Détection de la sélection

Formation phylogénie – Institut Pasteur

Guy Perrière

Laboratoire de Biométrie et Biologie Évolutive
UMR CNRS n° 5558
Université Claude Bernard – Lyon 1

1er octobre 2015

Le code génétique

I \ II	T	C	A	G	III
T	TTT Phe F	TCT Ser S	TAT Tyr Y	TGT Cys C	T
	TTC Phe F	TCC Ser S	TAC Tyr Y	TGC Cys C	C
	TTA Leu L	TCA Ser S	TAA Stop	TGA Stop	A
	TTG Leu L	TCG Ser S	TAG Stop	TGG Trp W	G
C	CTT Leu L	CCT Pro P	CAT His H	CGT Arg R	T
	CTC Leu L	CCC Pro P	CAC His H	CGC Arg R	C
	CTA Leu L	CCA Pro P	CAA Gln Q	CGA Arg R	A
	CTG Leu L	CCG Pro P	CAG Gln Q	CGG Arg R	G
A	ATT Ile I	ACT Thr T	AAT Asn N	AGT Ser S	T
	ATC Ile I	ACC Thr T	AAC Asn N	AGC Ser S	C
	ATA Ile I	ACA Thr T	AAA Lys K	AGA Arg R	A
	ATG Met M	ACG Thr T	AAG Lys K	AGG Arg R	G
G	GTT Val V	GCT Ala A	GAT Asp D	GGT Gly G	T
	GTC Val V	GCC Ala A	GAC Asp D	GGC Gly G	C
	GTA Val V	GCA Ala A	GAA Glu E	GGA Gly G	A
	GTG Val V	GCG Ala A	GAG Glu E	GGG Gly G	G

Substitutions synonymes et non synonymes

- Exemple d'une transversion C $\rightarrow$ A dans le codon CTC :
  - En position I :

    $$\mathrm{CTC}^{\mathrm{Leu}} \rightarrow \mathrm{ATC}^{\mathrm{Ile}}$$

    soit une substitution *non synonyme* (ou *non silencieuse*).
  - En position III :

    $$\mathrm{CTC}^{\mathrm{Leu}} \rightarrow \mathrm{CTA}^{\mathrm{Leu}}$$

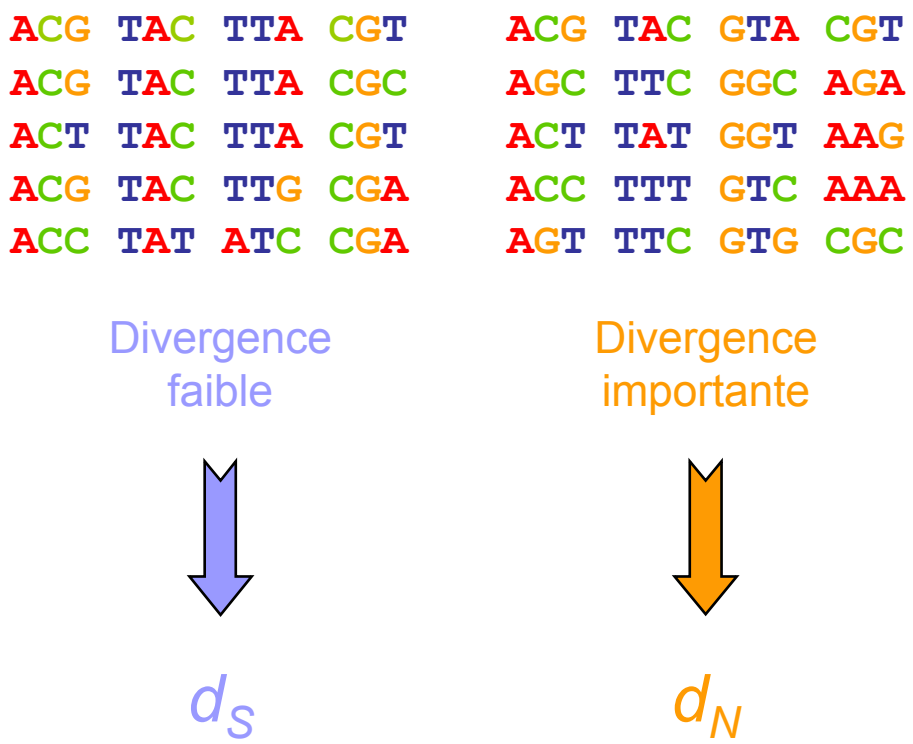
    soit une substitution *synonyme* (ou *silencieuse*).
- Toutes les substitutions touchant la position II des codons sont non synonymes.

