Distribution du score local pour la détection de régions atypiques au sein de séquences

Sabine Mercier

▶ **To cite this version:**

Sabine Mercier. Distribution du score local pour la détection de régions atypiques au sein de séquences. Statistiques [math.ST]. Université Toulouse III Paul Sabatier (UT3 Paul Sabatier), 2018. tel-01944432

HAL Id: tel-01944432
https://hal.science/tel-01944432

Submitted on 4 Dec 2018

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Université Fédérale de Toulouse

Manuscrit
présenté pour l'obtention de

l'Habilitation à Diriger des Recherches

Sabine Mercier

INSTITUT DE MATHÉMATIQUES de TOULOUSE

Local score distribution to highlight atypical segments in sequences

---

Soutenue le 3 décembre 2018,

après avis des rapporteurs

DR. Franck Picard	CNRS, Laboratoire Biométrie et Biologie Évolutive, UCB Lyon 1
Pr. Etienne Birmelé	Laboratoire MAP5, Université Paris Descartes
Pr. Pierre Pudlo	Institut de Mathématiques de Marseille, Université d'Aix Marseille

devant le jury formé de

Pr. Catherine Matias (Présidente)	CNRS, Laboratoire Probabilité Statistique et Modélisation, Sorbonne Université, Université Paris Diderot
Pr. Etienne Birmelé	Laboratoire MAP5, Université Paris Descartes
Pr. Patrick Cattiaux	Institut de Mathématiques de Toulouse, Université Paul Sabatier
Pr. Béatrice Laurent-Bonneau	Institut de Mathématiques de Toulouse, INSA de Toulouse
DR. Franck Picard	CNRS, Laboratoire Biométrie et Biologie Évolutive, UCB Lyon 1
Pr. Pierre Pudlo	Institut de Mathématiques de Marseille, Université d'Aix Marseille
Pr. Stéphane Robin	INRA, AgroParisTech

À Bernard Prum,
et
à ma fille Aurélie Mercier.

Remerciements

Je souhaite tout d'abord remercier sincèrement Stéphane Robin, mon parrain, d'avoir accepté de me représenter, mais aussi pour avoir pris le temps de m'apporter son regard sur mon travail il y a quelques années de cela.

Je remercie chaleureusement Franck Picard, Etienne Birmelé et Pierre Pudlo, les trois rapporteurs de cette habilitation, pour le soin apporter à leur rapport. J'en profite également pour remercier Madame Antonin pour son aide précieuse et sa souplesse à gérer mon dossier.

Un grand merci à Béatrice Laurent-Bonneau, Patrick Cattiaux et Catherine Matias pour s'être joints au jury.

Ce travail doit énormément à l'ensemble de mes collaborateurs. Que ces personnes soient assurées de ma reconnaissance pour avoir accepté de développer ces projets, enrichi nos échanges, cru en mes idées. Je leurs exprime ici outre mon amitié tout le plaisir que j'ai pu avoir à travailler avec eux.

Mes remerciements vont également à l'ensemble de mes collègues, de l'Université de Toulouse 2 Jean Jaurès, de l'Institut de Mathématiques de Toulouse ainsi que de l'INRA de Castanet Tolosan. Merci aux personnes de l'IMT qui malgré que j'ai un pied de chaque côté de la Garonne me rappellent que je suis bien un membre à part entière de cet institut.

Une pensée pour Grégory Nuel qui m'a témoigné de sa confiance et n'a pas hésité à me soutenir financièrement dans une période délicate. Pour cela et les magnifiques projets communs je le remercie chaleureusement.

Caroline Thierry a toujours su apporter un regard critique sur mon travail et ses questions ont régulièrement amélioré la présentation de la problématique pour un public non spécialiste. Un grand merci à elle pour avoir su se rendre disponible à chaque fois que j'ai pu la solliciter.

Enfin, j'adresse un remerciement particulier à Simona Grusea pour l'immense plaisir à partager avec elle nos séances de travail fructueuses, mélangeant plaisir et rigueur des mathématiques et dégustation de thés.

Contents

Introduction

La motivation principale du travail présenté dans ce document est l'extraction d'information des séquences biologiques, une séquence biologique pouvant être un brin d'ADN (succession de nucléotides), une protéine (succession d'acides aminés), ou bien de l'ARN messager ou encore une suite de site de gènes, pour ne citer que ces exemples. Plus précisément, il s'agit de mettre en évidence des segments atypiques au sein de ces séquences. Un segment atypique étant un segment qui possède une propriété exceptionnelle par rapport au reste de la séquence, la propriété en question étant quantifiée par ce qu'on appelle le score local. Il s'agit principalement d'établir la signification statistique du score local pour conclure à la présence ou non d'un segment atypique. Nous verrons également que ces travaux peuvent être portés à d'autres domaines d'études que la biologie moléculaire. Ils peuvent en fait être appliqués de manière très générale à tout domaine comportant des séquences longitudinales comme une succession de mesures ou d'observations.

Plusieurs choix sont possibles pour présenter les résultats sur la signification du score local que ce soient ceux déjà existants ou bien ceux auxquels j'ai contribués et que je vais exposer dans ce document. Nous pouvons par exemple les rassembler suivant les différentes théories et les outils mathématiques sur lesquelles reposent ces résultats : la théorie du renouvellement, des chaînes de Markov, du mouvement brownien ou encore théorie de Chein-Stein, des grandes déviations. J'ai souhaité privilégier l'organisation de ce document en me plaçant du côté de l'utilisateur et de l'application et de me reposer sur les critères et le contexte d'étude. J'espère par ce choix élargir le public lecteur de ce document. Dans ce même état d'esprit, j'ai opté pour une rédaction le plus souvent littérale et tout public. Les aspects techniques et mathématiques sont quant à eux largement développés dans les articles cités.

Les différents contextes ou critères d'étude rencontrés sont les suivants :

- **Le choix du modèle** utilisé pour représenter les séquences étudiées : Nous aborderons le cas des suites de variables aléatoires indépendantes et identiquement distribuées (i.i.d.); le cas des chaînes de Markov (CM); ainsi que le cas des chaînes de Markov cachées (HMM pour Hidden Markov Model) pour le calcul de distributions a posteriori.
- **Le choix des scores**, réels ou entiers naturels, et le signe du score moyen

sous le modèle choisi : En effet certains résultats reposent sur des hypothèses d'un score moyen négatif alors que d'autres méthodes nécessitent que les scores soient des entiers naturels.

- **La longueur des séquences** : parmi les résultats théoriques actuellement établis, il faudra distinguer les méthodes exactes de celles approchées reposant sur des résultats asymptotiques pour une longueur de séquence tendant vers l'infini. Ainsi, suivant que l'on étudie des protéines, d'une longueur moyenne de 350 acides aminés, des séquences nucléiques, environ 1000 nucléotides en moyenne, ou bien des génomes entiers ($\simeq 10^9$ nucléotides, ou 10^4 loci), certains résultats ne seront pas appropriés. Les questions de mise en pratique, implémentation, rapidité d'exécution, sont aussi à prendre en compte, notamment pour les méthodes exactes qui permettent d'établir théoriquement la distribution du score local quel que soit la longueur de la séquence étudiée, mais dont le temps de calcul peut s'avérer assez long pour de trop longues séquences. Les approximations asymptotiques sont performantes pour des longueurs de séquences moyennes à longues suivant les cas.

Dans la **Partie** I nous rappelons les résultats théoriques obtenus suivant les différents modèles de séquences cités ci-dessus (i.i.d., CM et HMM). Les applications du score local et les questions que ces applications soulèvent sont présentées dans la **Partie** II. La **Partie** III développe deux projets de recherche.

La **Partie** I comporte quatre chapitres. Nous commençons par exposer tout d'abord le problème et donnons les définitions principales dans le **Chapitre** 1.

Le **Chapitre** 2 porte sur la signification statistique du score local dans le cas de séquences i.i.d. Il rappelle tout d'abord les résultats obtenus lors de ma thèse pour un souci de synthèse et de cohérence : il s'agit de la méthode exacte ainsi que de l'amélioration de l'approximation de Karlin *et al.* dans le modèle i.i.d., que nous rappelons également. Ce chapitre comporte également les travaux de thèse d'Afshin Fayyaz Movagghar que j'ai co encadrée de 2004 à 2007 avec Louis Ferré et qui propose une nouvelle approximation du score local de comparaison de deux séquences i.i.d. avec décalage et insertions délétions (indels) de composants possibles. Nous abordons à la suite les travaux effectués en collaboration avec Pierre Valois (Institut Elie Cartan de Nancy) et Agnès Lagnoux (Institut de Mathématiques de Toulouse) utilisant principalement la théorie du mouvement brownien. Les résultats portent sur la distribution de la longueur de réalisation du score local, celle du couple (score local, longueur du score local). Suivent les résultats sur la position du score local. La question du choix des scores utilisés et de la longueur des séquences adaptées à une mise en application est également discutée.

Le cas de la modélisation par chaînes de Markov est abordé au **Chapitre** 3. Nous commençons par présenter la méthode exacte pour un modèle markovien portant sur la suite des scores, puis celui sur la suite des composants des séquences (nucléotides, acides aminés ou autres suivant le contexte d'étude). Ce travail a été effectué en collaboration avec Claudie Chabriac de l'université

de Toulouse 2 Jean Jaurès, et repose essentiellement sur la théorie des chaînes de Markov. Nous rappelons ensuite l'approximation de Karlin et Dembo dans le modèle markovien, résultat peu, voire pas, référencé. Une approximation plus précise de la distribution du score local dans le cas markovien et pour un choix de scores plus usuel est ensuite proposée, travaux en collaboration avec Simona Grusea (INSA de Toulouse, Institut de Mathématiques). A nouveau la question du choix des scores et des longueurs de séquences est discutée.

Dans le **Chapitre** 4 nous abordons avec Grégory Nuel (Laboratoire de Probabilité et Modèles Aléatoires, Universités Sorbonne, Pierre et Marie Curie) l'approche par chaînes de Markov cachées. Nous commençons par faire le lien entre une approche par segmentation reposant sur les HMM, et l'approche usuelle du score local. Les résultats des probabilités a posteriori de la présence d'un segment atypique au sein d'une séquence, les probabilités a posteriori sur la longueur du score local, et celles des indices de position du début et de la fin de segment, sont ensuite présentés. Nous verrons également différentes conséquences ou portées intéressantes que cette approche peut amener, comme par exemple la définition de scores pour composants ambigus de séquence ("A ou B").

La **Partie** II présente principalement 3 applications des résultats précédents. Le **Chapitre** 5 regroupe les applications biologiques : la mise en évidence d'une part de région de sélection de gènes dans les génomes de cailles, travaux de thèse de Marìa Inès Fariello en 2014 ; et d'autre part la recherche simultanée de plusieurs segments atypiques au sein d'une séquence par l'approche des HMM, travaux en collaboration avec Grégory Nuel et Hélène Chiapello, INRA de Jouy-en-Josas. Nous abordons au **Chapitre** 6 un travail de mise en évidence d'apparition plus fréquente d'événements rares en Maîtrise Statistique des Procédés (surveillance en agroalimentaire, risques médico-sanitaires, ...), travail réalisé avec François Bergeret, entreprise Ippon Innovation. D'autres idées ou projets d'applications sont proposés au **Chapitre** 7 .

La Partie III comporte deux chapitres correspondant à deux projets différents quoique liés. Ces sujets de recherche sont venus très naturellement suite à la constatation suivante que les biologistes étaient tout autant intéressés par le segment le plus atypique que par les N segments les plus atypiques, voire tous les segments atypiques d'une séquence.

Le **Chapitre** 8 présente un sujet de recherche reposant sur la théorie de Chen-Stein, théorie des chaînes de Markov et grandes déviations. Il fait suite aux travaux théoriques effectués avec Simona Grusea du Chapitre 3 et propose d'établir une approximation de la loi du nombre de régions d'une séquence, dont la hauteur dépasse un seuil donné, par une loi de Poisson.

Nous proposons dans le **Chapitre** 9 un travail en lien avec les tests multiples. Cette proposition provient d'une réflexion effectuée lors du suivi de la thèse de Marìa Inès Fariello et de discussions avec Pierre Neuvial (Institut de Mathématiques de Toulouse). Ce second projet contient deux aspects techniques. Tout d'abord, il s'agit d'établir la loi exacte ou approchée des scores de

régions de séquences dans les modèles i.i.d. et markovien ; puis de proposer une correction de type Bonferroni sur les tests multiples de ces régions en nombre certes fini mais aléatoire.

Par ailleurs, l'approximation proposée dans le premier projet, pour la loi du nombre de régions d'une séquence dépassant un seuil donné t, amène à se poser la question suivante : Pour une séquence donnée, nous testons pour tout t si la séquence admet un nombre significatif de régions dépassant t, comment alors prendre en compte la multiplicité de type continu de ce test, t prenant ses valeurs dans un intervalle.

Remarque 1 *Dans ce document, un travail d'uniformisation des notations a été effectué. Il en résulte une différence de notations utilisées dans les articles cités. Une liste récapitulative des principales notations est donnée en annexe.*

Part I

Significativité du score local - Résultats théoriques

Chapter 1

Position du problème, notations et définitions

Soit $\mathbb{A} = (A_i)_{1 \leq i \leq n}$, une suite de "composants" de séquences, des nucléotides par exemple dans le cas d'une séquence d'ADN. Soit $\mathbb{X} = (X_i)_{1 \leq i \leq n}$ une suite dite de "scores" valeurs dans $\mathbb{R}$ ou $\mathbb{Z}$. Avec s une fonction de scores, on note $X_i = s(A_i)$. Les suites $\mathbb{A}$ et $\mathbb{X}$ représentent des suites de variables aléatoires dont la distribution, sur $\mathbb{A}$ ou $\mathbb{X}$ sera précisée suivant les chapitres.

Soit $S_0 := 0$ et $S_m := \sum_{k=1}^{m} X_k = \sum_{k=1}^{m} s(A_k), m \geq 1$ les sommes partielles associées à la séquence des scores et $S^+ = \max_{1 \leq k \leq n}(S_k, 0)$, le maximum des sommes partielles.

Karlin et Altschul définissent en 1990 le score local d'une séquence de scores de la manière suivante

$$M_n := \max_{0 \leqslant i \leqslant j \leqslant n} \sum_{k=i}^{j} X_k \ . \tag{1.1}$$

En 2001, une nouvelle définition équivalente est proposée [17, 40]. Nous avons

$$M_n := \max_{0 \leqslant j \leqslant n} U_j \tag{1.2}$$

avec $(U_j)_{j \geqslant 0}$ le processus de Lindley définit récursivement par

$$U_0 := 0 \text{ et } U_j := (U_{j-1} + X_j)^+ \ . \tag{1.3}$$

La récurrence du processus peut également s'écrire

$$U_j = \sum_{k=1}^{j} X_k - \min_{1 \leqslant i \leqslant j} \sum_{k=1}^{i} X_k \ .$$

Le score local correspond alors à la plus grande hauteur des excursion au dessus de 0, ou "montagnes", que définit le processus de Lindley. Cela peut être

visuellement très pratique et est largement utilisée par les biologistes. Cette définition équivalente ouvre de nouvelles possibilités à la fois pour le travail mathématique mais aussi pour les applications pratiques grâce à certaines propriétés bien utiles.

Propriété 1 (Score local, processus de Lindley et excursions)

1. *Le score local est la hauteur de la plus grande excursion au dessus de 0 du processus de Lindley.*
2. *La valeur du score local d'une séquence* $\mathbb{X} = X_1, \ldots, X_n$ *est la même que celle de la séquence inverse* $\tilde{\mathbb{X}} = \tilde{X}_1, \ldots, \tilde{X}_n$ *avec* $\tilde{X}_i = X_{n+1-i}$ *:* $M_n(\mathbb{X}) = M_n(\tilde{X})$.
3. *Les segments qui réalisent ces valeurs correspondent aux mêmes composants de séquences. Si* $x_d, \ldots, x_f$ *et* $\tilde{x}_{\tilde{d}}, \ldots, \tilde{x}_{\tilde{f}}$ *sont les segments réalisant le score local dans la sequence lue directement et dans celle inverse, alors* $x_f = \tilde{x}_{\tilde{d}}$ *et* $x_d = \tilde{X}_{\tilde{f}}$
4. *Les deux propriétés précédentes permettent de mettre en évidence graphiquement le score local de manière efficace même dans le cas de séquences très longues et pouvant comporter de nombreuses excursions.*

La figure suivante 1.1 illustre les propriétés précédantes.

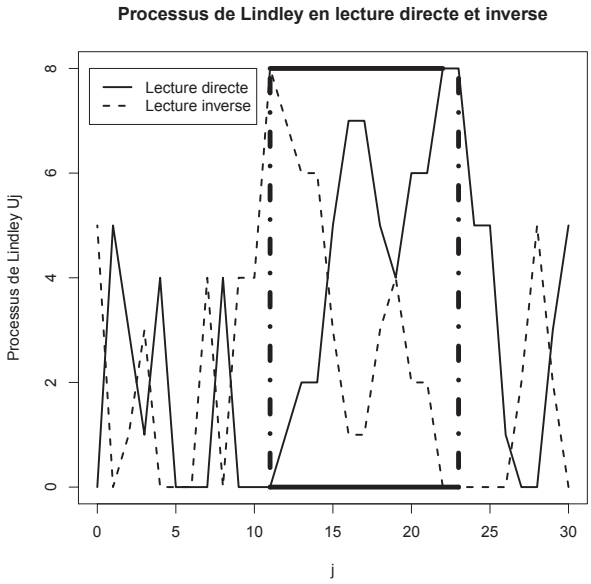


Figure 1.1: En trait continu, le processus de Lindley d'une séquence de longueur ici $n = 30$. La hauteur de la plus haute montagne est 8, commence à l'indice $d = 11$ et atteint son maximum en $f = 22$ et 23 (voir segment en gras). Dans le cas de la lecture inversée, de l'indice 30 à 1, le processus de Lindley est représenté en pointillé, la plus haute montagne a pour hauteur 8, commence à l'indice $\tilde{d} = 9 = 30+1-22 = n+1-f$ et se réalise en $\tilde{f} = 20 = 30+1-11 = n-f+1$.

Chapter 2

Loi du score local de séquences aléatoires i.i.d.

Dans ce chapitre nous pouvons distinguer trois parties. La première rappelle les résultats établis lors de ma thèse et publiés en 2001 et 2003 ainsi que le premier résultat sur la distribution du score local d'analyse de séquence, l'approximation de Karlin et Altschul de 1990 [30]. Il est important de rappeler ces résultats dans ce document afin d'une part que ce document recouvre l'ensemble des résultats permettant d'établir la signification statistique du score local et d'autre part nous nous y ramenons à plusieurs reprises lors des autres travaux présentés ici. La seconde partie plus récente, de 2014 à 2018, complète la problématique de la signification statistique du score local en prenant en compte, la longueur de réalisation du score local, et/ou sa position. Ces derniers travaux reposent en partie sur la théorie du mouvement brownien alors que les précédents utilisent la théorie des chaînes de Markov et/ou du renouvellement. La dernière partie aborde le cas de la comparaison de deux séquences i.i.d. et le travail de thèse d'Afshin Fayyaz Movaghar que j'ai co-encadré avec Louis Ferré en 2007.

D'un point de vue pratique, nous allons voir que pour une longueur de séquence et une fonction de scores données, il n'existe que peu de cas où un choix dans les différents résultats est possible. En effet, le contexte d'analyse de séquences impose la plupart du temps la méthode à utiliser.

2.1 Analyse de séquences courtes à moyennes : méthode exacte

Théorie des chaînes de Markov
Scores : $E[X]$ quelconque, scores entiers ou rationnels
Travail effectué pendant ma thèse
Mes collaborateurs : Jean-Jacques Daudin

Publications : [17], [40]

Résultat 1 (Méthode exacte [40]) *Pour un choix de scores entiers relatifs, la distribution du score local s'obtient, indépendamment du signe du score moyen $E[X]$, par la méthode suivante*

$$\mathbb{P}(M_n \geqslant a) = (1, 0, \ldots, 0) \cdot \Pi^n \cdot (0, \ldots, 0, 1)', \tag{2.1}$$

avec Π une matrice carrée $(a+1) \times (a+1)$ établie à l'aide de la distribution des scores (voir [40] pour plus de détails).

L'application de ce résultat est clairement dépendante d'une part de la longueur de la séquence étudiée n et d'autre part de la nécessité de scores entiers relatifs. Pour une longueur assez grande n, la valeur du score local a peut devenir lui même assez grand et l'on se retrouve avec une matrice de taille importante $a+1$ et une exponentiation à un niveau lui-même importante n. La rapidité des calculs est conditionnée par l'outil informatique utilisé. Pour un usage intensif et des calculs à répétition il n'est pas raisonnable d'utiliser des séquences de plusieurs milliers de composants, encore moins dans le cadre d'un score moyen positif qui augmente les chances d'obtenir un score local élevé et donc la taille de la matrice Π. Les scores entiers relatifs quant à eux entrainent un score local a fortiori entier et permettent la construction de la matrice Π ce qui ne serait pas faisable dans un quelconque autre cas.

Pour des séquences de longueurs inférieures à 10^4 et des scores entiers relatifs c'est l'outil à privilégier.

2.2 Analyse de séquences longues : approximations et amélioration

Théorie du renouvellement et théorie des chaînes de Markov

Scores : score moyen négatif, scores entiers ou rationnels pour les applications

Travail effectué pendant ma thèse

Mes collaborateurs : Dominique Cellier, François Charlot

Publications : [26] [14]

En 1990, Karlin et Alschul établissent une approximation de la distribution du score local par une loi de Gumbel lorsque que la longueur de la séquence tend vers l'infini. Cette approximation repose sur les deux paramètres de la Gumbel, λ et K^* qui dépendent tous deux de la distribution des scores. Si λ reste très simple à obtenir, cela est plus délicat pour K^*. L'existence de λ est conditionnée par les deux hypothèses ci-dessous (2.2). Nous retrouverons ces conditions dans le travail sur les chaînes de Markov (Chapitre 3), ainsi que sur les chaînes de Markov cachées pour le calcul des probabilités a posteriori (Chapitre 4).

