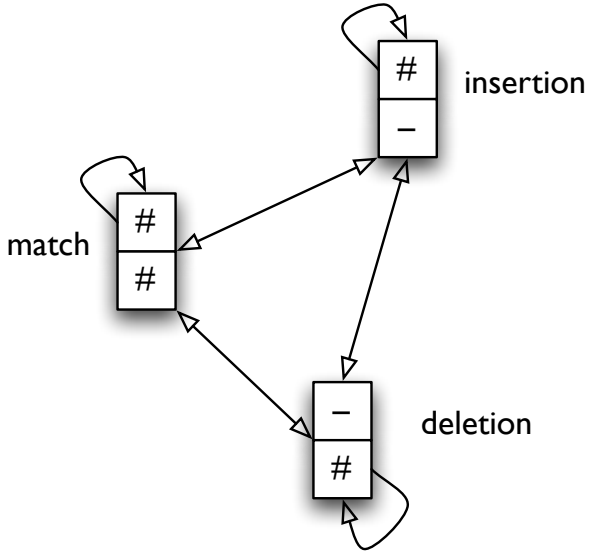
Alignement statistique

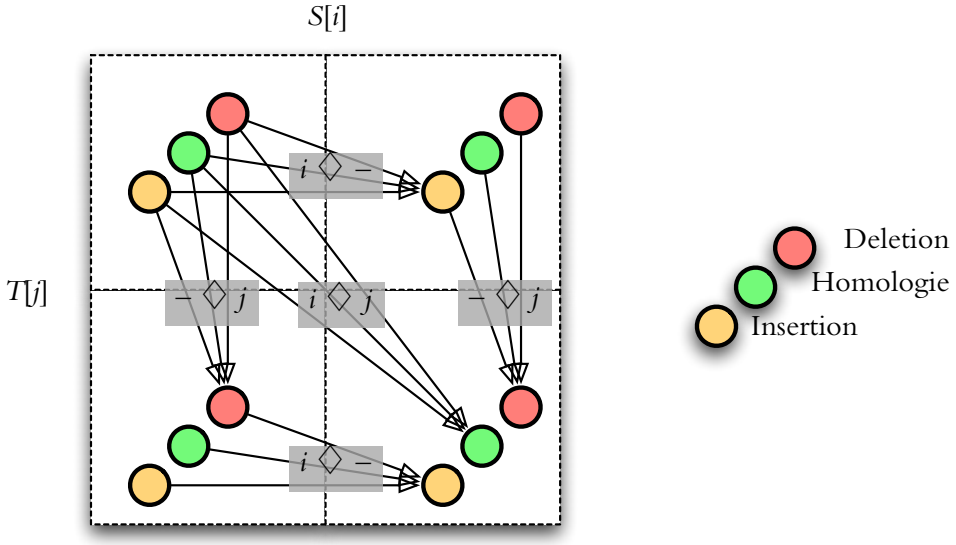
Paire-HMM

Modèle de Markov caché :
$\longrightarrow$ états = {insertion, suppression, substitution/identité}
$\longrightarrow$ émission : colonne de l'alignement

Problème d'inférence : on connaît les séquences sans trous — qu'est-ce qu'on peut dire sur l'alignment inconnu ?



