Extension des modèles stochastiques de substitution de nucléotides : approche Kronecker

Équipe de Bioinformatique théorique,
Fouille de données et Optimisation stochastique
Laboratoire des Sciences de l'Image, de l'Informatique et de la Télédétection

2012

Plan

1. Modèles d'évolution stochastiques classiques
2. Modèle d'évolution analytique avec Kronecker
3. Logiciel de recherche : Stochastic Evolution of Genetic Motifs
4. Références bibliographiques

Plan

1 Modèles d'évolution stochastiques classiques

2 Modèle d'évolution analytique avec Kronecker

3 Logiciel de recherche : Stochastic Evolution of Genetic Motifs

4 Références bibliographiques

Notion de motif

- Alphabet génétique : $\mathcal{A} = \{A,C,G,T\}$
- $\mathcal{A}^n$: ensemble des motifs de longueur n sur $\mathcal{A}$
- 4^n motifs de longueur n
- i, motif de longueur n : dans $\mathcal{A}^n$

Exemple : dinucléotides

Motifs de $\mathcal{A}^2$

motif	AA	AC	AG	...	TT
ordre	1	2	3	...	16

$\rightarrow i = \text{AG} \Leftrightarrow i = 3$

Modélisation stochastique d'évolution des gènes

Principe de nos modèles stochastiques d'évolution de motifs
Exemple avec une séquence de dinucléotides :

Modèle physique	Modèle stochastique

Séquence init.	Probabilités d'occurrence initiales à $t = 0$
AA,AA,AA,AA,AA	$P_{AA}(0) = 1, P_{AC}(0), \ldots$

Modélisation stochastique d'évolution des gènes

Principe de nos modèles stochastiques d'évolution de motifs
Exemple avec une séquence de dinucléotides :

Modèle physique	Modèle stochastique

Séquence init.	Probabilités d'occurrence initiales à $t = 0$
AA,AA,AA,AA,AA	$P_{AA}(0) = 1, P_{AC}(0), \ldots$

Évolution	Probabilités d'occurrence à $t = 1, 2, \ldots$
AA,AA,A**C**,AA,AA	$P_{AA}(1) < 1, P_{AC}(1) > 0, P_{AG}(1) = 0$

Modélisation stochastique d'évolution des gènes

Principe de nos modèles stochastiques d'évolution de motifs
Exemple avec une séquence de dinucléotides :

Modèle physique	Modèle stochastique

Séquence init.	Probabilités d'occurrence initiales à $t = 0$
AA,AA,AA,AA,AA	$P_{AA}(0) = 1, P_{AC}(0), \ldots$

Évolution	Probabilités d'occurrence à $t = 1, 2, \ldots$
AA,AA,A**C**,AA,AA	$P_{AA}(1) < 1, P_{AC}(1) > 0, P_{AG}(1) = 0$
AA,**G**A,A**C**,AA,AA	$P_{AA}(2) < 1, P_{AC}(2) > 0, P_{GA}(2) > 0, P_{AG}(2) = 0$

Modélisation stochastique d'évolution des gènes

Principe de nos modèles stochastiques d'évolution de motifs
Exemple avec une séquence de dinucléotides :

Modèle physique	Modèle stochastique

Séquence init.	Probabilités d'occurrence initiales à $t = 0$
AA,AA,AA,AA,AA	$P_{AA}(0) = 1, P_{AC}(0), \ldots$

Évolution	Probabilités d'occurrence à $t = 1, 2, \ldots$
AA,AA,A**C**,AA,AA	$P_{AA}(1) < 1, P_{AC}(1) > 0, P_{AG}(1) = 0$
AA,**G**A,A**C**,AA,AA	$P_{AA}(2) < 1, P_{AC}(2) > 0, P_{GA}(2) > 0, P_{AG}(2) = 0$
AA,**G**A,A**C**,AA,**T**A	$P_{AA}(3) < 1, P_{AC}(3) > 0, P_{GA}(3) > 0, P_{TA}(3) > 0, P_{AG}(3) = 0$

Modélisation stochastique d'évolution des gènes

Principe de nos modèles stochastiques d'évolution de motifs
Exemple avec une séquence de dinucléotides :

Modèle physique	Modèle stochastique

Séquence init.	Probabilités d'occurrence initiales à $t = 0$
AA,AA,AA,AA,AA	$P_{AA}(0) = 1, P_{AC}(0), \ldots$

Évolution	Probabilités d'occurrence à $t = 1, 2, \ldots$
AA,AA,A**C**,AA,AA	$P_{AA}(1) < 1, P_{AC}(1) > 0, P_{AG}(1) = 0$
AA,**G**A,A**C**,AA,AA	$P_{AA}(2) < 1, P_{AC}(2) > 0, P_{GA}(2) > 0, P_{AG}(2) = 0$
AA,**G**A,A**C**,AA,**T**A	$P_{AA}(3) < 1, P_{AC}(3) > 0, P_{GA}(3) > 0, P_{TA}(3) > 0, P_{AG}(3) = 0$
AA,**A**A,A**C**,AA,**T**A	$P_{AA}(4) < 1, P_{AC}(4) > 0, P_{GA}(4) = 0, P_{TA}(4) > 0, P_{AG}(4) = 0$

Équation différentielle d'évolution

Notations

i, j : deux motifs de taille n

$P_{\Delta t}(j \to i)$: probabilité que j mute en i pendant Δt

$P_i(t)$: probabilité d'occurrence de i au temps t

Δt suffisamment petit pour avoir au plus une substitution pendant cet intervalle de temps

Probabilité d'occurrence du motif i au temps $t + \Delta t$

$$P_i(t + \Delta t) = \sum_j P_j(t) \times P_{\Delta t}(j \to i)$$

Équation différentielle d'évolution

Dérivation

$$P_i'(t) = \lim_{\Delta t \to 0} \frac{P_i(t + \Delta t) - P_i(t)}{\Delta t}$$

$$\boxed{P_i'(t) = \sum_j P_j(t) \times P(j \to i) - P_i(t)}$$

$P(j \to i)$: probabilité de substitution instantanée d'un motif j en un motif i

Équation différentielle d'évolution

$$P_i'(t) = \sum_j P_j(t) \times P(j \to i) - P_i(t)$$

$P_n(t) = [P_i(t)]_{i=1}^{4^n}$: vecteur colonne des probabilités d'occurrence des 4^n motifs de taille n

$\Downarrow$ Notation matricielle

$$P_n'(t) = M_n \cdot P_n(t) - P_n(t) = \underbrace{(M_n - I_n)}_{A_n} \cdot P_n(t)$$