Résultat 2 (Approximation par une loi de Gumbel) *Pour*

$$\mathbb{E}[X] < 0 \text{ et } \mathbb{P}(X > 0) > 0 \ , \tag{2.2}$$

nous avons

$$\lim_{n\to\infty} \mathbb{P}(M_n \leqslant a + \frac{\log n}{\lambda}) = \exp\big\{ - K^* \cdot e^{-\lambda a}\big\}, \tag{2.3}$$

avec

Ce résultat est établi en différentes étapes : approximation de la loi du maximum des sommes partielles $S^+ = \max_k S_k$ avec $S_k = \sum_{j=1}^k X_j$ reposant sur la théorie du renouvellement. Ce résultat est ensuite porté à la distribution de la hauteur de la première excursion avant la première somme partielle négative notée $Q = Q_1$. Sous le modèle i.i.d. la hauteur des excursions Q_k définies par les temps d'arrêt d'échelle descendant successifs

$$T_0 := 0 \text{ et } T_k = \inf\{j > T_{k-1},\ S_{T_{k-1}} - S_j < 0\}$$

sont i.i.d. L'approximation sur la distribution du score local M_n vient ensuite du fait que $M_n \simeq \max_k Q_k$, et est donc un maximum de variables aléatoires i.i.d. Nous aurons l'occasion de revenir sur le fait que le nombre de montagne est aléatoire et qu'il faut également gérer la queue de la séquence, i.e. la dernière montagne incomplète.

L'amélioration de ce résultat repose sur le fait qu'il est possible d'obtenir la loi exacte du maximum des sommes partielles. En effet, cette loi est la (une ?) loi invariante de la chaîne de Markov U^x pour $x = \sup_{p\leq 0, p\in\mathbb{Z}} \left(\sum_{i=p+1}^0 X_i\right)$, $U_0^x := x$, $U_{k+1}^x := (U_k^x + X_{k+1})^+$, qui n'est autre que la récurrence du processus de Lindley.

En portant cette amélioration aux étapes suivantes, la distribution de Q, puis la distribution de M_n, nous obtenons l'approximation suivante

Résultat 3 (Approximation plus précise [15])

$$\lim_{n\to\infty} \mathbb{P}(M_n \leqslant a + \frac{\log n}{\lambda}) = \Big[1 - \sum_{k\geq 0} (-1)^{\alpha_k} \frac{K_k \cdot |R_k|^a}{n^{\log|R_k|/\log R}}\Big]^{\frac{n}{\mu}+1}, \tag{2.4}$$

où $\alpha_1 = 0$, α_k *est la partie entière de* $\left(a + \frac{\log(n)}{\lambda}\right)$ *quand* $k \geq 0$, $\mu = \mathbb{E}[T_1]$ *avec* $T_1 = \inf\{k : \sum_{i=1}^k X_i < 0\}$ *et* $(R_k, K_k)_{k\geqslant 0}$ *des constantes adéquates qui dépendent de la distribution des* X_i. *Remarquons que* λ *dans* (2.3) *est égale à* $\log(1/R_1)$ *et que le premier terme de la somme de* (2.4) *permet d'obtenir la limite de Karlin et al.*

Cette approximation s'avère une bonne alternative lorsque que la méthode exacte est trop longue en temps de calcul mais que les séquences ne sont pas suffisamment grandes pour que la loi de Gumbel soit utilisée : comme pour des séquences de l'ordre d'une centaine de composants et de grandes valeurs de score local par exemple.

Les idées principales des démonstrations de ces résultats, que ce soit pour la méthode exacte et les approximations seront reprises dans le cas du modèle des chaînes de Markov : L'utilisation du processus de Lindley dans le cas des méthodes exactes et la théorie des chaînes de Markov ; l'enchaînement loi du maximum des sommes partielles, loi de la hauteur de la première excursion puis celle du score local pour les approximations reposant sur la théorie du renouvellement.

2.3 Analyse de séquences très longues

Théorie du mouvement brownien

Scores : score moyen nul

Mes collaborateurs : Agnès Lagnoux, Pierre Vallois

Publications : score local et de sa longueur : [16] ; position du score local dans les excursions : [3] ; indice de position dans la séquence : article soumis [5]. Voir également [4].

La théorie du mouvement brownien permet de nouvelles perspectives de travail et des résultats combinant à la fois valeur du score local, et longueur ou position de réalisation du score local. Cela implique par ailleurs de prendre en compte les éléments suivants.

- Cela nécessite de travailler dans le cadre d'un score moyen centré réduit, $\mathbb{E}[X] = 0$ et $\mathbb{V}[X] = 1$.

- La théorie du mouvement brownien étudie des trajectoires continues. Le lien avec les séquences biologiques discrètes s'effectue par le biais d'une interpolation linéaire (voir juste ci-dessous) et des convergences. Les simulations ont montré que la vitesse de convergence est très faible et cela nécessite de très longues séquences pour utiliser concrètement les résultats.

- Afin de pouvoir utiliser les résultats du mouvement brownien, nous sommes contraints de ne prendre en compte que les montagnes complètes du processus de Lindley et de laisser de côté la dernière excursion incomplète. Nous travaillons donc sur un score local légèrement différent, défini sur la séquence tronquée aux excursions complètes, noté M_n^*. Le Résultat 6 permet de connaître la probabilité que les deux scores locaux soit identiques.

Interpolation linéaire

Soit $M > 0$ un paramètre d'échelle. Le processus continue classique noté $\left(B^M(t),\ t \geq 0\right)$ associés aux (S_n) et au facteur de normalisation M est défini par $B^M\left(\frac{k}{M}\right) = \frac{1}{\sqrt{M}} S_k$ et pour tout k tel que $\frac{k}{M} \leq t \leq \frac{k+1}{M}$

$$B^M(t) = B^M\left(\frac{k}{M}\right) + M\left(t - \frac{k}{M}\right)\left(B^M\left(\frac{k+1}{M}\right) - B^M\left(\frac{k}{M}\right)\right).$$

On définit ensuite le processus $(U^M(t), t \geq 0)$ par

$$U^M(t) = B^M(t) - \min_{s \leq t} B^M(s), \quad t \geq 0.$$

Remarquons que $U^M\left(\frac{k}{M}\right) = \frac{1}{\sqrt{M}} U_k$ pour $k \geq 0$ où (U_k) est le processus de Lindley associé aux sommes partielles (S_k). Le Théorème de Donsker, cf. Section 2.10 in [9], nous indique que le processus $\big(B^M(t),\ t \geq 0\big)$ converge faiblement vers le mouvement Brownian $\big(B(t),\ t \geq 0\big)$ où $M \to +\infty$ (cf. Section 3.2 de [16]). En utilisant de plus l'identité en loi $\big(|B(t)|,\ t \geq 0\big) \overset{(d)}{=} \big(B(t) - \min_{0 \leq u \leq t} B(u),\ t \geq 0\big)$ nous obtenons le résultat suivant (cf. [16])

$$\big(U^M(t),\ t \geq 0\big) \text{ converge faiblement vers } \big(U(t),\ t \geq 0\big) \tag{2.5}$$

avec $\big(U(t) := |B(t)|, \quad t \geq 0\big)$ le mouvement brownien réfléchi. Nous avons par ailleurs, pour $g(t) = \sup\{s \leq t,\ U(s) = 0\}$ l'indice du dernier 0,

$$M^*(t) := \sup_{0 \leq s \leq g(t)} U(s), \quad t \geq 0\ ,$$

l'équivalent du score local dans le cas continu pour une séquence tronquée aux montagnes complètes. Posons également pour $g_n := \max\{k \leq n\ ;\ U_k = 0\}$,

$$M_n^* := \max_{0 \leq k \leq g_n} U_k$$

le score local restreint à la séquence tronquée aux excursions complètes. On notera L_n^* la longueur du segment réalisant M_n^*, $L^*(t)$ la longueur du segment réalisant $M^*(t)$, et $g^*(t)$, l'indice de sa réalisation.

2.3.1 Score local et longueur du score local

Voyons tout d'abord ce peut apporter d'un point de vue pratique la prise en compte de la longueur ? Dans [4], les auteurs illustrent l'apport de la distribution du couple score local et longueur par rapport à celle de la distribution du score local seul, de différentes manières. En voici deux présentées ci-dessous. La première méthode consiste à étudier 606 séquences de protéines d'une base de données de SCOP. A calculer le score local et sa longueur pour chacune de ces séquences. Puis de calculer la p-valeur du score local par la méthode exacte et d'en déduire une liste ordonnée des séquences les plus significatives. Et de calculer par ailleurs la p-valeur du couple score local et longueur et d'en déduire une seconde liste ordonnée des séquences les plus significatives. Les 10 séquences les plus significatives pour chacune des listes ont ensuite été comparées. Il existe une différence dès la troisième séquence (cf. Table 2.1).

La seconde méthode repose sur des calculs de spécificité et de sensibilité. Pour cela nous avons généré $3 \cdot 10^4$ séquences sous l'hypothèse $\mathcal{H}_0$ que les X_i sont

Table 2.1: Liste des 10 séquences les plus significatives pour le score local (haut du tableau) et trois séquences qui ne sont pas considérées parmis les 10 plus intéressantes avec le score local seul, alors qu'elles le sont en prenant en compte la longueur (bas du tableau). On observe également un changement dans l'ordre dès la troisième séquence.

n_i	h_{n_i}	ℓ_{n_i}	p-valeur	H_n ordre	Couple Estimation	(H_n, L_n) ordre
173	185	169	10^{-6}	1	$.< 10^{-6}$	1
103	106	88	$3.13 \cdot 10^{-4}$	2	$5 \cdot 10^{-5}$	2
80	93	76	$4.17 \cdot 10^{-4}$	3	$3.10 \cdot 10^{-4}$	4
94	100	85	$4.03 \cdot 10^{-4}$	4	$2.50 \cdot 10^{-4}$	3
93	88	86	$1.68 \cdot 10^{-4}$	5	$1.24 \cdot 10^{-3}$	5
111	82	107	$6.41 \cdot 10^{-3}$	6	$5.81 \cdot 10^{-3}$	9
129	76	127	$1.75 \cdot 10^{-2}$	7	$1.69 \cdot 10^{-2}$	13
227	93	102	$1.84 \cdot 10^{-2}$	8	$2.94 \cdot 10^{-3}$	8
145	73	130	$2.98 \cdot 10^{-2}$	9	$2.64 \cdot 10^{-2}$	12
109	67	79	$2.56 \cdot 10^{-2}$	10	$1.37 \cdot 10^{-2}$	11
113	49	22	$1.26 \cdot 10^{-1}$	33	$1.37 \cdot 10^{-3}$	6
133	44	18	$2.28 \cdot 10^{-1}$	67	$1.53 \cdot 10^{-3}$	7
227	40	19	$4.96 \cdot 10^{-1}$	192	$8.21 \cdot 10^{-3}$	10

i.i.d. d'une distribution donnée $\mathcal{D}$. Nous avons choisi des scores moyens égaux à -0.01 puis -2.20. Nous avons également simulé $3 \cdot 10^4$ avec la distribution $\mathcal{D}$ puis remplacé un segment de longueur 30 par 30 valeurs i.i.d. simulées avec une distribution vérifiant $\mathbb{E}[X] = +0.5$. La position du segment modifié est choisi de manière aléatoire. Pour chacune de ces $6 \cdot 10^4$, le score local est calculé et sa longueur est relevée. La p-valeur de ces observations est ensuite calculée en utilisant les différentes méthodes disponibles dans le cas d'un score moyen négatif $\mathbb{E}[X] < 0$: méthode exacte, approchées, ainsi que le résultat établi dans la thèse de Marie-Pierre Etienne [19]. Une p-valeur du couple score local et longueur est calculée par une méthode de Monte Carlo afin d'avoir une valeur plus précise. Pour chaque méthode et différents seuils α les séquences vrai et faux positif ainsi que vrai et faux négatif sont comptabilisées et les calculs des spécificités et sensibilités déduits. Les tableaux ci-dessous regroupent les résultats pour les différentes possibilités de calculs de p-valeurs.

Pour le score moyen proche de 0, les sensibilités sont équivalentes et bonnes mais prendre en compte la longueur apporte une meilleure spécificité. Dans le cas d'un score moyen plus négatif, les spécificités sont toutes très bonnes et équivalentes, les sensibilités également toutes du même ordre, plus faibles que dans le cas précédent. La méthode du couple se distingue toutefois en donnant un résultat un peu meilleur.

Les simulations pour d'autres longueurs de segments ont donné des résultats similaires.

Table 2.2: Specificité (SPC) et la Sensibilité (TRP) pour différents seuils α pour un score moyen $\mathbb{E}_{\mathcal{H}_0}[X] = -0.01$

	α	Exact (2)	Karlin (3)	Karlin+ (4)	Etienne	Couple Emp.
SPC	5%	0.952	0.000	0.000	0.301	0.605
	1%	0.990	0.000	0.945	0.490	0.867
	0.5%	0.995	0.999	1	0.551	0.921
	0.1%	0.999	1	1	0.673	0.979
TPR	5%	0.073	1	1	0.760	0.410
	1%	0.016	1	0.080	0.578	0.135
	0.5%	0.008	0.001	0	0.518	0.077
	0.1%	0.002	0	0	0.389	0.020

Table 2.3: Specificité (SPC) et la Sensibilité (TRP) pour différents seuils α pour un score moyen $\mathbb{E}_{\mathcal{H}_0}[X] = -2.20$

	α	Exact (2)	Karlin (3)	Karlin+ (4)	Etienne	Couple Emp.
SPC	5%	0.960	0.946	0.960	0	0.889
	1%	0.990	0.989	0.992	0	0.977
	0.5%	0.996	0.995	0.996	0	0.987
	0.1%	0.999	0.999	0.999	0	0.996
TPR	5%	0.850	0.872	0.850	1	0.780
	1%	0.751	0.751	0.724	1	0.656
	0.5%	0.671	0.697	0.671	1	0.591
	0.1%	0.567	0.594	0.567	1	0.466

Nous avons comme résultat dans le cas continu

Résultat 4 (Score local et longueur, cas continu) *Soit $f : \mathbb{R}_+^2 \to \mathbb{R}$ une fonction de Borel encadrée. Soit $t > 0$. Alors*

$$\mathbb{E}\left(f(M^*(t), L^*(t))\right) = \sqrt{\frac{\pi}{2}}\mathbb{E}\left[f\left(\frac{\alpha_1\sqrt{t}}{\sqrt{Z}}, \frac{t\alpha_1^2\xi}{Z}\right)\frac{\alpha_2 e_0^{'2}}{\sqrt{Z}}\right] \tag{2.6}$$

où $Z = \xi + \xi' + e_0^{'2}\alpha_2^2\lambda(1/e_0')$, $(\xi_k)_{k\geq 1} \cup \{\xi, \xi'\}$ une famille de variables aléatoires indépendantes et identiquement distribuées telles que

$$\xi \overset{(d)}{=} \xi' \overset{(d)}{=} \xi_k \overset{(d)}{=} T_1(R)$$

avec $T_x(R) = \inf\{s \geq 0 \ ; \ R(s) = x\}$, $x > 0$ et $(R(s), s \geq 0)$ un processus de Bessel à 3 dimensions.

e_0', $(e_n)_{n\geq 0}$ une suite de variables exponentielles i.i.d., $(\lambda(x), x \geq 0)$ le processus défini par

$$\lambda(x) := x^2(\xi_1 + \xi_2) + \sum_{k\geq 1} \frac{\xi_{2k+1} + \xi_{2k+2}}{\left(\frac{1}{x} + e_1 + \cdots + e_k\right)^2}, \quad x \geq 0.$$

On supposera e'_0, $(e_n)_{n\geq 0}$, $(\xi_n)_{n\geq 1}$, ξ, ξ' *et* $(U(t))_{t\geq 0}$ *indépendants.*

Nous avons également

$$f_{L^*(t)}(x) = \frac{1}{x}\sum_{k\geq 1}(-1)^{k+1}\frac{\sinh\left(\pi k\sqrt{\frac{x}{t-x}}\right)}{\cosh^2\left(\pi k\sqrt{\frac{x}{t-x}}\right)}\mathbb{1}_{[0,t]}(x). \tag{2.7}$$

$$f_{M^*(t)}(x) = 4\sqrt{\frac{2}{\pi t}}\left(\sum_{k\geq 1}(-1)^{k-1}ke^{-\frac{2k^2x^2}{t}}\right), \quad x>0. \tag{2.8}$$

Remarque 2 *La densité* p_ξ *est connue explicitement*

$$\begin{aligned} p_\xi(u) &= \frac{1}{\sqrt{2\pi}u^{3/2}}\sum_{k\in\mathbb{Z}}\left(-1+\frac{(1+2k)^2}{u}\right)\exp\left(-\frac{(1+2k)^2}{2u}\right) \\ &= \frac{d}{du}\left(\sum_{k\in\mathbb{Z}}(-1)^k\exp\left(-\frac{k^2\pi^2u}{2}\right)\right) \end{aligned}$$

Dans le cas discret

Résultat 5 (Score local et longueur, cas discret) *Supposons* $\mathbb{E}[X]=0$ *et* $\mathbb{V}[X]=1$. *Pour* $a\geqslant 0$ *et* $0\leqslant \ell\leqslant 1$, *nous avons*

$$\mathbb{P}(M_n^*\geqslant\sqrt{n}a; L_n^*\leqslant n\ell)\approx\mathbb{P}(M^*(1)\geqslant\sqrt{n}a; L^*(1)\leqslant n\ell) \tag{2.9}$$

avec $(M^*(1), L^*(1))$ *definis dans Eq. (2.4) et (2.8) dans [16] ou bien précédemment avec* $t=1$.

La figure suivante 2.1 permet d'apprécier la qualité de l'approximation. Nous voyons que la vitesse de convergence du résultat est lente et que la longueur des séquences doit être vraiment grande au moins de l'ordre de plusieurs dizaines de milliers. Dans le cas de séquences biologiques ce résultat semble approprié pour des travaux portant sur des génomes entiers, voir éventuellement des suites de loci.

2.3.2 Score local et position du score local

Nous obtenons également un résultat présentant une convergence de la probabilité que les deux scores locaux M_n, défini sur la totalité de la séquence, et M_n^*, défini sur les montagnes complètes uniquement, soient identiques (cf. [3]).

Résultat 6 (Position du score local dans les excursions)

Soit $p_c^{(n)} := P\Big(\max_{0\leqslant k\leqslant n}U_k = \max_{0\leqslant k\leqslant g_n}U_k\Big)$ *où* $g_n := \max\{k\leqslant n, U_k = 0\}$, *la probabilité que le score local se réalise dans l'une des montagnes complètes du processus de Lindley, i.e. pour lesquelles on observe un retour en 0 en fin d'excursion.*

Nous avons $p_c^{(n)}$ *qui converge vers* $\psi(1/4)-\psi(1/2)+1+\pi/2\approx 0.3069$ *pour* $n\to\infty$.

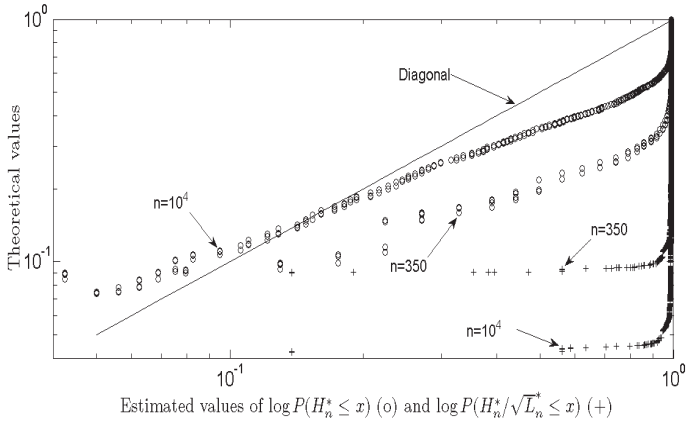


Figure 2.1: Log-log plot de la valeur empirique de $\mathbb{P}(M_n^* \leq h_{n,i}^*)$ (avec M_n^* noté H_n^* dans la figure) et $\mathbb{P}(M_n^*/\sqrt{L_n^*} \leq h_{n,i}^*/\sqrt{\ell_{n,i}^*})$ suivant la valeur de sa limite déduite de [16, Theorem 2.4] avec $n = 350$ et $n = 10^4$.

Cette probabilité est faible par rapport à ce que l'on pourrait s'attendre. En effet, plus la longueur de séquence est grande et plus l'on s'attend à un nombre de montagnes important. Naïvement, l'on pourrait penser que chaque montagne ait autant de chance de réaliser le score local et donc que l'unique dernière montagne incomplète ait donc de moins en moins de chance de réaliser le score local à mesure que la séquence grandit. Or cette probabilité reste élevée $1 - p_c \simeq 2/3$.

Le résultat précédent nous indique que le score local a environ 2 chances sur 3 de se situer dans la dernière montagne et donc plutôt en fin de séquence. Mais sans information sur l'indice de début de la dernière excursion l'information sur la position reste incomplète. Cette remarque a motivé le travail nous permettant d'obtenir le résultat suivant (article soumis [5]).

Résultat 7 (Densité de l'indice de réalisation du score local) *Notons g^* l'indice de réalisation du score local d'une séquence continue tronquée aux montagnes complètes et $h(x) = \sum_{k\geq 1}(-1)^{k+1}\frac{k}{\cosh^2(kx)}$. La fonction de distribution de g^* est alors donnée par*

(i)

$$p_{g^*}(y) = \frac{1}{2\pi\sqrt{y(1-y)}}\int_0^{+\infty} \ln\left|1 - \frac{\pi^2(1-y)}{4ys}\right| \frac{h(\sqrt{s})}{\sqrt{s}}ds,\ 0 \leq y \leq 1 \tag{2.10}$$

(ii) ou encore par

$$p_{g^*}(y) = \frac{1}{2\pi\sqrt{y(1-y)}}\int_0^{+\infty} \ln\left|\cot\sqrt{\frac{ty}{1-y}}\right| \frac{dt}{\sqrt{t}\cosh^2(\sqrt{t})},\ 0 \leq y \leq 1. \tag{2.11}$$

Le résultat peut se transporter aux séquences discrètes de manière similaire aux résultats sur la longueur vues précédemment.

