Journal de la Société Française de Statistique
Vol. 153 No. 2 (2012)

Chaînes de Markov et absorption. Application à l'algorithme de Fu en génomique

Title: Absorption of Markov chains. Application to Fu algorithm in genomics

Bernard Prum[1]

Résumé : Cet article est motivé par la recherche de mots ou de motifs exceptionnellement rares ou exceptionnellement présents dans une séquence d'ADN chromosomique. Cette approche permettra en effet de découvrir des motifs ayant un rôle biologique néfaste ou bénéfique pour l'organisme qui le porte. On modélise alors la séquence par une chaîne de Markov (CM) et l'approche classique cherche l'espérance et la variance du nombre $N(W)$ d'occurences du mot W. Nous développons ici une approche duale, déterminant l'espérance et la variance du temps $T(W)$ entre deux occurrences de W. Ceci s'appuie sur une CM auxilliaire dont les états sont les préfixes de W et $T(W)$ est alors le temps que met cette CM pour atteindre le mot complet W. L'étude de l'absoption d'une CM est, pour ce faire, présentée en détail.

Abstract: The motivation of this paper is the research of exceptionaly frequent (or rare) motifs or words W in a chromosomal DNA sequence : this approach often allows the identification of motifs playing a beeficial (or harmful) role for the organism carrying this chromosome. The sequence is modelized by a Markov chain (MC), and the classical approach consists in computing the expectation and the variance of the number $N(W)$ of occurrences of W. Here we develop a dual point of view dealing with the expectation and the variance of the time $T(W)$ between two occurences of W. This is done by introducing an auxiliary MC whose states are the prefixs of W, and $T(W)$ is the time this new MC needs to reach the complete word W. A detailed presentation of the absoption for a finite MC is therefore presented.

Mots-clés : Absorption de chaÃ®nes de Markov, Mots exceptionnels, Algorithme de Fu

Keywords: Absorption of Matkov chains, Exceptional words, Fu algorithm

Classification AMS 2000 : 60J10, 62P10, 92D99

Les résultats les plus connus et les plus utilisés sur les chaînes de Markov (CM) concernent les CM "récurrentes" (voir définition au paragraphe 1) : recherche d'un régime stationnaire, convergence vers celui-ci, etc. Dans cet article, nous considérons un problème moins connu, celui de l'absorption par un état absorbant – et tout particulièrement celui du temps (aléatoire) nécessaire pour cette absorption.

Nous donnons à la fin de l'article quelques exemples d'applications, en particulier celui qui nous a motivés : le temps d'attente nécessaire pour voir apparaître dans la réalisation d'une CM un mot donné W. Ce problème, qui se pose dans l'étude des génomes, se résout en introduisant une CM auxilliaire (dite "embedded" Markov chain), comme il est expliqué au paragraphe 5.

[1] genopole Université d'Evry.
E-mail : bernard.prum@genopole.cnrs.fr

 ISSN: 2102-6238

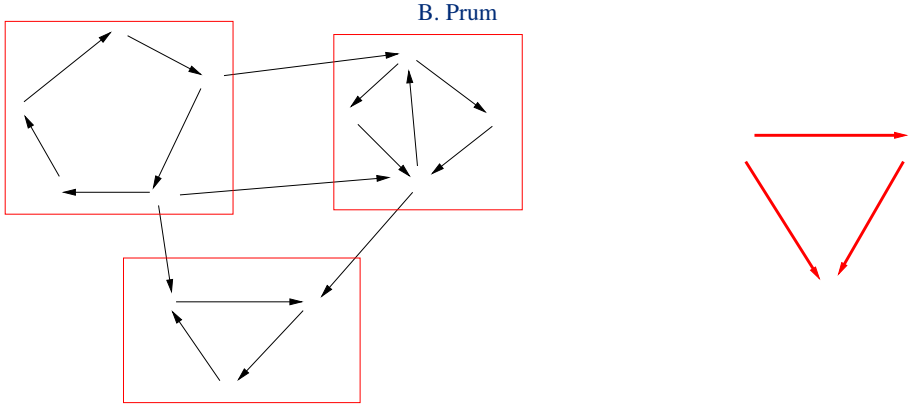

FIGURE 1. *À gauche, un exemple de graphe (seules les flèches ont été tracées) ; les classes d'équivalence ont été indiquées en rouge. À droite on a tracé le graphe quotient.*

1. Classification des états

Soit E un ensemble fini, composé de n "états" $\{1,2,\dots,n\}$ et X une CM sur E définie par :
• sa loi initiale, que l'on choisira ici dégénérée (X_0 fixé) ;
• sa matrice de transition :

$$\pi(u,v) = \mathbb{P}(X_{k+1} = v \mid X_k = u)$$

À une CM on associe un graphe orienté, dont les sommets sont les états ; il y a une flèche du sommet u vers le sommet v si et seulement si $\pi(u,v) > 0$.

On dit que la CM est *irréductible* si son graphe comporte une seule composante connexe (sans prendre en compte l'orientation des flèches). Nous supposerons désormais toujours la CM irréductible (sinon, on traite chaque composante connexe séparément).

Définition 1. On dit que u mène à v (et l'on note $u \longrightarrow v$) s'il existe une suite de sommets $s_0 = u, s_1, \dots, s_{k-1}, s_k = v$ telle que pour tout $j \in \{1,k\}$, on ait $\pi(s_{j-1}, s_j) > 0$.

