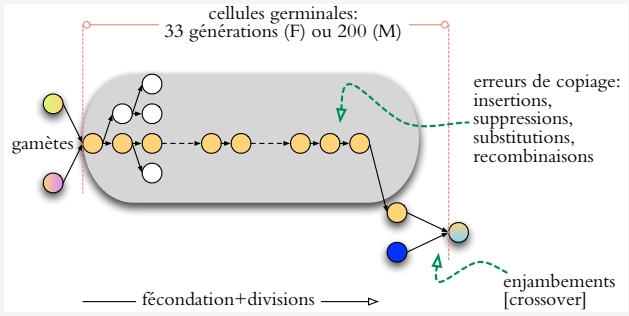
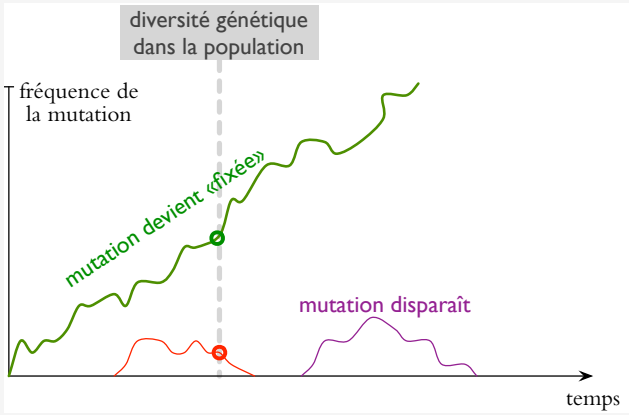
6 Modèles temporaires

6.1 Descendance avec modification



Les mutations lors d'héritage du matériel génétique entre générations introduisent de nouveaux allèles dans une population. Un allèle du parent (p.e., mutation nouvelle dans la lignée de cellules germinales) peut être hérité par l'enfant.

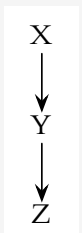


Ultimement, l'allèle devient fixé dans la population entière, ou disparaît complètement. La probabilité de fixation dépend (a) de la taille de la population, et (b) l'effet de l'allèle (avantage ou désavantage possible). Les mutations fixées s'accumulent → divergence de séquences d'origine commune. La divergence croît avec le nombre de générations depuis l'ancêtre.

Substitutions multiples. Évolution de caractères homologues est sans mémoire : états successifs forment une chaîne de Markov. Probabilités de substitution selon la matrice $\mathbf{M}_{X\to Y}$ où

$$\mathbf{M}_{X\to Y}[a,b] = \mathbb{P}\Big\{Y = b \;\Big|\; X = a\Big\}$$

avec valeurs $a, b \in \Sigma = \{1, \ldots, r\}$. (Exemple : $\Sigma = \{1,2,3,4\}$ encodant ADN)

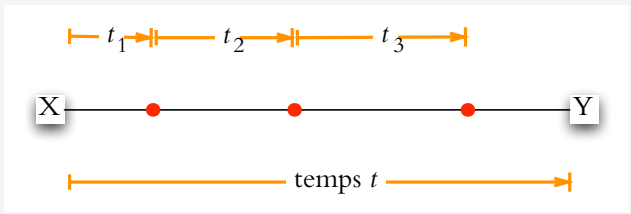


$$\mathbb{P}\Big\{Z = z \;\Big|\; X = x\Big\} = \sum_{y=1}^{r} \mathbb{P}\Big\{Z = z, Y = y \;\Big|\; X = x\Big\} = \sum_{y=1}^{r} \mathbb{P}\Big\{Z = z \;\Big|\; X = x, Y = y\Big\}\mathbb{P}\Big\{Y = y \;\Big|\; X = x\Big\}$$
$$= \sum_{y=1}^{r} \mathbb{P}\Big\{Z = z \;\Big|\; Y = y\Big\}\mathbb{P}\Big\{Y = y \;\Big|\; X = x\Big\} \qquad \textit{(propriété de Markov)}$$

Donc, on peut écrire que $\mathbf{M}_{X\to Z} = \mathbf{M}_{X\to Y} \cdot \mathbf{M}_{Y\to Z}$

6.2 Temps continu.

Comment peut-on ajouter une notion de temps ? Quelle est la matrice $\mathbf{M}$ qui correspond à un temps de divergence connu entre les séquences ?