2.4 Comparaison de deux séquences i.i.d avec décalages et indels

Théorie des chaînes de Markov
Scores entiers relatifs, score moyen quelconque
Mes collaborateurs : Afshin Fayaz Movaghar, Louis Ferré
Publications : [2] [1]

Louis Ferré ayant reçu une proposition d'IRAN d'encadrement de thèse auto financée en biostatistique dont le sujet était à définir, m'a proposé de co-encadrer cette thèse en apportant le sujet ce que j'ai accepté avec enthousiasme.

La méthode exacte établissant la signification statistique du score local d'analyse d'une séquence i.i.d. a également permis de proposer dans le cas

de séquences i.i.d., une approximation du score local de comparaison de deux séquences autorisant les décalages, mais sans prendre en compte les insertions délétions (indels ou gaps) (voir [40]). La thèse d'Afshin Fayyaz repose sur l'idée d'utiliser ce résultat et d'incorporer les insertions délétions par le biais d'un choix de fonction de scores judicieux.

Commençons tout d'abord par définir les différents outils mathématiques.

Différentes définitions du score local de comparaison de deux séquences $\mathbb{A}$ et $\mathbb{B}$ existent suivant le degré de complexité de l'évolution des séquences que l'on souhaite prendre en compte. La plus simple étant celle qui ne prend en compte que les décalages (ou shifts) possibles entre les deux séquences et ne tient pas compte des indels.

Sans insertions délétions

Soient $\mathbb{A}$ et $\mathbb{B}$ deux séquences de longueur n et m.

$$H_{n,m} = \max_{\substack{0\leq\ell\leq\min(n,m)\\ 1\leq i\leq n-\ell\\ 1\leq j\leq m-\ell}} \sum_{k=0}^{\ell} s(A_{i+\ell}, B_{j+\ell}) \tag{2.12}$$

avec s une fonction de similarité entre composants.

Avec insertions délétions

La définition du score local de comparaison qui prend en compte les insertions et/ou délétions possible de composants au cours de l'évolution donne lieu à un objet mathématique particulièrement complexe à étudier. Pour cela, il nous faut ici présenter le score global de deux séquences, ou de deux segments de séquences. Soient u (*respectivement* v) une sous-suite de $\{1,\dots,n\}$ (*resp.* $\{1,\dots,m\}$) de longueur ℓ. Ces deux sous-suites vont définir les "lettres alignées" des deux séquences. Les lettres non alignées font l'objet d'une insertion ou d'une délétion et sont pénalisées par $-\delta$.

Dans l'exemple de la figure qui suit $n = 8$, $m = 7$; il y a 6 lettres alignées et les suites u et v sont définies ainsi : $u(1) = 1,\ u(2) = 2,\ u(3) = 4,\ u(4) = 5,\ u(5) = 7,\ u(6) = 8$; $v(1) = 1,\ v(2) = 2,\ v(3) = 3,\ v(4) = 5,\ v(5) = 6,\ v(6) = 7$.

Le meilleur alignement global a donc pour valeur

$$S(\mathbb{A},\mathbb{B}) = \max_{\ell,u,v} \left\{ \sum_{k=1}^{\ell} s(A_{u(k)}, B_{v(k)}) - \delta(n-\ell+m-\ell) \right\} . \tag{2.13}$$

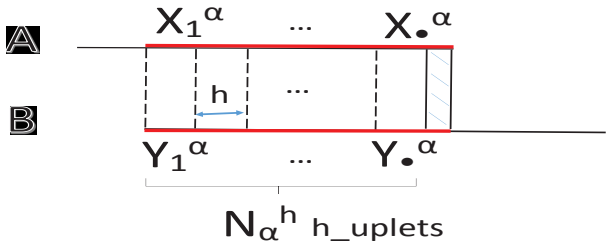
Le score local avec décalages et indels de Smith et Waterman pour deux séquences $\mathbb{A}$ et $\mathbb{B}$ de longueur n et m ([50]) est alors défini par

$$M_{n,m}(\mathbb{A},\mathbb{B}) = \max_{I\subset\mathbb{A};J\subset\mathbb{B}} S(I,J) . \tag{2.14}$$

Travail de thèse d'Afshin

L'objectif de la thèse d'Afshin fût de tirer partie au mieux de la méthode exacte tout en prenant en compte les indels. Pour cela nous découpons les séquences en segments de longueur donnée h (voir Figure 2.4). On définit ainsi un nouveau type de séquence en se ramenant à des suites de h-uplets : un h-uplets étant alors considéré comme un composant de ce nouveau type de séquence, de même que les acides aminés sont les composants d'une protéine. La fonction de scores est alors définie sur les couples de ces h-uplets en utilisant le score global. Plus précisément, pour un décalage α donné entre les deux séquences on définit à partir des *deux* séquences initiales *une* nouvelle séquence, suite de couples de h-uplets de longueur notée N_α^h, ou parfois N_α pour alléger l'écriture

$$(\mathbb{A},\mathbb{B})_\alpha^h = (U_i)_{1\le i\le N_\alpha^h} \text{ avec } U_i \in \mathcal{A}^h \times \mathcal{A}^h .$$



Pour h donné, chaque décalage α définit deux séquences de h-uplets, ou bien une séquence de couples de h-uplets. On peut alors calculer le score local de cette séquence ainsi définie en utilisant une fonction de scores adaptée $\mathcal{S}$ suivante

$$\mathcal{S} : \begin{array}{ccc} \mathcal{A}^h \times \mathcal{A}^h & \to & R \\ (I,J) & \to & S(I,J) \end{array}$$

où $S(I,J)$ est le score global des segments I et J.

Nous avons

$$H_{N_\alpha^h} = H_{N_\alpha^h}((\mathbb{A},\mathbb{B})_\alpha^h)$$

$H_\bullet$ défini en (1.1) et/ou (1.2) étant le score local d'analyse d'une séquence (ici une séquence de couple) pour lequel la loi exacte est connue.

Le score local avec décalages et indels proposé est alors pour h donné

$$\mathcal{M}_{n,m} = \max_{\alpha} H_{N_\alpha^h} \ . \tag{2.15}$$

et le résultat suivant présente une approximation de la distribution de ce score

Résultat 8 (Comparaison de deux séquences avec indels) *Indépendamment du score moyen et pour h donné,*

$$\mathbb{P}(\mathcal{M}_{n,m} \geq a) = 1 - \left[\left(\Pi_{r=h,2h,\ldots,(K/h-1)/h}\mathbb{P}[H_{N_r} < a]^{2h}\right)\mathbb{P}[H_{N_K} < a]^{2(n-K)}\right] \times \mathbb{P}[H_{N_n} < a]^{m-n+1} \tag{2.16}$$

où $K = [\frac{n-1}{n}]h$ et a est le score local $\mathcal{M}_{n,m}$ observé.

L'approximation vient ici du fait que les différents décalages créent des séquences qui ne sont pas indépendantes.

Nous considérons les deux indels-scores locaux de comparaison de deux séquences, celui que nous proposons, définis en (2.15) $\mathcal{M}_{n,m}$, et celui définit par Smith et Waterman défini en (2.14) $M_{n,m}$ et largement utilisé. Les étapes suivantes consistent à

- comparer les deux scores,
- comparer les couples de séquences significatifs par le biais des p-valeurs,
- étudier l'influence du choix de h sur les résultats.

La distribution de $M_{n,m}$ utilisée dans la comparaison des résultats est celle fournie par le logiciel FASTA. Celle de $\mathcal{M}_{n,m}$ est obtenue par (2.16). Ces deux p-valeurs sont ensuite comparées à une distribution empirique calculée par une méthode de Monte Carlo sur un ensemble de séquences : premièrement sur un ensemble de séquences simulées pour ne pas prendre en compte le biais de l'adéquation du modèle ; deuxièmement sur un ensemble de séquences réelles représentatives (SCOP 1.53) pour évaluer la méthode dans des cas réels.

L'influence du choix de h ainsi que de celle de la longueur des séquences sur l'adéquation des résultats est un problème théorique ardue et l'étude a donc été menée numériquement par simulations. Les résultats numériques mettent en avant que si globalement le choix de h influe peu, il est préférable d'utiliser $h = 2$ pour des petites séquences et un h plus grand ($h \geq 4$) pour des séquences de longueur > 500.

Afin d'optimiser le temps de calcul, l'algorithme proposé par Nuel G. et Prum B. [43] pour le calcul des p-valeurs exactes a été utilisé.

En conclusion, cette approche reposant sur les h-uplets des séquences et la méthode exacte pour établir la loi du score local, constitue une alternative appropriée à la p-valeur proposée par les logiciels usuels de comparaison de séquences

(BLAST, FASTA, ...). Elle présente par ailleurs, dans le cas de séquences courtes à moyennes (≤ 500), une solution pertinente pour dégager des séquences significativement ressemblantes.

Afshin est assistant professeur à l'université de Mazandaran au nord de l'Iran et vice-président de la faculté de Mathématiques (2017).

2.5 Conclusion sur le cas des séquences i.i.d.

Le tableau ci-dessous résume l'existence de résultats sur la distribution du score local d'analyse d'une séquence dans le cas du modèle i.i.d. suivant les critères de longueur de la séquence n et le signe du score moyen $\mathbb{E}[X]$ (voir [4]) E correspond à la méthode exacte, K, l'approximation asymptotique de Karlin *et al.*, A l'amélioration de cette approximation et $\mathcal{L}$ (*respectivement* $\mathcal{P}$) des résultats relatifs à la longueur (*resp.* position) de réalisation du score local.

$\mathbb{E}[X]$	$n \leq 10$	$n \leq 10^2$	$n \simeq 10^3$	$n > 10^4$
< 0	E	$A + E$	$A + E + K$	K
$= 0$	E	E	$\simeq E$	$\mathcal{L} + \mathcal{P}$
> 0	E	E	$\simeq E$	

Les séquences courtes à moyennes constituent le domaine d'application de la méthode exacte. Les méthodes asymptotiques sont à utiliser pour des scores moyens négatifs et des séquences moyennes pour A et longues pour A et K. Les résultats sur le mouvement brownien sont spécifiques à de très longues séquences et un score moyen nul.

Chapter 3

Loi du score local de séquences markoviennes

Ce chapitre concerne les résultats sur la signification statistique du score local dans le cas de séquences modélisées par une chaîne de Markov (notée CM). Comme dans le chapitre précédent sur le modèle de séquences i.i.d., nous distinguons deux types de résultats principaux. Les premiers proposant une méthode exacte pour établir la p-valeur du score local (cf. Hassenforder et Mercier 2007 [25]). Ces résultats sont adaptés à des séquences courtes voire même beaucoup plus courtes que dans le cas du modèle i.i.d. Le second type de résultats correspond à des approximations asymptotiques, *i.e.* pour une longueur de séquences tendant vers l'infini, adaptées à des séquences plus longues, de plusieurs centaines de composants au moins. Là encore, comme dans le chapitre précédent, nous distinguons une approximation par une loi de Gumbel (Karlin et Dembo 1992 [31]) ainsi qu'une nouvelle approximation (Grusea et Mercier, article soumis [23]).

Il peut paraître surprenant de parler d'une approximation par une loi de Gumbel de Karlin et Dembo 1992 pour le *modèle markovien* alors que ce résultat n'est quasiment pas référencé dans la littérature malgré les besoins réels d'un tel résultat. Cela peut s'expliquer par différents éléments. Le premier repose sur le choix des auteurs d'une fonction de scores inhabituelle qui rend la compréhension du contexte et du modèle plus délicate : alors que les fonctions de scores associent à chaque composant une valeur de score déterministe, le travail de Karlin et Dembo repose sur une fonction de score aléatoire pour chaque paire des composants consécutifs. Il nous semble que ce choix a été en partie guidé par les besoins de la démonstration du résultat. En effet, cette fonction de scores aléatoire permet d'attribuer des propriétés d'irréductibilité à des matrices définies par les auteurs. Quoiqu'il en soit, le résultat de l'approximation par une loi de Gumbel de Karlin et Dembo dans le cas du modèle markovien, n'est pas référencé. Exception faite par Hansen 2006 [24] qui en fait référence dans le corps de son article pour le calcul d'un paramètre dont lui-même se sert, mais pas dans l'état

de l'art proposé dans l'introduction. Nous présentons ici cette approximation. Nous proposons également une nouvelle approximation basée sur une fonction de scores plus usuelle, avec des résultats plus précis, ainsi qu'une méthodologie de calcul pratique des paramètres du résultat.

3.1 Séquences courtes : méthode exacte

Collaborateurs : Claudie Chabriac-Hassenforder, UT2J, IMT
Scores entiers relatifs, score moyen quelconque
Théorie des chaînes de Markov
Publications : [25, 41]

Le résultat sur la distribution exacte du score local établi dans le modèle i.i.d. peut facilement s'étendre au cas d'une suite $(X_k)_k$ sous un modèle de Markov. Pour $(X_k)_k$ i.i.d. et $(U_k)_k$ le processus de Lindley associé (cf. (1.3)), le processus de Lindley arrêté en a, $(U_k^*)_k$, défini par $U_0^* := 0$, $U_k^* = max(U_{k-1} + X_k, 0)$ pour $k \leq \inf\{i : U_i \geq a\}$ et $U_k^* = a$ sinon, est une chaîne de Markov. Pour $(X_k)_k$ markovien, $(U_k^*)_k$ ne l'est pas. Par contre si l'on considère le processus $(Y_{k+1})_k = (U_k^*, X_{k+1})_k$ le processus $(Y_{k+1})_k$ est quant à lui une chaîne de Markov.

Dans le cas d'un modèle markovien sur les lettres $(A_k)_k$ et non pas sur la suite des scores $(X_k)_k$, avec $X_k = s(A_k)$, le processus à considérer est alors $(U_k^*, A_k, A_{k+1})_k$. Le résultat théorique se déduit facilement. Les mises en application (implémentation) sont quant à elles plus délicates, coûteuses en temps et en place mémoire.

Résultat 9 (Loi exacte - modèles markoviens [25])

1) Modèle markovien sur les scores

Soit $\mathbf{P} = (\mathbf{P}_{uv})_{u,v\in\mathbb{Z}}$ *la matrice des probabilités de transition de la chaîne de Markov* $(X_k)_k$ *et* π *sa distribution initiale. Soit* $a \in \mathbb{N}$ *une valeur observée de score local.*

On définit $\tilde{\mathbf{P}} = \left(\tilde{\mathbf{P}}_{(i,u)(j,v)}\right)_{(i,u);(j,v)\in E^2}$ *avec* $E = \{0, ..., a\}\times\{s_{\min}, ..., 0, ..., s_{\max}\}$, $s_{\min}$ *et* $s_{\max}$ *étant les scores maximum et minimum, par :*

$$\tilde{\mathbf{P}}_{(a,u)(a,v)} = \mathbf{P}_{uv} \text{ , et pour } 0 \leq i \leq a-1$$

$$\begin{cases} \tilde{\mathbf{P}}_{(i,u)(0,v)} = \mathbf{P}_{uv} & \text{if } i+u \leq 0 \\ \tilde{\mathbf{P}}_{(i,u)(i+u,v)} = \mathbf{P}_{uv} & \text{if } 1 \leq i+u \leq a-1 \quad \text{et} \quad \tilde{\mathbf{P}}_{(i,u)(j,v)} = 0 \text{ sinon.} \\ \tilde{\mathbf{P}}_{(i,u)(a,v)} = \mathbf{P}_{uv} & \text{if } i+u \geq a \end{cases}$$

La p-valeur du score local M_n *est alors donnée par*

$$(\forall a \geq 0) \quad P[M_n \geq a] = \sum_{u,v} \pi_u \cdot \tilde{\mathbf{P}}^{n-1}_{(0,u)(a,v)}. \tag{3.1}$$

2) Modèle markovien sur les lettres

Soit $\mathcal{A}$ l'ensemble des lettres codant les composants de la séquence. Soit $\mathbb{A}$ la séquence des lettres, chaîne de Markov d'ordre 1 de matrice de transition $\mathbf{P} = (\mathbf{P}_{\alpha,\beta})_{\alpha,\beta\in\mathcal{A}}$ et π la distribution initiale. Soit $\breve{E} = \{0, ..., a\} \times \mathcal{A}^2$ avec $a \in \mathbb{N}$ le score local observé.

Soit $\breve{\mathbf{P}} = \left(\breve{\mathbf{P}}_{(i,\alpha,\beta),(j,\gamma,\delta)}\right)$ une matrice, (i, α, β) et (j, γ, δ) dans $\breve{E}$, définie par :

$\breve{\mathbf{P}}_{(a,\alpha,\beta)(a,\beta,\delta)} = \mathbf{P}_{\beta,\delta}$,
et pour $0 \leq i \leq a-1$

$$\begin{array}{ll} \breve{\mathbf{P}}_{(i,\alpha,\beta)(0,\beta,\delta)} = \mathbf{P}_{\beta,\delta} & si\ s(\beta)+i \leq 0 \\ \breve{\mathbf{P}}_{(i,\alpha,\beta)(s(\beta)+i,\beta,\delta)} = \mathbf{P}_{\beta,\delta} & si\ 1 \leq s(\beta)+i \leq a-1 \\ \breve{\mathbf{P}}_{(i,\alpha,\beta)(a,\beta,\delta)} = \mathbf{P}_{\beta,\delta} & si\ s(\beta)+i \geq a \end{array}$$

$\breve{\mathbf{P}}_{(i,\alpha,\beta),(j,\gamma,\delta)} = 0$ *sinon.*

La p-valeur de M_n est alors donnée par

$$(\forall a \geq 0) \quad \mathbb{P}[M_n \geq a] = \sum_{\alpha,\beta,\gamma,\delta} \pi_\alpha \cdot \breve{\mathbf{P}}^{n-1}_{((s(\alpha),0)^+,\alpha,\beta)(a,\gamma,\delta)}. \tag{3.2}$$

C'est l'implémentation qui est coûteuse en place mémoire car les matrices sont de grande taille, même en tenant compte du fait que les matrices en question sont très creuses. Par exemple, pour le cas markovien sur les lettres dans le cas de protéines et pour un score local de 10, la matrice à implémenter à la puissance d'en moyenne 350 est de taille 4000×4000.

Comparaison des différents résultats sur données réelles

Dans [25], une comparaison de l'adéquation des différents types de résultats (approximation ou méthode exacte) sous les différents modèles est présentée. Pour cela, il a été nécessaire de dégager des outils de comparaison de ces différentes distributions. Trois outils de comparaison, autres que les graphiques usuellement disponibles dans la littérature, ont été alors utilisés.

- La P-value Slope Error (PSE) proposé par Bailey and Gribskov (2002) $PSE = 1 - m$ avec m la pente de la droite des moindres carrés pour les points $(\log(p_{emp}); \log(p_{théo}))$ avec p_{emp} une p-valeur empirique de type Monte Carlo et $p_{théo}$ calculée à l'aide des différents résultats (cf. Résultat 2 pour l'approximation de Karlin *et al.* dans le modèle i.i.d., Résultat 1 pour la méthode exacte cas i.i.d., et Résultat 9 pour la méthode exacte cas markovien).

- Nous proposons également la moyenne du carré des erreurs (Mean Square Error (MSE))

$$MSE = \frac{1}{\#a} \sum_a [\log(p_{emp}(a)) - \log(p_{théo}(a))]^2$$

avec a les valeurs de score locaux observés et $\#a$ le nombre de scores locaux différents observés.

Table 3.1: **Moyenne des trois mesures différentes** entre les valeurs empiriques et les valeurs théoriques (K pour l'approximation de Karlin *et al.*, M_0, resp. M_{1-X}, pour la méthode exacte utilisant le modèle M_0, resp. M_{1-X}) pour les séquences indépendantes et identiquement distribuées (IID) et les chaînes de Markov (CM).

	Séquences simulées IID			Séquences simulées CM		
	MSE	PSE	d_{KL}	MSE	PSE	d_{KL}
K	$9,47e{-2}$	$0,182$	$8,32e{-1}$	$3,85e{-1}$	$0,391$	$1,76$
M_0	$5,88e{-3}$	$0,026$	$2,66e{-3}$	$9,87e{-2}$	$0,153$	$4,20e{-2}$
M_{1-X}	$5,07e{-3}$	$0,022$	$2,74e{-3}$	$7,98e{-3}$	$0,038$	$3,66e{-3}$

- Nous avons également utilisé la distance de Kullback : soit $p = (p_1, \ldots, p_\kappa)$ et $q = (q_1, \ldots, q_\kappa)$ deux distributions discrètes. Ici, $p_k = p_{emp}(k)$ et $q_k = q_{théo}(k)$. Nous avons,

$$d_{KB}(p, q) = \sum_{k=1}^{\kappa} \log_2 \left(\frac{p_k}{q_k} \right) .$$

Les résultats des comparaisons numériques sont rassemblés dans les tables 3.1 et 3.1.

La première correspond à des comparaisons sur des séquences simulées, i.i.d. et markoviennes. La seconde à des séquences réelles. Nous pouvons constater dans le premier tableau, regardons le cas des séquences simulées sous le modèle i.i.d. (partie gauche du tableau) : nous pouvons constater que les différentes mesures MSE, PSE, d_{KL} calculées pour les méthodes théoriques reposant sur le modèle i.i.d. et celles sous le modèle markovien sont similaires comme attendues. Dans le cas des séquences simulées sous le modèle markovien (partie droite du tableau), les mesures de comparaison sont nettement inférieures dans le cas des valeurs théoriques reposant sur le modèle markovien : le gain de la prise en compte d'un modèle de Markov est important ; les méthodes étant exactes, elles nous permettent de plus d'évaluer l'erreur effectuée pour calculer une p-valeur du score local lorsque l'on utilise un modèle i.i.d. alors que le modèle réel est markovien, cette erreur n'étant due qu'au choix de modèle et non aux approximations des méthodes. De la même façon, les valeurs des différentes mesures reposant sur les méthodes exactes nous permettent d'estimer le degré de précision de nos critères de comparaison : une distance de Kullback est calculée à 10^{-3} près ; une PSE de l'ordre de 10^{-2} doit être considérée comme très proche de 0, voire nulle...

Les données réelles utilisées sont des fichiers de séquences de SCOP (Structural Classification Of Proteins, http://scop2.mrc-lmb.cam.ac.uk/downloads/). La fonction de scores utilisée prend ses valeurs de -5 à +5 et vérifie $\mathbb{E}[X] < 0$ pour que l'utilisation de l'approximation de Karlin *et al.* (cf. Section 2.2) soit possible.