Distances d_N et d_S

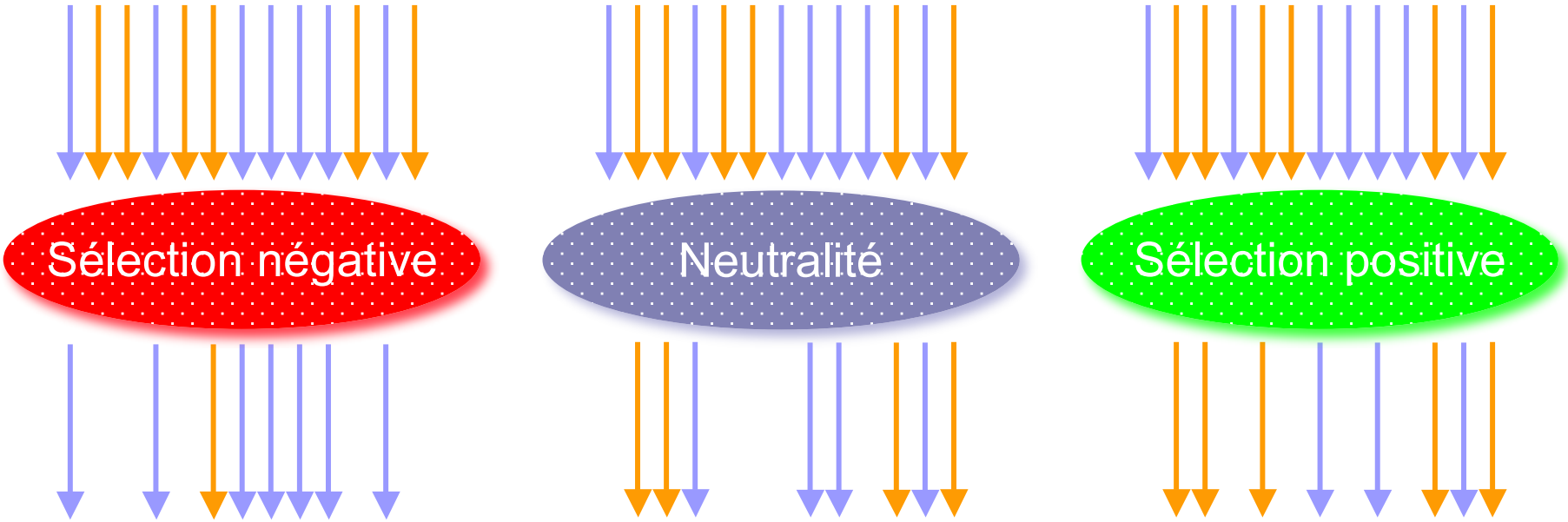
- Dans les gènes protéiques, il existe deux classes de sites ayant des vitesses évolutives différentes :
  - Substitutions non synonymes lentes.
  - Substitutions synonymes rapides.
  - L'hypothèse faite par les modèles d'évolution « classiques » que chaque site évolue en suivant le même processus est fausse.
- Calcul de deux distances évolutives différentes :
  - Distance non synonyme (d_N) :
    - Calcul à partir de p_N = nb. de substitutions non synonymes / nb. de sites non synonymes.
  - Distance synonyme (d_S) :
    - Calcul à partir de p_S = nb. de substitutions synonymes / nb. de sites synonymes.

Utilisation

- On se trouve fréquemment dans l'une ou l'autre de ces deux situations :
  - Séquences évolutivement peu distantes :
    - d_S est informatif, d_N ne l'est pas.
  - Séquences évolutivement très distantes :
    - d_S est saturé, d_N est informatif.



Sélection et neutralité



$d_N/d_S < 1$ $d_N/d_S \simeq 1$ $d_N/d_S > 1$

Substitutions synonymes
Substitutions non synonymes

Principe général

- Proposée par Nei et Gojobori (1986).
- Simplification des premières méthodes développées par Miyata et Yasunaga (1980) et Perler *et al.* (1980).
- Pour chaque paire de séquences A et B alignées sur l'ensemble de leurs m codons il faut :
  1. Compter le nombre de *sites* synonymes et non synonymes contenus dans chacune des séquences.
  2. Compter le nombre de *différences* synonymes et non synonymes entre les deux séquences.
  3. Calculer les p-distances synonymes et non synonymes p_S et p_N.
  4. Calculer les distances corrigées d_S et d_N pour tenir compte des substitutions multiples.

Comptage des sites synonymes et non synonymes

- Soit une séquence A comprenant m codons.
- Soit s_j le nombre de sites synonymes et n_j le nombre de sites non synonymes pour le codon j $(1 \leq j \leq m)$, dans ce cas :

$$s_j = \sum_{i=1}^{3} f_i \quad \text{et} \quad n = 3 - s$$

avec f_i la fraction de changements synonymes à chaque position $i = \{1, 2, 3\}$ du codon j.

Exemple pour le codon AGG

- Position I :

$$\left.\begin{array}{l} \mathrm{A} \rightarrow \mathrm{C} \Rightarrow \mathrm{AGG}^{\mathrm{Arg}} \rightarrow \mathrm{CGG}^{\mathrm{Arg}} \\ \mathrm{A} \rightarrow \mathrm{T} \Rightarrow \mathrm{AGG}^{\mathrm{Arg}} \rightarrow \mathrm{TGG}^{\mathrm{Trp}} \\ \mathrm{A} \rightarrow \mathrm{G} \Rightarrow \mathrm{AGG}^{\mathrm{Arg}} \rightarrow \mathrm{GGG}^{\mathrm{Gly}} \end{array}\right\} \Rightarrow f_1 = 1/3$$

- Position II :

Toutes les substitutions sont non synonymes $\Rightarrow f_2 = 0$

- Position III :

$$\left.\begin{array}{l} \mathrm{G} \rightarrow \mathrm{A} \Rightarrow \mathrm{AGG}^{\mathrm{Arg}} \rightarrow \mathrm{AGA}^{\mathrm{Arg}} \\ \mathrm{G} \rightarrow \mathrm{C} \Rightarrow \mathrm{AGG}^{\mathrm{Arg}} \rightarrow \mathrm{AGC}^{\mathrm{Ser}} \\ \mathrm{G} \rightarrow \mathrm{T} \Rightarrow \mathrm{AGG}^{\mathrm{Arg}} \rightarrow \mathrm{AGT}^{\mathrm{Ser}} \end{array}\right\} \Rightarrow f_3 = 1/3$$

On en déduit $s_j = 1/3 + 0 + 1/3 = 2/3$ et $n = 3 - 2/3 = 7/3$.

