Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Acad. Sci. Paris, Ser. I 336 (2003) 863–868
COMPTES RENDUS MATHEMATIQUE

Statistique/Probabilités

Distribution exacte du score local, cas markovien

Sabine Mercier, Claudie Hassenforder

Université de Toulouse II, équipe GRIMM, dpt Math-Info, UFR SES, 31100 Toulouse cedex 9, France

Reçu le 9 janvier 2003 ; accepté après révision le 15 avril 2003

Présenté par Paul Deheuvels

Résumé

Soit $\mathbb{X} = (X_k)_{k\geqslant 1}$ une suite de variables à valeurs dans $\{-v, \ldots, 0, \ldots, +u\}$. On définit le score local d'une séquence par $H_n = \max_{1\leqslant i\leqslant j\leqslant n}(\sum_{k=i}^{j} X_k)$. Le score local est utilisé notamment dans l'analyse des séquences biologiques afin de mettre en évidence des régions de séquences ayant des propriétés biologiques intéressantes. La signification statistique des scores locaux calculés permet alors de mettre en évidence ce qui est réellement intéressant et il est donc nécessaire de connaître la distribution du score local. Nous établissons ici la loi exacte du score local dans le cas où la suite des X_i est une chaîne de Markov d'ordre 1. ***Pour citer cet article : S. Mercier, C. Hassenforder, C. R. Acad. Sci. Paris, Ser. I 336 (2003).***

Abstract

Exact distribution for the local score of a Markov chain. Given a sequence $\mathbb{X} = (X_k)_{k\geqslant 1}$ of random variables taking values in $\{-v, \ldots, 0, \ldots, +u\}$, let's define the local score of the sequence by $H_n = \max_{1\leqslant i\leqslant j\leqslant n}(\sum_{k=i}^{j} X_k)$. The local score is used to analyze biological sequences pointing out regions of the sequences with interesting biological properties. In order to separate randomly events from really interesting segments, we establish here the distribution of the local score of H_n when the sequence $\mathbb{X}$ is a Markov chain of order 1. ***To cite this article: S. Mercier, C. Hassenforder, C. R. Acad. Sci. Paris, Ser. I 336 (2003).***

Abridged English version

Introduction

The Human Genome Project started in 1990, focusing on developing tools used to analyze the numerous biological sequences filling data banks. In order to point out segments with interesting biological properties, each component of the sequence is assigned a numerical value representing physicochemical properties as hydrophobicity, antigeniticity, ... These values are called scales or scores and are often used to analyze proteins. Let $s(k)$ be the score of the k-th component of a given sequence of length n. The local score is then defined by

Adresses e-mail : mercier@univ-tlse2.fr (S. Mercier), chabriac@univ-tlse2.fr (C. Hassenforder).

1631-073X/03/$ – see front matter
doi:10.1016/S1631-073X(03)00208-5

$H_n = \max_{1\leqslant i\leqslant j\leqslant n} \sum_{k=i}^{j} s(k)$. Let us specify that the length of the segment which realizes the local score is not *a priori* known. The statistical problem is to establish the distribution of the local score in order to separate random events from really interesting segments. The distribution of H_n was already established when the sequence of the scores $\mathbb{X}$: $X_1, \ldots, X_n$ is modeled as a sequence of independent and identically distributed random variables taking values in $\{-v, \ldots, 0, \ldots, +u\}$. See [4,6] for asymptotic formula of the distribution when the average score is non positive, and [7] for the exact distribution in any average score. A Markov chain is a more adequate model for biological sequence analysis and we give here the distribution of H_n when $\mathbb{X}$ is a 1-order Markov chain.

Results

Let $(X_k)_{k\geqslant 1}$ be a 1-order Markov chain of probability matrix $\Lambda = (\Lambda_{ij})_{i,j\in\mathbb{Z}}$ and γ the initial distribution. The Markov chains will be implicitly of order 1.

