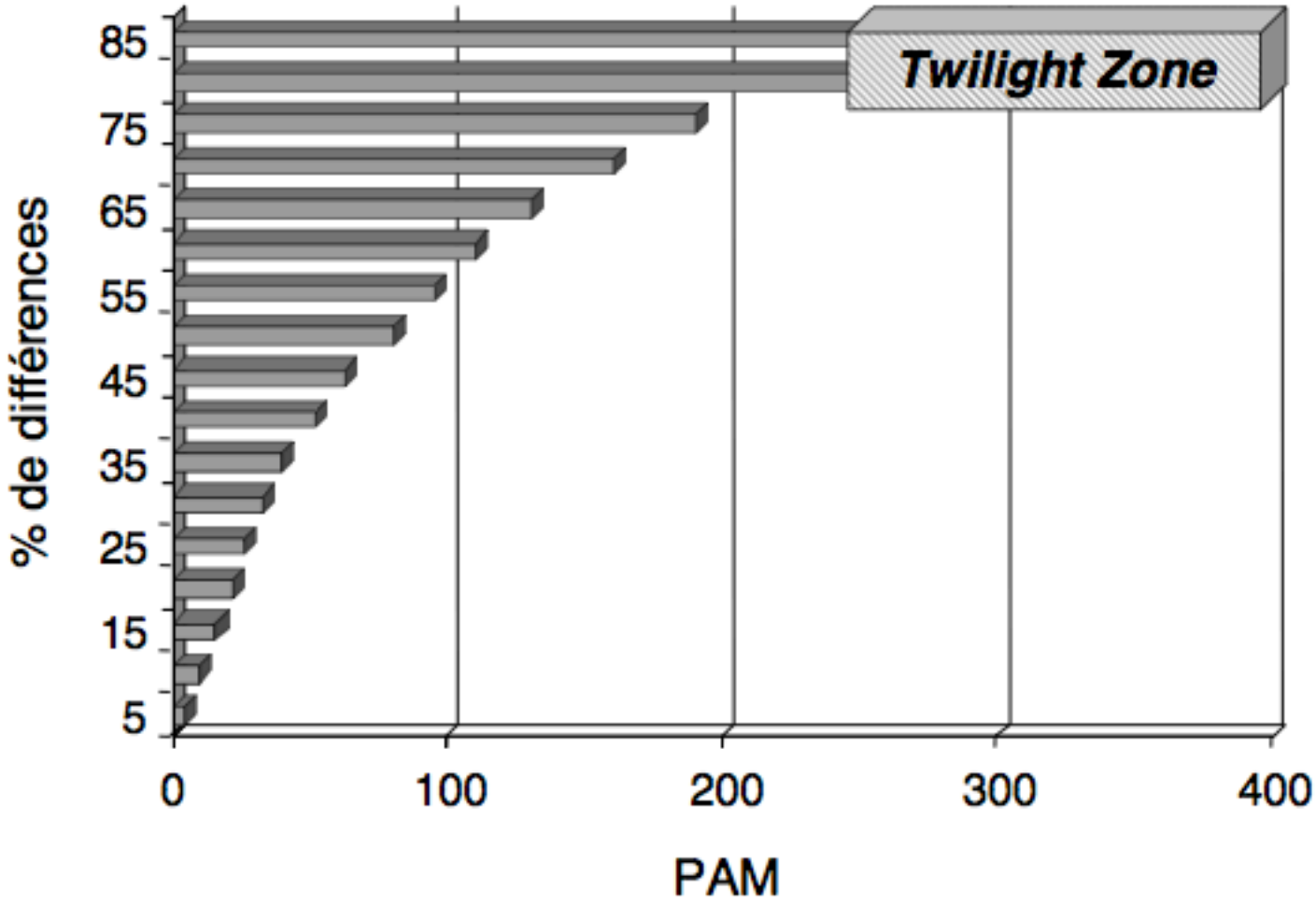
Modèles de substitutions : évolution protéique