Les séquences réelles de SCOP ont été tronquées pour une longueur homogène sur toutes les séquences de la base de données de $n = 100$, afin de pouvoir calculer une p-valeur empirique pour une longueur n fixée et effectuer des comparaisons.

Le gain du modèle markovien par rapport au modèle i.i.d. est assez net dans le cas de vraies séquences. Par contre, il semble que l'intérêt d'un modèle markovien sur les lettres par rapport au modèle markovien sur les scores ne semble pas nettement justifié d'autant que les temps de calculs sont beaucoup plus importants pour le modèle sur les lettres. La méthode exacte devance l'approximation de Karlin *et al.* dans le cas de petites séquences, ici $n = 100$.

	SCOP database ($n = 100$)					
	$a \leq 35$			$a \leq 9$		
	MSE	PSE	d_{KL}	MSE	PSE	d_{KL}
K	2.91e-2	0.102	6.44e-2	5.16e-4	-0.169	3.84e-2
i.i.d.	2.89e-2	0.095	6.91e-2	3.18e-4	0.267	1.94e-2
M_{1-X}	1.58e-2	0.061	6.14e-2	1.60e-4	0.138	1.19e-2
M_{1-A}	-	-	-	1.18e-4	0.105	9.91e-3

Table 3.2: PSE correspond à la P-value Slope Error, MSE à la Mean Square Error, d_{KL} à la distance de Kullback. Les différentes valeurs théoriques sont notées K pour l'approximation de Karlin *et al.* dans le modèle i.i.d., i.i.d. (*respectivement* M_{1-X}, M_{1-A}) les p-valeurs exactes suivant les modèles i.i.d. (*resp.* modèle markovien sur les scores, modèle markovien sur les lettres). Pour le cas du modèle markovien sur les lettres les calculs ne sont effectués que pour des petites valeurs a de score observés ($a \leq 9$) pour une raison de temps de calculs.

Les implémentations en langage C++ des méthodes markoviennes ont été effectuées par des stagiaires : Périès Laurent de l'INSA, 4-ième année en Génie Informatique 2004, pour la partie modèle markovien sur les lettres ; Emmanuel en seconde année de l'IUP SID, Université de Paul Sabatier, en 2004 pour le modèle markovien pour les scores ; Bénédicte Urbano, L3 MIASHS Université de Toulouse 2 Jean Jaurès 2002, pour la méthode exacte, l'approximation de Karlin *et al.* et son amélioration dans le cas du modèle i.i.d.

3.2 Séquences longues : approximations

Collaboratrice : Simona Grusea, INSA Toulouse, IMT

Hypothèses : scores entiers relatifs, score moyen négatif

Approche par la théorie des chaînes de Markov, du renouvellement et la théorie des grandes déviations

Publications : article soumis en mars 2018

Nous abordons dans cette section les approximations de la distribution du score local d'une chaîne de Markov. Nous commençons par rappeler les résultats de Karlin et Dembo [31] dans le cas d'une fonction de scores plus générale. Puis nous présentons une nouvelle approximation dans le cas d'une fonction de scores lattice et à valeurs dans un ensemble fini (cf. Définition 3.3).

Une variable aléatoire X est dite lattice si

$$\exists d \geq 0 : \sum_{k \in \mathbb{Z}} \mathbb{P}(X = kd) = 1 \ . \tag{3.3}$$

3.2.1 Approximation par une loi de Gumbel (Karlin et Dembo 92, [31])

Karlin et Dembo dans [31] présentent leur modèle de la façon suivante, seules les notations ont été changées pour correspondre à celles utilisées dans le document présent :

"Let $\mathbb{A} = (A_i)_{1 \leq i \leq n}$ be generated governed by an r-state irreducible aperiodic Markov chain. The partial sum process $S_{\alpha,k} = \sum_{i=0}^{k-1} X_{A_i A_{i+1}}$, $k = 1, 2, \ldots$ is determined by a realization $(a_i)_{i=0}^{n}$ of states with $a_0 = \alpha$ and the real-valued i.i.d. bounded variables $X_{\alpha\beta}$ associated with the transitions $a_i = \alpha$ et $a_{i+1} = \beta$. Assume $X_{\alpha\beta}$ has negative stationary mean. The explicit limit distribution of the maximal segmental sum $M_n := \max_{0 \leqslant k \leqslant \ell \leqslant n}(S_{\alpha,\ell} - S_{\alpha,k})$is derived..."

Prenons le temps de bien comprendre ce modèle. Les auteurs associent à chaque transition $(A_{i-1}, A_i) = (\alpha, \beta)$ un score qui est une v.a.r. bornée $X_{\alpha\beta}$ et dont la loi dépend de (α, β). Ils supposent de plus que pour deux couples $(A_{i-1}, A_i) = (\alpha, \beta)$ et $(A_{j-1}, A_j) = (\alpha, \beta)$, les scores associés $X_{\alpha\beta}$ (en i, respectivement en j) sont i.i.d.

Usuellement, on associe à A_i un score $s(A_i)$ ce qui est un cas particulier du modèle de Karlin et Dembo. En effet, la loi $\mathcal{L}(X_{\alpha\beta}) = \delta_{s(\beta)}$ pour une transition (α, β) donnée, est un score déterministe et dépend seulement de β via la fonction s. On a en fait $s(A_i) = X_{A_{i-1}A_i}$. Cette approche d'un score qui dépend de paires de sites consécutifs et qui est en plus aléatoire n'a pas aidé à la diffusion du résultat des auteurs et nous pensons qu'il s'agit d'une des raisons pour lesquelles l'approximation par une loi de Gumbel dans le cas markovien n'a pas été référencée. Nous pensons que ce choix de fonction de score a été en partie motivé par le fait d'assurer certaines propriétés aux matrices utilisées dans la démonstration pour en faciliter cette dernière (irréductibilité de la matrice $G(\infty)$ par exemple).

La démonstration du résultat utilise le schéma identique à celui du modèle i.i.d. c'est-à-dire que les auteurs établissent une approximation de la loi du maximum des sommes partielles S^+. Ils portent ensuite ce résultat à la loi de la hauteur de la première excursion $Q_1 = S_{T_1}$; gèrent le passage à n'importe quelle excursion, puis au score local.

La section suivante qui présente une nouvelle approximation reprendra ce schéma mais reposera, elle, tout comme dans le chapitre précédent dans le modèle i.i.d., sur la loi *exacte* du maximum des sommes partielles.

Nous donnons ci-dessous les résultats des trois étapes de l'article de Karlin et Dembo [31]. De cette manière, les apports de notre travail exposés dans la section suivante apparaîtrons plus clairement. Rappelons tout d'abord les hypothèses de travail ainsi que certaines notations nécessaires.

Notations, définitions : Soit $(A_i)_i$ une CM irréductible, apériodique et stationnaire à valeurs dans un espace d'état fini $\mathcal{A}$ à r états $(\alpha, \beta, \dots)$, de matrice de transition $\mathbf{P} = (p_{\alpha\beta})_{\alpha,\beta}$, avec $p_{\alpha\beta} > 0$ pour tout α, β, et de vecteur de fréquences stationnaires $(\pi_\alpha)_\alpha$. Soit $s : \mathcal{A} \to \mathbb{R}$ une fonction de scores.
Hypothèses : Score moyen négatif et score positif possible :

$$\mathbb{E}[s(A_i)] = \sum_\alpha s(\alpha)\pi_\alpha < 0. \tag{3.4}$$

$$\forall \alpha,\ \mathbb{P}_\alpha(s(A_1) > 0) > 0, \quad \mathbb{P}_\alpha(s(A_1) < 0) > 0. \tag{3.5}$$

Rappelons que la CM est stationnaire car irréductible sur un espace d'états fini. La loi de A_i est π pour tout i. On notera $\mathbb{E}[s(A)]$ le score moyen.

On considère pour θ réel, $\mathbf{\Phi}(\theta) := (p_{\alpha\beta} \cdot \exp(\theta f(\beta)))_{\alpha,\beta}$. Cette matrice est strictement positive et par le Théorème de Frobénius son rayon spectral noté $\rho(\theta)$ admet un vecteur propre à droite normalisé strictement positif $u(\theta) = (u_1(\theta), \dots, u_r(\theta))$.

Soit $\sigma^- := \inf\{k \geqslant 1 : S_k < 0\}$, temps d'arrêt fini *p.s.* grâce à l'hypothèse d'un score moyen négatif. Soit $Q_1 := \max_{0 \leqslant k \leqslant \sigma^-} S_k$ le maximum de la première excursion positive. On note $q_{\alpha\beta} := \mathbb{P}_\alpha(A_{\sigma^-} = \beta)$ et $\mathbf{Q} = (q_{\alpha\beta})_{\alpha,\beta}$. Les hypothèses assurent que c'est une matrice stochastique irréductible.

Soit $T_0 := 0$ et pour tout $i \geq 1$,

$$T_i := \inf\{k > T_{i-1} : S_k - S_{T_{i-1}} < 0\}.$$

Le temps d'arrêt T_i détermine la fin de la i-ème excursion positive. Notons que $T_1 = \sigma^-$. A chaque excursion positive $i \geq 1$ on associe un score maximal Q_i défini par

$$Q_i := \max_{T_{i-1} \leq k \leq T_i} (S_k - S_{T_{i-1}}).$$

Les états $(A_{T_i})_{i \geq 1}$ de la chaîne de Markov à la fin de chaque excursion positive forment également une chaîne de Markov irréductible de matrice de transition $\mathbf{Q}$. On a donc

$$q_{\alpha\beta} = \mathbb{P}(A_{T_i} = \beta \mid A_{T_{i-1}} = \alpha)$$

Grâce à la propriété de Markov forte, conditionnellement aux $(A_{T_i})_{i \geq 1}$, les v.a. $Q_1, Q_2, \dots$ sont indépendantes. Pour tout α, β et pour tout $i \geq 1$, on note

$F_{\alpha\beta}$ la fonction de répartition de la "hauteur d'une montagne" sachant qu'elle commence en α et qu'elle finit en β, i.e.

$$F_{\alpha\beta}(y) := \mathbb{P}(Q_i \leq y \mid A_{T_i} = \beta, A_{T_{i-1}} = \alpha) = \mathbb{P}_\alpha(Q_1 \leq y \mid A_{\sigma^-} = \beta).$$

Pour tout α, soit

$$F_\alpha(y) := \mathbb{P}(Q_i \leq y \mid A_{T_{i-1}} = \alpha) = \mathbb{P}_\alpha(Q_1 \leq y)$$

la fonction de répartition de la "hauteur d'une montagne" sachant qu'elle commence en α. Par la formule des probabilités totales, on a $F_\alpha(y) = \sum_\beta F_{\alpha\beta}(y) q_{\alpha\beta}$.

On définit ensuite $\sigma^+ := \inf\{k \geqslant 1 : S_k > 0\}$, qui est un temps d'arrêt à valeurs dans $\mathbb{N} \cup \{\infty\}$. A cause de la dérive négative (score moyen négatif), on a $\mathbb{P}_\alpha(\sigma^+ < \infty) < 1$ pour tout α.

Soit

$$L_{\alpha\beta}(\xi) := \mathbb{P}_\alpha(S_{\sigma^+} \leqslant \xi, \sigma^+ < \infty, A_{\sigma^+} = \beta)$$

la probabilité que la première excursion positive, commençant en α, finisse en β et ne dépasse pas ξ. Notons que $L_{\alpha\beta}(\infty) \leq \mathbb{P}_\alpha(\sigma^+ < \infty) < 1$, et donc

$$\int_0^\infty dL_{\alpha\beta}(\xi) = 1 - L_{\alpha\beta}(\infty) > 0. \tag{3.6}$$

Soit aussi

$$L_\alpha(\xi) := \sum_\beta L_{\alpha\beta}(\xi) = \mathbb{P}_\alpha(S_{\sigma^+} \leqslant \xi, \sigma^+ < \infty).$$

De même que dans le cas du modèle i.i.d., les hypothèses (3.4) et (3.5) nous assurent qu'il existe un unique $\theta^* > 0$ tel que

$$\rho(\theta^*) = 1\ . \tag{3.7}$$

On définit alors pour tout $x > 0$

$$G_{\alpha\beta}(x) := \frac{u_\beta(\theta^*)}{u_\alpha(\theta^*)} \int_0^x \exp(\theta^* \xi) dL_{\alpha\beta}(\xi) = \int_0^x dG_{\alpha\beta}(\xi),$$

avec $dG_{\alpha\beta}(\xi) = \frac{u_\beta(\theta^*)}{u_\alpha(\theta^*)} dL_{\alpha\beta}(\xi)$. On a donc

$$G_{\alpha\beta}(\infty) = \frac{u_\beta(\theta^*)}{u_\alpha(\theta^*)} \int_0^\infty \exp(\theta^* \xi) dL_{\alpha\beta}(\xi)$$

et la matrice $G(\infty) := (G_{\alpha\beta}(\infty))_{\alpha,\beta}$ est stochastique strictement positive, donc irréductible (la stricte positivité découle de la relation (3.6) et du fait que $u(\theta^*) > 0$).

Soit aussi $w = (w_1, \ldots, w_r) > 0$ le vecteur des fréquences stationnaires de la matrice $G(\infty)$.

Lemme 1 (Limite pour loi de S^+ - Lemma 4.3 de [31])
On a

$$\lim_{x\to+\infty} \frac{e^{\theta^* x}\mathbb{P}_\alpha(S^+ > x)}{u_\alpha(\theta^*)} = c(\infty),$$

avec

$$c(\infty) = \frac{\sum_\gamma w_\gamma \hat{a}_\gamma}{\sum_{\gamma,\beta} w_\gamma \int_0^\infty \xi \; dG_{\gamma\beta}(\xi)},$$

où $w = (w_1, \ldots, w_r)$ est le vecteur des fréquences stationnaires de la matrice stochastique $G(\infty)$ (on a $w > 0$ car $G(\infty)$ est irréductible) et

$$\hat{a}_\alpha = \frac{1}{u_\alpha(\theta^*)} \int_0^\infty (L_\alpha(\infty) - L_\alpha(x)) e^{\theta^* x} dx.$$

Lemme 2 (Limite pour la loi de Q_1 - Lemma 4.4 de [31])
On a

$$\lim_{y\to+\infty} e^{\theta^* y}\mathbb{P}_\alpha(Q_1 > y) = \lim_{y\to+\infty} e^{\theta^* y}(1 - F_\alpha(y)) = e_\alpha(\infty)$$

avec

$$e_\alpha(\infty) := c(\infty) \cdot \left[1 - \sum_\beta u_\beta(\theta^*) \int_{-\infty}^0 e^{\theta^* \eta} dK_{\alpha\beta}(\eta)\right]$$

$$\text{et } K^{(y)}_{\alpha\beta}(\eta) := \mathbb{P}_\alpha(A_{\sigma^-} = \beta, S_{\sigma^-} \leqslant \eta, Q_1 \leqslant y) \; .$$

Résultat 10 (Loi de Gumbel pour le score local - Formule (1.27) de [31])
On a

$$\lim_{n\to+\infty} \mathbb{P}\left(M_n - \frac{\ln n}{\theta^*} \leqslant x\right) = \exp\left(-K^* \cdot e^{\theta^* x}\right)$$

où θ^* est défini par (3.7), $K^* = C^*/A^*$ par

$$C^* = \sum_{\beta=1}^r z_\beta \cdot e_\beta(\infty)$$

$$\text{et } A^* = \frac{1}{\mathbb{E}[X] \cdot \sum_{\beta=1}^r z_\beta \cdot \mathbb{E}[S_{\sigma^-} | A_0 = \beta]} \; ,$$

et z est le vecteur de probabilité invariant pour la matrice Q.

Dans la Section 5 de [31], les auteurs proposent une méthodologie pour implémenter les résultats précédents dans le cas lattice de pas $d = 1$.

3.2.2 Nouvelle approximation

Nous nous plaçons à nouveau sous les hypothèses (3.4) et (3.5). Supposons de plus la fonction de scores s à valeurs dans $\{-u,\dots,0,\dots,+v\}$, et donc lattice de pas $d=1$. Nous établissons par récurrence la loi exacte du maximum des sommes partielles S^+ puis proposons une approximation pour la loi de Q_1 et de M_n. Notons

$$L_{\alpha\beta}^{(\ell)} := \mathbb{P}_\alpha(S_{\sigma^+}=\ell d, \sigma^+<\infty, A_{\sigma^+}=\beta)\ ,$$

et

$$F_{S^+,\alpha}(\ell d) = \mathbb{P}_\alpha(S^+\le \ell d).$$

Résultat 11 (Distribution exacte de S^+) *$\forall \ell \ge 1$ et $\alpha\in\mathcal{A}$ nous avons*

$$F_{S^+,\alpha}(0) = \mathbb{P}_\alpha(\sigma^+=\infty) = 1-L_\alpha(\infty),$$
$$F_{S^+,\alpha}(\ell d) = 1-L_\alpha(\infty) + \sum_\beta \sum_{k=1}^{\ell} L_{\alpha\beta}^{(k)}\ F_{S^+,\beta}((\ell-k)d).$$

Remarque 3 *Pour $\alpha\in\mathcal{A}$, les $F_{S^+,\alpha}(\ell)$ peuvent être déterminés récursivement en ℓ et les valeurs $(F_{S^+,\alpha}(0))_\alpha$ peuvent être établies par résolution de r équations à r inconnus.*

Pour $\alpha,\beta\in\mathcal{A}$, les $L_{\alpha\beta}^{(\ell)} = \mathbb{P}_\alpha = (\sigma^+<\infty, A_{\sigma^+}=\beta, S_{\sigma^+}=\ell)$ peuvent également être déterminés récursivement en ℓ.

Résultat 12 (Nouvelle approximation de la distribution de Q_1)

$$\mathbb{P}_\alpha(Q_1>y) \underset{y\to\infty}{\sim} \mathbb{P}_\alpha(S^+>y) - \sum_{x<0}\sum_y \mathbb{P}_\beta(S^+>y-x)\cdot\mathbb{P}_\alpha(S_{\sigma^-}=x, A_{\sigma^-}=\beta)$$

avec les $\mathbb{P}_\alpha(S_{\sigma^-}=x, A_{\sigma^-}=\beta) := Q_{\alpha\beta}^{(x)}$, $\alpha,\beta\in\mathcal{A}$, $x\in\{-1,\dots,-u\}$, déterminés comme dans [31].

Résultat 13 (Nouvelle approximation de la distribution de M_n)

$$\mathbb{P}_\alpha\left(M_n\le\frac{\log(n)}{\theta^*}+x\right) \underset{n\to\infty}{\sim} \exp\left\{-\frac{n}{A^*}\sum_\beta z_\beta \mathbb{P}_\beta\left(S^+>\left\lfloor\frac{\log(n)}{\theta^*}+x\right\rfloor\right)\right\}$$
$$\times\exp\left\{\frac{n}{A^*}\sum_{k<0}\sum_\gamma \mathbb{P}_\gamma\left(S^+>\left\lfloor\frac{\log(n)}{\theta^*}+x\right\rfloor - kd\right)\cdot\sum_\beta z_\beta Q_{\beta\gamma}^{(k)}\right\}, \qquad (3.8)$$

avec

$$A^* := \lim_{m\to+\infty}\frac{T_m}{m} = \frac{1}{\mathbb{E}[f(A)]}\sum_\beta z_\beta \mathbb{E}_\beta[S_{\sigma^-}]\ \ a.s.$$

et $z=(z_\alpha)_{\alpha\in\mathcal{A}}$ la mesure de probabilité invariante associée à la matrice $\mathbf{Q}$. Notez que le terme de droite de (3.8) ne dépend pas de l'état initial α.

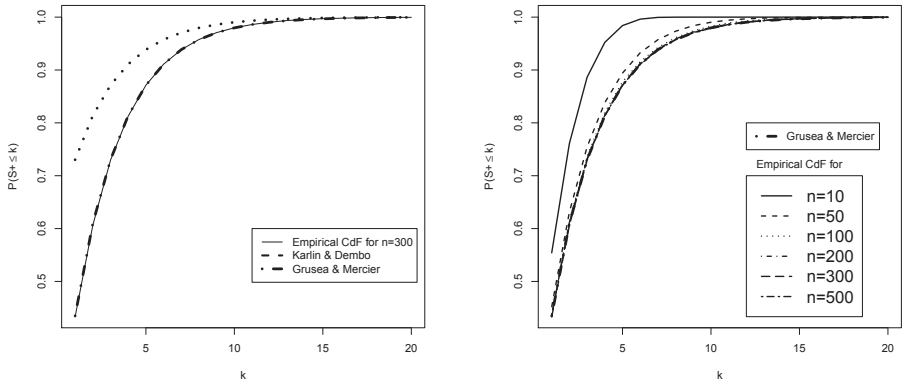


Figure 3.1: Fonction de répartition de S^+, pour une séquence génomique avec $\mathcal{A} = \{A, C, G, T\}$, une échelle de scores $(-1, 0, +1)$, $A_0 =$"A". Paneau gauche : comparaison entre l'approximation de Karlin et Dembo proposée dans [31], une estimation de Monte Carlo pour 10^5 répétitions et notre formule proposée dans le Résultat 11. Paneau droit : comparaison pour différentes valeurs de n entre la fonction de répartition empirique obtenue par Monte Carlo et le résultat exact 11.

Une méthodologie est proposée dans [23] pour implémenter ces différents résultats. Une application numérique, reposant sur une écriture matricielle des résultats, est donnée dans le cas de séquences génomiques avec $\mathcal{A} = \{A, C, G, T\}$, une échelle de scores simples $(-1, 0, +1)$ et une matrice de probabilités de transition suivante

$$\mathbf{P} = \begin{pmatrix} 1/2 & 1/6 & 1/6 & 1/6 \\ 1/4 & 1/4 & 1/4 & 1/4 \\ 1/6 & 1/6 & 1/6 & 1/2 \\ 1/6 & 1/6 & 1/2 & 1/6 \end{pmatrix} .$$

Une comparaison des différentes valeurs des distributions de S^+, Q_1 et M_n de Karlin et Dembo [31] est également proposée (voir Figures 3.1 et 3.2). C'est à notre connaissance la première fois qu'une implémentation des résultats markoviens de Karlin et Dembo est présentée.