Extension à tous les codons

- Nombre total de sites synonymes de A :

$$S_{\mathrm{A}} = \sum_{j=1}^{m} s_j$$

- Nombre total de sites non synonymes de A :

$$N_{\mathrm{A}} = 3m - S_{\mathrm{A}}$$

- Le même calcul est effectué pour B.
- Obtention du nombre sites synonymes et non synonymes en faisant la moyenne des valeurs obtenues :

$$S = (S_{\mathrm{A}} + S_{\mathrm{B}})/2 \quad \text{et} \quad N = (N_{\mathrm{A}} + N_{\mathrm{B}})/2$$

Comptage des différences

- Soit s_{d_j} le nombre de différences synonymes et n_{d_j} le nombre de différences non synonymes entre deux codons alignés de A et B.
- Simple à calculer dans le cas où l'on a une seule substitution :

$$\left.\begin{array}{l} \text{A} : \text{TGT}^{\text{Cys}} \\ \text{B} : \text{TGG}^{\text{Trp}} \end{array}\right\} \Rightarrow s_{d_j} = 0 \text{ et } n_{d_j} = 1$$

$$\left.\begin{array}{l} \text{A} : \text{CCC}^{\text{Pro}} \\ \text{B} : \text{CCT}^{\text{Pro}} \end{array}\right\} \Rightarrow s_{d_j} = 1 \text{ et } n_{d_j} = 0$$

- Cas où l'on a plusieurs substitutions :
  - Détermination de tous les chemins évolutifs possibles :
    - Chemins passant par un codon Stop ignorés.
    - Deux chemins pour deux substitutions.
    - Six chemins pour trois substitutions.
    - Utilisation des valeurs moyennes de s_{d_j} et n_{d_j}.

Exemple avec deux substitutions

- Alignement de $\mathrm{AAA}^{\mathrm{Lys}}$ avec $\mathrm{ATG}^{\mathrm{Met}}$:

Chemin	s_{d_j}	n_{d_j}
$\mathrm{AAA}^{\mathrm{Lys}} \leftrightarrow \mathrm{AAG}^{\mathrm{Lys}} \leftrightarrow \mathrm{ATG}^{\mathrm{Met}}$	1	1
$\mathrm{AAA}^{\mathrm{Lys}} \leftrightarrow \mathrm{ATA}^{\mathrm{Ile}} \leftrightarrow \mathrm{ATG}^{\mathrm{Met}}$	0	2
Moyenne	0.5	1.5

- Intuitivement, le premier chemin apparaît plus probable que le deuxième :
  - Mise en place d'une pondération en faveur des substitutions synonymes dans les premières versions de la méthode :
    - Biais en cas de sélection positive !
  - Absence de pondération :
    - Fixation indirecte de ω vers 1.
    - Biais en cas d'écart à la neutralité !

Calcul des distances évolutives

- Soit S_d le nombre total de différences synonymes et N_d le nombre total de différences non synonymes entre A et B :

$$S_\mathrm{d} = \sum_{j=1}^{m} s_{\mathrm{d}_j} \quad \text{et} \quad N_\mathrm{d} = \sum_{j=1}^{m} n_{\mathrm{d}_j}$$

- Les divergences observées synonymes p_S et non synonymes p_N entre les séquences A et B sont définies comme :

$$p_\mathrm{S} = S_\mathrm{d}/S \quad \text{et} \quad p_\mathrm{N} = N_\mathrm{d}/N$$

- Application du modèle JC aux deux divergences observées :

$$d_\mathrm{S} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p_\mathrm{S}\right) \quad \text{et} \quad d_\mathrm{N} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}p_\mathrm{N}\right)$$

Limitations de la méthode

- Fait l'hypothèse que les substitutions se produisent aléatoirement à chaque site avec la même probabilité :
  - Exemple de la position III du codon AGT :

$$\left.\begin{array}{l} \mathrm{T} \rightarrow \mathrm{A} \Rightarrow \mathrm{AGT}^{\mathrm{Ser}} \rightarrow \mathrm{AGA}^{\mathrm{Arg}} \\ \mathrm{T} \rightarrow \mathrm{C} \Rightarrow \mathrm{AGT}^{\mathrm{Ser}} \rightarrow \mathrm{AGC}^{\mathrm{Ser}} \\ \mathrm{T} \rightarrow \mathrm{G} \Rightarrow \mathrm{AGT}^{\mathrm{Ser}} \rightarrow \mathrm{AGG}^{\mathrm{Arg}} \end{array}\right\} \Rightarrow f_3 = 1/3$$

- Or les transitions sont généralement plus fréquentes que les transversions :
  - Substitution $\mathrm{AGT}^{\mathrm{Ser}} \rightarrow \mathrm{AGC}^{\mathrm{Ser}}$ plus fréquente.
- De plus, les transitions en position III sont presque toutes synonymes, ce qui n'est pas le cas des transversions :
  - Sous-estimation de certaines valeurs de s_j et donc de S.
  - Surestimation de p_{S} et d_{S} par rapport à p_{N} et d_{N}.

Prise en compte du biais

- Modification de la méthode de Nei et Gojobori (1986) proposée par Ina (1995) :
  - Différenciation entre les transitions et les transversions.
- Utilisation du modèle K2P pour l'estimation de S et de N :
  - Utilisation des écheangeabilités pour les transitions (α) et pour les transversion (β).