M_n : matrice de substitution $(4^n, 4^n)$ telle que $m_{i,j} = P(j \to i)$

I_n : matrice identité $(4^n, 4^n)$

A_n : matrice des taux de substitution instantanée $(4^n, 4^n)$

Équation différentielle d'évolution

$$P_i'(t) = \sum_j P_j(t) \times P(j \to i) - P_i(t)$$

$P_n(t) = [P_i(t)]_{i=1}^{4^n}$: vecteur colonne des probabilités d'occurrence des 4^n motifs de taille n

$\Downarrow$ Notation matricielle

$$\boxed{P_n'(t) = M_n \cdot P_n(t) - P_n(t) = \underbrace{(M_n - I_n)}_{A_n} \cdot P_n(t)}$$

M_n : matrice de substitution $(4^n, 4^n)$ telle que $m_{i,j} = P(j \to i)$

I_n : matrice identité $(4^n, 4^n)$

A_n : matrice des taux de substitution instantanée $(4^n, 4^n)$

Matrice des taux de substitution instantanée

Structure de la matrice A_1 (nucléotides)

$$\begin{array}{c} \\ \mathrm{A} \\ \mathrm{C} \\ \mathrm{G} \\ \mathrm{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} \end{array} \\ \begin{pmatrix} * & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{2,1} & * & a_{2,3} & a_{2,4} \\ a_{3,1} & a_{3,2} & * & a_{3,4} \\ a_{4,1} & a_{4,2} & a_{4,3} & * \end{pmatrix} \end{array}$$

- Éléments diagonaux : $a_{i,i} = -\sum_{j=1}^{4} a_{j,i}$
- Ainsi les colonnes somment à 0

Matrice des taux de substitution instantanée

Structure de la matrice A_1 (nucléotides)

$$\begin{array}{c} \\ \mathrm{A} \\ \mathrm{C} \\ \mathrm{G} \\ \mathrm{T} \end{array} \begin{array}{c} \begin{array}{cccc} \mathrm{A} & \mathrm{C} & \mathrm{G} & \mathrm{T} \end{array} \\ \begin{pmatrix} * & a_{1,2} & a_{1,3} & a_{1,4} \\ a_{2,1} & * & a_{2,3} & a_{2,4} \\ a_{3,1} & a_{3,2} & * & a_{3,4} \\ a_{4,1} & a_{4,2} & a_{4,3} & * \end{pmatrix} \end{array}$$

- Éléments diagonaux : $a_{i,i} = -\sum_{j=1}^{4} a_{j,i}$
- Ainsi les colonnes somment à 0

Modèles de substitution de nucléotides utilisés

Modèle de substitution **1P** : Jukes et Cantor, 1969

$$\begin{pmatrix} * & \alpha & \alpha & \alpha \\ \alpha & * & \alpha & \alpha \\ \alpha & \alpha & * & \alpha \\ \alpha & \alpha & \alpha & * \end{pmatrix}$$

Toutes les substitutions ont un taux identique α

Modèle de substitution **2P** : Kimura, 1980

$$\begin{pmatrix} * & \beta & \alpha & \beta \\ \beta & * & \beta & \alpha \\ \alpha & \beta & * & \beta \\ \beta & \alpha & \beta & * \end{pmatrix}$$

Distinction entre les transitions α et les transversions β

Modèles de substitution de nucléotides utilisés

Modèle de substitution **3P** : Kimura,1981

$$\begin{pmatrix} * & \gamma & \alpha & \beta \\ \gamma & * & \beta & \alpha \\ \alpha & \beta & * & \gamma \\ \beta & \alpha & \gamma & * \end{pmatrix}$$

Distinction entre :

- les transitions α
- les transversions de type I β (A$\leftrightarrow$T et C$\leftrightarrow$G)
- les transversions de type II γ (A$\leftrightarrow$C et G$\leftrightarrow$T)

Notation des modèles de substitution

1P, 2P, 3P : Modèles de substitution de nucléotides

1PS, 2PS, 3PS : Modèles de substitution de motifs

Résolution analytique de l'équation différentielle d'évolution

Rappel : Équation différentielle d'évolution

$$P'_n(t) = A_n \cdot P_n(t)$$

Quand la matrice des taux de substitution instantanée A_n est diagonalisable :

$$A_n = Q_n \cdot D_n \cdot Q_n^{-1}$$

D_n : matrice diagonale $(4^n, 4^n)$ des valeurs propres

Q_n : matrice $(4^n, 4^n)$ des vecteurs propres

Q_n^{-1} : matrice inverse de Q_n

Résolution analytique de l'équation différentielle d'évolution

$$P'_n(t) = Q_n \cdot D_n \cdot Q_n^{-1} \cdot P_n(t)$$

$\Downarrow$ Solution classique (Lange, 2005)

$$\boxed{P_n(t) = Q_n \cdot e^{D_n \times t} \cdot Q_n^{-1} \cdot P_n(0)}$$

$e^{D_n \times t}$: matrice diagonale des exponentielles des valeurs propres de $A_n \times t$

$P_n(0)$: vecteur des probabilités d'occurrence initiales des 4^n motifs de taille n

Inversion du sens d'évolution

Sens d'évolution direct

$$P_n(t) = Q_n \cdot e^{D_n \times t} \cdot Q_n^{-1} \cdot P_n(0)$$

$P_n(0)$: Probabilités d'occurrence actuelles des motifs

$P_n(t)$: Probabilités d'occurrence futures des motifs

$\Downarrow$ Inversion

Sens d'évolution inverse

$$\tilde{P}_n(t) = Q_n \cdot e^{-D_n \times t} \cdot Q_n^{-1} \cdot \tilde{P}_n(0)$$

$\tilde{P}_n(t)$: Probabilités d'occurrence passées des motifs

$\tilde{P}_n(0)$: Probabilités d'occurrence actuelles des motifs

Comment déterminer les valeurs et vecteurs propres de A_n ?