3.2.3 Perspectives de travail dans le cas markovien

Nombre d'excursions dépassant un seuil donné (cf. Hansen 2006, [24])

Dans [24], Hansen considère le score local de comparaison de deux CM sans insertions/délétions. Ce travail généralise celui de Karlin *et al.* [18] dans le cas de deux séquences i.i.d. Hansen met en évidence des conditions suffisantes pour que la distribution d'alignements locaux de score dépassant un certain seuil donné soit asymptotiquement distribuée selon une loi de Poisson. Il nous semble

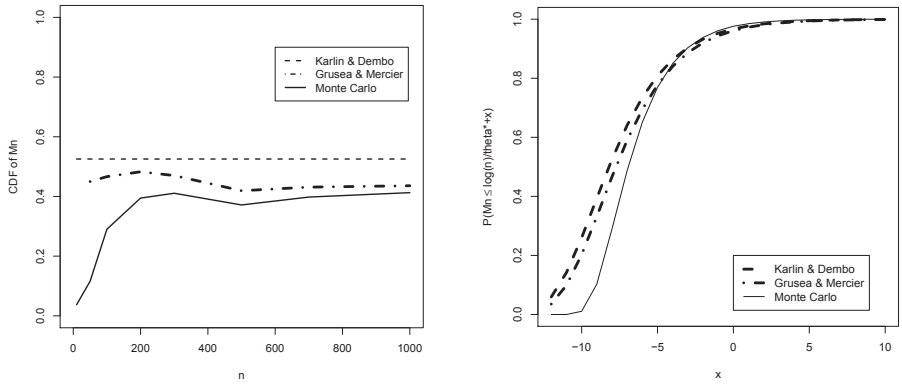


Figure 3.2: Paneau gauche : Comparaison entre différentes approximations pour $\mathbb{P}\left(M_n \leq \frac{\log(n)}{\theta^*} + x\right)$: le résultat de Karlin et Dembo [31], l'approximation proposée dans le Résultat 13 et une approximation de Monte Carlo. Schéma de scores $-1, 0, +1$, $x = -8$ fixé et n variant. Paneau droit : Differentes approximations pour $\mathbb{P}\left(M_n \leq \frac{\log(n)}{\theta^*} + x\right)$: le résultat de Karlin et Dembo [31], celui proposé dans le Résultat 13 et une approximation de Monte Carlo, en fonction de x, pour $n = 100$ et la fonction de scores $-1, 0, +1$.

intéressant d'étudier le travail de Hansen afin de le porter au cas d'analyse des scores de montagnes d'une CM et non au cas de la comparaison de deux CM. Cela permettrait une approche portant non plus sur uniquement la plus haute des excusions, *i.e.* celle qui réalise le score local, mais une excursion quelconque. En effet, cette approche est parfois celle privilégiée par les biologistes car ils souhaitent détecter toute région significative, et non pas seulement la région la plus significative (cf. Chapitre 5).

Cette perspective constituerait une première étape pour prendre en compte des corrections nécessaires dues aux tests multiples effectués. Il est facile de montrer que dans le modèle i.i.d. effectuer une correction classique de type Bonferroni sur le test de significativité de toutes les excursions revient à utiliser un seuil de la loi du score local (voir Section 5.1 pour le cas i.i.d.). Mais cela n'est pas aussi évident dans le cas markovien.

Ce projet de recherche est plus largement développé dans la Partie III de ce document.

3.3 Synthèse pour le modèle markovien

Le tableau ci-dessous résume la disponibilité et l'application possible de résultats sur la distribution du score local d'analyse d'une séquence dans le cas d'une

chaîne de Markov suivant les critères de longueur de la séquence n et le signe du score moyen $\mathbb{E}[X]$: E_{CM-A} correspond à la méthode exacte pour un modèle markovien sur les lettres , E_{CM-X} correspond à la méthode exacte pour un modèle markovien sur la séquence des scores ; K_{CM}, l'approximation asymptotique de Karlin *et al.* dans le cas markovien, A_{CM} l'amélioration de cette approximation.

$\mathbb{E}[X]$	$n \leq 10$	$n \leq 10^2$	$n \simeq 10^3$	$n > 10^4$
< 0	$E_{CM-X} + E_{CM-A}$	$A_{CM} + E_{CM-X}$	$A_{CM} + K_{CM}$	$K_{CM} + A_{CM}$
$= 0$	$E_{CM-X} + E_{CM-A}$	E_{CM-X}		
> 0	$E_{CM-X} + E_{CM-A}$	E_{CM-X}		

Les séquences courtes à moyennes constituent de même que dans le modèle i.i.d. le domaine d'application des méthodes exactes. La méthode exacte sur les lettres n'étant raisonnable que pour des longueur de séquences petites pour des usages répétitifs. Les méthodes asymptotiques A_{CM} et K_{CM} sont à utiliser pour des scores moyens négatifs et des séquences moyennes à longues. Nous avons noté que pour des cas de fonctions de scores très simples, comme $(-1, 0 + 1)$, l'approximation K_{CM} donnait déjà de bons résultats.

Chapter 4

Score local et modèle de Chaînes de Markov cachées

Collaborateur : Grégory Nuel (Laboratoire de Probabilités et Modèles Aléatoires (LPMA), Paris 6)
Scores réels, score moyen négatif
Publications : Article soumis [42]

Ce travail correspond à une approche différente de celle présentée précédemment. En effet, dans les chapitres précédents des résultats sur la distribution du score local d'analyse et de comparaison, ou bien sur des informations relatives autour du score local (longueur, position) sont établis sur des séquences aléatoires. Il s'agit dans ce chapitre d'établir des probabilités a posteriori conditionnellement à l'observation de la séquence étudiée. Par ailleurs, nous allons étendre l'intérêt non plus exclusivement au segment de plus grand score mais à l'information de la séquence dans son ensemble.

Dans un premier temps, définissons la notion de segmentation et le lien avec le score local, ainsi que le modèle génératif.

4.1 Définitions et notations

4.1.1 Segmentation et score local

Pour une séquence donnée $\mathbb{A} = (A_i)_{1 \leq i \leq n} \in \mathcal{A}^n$, avec $\mathcal{A} = \{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}, \}$ par exemple, $s : \mathcal{A} \to \mathbb{R}$ une fonction de score, nous noterons dans ce chapitre $M_s(\mathbb{A})$ le score local associé à s afin de mettre en évidence la fonction de scores utilisée. De plus, on écrit

$$M_s(\mathbb{A}) := \max_{[i,j]} \sum_{k \in [i,j]} s(A_k)$$

où $[i,j]$ peut éventuellement être vide et ainsi $M_s(\mathbb{A}) \geq 0$. On définit l'ensemble des segmentations d'intérêt $\mathcal{S}$ de la manière suivante :

$$\begin{aligned}\mathcal{S} \quad := \quad & \{S=(S_i)_{0\leq i\leq n+1} \in \{1,2,3\}^{n+2}, \quad \text{telle que} \\ & S_0 = 1, S_{n+1} \neq 2, S_i - S_{i-1} \in \{0,1\} \text{ pour } i = 1,\ldots,n+1\} \ .\end{aligned}$$

L'état 1 correspond à "avant un segment atypique", l'état 2 à "dans un segment atypique" et l'état 3 à "après un segment atypique".

Remarque 4 *On peut établir une bijection entre $\mathcal{S}$ et $\mathcal{I} = \{[i,j] : 1 \leq i \leq j \leq n\} \cup \emptyset$. Pour tout $S \in \mathcal{S}$, on définit $I = I_S$ tel que $I_S = \{i \in \{1,\ldots,n\} : S_i = 2\}$. Inversement, pour tout intervalle éventuellement vide $I = [i,j] \in \mathcal{I}$, $\exists S_I \in \mathcal{S}$ tel que $S_k = 1 \times \mathbf{1}_{k<i} + 2 \times \mathbf{1}_{k\in[i,j]} + 3 \times \mathbf{1}_{k>j}$.*

Dans l'exemple suivant, $I = [2,4]$ et $n = 5$.

Indice	=	0	1	*2*	*3*	*4*	5	6
S	=	1	1	*2*	*2*	*2*	3	3

Pour toute séquence $\mathbb{A} \in \mathcal{A}^n$ et toute segmentation $S \in \mathcal{S}$, on définit

$$M_s(S|\mathbb{A}) := \sum_{i=1}^{n} \mathbb{1}_{\{S_i=2\}} \cdot s(A_i) = \sum_{i\in I_S} s(A_i)$$

Remarque 5 *Ainsi nous avons*

$$M_s(\mathbb{A}) = \max_{S\in\mathcal{S}} M_s(S|\mathbb{A}). \tag{4.1}$$

4.1.2 Espace probabilisé des segments d'une séquence

L'approche par score local se focalise sur le (voire les) segment(s) réalisant le score additif maximal. L'idée ici est de s'intéresser à l'ensemble des segments, ensemble que nous allons probabiliser de la manière suivante.

$$\forall I = [i,j] \in \mathcal{I}, \quad \mathbb{P}_{(f,T)}(I|\mathbb{A}) \propto \exp\left(\frac{1}{T}\sum_{k=i}^{j} f(A_k)\right) \tag{4.2}$$

Cette distribution est connue comme la distribution de Gibbs largement utilisée en statistique physique. Le paramètre T est alors appellé température. Remarquons que $\mathbb{P}_{(f,T)}(I|\mathbb{A})$ tend vers une distribution de Dirac pour $T \to 0$, pour lequel nous retrouvons l'approche du score local qui se focalise sur le(s) segment(s) réalisant le maximum sans s'intéresser aux segments subopitmaux. Et $\mathbb{P}_{(f,T)}(I|\mathbb{A})$ tend vers une loi uniforme pour $T \to \infty$. La température est un paramètre de contraste. Afin d'illustrer cela, considérons une séquence

simulée de longueur $n = 40$ à valeurs dans $\{A, C, G, T\}$ et les scores correspondants $\{-2; -2; 1; 1\}$. La Figure 4.1 représente les probabilités des 821 segments différents possibles, pour une température $T = 0.1$ (panneau du haut), $T = 0.6$ (panneau du milieu) and $T = 5$ (panneau du bas). Le segment réalisant le score local est mise en évidence dans le panneau de gauche comme le seul segment significatif. Dans le panneau du milieu, des segments suboptimaux apparaissent. Pour une température plus élevée, la plupart des segments ont une probabilité d'un même ordre de grandeur. La question est ici de choisir une valeur de T adaptée pour mettre en évidence l'information intéressante dans la séquence dans son entier.

Nous allons voir dans ce chapitre qu'il est possible de déterminer, sous des hypothèses qui seront précisées, une valeur *canonique* de T.

4.1.3 Modèle génératif

Notons $\mathcal{D}_{\mathcal{A}}$ l'ensemble des distributions multinomiales sur $\mathcal{A}$. Soient $q_0(\cdot)$ et $q_1(\cdot) \in \mathcal{D}_{\mathcal{A}}$. Pour $\mathbb{A} \in \mathcal{A}^n$ et $S \in \mathcal{S}$ on définit

$$\mathbb{P}(\mathbb{A}|S) := \prod_{i=1}^{n} \mathbb{P}(A_i|S_i) = \prod_{i=1}^{n} q_0(A_i)^{\mathbb{1}_{\{S_i \neq 2\}}} q_1(A_i)^{\mathbb{1}_{\{S_i = 2\}}}$$

avec $\mathbb{P}(A_i|S_i \neq 2) = q_0(A_i)$ et $\mathbb{P}(A_i|S_i = 2) = q_1(A_i)$ (q_0 pour générer A_i quand $S_i \in \{1, 3\}$ et q_1 pour générer A_i quand $S_i = 2$).

Supposons de plus toutes les segmentations équiprobables. Alors

$$\mathbb{P}(S|\mathbb{A}) = \mathbb{P}(\mathbb{A}|S)/Z \tag{4.3}$$

avec

$$Z := \sum_{S \in \mathcal{S}} \mathbb{P}(\mathbb{A}|S) = \sum_{S \in \{1,2,3\}^{n+2}} \mathbb{1}_{\{S \in \mathcal{S}\}} \mathbb{P}(\mathbb{A}|S)$$

On remarquera que le modèle résultant est une chaîne de Markov cachée où $\mathbb{A}$ sont les observations et S est la séquence cachée (contrainte).

Une des nombreuses possibilités de cette approche est de permettre le calcul de $\mathbb{P}(S_j = 2|\mathbb{A})$, probabilité conditionnellement à la séquence observée d'être à l'indice i de la séquence dans un segment atypique. Mais aussi :

- Probabilité conditionnelle que l'indice i soit le début d'un segment,
- Probabilité conditionnelle que l'indice i soit la fin d'un segment,
- Probabilité conditionnelle que le segment contienne un segment de longueur ℓ,

ce qui sera abordé et illustré dans les sections suivantes.

La Section 4.2 qui suit présente les résultats théoriques établissant les conditions d'équivalence des deux approches permettant de mettre en évidence un

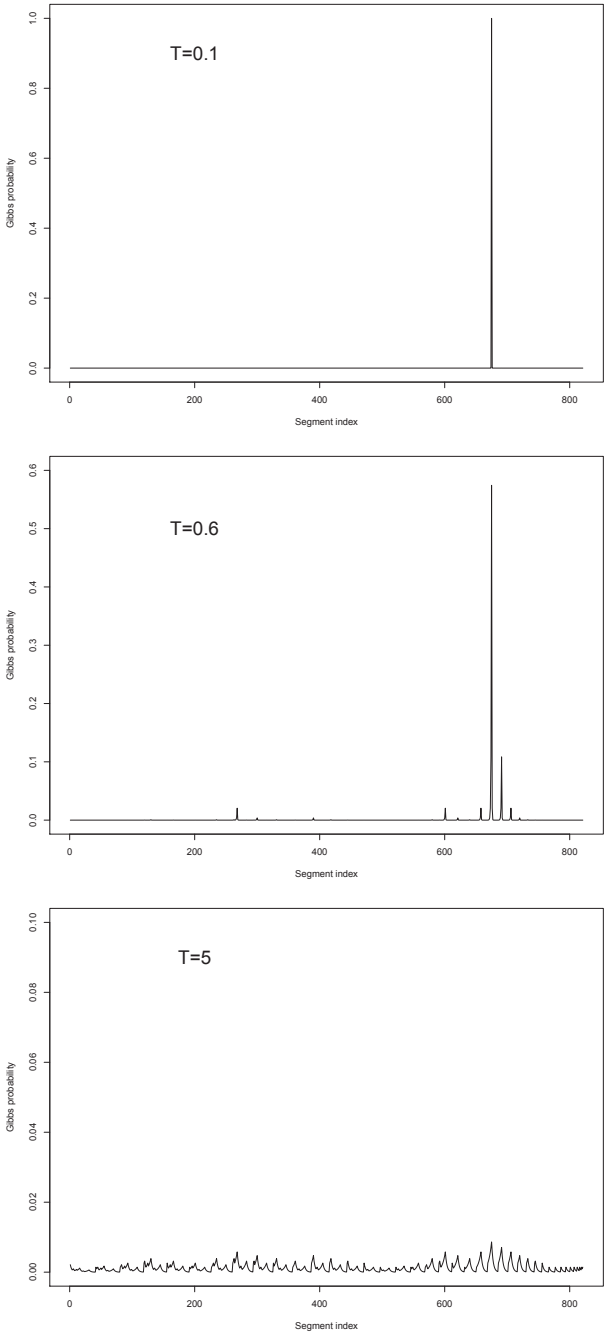


Figure 4.1: Distributions de Gibbs des 821 segments possibles pour différentes valeurs de T : panneau du haut pour $T = 0.1$; panneau du milieu pour $T = 0.6$; panneau du bas pour $T = 5$. Noter le changement d'échelle pour l'axe des y.

segment atypique, c'est-à-dire celle reposant sur la segmentation et le modèle génératif et celle reposant sur l'utilisation d'une fonction de scores. Nous abordons dans la Section 4.3 des applications que cette dualité des deux modèles impliquent. Une adaptation de l'algorithme Forward Backward est présentée en Sous-section 4.3.1, qui permet de calculer les quantités d'intérêt comme par exemple $\mathbb{P}(S_j = 2|\mathbb{A})$. Une application des résultats aux composants ambigus, incertains ou de profils est présentée en Sous-section 4.3.2 établissant ainsi des scores cohérents pour ce type de composants ou d'objets. Les résultats théoriques de la Section 4.2 permettent également d'apprendre de manière supervisée, ou non, une fonction de score à partir d'un ensemble de données présentant des segments atypiques. Une application au cas de segments transmembranaires pour chacun des aspects, calculs des probabilités conditionnelles, apprentissage supervisé ou non d'une fonction de scores, est présentée dans la Sous-section 4.3.3. Nous proposons une conclusion et des perspectives de travail dans la Section 4.4. La Section 4.5 reprend le type de tableaux récapitulatifs des chapitres précédants et synthétise les différentes possibilités de cette nouvelle approche suivant le contexte (longueur de séquence, type de scores).

4.2 Equivalence des deux espaces probabilisés

Résultat 14 (Du modèle génératif vers la fonction de scores)

$$\forall (q_0, q_1) \in \mathcal{D}_{\mathcal{A}}^2, \quad \exists!\ \sigma : \mathcal{A} \to \mathbb{R} \text{ tel que } \mathbb{P}(S|\mathbb{A}) \overset{\forall S \in \mathcal{S}}{\propto} \exp\left(\sum_{k=i, S_i=2} \sigma(A_i)\right) .$$

Preuve 1 *En supposant la distribution uniforme sur $\mathcal{S}$ tel que $\forall S \in \mathcal{S}$, $\mathbb{P}(S) = 1/|\mathcal{S}|$, alors nous avons*

$$\mathbb{P}(S|\mathbb{A}) = \frac{1}{Z}\mathbb{P}(\mathbb{A}|S) \propto \frac{\mathbb{P}(\mathbb{A}|S)}{\mathbb{P}(\mathbb{A}|S = 1 \dots 1)}$$

et

$$\begin{aligned} \frac{\mathbb{P}(\mathbb{A}|S)}{\mathbb{P}(\mathbb{A}|S = 1 \dots 1)} &= \frac{\prod_{i=1}^n q_0(A_i)^{\mathbb{1}_{S_i \neq 2}} q_1(A_i)^{\mathbb{1}_{S_i=2}}}{\prod_{i=1}^n q_0(A_i)} \\ &= \prod_{i, S_i=2} \frac{q_1(A_i)}{q_0(A_i)} = \exp\left(\sum_{i, S_i=2} \log \frac{q_1(A_i)}{q_0(A_i)}\right) . \end{aligned}$$

La fonction definie par $\sigma(a) \overset{\forall a \in \mathcal{A}}{=} \log(q_1(a)/q_0(a))$ verifie les propriétés. L'unicité se montre facilement en définissant des segmentations appropriées.

Résultat 15 (De la fonction de scores vers le modèle génératif)
$\forall q_0 \in \mathcal{D}_{\mathcal{A}}$ *et* $f : \mathcal{A} \to \mathbb{R}$ *verifiant* $(\sum_a q_0(a) f(a) < 0)$ *et* $(\exists a, f(a) > 0)$, *alors*

$$\exists!\ T > 0, \text{ tel que } q_1(a) \overset{\forall a \in \mathcal{A}}{=} q_0(a) \exp(f(a)/T) \in \mathcal{D}_{\mathcal{A}}$$

$$et\ \mathbb{P}(S|\mathbb{A}) \overset{\forall S\in\mathcal{S}}{\propto} \exp\left(\frac{1}{T}\sum_{k=i,S_i=2} f(A_i)\right) .$$

Preuve 2 *La démonstration repose sur le lemme suivant.*

Lemme 3 *$\forall f : \mathcal{A} \to \mathbb{R}$, et $\forall q_0 \in \mathcal{D}_{\mathcal{A}}$ les deux assertions suivantes sont équivalentes.*

i) $\exists! \rho > 0, \sum_a q_0(a)\exp(\rho f(a)) = 1$

ii) $\sum_a q_0(a) f(a) < 0$ *et* $\exists a, f(a) > 0$

Preuve 3 *La démonstration repose sur l'étude de la fonction*

$$g(r) = \sum_a q_0(a)\exp(rs(a)) = \mathbb{E}_{q_0}[e^{rf}]$$

et des éléments de convexité.

On a alors $T = 1/\rho$ et $q_1 = q_0\exp(f/T) \in \mathcal{D}_{\mathcal{A}}$ qui nous ammène au résultat.

4.3 Conséquences et applications

4.3.1 Calcul des quantités d'intérêt

Une adaptation de la procédure Forward-Backward [45] nous permet d'établir dans un temps linéaire les probabilités a posteriori pour chaque indice i de la séquence de se trouver dans chacun des états de la segmentation.

Posons

$$\mathrm{ev}_0(k) := \mathbb{1}_{\{k=1\}},\ \mathrm{ev}_{n+1}(k,\ell) := \mathbb{1}_{\{\ell-k\}} \in \{0,1\} \times \mathbb{1}_{\{\ell\neq 2\}} \quad \text{et}$$

$\mathrm{ev}_i(k,\ell) := \mathbb{1}_{\{\ell-k\in\{0,1\}\}} q_0(A_i)^{\mathbb{1}_{\{S_i\neq 2\}}} q_1(A_i)^{\mathbb{1}_{\{S_i=2\}}}$ pour $i = 1,\dots,n$.