Let $P = (P_{(i,j)(k,\ell)})$ be a matrix such that (i, j) and (k, ℓ) belong to

$$E = \{0, \ldots, a\} \times \{-v, \ldots, 0, \ldots, +u\} \quad \text{with } a \in \mathbb{N}^*, \tag{1}$$

and defined by

$$\begin{cases} P_{(i,j)(0,\ell)} = \Lambda_{j\ell} & \text{if } i + j \leqslant 0, \\ P_{(i,j)(i+j,\ell)} = \Lambda_{j\ell} & \text{if } 0 < i + j \leqslant a - 1, \\ P_{(i,j)(a,\ell)} = \Lambda_{j\ell} & \text{if } i + j \geqslant a, \\ P_{(i,j)(k,\ell)} = 0 & \text{else.} \end{cases} \tag{2}$$

Theorem 0.1. *The statistical significance of the local score H_n is given by*

$$(\forall a \geqslant 0) \quad P[H_n \geqslant a] = \sum_{j,\ell} \gamma_j \cdot P^n_{(0,j)(a,\ell)}.$$

Let S_k be the partial sums of the sequence $\mathbb{X}$: $S_0 = 0$ and $S_k = X_1 + \cdots + X_k$. Let T_k be the following stopping times: $T_0 = 0$ and $T_{k+1} = \inf\{i > T_k;\ S_i - S_{T_k} < 0\}$. Consider the process U defined by: $U_0 = 0$ and for $T_k \leqslant j < T_{k+1}$, $U_j = S_j - S_{T_k}$. We have

Lemma 0.2. $U_j = \max(U_{j-1} + X_j, 0) = (U_{j-1} + X_j)^+$ *and* $H_n = \max_{1\leqslant k\leqslant n} U_k$.

Let U^* be the process stopped in a, with $a \in \mathbb{N}^*$. We get $U^*_j = U_j$ if $j < \tau_a$ and $U^*_j = a$ if $j \geqslant \tau_a$ with $\tau_a = \inf\{j \geqslant 1;\ U_j \geqslant a\}$. And finally, let us define the sequence $\mathbb{Y}$ by: $Y_{n+1} = (U^*_n, X_{n+1})$ for $n \geqslant 0$.

Lemma 0.3. $\mathbb{Y}$ *is a Markov chain with probability matrix* $P = (P_{(i,j)(k,\ell)})_{(i,j)(k,\ell)\in E}$, *and* $P_{(i,j)(k,\ell)} = P[(U^*_n = k) \cap (X_{n+1} = \ell) \mid (U^*_{n-1} = i) \cap (X_n = j)]$, *determined in* (2).

Lemma 0.4. *The distribution of U^* is given by*

$$P[U^*_n = k] = \sum_{j,\ell} \gamma_j \cdot P^n_{(0,j)(k,\ell)}.$$

From Lemma 0.2, we deduce $P[H_n \geqslant a] = P[U^*_n = a]$ and using Lemma 0.4 and the explicitation of the $P_{(i,j)(k,\ell)}$, Theorem 0.1 is proved. The determination of the $P_{(i,j)(k,\ell)}$ is easy, using elementary notions of Markov Chain Theory.

Conclusion and perspectives

Using both the exact distribution of H_n in i.i.d. and Markovian case, we are going to compare these two models with an empirical distribution calculated on real biological sequences. This comparison will stand on exact formulas and thus will focus on the models only.

1. Introduction

La biostatistique est une discipline en pleine expansion de même que toutes les disciplines en rapport avec le génome et les bases de données de séquences biologiques croissent de façon exponentielle. Nous sommes maintenant dans l'ère de l'après séquençage où l'information doit être extraite de ces nouvelles données. Un des objectifs du projet Génome Humain, débuté en 1990, consiste à développer et à améliorer les outils d'analyse de séquences. Il existe actuellement de nombreux outils permettant d'extraire de l'information de ces séquences. Plusieurs portent sur l'analyse de la structure primaire, c'est-à-dire sur la connaissance de la succession des bases azotées (ou nucléotides) ou des acides aminés (pour les protéines), et déterminent des profils de protéines ou d'ADN à partir d'échelles. Une échelle d'acides aminés associe à chaque type d'acides aminés une valeur numérique également appelée score et déterminée expérimentalement ou statistiquement. Les échelles d'acides aminés les plus fréquemment utilisées sont les échelles d'hydrophobicité et celles correspondant aux paramètres de conformation de structure secondaire. Ces profils mettent en évidence des parties ou segments de séquences les plus pondérées suivant l'échelle utilisée. Soit H_n le score local correspondant au score maximal observé, en considérant tous les segments de la séquence de longueur n, à n'importe quelle position, de toutes les longueurs possibles. Nous avons

$$H_n = \max_{1 \leqslant i \leqslant j \leqslant n} \sum_{k=i}^{j} s(k),$$

avec $s(k)$ le score du $k$ème élément de la séquence.