Problème

Taille des matrices de substitution A_n

Exemple

Matrice de substitution A_5 :
$4^{10} = 1048576$ termes

Approche classique pour le modèle analytique 3PS

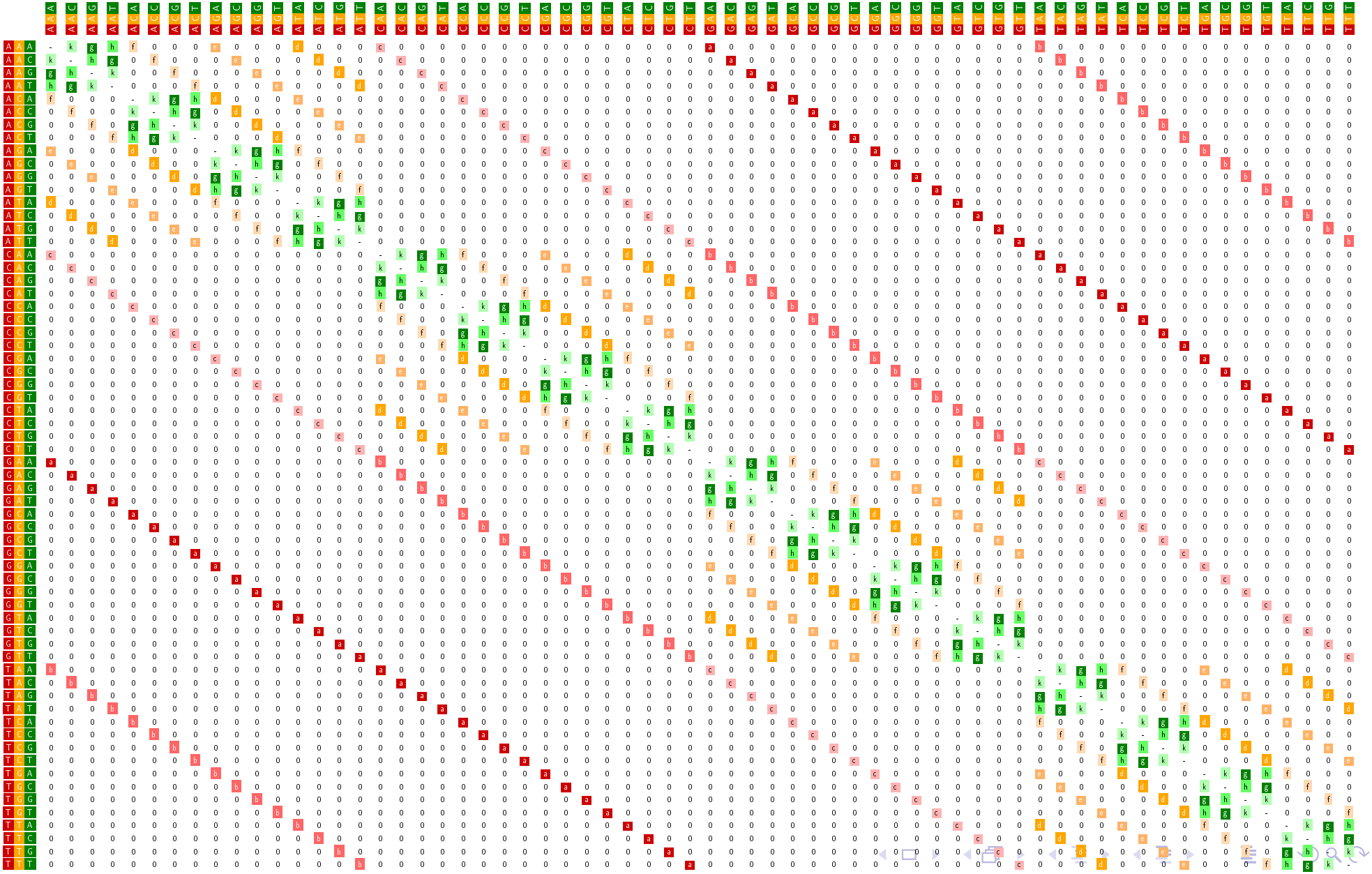
Propriétés de la matrice de substitution A_n de type 3PS

- Décomposable en blocs
- Décomposition particulière

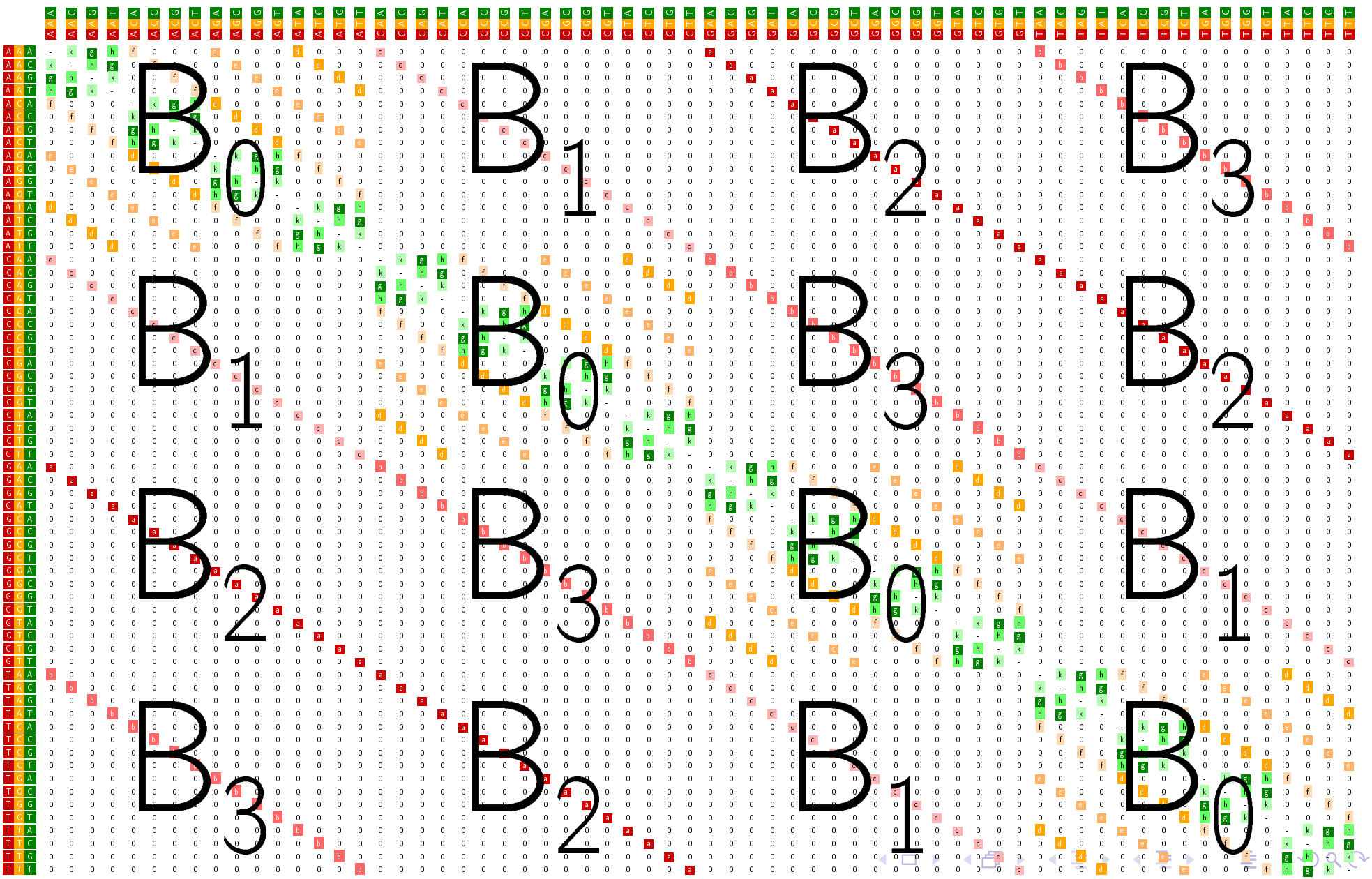
Approche utilisée par Frey et Michel (2006), Michel (2007)