Exemple pour le codon AGT

- Position I :

$$\left.\begin{array}{l} \mathrm{A} \rightarrow \mathrm{C} \Rightarrow \mathrm{AGT}^{\mathrm{Ser}} \rightarrow \mathrm{CGT}^{\mathrm{Arg}} \\ \mathrm{A} \rightarrow \mathrm{G} \Rightarrow \mathrm{AGT}^{\mathrm{Ser}} \rightarrow \mathrm{GGT}^{\mathrm{Gly}} \\ \mathrm{A} \rightarrow \mathrm{T} \Rightarrow \mathrm{AGT}^{\mathrm{Ser}} \rightarrow \mathrm{TGT}^{\mathrm{Cys}} \end{array}\right\} \Rightarrow f_1 = 0$$

- Position II :

Toutes les substitutions sont non synonymes $\Rightarrow f_2 = 0$

- Position III :

$$\left.\begin{array}{l} \mathrm{T} \rightarrow \mathrm{A} \Rightarrow \mathrm{AGT}^{\mathrm{Ser}} \rightarrow \mathrm{AGA}^{\mathrm{Arg}} \\ \mathrm{T} \rightarrow \mathrm{C} \Rightarrow \mathrm{AGT}^{\mathrm{Ser}} \rightarrow \mathrm{AGC}^{\mathrm{Ser}} \\ \mathrm{T} \rightarrow \mathrm{G} \Rightarrow \mathrm{AGT}^{\mathrm{Ser}} \rightarrow \mathrm{AGG}^{\mathrm{Arg}} \end{array}\right\} \Rightarrow f_3 = \frac{\alpha}{\alpha + 2\beta}$$

on en déduit $s = 0 + 0 + \dfrac{\alpha}{\alpha + 2\beta} = \dfrac{\alpha/\beta}{2 + \alpha/\beta}$

Exemple pour le codon CTA

- Position I :

$$\left.\begin{array}{l} \mathrm{C} \to \mathrm{A} \Rightarrow \mathrm{CTA}^{\mathrm{Leu}} \to \mathrm{ATA}^{\mathrm{Ile}} \\ \mathrm{C} \to \mathrm{G} \Rightarrow \mathrm{CTA}^{\mathrm{Leu}} \to \mathrm{GTA}^{\mathrm{Val}} \\ \mathrm{C} \to \mathrm{T} \Rightarrow \mathrm{CTA}^{\mathrm{Leu}} \to \mathrm{TTA}^{\mathrm{Leu}} \end{array}\right\} \Rightarrow f_1 = \frac{\alpha}{\alpha + 2\beta}$$

- Position II :

$$\text{Toutes les substitutions sont non synonymes} \Rightarrow f_2 = 0$$

- Position III :

$$\left.\begin{array}{l} \mathrm{A} \to \mathrm{C} \Rightarrow \mathrm{CTA}^{\mathrm{Leu}} \to \mathrm{CTC}^{\mathrm{Leu}} \\ \mathrm{A} \to \mathrm{G} \Rightarrow \mathrm{CTA}^{\mathrm{Leu}} \to \mathrm{CTG}^{\mathrm{Leu}} \\ \mathrm{A} \to \mathrm{T} \Rightarrow \mathrm{CTA}^{\mathrm{Leu}} \to \mathrm{CTT}^{\mathrm{Leu}} \end{array}\right\} \Rightarrow f_3 = 1$$

$$\text{on en déduit } s = \frac{\alpha}{\alpha + 2\beta} + 0 + 1 = \frac{\alpha/\beta}{2 + \alpha/\beta} + 1$$

Calcul des valeurs de S et N

- Le nombre de sites synonymes s dépend du rapport $\kappa = \alpha/\beta$ ou ratio transitions/transversions :
  - Estimation au moyen de la formule de Kimura (1980) :

$$\kappa = \frac{\alpha}{\beta} = \frac{2\ln(1-2r-v)}{\ln(1-2v)} - 1$$

- Calcul des valeurs de S et N en utilisant les mêmes équations que celles de la Diapo. 10.

Comptage des différences

- Différences synonymes :
  - Nombre de différences synonymes liées à une transition (s_r) ou à une transversion (s_v).
- Différences non synonymes :
  - Nombre de différences non synonymes liées à une transition (n_r) ou à une transversion (n_v).
- Exemples à une seule substitution :

$$\left.\begin{array}{l} \text{A : TGT}^{\text{Cys}} \\ \text{B : TGG}^{\text{Trp}} \end{array}\right\} \Rightarrow s_r = s_v = n_r = 0 \text{ et } n_v = 1$$

$$\left.\begin{array}{l} \text{A : CCC}^{\text{Pro}} \\ \text{B : CCT}^{\text{Pro}} \end{array}\right\} \Rightarrow s_r = 1 \text{ et } s_v = n_r = n_v = 0$$

Calcul des divergences observées

- Soit S_{r} le nombre total de différences synonymes liées à une transition et S_{v} le nombre total de différences synonymes liées à une transversion :

$$S_{\mathrm{r}} = \sum_{j=1}^{m} s_{\mathrm{r}_j} \quad \text{et} \quad S_{\mathrm{v}} = \sum_{j=1}^{m} s_{\mathrm{v}_j}$$

- Soit N_{r} le nombre total de différences non synonymes liées à une transition et N_{v} le nombre total de différences non synonymes liées à une transversion :

$$N_{\mathrm{r}} = \sum_{j=1}^{m} n_{\mathrm{r}_j} \quad \text{et} \quad N_{\mathrm{v}} = \sum_{j=1}^{m} n_{\mathrm{v}_j}$$

Calcul des distances évolutives

- Les divergences observées synonymes (r_S) et non synonymes (r_N) liées aux transitions sont définies comme :

$$r_S = S_r/S \text{ et } r_N = N_r/N$$

- Les divergences observées synonymes (v_S) et non synonymes (v_N) liées aux transversions sont définies comme :

$$v_S = S_v/S \text{ et } v_N = N_v/N$$

- Enfin, les distances synonyme et non synonyme sont données par :

$$d_S = -\frac{1}{2}\ln(1 - 2r_S - v_S) - \frac{1}{4}\ln(1 - 2v_S)$$

$$d_N = -\frac{1}{2}\ln(1 - 2r_N - v_N) - \frac{1}{4}\ln(1 - 2v_N)$$

Limitations de la méthode

- Peu efficace lorsque le biais de transition est faible :
  - Surestimation de ω quand la valeur réelle est < 1 et sous-estimation quand elle est > 1.
- Pas de pondération pour les différents chemins évolutifs pour passer d'un codon à un autre :
  - Mêmes problèmes qu'avec la méthode Nei et Gojobori.
- Surestimation de ω lorsque les fréquences des nucléotides sont différentes entre les séquences.

