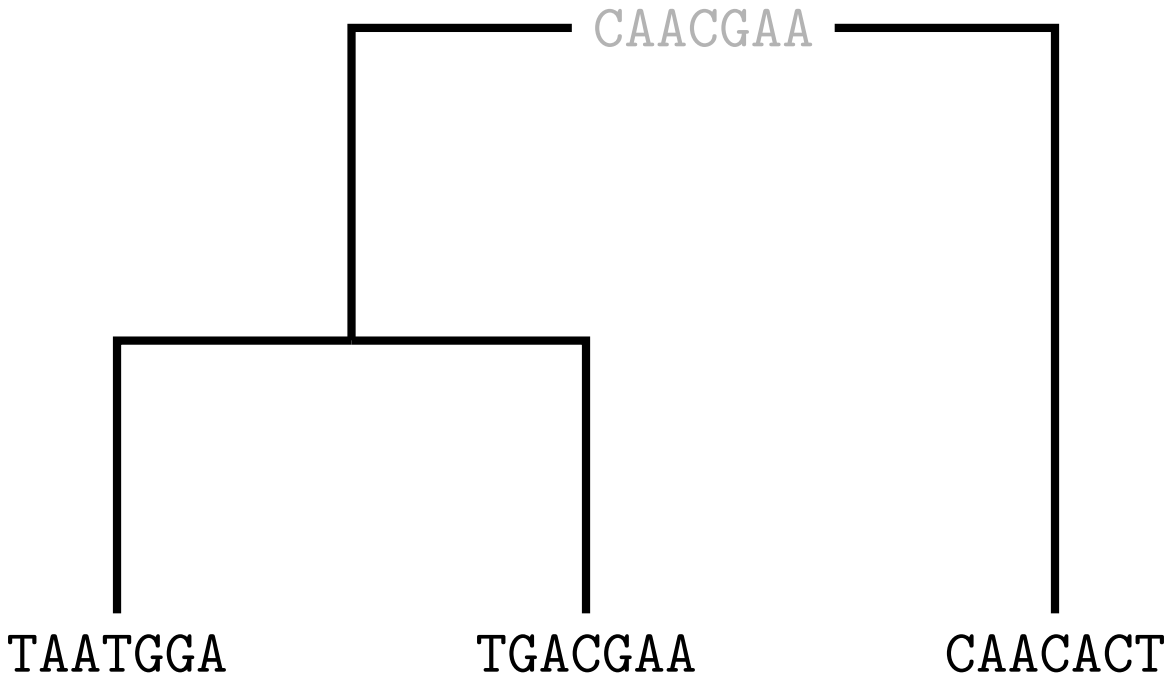
# Utilisation de méthodes particulaires pour l'inférence de modèles d'évolution avec dépendance au contexte

Alexis Huet

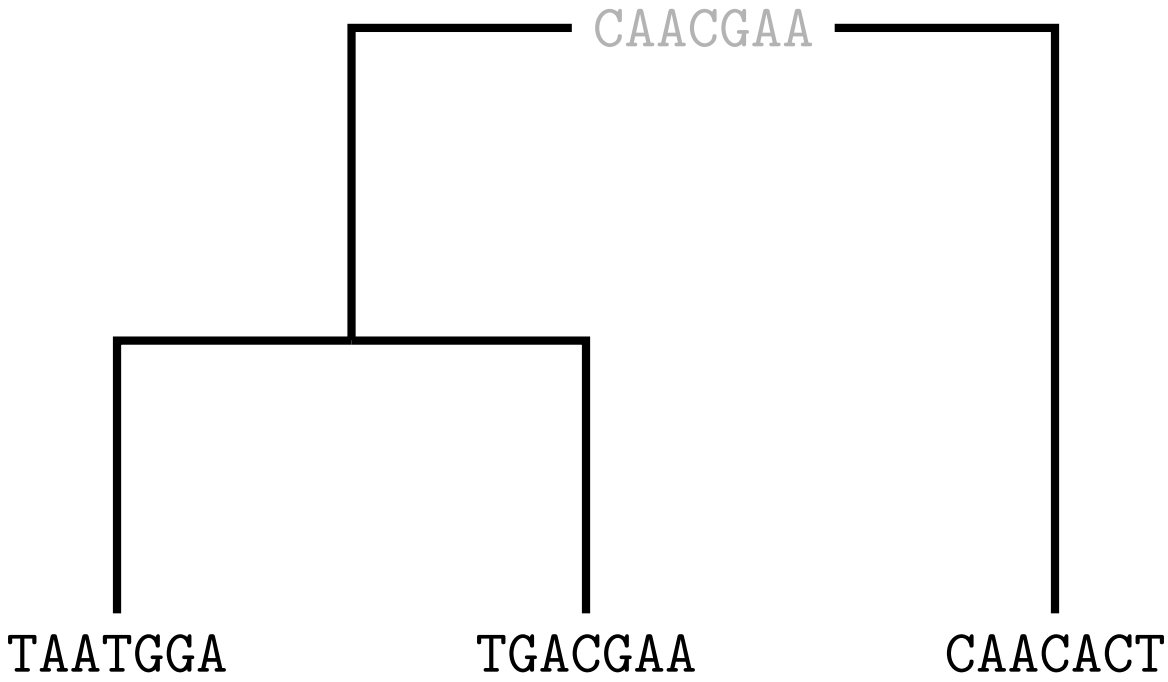
28 août 2014

# Plan

# Plan




$\rightarrow$ Problématique : étant donné un modèle d'évolution et une loi pour la séquence ancestrale, calculer la vraisemblance d'un alignement de séquences actuelles.




$\rightarrow$ Problématique : étant donné un modèle d'évolution et une loi pour la séquence ancestrale, calculer la vraisemblance d'un alignement de séquences actuelles.

# Modèles à sites indépendants

- chaque site évolue de façon indépendante selon la même loi,
- chaîne de Markov en temps continu sur $\{A, C, G, T\}$,

Exemple : générateur pour les modèles RN95 (Rzhetsky / Nei)

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} . & v_C & w_G & v_T \\ v_A & . & v_G & w_T \\ w_A & v_C & . & v_T \\ v_A & w_C & v_G & . \end{pmatrix} \end{array}$$

Deux types de bases : $A, G$=purines=$R$ et $C, T$=pyrimidines=$Y$.
Transversion : substitution $R \to Y$ ou $Y \to R$.
Transition : substitution $R \to R$ ou $Y \to Y$.