À chaque mutation, le changement d'état est déterminé par la matrice $\mathbf{M}$. Si k événements en temps t, alors $\mathbf{M}_{X\to Y} = \mathbf{M}^k$. Donc la matrice pour temps t devrait être $\mathbf{M}^{N(t)}$ où $N(t)$ est le nombre de mutations pendant temps t. Problème : $N(t)$ est aléatoire...

Procéssus de Poisson. Supposons que $\{N(t)\colon t \in [0,\infty)\}$ est un procéssus qui compte l'occurrence d'événements de quelque sort. Supposons que le procéssus satisfait les critères suivants

☞ **[«orderly»]** La probabilité qu'il y ait plus d'une occurrence dans un petit intervalle de temps est négligeable

$$\lim_{\delta\to 0} \mathbb{P}\Big\{N(t+\delta) > 1 \;\Big|\; N(t+\delta) \geq 1\Big\} = 0$$

☞ **[«memoryless»]** Le nombre d'occurrences pendant un intervalle ne dépend pas des événements précédents : $N(t+s) - N(t)$ est indépendant de $N(t)$

Si le procéssus est *orderly* et *memoryless*, alors c'est un **procéssus de Poisson** avec les propriétés suivantes.

⋆ nombre d'arrivées pendant temps s est une v.a. Poisson :

$$\forall t, s\colon \mathbb{P}\{N(t+s) - N(t) = k\} = e^{-\lambda s}\frac{(\lambda s)^k}{k!}$$

Le paramètre λ est l'intensité ou le **taux** du procéssus.

⋆ temps d'attente est une v.a. exponentielle : $\forall t, s\colon \mathbb{P}\big\{N(t+s) = N(t)\big\} = e^{-\lambda t}$

Procéssus de Markov. En retournant à ce qui se passe pendant temps t :

$$\sum_{k=0}^{\infty} e^{-\lambda t}\frac{(\lambda t)^k}{k!}\mathbf{M}^k = \exp\Big(\lambda t(\mathbf{M}-\mathbf{I})\Big) = \exp\Big(\lambda t\mathbf{Q}\Big)$$

où $\mathbf{I}$ est la matrice d'unité. $\mathbf{Q} = \mathbf{M} - \mathbf{I}$ est la *matrice instantanée de taux de substitutions*. Calcul de $\exp(\lambda\mathbf{Q}t)$: décomposition de la matrice $\mathbf{Q} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}$, et faire $e^{\lambda\mathbf{Q}t} = \mathbf{U}e^{\lambda t\mathbf{\Lambda}}\mathbf{V}$.

$$\begin{pmatrix} -\eta & \eta \\ \nu & -\nu \end{pmatrix}$$
perte-gain (binaire)

$$\begin{pmatrix} \cdot & \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \cdot & \frac{\mu}{4} & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \cdot & \frac{\mu}{4} \\ \frac{\mu}{4} & \frac{\mu}{4} & \frac{\mu}{4} & \cdot \end{pmatrix}$$
Jukes-Cantor (DNA)

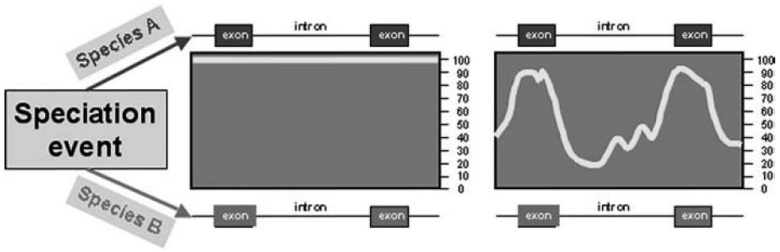
$$\begin{pmatrix} \cdot & \pi_{\mathtt{C}}\alpha & \pi_{\mathtt{G}}\beta & \pi_{\mathtt{T}}\alpha \\ \pi_{\mathtt{A}}\alpha & \cdot & \pi_{\mathtt{G}}\alpha & \pi_{\mathtt{T}}\beta \\ \pi_{\mathtt{A}}\beta & \pi_{\mathtt{C}}\alpha & \cdot & \pi_{\mathtt{T}}\alpha \\ \pi_{\mathtt{A}}\alpha & \pi_{\mathtt{C}}\beta & \pi_{\mathtt{G}}\alpha & \cdot \end{pmatrix}$$
Hasegawa-Kishino-Yano (DNA)