Afin de distinguer l'information pertinente du simple hasard, il est important d'établir la signification statistique du score local. De nombreux travaux portent sur l'établissement de la loi de H_n, qui nécessite avant tout de définir un modèle sur les séquences biologiques, ou bien sur les séquences des scores correspondantes.

Soit donc $X_1, \ldots, X_n$ une suite de variables aléatoires à valeurs dans $\{-v, \ldots, 0, \ldots, +u\}$, $-v$ et $+u$ étant respectivement les minimum et maximum de l'échelle choisie pour l'étude. Le cas où les variables X_k sont indépendantes et identiquement distribuées a fait l'objet de nombreux travaux. Voir par exemple [8,2] pour une revue des problèmes et résultats sur le sujet. La distribution exacte du score local a été établie dans le cas i.i.d. dans [1] et [7]. Ce résultat a trois avantages par rapport aux précédents travaux portant sur la loi du score local. Le premier est dû au caractère exact du résultat, contrairement aux approximations utilisées jusque là, [3,4]. Le deuxième intérêt, repose sur le fait que le score moyen $E(X_k)$ n'a pas besoin d'être négatif. En effet, les travaux de Karlin et al. [3,4], de Mercier et al. [6,5], ainsi que ceux de Waterman et al. se placent dans le cas où l'espérance du score est négative (on a alors $H_n = \mathcal{O}(\ln n)$), mais ne sont pas valables si cette espérance est positive (dans ce cas $H_n = \mathcal{O}(n)$). Le résultat de Daudin et al. [1,7] est quant à lui valable indépendamment du signe du score moyen. Enfin, un troisième aspect particulièrement intéressant concerne la longueur des séquences étudiées. En effet, les approximations de la loi du score local sont asymptotiques par rapport à la longueur des séquences. Or, les protéines ont une longueur moyenne de 350 acides aminés et ne rentrent pas dans ce cadre asymptotique. Certes, le caractère exact de [7] résout ce problème, mais l'implémentation du résultat sera d'autant plus précise et rapide à obtenir (cumul d'arrondis de calculs) que les séquences seront courtes.

Le but de cette Note consiste à établir la loi de H_n dans le cas où les séquences sont modélisées par une chaîne de Markov d'ordre 1. Nous travaillons directement sur la suite des scores c'est-à-dire sur des suites à valeurs dans $\mathbb{Z}$. La loi du score local dans le cas markovien est énoncée dans le Théorème 2.1. Les démonstrations utilisent des notions élémentaires de la théorie des chaînes de Markov.

Toutes les chaînes de Markov sont d'ordre 1, même si cela n'est pas précisé chaque fois.

2. Résultats et démonstations

Soit $(X_k)_{k\geqslant 1}$ une chaîne de Markov d'ordre 1, de matrice de transition $\Lambda = (\Lambda_{ij})_{i,j\in\mathbb{Z}}$ et de loi initiale γ. Le score local de la suite (X_k) est défini par $H_n = \max_{1\leqslant i\leqslant j\leqslant n}(\sum_{k=i}^{j} X_k)$. Soit

$$E = \{0,\dots,a\}\times\{-v,\dots,0,\dots,+u\} \quad \text{avec } a\in\mathbb{N}^*. \tag{3}$$

Introduisons la matrice $P = (P_{(i,j),(k,\ell)})$, où (i,j) et (k,ℓ) sont dans E, définie par

$$\begin{cases} P_{(i,j)(0,\ell)} = \Lambda_{j\ell} & \text{si } i+j\leqslant 0,\\ P_{(i,j)(i+j,\ell)} = \Lambda_{j\ell} & \text{si } 0<i+j\leqslant a-1,\\ P_{(i,j)(a,\ell)} = \Lambda_{j\ell} & \text{si } i+j\geqslant a,\\ P_{(i,j)(k,\ell)} = 0 & \text{sinon.}\end{cases} \tag{4}$$