De là,

$$\mathbb{P}(\mathbb{A}|S) := \mathrm{ev}_0(S_0)\cdot\prod_{i=1}^{n+1}\mathrm{ev}_i(S_{i-1},i)$$

$$\text{et } Z = \sum_{S\in\{1,2,3\}^{n+2}}\left(\mathrm{ev}_0(S_0)\prod_{i=1}^{n+1}\mathrm{ev}_i(S_{i-1},S_i)\right) .$$

$\forall j \in \{1,\dots,n\}$ notons $S_{<j}$ (*respectivement* $S_{>j}$) les sous-séquences $S_0\dots S_{j-1}$ (*resp.* $S_{j+1}\dots S_{n+1}$). Pour tout $k\in\{1,2,3\}$, introduisons les quantités Forward : $F_0(k) := \mathrm{ev}_0(S_0)$ et

$$F_j(k) := \sum_{S<j,S_j=k}\mathrm{ev}_0(S_0)\prod_{i=1}^{j}\mathrm{ev}_i(S_{i-1},S_i) \quad \text{for } j = 1,\dots,n+1$$

ainsi que les quantités backward: $B_{n+1}(k) := 1$ et

$$B_j(k) := \sum_{S>j, S_j=k} \prod_{i=j+1}^{n+1} \mathrm{ev}_i(S_{i-1}, S_i) \quad \text{for } j = 0, \ldots, n.$$

Résultat 16 *$\forall j \in \{0, \ldots, n+1\}$ et $\forall \ell \in \{1, 2, 3\}$ nous avons*

$$F_j(\ell) B_j(\ell) = \sum_{S<j, S_j=\ell, S>j} \mathrm{ev}_0(S_0) \prod_{i=1}^{n+1} \mathrm{ev}_i(S_{i-1}, S_i) = \sum_{S \in \mathcal{S}} \mathbb{1}_{\{S_j=\ell\}} \mathbb{P}(\mathbb{A}|S)$$

et $\forall j \in \{1, \ldots, n+1\}$ et $\forall k, \ell \in \{1, 2, 3\}$ nous avons

$$\begin{aligned} F_{j-1}(k)\mathrm{ev}_i(k, \ell) B_j(\ell) &= \sum_{S<j-1, S_{j-1}=k, S_j=\ell, S>j} \mathrm{ev}_0(S_0) \prod_{i=1}^{n+1} \mathrm{ev}_i(S_{i-1}, S_i) \\ &= \sum_{S \in \mathcal{S}} \mathbb{1}_{\{S_{j-1}=k, S_j=\ell\}} \mathbb{P}(\mathbb{A}|S). \end{aligned}$$

$\forall j \in \{1, \ldots, n+1\}$ et $\forall k \in \{1, 2, 3\}$:

$$F_j(\ell) = \sum_k F_{j-1}(k)\mathrm{ev}_j(k, \ell) \quad et \quad B_{j-1}(k) = \sum_\ell \mathrm{ev}_j(k, \ell) B_j(\ell).$$

De plus $\forall j \in \{1, \ldots, n+1\}$ et $\forall k, \ell \in \{1, 2, 3\}$ nous avons

$$Z = \sum_k F_j(k) B_j(k) = \sum_{k,\ell} F_{j-1}(k)\mathrm{ev}_j(k, \ell) B_j(\ell)$$

$$\mathbb{P}(S_j = k|\mathbb{A}) = \frac{F_j(k) B_j(k)}{Z}, \quad \mathbb{P}(S_{j-1} = k, S_j = \ell|\mathbb{A}) = \frac{F_{j-1}(k)\mathrm{ev}_j(k, \ell) B_j(\ell)}{Z}$$

$$\mathbb{P}(S_j = \ell|S_{j-1} = k, \mathbb{A}) = \frac{\mathrm{ev}_j(k, \ell) B_j(\ell)}{B_{j-1}(k)}$$

$$\mathbb{P}(S_{j-1} = k|S_j = \ell, \mathbb{A}) = \frac{F_{j-1}(k)\mathrm{ev}_j(k, \ell)}{F_j(\ell)}$$

Notons les éléments suivants :

- $\mathbb{P}(S_j = k|\mathbb{A})$ est particulièrement intéressante en $j = n+1$ correspondant à la probabilité a posteriori d'avoir ou non un segment généré sous q_1.

$$\mathbb{P}(\text{le segment est vide}|\mathbb{A}) = \mathbb{P}(S_{n+1} = 1|\mathbb{A}) = \frac{F_{n+1}(3)}{Z}$$

$$\mathbb{P}(\text{le segment n'est pas vide}|\mathbb{A}) = \mathbb{P}(S_{n+1} = 3|X) = \frac{F_{n+1}(1)}{Z}$$

- $\mathbb{P}(S_{j-1} = k, S_j = \ell|\mathbb{A})$ avec $k = 1,\ \ell = 2$ et $k = 2,\ \ell = 3$) correspond à la probabilité a posteriori que le segment débute (*respectivement* finit) en position j (*resp.* $j-1$) mais cela nécessite de normaliser par la probabilité a posteriori qu'il existe bien un segment atypique :

$$\begin{aligned}\mathbb{P}(\text{le segment démarre en } j|\mathbb{A}) &= \frac{\mathbb{P}(S_{j-1}=1, S_j=2|\mathbb{A})}{\mathbb{P}(S_{n+1}=3|\mathbb{A})} \\ &= \frac{F_{j-1}(1)\mathrm{ev}_j(1,2)B_j(2)}{F_{n+1}(3)}\ ;\end{aligned}$$

$$\begin{aligned}\mathbb{P}(\text{le segment finit en } j-1|\mathbb{A}) &= \frac{\mathbb{P}(S_{j-1}=2, S_j=3|\mathbb{A})}{\mathbb{P}(S_{n+1}=3|\mathbb{A})} \\ &= \frac{F_{j-1}(2)\mathrm{ev}_j(2,3)B_j(3)}{F_{n+1}(3)}\ .\end{aligned}$$

Cette méthode Forward-Backward reste performante en temps de calcul pour de très longues séquences de longueur supérieure à 10^5 voire 10^6. La précision des calculs de probabilités peut être assurée par des calculs en échelle logarithmique par exemple.

Exemple
Considérons une sequence simulée de longueur $n = 300$ à valeurs dans $\mathcal{A} = \{1, 2, 3, 4\}$ avec un segment atypique inséré en position $I = [100, 150]$. La Figure 4.2 représente les probabilités que les composants de la séquence soient dans l'état 2 (segment inséré), $\mathbb{P}(S_j = 2|X) = F_j(2)B_j(2)/Z$: Le segment inséré est parfaitement retrouvé ; la position (Figure 4.3, panneau gauche) ainsi que la longueur (Figure 4.3, panneau droit) sont correctes.

4.3.2 Lettres indécises, incertaines, scores de profils

Le résultat 14 implique que les scores de composants ambigus sont précifiés.

Corollaire 1 *La dualité du modèle génératif et de la fonction de scores implique que le score de "a ou b" ($a \neq b$) doit être*

$$\log \frac{q_1(a) + q_1(b)}{q_0(a) + q_0(b)}\ .$$

Remarquons que le score déduit du modèle génératif et qui permet une interprétation mathématique correcte n'est ni $\sigma(a)+\sigma(b)$ ni $q_0(a)\sigma(a)+q_0(b)\sigma(b)$ que l'on peut parfois voir utilisé.

Illustrons cela sur l'IUPAC (ambiguous DNA code). Considérons les quatre nucléotides $\{A, C, G, T\}$ et les scores correspondant $\{-2, -1, 0, +1\}$ ainsi que deux distributions de fond $q_0^U = (0.25, 0.25, 0.25, 0.25)$, distribution équiprobable, et $q_0^{GC} = (0.1, 0.4, 0.4, 0.1)$, distribution riche en G et C. La table 4.1 nous donne les scores correspondant pour chacune des distributions q_0.

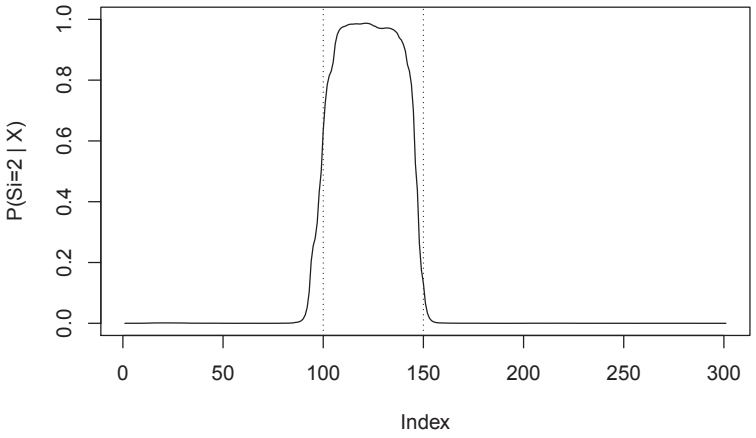


Figure 4.2: Exemple d'une séquence simulée avec $n = 300$ et $I = [150, 200]$: Distribution marginale a posteriori (conditionnellement à l'observation de la séquence X) de la segmentation pour l'état 2, $\mathbb{P}(S_j = 2|X)$.

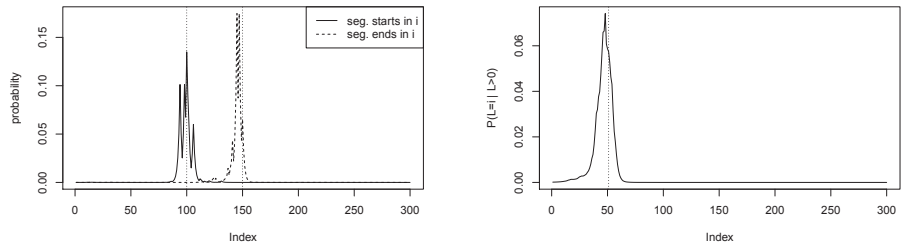


Figure 4.3: Exemple d'une séquence simulée avec $n = 300$ et $I = [150, 200]$: Distribution a posteriori du début et fin de segment (panneau gauche) et longueur (panneau droit).

Table 4.1:

IUPAC Code	A	C	G	T/U		
Meaning	A	C	G	T		
f	-2	-1	0	+1		
σ for q_0^U	-1.76	-0.88	0.00	0.88		
σ for q_0^{GC}	-3.29	-1.65	0.00	1.65		
IUPAC Code	M	R	W	S	Y	K
Meaning	A or C	A or G	A or T	C or G	C or T	G or T
σ for q_0^{GC}	-1.82	-0.21	0.96	-0.52	0.18	0.61
σ for q_0^U	-1.23	-0.54	0.26	-0.35	0.34	0.54
IUPAC Code	V	H	D	B	N	
Meaning	no T	no G	no C	no A	anyone	
σ for q_0^{GC}	-0.63	0.00	0.43	0.1	0.00	
σ for q_0^U	-0.64	0.00	0.18	0.24	0.00	

Le caractère incertain, *e.g.* en crystallisation de protéines, peut aussi être prise en compte. Pour une incertitude exprimée par des poids $w_a, w_b > 0$ des observations a et b, les scores suivants doivent être utilisés.

$$\log \frac{w_a q_1(a) + w_b q_1(b)}{w_a q_0(a) + w_b q_0(b)}$$

ce qui est à nouveau clairement différent de $w_a\sigma(a) + w_b\sigma(b)$.

De même, on peut établir les scores de profil (d'alignement multiples par exemple). Soit $w = (w_a)_{a\in\mathcal{A}}$ avec $w_a > 0$. Alors

$$\sigma(w) = \log\left(\frac{\sum_a w_a q_1(a)}{\sum_a w_a q_0(a)}\right) \neq \sum_a w_a\sigma(a) \ .$$

4.3.3 Apprendre une fonction de score

Dans la sous-section précédente nous voyons comment à partir d'une fonction de scores initiale, effectuer un changement d'échelle adapté permet la recherche du segment atypique de manière équivalente à l'aide de la segmentation et des algorythmes usuels des HMM.

Dans cette sous-section nous proposons de construire une échelle de scores σ à partir d'un ensemble de séquences possédant des segments reconnus atypiques dont on connaît la position (apprentissage supervisé) ou non (apprentissage non-supervisé). Cela s'effectue sur l'apprentissage de q_0 et q_1, avec $\sigma = \log(q_1/q_0)$.

Table 4.2: Echelle de scores σ calculée à partir de 56 TM protéines.

A	C	D	E	F	G	H	I	K	L
0.751	0.480	-3.834	-2.123	0.877	0.226	-0.306	0.625	-1.736	0.772
M	N	P	Q	R	S	T	V	W	Y
0.600	-1.694	-1.592	-1.086	-1.885	-0.583	-0.545	0.653	0.645	-0.107

Apprentissage supervisé

Ici, les deux distributions sont déduites d'un ensemble de 56 séquences transmembranaires (TM) (request "name:"transmembrane protein" soluble helical AND reviewed:yes" dans la base de données UniProtKB/Swiss-Prot (http://www.uniprot.org/uniprot) pour lesquelles les régions transmembranaires (TM) sont précisées. L'ensemble de ses séquences totalise 29,494 acides aminés en dehors des regions TM et 2,535 dans des régions TM. L'échelle de scores déduites est donnée en Table 4.2. Etudions maintenant une protéine TM supplémentaire sp|O15393|TMPS2_HUMAN avec une région TM en position 85-105 à l'aide de cette fonction de scores apprise. La Figure 4.4 permet de mettre en valeur la qualité de détection. La Figure 4.5 correspond à la même recherche de segment en utilisant la fonction de scores d'hydrophobicité usuelle de Kyte et Doolitle après le changement d'échelle approprié vu en section précédente. La comparaison de ces deux figures met clairement en évidence la pertinence de la fonction de scores apprises. D'autres figures notamment sur les longueurs et positions sont proposées dans l'article [42]. Une comparaison des deux échelles est également proposée.

Apprentissage non-supervisé

Pour un apprentissage non-supervisé l'utilisation de l'algorithme classique EM (Estimation Maximisation) permet de retrouver la segmentation. Pour l'étape Estimation nous calculons $\eta_i := \mathbb{P}(S_i = 2|X, \theta)$ avec $\theta = (q_0, q_1)$ et pour l'étape de Maximization, θ est mise à jour pour les acides aminés a par $q_1(a) = \sum_i \eta_i 1_{X_i=a} / \sum_i \eta_i$ et $q_0 = \sum_i (1-\eta_i) 1_{X_i=a} / \sum_i (1-\eta_i)$. Le modèle de cette section repose sur le fait qu'il n'y a au plus qu'un segment atypique. Nous n'avons pas pu extraire suffisamment de séquences correspondant à l'analyse et possédant un unique segment pour effectuer un apprentissage non supervisé qui nécessite un ensemble d'apprentissage plus grand que dans le cas supervisé. Nous avons donc effectué l'apprentissage sur des séquences simulées en utilisant la distribution q_0 apprise sur le jeu de données des 56 TM pour simuler les parties non TM et la distribution q_1 pour les segments TM. Différentes longueurs de segment ont été envisagées. Un ensemble de 300 séquences de longueur allant de 100 à 300 suffisent à retrouver la distribution initiale q_0. Pour apprendre correctement la distribution q_1 il est nécessaire que les segments atypiques insérés ne soit pas trop petits. Un ensemble de séquences d'apprentissage plus

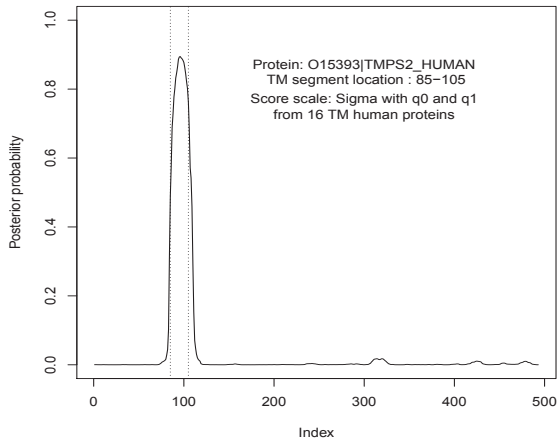


Figure 4.4: Probabilité a posteriori, $\mathbb{P}(S_j = 2|\mathbb{A})$, que l'indice S_i soit dans l'état 2 (dans le segment atypique) à partir de l'échelle σ calculée via q_0 and q_1 apprises des 56 protéines TM.

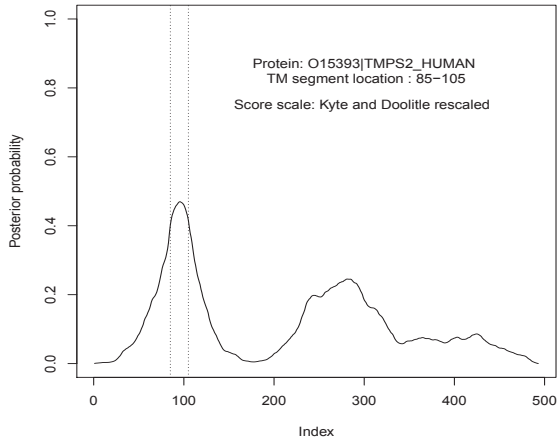


Figure 4.5: Probabilité a posteriori, $\mathbb{P}(S_j = 2|\mathbb{A})$, que l'indice S_i soit dans l'état 2 (dans le segment atypique) à partir de l'échelle de Kyte and Doolittle après un changement d'échelle adapté.

grand ou un nombre d'itérations de l'algorithme EM plus important ne permettent pas un apprentissage correcte pour des longueurs de segments trop petits (longueur $\ell = 25$ pour des séquences de longueur $n = 300$ par exemple). Considérons deux jeux de données différents tels que : Jeu 1, nombre de séquences $= NB$, longueur des séquences n, longueur des segments ℓ ; Jeu 2 : nombre de séquences $NB' = NB/10$, longueur des séquences $n' = 10 \cdot n$, longueur des segments $\ell' = 10 \cdot \ell$. Ces deux jeux de données totalisant chacun le même nombre d'acides aminées simulés sous q_0 et sous q_1, mais nous observons que la distribution q_1 est moins bien apprise par le jeu de données ayant les segments les plus courts (jeu 1).

4.4 Conclusion et perpectives de travail

Nous avons établi une dualité entre l'approche par HMM et celle par fonction de scores pour mettre en évidence des segments atypiques. Pour une HMM constrainte donnée, une fonction de scores peut être déduite rendant les deux espaces probabilisés équivalents. Nous avons montré que sous certaines hypothèses très classiques sur la fonction de scores, cette dernière peut être ajustée par homothétie et une distribution des segments atypiques établie, permettant ainsi de créer un modèle génératif équivalent. Cette dualité et ces conditions sur la fonction de scores permettent une interprétabilité des résultats de recherche de segment atypique par fonction de scores. En effet, notons que le test statistique sous-jacent et largement non présenté dans la littérature, reste très vague quant à l'hypothèse alternative.

H_0 : la séquence $\mathbb{A}$ est une observation des A_i i.i.d. de distribution q_0.

H_1 : non H_0.

Intuitivement, il s'agit d'une alternative exprimant le fait qu'il existe un segment I et une distribution $q_1 \neq q_0$ pour lesquels les $(A_i)_{i \in I}$ sont distribuées sous q_1. Nous montrons ici que cette distribution existe pour une fonction de scores vérifiant les propriétés d'une score moyen négatif et de scores positifs possibles.

La dualité possède de nombreuses conséquences.

L'approche par HMM permet une implémentation linéaire, avec des scores qui peuvent être réels, par le biais d'une adaptation de l'algorithme Forward and Backward. Cette implémentation permet le calcul de probabilités a posteriori très intéressantes pouvant prendre en compte la position ou la longueur du segment atypique.

Les scores de composants ambigus, incertains et ceux de profils peuvent être établis de manière cohérente.

Des fonctions de scores peuvent être apprises à partir de données réelles.

Un autre avantage de l'approche par segmentation est de rendre possible la recherche simultanée de plusieurs segments atypiques au sein de la séquence. Ce travail, en collaboration avec Hélène Chiapello (INRA de Jouy-en-Josas), est développé dans le Chapitre 5.

Il est également possible d'établir des intervalles de confiance sur le score de segments en s'appuyant sur le travail d'Alexandra Lefèvre en thèse au LPMA dirigée par Grégory Nuel.

4.5 Synthèse

Cette nouvelle approche apporte de nombreuses possibilités et est très performante. Notons que cette méthode peut-être utilisée dans l'objectif de détecter des régions atypiques indépendamment d'une fonction de scores. Dans l'objectif d'être dans une recherche de régions atypiques équivalente à celle du score local, alors il est nécessaire d'utiliser une fonction de scores de moyenne négative. Cette méthode peut par ailleurs être utilisée pour de petites à de très grandes séquences. Dans le tableau suivant, HMM correspond à un résultat de type probabilité a postériori d'une région atypique, $\mathcal{L}$ (*respectivement* $\mathcal{P}$) un résultat sur la longueur (*resp.* position) du segment atypique.

$\mathbb{E}[X]$	$n \leq 10$	$n \leq 10^2$	$n \simeq 10^3$	$n > 10^4$
< 0	$HMM + \mathcal{L} + \mathcal{P}$	$HMM + \mathcal{L} + \mathcal{P}$	$HMM + \mathcal{L} + \mathcal{P}$	$HMM + \mathcal{L} + \mathcal{P}$
$= 0$				
> 0				

Part II

Applications du score local

Dans de nombreuses applications, il est crucial de détecter une rupture dans la dynamique d'un phénomène le plus tôt possible. C'est le cas dans la recherche de régions atypiques dans les séquences génomiques (voir Chapitre 5.1 et 5.2). C'est aussi un des enjeux importants de la gestion de production et la maîtrise statistique des procédés (voir Chapitre 6) avec des applications dans des domaines très variés comme celui de la santé publique, le suivi médical, la géostatistique, etc...

Nous présentons dans cette partie différentes utilisations du score local afin de détecter des régions atypiques et/ou des points de rupture.

Chapter 5

Application aux séquences biologiques

5.1 Régions cibles pour la sélection

Collaborateurs principaux : Magali San Cristobal (INRA de Castanet Tolozan), Marìa Inès Fariello, Simon Boitard (INRA de Castanet Tolozan), David Robelin (INRA de Castanet Tolozan)
Publication : [38]

5.1.1 Réflexion sur la distribution du score local

Lors de la thèse de Marìa Inès Fariello, encadrée par Magali San Cristobal, Simon Boitard et Naya Hugo (Institut Pasteur de Montevideo), il a été question de mettre en évidence des régions de gènes dans des génomes de cailles ayant été la cible de sélection au cours de l'évolution. Les séquences étudiées sont alors une suite de locii (positions de gènes). Les scores choisis et utilisés correspondent à la p-valeur d'un test usuel de sélection appelé $\mathcal{F}_{ST}$. Les scores initiaux étaient donc réels à valeurs dans $]0,1[$. Les séquences sont de l'ordre de plusieurs milliers à plusieurs dizaines de milliers (10^3–10^4). La dépendance des sites était également un point important à prendre en compte.