(somme est 0 dans toute rangée)

Le modèle de Yang et Nielsen (1998) pour codons utilise un paramètre ω qui mesure sélection, et un paramètre κ pour rapport de transitions/transversions :

$$\mathbf{Q}[a,b] = \begin{cases} \mu\pi_b & 1 \text{ transversion synonyme} \\ \mu\kappa\pi_b & 1 \text{ transition synonyme} \\ \mu\omega\pi_b & 1 \text{ transversion non-synonyme} \\ \mu\kappa\omega\pi_b & 1 \text{ transition non-synonyme} \end{cases}$$

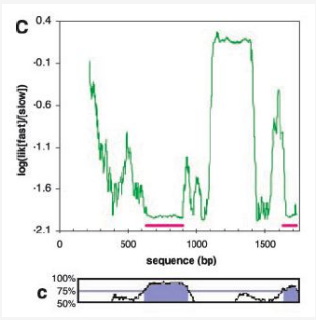
6.3 Conservation de séquence



Miller & al. *Annu Rev Genomics Hum Genet* 5 :15 (2004)

Principe de génomique comparative : éléments fonctionnels sont plus conservés (séléction négative) que les éléments non-fonctionnels (évolution neutre)

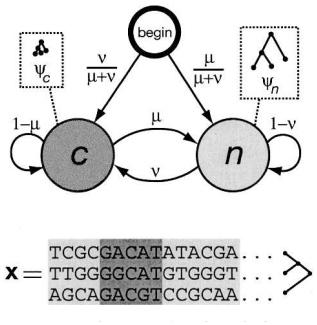
Phylogenetic shadowing. Méthode : modèle de substitutions dans les cas de sélection négative et neutre



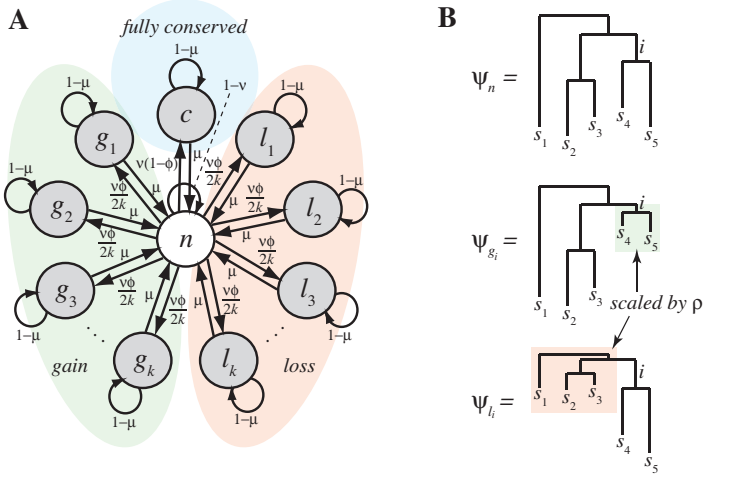
Comparaison de séquences entre des espèces proches : évolution rapide/lente $e^{\mu \mathbf{Q} t}$ où $\mu < 1$ pour des régions sous sélection purificatrice.
Exemple : fenêtre glissant de 50 pb sur alignement multiple de primates. Axis Y : score LODS (en haut) et % d'identité (en bas)

Boffelli & al. *Science* 299 :1391 (2003)