# Modèles à sites indépendants

- chaque site évolue de façon indépendante selon la même loi,
- chaîne de Markov en temps continu sur $\{A, C, G, T\}$,

Exemple : générateur pour les modèles RN95 (Rzhetsky / Nei)

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} . & v_C & w_G & v_T \\ v_A & . & v_G & w_T \\ w_A & v_C & . & v_T \\ v_A & w_C & v_G & . \end{pmatrix} \end{array}$$

Deux types de bases : $A, G$=purines=$R$ et $C, T$=pyrimidines=$Y$.
Transversion : substitution $R \to Y$ ou $Y \to R$.
Transition : substitution $R \to R$ ou $Y \to Y$.

# Modèles à sites indépendants

- chaque site évolue de façon indépendante selon la même loi,
- chaîne de Markov en temps continu sur $\{A, C, G, T\}$,

Exemple : générateur pour les modèles RN95 (Rzhetsky / Nei)

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} . & v_C & w_G & v_T \\ v_A & . & v_G & w_T \\ w_A & v_C & . & v_T \\ v_A & w_C & v_G & . \end{pmatrix} \end{array}$$

Deux types de bases : $A, G =$ purines $= R$ et $C, T =$ pyrimidines $= Y$.

Transversion : substitution $R \to Y$ ou $Y \to R$.

Transition : substitution $R \to R$ ou $Y \to Y$.

# Modèles à sites indépendants

- chaque site évolue de façon indépendante selon la même loi,
- chaîne de Markov en temps continu sur $\{A, C, G, T\}$,

Exemple : générateur pour les modèles RN95 (Rzhetsky / Nei)

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} . & v_C & w_G & v_T \\ v_A & . & v_G & w_T \\ w_A & v_C & . & v_T \\ v_A & w_C & v_G & . \end{pmatrix} \end{array}$$

Deux types de bases : $A, G =$ purines $= R$ et $C, T =$ pyrimidines $= Y$.
Transversion : substitution $R \to Y$ ou $Y \to R$.
Transition : substitution $R \to R$ ou $Y \to Y$.

# Modèles à sites indépendants

- chaque site évolue de façon indépendante selon la même loi,
- chaîne de Markov en temps continu sur $\{A, C, G, T\}$,

Exemple : générateur pour les modèles RN95 (Rzhetsky / Nei)

$$Q = \begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} . & v_C & w_G & v_T \\ v_A & . & v_G & w_T \\ w_A & v_C & . & v_T \\ v_A & w_C & v_G & . \end{pmatrix} \end{array}$$

Deux types de bases : $A, G$=purines=$R$ et $C, T$=pyrimidines=$Y$.
Transversion : substitution $R \to Y$ ou $Y \to R$.
Transition : substitution $R \to R$ ou $Y \to Y$.

# Hypermutabilité des dinucléotides CpG

Biochimiquement (par exemple chez les mammifères) :

- Taux $C \to T$ accru si le nucléotide à droite est $G$,
- Taux $G \to A$ accru si le nucléotide à gauche est $C$.

$\to$ Nécessité de définir des taux de substitution prenant en compte le contexte local.

# Hypermutabilité des dinucléotides CpG

Biochimiquement (par exemple chez les mammifères) :

- Taux $C \to T$ accru si le nucléotide à droite est $G$,
- Taux $G \to A$ accru si le nucléotide à gauche est $C$.

$\to$ Nécessité de définir des taux de substitution prenant en compte le contexte local.

# Taux de saut avec renforcement

$$P\left(\begin{array}{ll} \text{...ACGTA ...} & \text{temps t} \\ \downarrow & \\ \text{...ACATA ...} & \text{temps t+dt} \end{array}\right) = (w_A + r_{CG\to CA})dt + o(dt),$$

$$P\left(\begin{array}{ll} \text{...AGGTA ...} & \text{temps t} \\ \downarrow & \\ \text{...AGATA ...} & \text{temps t+dt} \end{array}\right) = w_A dt + o(dt).$$

# Taux de saut avec renforcement

$$P\begin{pmatrix} \text{...ACGTA ...} & \text{temps t} \\ \downarrow & \\ \text{...ACATA ...} & \text{temps t+dt} \end{pmatrix} = (w_A + r_{CG\to CA})dt + o(dt),$$

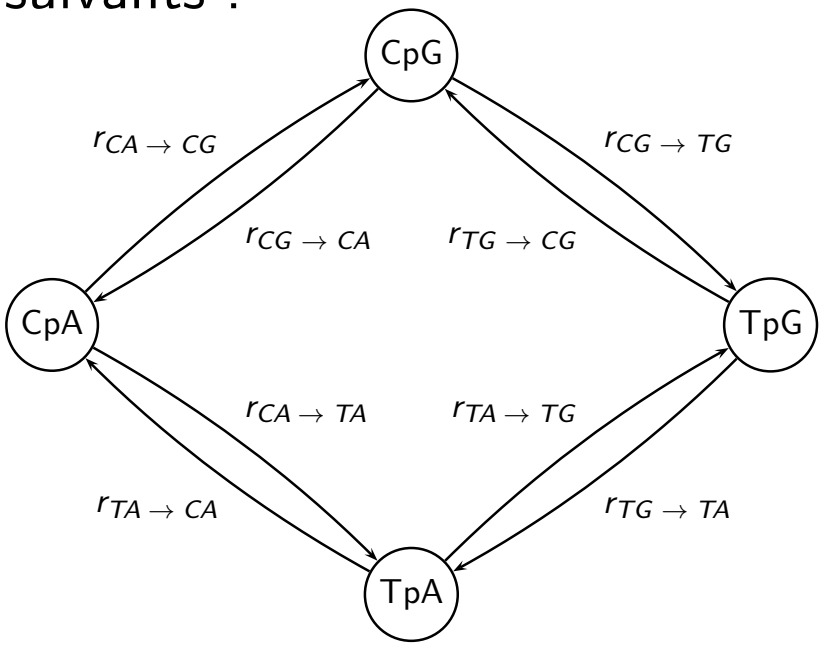
$$P\begin{pmatrix} \text{...AGGTA ...} & \text{temps t} \\ \downarrow & \\ \text{...AGATA ...} & \text{temps t+dt} \end{pmatrix} = w_A dt + o(dt).$$