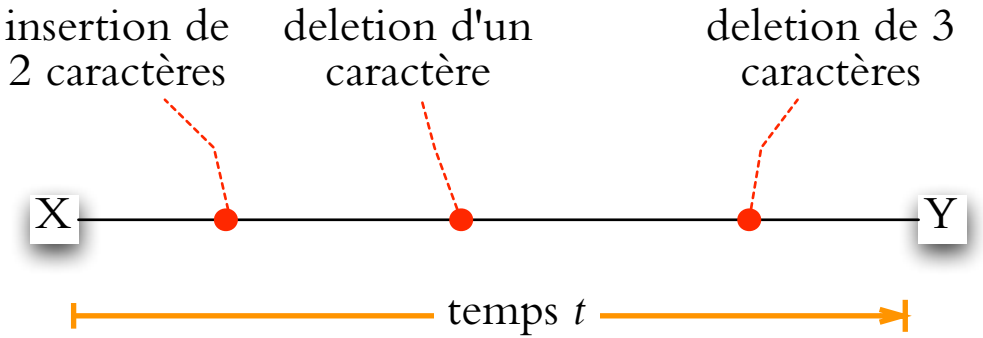
Paire-HMM : alignement



Viterbi pour trouver le meilleur alignement . . .

Procéssus d'indel

On a vu les modèles de Markov pour substitutions — est-ce qu'on peut modéliser aussi 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

Longueur d'insertsion/suppressions ? Distributions $p_{\mathsf{Ins}}(\ell)$ et $p_{\mathsf{Del}}(\ell)$:

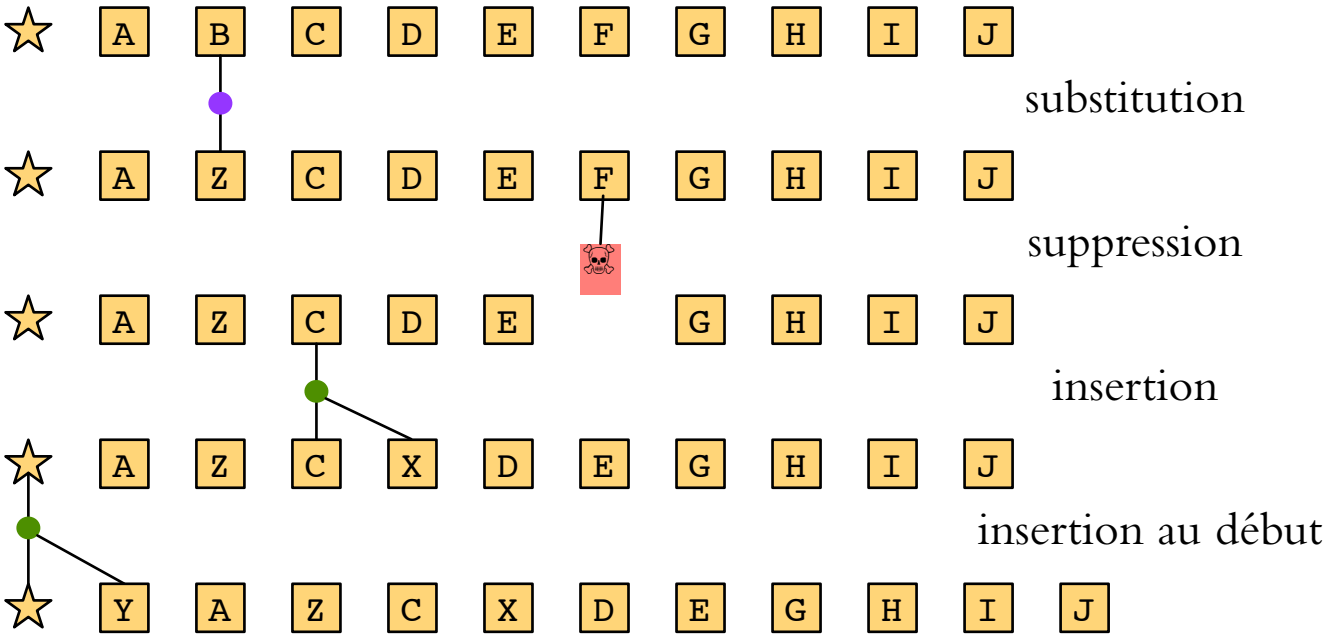
$$\sum_{\ell=1}^{\infty} p_{\mathsf{Ins}}(\ell) + \sum_{\ell=1}^{\infty} p_{\mathsf{Del}}(\ell) = 1$$