		A ala	R arg	N asn	D asp	C cys	Q gln	E glu	G gly	H his	I ile	L leu	K lys	M met	F phe	P pro	S ser	T thr	W trp	Y tyr	V val
A	ala	9867	2	9	10	3	8	17	21	2	6	4	2	6	2	22	35	32	0	2	18
R	arg	1	9913	1	0	1	10	0	0	10	3	1	19	4	1	4	6	1	8	0	1
N	asn	4	1	9822	36	0	4	6	6	21	3	1	13	0	1	2	20	9	1	4	1
D	asp	6	0	42	9859	0	6	53	6	4	1	0	3	0	0	1	5	3	0	0	1
C	cys	1	1	0	0	9973	0	0	0	1	1	0	0	0	0	1	5	1	0	3	2
Q	gln	3	9	4	5	0	9876	27	1	23	1	3	6	4	0	6	2	2	0	0	1
E	glu	10	0	7	56	0	35	9865	4	2	3	1	4	1	0	3	4	2	0	1	2
G	gly	21	1	12	11	1	3	7	9935	1	0	1	2	1	1	3	21	3	0	0	5
H	his	1	8	18	3	1	20	1	0	9912	0	1	1	0	2	3	1	1	1	4	1
I	ile	2	2	3	1	2	1	2	0	0	9872	9	2	12	7	0	1	7	0	1	33
L	leu	3	1	3	0	0	6	1	1	4	22	9947	2	45	13	3	1	3	4	2	15
K	lys	2	37	25	6	0	12	7	2	2	4	1	9926	20	0	3	8	11	0	1	1
M	met	1	1	0	0	0	2	0	0	0	5	8	4	9874	1	0	1	2	0	0	4
F	phe	1	1	1	0	0	0	0	1	2	8	6	0	4	9946	0	2	1	3	28	0
P	pro	13	5	2	1	1	8	3	2	5	1	2	2	1	1	9926	12	4	0	0	2
S	ser	28	11	34	7	11	4	6	16	2	2	1	7	4	3	17	9840	38	5	2	2
T	thr	22	2	13	4	1	3	2	2	1	11	2	8	6	1	5	32	9871	0	2	9
W	trp	0	2	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	9976	1	0
Y	tyr	1	0	3	0	3	0	1	0	4	1	1	0	0	21	0	1	1	2	9945	1
V	val	13	2	1	1	3	2	2	3	3	57	11	1	17	1	3	2	10	0	2	9901

⇒ La matrice PAM 250 est couramment utilisée pour des séquences qui différent de ~ 80%

Modèles de substitutions : évolution protéique

Seuil pour les matrices PAM



Modèles de substitutions : évolution protéique

- **Matrices PAM / JTT / BLOSUM**

PAM (*Point Accepted Mutation*):

✓ 71 familles de gènes nucléaires correspondant à 1300 séquences :

✓ Séquences peu divergentes entre elles(identité ≥ 85% entre chaque paire possible dans une famille).

✓ Alignements globaux.

JTT (*Jones,Taylor and Thornton*):

✓ Construites à partir de 59 190 mutations ponctuelles observées dans 16 300 protéines.

✓ Alignements globaux.

Modèles de substitutions : évolution protéique

- **Matrices PAM / JTT / BLOSUM**

BLOSUM (*Blocks Substitution Matrices*) :

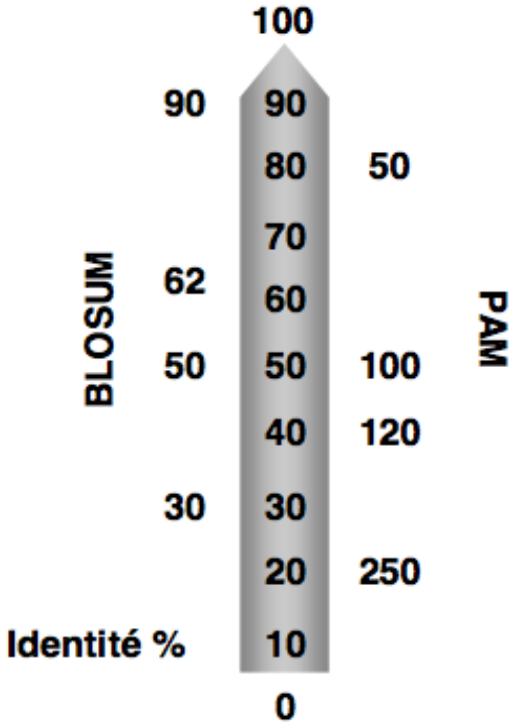
- Utilisation de ~2000 domaines conservés provenant de 500 familles de protéines.
- Comparaisons effectuées dans les domaines alignés (banque BLOCKS).
- Matrices créées à partir de domaines comprenant des séquences ± divergentes :
  - Toutes les paires ayant servi à construire une matrice BLOSUM*k* ont une identité ≥ à *k* %.
  - Matrices plus adaptées pour des protéines distantes du point de vue évolutif.