PhyloHMM. Modèle HMM pour segmentation : émissions = colonnes de l'alignement multiple, avec probabilités $e^{\mathbf{Q}t}$ (neutre) ou $e^{\rho \mathbf{Q} t}$ (sélection négative avec $\rho < 0$)



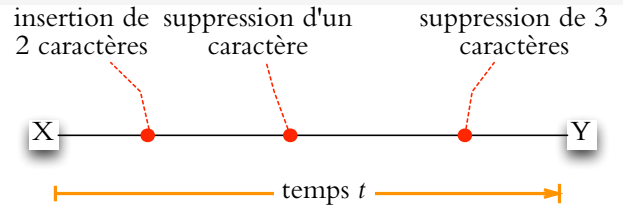
deux états : conservé ou non
Siepel & al. *Genome Res* 15 :1034 (2005)



DLESS : états correspondent à conservation dans un clade phylogénétique, ou à la perte de fonction
Siepel & al. *RECOMB 2006*

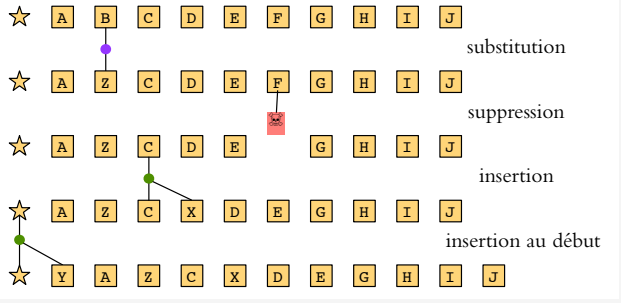
6.4 Modèle temporaire pour suppressions et insertions

On a vu le modèle de Markov pour substitutions — est-ce qu'on peut modéliser les indels par un procéssus stochastique ?

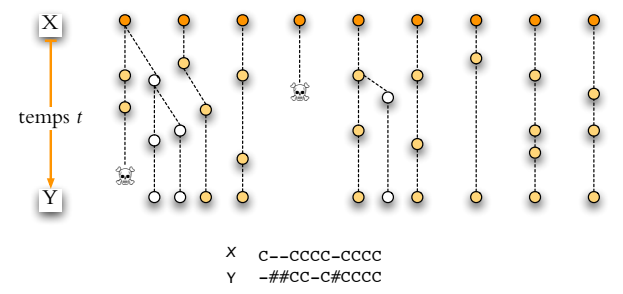


Idée : événements forment un procéssus (p.e., Poisson de taux θ), il faut juste spécifier ce qui se passe lors d'un événement. Un modèle assez général : insertions de longueur ℓ avec probabilité $p_{\mathrm{Ins}}(\ell)$ et suppressions de longueur ℓ avec probabilité $p_{\mathrm{Del}}(\ell)$ — très difficile (ou impossible) de travailler avec...

Thorne-Kishino-Felsenstein [1991] Un modèle simple : insertions et suppressions de longueur 1.



Un procéssus indel (en parallèle avec les procéssus de substitution) à chaque caractère + au début.
Événement d'insertion : choix d'un caractère selon la distribution stationnaire.
Taux d'insertion λ à chaque caractère ainsi qu'au début ; taux de suppression μ à chaque caractère.



But : étant donné une séquence ancestrale et une séquence déscendante, établir la homologie au niveau des caractères (représentée par un alignement). Les caractères s'évoluent indépendamment : on peut considérer le sort de chaque caractère ancestral séparamment.

Sort	Description	Bloc d'alignement
$C\underbrace{\#\cdots\#}_{n \text{ fois}}$	ancêtre survivant, avec $n \geq 0$ caractères inserés à côté	`C---` `C###`
$\square\underbrace{\#\cdots\#}_{n \text{ fois}}$	ancêtre mort, avec $n > 0$ caractères insérés à côté	`C---` `-###`
$\square$	ancêtre mort, aucune insertion à côté	`C` `-`
$\star\underbrace{\#\cdots\#}_{n \text{ fois}}$	insertion de $n \geq 0$ caractères au début	`---` `###`