La longueur des séquences ne permettait pas l'utilisation des méthodes exactes. Les approximations de type loi de Gumbel nécessitaient quant à elles un score moyen négatif ce qui était en contradiction avec des scores à valeurs dans $]0,1[$. De plus, à cette époque l'approximation par une loi de Gumbel dans le cas Markovien ne semblait pas accessible. Pour cela, il s'est avéré nécessaire à cette époque d'établir la p-valeur des scores locaux calculés par une méthode de Monte Carlo. Une transformation de l'échelle des scores a également été effectuée, $s = -\log\left(p_{valeur}(\mathcal{F}_{ST})\right) + \xi$. De cette manière, un locus ayant une faible p_{valeur} dans le test $\mathcal{F}_{ST}$ avait un score important et inversement. La

translation de ξ c'est avérée nécessaire afin de rendre la recherche plus local et d'autre part de distinguer des régions plus marquées, dénuées d'un bruit de fond. De plus, il s'est avéré par la suite que la valeur de $\xi = -1$ ou -2 étaient en cohérence avec un ensemble d'observations/connaissances prélables sur des séquences déjà bien étudiées. Ce travail a soulevé notamment les différentes questions théoriques suivantes :

- Dans ce travail, nous avons été amené à démontrer que le score local d'un séquence "lue" dans un sens (de droite à gauche) était le même que celui de la séquence lue dans l'autre sens (de gauche à droite), et que le segment qui réalisait ces deux scores locaux correspondait à la même région de gènes. Cela a permis une détection visuelle des régions par la représentation graphiques des processus de Lindley dans les deux sens de lecture sur le même graphique.
- Le point le plus important a été de répondre à la question, quelle est la p-valeur dont nous avons besoin : celle correspondant à la hauteur observée de n'importe quelle région ou bien la valeur du score local, associée à la région de plus haut score. Ce dernier point fait l'objet de la sous-section suivante.

La Figure 5.1 met en évidence sur les données de génome de caille la qualité de détection de la région cible en comparaison avec différentes méthodes usuelles.

5.1.2 Réflexion autour d'une correction due aux tests multiples

Il est apparu clairement dans ce travail que la question de la significativité d'une région ne se posait pas uniquement sur la région réalisant le score local mais sur toutes les régions réalisant une excusion de score positif. En effet, la problématique consistait à mettre en évidence les régions de gènes ayant été significativement impliquées dans la sélection. Il s'agissait donc de mettre en place une méthode pour choisir un score seuil, noté s_{seuil} tel que toute région ayant un score dépassant cette valeur à déterminer était intéressante. Nous sommes donc ici dans un cas particulier de tests multiples ou le nombre de tests effectués est aléatoire puisque correspondant au nombre d'excursions dans la séquence et a priori inconnu. Nous avons pu montrer que dans le modèle i.i.d. choisir un score seuil s_{seuil} permettant de corriger le test multiple de manière similaire à Bonferroni [12] correspondait à prendre le seuil correspondant à la loi du score local et de s'affranchir du nombre d'excursions a priori inconnu. Notons $m(n)$ la variable aléatoire du nombre d'excursions dans la séquence de longueur n, Q_i le score de la i-ème excursion et M_n le score local de la séquence. Nous avons

$$\forall i = 1, \dots, m(n), \qquad \mathbb{P}(Q_i > s_{seuil}) < \alpha/m(n) \Leftrightarrow \mathbb{P}(M_n > s_{seuil}) < \alpha \ .$$

A priori, la démonstration dans le cas Markovien ne nous est pas apparue aussi évidente. Cependant les travaux plus récents avec Simona Grusea dans le cas

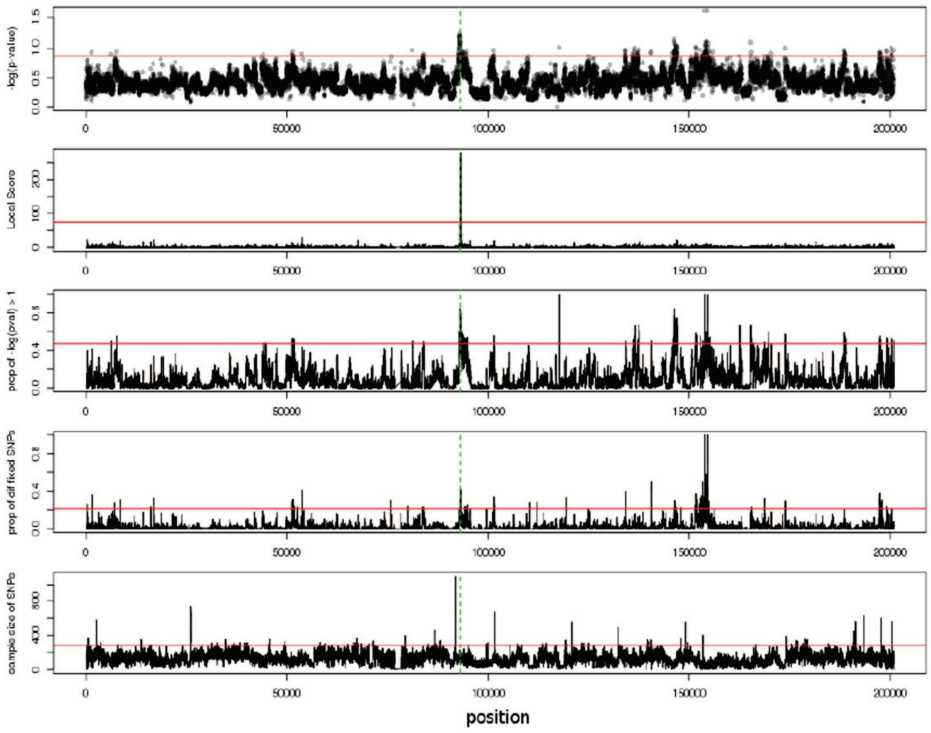


Figure 5.1: Empreinte de sélection pour des données de génome de cailles (plus précisément GGA1). Le processus de Lindley basé sur la fonction de scores $-\log\left(p_{valeur}(\mathcal{F}_{ST})\right) - 1$ correspond à la deuxième fenêtre en partant du haut. Les autres fenêtres correspondent à différentes méthodes usuelles. Les positions sont indiquées en Mb. La ligne rouge indique le seuil correspondant aux 99ième centile. La ligne verte indique le centre de la région détectée par l'approche reposant sur le score local.

Markovien (voir Section 3.2.3) ouvrent de nouvelles possibilités qu'il serait judicieux d'exploiter. Voir également Chapitre 9 pour le développement de cette ouverture.

5.1.3 Conclusion

La méthode proposée s'est avérée performante par rapport aux méthodes plus classiques utilisées jusque là pour mettre évidence des régions ayant été la cible de sélection.

La question de correction de la multiplicité des tests et mais aussi du besoin de résultats sur la loi du score des excursions et du score local dans des modèles de dépendance a clairement été mise en évidence dans ce travail. Cela a constitué une des motivations de travail avec Simona Grusea ainsi que le projet proposé au Chapitre 9.

5.2 Recherche simultanée de plusieurs segments

Collaborateurs : Grégory Nuel et Hélène Chiapello (INRA de Jouy-en-Josas)
Travail en cours [39].

Un des nombreux avantages de l'approche par segmentation abordée au Chapitre 4 est de rendre possible la recherche simultanée de plusieurs segments atypiques au sein de la séquence. Les états de la segmentation possibles sont alors à valeurs dans $\{1, \dots, 2K+1\}$ avec K le nombre de segments dans la séquence. Les états impaires correspondant aux parties hors segments atypiques et générées avec la distribution de référence q_0 et les états paires aux segments atypiques générés sous q_1. L'adaptation des programmes de l'algorithme Forward - Backward explicité au Chapitre 4 est quasi immédiate.

Pour $K \in \{1, \dots, K_{max}\}$, le nombre inconnu de segments atypiques dans la séquence, l'objectif est de déterminer $\mathbb{P}(K = k|\mathbb{A})$ avec $\mathbb{A}$ la séquence biologique observée. Nous avons alors la quantité Forward qui nous donne

$$\mathbb{P}(\mathbb{A}|K = k) = F_{n+1}(2k+1) \ .$$

Pour une séquence de longueur n, nous avons C_{n+1}^{2K} segmentations possibles possédant K segments. Suivant l'idée sous-jacente au critère EBIC (Enhance BIC), nous souhaitons que la probabilité a priori $\mathbb{P}(K = k)$ tienne compte de ce nombre différent de segmentations possibles suivant la valeur k

$$\mathbb{P}(K = k) \propto \frac{p_k}{C_{n+1}^{2k}}$$

avec $p_k = 1$ s'il n'y a pas d'information a priori donnée. D'autres distributions (Poisson, une distribution déduites d'un apprentissage supervisé, ...) peuvent être envisagées pour $(p_k)_k$.

Prenons par exemple $K = 1$ et 2, et $p_1 = p_2 = 1$. Cela donne 55 segmentations possédant 1 segment atypique et 330 segmentations possédant 2

segments atypiques. Si toutes ses segmentations ont le même poids, alors $\mathbb{P}(K = 1) = 55/385 = 0.14$ et $\mathbb{P}(K = 2) = 330/385 = 0.86$. Afin de donner autant de poids à $K = 1$ que $K = 2$ nous allons corriger la probabilité des segmentations a priori de manière proportionnelle à l'inverse du nombre de segmentations correspondantes, soit donc $1/C_{n+1}^{2k}$. Nous avons finalement

$$\log(\mathbb{P}(K = k|\mathbb{A})) = \text{Cste} + \log(F_{n+1}(2k + 1)) + \log(p_k) - \log(C_{n+1}^{2k}) \ .$$

Nous montrons que

- Plus la longueur n de la séquence est grande et plus le nombre réel de segments est retrouvé de manière correcte. La Figure 5.2 représente la probabilité a posteriori du nombre de segments atypiques $\mathbb{P}(K = k|\mathbb{A})$ pour différentes longueurs n de séquences simulées ayant 4 composants différents possibles, avec pour distributions $q_0 = (0.25, 0.25, 0.25, 0.25)$ et $q_1 = (0.05, 0.45, 0.45, 0.05)$, et un nombre réel de segments égale à 5. Pour une longueur $n \geq 5 \cdot 10^3$ la plus grande valeur de $\mathbb{P}(K = k|\mathbb{A})$ correspond bien à $k = 5$. Pour $n \leq 2 \cdot 10^3$ le nombre de segments proposé par la méthode n'est pas correct.

- Plus les segments sont petits et plus il est difficile de les retrouver tous. Cela est illustré dans la Figure 5.3 pour laquelle $n = 5 \cdot 10^5$, $q_0 = (0.25, 0.25, 0.25, 0.25)$, $q_1 = (0.2, 0.3, 0.3, 0.2)$ et le nombre réel de segments est égal à 5. Les courbes correspondant à $\ell = 5$ et 10 se superposent et propose à tord 1 pour le nombre de segments. Pour $\ell = 500$, le maximum correspond à 3 segments. Les trois courbes relatives à $\ell = 1000,\ 1500,\ 2000$ atteignent correctement leur maximum en $k = 5$.

- Plus les segments sont atypiques et plus il est facile de retrouver le nombre de segments atypiques. Dans la Figure 5.4, $q_0 = (0.24, 0.26, 0.26, 0.24)$ et $q_1 = (0.28, 0.22, 0.22, 0.28)$. Ces distributions correspondent à celles apprises sur un ensemble de séquences réelles SentTyphi possédant des régions Backbone (sous la distribution de background q_0) et des régions dites variables se distinguant des régions Backbone (régions atypiques sous la distribution q_1). Chaque séquence contient 5 segments dont les longueurs ℓ varient de $n/200$ à $n/40$. Dans le panneau gauche, les longueurs n de séquences considérées vont de 10^4 à 10^5 et l'estimation du nombre de segments la plus proche de la réalité est pour $n = 10^5$ avec $\mathbb{P}(K = k|\mathbb{A}) \simeq 0.4$. Dans le panneau droit, les longueurs de séquence vont de $n = 5 \cdot 10^4$ à $3.5 \cdot 10^5$. Pour $n \geq 2 \cdot 10^5$ l'estimation est correcte avec une probabilité $\mathbb{P}(K = k|\mathbb{A})$ maximale de 0.8 pour $n = 2 \cdot 10^5,\ 3 \cdot 10^5$ et $3 \cdot 10^5$.

Nous souhaitons maintenant tester la méthode sur des données réelles de type SentTyphi et étudier la qualité de détection du nombre de segments ainsi que leur position.

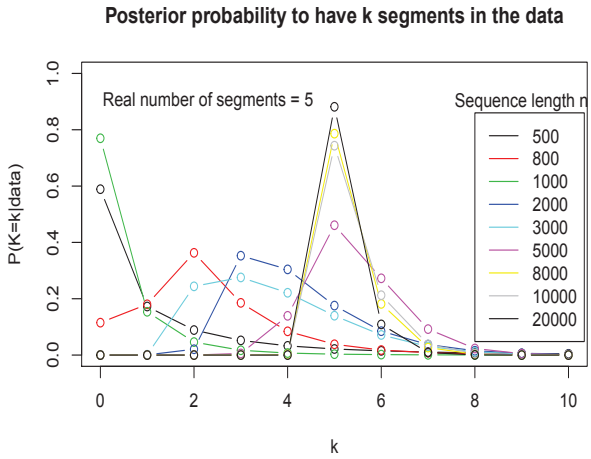


Figure 5.2: Probabilités a posteriori pour le nombre de segments atypiques pour différentes longueurs n de séquence d'ADN $\mathbb{P}(K = k|\mathbb{A})$. Distribution $q_0 = (0.25, 0.25, 0.25, 0.25)$, distribution $q_1 = (0.05, 0.45, 0.45, 0.05)$ et un nombre réel de segments égal à 5.

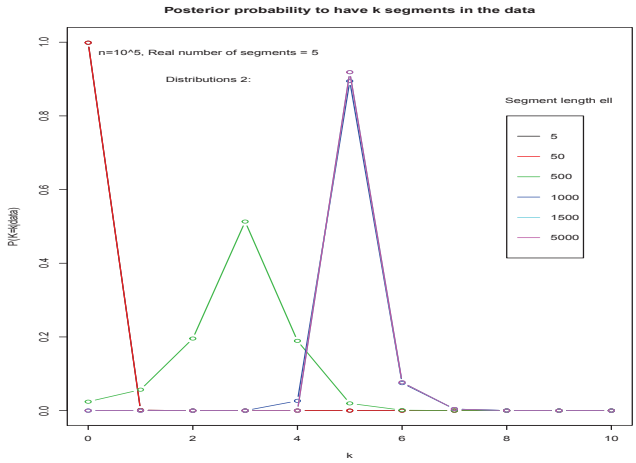


Figure 5.3: Probabilités a posteriori pour le nombre de segments atypiques $\mathbb{P}(K = k|\mathbb{A})$ pour différentes longueurs ℓ de segments. Longueur de la séquence $n = 5 \cdot 10^5$, distribution $q_0 = (0.25, 0.25, 0.25, 0.25)$, distribution $q_1 = (0.2, 0.3, 0.3, 0.2)$ et un nombre réel de segments égal à 5.

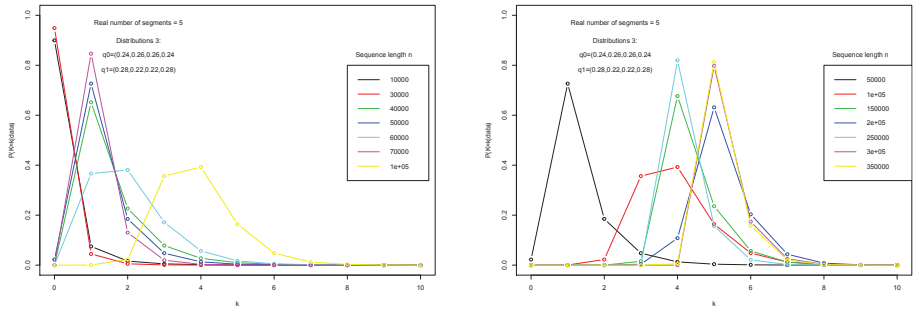


Figure 5.4: Probabilités a posteriori pour le nombre de segments atypiques $\mathbb{P}(K = k|\mathbb{A})$ pour différentes longueurs n de segments, les distributions $q_0 = (0.24, 0.26, 0.26, 0.24)$ et distribution $q_1 = (0.28, 0.22, 0.22, 0.28)$, un nombre réel de segments égal à 5 dont les longueurs ℓ varient de $n/200$ à $n/40$. Panneau gauche : n, la longueur des séquences varie de 10^4 à 10^5. Panneau droit : n varie de $n = 5 \cdot 10^4$ à $3.5 \cdot 10^5$.

Chapter 6

Maîtrise statistique des procédés (MSP)

Mes collaborateurs : François Bergeret (Entreprise Ippon Innovation).
Publications : [8], nouvelle édition en cours avec de nouvelles études de cas et un chapitre supplémentaire sur les cartes de contrôle spécifiques aux événements rares.
Obtention en février 2018 d'un PEPS1 d'un montant de $6.5K$ euros pour 2 ans.

6.1 Introduction et motivation

L'utilisation des fenêtres glissantes ou moyennes mobiles de taille donnée se retrouve dans de très nombreux domaines. C'est le cas de la gestion de production, mais aussi la pharmaco vigilance, le domaine médical ou encore la santé publique...

Exemple 1 (Fenêtres glissantes et suivi médical) *Dans les travaux d'Alice Cleynen par exemple (cf. exposé au séminaire de l'Institut de Mathématiques de Toulouse le 13 juin 2017 sur la détection de rupture dans la dynamique d'un processus de Markov déterministe par morceaux, article en cours de préparation) le domaine d'application concerne des patients ayant été atteints d'un cancer, en rémission mais susceptibles de rechuter. Des mesures de la quantité de cellules tumorales sont effectuées à intervalles réguliers. Le suivi présenté est effectué à partir de fenêtres glissantes de taille fixée. Un enjeu naturel consiste à détecter le plus tôt possible, à partir du suivi de ces mesures, une rechute.*

Exemple 2 (Surveillance dans le domaine de la santé publique) *Les fenêtres glissantes de taille fixe sont également utilisées dans le domaine de la santé publique. Voir les nombreuses références proposées dans [13] avec par exemple la surveillance effectuée par le Swedish Radiation Protection Institute*

pour la détection d'une augmentation du niveau de radiation γ en Suède [32], ou les documents relatifs à "Food and Drug Administration" [20]. Les fenêtres glissantes de taille fixe ont ensuite été délaissées car reconnues comme peu performantes pour détecter à temps ou correctement les ruptures [21, 49].

Il nous a paru naturel de proposer et tester une nouvelle statistique de test dans le suivi de ces mesures dans le but de s'affranchir de la taille de la fenêtre à considérer. Les travaux autour du score local nous semblent tout à fait indiqués pour cela. Nous nous intéressons plus particulièrement aux cas des événements rares, comme l'apparition de produits non conformes (présence de bactéries dangereuses dans des produits laitiers, ou encore du nombre d'infections post opératoires).

6.2 Quelques cartes de contrôle actuelles

De manière générale une carte de contrôle est une visualisation de tests successifs avec H_0 ="Le procédé est sous contrôle" et H_1 ="Le procédé n'est pas sous contrôle". Elle représente pour chaque échantillon l'intervalle de confiance dont les bornes sont les limites de contrôle (LC) de la carte. Un échantillon hors des LC produit une alarme, rejet de H_0.

Plusieurs approches sont possibles pour effectuer le contrôle du processus. Le suivi par cartes "p" ou "np" ou encore cartes "c" ou "u", est celui le plus répandu dans les logiciels pour le suivi dans le cas des attributs (nombre de défectueux ou nombre de défauts) comme pour le nombre d'infections post opératoires ou le nombre de fois où la présence de bactéries a été décelée. On considère l'individu statistique i, par exemple une opération chirurgicale, et soit $X_i \hookrightarrow \mathcal{B}(p)$ avec $p = \mathbb{P}(X = 1) = \mathbb{P}$(on observe un cas d'infection). On comptabilise alors le nombre de cas observés sur une période déterminée, constante ou non, pour toutes les opérations ayant eu lieu au cours du mois m. La carte np représente alors le nombre de cas par mois, $Y_m \hookrightarrow \mathcal{B}(n, p)$ avec n le nombre d'opérations par mois (ou bien la proportion de cas pour la carte "p" si le nombre d'opérations n'est pas constant). Les limites de contrôle classiquement posées sont les suivantes

$$LC_Y = \text{cible} \pm 3\sigma_Y \text{ ou bien } LC_Y = \mathbb{E}[Y] \pm 3\sigma_Y \tag{6.1}$$

ce qui dans le cas de la carte np correspond à

$$LC_{np} = np \pm 3\sqrt{np(1-p)} \ .$$

Le $3\sigma_Y$ de la formule (6.1) provient de l'analogie à la carte de Shewhart au cas des mesures s'appuyant sur l'hypothèse de normalité et dans un objectif d'assurer un risque de fausse alarme (risque de type I) de 0.27%.

Dans le cas des événements rares, $p \simeq 10^{-3}$ ou 10^{-5} pour un processus sous contrôle. Le type de suivi et les cartes de contrôle précédants et ne sont pas adaptés. Il est facile de montrer que pour ces cartes de contrôle la probabilité de recouvrement des intervalles de confiance relatives est différente de la valeur

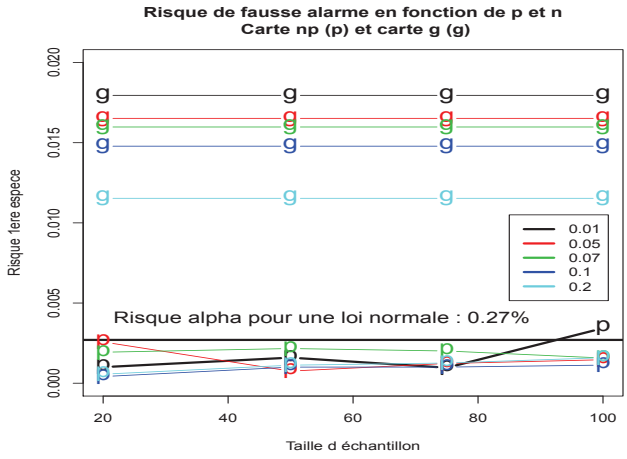


Figure 6.1: Risque de fausse alarme (risque de type I) pour la carte "np" et la carte "g"

nominale, i.e. le risque de fausse alarme (risque de type I) effectif est différent de la valeur nomimale α usuelle et attendue de 0.27% (voir Figure 6.1). On constate également que le risque client (risque de seconde espèce) peut-être également élevé (voir Figure 6.2). Cela semble logique, puisque l'approximation de la loi Binomiale par une loi normale est adaptée à condition que n soit assez grand et pour p tel que $np \geq 10$ et $n(1-p) \geq 10$. Les deux dernières conditions correspondant à dire que l'apparition d'un cas n'est ni trop rare ni trop fréquente.