On définit la relation d'équivalence entre états :

$$u \sim v \quad \Longleftrightarrow \quad u \longrightarrow v \text{ et } v \longrightarrow u$$

On dira alors qu'il existe une *boucle* passant par u et v.

On notera $\tilde{E}$ l'ensemble quotient ; on définit un graphe orienté sur $\tilde{E}$ en traçant une flèche de $U \in \tilde{E}$ vers $V \in \tilde{E}$ si et seulement si :

$$\exists u \in U\ , \exists v \in V, \qquad \pi(u,v) > 0$$

Le graphe ainsi obtenu ne comporte pas de boucle.

Les CM (sur un espace fini) se divisent alors en deux types :
• celles pour lesquelles le graphe quotient se réduit à un seul point. Tout sommet mène à tout sommet, deux sommets quelconques sont sur une même boucle. Partant de n'importe quel état, la CM passera alors une infinité de fois par chacun des états (cf. Borel-Cantelli). Les états sont

 ISSN: 2102-6238

qualifiés de *récurrents* (et la CM de récurrente). C'est (peut-être) le type de CM le plus souvent considéré [1].

• les autres − auxquelles nous nous intéressons ici. Le graphe quotient associé, dépourvu de boucle, contient donc des sommets (i.e. des classes) dont ne part aucune flêche. On dira qu'il s'agit de "classes absorbantes". Les autres classes seront qualifiées de "transitoires"

Définition 2. On appelle état transitoire les états appartenant à un classe transitoire.

On appelle état absorbant un état constituant à lui seul une classe absorbante. Un état $w \in E$ est donc *absorbant* si et seulement si $\pi(w,w) = 1$.

L'exemple de la figure 1 contient une seule classe absorbante, qui est composée de trois états, qui ne sont donc pas (chacun) absorbants.

Une CM peut avoir plusieurs états absorbants ; on notera A l'ensemble des états absorbants, s leur nombre et $t = n - s$ le nombre d'états transitoires :

Exemple 3. Pour la CM de transition :

$$\pi = \begin{pmatrix} .5 & .4 & .1 & 0 & 0 \\ .5 & .4 & 0 & 0 & .1 \\ 0 & .1 & .7 & .1 & .1 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

les états 4 et 5 sont absorbants.

2. Temps d'absorption

Plaçons nous désormais dans une CM dont les états se répartissent en états absorbants et états transitoires.

Problème : Partant de $X_0 = u$ transitoire, appelons T le nombre de pas jusqu'à absorption ? ($T = \inf\{k, X_k \in A\}$) ;
1) Quelle est l'espérance de T ?
2) Quelle est la loi de T ?
3) Si A comporte plus d'un élément, disons $A = \{w_1, \dots w_s\}$, quelle est la probabilité que la CM soit absorbée par chacun des w_j ?
4) Conditionnellement au fait que la CM est absorbée par w_i, quelle est la loi de T, l'espérance de $T, \dots$

Quitte à ré-ordonner les points de E, on peut toujours supposer que les points absorbants sont les derniers : $A = \{t+1, \dots, n\}$.

[1] la CM induite, dans le cas général, sur chaque "classe absorbante" est elle aussi une CM récurrente.

 ISSN: 2102-6238

On peut alors découper la matrice π en quatre sous-matrices :

$$\begin{array}{rl} Q = \pi(u,v) & \text{si } u \text{ et } v \text{ sont transitoires} \\ P = \pi(u,w) & \text{si } u \text{ est transitoire et } w \text{ absorbant} \\ 0 = \pi(w,v) & \text{si } w \text{ est absorbant et } v \text{ transitoire} \\ I = \pi(w,w') & \text{si } w \text{ et } w' \text{ sont absorbants.} \end{array}$$

et l'on peut adopter la notation par blocs : $\pi = \begin{pmatrix} Q & P \\ 0 & I \end{pmatrix}$.

Exemple 4. Suite de l'exemple 3

$$\pi = \left(\begin{array}{ccc|cc} .5 & .4 & .1 & .0 & 0 \\ .5 & .4 & 0 & 0 & .1 \\ 0 & .1 & .7 & .1 & .1 \\ \hline 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{array}\right)$$

Q est une matrice $t \times t$ sous stochastique ($Q(u,v) \geq 0$; $\sum_v Q(u,v) \leq 1$) ;
P est une matrice $t \times s$ sous stochastique ;
0 est la matrice $s \times t$ composée de zéros ;
I est la matrice unitaire [2] $s \times s$.

2.1. Résultat préliminaire

On sait que $\mathbb{P}(X_k = v \mid X_0 = u) = \pi^k(u,v)$.

Propriété 5. Les puissances de π sont données par :

$$\pi^2 = \begin{pmatrix} Q^2 & P+QP \\ 0 & I \end{pmatrix} \qquad \pi^3 = \begin{pmatrix} Q^3 & P+QP+Q^2P \\ 0 & I \end{pmatrix}$$

et, de façon générale :

$$\pi^k = \begin{pmatrix} Q^k & P+QP+Q^2P+\cdots+Q^{k-1}P \\ 0 & I \end{pmatrix}$$

La démonstration est immédiate, par récurrence.