# Classe de modèles RN95+YpR

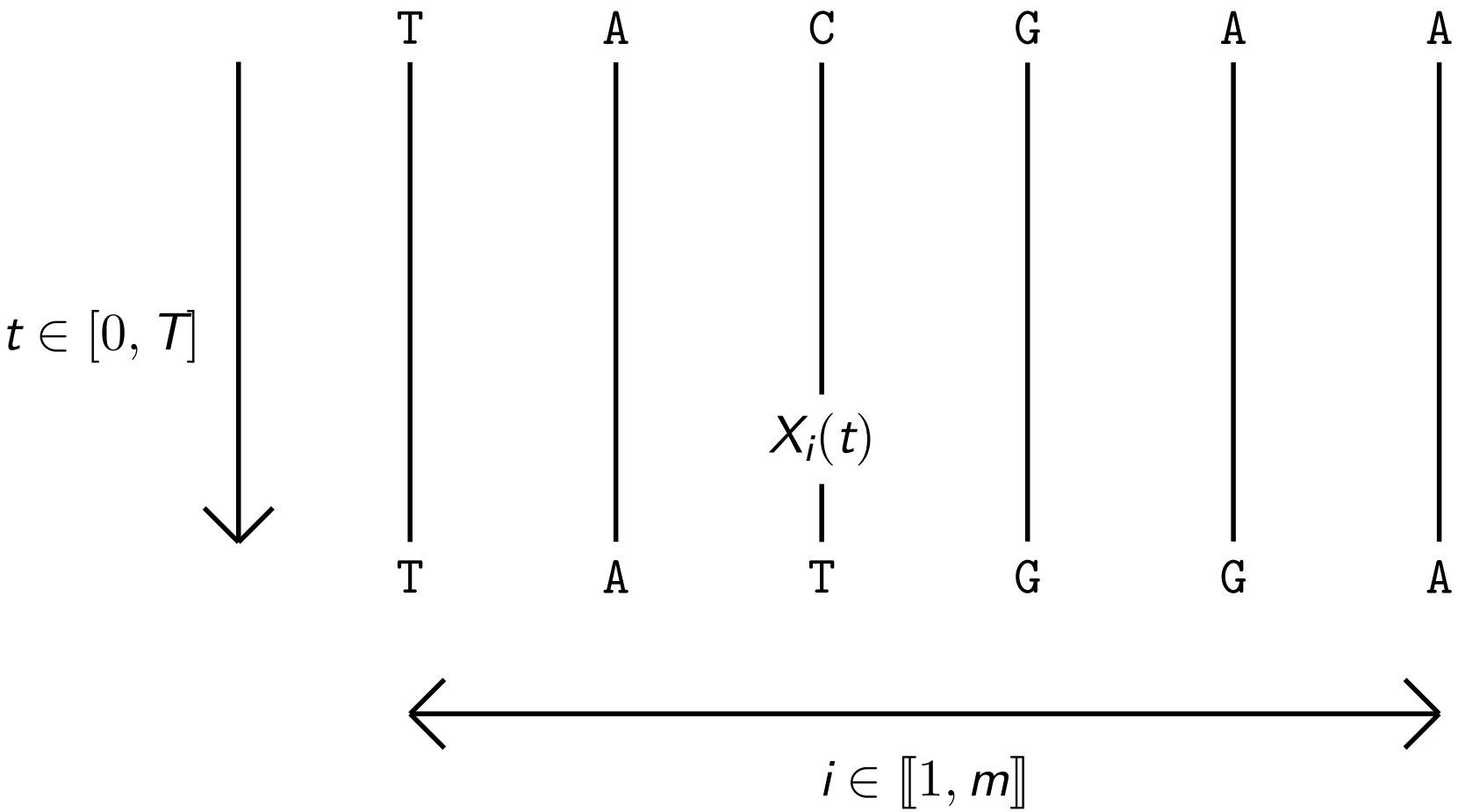
- modèle RN95 de matrice de sauts

$$\begin{array}{c} \\ A \\ C \\ G \\ T \end{array} \begin{array}{c} \begin{array}{cccc} A & C & G & T \end{array} \\ \begin{pmatrix} . & v_C & w_G & v_T \\ v_A & . & v_G & w_T \\ w_A & v_C & . & v_T \\ v_A & w_C & v_G & . \end{pmatrix} \end{array},$$

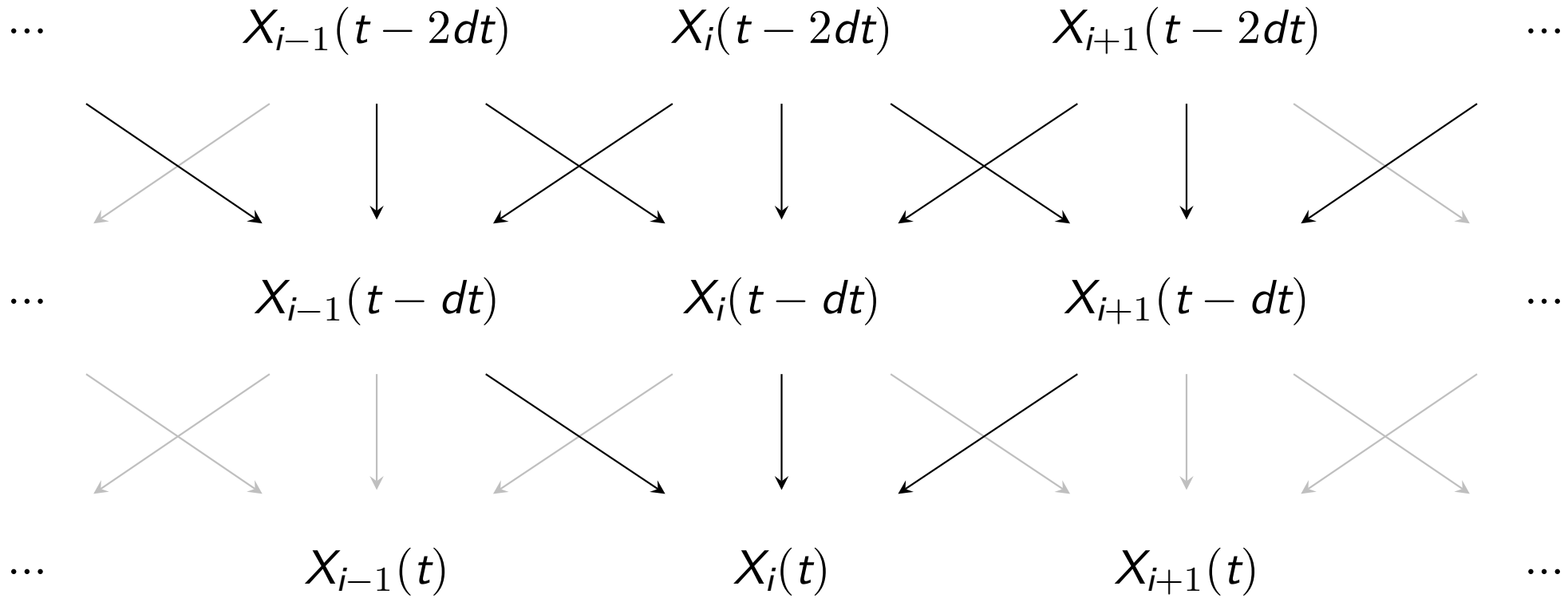
- renforcements suivants :



# Évolution de séquence à séquence



# Chaînes de dépendance



# Plan

# Encodages

Encodages de nucléotides :

- $\pi(A) := R$; $\pi(G) := R$; $\pi(C) := Y$; $\pi(T) := Y$,
- $\rho(A) := R$; $\rho(G) := R$; $\rho(C) := C$; $\rho(T) := T$,
- $\eta(A) := A$; $\eta(G) := G$; $\eta(C) := Y$; $\eta(T) := Y$.