$\longrightarrow$ très difficile (ou impossible) de travailler avec un tel modèle générique

Thorne-Kishino-Felsenstein 1991

Modèle simple : insertions et suppressions de longueur 1

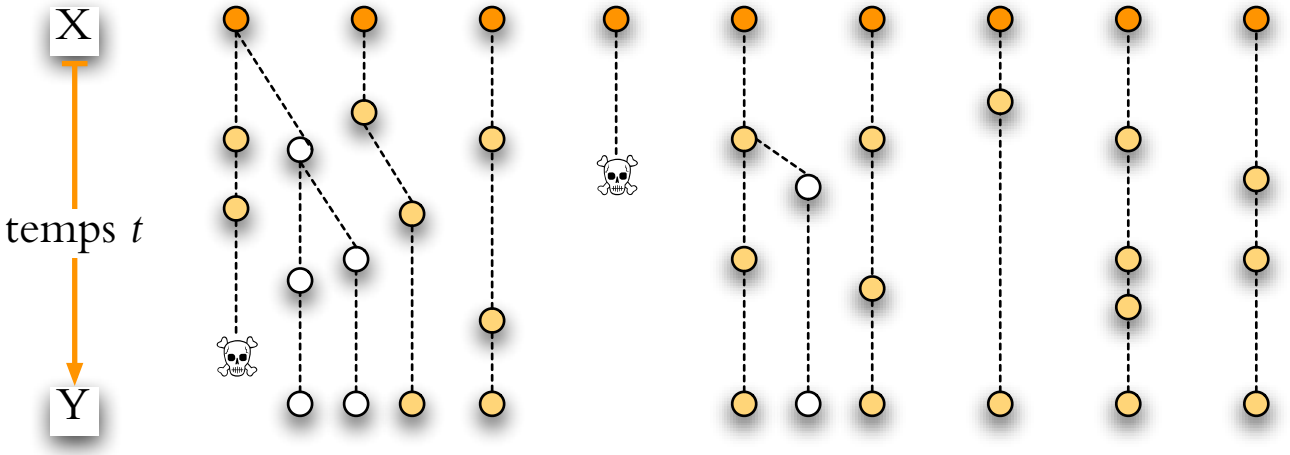
Un procéssus indel à chaque caractère + au début, en parallèle avec les procéssus de substitution (iid à chaque résidue)



Événement d'insertion : choix d'un caractère selon la distribution stationnaire

TKF 2

But : étant donné une séquence ancestrale et une séquence déscendante, établir de la homologie au niveau des caractères (représenté par un alignement)




```
X  c--cccc-cccc
Y  -##cc-c#cccc
```

C=ancêtre ou son homologue (match/mismatch)

#=aucune homologue à l'ancêtre (insertion)

TKF 3

Les caractères s'évoluent indépendamment : on peut considérer le sort de chaque caractère ancestral séparamment

$C \underbrace{\# \cdots \#}_{n \text{ fois}}$	ancêtre survivant, avec $n \geq 0$ caractères inserés à côté	$\begin{smallmatrix} \mathtt{C} & - & - & - \\ \mathtt{C} & \# & \# & \# \end{smallmatrix}$
$\square \underbrace{\# \cdots \#}_{n \text{ fois}}$	ancêtre mort, avec $n > 0$ caractères inserés à côté	$\begin{smallmatrix} \mathtt{C} & - & - & - \\ - & \# & \# & \# \end{smallmatrix}$
$\square$	ancêtre mort, aucune insertion à côté	$\begin{smallmatrix} \mathtt{C} \\ - \end{smallmatrix}$
$\star \underbrace{\# \cdots \#}_{n \text{ fois}}$	insertion de $n \geq 0$ caractères au début	$\begin{smallmatrix} - & - & - \\ \# & \# & \# \end{smallmatrix}$

TKF 4

Taux d'insertion λ à chaque caractère ainsi qu'au début
taux de suppression μ à chaque caractère