Modèles de substitutions : évolution protéique

Blosum-80
PAM-1
Faible divergence

Blosum-62
PAM-120

Blosum-45
PAM-250
Forte divergence



% ≠ observé	dist. évolutive PAM
1	1
5	5
10	11
15	17
20	23
25	30
30	38
34	47
40	56
45	67
50	80
55	94
60	112
65	133
70	159
75	195
80	246
85	328

Modèles de substitutions : Choix du modèle

Model name	Base composition	Different transition and transversion rates	All transition rates identical	All transversion rates identical	Reference
JC (JC69)	1:1:1:1	No	Yes	Yes	Jukes and Cantor (1969)
Felsenstein 81 (F81)	Variable	No	Yes	Yes	Felsenstein (1981)
K2P (K80)	1:1:1:1	Yes	Yes	Yes	Kimura (1980)
HKY85	Variable	Yes	No	No	Hasegawa et al. (1985)
Tamura-Nei (TN)	Variable	Yes	No	Yes	Tamura and Nei (1993)
K3P (K81)	Variable	Yes	No	Yes	Kimura (1981)
SYM	1:1:1:1	Yes	No	No	Zharkikh (1994)
REV (GTR)	Variable	Yes	No	No	Rodriguez et al. (1990)

+ Vitesse d'évolution constante ou non entre les sites
+ Nombre de sites invariants

Logiciels Modeltest- Protest

http://darwin.uvigo.es/software/modeltest.html

Modèles de substitutions : choix du modèle

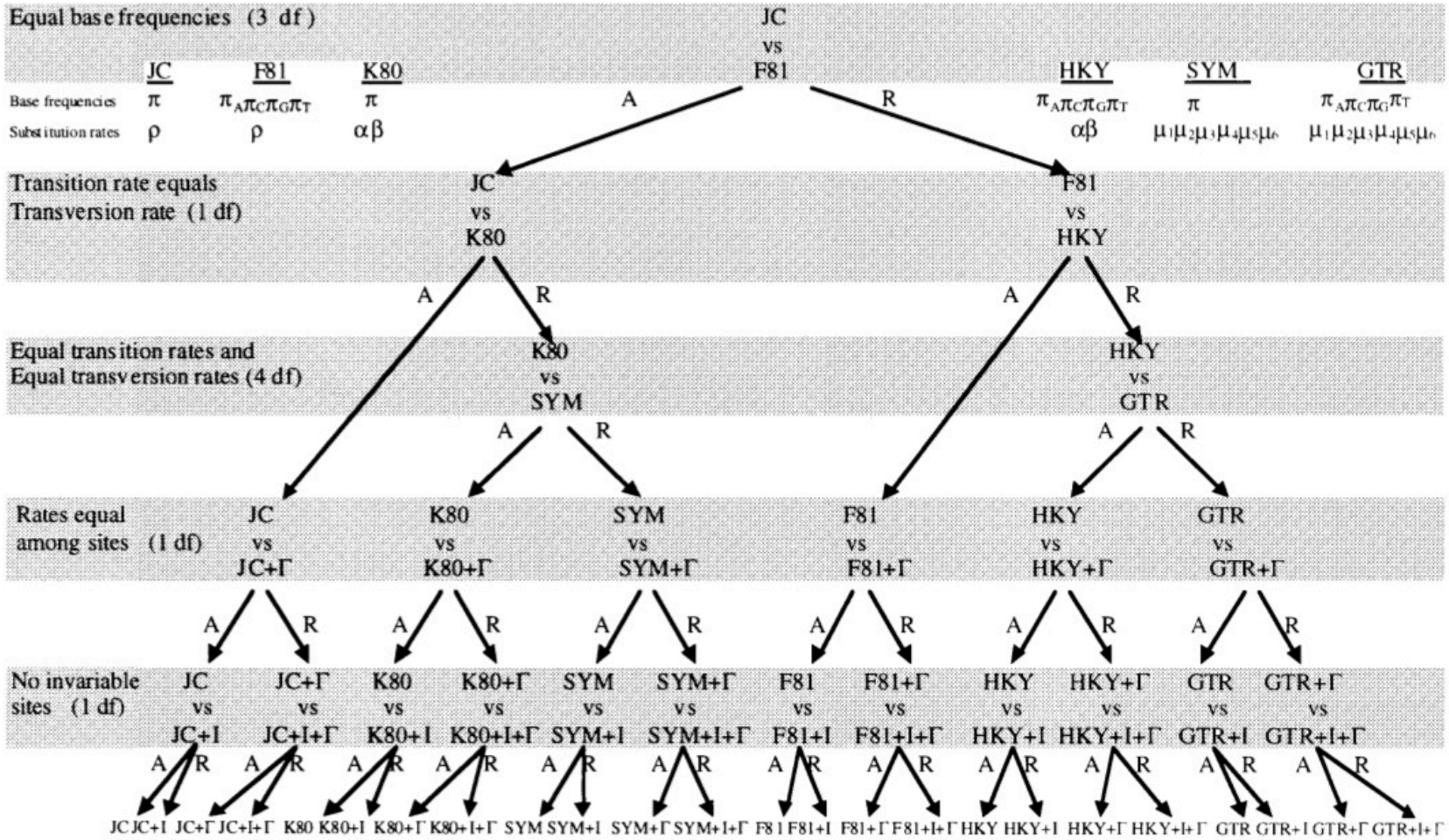
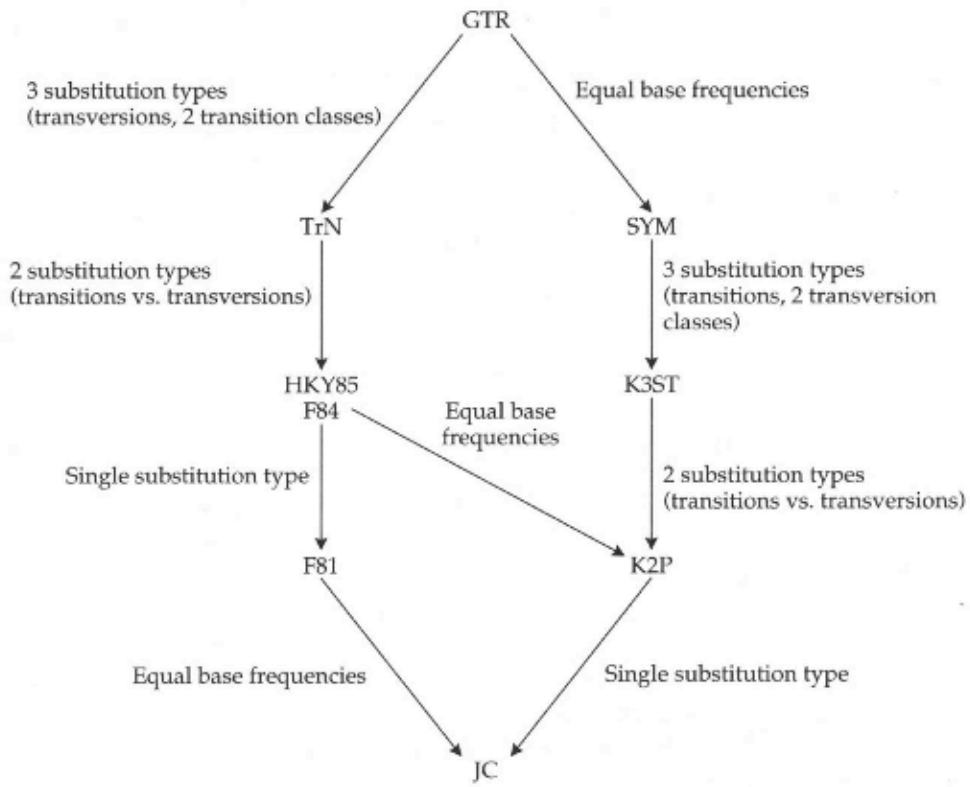