$\Phi$-encodage d'une séquence de nucléotides :

$$\Phi(x_1(t), \ldots, x_m(t)) = (\rho(x_1(t)), x_2(t), \ldots, x_{m-1}(t), \eta(x_m(t))).$$

# Encodages

Encodages de nucléotides :

- $\pi(A) := R$; $\pi(G) := R$; $\pi(C) := Y$; $\pi(T) := Y$,
- $\rho(A) := R$; $\rho(G) := R$; $\rho(C) := C$; $\rho(T) := T$,
- $\eta(A) := A$; $\eta(G) := G$; $\eta(C) := Y$; $\eta(T) := Y$.

$\Phi$-encodage d'une séquence de nucléotides :

$$\Phi(x_1(t), \ldots, x_m(t)) = (\rho(x_1(t)), x_2(t), \ldots, x_{m-1}(t), \eta(x_m(t))).$$

# Théorème pour l'évolution $\Phi$-encodée

Évolution $\Phi$-encodée d'une séquence

$$(\rho(X_1)(t), X_2(t), \ldots, X_{m-1}(t), \eta(X_m)(t))_{t\in[0,T]}$$

issue d'un modèle RN95+YpR :

**Théorème [BGP08]**

Une séquence $\Phi$-encodée évolue dans le temps selon une chaîne de Markov explicite.

# Intérêt pour le calcul de la vraisemblance

$\rightarrow$ Calcul de la vraisemblance pour des séquences $\Phi$-encodées de longueurs $m = 2, 3, 4, 5$.

$\rightarrow$ Calcul de la vraisemblance pour des observations générales ?

- Vraisemblances composites.
- Approximation de type Monte Carlo.

# Intérêt pour le calcul de la vraisemblance

$\rightarrow$ Calcul de la vraisemblance pour des séquences $\Phi$-encodées de longueurs $m = 2, 3, 4, 5$.

$\rightarrow$ Calcul de la vraisemblance pour des observations générales ?

- Vraisemblances composites.
- Approximation de type Monte Carlo.

# Structure de chaîne de Markov explicite

Évolution $\Phi$-encodée d'une séquence issue d'un modèle RN95+YpR :

$$\Phi(X) = (\rho(X_1)(t), X_2(t), \ldots, X_{m-1}(t), \eta(X_m)(t))_{t\in[0,T]}$$

Plutôt que de regarder l'évolution site par site, on regarde l'évolution de chaque dinucléotide $\Phi$-encodé avec chevauchement.

**Définition**

$$\rho_i = (\rho(X_i(t)))_t \text{ et } \eta_i = (\eta(X_i(t)))_t,$$

$$Z_i = (\rho_i, \eta_{i+1}).$$

Alphabet associé : $\{C, T, R\} \times \{A, G, Y\}$.

$$\begin{aligned}\Phi(X) &= (\rho_1, X_2, \ldots, X_{m-1}, \eta_m)\\ &\equiv (\rho_1, \eta_2, \rho_2, \ldots, \eta_{m-1}, \rho_{m-1}, \eta_m)\\ &= (Z_1, \ldots, Z_{m-1}).\end{aligned}$$

# Structure de chaîne de Markov explicite

Évolution $\Phi$-encodée d'une séquence issue d'un modèle RN95+YpR :

$$\Phi(X) = (\rho(X_1)(t), X_2(t), \ldots, X_{m-1}(t), \eta(X_m)(t))_{t\in[0,T]}$$

Plutôt que de regarder l'évolution site par site, on regarde l'évolution de chaque dinucléotide $\Phi$-encodé avec chevauchement.

**Définition**

$$\rho_i = (\rho(X_i(t)))_t \text{ et } \eta_i = (\eta(X_i(t)))_t,$$

$$Z_i = (\rho_i, \eta_{i+1}).$$

Alphabet associé : $\{C, T, R\} \times \{A, G, Y\}$.

$$\begin{aligned}\Phi(X) &= (\rho_1, X_2, \ldots, X_{m-1}, \eta_m)\\ &\equiv (\rho_1, \eta_2, \rho_2, \ldots, \eta_{m-1}, \rho_{m-1}, \eta_m)\\ &= (Z_1, \ldots, Z_{m-1}).\end{aligned}$$

# Structure de chaîne de Markov explicite

**Théorème (thèse H.)**

On suppose la racine fixée. Alors $(Z_i)_i$ est une chaîne de Markov.

**Théorème (thèse H.)**

Conditionnellement à $Z_{1:i-1}$ et $Z_i([0,t[)$, la description de la loi de transition de $Z_i$ à l'instant $t$ est explicite.

1. Si $\pi_i(t^-) = \pi_i(t) \in \{R, Y\}$, matrice de taux de sauts $W_R$ ou $W_Y$.
2. Si $\pi_i(t^-) \neq \pi_i(t)$, substitution obligatoire régie par des matrices instantanées $U_{Y\to R}$ et $U_{R\to Y}$.

Démonstration basée sur les structures spécifiques de la classe RN95+YpR.

# Structure de chaîne de Markov explicite

**Théorème (thèse H.)**

On suppose la racine fixée. Alors $(Z_i)_i$ est une chaîne de Markov.

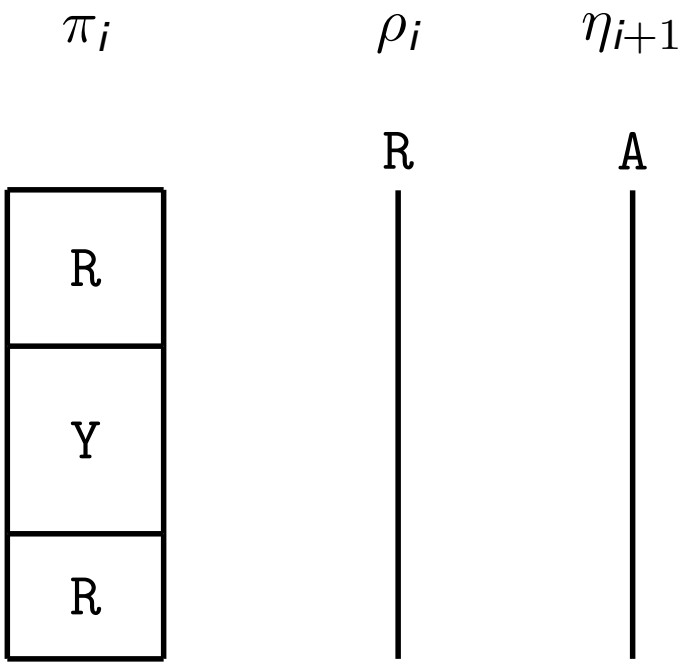
**Théorème (thèse H.)**

Conditionnellement à $Z_{1:i-1}$ et $Z_i([0,t[)$, la description de la loi de transition de $Z_i$ à l'instant $t$ est explicite.