Principe général

- Utilise la dégénérescence du code génétique et le modèle K2P.
- Fondée sur le calcul de trois valeurs :
  - L_0, le nombre moyen de sites non dégénérés :
    - Sites pour lesquels aucune substitution n'est synonyme.
  - L_2, le nombre moyen de sites dégénérés deux fois :
    - Sites pour lesquels une substitution est synonyme et les deux autres non synonymes.
  - L_4, le nombre moyen de sites dégénérés quatre fois :
    - Sites pour lesquels toutes les substitutions sont synonymes.
  - Il existe des sites dégénérés trois fois :
    - Classement parmi les sites dégénérés deux fois.

Exemples sur des codons

- Sites non dégénérés :
  - Position I des codons NTT :
    - $\mathrm{ATT}^{\mathrm{Ile}}$ | $\mathrm{CTT}^{\mathrm{Leu}}$ | $\mathrm{GTT}^{\mathrm{Val}}$ | $\mathrm{TTT}^{\mathrm{Phe}}$
- Sites dégénérés deux fois :
  - Position I des codons YTA :
    - $\mathrm{CTA}^{\mathrm{Leu}}$ | $\mathrm{TTA}^{\mathrm{Leu}}$ | $\mathrm{ATA}^{\mathrm{Ile}}$ | $\mathrm{GTA}^{\mathrm{Val}}$
- Sites dégénérés trois fois :
  - Position III des codons ATH :
    - $\mathrm{ATA}^{\mathrm{Ile}}$ | $\mathrm{ATC}^{\mathrm{Ile}}$ | $\mathrm{ATT}^{\mathrm{Ile}}$ | $\mathrm{ATG}^{\mathrm{Met}}$
- Sites dégénérés quatre fois :
  - Position III des codons CCN :
    - $\mathrm{CCA}^{\mathrm{Pro}}$ | $\mathrm{CCC}^{\mathrm{Pro}}$ | $\mathrm{CCG}^{\mathrm{Pro}}$ | $\mathrm{CCT}^{\mathrm{Pro}}$

Exemple pour deux séquences alignées

- Soit deux séquences alignées A et B :

```
A: AAA GCG
B: ACA GCC
   123 456
```

- Séquence A :
  - Les nucléotides 1, 2, 4 et 5 sont dégénérés zéro fois $\Rightarrow L_0^{(A)} = 4$
  - Le nucléotide 3 est dégénéré deux fois $\Rightarrow L_2^{(A)} = 1$
  - Le nucléotide 6 est dégénéré quatre fois $\Rightarrow L_4^{(A)} = 1$
- Séquence B :
  - Les nucléotides 1, 2, 4 et 5 sont dégénérés zéro fois $\Rightarrow L_0^{(B)} = 4$
  - Aucun nucléotide n'est dégénéré deux fois $\Rightarrow L_2^{(B)} = 0$
  - Les nucléotides 3 et 6 sont dégénérés quatre fois $\Rightarrow L_4^{(B)} = 2$
- Pour l'alignement, utilisation des valeurs moyennes :
  - $L_0 = (4+4)/2 = 4$, $L_2 = (1+0)/2 = 0.5$ et $L_4 = (1+2)/2 = 1.5$

Calcul de S et N

- Toutes les substitutions se produisant au niveau de sites non dégénérés entrainent un changement d'acide aminé :
  - Sites non synonymes.
- Aucune substitution se produisant au niveau de sites dégénérés quatre fois ne provoque de changement d'acide aminé :
  - Sites synonymes.
- Les transversions se produisant au niveau de sites dégénérés deux fois sont non synonymes alors que les transitions sont synonymes :
  - 1/3 de changements synonymes et 2/3 de changements non synonymes.
- Sous ces hypothèses :

$$S = \frac{1}{3}L_2 + L_4 \quad \text{et} \quad N = \frac{2}{3}L_2 + L_0$$

Estimation de S_d et N_d

- Soit r_i et v_i ($i \in \{0, 2, 4\}$), les fréquences des transitions et des transversions observées pour chacune des trois catégories de sites.
- Exemple de l'alignement précédent :

```
A: AAA GCG
B: ACA GCC
```

  - AAA et ACA diffèrent par une transversion s'étant produite à un site dégénéré zéro fois.
  - GCG et GCC diffèrent par une transversion à un site dégénéré quatre fois :
    - $v_0 = 1/L_0 = 1/4$, $v_4 = 1/L_4 = 1/1.5$, $v_2 = r_0 = r_2 = r_4 = 0$
- Lorsque des codons diffèrent par plus d'une position :
  - Exploration de tous les chemins possibles comme pour la méthode de Nei et Gojobori.