Longueur stationnaire ($t \to \infty$) de la séquence ? Soit $p_\ell = \mathbb{P}\{\text{longueur} = \ell\}$.
En équilibre, on a pour tout $\ell \geq 0$:

$$\underbrace{\lambda(\ell+1) \cdot p_\ell}_{\text{taux } \ell \to \ell+1} = \underbrace{\mu(\ell+1) \cdot p_{\ell+1}}_{\text{taux } \ell+1 \to \ell}$$

d'où $\frac{p_{\ell+1}}{p_\ell} = \frac{\lambda}{\mu}$

Si $\lambda \geq \mu$, alors $p_\ell \equiv 0$ (la séquence croît sans borne)

Si $\lambda < \mu$, alors $p_\ell \propto \left(\frac{\lambda}{\mu}\right)^\ell$
$\Rightarrow$ distribution géométrique ($\sum_{\ell=0}^{\infty} p_\ell = 1$)

TKF 5

Il est possible de trouver les formules exactes pour les probabilités des sorts :

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
$\star \to \star\#^n$	$I_t B_t^n$

avec

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

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

TKF 6

Maintenant on peut

- trouver le meilleur alignment
- maximiser la vraisemblance pour deux séquences homologoues 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

TKF 7

Calculer la vraisemblance — utiliser de la PD (en se servant des queues géométriques $\propto B^n$)

$L(i,j)$ probabilité de l'alignement de préfixes $X[1..i]$ et $Y[1..j]$;
$Z(i,j)$ variables auxiliaires

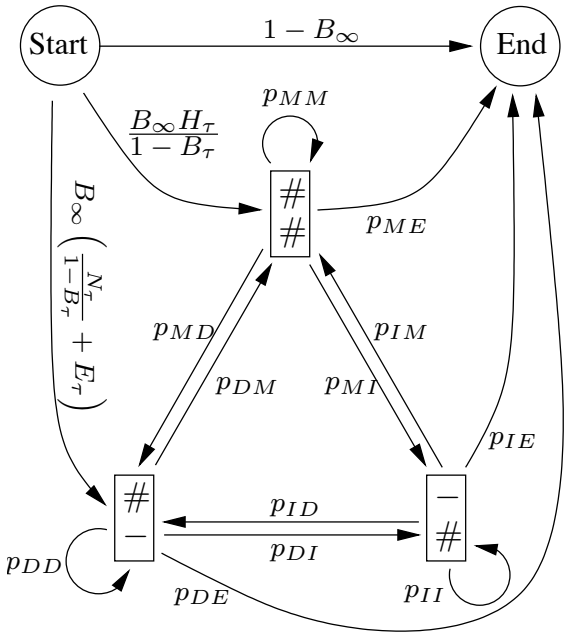
$$\begin{aligned} Z(i,j) = \quad & p_{i,j} \cdot H \cdot Z(i-1,j-1) && \left\{ H \times B^{n-1} \right\} \\ & + p_j \cdot N \cdot Z(i-1,j-1) && \left\{ N \times B^{n-1} \right\} \\ & + p_j \cdot B \cdot Z(i,j-1) && \left\{ N \text{ ou } H \times B^{n-1} \right\} \\ L(i,j) = & \, Z(i,j) + E \cdot L(i-1,j) \end{aligned}$$

où $p_{i,j}$ est la probabilité $X[i] \rightarrow Y[j]$ dans la matrice de substitution et p_j est la probabilité stationnaire de $Y[j]$;
initialisation :

$$Z(0,0) = L(0,0) = I$$

TKF 8

Il est possible de trouver un paire-HMM équivalent



(p_{MM}, p_{MI}, etc. sont calculés à partir de I_t, B_t, H_t, N_t, E_t)

Lunter et al. in Nielsen (ed.) *Statistical Methods in Molecular Evolution* (2005)