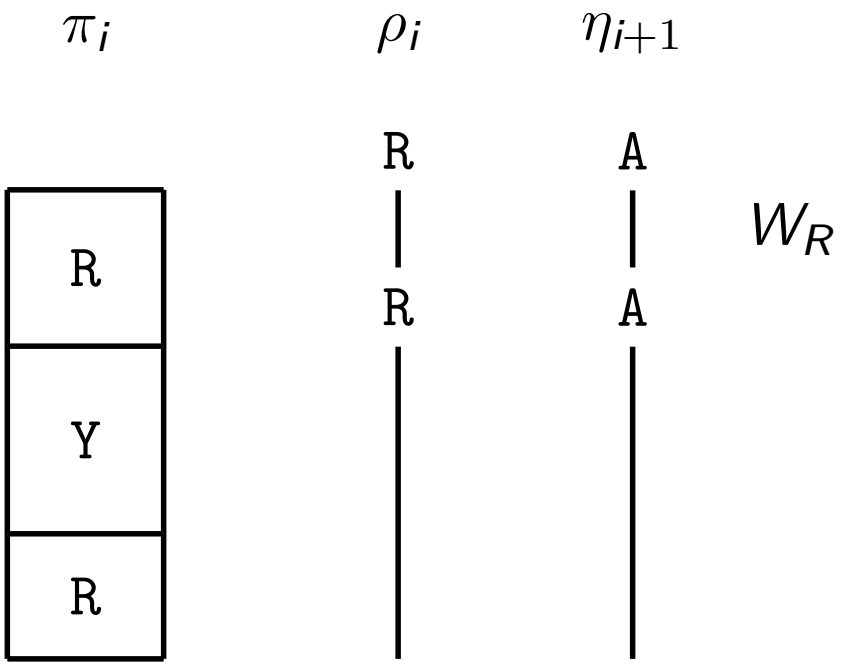
1. Si $\pi_i(t^-) = \pi_i(t) \in \{R, Y\}$, matrice de taux de sauts $W_R$ ou $W_Y$.
2. Si $\pi_i(t^-) \neq \pi_i(t)$, substitution obligatoire régie par des matrices instantanées $U_{Y\to R}$ et $U_{R\to Y}$.

Démonstration basée sur les structures spécifiques de la classe RN95+YpR.