Nous avons le résultat suivant

Théorème 2.1. *La signification statistique du score local H_n est donnée par la formule suivante*

$$(\forall a\geqslant 0)\quad P[H_n\geqslant a] = \sum_{-v\leqslant j,\ell\leqslant +u} \gamma_j\cdot P^n_{(0,j)(a,\ell)}. \tag{5}$$

Démonstration. Notons S_k les sommes partielles associées à la suite des X_k : $S_0 = 0$ et $S_k = X_1+\cdots+X_k$. Considérons la suite des temps d'arrêt T_k définie par : $T_0 = 0$ et $T_{k+1} = \inf\{i > T_k;\ S_i - S_{T_k} < 0\}$.

Par définition des T_k, la suite des S_{T_k} est strictement décroissante, et les T_k sont appelés les temps successifs des records négatifs.

Soit U la suite définie de manière récurrente par : $U_0 = 0$ et pour $T_k\leqslant j < T_{k+1}$, $U_j = S_j - S_{T_k} = X_{T_k+1} + \cdots + X_j$. On a en particulier $U_{T_k} = 0$ pour tout $k\geqslant 0$. Les (U_j) forment une suite positive non nécessairement bornée. Comme démontré dans [1] et [7], nous avons les résultats suivants :

Lemme 2.2.

$$U_j = \max(U_{j-1}+X_j,0) = (U_{j-1}+X_j)^+ \quad \textit{et} \quad H_n = \max_{1\leqslant k\leqslant n} U_k.$$

Notons U^* le processus arrêté de U en a, où a est dans $\mathbb{N}^*$.

$$U^*_j = U_j \quad \text{si } j<\tau_a \quad \text{et} \quad U^*_j = a \quad \text{si } j\geqslant\tau_a \quad \text{avec } \tau_a = \inf\{j\geqslant 1;\ U_j\geqslant a\}. \tag{6}$$

Le processus (U^*_j) est une chaîne de Markov non homogène à valeurs dans $\{0,1,\dots,a\}$.

Dans le cas d'une suite (X_k) i.i.d. (voir [1] et [7]), U^* est une chaîne de Markov homogène et il est alors facile d'établir la loi de U^*_n, ce qui n'est plus vrai dans le cas d'une suite (X_k) markovienne. Afin d'expliciter la loi de U^*_n, considérons la chaîne (Y_n) définie par

$$(\forall n\geqslant 0)\quad Y_{n+1} = (U^*_n, X_{n+1}), \tag{7}$$

qui est homogène et dont l'ensemble des états est E défini en (3).

Lemme 2.3 (Matrice de transition de Y). *$(Y_n)_{n\geqslant 1}$ est une chaîne de Markov de matrice de transition $P = (P_{(i,j)(k,\ell)})$, avec (i,j) et (k,ℓ) dans E, où les $P_{(i,j)(k,\ell)}$ sont donnés par* (4).

$$P\big[(U^*_n = k)\cap(X_{n+1} = \ell)\mid (U^*_{n-1} = i)\cap(X_n = j)\big] = P_{(i,j)(k,\ell)}.$$

Les $P_{(i,j)(k,\ell)}$ sont déterminés à l'aide de Λ comme indiqué dans le Théorème 2.1.

Lemme 2.4 (Loi de U_n^*).

$$P[U_n^* = k] = \sum_{j,\ell} \gamma_j \cdot P^n_{(0,j)(k,\ell)}. \tag{8}$$

Du Lemme 2.2, on tire alors $P[H_n \geqslant a] = P[U_n^* = a]$. En utilisant ensuite le Lemme 2.4 et l'explicitation des $P_{(i,j)(k,\ell)}$, le Théorème 2.1 est alors démontré.