Propriété 6. Donc, si u et v sont deux états transitoires on a :

$$\mathbb{P}(X_k = v \mid X_0 = u) = Q^k(u,v) \tag{1}$$

Un "chemin" joignant deux états transitoires, ne passe, bien sûr, que par des états transitoires.

Propriété 7 (admise)**.** Toutes les valeurs propres de Q ont un module inférieur à 1 ; Q^k tend vers zéro quand k tend vers l'infini et l'on a :

$$I+Q+Q^2+\cdots+Q^n = (I-Q^{n+1})(I-Q)^{-1} \to (I-Q)^{-1}$$

[2] on utilisera la notation I pour les matrices unitaires de tailles diverses sans préciser.

 ISSN: 2102-6238

$R = I - Q$ est parfois appelée la matrice "résolvante" du problème. Nous poserons $N = R^{-1}$. Nous admettrons aussi les formules suivantes [3] :

$$\sum_{k=0}^{\infty} Q^k = N \tag{2}$$

$$\sum_{k=0}^{\infty} k\,Q^{k-1} = N^2 \tag{3}$$

$$\sum_{k=0}^{\infty} k(k-1)\,Q^{k-2} = 2N^3 \tag{4}$$

On déduit alors de l'équation (1) que l'espérance du nombre [4] de passages en v partant de u vaut :

$$\sum_{k=0}^{\infty} Q^k(u,v) = N(u,v) \qquad (\text{soit } (I-Q)^{-1}(u,v)) \tag{5}$$

3. Un seul état absorbant

u et v noteront toujours des états transitoires, et w un état absorbant. Pour plus de clarté, commençons par le cas où il y a un seul état absorbant w. P n'a donc qu'une colonne, nous noterons néanmoins ses termes $P(u,w)$.

Il se peut que le problème posé ne comporte d'emblée qu'un seul état absorbant. Il se peut aussi, qu'il en ait plusieurs, mais que l'on ne s'intéresse qu'au temps jusqu'à absorption, sans être intéressé par "lequel des états absorbants a été atteint" ; on regroupe alors tous les états absorbants en un seul, remplaçant la matrice P par le vecteur de taille $s = n-1$ dont le terme de la ligne u vaut $\sum_w P(u,w)$

Exemple 8. Construisons de la sorte l'exemple de CM à un seul état absorbant à partir de l'exemple 3 ci-dessus :

$$\pi = \begin{pmatrix} .5 & .4 & .1 & 0 \\ .5 & .4 & 0 & .1 \\ 0 & .1 & .7 & .2 \\ 0 & 0 & 0 & 1 \end{pmatrix}$$

Seul l'état $w = 4$ est absorbant. Dans cet exemple :

$$Q = \begin{pmatrix} .5 & .4 & .1 \\ .5 & .4 & 0 \\ 0 & .1 & .7 \end{pmatrix} \qquad \text{et} \qquad P = \begin{pmatrix} 0 \\ .1 \\ .2 \end{pmatrix}$$

[3] dans le cas où Q est un réel, les équations (3) et (4) s'obtiennent par dérivation. Dans tous les cas, une vérification formelle, en multipliant pas R^2 ou R^3, est possible.

[4] comptant 1 pour l'instant $k = 0$ si nécessaire.

 ISSN: 2102-6238

3.1. Ordre zéro

La probabilité d'être absorbé par w en exactement k pas vaut donc :

$$\mathbb{P}(u \to w \text{ en } k \text{ pas}) = \sum_v Q^{k-1}(u,v)\,P(v,w)$$

soit

$$\mathbb{P}(u \to w \text{ en } k \text{ pas}) = Q^{k-1}P(u,w) \tag{6}$$

Propriété 9. Partant de u, la probabilité d'être absorbé par w vaut $NP(u,w)$.

Théorème 10. Quel que soit l'état initial, la CM est presque sûrement absorbée.

C'est à nouveau une conséquence de Borel-Cantelli : à chaque pas, la CM "court un risque" r strictement positif d'être absorbée en moins de t pas, donc (nombre fini d'états) un risque $r > p$, où $p > 0$.

Donnons aussi de ce théorème une démonstration algébrique : notons 1_t le vecteur de longueur t, constitué que de 1.

π est une matrice stochastique, donc $\pi \times 1_n = 1_n$; les t premières lignes de cette équation s'écrivent $(Q \quad P) \times 1_n = 1_t$; retranchant à cette équation l'équation (triviale) $(I \quad 0) \times 1_n = 1_t$ on obtient $(Q-I \quad P) \times 1_n = 0$; multiplier cette équation à gauche par $N = (I-Q)^{-1}$ donne $I \times 1_{n-1} = NP$ (fois 1) ; qui dit bien que chaque élément de NP vaut 1.

Propriété 11. Quand il y a un seul état absorbant, $NP = 1_t$, où l'on a noté 1_t le vecteur de longueur t, constitué que de 1.