Exceptions

- Il existe deux exceptions à la règle liant le niveau de dégénérescence d'un site et le type de substitution s'y produisant :
  - Position III des codons $\mathrm{ATH}^{\mathrm{Ile}}$:
    - Considérée comme un site dégénéré deux fois bien que la transition $\mathrm{ATA}^{\mathrm{Ile}} \rightarrow \mathrm{ATG}^{\mathrm{Met}}$ conduise à une substitution non synonyme.
  - Transversion C $\rightarrow$ A en position I des codons $\mathrm{CGR}^{\mathrm{Arg}}$ et transversion A $\rightarrow$ C en position I des codons $\mathrm{AGR}^{\mathrm{Arg}}$:
    - Substitutions synonymes alors que tous les autres changements en position I pour ces deux codons sont non synonymes.
    - Contournement en considérant que les changements synonymes sont des transitions et les changements non synonymes des transversions.

Estimation de R_i et V_i

- Soit R_i et V_i ($i \in \{0, 2, 4\}$) les distances évolutives en termes de transitions et de transversions.
- Estimation en appliquant le modèle K2P aux valeurs de r_i et v_i :

$$R_i = \frac{1}{2} \ln \left(\frac{1}{1 - 2r_i - v_i} \right) - \frac{1}{4} \ln \left(\frac{1}{1 - 2v_i} \right)$$

$$V_i = \frac{1}{2} \ln \left(\frac{1}{1 - 2v_i} \right)$$

Fréquences de substitutions par type de site

- Connaissant R_i et V_i, il est possible de déduire les fréquences de substitutions synonymes et non synonymes par type de site :
  - Sites dégénérés zéro fois :
    - Fréquence de substitutions synonymes égale à 0.
    - Fréquences des substitutions non synonymes égale à $R_0 + V_0$.
  - Sites dégénérés deux fois :
    - Fréquence des substitutions synonymes égale à R_2.
    - Fréquence des substitutions non synonymes égale à V_2.
  - Sites dégénérés quatre fois :
    - Fréquence des substitutions synonymes égale à $R_4 + V_4$.
    - Fréquence des substitutions non synonymes égale à 0.

Calcul des distances évolutives

- Calcul S_d et N_d en multipliant les fréquences des substitutions par le nombre de sites de chaque catégorie :

$$S_\mathrm{d} = L_2 R_2 + L_4(R_4 + V_4) \quad \text{et} \quad N_\mathrm{d} = L_2 V_2 + L_0(R_0 + V_0)$$

- Calcul de d_S et d_N :

$$d_\mathrm{S} = \frac{S_\mathrm{d}}{S} = \frac{L_2 R_2 + L_4(R_4 + V_4)}{2L_2/3 + L_4}$$

$$d_\mathrm{N} = \frac{N_\mathrm{d}}{N} = \frac{L_2 V_2 + L_0(R_0 + V_0)}{L_2/3 + L_0}$$

Problème dans le calcul de S et N

- Comptage des sites dégénérés deux fois comme étant dans 1/3 des cas synonymes et dans 2/3 des cas non synonymes :
  - Ne prend pas en compte le fait que les transitions se produisent plus fréquemment que les transversions :
    - Sous-estimation de $S \Rightarrow$ surestimation de d_S.
  - Amélioration de Pamilo et Bianchi (1993) :
    - Proportions de substitutions synonymes et non synonymes au niveau des sites dégénérés deux fois = celles se produisant aux sites dégénérés quatre fois.
    - Partage des sites synonymes et non synonymes en fonction du ratio R_4/V_4, plutôt qu'en fonction du ratio 1/2.

Calcul des distances évolutives

- La fréquence moyenne du taux de transition aux sites dégénérés deux et quatre fois est égale à $(L_2 R_2 + L_4 R_4)/(L_2 + L_4)$.
- Calcul de d_S et d_N :

$$d_S = \frac{L_2 R_2 + L_4 R_4}{L_2 + L_4} + V_4$$

$$d_N = R_0 + \frac{L_0 V_0 + L_2 V_2}{L_2 + L_0}$$

- Plusieurs autres améliorations de la méthode ont été proposées depuis la publication originale.