Pour l'extension des modèles 2P et 3P aux dinucléotides et trinucléotides :
BMF : Factorisation par Matrices Blocs (Tian et Styan, 2001)

Exemple : matrice des taux de substitution instantanée A_3



Exemple : matrice des taux de substitution instantanée A_3



Factorisation de la matrice de substitution A_n par BMF

BMF : Factorisation par Matrices Blocs (Tian et Styan, 2001)

Structure en blocs de la matrice de substitution A_n

$$A_n = \begin{pmatrix} B_0 & B_1 & B_2 & B_3 \\ B_1 & B_0 & B_3 & B_2 \\ B_2 & B_3 & B_0 & B_1 \\ B_3 & B_2 & B_1 & B_0 \end{pmatrix}$$

Factorisation de la matrice de substitution A_n par BMF

$$A_n = F_{n-1} \cdot \begin{pmatrix} C_0 & 0 & 0 & 0 \\ 0 & C_1 & 0 & 0 \\ 0 & 0 & C_2 & 0 \\ 0 & 0 & 0 & C_3 \end{pmatrix} \cdot F_{n-1}$$

avec :

- $C_0 = B_0 + B_1 + B_2 + B_3$
- $C_1 = B_0 + B_1 - B_2 - B_3$
- $C_2 = B_0 - B_1 + B_2 - B_3$
- $C_3 = B_0 - B_1 - B_2 + B_3$

et

- $F_{n-1} = \frac{1}{2} \cdot \begin{pmatrix} I_{n-1} & I_{n-1} & I_{n-1} & I_{n-1} \\ I_{n-1} & I_{n-1} & -I_{n-1} & -I_{n-1} \\ I_{n-1} & -I_{n-1} & I_{n-1} & -I_{n-1} \\ I_{n-1} & -I_{n-1} & -I_{n-1} & I_{n-1} \end{pmatrix}$

Bilan de l'approche classique avec BMF

Avantage de la technique BMF

- Calcul des valeurs propres relativement simple

Inconvénients de la technique BMF

- Ne donne pas les vecteurs propres
- Spécifique au modèle 3PS

Plan

1 Modèles d'évolution stochastiques classiques

2 Modèle d'évolution analytique avec Kronecker

3 Logiciel de recherche : Stochastic Evolution of Genetic Motifs

4 Références bibliographiques

Approche Kronecker

Avec deux matrices carrées $X(m, m)$ et $Y(n, n)$

Produit de Kronecker

$$X \otimes Y = \begin{pmatrix} x_{1,1} \times Y & x_{1,2} \times Y & \dots & x_{1,m} \times Y \\ x_{2,1} \times Y & x_{2,2} \times Y & \dots & x_{2,m} \times Y \\ \vdots & \vdots & \ddots & \vdots \\ x_{m,1} \times Y & x_{m,2} \times Y & \dots & x_{m,m} \times Y \end{pmatrix}$$

Somme de Kronecker

$$X \oplus Y = X \otimes I_n + I_m \otimes Y$$

avec I_n et I_m deux matrices identité de tailles respectives (n, n) et (m, m)

Construction récursive de A_n de type 3PS avec Kronecker

Construcion récursive pour k de n à 1 :

$$A_k = \begin{pmatrix} A_{k-1} & c_{n-k+1} I_{k-1} & a_{n-k+1} I_{k-1} & b_{n-k+1} I_{k-1} \\ c_{n-k+1} I_{k-1} & A_{k-1} & b_{n-k+1} I_{k-1} & a_{n-k+1} I_{k-1} \\ a_{n-k+1} I_{k-1} & b_{n-k+1} I_{k-1} & A_{k-1} & c_{n-k+1} I_{k-1} \\ b_{n-k+1} I_{k-1} & a_{n-k+1} I_{k-1} & c_{n-k+1} I_{k-1} & A_{k-1} \end{pmatrix}$$

$$\Downarrow$$

$$A_k = \begin{pmatrix} 0 & c_{n-k+1} I_{k-1} & a_{n-k+1} I_{k-1} & b_{n-k+1} I_{k-1} \\ c_{n-k+1} I_{k-1} & 0 & b_{n-k+1} I_{k-1} & a_{n-k+1} I_{k-1} \\ a_{n-k+1} I_{k-1} & b_{n-k+1} I_{k-1} & 0 & c_{n-k+1} I_{k-1} \\ b_{n-k+1} I_{k-1} & a_{n-k+1} I_{k-1} & c_{n-k+1} I_{k-1} & 0 \end{pmatrix} \oplus A_{k-1}$$

Construction récursive de A_n de type 3PS avec Kronecker

$$\boxed{A_n = \oplus_{k=1}^{n} N_k}$$

Avec N_k la matrice de substitution de nucléotides de type 3P associée au site k :

$$N_k = \begin{pmatrix} s_k & c_k & a_k & b_k \\ c_k & s_k & b_k & a_k \\ a_k & b_k & s_k & c_k \\ b_k & a_k & c_k & s_k \end{pmatrix}$$

$$s_k = -(a_k + b_k + c_k)$$

Calcul des valeurs et vecteurs propres de la matrice A_n (3PS)

Diagonalisation : matrice de substitution de nucléotides N_k (3P)

$$N_k = R \cdot S_k \cdot R^{-1}$$

avec S_k la matrice diagonale (4,4) des valeurs propres de N_k :

$$S_k = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & -2(a_k + b_k) & 0 & 0 \\ 0 & 0 & -2(a_k + c_k) & 0 \\ 0 & 0 & 0 & -2(b_k + c_k) \end{pmatrix}$$

et R la matrice (4,4) des vecteurs propres de N_k :

$$R = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \\ 1 & -1 & 1 & -1 \end{pmatrix}$$