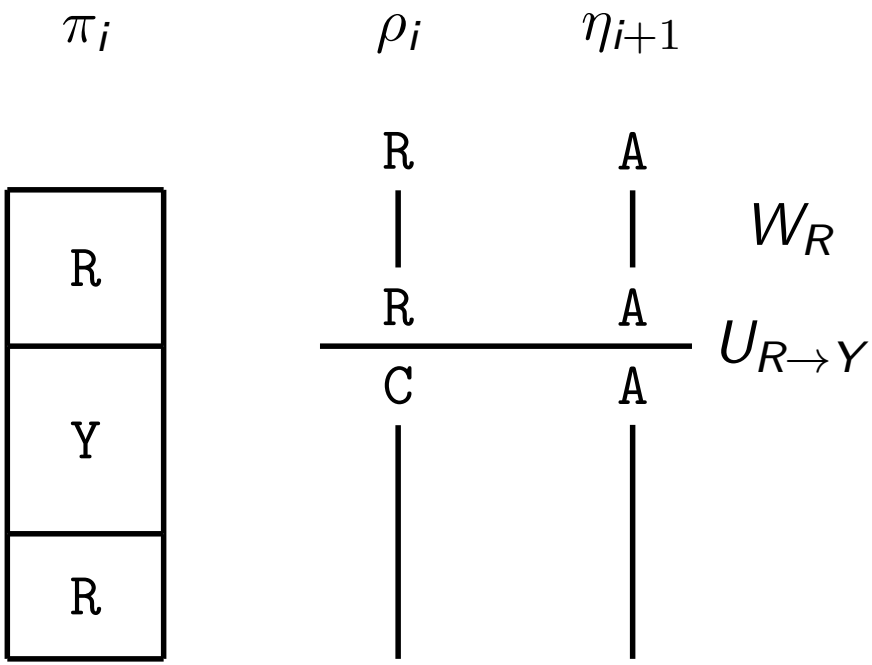
# Exemple



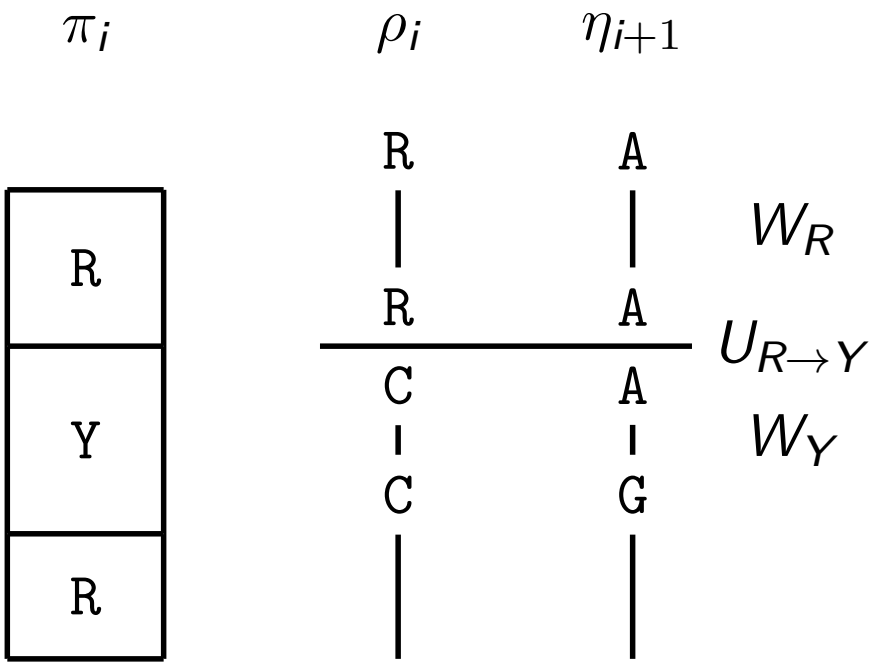
# Exemple



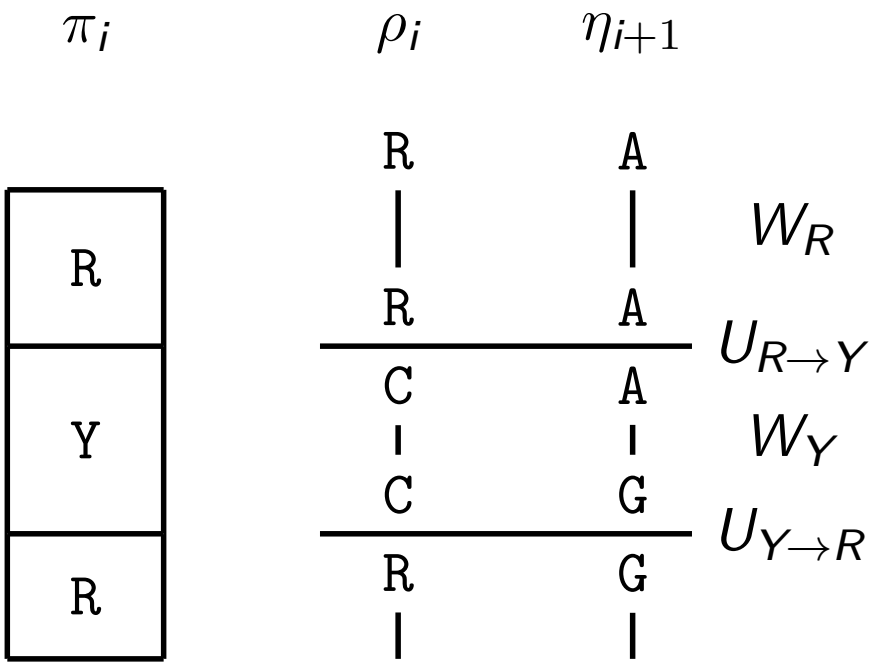
# Exemple



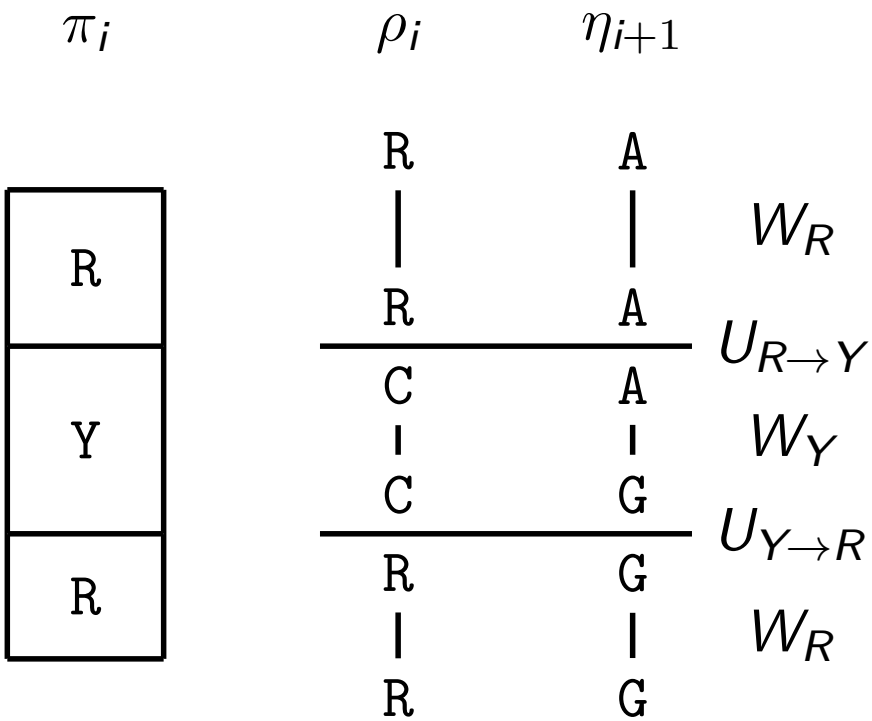
# Exemple



# Exemple



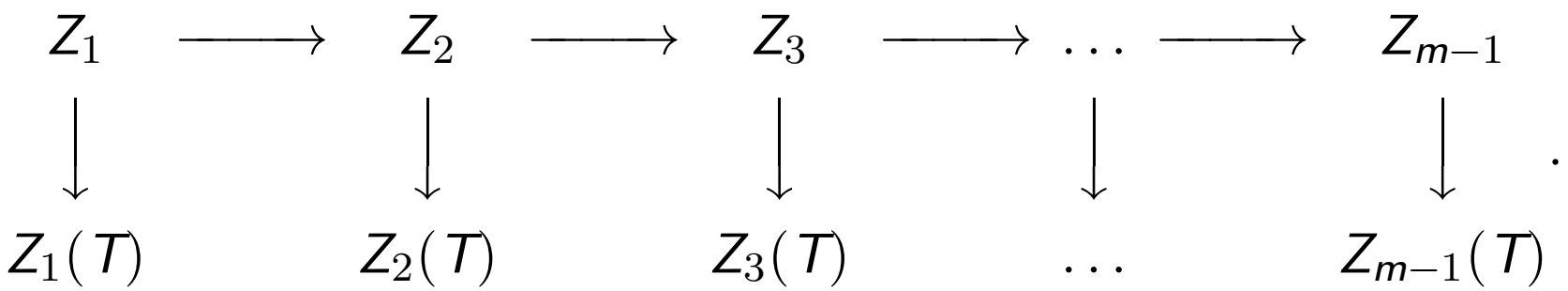
# Exemple



# Plan

# Comment approcher la vraisemblance ?

Pour l'évolution markovienne de l'historique $(Z_i)_i$, on a :



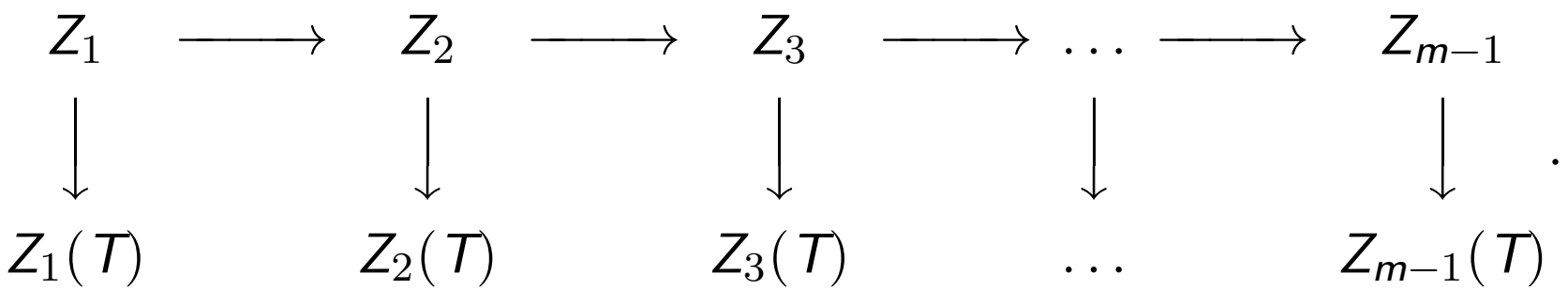

- Connaissance de $z_{1:m-1}(T)$ mais pas de $z_{1:m-1}$.
- But : calculer la vraisemblance des observations $z_{1:m-1}(T)$ par produit :

$$p(z_{1:m-1}(T)) = \prod_{i=0}^{m-2} p(z_{i+1}(T)|z_{1:i}(T)).$$

- Méthode : filtre particulaire auxiliaire (APF).