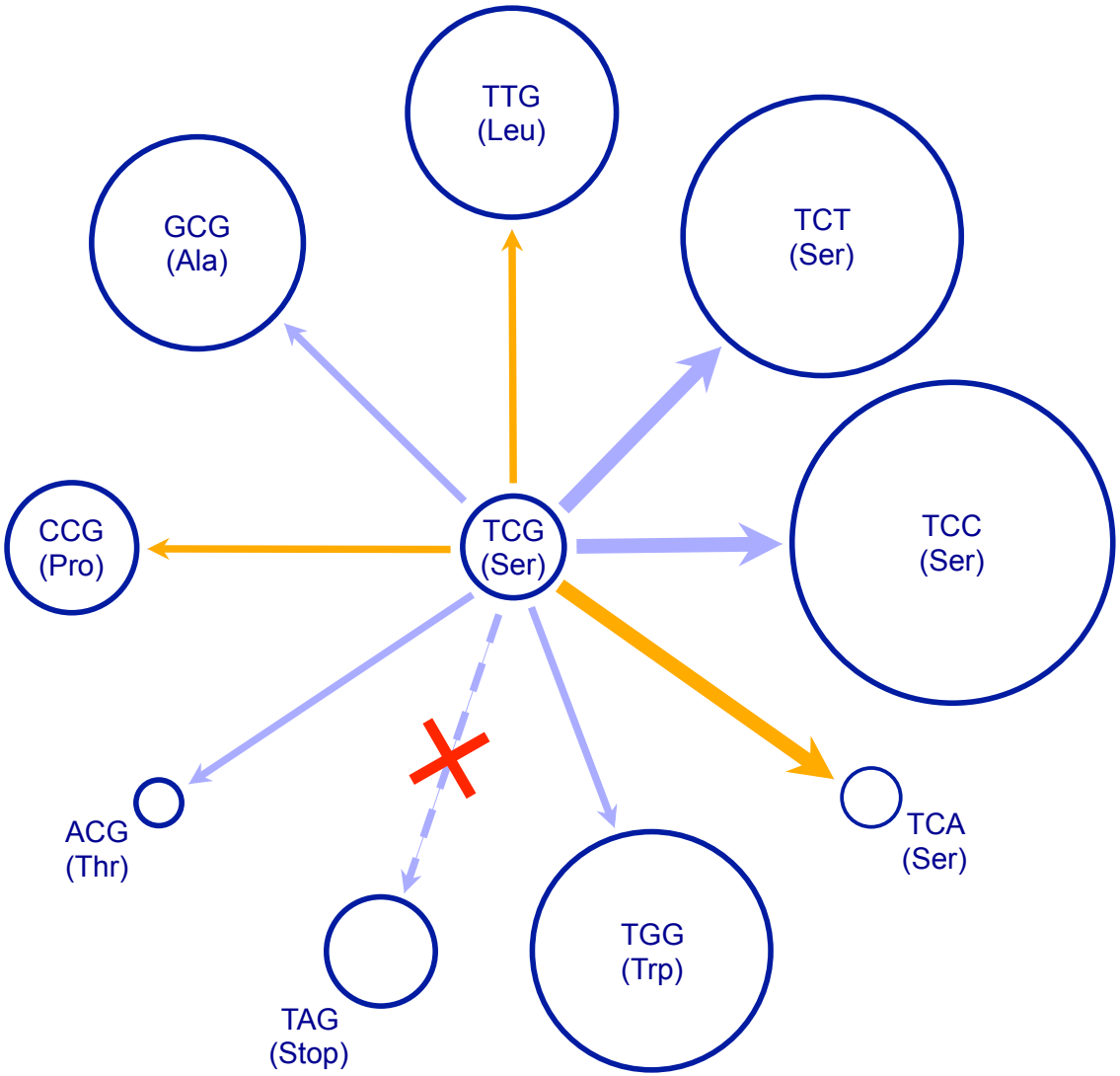
Introduction de ces méthodes

- Justification par le fait que la vraisemblance des valeurs des paramètres des modèles utilisés peut être testée :
  - Il est possible de les optimiser en cherchant les valeurs pour lesquelles leur vraisemblance est maximale.
- Publication la même année de deux méthodes similaires :
  - Goldman et Yang (1994).
  - Muse et Gaut (1994).
- Définition d'un modèle de substitution spécifique aux codons.

Hypothèses du modèle

- Indépendance des sites à l'intérieur d'un codon.
- À chaque intervalle de temps dt, une seule des trois positions est susceptible de muter.
- À chaque instant t, tout codon est susceptible de se substituer vers l'un de ses voisins :
  - Un voisin est un codon différant du codon d'intérêt par une seule substitution :
    - Position I, II ou III.
    - Codons Stop non considérés.
  - Chaque codon possède au plus neuf voisins.

Exemple du codon TCG^{Ser}



Paramètres du modèle

- Matrice $\mathbf{Q} = (q_{ij})$ des taux instantanés ($i, j \in \{\text{AAA, AAC, AAG, } \ldots, \text{TTT}\}$) :

$$\mathbf{Q} = \begin{pmatrix} -\lambda_{\text{AAA}} & q_{\text{AAA,AAC}} & \cdots & q_{\text{AAA,TTT}} \\ q_{\text{AAC,AAA}} & -\lambda_{\text{AAC}} & \cdots & q_{\text{AAC,TTT}} \\ \vdots & \vdots & \ddots & \vdots \\ q_{\text{TTT,AAA}} & q_{\text{TTT,AAC}} & \cdots & -\lambda_{\text{TTT}} \end{pmatrix}$$

  Codons Stop exclus de la matrice.
- Temps t (= longueur des branches).
- Ratio des taux de transitions/transversions κ.
- Ratio des distances non synonymes/synonymes ω.

Taux de substitutions instantanés

- Les valeurs de q_{ij} sont telles que :

$$q_{ij} = \begin{cases} 0, & \text{si } i \text{ et } j \text{ diffèrent en plus d'une position} \\ \pi_j, & \text{pour une transversion synonyme} \\ \kappa\pi_j, & \text{pour une transition synonyme} \\ \omega\pi_j, & \text{pour une transversion non synonyme} \\ \omega\kappa\pi_j, & \text{pour une transition non synonyme} \end{cases}$$

avec normalisation de telle façon que le taux moyen de substitutions soit égal à un :

$$-\sum_i \pi_i \, q_{ij} = 1$$

Probabilités de transition

- Comme pour les modèles standards, les valeurs des probabilités de transition $p_{ij}(t)$ sont données en résolvant $\mathbf{P}(t) = \mathrm{e}^{\mathbf{Q}t}$.
- À partir des valeurs de $p_{ij}(t)$, calcul des valeurs de $f_{ij}(t)$:

$$f_{ij}(t) = \pi_i p_{ij}(t)$$

soit la probabilité d'observer le codon i de la séquence A aligné avec le codon j de la séquence B.

Fonction de vraisemblance

- La fonction de vraisemblance est définie par :

$$\ln L(t, \kappa, \omega) = \sum_{i,j} n_{ij} \ln[f_{ij}(t)]$$
$$= \sum_{i,j} n_{ij} \ln[\pi_i p_{ij}(t)]$$

  avec n_{ij} le nombre de fois où un codon i est aligné avec un codon j.
- Valeurs de π_i :
  - Uniforme ($\pi_i = 1/61, \forall i$).
  - À partir des fréquences des 61 codons dans le jeu de données (F61).
  - À partir des fréquences des nucléotides, toutes positions confondues (F1×4).
  - À partir des fréquences des nucléotides à chacune des trois positions des codons (F3×4).