Notons que la formule (6) donne le résultat le plus fort sur T, à savoir sa loi :

Théorème 12. $\mathbb{P}(T(u) = k) = Q^{k-1}P(u,w)$

3.2. Ordre 1

3.2.1. Première approche

De l'équation (5), on déduit que, partant de u, le nombre (aléatoire) de fois $T(u)$ où la CM passe par un état transitoire (quel qu'il soit) avant absorption a pour espérance :

$$\sum_v N(u,v) \qquad \text{soit} \qquad N \times 1_t(u)$$

C'est bien sûr le nombre moyen de pas conduisant à l'absorption, partant de u :

$$\mathbb{E}(T(u)) = N\,1_t(u) \tag{7}$$

Exemple 13. Dans l'exemple 8,

$$Q = \begin{pmatrix} .5 & .4 & .1 \\ .5 & .4 & 0 \\ 0 & .1 & .7 \end{pmatrix} \quad Q^2 = \begin{pmatrix} .45 & .37 & .12 \\ .45 & .36 & .05 \\ .05 & .11 & .49 \end{pmatrix} \quad Q^3 = \begin{pmatrix} .410 & .340 & .129 \\ .405 & .329 & .080 \\ .080 & .113 & .348 \end{pmatrix}$$

 ISSN: 2102-6238

et

$$N = \begin{pmatrix} 7.2 & 5.2 & 2.4 \\ 6.0 & 6.0 & 2.0 \\ 2.0 & 2.0 & 4.0 \end{pmatrix} \qquad \mathbb{E}(T) = N \begin{pmatrix} 1 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} 14.8 \\ 14.0 \\ 8.0 \end{pmatrix}$$

3.2.2. Seconde approche

Par définition de l'espérance, en utilisant la formule (6), on a

$$\begin{aligned} \mathbb{E}T(u) &= \sum_{k=0}^{\infty} k\,\mathbb{P}(T(u)=k) \\ &= \sum_{k=0}^{\infty} k \sum_{v} Q^{k-1}(u,v)\,P(v,w) \end{aligned}$$

Après interversion des deux sommes, la formule (2) donne

$$\mathbb{E}(T(u)) = N^2\,P(u) \tag{8}$$

Rapprochant (7) et (8), on a écrit $N^2P = N1_t$; ce qui est une conséquence directe de la propriété 11 : $NP = 1_t$.

3.3. Ordre 2

Les calculs sont similaires (on omet ici de noter le point de départ, u) :

$$\begin{aligned} \mathbb{E}(T^2) &= \sum k^2\,\mathbb{P}(T=k) \quad = \quad \sum k^2\,Q^{k-1}\,P \\ &= \sum (k^2-k)\,Q^{k-1}\,P \;+\; \sum k\,Q^{k-1}\,P \\ &= Q\sum (k^2-k)\,Q^{k-2}\,P \;+\; \sum k\,Q^{k-1}\,P \\ &= Q\times(2N^3P) + N^2P = (2QN^2+N)\times 1_t \end{aligned}$$

d'où la variance

$$\mathbb{V}(T(u)) = \mathbb{E}(T(u)^2) - \mathbb{E}(T(u))^2 \tag{9}$$

Exemple 14. Dans l'exemple 3, $Var(T) = \begin{pmatrix} 163.28 \\ 167.60 \\ 107.20 \end{pmatrix}$ $\sigma(T) = \begin{pmatrix} 12.778 \\ 12.946 \\ 10.354 \end{pmatrix}$

3.4. Loi de T

Nous avons vu (théorème 12) que l'on avait même la loi de T :

$$\mathbb{P}(T(u)=k) = Q^{k-1}P(u,w) \tag{10}$$

Face à un échantillon [5], un test de modèle (test sur les paramètres de la matrice π, test sur le fait que T soit statistiquement plus petit – disons – que ce que le modèle prévoit) peut donc être

[5] voir section 5.

 ISSN: 2102-6238

fondé sur cette vraie loi – utilisant par exemple un test de Kolmogorov-Smirnof, éventuellement unilatère –, plutôt que sur des approximations, gaussiennes par exemple, de la loi de la moyenne des t_i observés.

Plus précisément, le théorème 12 dit

Exemple 15. Le tableau suivant donne (en %), dans le cas de l'exemple 8, les probabilités pour que la CM soit absorbée après k pas pour $k = 1, \ldots, 13$, partant de $u = 1$, 2 ou 3 :

u	1	2	3	4	5	6	7	8	9	10	11	12	13
1	0.00	6.00	6.10	5.98	5.76	5.47	5.16	4.83	4.51	4.20	3.90	3.62	3.35
2	10.0	4.00	4.60	4.89	4.95	4.86	4.68	4.45	4.20	3.94	3.68	3.42	3.18
3	20.0	15.0	10.9	8.09	6.15	4.80	3.84	3.16	2.66	2.28	1.99	1.76	1.57

La figure 2 donne elle aussi ces lois.

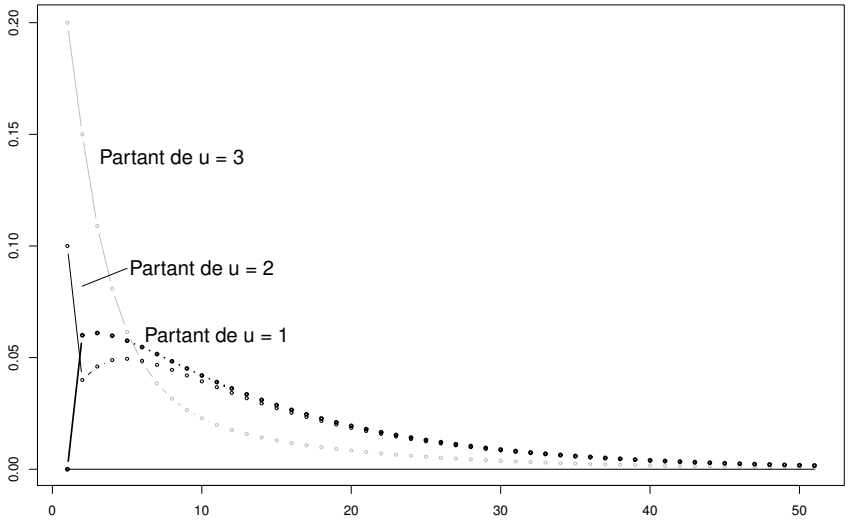

FIGURE 2. *Lois des temps d'absorption $T(u^k)$ partant des trois états initiaux possibles.*