Calcul des valeurs et vecteurs propres de la matrice A_n (3PS)

Valeurs propres

D_n matrice diagonale des valeurs propres de A_n :

$$D_n = \oplus_{k=1}^{n} S_k$$

Vecteurs propres

Q_n matrice des vecteurs propres de A_n :

$$Q_n = \otimes_{k=1}^{n} R, \text{ et } Q_n^{-1} = \otimes_{k=1}^{n} R^{-1}$$

$$\Downarrow$$

Matrice des taux de substitution instantanée

$$A_n = \otimes_{k=1}^{n} R \cdot \oplus_{k=1}^{n} S_k \cdot \otimes_{k=1}^{n} R^{-1}$$

Solutions analytiques avec l'approche Kronecker

Probabilités d'occurrence analytiques des motifs de taille n

$$\boxed{P_n(t) = \otimes_{k=1}^{n}(R \cdot e^{S_k t} \cdot R^{-1}) \cdot P_n(0)}$$

R : matrice (4,4) des vecteurs propres de N_k

$e^{S_k t}$: matrice diagonale (4,4) des exponentielles des valeurs propres de N_k

R^{-1} : matrice inverse (4,4) de R

$P_n(0)$: vecteur des probabilités d'occurrence initiales des 4^n motifs de taille n

Bilan de l'approche Kronecker

Avantages de l'approche Kronecker

- Calcul rapide des valeurs propres
- Calcul rapide des vecteurs propres
- Valable pour des matrices de substitution A_n différentes du type 3PS

Solution pour un motif i_1 donné de taille n avec 3PS

- Calculs longs pour les motifs de grande taille
- Cas où l'on ne s'intéresse qu'à l'évolution de certains motifs

$$P_{i_1}(t) = \frac{1}{4^n} \sum_{i_2=1}^{4^n} e^{\sum_{k=1}^{n} L_k[\delta(i_2,k)]t} \times \sum_{i_3=1}^{4^n} \left(\prod_{k=1}^{n} \left(R\left[\delta(i_1,k), \delta(i_2,k)\right] \times R\left[\delta(i_2,k), \delta(i_3,k)\right] \right) \times P_{i_3}(0) \right)$$

Avec :

- $L_k = \{0, -2(a_k + b_k), -2(a_k + c_k), -2(b_k + c_k)\}$
- R, matrice (4,4) des vecteurs propres de la matrice de substitution de nucléotides (3P)
- $\delta(i_1, k) = \lfloor \frac{i_1 - 1}{4^{n-k}} \rfloor [4] + 1$

Solution pour un motif i_1 donné de taille n avec 3PS

- Calculs longs pour les motifs de grande taille
- Cas où l'on ne s'intéresse qu'à l'évolution de certains motifs

$$P_{i_1}(t) = \frac{1}{4^n} \sum_{i_2=1}^{4^n} e^{\sum_{k=1}^{n} L_k[\delta(i_2,k)]t} \times \sum_{i_3=1}^{4^n} \left(\prod_{k=1}^{n} \left(R\left[\delta(i_1,k), \delta(i_2,k)\right] \times R\left[\delta(i_2,k), \delta(i_3,k)\right] \right) \times P_{i_3}(0) \right)$$

Avec :

- $L_k = \{0, -2(a_k + b_k), -2(a_k + c_k), -2(b_k + c_k)\}$
- R, matrice (4,4) des vecteurs propres de la matrice de substitution de nucléotides (3P)
- $\delta(i_1, k) = \lfloor \frac{i_1 - 1}{4^{n-k}} \rfloor [4] + 1$

Solution pour un motif i_1 donné de taille n avec 3PS

- Calculs longs pour les motifs de grande taille
- Cas où l'on ne s'intéresse qu'à l'évolution de certains motifs

$$P_{i_1}(t) = \frac{1}{4^n} \sum_{i_2=1}^{4^n} e^{\sum_{k=1}^{n} L_k[\delta(i_2,k)]t} \times \sum_{i_3=1}^{4^n} \left(\prod_{k=1}^{n} \left(R\left[\delta(i_1,k), \delta(i_2,k)\right] \times R\left[\delta(i_2,k), \delta(i_3,k)\right] \right) \times P_{i_3}(0) \right)$$

Avec :

- $L_k = \{0, -2(a_k + b_k), -2(a_k + c_k), -2(b_k + c_k)\}$
- R, matrice (4,4) des vecteurs propres de la matrice de substitution de nucléotides (3P)
- $\delta(i_1, k) = \lfloor \frac{i_1 - 1}{4^{n-k}} \rfloor [4] + 1$

Exemple : solution analytique pour le dinucléotide AG

$$
\begin{aligned}
P_{AG}(t) = P_3(t) = \frac{1}{16}\, & e^{0}(P_1(0) + P_2(0) + P_3(0) + P_4(0) + P_5(0) + P_6(0) + P_7(0) + P_8(0) + \\
& P_9(0) + P_{10}(0) + P_{11}(0) + P_{12}(0) + P_{13}(0) + P_{14}(0) + P_{15}(0) + P_{16}(0)) \\
& + e^{-2(a_2+b_2)t}(-P_1(0) - P_2(0) + P_3(0) + P_4(0) - P_5(0) - P_6(0) + P_7(0) + P_8(0) \\
& - P_9(0) - P_{10}(0) + P_{11}(0) + P_{12}(0) - P_{13}(0) - P_{14}(0) + P_{15}(0) + P_{16}(0)) \\
& + e^{-2(a_2+c_2)t}(-P_1(0) + P_2(0) + P_3(0) - P_4(0) - P_5(0) + P_6(0) + P_7(0) - P_8(0) \\
& - P_9(0) + P_{10}(0) + P_{11}(0) - P_{12}(0) - P_{13}(0) + P_{14}(0) + P_{15}(0) - P_{16}(0)) \\
& + e^{-2(b_2+c_2)t}(P_1(0) - P_2(0) + P_3(0) - P_4(0) + P_5(0) - P_6(0) + P_7(0) - P_8(0) \\
& + P_9(0) - P_{10}(0) + P_{11}(0) - P_{12}(0) + P_{13}(0) - P_{14}(0) + P_{15}(0) - P_{16}(0)) \\
& + e^{-2(a_1+b_1)t}(P_1(0) + P_2(0) + P_3(0) + P_4(0) + P_5(0) + P_6(0) + P_7(0) + P_8(0) \\
& - P_9(0) - P_{10}(0) - P_{11}(0) - P_{12}(0) - P_{13}(0) - P_{14}(0) - P_{15}(0) - P_{16}(0)) \\
& + e^{-2(a_1+b_1+a_2+b_2)t}(-P_1(0) - P_2(0) + P_3(0) + P_4(0) - P_5(0) - P_6(0) + P_7(0) + P_8(0) \\
& + P_9(0) + P_{10}(0) - P_{11}(0) - P_{12}(0) + P_{13}(0) + P_{14}(0) - P_{15}(0) - P_{16}(0)) \\
& + e^{-2(a_1+b_1+a_2+c_2)t}(-P_1(0) + P_2(0) + P_3(0) - P_4(0) - P_5(0) + P_6(0) + P_7(0) - P_8(0) \\
& + P_9(0) - P_{10}(0) - P_{11}(0) + P_{12}(0) + P_{13}(0) - P_{14}(0) - P_{15}(0) + P_{16}(0)) \\
& + e^{-2(a_1+b_1+b_2+c_2)t}(P_1(0) - P_2(0) + P_3(0) - P_4(0) + P_5(0) - P_6(0) + P_7(0) - P_8(0) \\
& - P_9(0) + P_{10}(0) - P_{11}(0) + P_{12}(0) - P_{13}(0) + P_{14}(0) - P_{15}(0) + P_{16}(0)) \\
& + \ldots \\
& + e^{-2(b_1+c_1+b_2+c_2)t}(P_1(0) - P_2(0) + P_3(0) - P_4(0) - P_5(0) + P_6(0) - P_7(0) + P_8(0) \\
& + P_9(0) - P_{10}(0) + P_{11}(0) - P_{12}(0) - P_{13}(0) + P_{14}(0) - P_{15}(0) + P_{16}(0))]
\end{aligned}
$$