Proportion de substitutions

- Proportion de substitutions synonymes ρ_S par codon :

$$\rho_S = \sum_i \sum_{j \neq i} \pi_i \, q_{ij}$$

pour les codons i et j correspondants à des acides aminés identiques.

- Proportion de substitutions non synonymes ρ_N par codon :

$$\rho_N = \sum_i \sum_{j \neq i} \pi_i \, q_{ij}$$

pour les codons i et j correspondants à des acides aminés différents.

Calcul de S_d et N_d

- Hypothèse du modèle qu'une seule substitution par codon se produit par unité de temps :

$$\rho_\mathrm{N} + \rho_\mathrm{S} = 1$$

- On en déduit les valeurs du nombre total de différences synonymes et non synonymes comme étant égales à :

$$S_\mathrm{d} = \rho_\mathrm{S} t \quad \text{et} \quad N_\mathrm{d} = \rho_\mathrm{N} t$$

avec $N_\mathrm{d} + S_\mathrm{d} = t$.

Calcul des distances évolutives

- Soit ρ_{S}^1 et ρ_{N}^1 les proportions de *sites* synonymes et non synonymes par codon :
  - Proportion de « mutations » synonymes et non synonymes avant le filtre de la sélection naturelle au niveau des acides aminés.
  - Calcul de la même façon que ρ_S^1 et ρ_N^1 par estimation au maximum de vraisemblance en fixant $\omega = 1$.
- Nombre de sites synonymes et non synonymes par codon :

$$S = 3\rho_{\mathrm{S}}^1 \quad \text{et} \quad N = 3\rho_{\mathrm{N}}^1$$

- Pour finir, les distances d_{S} et d_{N} sont données par :

$$d_{\mathrm{S}} = S_{\mathrm{d}}/S = \rho_{\mathrm{S}}\, t/3\rho_{\mathrm{S}}^1$$
$$d_{\mathrm{N}} = N_{\mathrm{d}}/N = \rho_{\mathrm{N}}\, t/3\rho_{\mathrm{N}}^1$$

Modélisation de l'hétérogénéité

- Dans une phylogénie, certaines lignées peuvent être soumises à de la sélection et d'autres non.
- Utilisation de plusieurs modèles afin de pouvoir détecter ces phénomènes :
  - Même valeur de ω pour toutes les branches de l'arbre (homogénéité).
  - Autant de valeurs de ω qu'il existe de branches dans l'arbre (hétérogénéité maximale).
  - Plusieurs intermédiaires entre ces deux extrêmes.
- Comparaison des différents modèles au moyen de tests LRT afin de déterminer quel est le scénario le plus vraisemblable.

Sélection site spécifique

- Les contraintes fonctionnelles peuvent varier le long d'un gène :
  - Un site donné peut être sous sélection positive alors que le reste de la séquence évolue de façon neutre.
  - Les méthodes précédentes ne sont pas adaptées à détecter ce genre de phénomènes :
    - Considèrent que l'ensemble de la séquence est soumise à la même pression de sélection.
- Développement de méthodes permettant de détecter les pressions à l'échelle d'un site :
  - Fondées sur l'utilisation d'alignements multiples plutôt que de comparaisons deux à deux.
  - Typologie :
    - Méthodes de comptage.
    - Méthodes à effet aléatoire.
    - Méthodes à effets fixe.

Codes IUPAC pour les nucléotides

Code	Signification	Compl.
A	A	T/U
C	C	G
G	G	C
T/U	T/U	A
M	A/C	K
R	A/G	Y
S	C/G	S
W	A/T/U	W
Y	C/T/U	R
K	G/T/U	M
V	A/C/G	B
H	A/C/T/U	D
D	A/G/T/U	H
B	C/G/T/U	V
N	A/C/G/T/U	N

Extrait de la table du χ^2

d.d.l./α	**0.9**	**0.5**	**0.3**	**0.2**	**0.1**	**0.05**	**0.02**	**0.01**	**0.001**
1	0.016	0.455	1.074	1.642	2.706	3.841	5.412	6.635	10.827
2	0.211	1.386	2.408	3.219	4.605	5.991	7.824	9.21	13.815
3	0.584	2.366	3.665	4.642	6.251	7.815	9.837	11.345	16.266
4	1.064	3.357	4.878	5.989	7.779	9.488	11.668	13.277	18.467
5	1.61	4.351	6.064	7.289	9.236	11.07	13.388	15.086	20.515
6	2.204	5.348	7.231	8.558	10.645	12.592	15.033	16.812	22.457
7	2.833	6.346	8.383	9.803	12.017	14.067	16.622	18.475	24.322
8	3.49	7.344	9.524	11.03	13.362	15.507	18.168	20.09	26.125
9	4.168	8.343	10.656	12.242	14.684	16.919	19.679	21.666	27.877
10	4.865	9.342	11.781	13.442	15.987	18.307	21.161	23.209	29.588
11	5.578	10.341	12.899	14.631	17.275	19.675	22.618	24.725	31.264
12	6.304	11.34	14.011	15.812	18.549	21.026	24.054	26.217	32.909
13	7.042	12.34	15.119	16.985	19.812	22.362	25.472	27.688	34.528
14	7.79	13.339	16.222	18.151	21.064	23.685	26.873	29.141	36.123
15	8.547	14.339	17.322	19.311	22.307	24.996	28.259	30.578	37.697
16	9.312	15.338	18.418	20.465	23.542	26.296	29.633	32	39.252
17	10.085	16.338	19.511	21.615	24.769	27.587	30.995	33.409	40.79