4. Plusieurs états absorbants

S'il y a plus d'un état absorbant, $w_1, \ldots, w_s$, les questions nouvelles sont :

- Partant de u, avec quelle probabilité $\alpha(u, w_j)$ la CM est-elle absorbée par chaque w_j ? (on sait déjà que $\sum_j \alpha(u, w_j) = 1$.)
- Sâchant que la CM est absorbée par w_j [on notera cet événement $\{\rightarrow w_j\}$], quelle est la loi du temps d'absorption ? son espérance ? sa variance ?

 ISSN: 2102-6238

4.1. *Répartition entre états absorbants*

Partant de $X_0 = u$, la CM est absorbée par w s'il existe un instant k et un état transitoire v, tels que $X_k = v$, suivi d'une transition de v vers w :

$$\begin{aligned}\mathbb{P}(X \text{ absorbé par } w \mid X_0 = u) = & \sum_k \sum_v Q^k(u,v)\, P(v,w) \\ = & \sum_v N(u,v)\, P(v,w) \;=\; NP(u,w)\end{aligned}$$

Théorème 16. $\mathbb{P}(X \text{ absorbé par } w \mid X_0 = u) \;=\; NP(u,w)$

Autre démonstration :

Notons $\mathbb{P}(u \to w)$ la probabilité, partant de u, d'être absorbé par w.

Partant de u, la CM est absorbée par w en 1 pas avec probabilité $P(u,w)$ ou bien, avec probabilité $Q(u,v)$ elle transite vers v transitoire et est ensuite absorbée par w avec probabilité $\mathbb{P}(v \to w)$:

$$\mathbb{P}(u \to w) = P(u,w) \;+\; \sum_v Q(u,v)\, \mathbb{P}(v \to w)$$

Pour chaque w fixé, le vecteur V des $\mathbb{P}(\bullet \to w)$ vérifie bien $V - Q\,V = P(\bullet, w)$.

Exemple 17. Dans l'exemple 3,

$$NP = \begin{pmatrix} 7.2 & 5.2 & 2.4 \\ 6.0 & 6.0 & 2.0 \\ 2.0 & 2.0 & 4.0 \end{pmatrix} \begin{pmatrix} 0 & 0 \\ 0 & 0.1 \\ 0.1 & 0.1 \end{pmatrix} = \begin{pmatrix} 0.24 & 0.76 \\ 0.20 & 0.80 \\ 0.40 & 0.60 \end{pmatrix}$$

On a bien, sur chaque ligne, une loi de probabilité, i.e. $\mathbb{P}$(être absorbé par état 4) + $\mathbb{P}$(être absorbé par état 5) = 1.

4.2. *Conditionnement par l'état absorbant*

Les calculs sont similaires ; le théorème 12 s'écrit encore :

$$\mathbb{P}(T(u) = k \text{ et } \to w) = Q^{k-1}P(u,w)$$

Cependant la somme sur k de ces quantités ne vaut plus 1 mais $\alpha(u,w) = NP(u,w)$. La formule de Bayes donne donc la loi de $T(u) \mid \to w$:

Propriété 18. $\mathbb{P}(T(u) = k \mid \to w) = \dfrac{1}{\alpha(u,w)}\, Q^{k-1}P(u,w)$

Et, par exemple : $\mathbb{E}(T(u) \mid \to w) = \dfrac{1}{\alpha(u,w)}\, N^2 P(u,w)$.

 ISSN: 2102-6238

5. Application aux occurences d'un mot dans une CM

Un problème rencontré, en particulier en génomique, est celui du nombre d'occurrences d'un mot W, ou plus généralement d'un motif, dans une chaîne de Markov. Une vaste littérature est consacrée à ce sujet ; dans le seul cadre de la génomique, où l'on s'intéresse aux mots "exceptionnellement fréquents" ou "exceptionnellement rares" [6], citons [4][5][2][3].

Une approche largement développée consiste à considérer le nombre d'occurences $N(W)$ du mot W dans la séquence observée et à le comparer à son espérance $\mathbb{E}(N(W))$ dans le modéle choisi.

Une approche alternative consiste à considérer les "distances entre deux occurences consécutives" de W et à comparer leurs valeurs $T_1, T_2, \dots$ sur la séquence observée à la loi théorique de T dans le même modèle choisi.

Ces deux approches sont en quelque sorte duales, puisque :

$N(W)$ significativement trop grand $\iff$ T significativement trop petit

$N(W)$ significativement trop petit $\iff$ T significativement trop grand.

James Fu ([1]) a proposé une formulation élégante, amplement développée depuis (voir par exemple [2]).