Fig. 1. Hierarchical hypothesis testing in MODELTEST. At each level the null hypothesis (upper model) is either accepted (A) or rejected (R). The models of DNA substitution are: JC (Jukes and Cantor, 1969), K80 (Kimura, 1980), SYM (Zharkikh, 1994), F81 (Felsenstein, 1981), HKY (Hasegawa *et al.*, 1985), and GTR (Rodríguez *et al.*, 1990). Γ: shape parameter of the gamma distribution; I: proportion of invariable sites. df: degrees of freedom. 1: equal base frequencies (0.25), π_A: frequency of adenine, π_C: frequency of cytosine, π_G: frequency of guanine, π_T: frequency of thymine. ρ: equal substitution rate, α: transition rate, β: transversion rate; μ_1: A⇒C rate, μ_2: A⇒G rate, μ_3: A⇒T rate, μ_4: C⇒G rate, μ_5: C⇒T rate, μ_6: G⇒T rate.

Modèles de substitutions : Choix du modèle

Logiciels Modeltest- Protest

http://darwin.uvigo.es/software/modeltest.html



Document : Model selection criteria

Alignements
Arbre
Taille de l'échantillon

hLRT : likelihood ratio test

2(L1-L0) avec L1 et L0 les vraisemblances des deux modèles à comparer

AIC : akaike information criterion

2L+2k avec L log de la vraisemblance et k le nombre de paramètres libres du modèle

BIC : bayesian information criterion

-2L+k log(n) avec n le nombre de sites dans l'échantillon

Ranking of models
« Best » tree
Relative importance of parameters

On préférera le modèle dont le(s) critère(s) à la valeur la plus faible