Sort	probabilité $X \to Y$
$C \to C\#^{n-1}$	$H_t B_t^{n-1}$
$C \to \square\#^n$	$N_t B_t^{n-1}$
$C \to \square$	E_t
$\star \to \star\#^n$	$I_t B_t^n$

$$\beta_t = \frac{1-e^{-(\mu-\lambda)t}}{\mu-\lambda e^{-(\mu-\lambda)t}}$$

$$B_t = \lambda\beta_t \qquad E_t = \mu\beta_t$$
$$I_t = 1-\lambda\beta_t \qquad H_t = e^{-\mu t}(1-\lambda\beta_t)$$
$$N_t = (1-e^{-\mu t}-\mu\beta_t)(1-\lambda\beta_t);$$

Maintenant on peut

⋆ trouver le meilleur alignment

⋆ maximiser la vraisemblance pour deux séquences homologues et ainsi trouver les valeurs de λ, μ, t

⋆ concevoir des tests de homologie basés sur la valeur de la vraisemblance

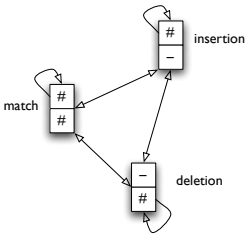
⋆ évaluer la fiabilité de l'alignment optimal

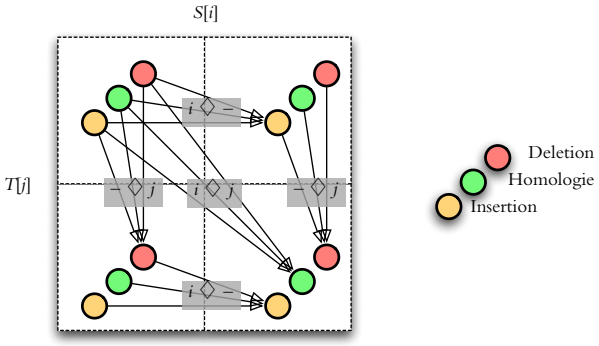
Exemple : calcul de la vraisemblance — utiliser la programmation dynamique (en se servant des queues géométriques $\propto B^n$). $L(i,j)$ probabilité que $X[1..i]$ est devenu $Y[1..j]$; $Z(i,j)$ variables auxiliaires. Récurrences :

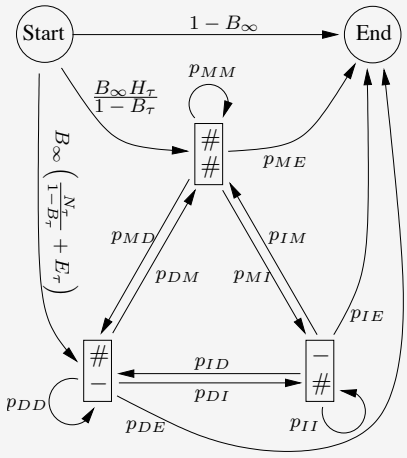
$$Z(i,j) = p_{i,j} \cdot H_t \cdot Z(i-1,j-1) + p_j \cdot N_t \cdot Z(i,j-1) + p_j \cdot B_t \cdot Z(i,j-1)$$
$$L(i,j) = Z(i,j) + E_t \cdot L(i-1,j),$$

où $p_{i,j} = \mathbf{M}\big[X[i], Y[j]\big]$ est la probabilité $X[i] \to Y[j]$ dans la matrice de substitution, et $p_j = \pi_{Y[j]}$ est la probabilité stationnaire de $Y[j]$. Initialisation : $Z(0,0) = L(0,0) = I_t$.

Paire-HMM. Modèle de Markov caché où les états correspondent à des insertions, suppressions et substitutions/identités dans l'alignement. La machine *émet* un alignment, mais on obsèrve seulement les séquences sans trous. Problème d'inférence : on connaît les séquences — qu'est-ce qu'on peut dire sur l'alignment inconnu ?