Notre but ici n'est nullement d'introduire cet algorithme dans sa généralité (cas de "motifs" composés de plusieurs mots), mais de montrer sur un **exemple** l'idée de base de cet algorithme, montrant qu'il étudie l'absorption d'une chaîne de Markov par un état absorbant, et s'appuie donc sur la théorie développée ci-dessus.

5.1. Embedded Markov Chain

Exemple 19. Considérons la CM $X = (X_0 X_1 X_2 \dots X_n \dots)$ sur l'alphabet $\mathscr{A}$ à deux lettres `a` et `b` dont la matrice de transition est :

$$\tilde{\pi} = \begin{pmatrix} .4 & .6 \\ .3 & .7 \end{pmatrix}$$

et donnons nous un mot $W =$ `abaab`.

Question 1 : partant de X_0 donné que dire de la variable aléatoire T_0, nombre de pas nécessaires pour rencontrer W ?
Question 2 : après une occurence de W, quel est le nombre (aléatoire) de pas nécessaires pour trouver une autre occurence de W ?

L'idée est que le mot "s'écrit" petit à petit lorsque la CM se réalise, autrement dit qu'il est judicieux d'introduire les "préfixes" de W :

[6] c'est à dire apparaissant significativement plus fréquemment ou significativement moins fréquemment qu'il n'est attendu dans le modèle de CM utilisé pour modéliser la séquence. L'idée est que si un mot apparaît plus souvent que ne le prédit le modèle, il a sans doute un rôle biologique positif pour l'organisme qui porte ce génome – et, inversement, s'il apparaît significativement mois souvent que prévu, il est suspecté d'avoir un rôle néfaste.

 ISSN: 2102-6238

$\mathscr{S}$ = {Ø , a , ab , aba , abaa , W=abaab }

Par commodité, nous numéroterons ces préfixes :

$$0 = \emptyset \qquad 1 = \texttt{a} \qquad 2 = \texttt{ab}$$
$$3 = \texttt{aba} \quad 4 = \texttt{abaa} \quad 5 = \texttt{W}$$

À la CM X, il est alors possible d'associer une autre chaîne Y en définissant :

$$Y_n = \text{le plus long préfixe de } W \text{ qui soit un suffixe de } X_1X_2\ldots X_n$$

Exemple 20. La chaîne X ci-dessous est ainsi transformée en la chaîne Y :

```
X = baaabbababaaabbbabaabaaabbabbaabaab
Y = 01112012323412001234531120120112345
```

Propriété 21. Si X est une CM (sur l'alphabet $\mathscr{A}$), alors Y est une CM sur l'alphabet des préfixes, $\mathscr{S}$. La CM Y sera qualifiée de "embedded".

Cette propriété est évidente. Qui plus est, il est facile de calculer la matrice de transition π qui la régit à partir de $\tilde{\pi}$, matrice des transitions de X :

Exemple 22. Dans l'exemple 19 ci dessus, on a :

$$\pi = \begin{pmatrix} .7 & .3 & 0 & 0 & 0 & 0 \\ 0 & .4 & .6 & 0 & 0 & 0 \\ .7 & 0 & 0 & .3 & 0 & 0 \\ 0 & 0 & .6 & 0 & .4 & 0 \\ 0 & .4 & 0 & 0 & 0 & .6 \\ .7 & 0 & 0 & .3 & 0 & 0 \end{pmatrix}$$

Pour formaliser la question "au bout de combien de pas atteint-on W ?", donc exprimer le fait que, une fois W rencontré, le problème est terminé, on "arrête" la CM lorsqu'elle atteint W, on rend l'état $5 = W$ absorbant, autrement dit on modifie π en posant :

$$\pi(5,5) = 1 \qquad \text{et} \qquad \forall j \neq 5,\ \pi(5,j) = 0$$

π s'écrit donc maintenant :

$$\pi = \left(\begin{array}{ccccc|c} .7 & .3 & 0 & 0 & 0 & 0 \\ 0 & .4 & .6 & 0 & 0 & 0 \\ .7 & 0 & 0 & .3 & 0 & 0 \\ 0 & 0 & .6 & 0 & .4 & 0 \\ 0 & .4 & 0 & 0 & 0 & .6 \\ \hline 0 & 0 & 0 & 0 & 0 & 1 \end{array}\right)$$

 ISSN: 2102-6238

5.2. *Solution*

On a donc affaire à une CM possédant un seul état absorbant et, avec les notations du paragraphe 3, on a :

$$Q = \begin{pmatrix} .7 & .3 & 0 & 0 & 0 \\ 0 & .4 & .6 & 0 & 0 \\ .7 & 0 & 0 & .3 & 0 \\ 0 & 0 & .6 & 0 & .4 \\ 0 & .4 & 0 & 0 & 0 \end{pmatrix}$$

d'où $N = (I - Q)^{-1}$:

$$N = \begin{pmatrix} 35.75 & 19.00 & 13.89 & 4.17 & 1.67 \\ 32.41 & 19.00 & 13.89 & 4.17 & 1.67 \\ 32.41 & 17.31 & 13.89 & 4.17 & 1.67 \\ 24.63 & 13.42 & 10.55 & 4.17 & 1.67 \\ 12.96 & 7.58 & 5.55 & 1.67 & 1.67 \end{pmatrix} \qquad \mathbb{E}(T) = \begin{pmatrix} 74.44 \\ 71.11 \\ 68.44 \\ 54.44 \\ 29.44 \end{pmatrix}$$

on a aussi indiqué le vecteur $\mathbb{E}(T)$, contenant les espérances du nombre de pas nécessaires pour atteindre W, partant de chacun de ses préfixes.