Démonstration du Lemme 2.3. Nous avons pour $i = a$, $P_{(a,j)(k,\ell)} = 0$ si $k \leqslant a-1$ car U^* est un processus arrêté en a, et $P_{(a,j)(k,\ell)} = \Lambda_{j\ell}$ pour $k = a$. Pour $i \neq a$, alors : $P_{(i,j)(k,\ell)} = P[(U_n^* = k) \cap (X_{n+1} = \ell) \mid (U_{n-1} = i) \cap (X_n = j)]$.

- Si $k = 0$, alors

$$\begin{aligned} P_{(i,j)(0,\ell)} &= P\big[(X_n \leqslant -U_{n-1}) \cap (X_{n+1} = \ell) \mid (U_{n-1} = i) \cap (X_n = j)\big] \\ &= P\big[(j \leqslant -i) \cap (X_{n+1} = \ell) \mid (U_{n-1} = i) \cap (X_n = j)\big] \\ &= P\big[(j \leqslant -i) \cap (X_{n+1} = \ell) \mid (X_n = j)\big], \end{aligned}$$

  car U_{n-1} ne dépend que de $X_1, \ldots, X_{n-1}$ et X_n est une chaîne de Markov d'ordre 1 par hypothèse. D'où, $P_{(i,j)(0,\ell)} = \Lambda_{j\ell}$ si $j \leqslant -i$ et 0 sinon.
- Si $0 < k \leqslant a-1$, alors

$$\begin{aligned} P_{(i,j)(k,\ell)} &= P\big[(U_{n-1} + X_n = k) \cap (X_{n+1} = \ell) \mid (U_{n-1} = i) \cap (X_n = j)\big] \\ &= P\big[(i + j = k) \cap (X_{n+1} = \ell) \mid (U_{n-1} = i) \cap (X_n = j)\big] \\ &= \Lambda_{j\ell} \text{ si } k = i + j \quad \text{et} \quad 0 \text{ sinon.} \end{aligned}$$

- Pour $k = a$, nous avons

$$\begin{aligned} P_{(i,j)(a,\ell)} &= P\big[(X_n \geqslant a - U_{n-1}) \cap (X_{n+1} = \ell) \mid (U_{n-1} = i) \cap (X_n = j)\big] \\ &= P\big[(j \geqslant a - i) \cap (X_{n+1} = \ell) \mid (U_{n-1} = i) \cap (X_n = j)\big] \\ &= \Lambda_{j\ell} \text{ si } i + j \geqslant a \quad \text{et} \quad 0 \text{ sinon.} \end{aligned}$$

3. Perspectives

Ayant à notre disposition la loi exacte du score local dans le cas où les séquences sont modélisées par une suite de variables aléatoires indépendantes et identiquement distribuées, cf. [1] et [7], ainsi que maintenant dans le cas markovien, nous avons ici la possibilité de comparer des modèles sans que les conclusions ne soient altérées par le fait que l'on aura utilisé des approximations.

Références

[1] J.-J. Daudin, S. Mercier, Distribution exacte du score local d'une suite de variables indépendantes et identiquement distribuées, C. R. Acad. Sci. Paris 329 (1) (1999) 815–820.

[2] R. Durbin, S. Eddy, A. Krogh, G. Mitchison, Biological Sequence Analysis. Probabilistic Models of Proteins and Nucleic Acids, Cambridge University Press, Cambridge, UK, 1998.

[3] S. Karlin, S.F. Altschul, Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes, Proc. Nat. Acad. Sci. USA 87 (1990) 2264–2268.

[4] S. Karlin, A. Dembo, Limit distributions of maximal segmental score among Markov-dependent partial sums, Adv. Appl. Probab. 24 (1992) 113–140.

[5] S. Mercier, Statistiques des scores pour l'analyse et la comparaison de séquences biologiques, Thèse de doctorat d'Université, Rouen, 1999.
[6] S. Mercier, D. Cellier, F. Charlot, J.-J. Daudin, Exact and asymptotic distribution for the local score of one i.i.d. random sequence, in: JOBIM 2000, in: Lecture Notes in Comput. Sci., Vol. 2066, 2001, pp. 74–85.
[7] S. Mercier, J.-J. Daudin, Exact distribution for the local score of one i.i.d. random sequence, J. Comp. Biol. 8 (4) (2001) 373–380.
[8] M.S. Waterman, Introduction to Computational Biology, Chapman and Hall, London, 1995.