Exemple : solution analytique pour le dinucléotide AG

$$
\begin{aligned}
P_{AG}(t) = P_3(t) = \frac{1}{16}\, & e^{0}(P_1(0) + P_2(0) + P_3(0) + P_4(0) + P_5(0) + P_6(0) + P_7(0) + P_8(0) + \\
& P_9(0) + P_{10}(0) + P_{11}(0) + P_{12}(0) + P_{13}(0) + P_{14}(0) + P_{15}(0) + P_{16}(0)) \\
& + e^{-2(a_2+b_2)t}(-P_1(0) - P_2(0) + P_3(0) + P_4(0) - P_5(0) - P_6(0) + P_7(0) + P_8(0) \\
& - P_9(0) - P_{10}(0) + P_{11}(0) + P_{12}(0) - P_{13}(0) - P_{14}(0) + P_{15}(0) + P_{16}(0)) \\
& + e^{-2(a_2+c_2)t}(-P_1(0) + P_2(0) + P_3(0) - P_4(0) - P_5(0) + P_6(0) + P_7(0) - P_8(0) \\
& - P_9(0) + P_{10}(0) + P_{11}(0) - P_{12}(0) - P_{13}(0) + P_{14}(0) + P_{15}(0) - P_{16}(0)) \\
& + e^{-2(b_2+c_2)t}(P_1(0) - P_2(0) + P_3(0) - P_4(0) + P_5(0) - P_6(0) + P_7(0) - P_8(0) \\
& + P_9(0) - P_{10}(0) + P_{11}(0) - P_{12}(0) + P_{13}(0) - P_{14}(0) + P_{15}(0) - P_{16}(0)) \\
& + e^{-2(a_1+b_1)t}(P_1(0) + P_2(0) + P_3(0) + P_4(0) + P_5(0) + P_6(0) + P_7(0) + P_8(0) \\
& - P_9(0) - P_{10}(0) - P_{11}(0) - P_{12}(0) - P_{13}(0) - P_{14}(0) - P_{15}(0) - P_{16}(0)) \\
& + e^{-2(a_1+b_1+a_2+b_2)t}(-P_1(0) - P_2(0) + P_3(0) + P_4(0) - P_5(0) - P_6(0) + P_7(0) + P_8(0) \\
& + P_9(0) + P_{10}(0) - P_{11}(0) - P_{12}(0) + P_{13}(0) + P_{14}(0) - P_{15}(0) - P_{16}(0)) \\
& + e^{-2(a_1+b_1+a_2+c_2)t}(-P_1(0) + P_2(0) + P_3(0) - P_4(0) - P_5(0) + P_6(0) + P_7(0) - P_8(0) \\
& + P_9(0) - P_{10}(0) - P_{11}(0) + P_{12}(0) + P_{13}(0) - P_{14}(0) - P_{15}(0) + P_{16}(0)) \\
& + e^{-2(a_1+b_1+b_2+c_2)t}(P_1(0) - P_2(0) + P_3(0) - P_4(0) + P_5(0) - P_6(0) + P_7(0) - P_8(0) \\
& - P_9(0) + P_{10}(0) - P_{11}(0) + P_{12}(0) - P_{13}(0) + P_{14}(0) - P_{15}(0) + P_{16}(0)) \\
& + \ldots \\
& + e^{-2(b_1+c_1+b_2+c_2)t}(P_1(0) - P_2(0) + P_3(0) - P_4(0) - P_5(0) + P_6(0) - P_7(0) + P_8(0) \\
& + P_9(0) - P_{10}(0) + P_{11}(0) - P_{12}(0) - P_{13}(0) + P_{14}(0) - P_{15}(0) + P_{16}(0))]
\end{aligned}
$$

Plan

1 Modèles d'évolution stochastiques classiques

2 Modèle d'évolution analytique avec Kronecker

3 Logiciel de recherche : Stochastic Evolution of Genetic Motifs

4 Références bibliographiques

Stochastic Evolution of Genetic Motifs

- Calcul des probabilités d'occurrence analytiques de motifs génétiques
- Courbes d'évolution
- Motifs de taille 1 à 5
- Sens d'évolution direct (présent-futur) et inverse (présent-passé)
- Solutions analytiques formelles et numériques
- Modèles d'évolution disponibles : 1PS, 2PS et 3PS

n	**nom**	**taille solutions analytiques formelles**
1	nucléotides	$4^{2\times 1} = 16$
2	*di*nucléotides	$4^{2\times 2} = 256$
3	*tri*nucléotides	$4^{2\times 3} = 4096$
4	*tetra*nucléotides	$4^{2\times 4} = 65536$
5	*penta*nucléotides	$4^{2\times 5} = 1048576$

Page d'accueil

Stochastic Evolution of Genetic Motifs

Emmanuel Benard and Christian J. Michel

Theoretical Bioinformatics, LSIIT/CNRS UMR7005 - University of Strasbourg

---

1. Choose the motif size:

Nucleotides (1)

---

2. Upload the initial occurrence probabilities file:

Enter a XLS file containing the 4 initial occurrence probabilities of motifs of size 1:

Parcourir...