On peut alors répondre aux deux questions posées dans l'exemple 4 :

Question 1 Il est raisonnable de supposer la première lettre de la séquence X obtenue selon la loi stationnaire de $\tilde{\pi}$, à savoir [7] $\mathbb{P}(X_0 = \texttt{a}) = .3/.9 = 1/3$, $\mathbb{P}(X_0 = \texttt{b}) = .6/.9 = 2/3$.

Le nombre de pas nécessaires pour atteindre W aura alors pour espérance

$$E(T) = \frac{2}{3}74.44 + \frac{1}{3}71.11 = 73.33$$

Question 2 Il est sans doute plus intéressant de s'interroger sur le nombre de pas permettant de passer d'une occurence de W à la suivante (voir l'application statistique ci-dessous).

$W = \texttt{abaab}$ étant périodique, une occurence en position k peut tout à fait être suivie d'une occurence en position $k+3$: le mot W crée pour l'occurence suivante le préfixe $\texttt{ab}$.

.....$\underline{\texttt{aba}}\overline{\underline{\texttt{ab}}\texttt{aab}}$.....

Si T désigne le nombre aléatoire de pas pour passer d'une occurence de W à la suivante (dans l'exemple ci-dessus, $T = 3$), on trouve donc $\mathbb{E}(T)$, dans la ligne $3 = \texttt{ab}$ de N :

$$\mathbb{E}(T) = \mu = 68.44 \tag{11}$$

La formule de la sous-section 3.3 donne $\mathbb{V}(T) = \sigma^2 = 5014.432$; l'écart type de T vaut donc $\sigma = 70.8$.

Application statistique Soit H_0 l'hypothèse selon laquelle une séquence est régie par la CM de transition $\tilde{\pi}$.

[7] on sait (ou on vérifie) que la CM de transition $\tilde{\pi} = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$ a pour loi stationnaire $\left(\frac{q}{p+q}, \frac{p}{p+q}\right)$

 ISSN: 2102-6238

Sur une séquence, on mesure les nombres de pas séparant $n = 100$ occurences consécutives de W. La moyenne de ces mesures est notée $\bar{T}$. Quel intervalle $[u, v]$ sera la zone de non-rejet bilatère du test de niveau $\alpha = 5\%$ (disons) de H_0 ?

Le TCL [8] indique que, sous H_0, la loi de $\bar{T}$ peut être approchée par la loi gaussienne $\mathcal{N}(\mu\,;\,\sigma^2/n)$, soit $\mathcal{N}(68.44\,;\,51.0144)$. on rejettera H_0 si et seulement si $\bar{T}$ est observé en dehors de l'intervalle $[u, v] = [68.44 \pm 1.96 \times 7.08] = [54.56\,;\,82.32]$.

6. Un exemple en génomique

Un exemple prototypique pour ce qui concerne les "mots exceptionnels" en génomique est celui du Chi d'*E. coli* : cette bactérie comporte un complexe protéique qui détruit les virus qui pénètrent dans la cellule. Pour que ce complexe ne s'attaque pas à son propre génome, celui-ci porte (un grand nombre de fois) une signature que reconnaît le complexe, qui peut ainsi faire la différence entre de l'ADN étranger (un virus) et l'ADN cellulaire.

Pour être efficace, le Chi (`gctggtgg`) doit être abondamment présent dans la cellule [9]. Autrement dit les intervalles entre deux occurences successives de Chi (les longueurs des segments fragiles face à l'action du complexe) doivent être courts.

Une approche pour vérifier cette propriété – ou pour rechercher les Chi d'autres bactéries – est celle de Fu.

1) On modélise le génome d'*E. coli* par une C.M., dont la matrice de transition (sur l'alphabet $\mathscr{A}$ = `{a,c,g,t}`, dans cet ordre) est estimée sur la séquence observée :

$$\tilde{\pi} = \begin{pmatrix} .295724 & .224614 & .208587 & .271075 \\ .275621 & .230135 & .293517 & .200727 \\ .227332 & .325986 & .229525 & .217157 \\ .185630 & .234349 & .282581 & .297440 \end{pmatrix}$$

2) Il est alors facile de construire la matrice de transition π, à 9 lignes et 9 colonnes entre les préfixes de Chi. Les matrices Q, R et N ont 8 lignes et 8 colonnes. La ligne intéressante de N est celle associée au préfixe g, préfixe "dont on repart" après l'observation d'un Chi :

$$N(\mathrm{g}) = (32887.56\ 16729.24\ 5553.49\ 1094.66\ 309.33\ 71.00\ 15.42\ 4.36)$$

La somme de ses termes est l'espérance $\mathbb{E}(T)$ de la distance entre deux occurences successives de Chi : $\mathbb{E}(T) = 63565$. L'équation (9) donne l'écart-type $\sigma(T) = \sqrt{\mathbb{V}(T)} = 63561.21$.