# Comment approcher la vraisemblance ?

Pour l'évolution markovienne de l'historique $(Z_i)_i$, on a :




- Connaissance de $z_{1:m-1}(T)$ mais pas de $z_{1:m-1}$.
- But : calculer la vraisemblance des observations $z_{1:m-1}(T)$ par produit :

$$p(z_{1:m-1}(T)) = \prod_{i=0}^{m-2} p(z_{i+1}(T)|z_{1:i}(T)).$$

- Méthode : filtre particulaire auxiliaire (APF).

# Algorithme APF

Idée de l'algorithme : approcher les lois de $z_{1:i}$ conditionnellement à $z_{1:i}(T)$ par la loi empirique associée à un nuage de particules.

- Site $i-1$ :

  Nuage $(z_{1:i-1}^{(j)})_j$ approche loi$(z_{1:i-1}|z_{1:i-1}(T))$.

- Pour toute particule $j \in 1:N$, conditionnellement à $z_{i-1}^{(j)}$, simuler selon $p(dz_i|z_{i-1}, z_i(T))$ :

  $$z_i^{(1)}, \ldots, z_i^{(N)}.$$

- Site $i$ :

  Nuage $(z_{1:i}^{(j)})_j$ approche loi$(z_{1:i}|z_{1:i}(T))$.

Puis : $p(z_{i+1}(T)|z_{1:i}(T)) = \int p(z_{i+1}(T)|z_i)\, p(dz_{1:i}|z_{1:i}(T))$.

Littérature : théorèmes de convergences et de normalité asymptotique lorsque le nombre de particules $N$ tend vers l'infini.

# Algorithme APF

Idée de l'algorithme : approcher les lois de $z_{1:i}$ conditionnellement à $z_{1:i}(T)$ par la loi empirique associée à un nuage de particules.

- Site $i-1$ :
  Nuage $(z_{1:i-1}^{(j)})_j$ approche loi$(z_{1:i-1}|z_{1:i-1}(T))$.
- Pour toute particule $j \in 1:N$, conditionnellement à $z_{i-1}^{(j)}$, simuler selon $p(dz_i|z_{i-1}, z_i(T))$ :
  $z_i^{(1)}, \ldots, z_i^{(N)}$.
- Site $i$ :
  Nuage $(z_{1:i}^{(j)})_j$ approche loi$(z_{1:i}|z_{1:i}(T))$.

Puis : $p(z_{i+1}(T)|z_{1:i}(T)) = \int p(z_{i+1}(T)|z_i)\, p(dz_{1:i}|z_{1:i}(T))$.

Littérature : théorèmes de convergences et de normalité asymptotique lorsque le nombre de particules $N$ tend vers l'infini.

# Algorithme APF

Idée de l'algorithme : approcher les lois de $z_{1:i}$ conditionnellement à $z_{1:i}(T)$ par la loi empirique associée à un nuage de particules.

- Site $i-1$ :
  Nuage $(z_{1:i-1}^{(j)})_j$ approche loi$(z_{1:i-1}|z_{1:i-1}(T))$.
- Pour toute particule $j \in 1:N$, conditionnellement à $z_{i-1}^{(j)}$, simuler selon $p(dz_i|z_{i-1}, z_i(T))$ :
  $$z_i^{(1)}, \dots, z_i^{(N)}.$$
- Site $i$ :
  Nuage $(z_{1:i}^{(j)})_j$ approche loi$(z_{1:i}|z_{1:i}(T))$.

Puis : $p(z_{i+1}(T)|z_{1:i}(T)) = \int p(z_{i+1}(T)|z_i)\, p(dz_{1:i}|z_{1:i}(T))$.

Littérature : théorèmes de convergences et de normalité asymptotique lorsque le nombre de particules $N$ tend vers l'infini.

# Algorithme APF

Idée de l'algorithme : approcher les lois de $z_{1:i}$ conditionnellement à $z_{1:i}(T)$ par la loi empirique associée à un nuage de particules.

- Site $i-1$ :
  Nuage $(z_{1:i-1}^{(j)})_j$ approche loi$(z_{1:i-1}|z_{1:i-1}(T))$.

- Pour toute particule $j \in 1:N$, conditionnellement à $z_{i-1}^{(j)}$, simuler selon $p(dz_i|z_{i-1}, z_i(T))$ :
  $$z_i^{(1)}, \dots, z_i^{(N)}.$$

- Site $i$ :
  Nuage $(z_{1:i}^{(j)})_j$ approche loi$(z_{1:i}|z_{1:i}(T))$.

Puis : $p(z_{i+1}(T)|z_{1:i}(T)) = \int p(z_{i+1}(T)|z_i)p(dz_{1:i}|z_{1:i}(T))$.

Littérature : théorèmes de convergences et de normalité asymptotique lorsque le nombre de particules $N$ tend vers l'infini.

# Algorithme APF

Idée de l'algorithme : approcher les lois de $z_{1:i}$ conditionnellement à $z_{1:i}(T)$ par la loi empirique associée à un nuage de particules.

- Site $i-1$ :
  Nuage $(z_{1:i-1}^{(j)})_j$ approche loi$(z_{1:i-1}|z_{1:i-1}(T))$.

- Pour toute particule $j \in 1:N$, conditionnellement à $z_{i-1}^{(j)}$, simuler selon $p(dz_i|z_{i-1}, z_i(T))$ :
  $$z_i^{(1)}, \dots, z_i^{(N)}.$$

- Site $i$ :
  Nuage $(z_{1:i}^{(j)})_j$ approche loi$(z_{1:i}|z_{1:i}(T))$.

Puis : $p(z_{i+1}(T)|z_{1:i}(T)) = \int p(z_{i+1}(T)|z_i)p(dz_{1:i}|z_{1:i}(T))$.

Littérature : théorèmes de convergences et de normalité asymptotique lorsque le nombre de particules $N$ tend vers l'infini.