Example of a valid XLS file containing 4 initial occurrence probabilities of Nucleotides available here

SUBMIT

Fonctionnalités

Stochastic Evolution of Trinucleotides HOME

Emmanuel Benard and Christian J. Michel

Theoretical Bioinformatics, LSIIT/CNRS UMR7005 - University of Strasbourg

---

Upload new initial occurrence probabilities?

Uploaded file informations:

✓ Initial occurrence probabilities file valid

---

1. Evolutionary time direction:

Inverse (present -> past)

---

2. Number of substitution parameters per motif site:

3 parameters: 1 transition rate (A↔G = C↔T), 1 transversion I rate (A↔T = C↔G), 1 transversion II rate (A↔C = G↔T).

2 parameters: 1 transition rate (A↔G = C↔T), 1 transversion rate (A↔T = A↔C = C↔G = G↔T).
u[x]=a[x], v[x]/2=b[x]=c[x]

1 parameter: 1 substitution rate (A↔C = A↔G = A↔T = C↔G = C↔T = G↔T).
p[x]/3=a[x]=b[x]=c[x]

3 parameters More about mutation matrices and substitution parameters

---

Fonctionnalités

3. Substitution parameters:

Enter values for the substitution parameters.
Non numerical or rational values will be replaced by the name of the corresponding parameter.

All the substitution parameters must have a numerical value to get plots.
All the substitution parameters and their sum must be ≥ 0 and < 1.

Site 0	a[0]: a0	b[0]: b0	c[0]: c0
Site 1	a[1]: a1	b[1]: b1	c[1]: c1
Site 2	a[2]: a2	b[2]: b2	c[2]: c2

Substitution parameters statut:

Parameters sum = "a0" + "a1" + "a2" + "b0" + "b1" + "b2" + "c0" + "c1" + "c2"

---

4. Choice of the probabilities to study and plot:

Choose up to 4 analytical solutions.
By default, only the analytical solution of the motif AAA is displayed and plotted.

motif AAA | ------ | ------ | ------

4b. Choice of the analytical solutions output format:

The analytical solutions can be displayed in 4 formats: Standard, C, Fortran and TeX.
By default, the analytical solutions are displayed in Standard format.

Standard

SUBMIT

Sorties

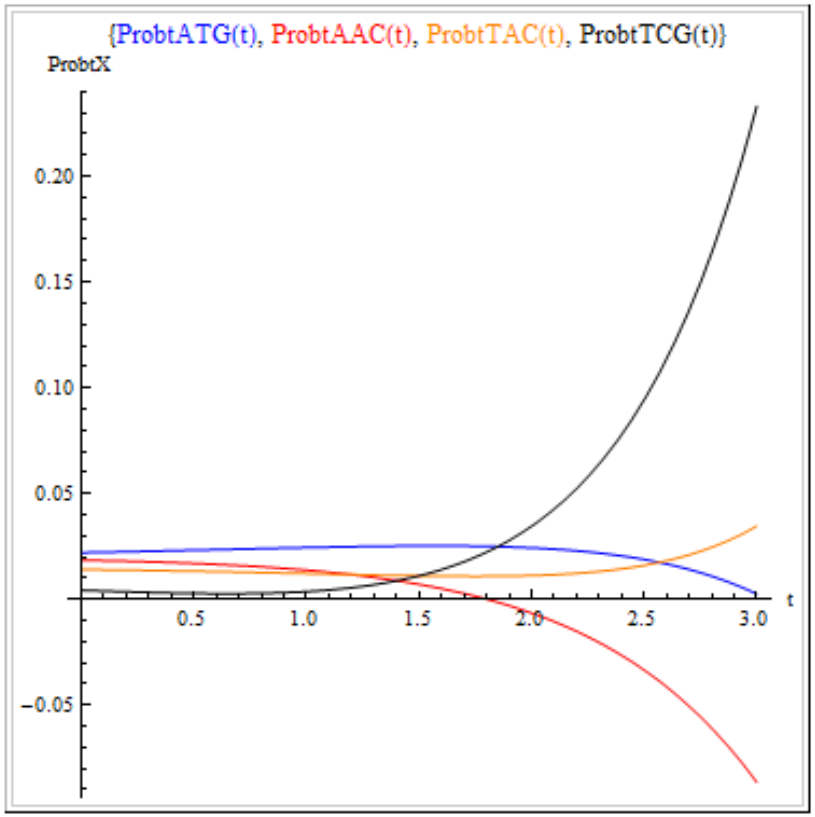
Results

Analytical solutions (Standard format):