3) Sur une souche donnée d'*E. Coli* (la souche K12), on observe (sur un des deux brins d'ADN) 503 occurences du Chi. Les 502 intervalles ont une longueur moyenne de $\bar{T} = 9319.70$. Dans la C.M. ajustée,

$$\bar{T} \sim \mathcal{N}\left(\mathbb{E}(T)\,;\,\frac{\mathbb{V}(T)}{n}\right) = \mathcal{N}(63565\,;\,2836.88^2)$$

Le test unilatère à gauche a un degré de significativité de 10^{-81}. Les Chi sont beaucoup plus proches les uns des autres que ne le voudrait le seul hasard.

[8] le caractère markovien de la CM de transition $\tilde{\pi}$ implique que les longueurs des intervalles successifs séparant des occurences de W sont indépendantes.

[9] Le vrai enjeu de la recherche est d'ailleurs de trouver quel est "le Chi" d'autres bactéries en se fondant sur son caractère "exceptionnellement" fréquent.

 ISSN: 2102-6238

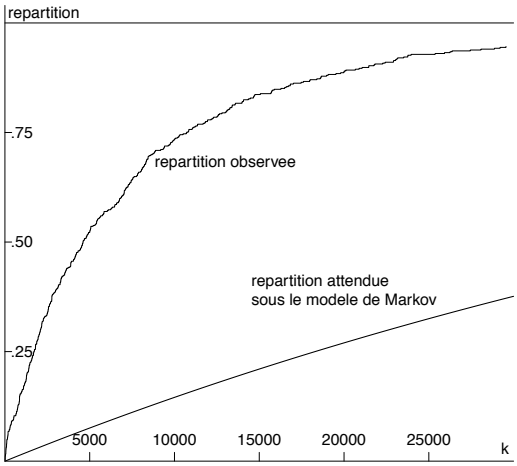

FIGURE 3. *Fonction de répartition observée et fonction de répartition sous le modèle de Markov de l'intervalle entre deux Chi.*

4) On peut aussi tester directement la loi de T : la formule (10) entraîne que, sous le modèle de Markov :

$$\mathbb{P}(T(u) \leq k) = R\,(I - Q^k)\,P(u,w)$$

La figure 3 donne dans le cas présent la loi observée et la loi théorique. On peut donc faire tout test souhaité (ici bien sûr extrêmement significatif).

7. Exemples d'autres applications

1 Deux joueurs A et B commencent une partie, A avec a euros, B avec b euros. Ils lancent une pièce de monnaie, si Pile sort, A donne un euro à B ; si Face sort c'est le contraire. Ils répètent l'opération jusqu'à la ruine de l'un d'eux. Quelle probabilité a chacun de gagner ; combien dure la partie ?.

Que deviennent les résultats si à chaque coup A gagne avec probabilité θ alors que B gagne avec probabilité $1-\theta$?

2 Un joueur de tennis gagne chaque "échange" avec probabilité p (p. ex. $p = .51$) (+ indépendance). Décrivant un "jeu" comme une CM sur :

$$E = \{0,15,30,40\}^2 \cup \{\text{avantage, désavantage, gain, perte}\}$$

quelle probabilité a-t-il de gagner un "jeu" ?

Hint : on pourra avec profit partitionner E en $E = A \cup B$, avec $B = \{$égalité,avantage, désavantage, gain, perte$\}$

Quelle probabilité a-t-il de gagner un "set" ? un "match" ? Combien dure un "jeu" ? un "set" ? un "match" ?

La figure 4 donne un aperçu du résultat.

 ISSN: 2102-6238

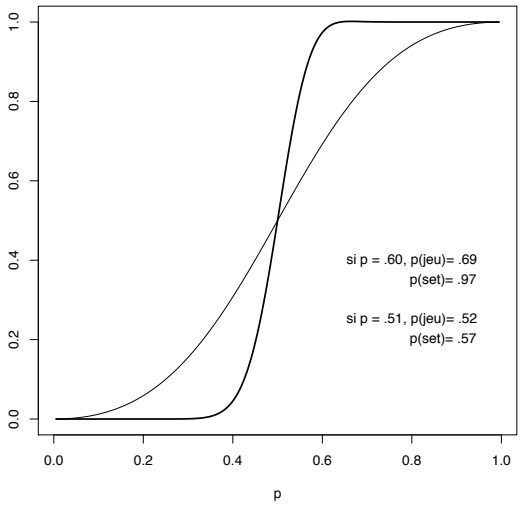

FIGURE 4. *Probabilité, sous le modèle du texte, de gagner un jeu ou un set, en fonction de la probabiliré p de gagner chaque échange..*

Je remercie Grégory Nuel pour avoir relu le manuscrit et pour avoir fourni les données numériques relatives au Chi d' E. coli.

Références

[1] J.C. FU et M.W. KOUTRAS : Distribution theory of runs : a markov chain aopproach. *J. Amer. Statist. Assoc.*, 89:1050–1058, 1994.

[2] G. NUEL : Pattern markov chains : optimal markov chain embedding through deterministic finite automata. *J. Appl. Proba.*, 45:226–243, 2008.

[3] G. NUEL et B. PRUM : *Analyse statistique des séquences biologiques*. 2007.

[4] B. PRUM, F. RODOLPHE et E. DE TURCKHEIM : Finding words with unexpected frequencies in dna sequences. *J. Royal Statistical Society*, 57:205–220, 1995.

[5] G. REINERT, S. SCHBATH et S. WATERMAN : Probabilistic and statistical properties of finite words in sequences. *In Applied Combinatorics in Words*. Cambridge University Press, 2005.

 ISSN: 2102-6238