Il existe une paire-HMM équivalent au TKF91 :

(transitions entre états p_{MM}, p_{MI}, etc. calculées à partir de I_t, B_t, H_t, N_t, E_t)

6.5 Alignement ?

alignement = hypothèse de homologies

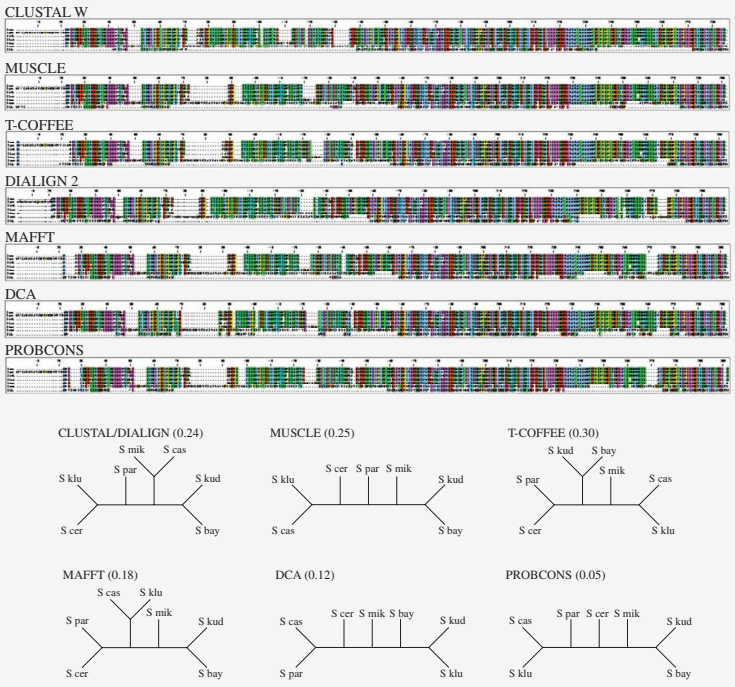
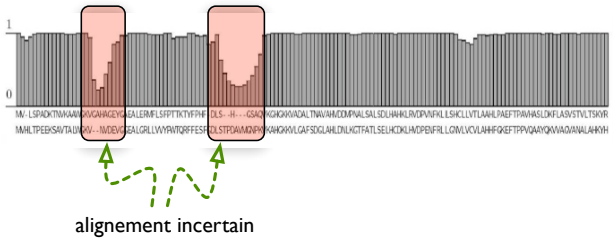


Illustration : sept logiciels construisent sept alignements différents qui mènent à 6 phylogénies différentes... [Wong, Suchard & Huelsenbeck. *Science* 319 :473–476, 2008]
Il est difficile d'aligner les séquences parce que l'information sur homologie n'est pas directe. Il n'existe pas d'alignements certains !

Fiabilité. On a la probabilité postérieure de homologie ($i \diamondsuit j$), et celle de manque de homologie ($i \diamondsuit -$, $- \diamondsuit j$) pour juger les colonnes de l'alignement. Une partie bien alignée correspond à une séquence de colonnes avec de grandes probabilités postérieures.



Maximal Expected Accuracy. l'alignement qui maximise le nombre de paires alignés correctement — utiliser $\mathbb{P}\{i \diamondsuit j\}$ comme le score du match de la colonne $\begin{matrix} X[i] \\ Y[j] \end{matrix}$, pas de pénalité pour trous.

AMAP. Problème avec MEA — qu'est-ce qui se passe avec deux séquence non-reliées ? Maximiser plutôt le nombre (en espérance) d'accords entre le vrai alignement (inconnu) et l'alignement qu'on calcule :

$$\text{désaccords} = n + m - 2M - D - I,$$

où M, D, I sont les nombres de matches, suppressions et insertions en commun entre deux alignements. En espérance, on a $\mathbb{E}M = \sum_{i,j} \mathbb{P}\{i \diamondsuit j\}$, $\mathbb{E}D = \sum_i \mathbb{P}\{i \diamondsuit -\}$, $\mathbb{E}I = \sum_j \mathbb{P}\{- \diamondsuit j\}$ (où $\mathbb{P}$ dénote probabilité postérieure, conditionnée sur les deux séquences).