# Implémentation

Entrées (sous forme de fichier texte) :

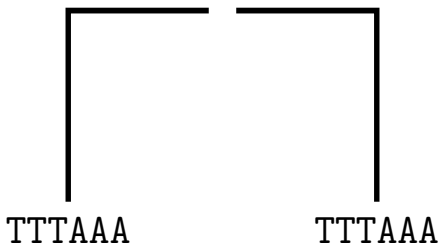
- le jeu de séquences, arbre, paramètres du modèle
- la loi à la racine (modèle markovien),
- $N$ le nombre de particules utilisées.

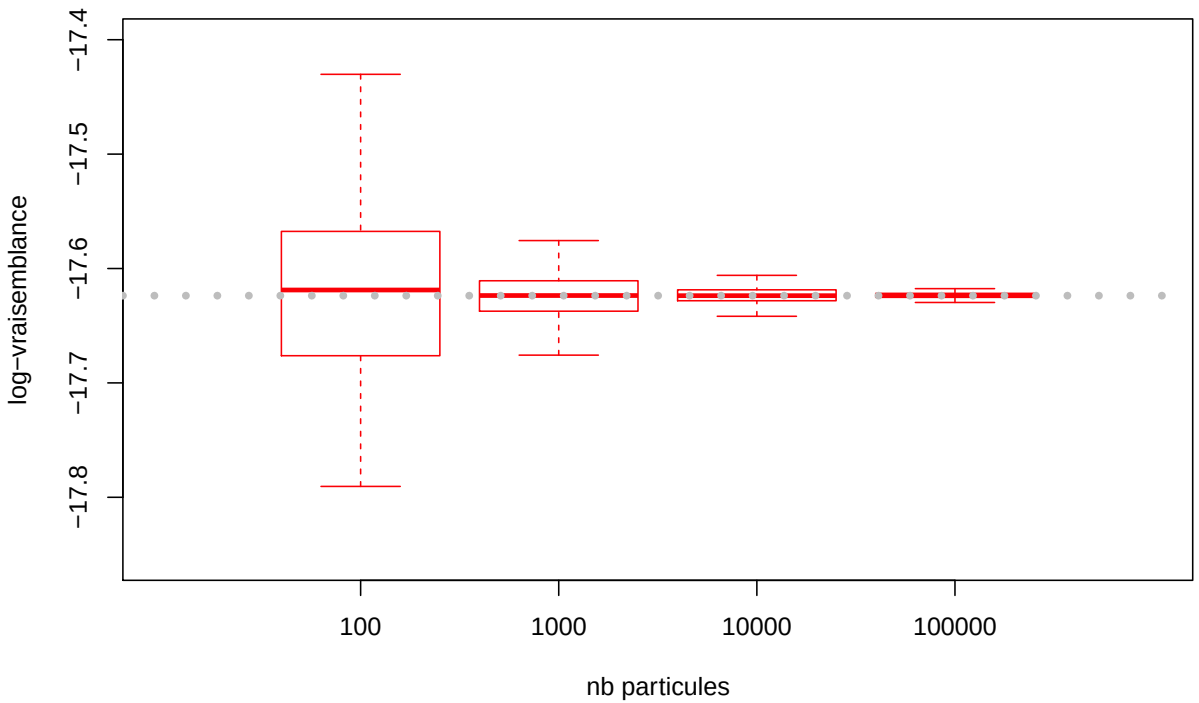
Sortie :

- la log-vraisemblance approchée du jeu de séquences observé.

# Exemple

- Modèle d'évolution $v_A = 7.1, \ \ldots \ r_{TG \to CG} = 7.7.$
- Arbre constitué de deux arêtes de longueur $1$.
- Loi à la racine loi stationnaire du modèle.
- Séquences observées : $(TTTAAA, TTTAAA)$.



# Plan

# Algorithme de calcul approché de la vraisemblance

## Algorithme pratique

1. Estimer le maximum de vraisemblance des observations par la méthode des triplets encodés (`bppml` de Bio++ [BG12, DB08, DGB+06]).
2. Calculer une approximation particulaire de la vraisemblance au maximum de vraisemblance.

# Comparaison de modèles

Exemple d'application : on dispose d'un alignement de trois séquences biologiques de $2215$ nucléotides.

| modèle | T92 | T92+CpGs | GTR |
|---|---|---|---|
| nombre de paramètres | 2 | 3 | 8 |
| log-vrais. | $-3432$ | ? | $-3428$ |
| AIC | $6868$ | ? | $6872$ |
| BIC | $6879$ | ? | $6918$ |

$\rightarrow$ Algorithme pratique appliqué pour le modèle T92+CpGs.

# Comparaison de modèles

Exemple d'application : on dispose d'un alignement de trois séquences biologiques de $2215$ nucléotides.

| modèle | T92 | T92+CpGs | GTR |
|---|---|---|---|
| nombre de paramètres | 2 | 3 | 8 |
| log-vrais. | $-3432$ | $-3389$ | $-3428$ |
| AIC | $6868$ | $6784$ | $6872$ |
| BIC | $6879$ | $6801$ | $6918$ |

$\rightarrow$ Algorithme pratique appliqué pour le modèle T92+CpGs.

# Autres applications

- Comparaison des approximations particulaires avec d'autres approximations composites
- Inférence d'un nucléotide à la racine
- Comparaison des estimateurs du maximum de vraisemblance

# Conclusion

$\rightarrow$ Étude d'un modèle d'évolution de l'ADN : RN95+YpR.

$\rightarrow$ Approximation consistante par approximations particulaires quand le nombre de particules tend vers l'infini de la vraisemblance pour des séquences issues de ces modèles.

$\rightarrow$ Mise en œuvre de cette approche.

$\rightarrow$ Applications.

Merci de votre attention !

# Références


- Jean Bérard and Laurent Guéguen, *Accurate estimation of substitution rates with neighbor-dependent models in a phylogenetic context*, Systematic Biology **61** (2012), no. 3, 510–521.
- Jean Bérard, Jean-Baptiste Gouéré, and Didier Piau, *Solvable models of neighbor-dependent substitution processes*, Mathematical Biosciences **211** (2008), no. 1, 56–88. MR 2392414 (2009h :92032)
- Julien Dutheil and Bastien Boussau, *Non-homogeneous models of sequence evolution in the Bio++ suite of libraries and programs*, BMC Evolutionary Biology **8** (2008), no. 1, 255.
- Julien Dutheil, Sylvain Gaillard, Eric Bazin, Sylvain Glémin, Vincent Ranwez, Nicolas Galtier, and Khalid Belkhir, *Bio++ : a set of C++ libraries for sequence analysis, phylogenetics, molecular evolution and population genetics*, BMC Bioinformatics **7** (2006), no. 1, 188.