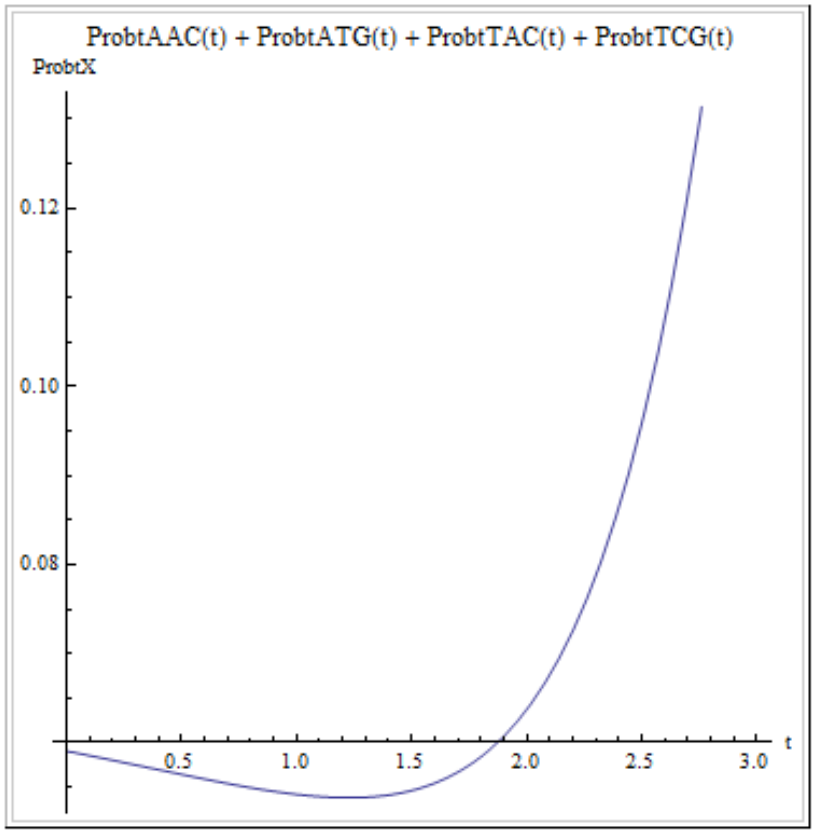
ProbtATG(t)	0.015625 + 0.000540842/E^2*(a0 + b0)*t - 0.00152927/E^2*(a1 + b1)*t - 0.0000528894 /E^2*(a0 + a1 + b0 + b1)*t + 0.000469227/E^2*(a0 + a2 + b0 + b2)*t + 0.00120317/E^2*(a1 + a2 + b1 + b2)*t + 0.00125916 /E^2*(a0 + a1 + b1 + c0)*t - 0.00117735/E^2*(a1 + b0 + b1 + c0)*t - 0.000459652/E^2*(a0 + a2 + b2 + c0)*t - 0.000696165 /E^2*(a2 + b0 + b2 + c0)*t + 0.00100502/E^2*(a0 + a1 + b0 + c1)*t + 0.000166589/E^2*(a0 + b0 + b1 + c1)*t + 0.00231898 /E^2*(a1 + a2 + b2 + c1)*t - 0.000283983/E^2*(a2 + b1 + b2 + c1)*t - 0.000835963/E^2*(a0 + a1 + c0 + c1)*t + 0.000572985 /E^2*(a1 + b0 + c0 + c1)*t + 0.000552281/E^2*(a0 + b1 + c0 + c1)*t - 0.00263804/E^2*(b0 + b1 + c0 + c1)*t + 0.000495397 /E^2*(a0 + a2 + b0 + c2)*t + 0.00157556/E^2*(a1 + a2 + b1 + c2)*t + 0.00121722/E^2*(a0 + b0 + b2 + c2)*t + 0.00013649 /E^2*(a1 + b1 + b2 + c2)*t - 0.000945947/E^2*(a0 + a2 + c0 + c2)*t - 0.000162183/E^2*(a2 + b0 + c0 + c2)*t - 0.00164162 /E^2*(a0 + b2 + c0 + c2)*t + 0.000777826/E^2*(b0 + b2 + c0 + c2)*t + 0.00176303/E^2*(a1 + a2 + c1 + c2)*t - 0.000420078 /E^2*(a2 + b1 + c1 + c2)*t + 0.00184492/E^2*(a1 + b2 + c1 + c2)*t - 0.00137639 /E^2*(b1 + b2 + c1 + c2)*t + (0.00036714 - 0.000178117/E^2*(a0 + a1 + b0 + b1)*t)/E^2*(a2 + b2)*t + (0.0026192 - 0.000264094 /E^2*(a1 + a2 + b1 + b2)*t)/E^2*(b0 + c0)*t + (-0.00191735 + 0.0000447409/E^2*(a1 + a2 + b1 + b2)*t) /E^2*(a0 + c0)*t + (-0.0000772344 + 0.0000866094/E^2*(a0 + a2 + b0 + b2)*t - 0.000258935 /E^2*(a0 + a2 + b2 + c0)*t + 0.000361008/E^2*(a2 + b0 + b2 + c0)*t)/E^2*(b1 + c1)*t + (0.00230261 + 0.000637553 /E^2*(a0 + a2 + b0 + b2)*t - 0.00112972/E^2*(a0 + a2 + b2 + c0)*t + 0.00052303/E^2*(a2 + b0 + b2 + c0)*t) /E^2*(a1 + c1)*t + (-0.000512409 - 0.000698465/E^2*(a0 + a1 + b0 + b1)*t - 0.00038629/E^2*(a0 + a1 + b1 + c0)*t - 0.000655334 /E^2*(a1 + b0 + b1 + c0)*t + 0.00061581/E^2*(a0 + a1 + b0 + c1)*t - 0.000362639/E^2*(a0 + b0 + b1 + c1)*t - 0.00126494 /E^2*(a0 + a1 + c0 + c1)*t + 0.000910484/E^2*(a1 + b0 + c0 + c1)*t + 0.000596409/E^2*(a0 + b1 + c0 + c1)*t - 0.000747581 /E^2*(b0 + b1 + c0 + c1)*t)/E^2*(b2 + c2)*t + (0.00224861 + 0.0000702462/E^2*(a0 + a1 + b0 + b1)*t - 0.00009901 /E^2*(a0 + a1 + b1 + c0)*t - 0.000181843/E^2*(a1 + b0 + b1 + c0)*t + 0.000928941/E^2*(a0 + a1 + b0 + c1)*t + 0.0000760541 /E^2*(a0 + b0 + b1 + c1)*t - 0.000931705/E^2*(a0 + a1 + c0 + c1)*t - 0.000225394/E^2*(a1 + b0 + c0 + c1)*t - 0.000123967 /E^2*(a0 + b1 + c0 + c1)*t + 0.000407291/E^2*(b0 + b1 + c0 + c1)*t)/E^2*(a2 + c2)*t

Sorties

Plots:



Plot sum:



SEGM

http://lsiit-bioinfo.u-strasbg.fr:8080/webMathematica/SEGM/SEGM.html

Plan

1 Modèles d'évolution stochastiques classiques

2 Modèle d'évolution analytique avec Kronecker

3 Logiciel de recherche : Stochastic Evolution of Genetic Motifs

4 Références bibliographiques

- E. Benard, C.J. Michel. *Computation of direct and inverse mutations with the SEGM web server (Stochastic Evolution of Genetic Motifs) : an application to splice sites of human genome introns*. Computational Biology and Chemistry, Vol. 33, p. 245-252, 2009.
- E. Benard, C.J. Michel. *A generalization of substitution evolution models of nucleotides to genetic motifs.* Journal of Theoretical Biology, Vol. 288, p. 73-83, 2011.
- C.J. Michel. *Evolution probabilities and phylogenetic distance of dinucleotides.* Journal of Theoretical Biology, Vol. 249, p. 271-277, 2007.
- C.J. Michel. *Codon phylogenetic distance.* Journal of Computational Biology and Chemistry, Vol. 31, p. 36-43, 2007.
- C.J. Michel. *An analytical model of gene evolution with 9 mutation parameters : an application to the amino acids coded by the common circular code.* Bulletin of Mathematical Biology, Vol. 69, p. 677-698, 2007.
- G. Frey, C.J. Michel. *An analytical model of gene evolution with 6 mutation parameters : an application to archaeal circular codes.* Journal of Computational Biology and Chemistry, Vol. 30, p. 1-11, 2006.