Désormais, des cartes un peu plus adaptées sont proposées, appelées cartes "g" (logiciels Excel, Jmp, Minitab par exemple) développées par Beneyan [27, 29]. Pour ce type de suivi, la variable usuellement associée X_i est le nombre d'observations sans réalisation d'un cas (nombre d'opérations sans infection) avant d'observer un cas. Cette variable suit une loi $\mathcal{G}(p)$ (d'où le nom de carte "g") avec comme dans le cas des cartes "np" $p = \mathbb{P}(\text{on observe un cas})$. Les limites de contrôle reposent à nouveau sur l'analogie aux cartes de Shewhart, voir formule (6.1). Ces cartes de contrôle permettent de détecter plus rapidement une dérive ou une modification des chances d'oberver un "défecteux" p. Les limites de contôle proposées par Beneyan [27] sont

$$LC_g = \frac{1}{p} \pm 3 \cdot \sqrt{(\frac{1-p}{p^2})} \ .$$

Dans le cas des cartes "g", la limite de contrôle inférieure tient un rôle plus important, une alarme correspondant à un temps entre deux cas plus petit et donc des cas plus fréquents, ce qui correspond à ce que l'on souhaite détecter. Un dépassement de la limite supérieure correspondant elle à des événements que l'on souhaite reproduire. Un inconvénient majeur de ces cartes est de procurer

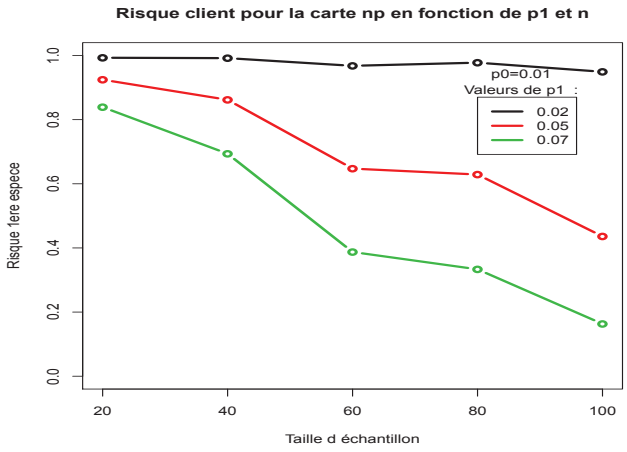


Figure 6.2: Risque client (de type II) pour la carte "np" pour différentes tailles d'échantillon n pour p_0 la probabilité d'apparition d'un cas pour un processus sous contrôle, utilisé pour le calcul des limites de contrôle, et différentes valeurs de p_1, probabilité d'apparition d'un cas pour un processus hors contrôle.

très souvent des limites de contrôle inférieures classiques négatives et donc peu exploitables.

Beaucoup d'autres cartes sont proposées dans des articles pour essayer de palier aux problèmes de performance ou bien pour prendre en compte le fait que la probabilité d'apparition d'un cas p n'est pas connu mais estimé ; que l'on ne souhaite pas attendre, étant donné les contextes d'études, d'avoir observé suffisamment de cas pour estimer correctement p. On peut noter par exemple, la carte "CCC" Cumulative Count of Conforming Chart [36], il s'agit d'une carte "g" mais dont les valeurs observées sont cumulées, les c-charts avec approche bayésienne [35] qui ont l'avantage de ne pas avoir besoin d'estimer p ni d'avoir observé des défectueux avant de mettre en place la carte mais qui nécessitent toutefois d'avoir une idée de la loi du paramètre p ; les cartes de type Shewhart pour les Time Between Event [33, 34].

Les cartes de contrôle comme les classiques CUSUM, EWMA reposent elles sur des fenêtres glissantes.

Notons également les travaux récents de Sotiris Bersimis qui conjugue les cartes de contrôle et les scans statistiques pour optimiser le contrôle de qualité (travaux non publiés actuellement, présentation congrès International Workshop in Applied Probability, Budapest juin 2018).

6.3 Une carte de contrôle "Score local"

Nous souhaitons travailler sur la suite des nombres d'observation entre deux cas. Nous proposons une échelle de scores de type

$$s(i) = E \cdot \frac{\mathbb{E}(X) - x_i}{\sigma(X)}$$

avec $X \hookrightarrow \mathcal{G}(p)$, $p = \mathbb{P}$(observer un cas) et E un paramètre d'ajustement $E = 10^k$ avec k dépendant de p. On a donc $s(i) > 0$ pour $x_i < \mathbb{E}[X]$, ie que le cas arrive plus tôt que la moyenne attendue. Nous proposons la carte de contrôle dite "Score local" définie de la manière suivante. Pour chaque i, nous proposons, dans un premier temps, de calculer le score local $h(i)$ de la séquence $x_1, \ldots, x_i$, avec x_i le nombre de réalisations avant le cas suivant. La carte de contrôle représente l'ensemble des points (i,p-valeur de $h(i)$). Elle comporte une seule limite de contrôle qui sera inférieure et qui a pour valeur le risque client α que l'on souhaite. L'alarme est donc donnée pour une p-valeur $< \alpha$.

La p-valeur du score local est calculée en utilisant les différentes méthodes vues dans la Partie I de ce document.

Remarque 6

- *Il est possible que les processus possèdent une dépendance temporelle (autocorrélation, questions liées à la saisonnalité, ...). Les résultats établissant la distribution du score local (exacte et/ou approchée) dans le cas d'un modèle markovien présentés dans le Chapitre 3 de la première partie de ce document peuvent être utilisés.*

- *Les séquences biologiques sont elles étudiées a postériori et non découvertes au fur et à mesure composant par composant comme cela peut se produire dans le suivi de production, de pharmaco vigilance de suivi médical ou de santé publique. On souhaite dans ces contextes réagir au plus tôt, une fois la rupture dans le processus réalisé, aussi on peut se poser la question si le score local est une statistique adaptée pour cela. Supposons une suite de valeurs observées $x_1, \ldots, x_j$ de score local h réalisé de k à $\ell < j$. On a alors le score local de la séquence $x_1, \ldots, x_\ell$ qui est également h. On sait que M_i est une fonction croissante en i de là $\mathbb{P}(M_\ell \leqslant h) \leqslant \mathbb{P}(M_j \leqslant h)$. Ainsi, si alarme il y a, ie $\mathbb{P}(M_j \leqslant h) < \alpha$ alors l'alarme aura également été donnée en ℓ dès la réalisation d'un score local exceptionnel.*

Par ailleurs, si dans un premier temps nous avons pensé au score local comme statistique de test, il nous semble important de coupler ce suivi avec la hauteur de la montagne en cours, i.e. l'excursion du processus de Lindley associée à la suite des $(X_i)_i$. En effet, il peut y avoir un phénomène de "noyade" du signal du score local qui dépend plus de la longueur totale de la suite que la hauteur d'une excursion. Cela revient à se poser la question "Quelle est le début de la séquence ?" ou encore "Quelle est la longueur de la séquence à prendre en considération ?". En effet sous H_0 = "Le processus est sous contrôle", la suite d'observations $x_1 \ldots x_n$ sans anomalie peut-être longue. Supposons qu'une

rupture arrive et que se réalise localement une suite d'observations de longueur ℓ sous H_1 ="Le processus n'est pas sous contrôle". Soit h la valeur du score de segment, $h = \sum_{i=n+1}^{n+\ell} s(x_i)$ qui correspond également au score local de la suite $x_1, \ldots, x_{n+\ell}$. La valeur h est aussi la hauteur de l'excursion en cours, $Q = h$. On peut avoir un événement tel que

$$\mathbb{P}(M_{n+\ell} > h) \geq \alpha \quad \text{et} \quad \mathbb{P}(Q > h) < \alpha \ .$$

Cela se réalisera d'autant plus souvent que la période sous contrôle observée précédant la rupture n sera grande. En effet, cela vient du fait de la croissance du score local en fonction de la longueur de la séquence n. Des contextes d'études pourrait impliquer l'existence d'une longueur maximale d'observations et/ou d'une ré-initialisation naturelle de la séquence : au lieu de $x_1 \ldots x_{n_1+n_2}$ on observe $x_1 \ldots x_{n_1}$ puis $x'_1 \ldots x'_{n_2}$ avec $x'_k = x_{n_1+k}$ pour $k = 1, \ldots, n_2$. Si le contexte ne permet pas cela, il est alors nécessaire de considérer la hauteur de l'excursion plutôt que le score local comme statistique de test. Des résultats sont disponibles pour la distribution de la hauteur d'une excursion quelconque dans le modèle d'indépendance. La distribution de la hauteur de la première montagne peut-être facilement établie dans le cas markovien (voir Partie III de ce document Chapitre 9). Cela revient à proposer le calcul de la hauteur de la première montagne en lecture inverse de la séquence i.e. $x_i, x_{i-1}, \ldots$. Dans le cas où deux statistiques, score local et score d'excursion, seraient utilisées conjointement, se pose la question de tests multiples avec deux statistiques différentes mais pas indépendantes.

6.4 Performance d'une carte de contrôle

Il est ensuite nécessaire de comparer les cartes nouvellement proposées aux différentes cartes usuelles. Voici différents critères que nous avons relevés dans la littérature.

- FAR : False Alarm Rate qui doit être le plus proche de α.
- ARL: Average Run Length between two non-conform items. Dans [13], deux ARL sont définis suivant que le process est sous ou bien hors control : ARL_0, calculé lorsque le process est in-control et qui doit est grand ; et ARL_1, calculé lorsque le process est out-of-control depuis le début de la surveillance et qui doit être petit pour une détection rapide de la rupture dans le process.
- $SDRL$: La moyenne ARL ne suffit pas. D'après [37], il faut également prendre en compte l'écart-type, $SDRL$.
- Certains auteurs distinguent le nombre de points notés sur la carte jusqu'à ce qu'une alarme soit donnée, mais également le nombre d'entités inspectées jusqu'à ce qu'une alarme soit donnée. La différence est présente lorsque la carte utilisée n'est pas une carte représentant les réalisations de

la loi géométrique des temps d'attente entre deux non-conformes. Les notations diffèrent suivant les auteurs : $ARL*_{avg}$ et $SDRL*$ [37] ; dans [36] les auteurs distinguent ANI (Average Number of Items) le nombre attendu d'entités inspectées avant une alarme, également appelé $ANOS$ (Average Number of Observations), et ATS (Average Time to Signal) le temps moyen observé sur la carte pour produire une alarme.

Les critères portant sur la détection d'une rupture sont ARL_1 [13]. Dans [13], les auteurs proposent également

- $CED(t) : \mathbb{E}[t_A - \tau | t_A \geq \tau - t]$ avec t_A le temps d'une alarme, et τ le temps de la rupture.

6.5 Projet en cours

Pour effectuer l'ensemble de ce travail nous avons obtenu, François Bergeret et moi-même, un financement PEPS1 nous permettant de rémunérer un stagiaire pour les implémentations nécessaires à l'évaluation des performances ainsi que divers déplacements et financer une mission à un congrès dans le domaine.

Chapter 7

D'autres projets d'application

7.1 Application au processus de polymérisation

Collaborateurs principaux : Céline Delmas (INRA de Castanet Tolozan), Jean-Daniel Marty et Simon Harrison (LIMRCP, Laboratoire des Interactions Moléculaires et Réactivité Chimique et Photochimique, UT3)

Travail démarré en novembre 2017.

7.2 Ouverture : score local à 2 dimensions

Ce document ne peut se terminer sans parler du score local à plusieurs dimensions. Les fenêtres glissantes de taille fixée restent un outil très utilisé dans bien des domaines, comme l'analyse des séquences biologiques, celui du suivi de production, mais aussi la géostatistique, et l'analyse d'images, pour ne cités que ceux-là. Je souhaiterais développer une collaboration avec les géographes de l'université Toulouse 2 Jean Jaurès, afin d'ancrer également ma recherche dans les Sciences Humaines et Sociales. Les contextes où la longueur de la fenêtre n'est apriori pas connue doivent exister et j'aimerais étudier les avantages de l'utilisation du score local dans ce domaine par rapport aux fenêtres glissantes et créer de nouveaux liens dans mon université d'accueil. De plus, les problématiques qui peuvent se poser dans le domaine de la géographie ou géostatistique ont un aspect multi dimensionnel, comme dans le cas de l'analyse d'image de manière générale (2 ou 3 dimensions). Des travaux sur le score local à deux dimensions existent. Voir par exemple les travaux de James Sharpnack [47, 48] ou encore de Zakhar Kabluchko [28].

Cette idée est certes encore très peu élaborée mais constitue un souhait bien réel de collaboration avec des enseignants chercheurs de l'UT2J. En effet, celle commencée avec des psycho-linguistes lors de mon arrivée à Toulouse n'avait

pu aboutir faute de moyens financiers pour réaliser les expériences nécessaires à l'estimation des paramètres des modèles dans le cas d'application du score local à des séquences "marquées" (lecture de syllabes sous différents stress) pour mettre en évidence des situations émotionnelles atypiques via le discours (dialogue d'un pilote d'avion à la tour de contrôle par exemple).

Part III

Projets de recherche - Ouvertures

La lecture des deux chapitres suivants a été souhaitée autonome, aussi un certain nombre de notations présentées dans les chapitres précédents sont à nouveau présentées.

Chapter 8

Projet "Chen-Stein"

Collaboratrice : Simona Grusea.

Karlin *et al.* [30, 31] établissent une approximation asymptotique de la loi du score local M_n d'une séquence dans le cas des modèles de séquence i.i.d. et markovien par une loi de Gumbel. Ils se reposent sur la théorie du renouvellement, des chaînes de Markov et des outils de larges déviations. Dans [51] Waterman et Vingron proposent une autre approche pour établir une approximation asymptotique de la loi du score local mais dans le cas de comparaison de deux séquences sans insertions-délétions par une loi de Gumbel dans le cas du modèle de deux séquences i.i.d. et indépendantes entre elles. Ils utilisent pour cela l'approximation de Chen-Stein [6] appliquée au nombre d'excursions (voir définition précise plus bas) dépassant un seuil t et établissent ainsi une approximation de la loi du nombre d'excursions par une loi de Poisson, avec un contrôle de l'approximation par la distance en variation totale. Le paramètre de la loi de Poisson étant sous la forme d'une exponentielle, cela vérifie la conjecture de l'époque que le résultat de Karlin *et al.* [30, 31] de l'approximation par une loi de Gumbel pour le cas d'une séquence i.i.d. s'étend bien au cas de la comparaison sans insertions-délétions de deux séquences i.i.d. (voir ci-dessous pour plus de détails). Hansen [24] s'intéresse au modèle markovien sur les séquences. Il établit une approximation de la loi du score local dans le cas de la comparaison sans insertions-délétions de deux chaînes de Markov de même longueur n, que nous noterons $\mathcal{M}_n$ pour le distinguer du score local d'analyse d'une séquence, par une loi de Poisson en se reposant sur la théorie de Chen-Stein.

Pour [51] et [24] qui utilisent la théorie de Chen-Stein, la démonstration consiste en deux étapes principales. La première correspond à obtenir une approximation du nombre d'excursions dépassant un certain seuil t, noté $C_n(t)$, par une loi de Poisson via le contrôle des bornes $(\beta_i(n))_i$ de Chen-Stein (voir (8.3-8.6)). La seconde consiste ensuite à faire le lien entre nombre d'excursions et le score local, étape immédiate par le fait que $\{C_n(t) = 0\} = \{\mathcal{M}_n \leqslant t\}$. Ainsi

$$\mathbb{P}(\mathcal{M}_n \leqslant t\}) \simeq \exp(-\lambda) \ .$$

Avec $\lambda = \lambda_n = \exp(u_n)$, u_n adapté (voir Équations (8.8)), nous retombons sur une double exponentielle de type loi de Gumbel.

Nous proposons dans ce projet "Chen-Stein" de transposer le résultat de Hansen [24] au nombre d'excursions dépassant un seuil donné, au cas de l'analyse d'une chaîne de Markov et non pour la comparaison de deux chaînes de Markov, ce qui a priori peut sembler un cas plus simple.

Soit $\mathbb{A} = (A_i)_{i\geqslant 1}$ une chaîne de Markov à valeurs dans un ensemble $\mathcal{A}$ de probabilité de transition P, irréductible, apériodique, de vecteur de probabilité invariant à gauche π_P. ($\mathbb{A}$ sera supposée stationnaire par la suite). Soit s la fonction de scores : on note $X_i = s(A_i)$. Soit $(X_i)_i$ la séquence des scores, $S_0 := 0$, $S_i := \sum_{k=1}^{i} X_k$ les sommes partielles et $S_i^\delta = \sum_{k=1}^{\delta} X_{i+k}$ le score du segment allant de $i+1$ à $i+\delta$. Soit $(T_i)_i$ le processus des temps échelles descendant : $T_0 = 0$ et $T_{i+1} = \min\{k \geq T_i,\ X_{T_i} + \cdots + X_k < 0\}$. La i-ième excursion (également appelée montagne) correspond à la portion de séquence $X_{T_i}, \ldots, X_{T_{i+1}-1}$. Notons Q_i le score, ou hauteur, de la i-ième excursion :

$$Q_i = \max_{T_i \leq k \leq T_{i+1}-1} \sum_{j=T_i}^{k} X_j \ .$$

Notons que l'hypothèse usuelle d'un score moyen négatif $\mathbb{E}[s(A)] < 0$, assure l'existence des excursions car dans ce cas, $\lim_{n\to\infty} = -\infty$ et $\mathbb{E}[T_i] < \infty$.

En pratique, on considère $(A_i)_{1\leq i\leq n}$ avec n la longueur de la séquence. Posons également les définitions de Hansen [24] transposées au cas d'analyse d'une séquence et notons $\Delta(i) = \inf\{\delta \geq 0 : S_i^\delta \leqslant 0 \text{ ou } i+\delta = n\}$ la longueur allant de i à la fin de l'excursion contenant i. Pour i tel que $\exists k, i = T_k$, $\Delta(i)$ correspond à la longueur de l'excursion dans sa totalité.

$$V_i(t) = V_i(t, \ell, \eta) = \mathbb{1}_{\{\max_{\delta;\delta\leqslant\Delta(i)\wedge\ell;d(\epsilon_i^\delta;\hat{\pi})\leqslant\eta\}} S_i^\delta > t; \exists k, i=T_k\}}$$

avec $\epsilon_i^\delta(x, x') = \frac{1}{\delta}\sum_{k=1}^{\delta} \mathbb{1}_{(x,x')}(X_{i+k-1}, X_{i+k})$ et $\hat{\pi}$ une distribution de couples en lien avec la stationnarité de la chaîne de Markov. On note $C_n(t)$ le nombre d'excursions de hauteur dépassant t.

$$C_n(t) = \sum_{1\leqslant i\leqslant n} \mathbb{1}_{\{\exists k, i=T_k; \max_{\delta\leqslant\Delta(i),}\ S_i^\delta > t\}} = \sum_{i=1}^{m(n)} \mathbb{1}_{\{Q_i > t\}} + \mathbb{1}_{\{Q^* > t\}}.$$

Le Théorème 4.1 de [24] pour $\tilde{C}_n(t)$ le nombre d'excursions, de hauteur dépassant t, dans le cas de la comparaison de deux chaînes de Markov sans insertions-délétions nous dit que pour t_n et λ_n adaptés,

$$||\mathcal{D}\left(\tilde{C}_n(t_n)\right) - \mathcal{P}(\lambda_n)||_{VT} \overset{n\to\infty}{\longrightarrow} 0, \tag{8.1}$$

avec $||.||_{VT}$ la distance en variation totale.

L'objectif de ce projet "Chen-Stein" est de montrer un résultat analogue dans le cas de l'analyse d'une chaîne de Markov pour $C_n(t)$. On peut penser rapidement qu'il s'agit en fait d'une simple adaptation mais cela n'est pas le cas.

Définissons le *voisinage de dépendance* associé à l'indice i de la manière suivante

$$\mathcal{B}_i = \{k \in I : |k - i| \leqslant 2 \cdot \ell_n\} \text{ avec } I = \{1, \ldots, n\} \text{ et } \ell_n \text{ à définir.} \tag{8.2}$$

Notons

$$\beta_{1,n} = |\sum_{i \in I} \mathbb{E}[V_i(t)] - \lambda_n| \tag{8.3}$$

$$\beta_{2,n} = |\sum_{i \in I; j \in \mathcal{B}_i} \mathbb{E}[V_i(t)] \cdot \mathbb{E}[V_j(t)]| \tag{8.4}$$

$$\beta_{3,n} = |\sum_{i \in I; j \in \mathcal{B}_i; j \neq i} \mathbb{E}[V_i(t) V_j(t)]| \tag{8.5}$$

$$\beta_{4,n} = \sum_{i \in I} \mathbb{E}|\mathbb{E}[V_i(t)|\mathcal{F}_i] - \mathbb{E}[V_i(t)]| \tag{8.6}$$

avec $(\mathcal{F}_i)_{i \geq 0}$ la filtration des σ-algèbres générées par la chaîne de Markov.

Conjecture 1 *Pour x réel, soient*

$$t_n = \frac{\log K^* + \log(n) + x}{\theta^*} \tag{8.7}$$

$$\lambda_n = \exp\left(-x + x_n\right) \text{ avec } x_n = \theta^*(t_n - \lfloor t_n \rfloor) \ . \tag{8.8}$$

Pour $(\ell_n)_{n \geqslant 1}$ et $\eta > 0$ choisis tels que $(\lambda_n)_n \geqslant 1$ vérifie $\beta_{1,n} \overset{n \to \infty}{\longrightarrow} 0$ et certaines conditions à mettre en évidence (voir par exemple (8.9)), alors pour $i = 1, \ldots, 4$, nous avons $\beta_{i,n} \overset{n \to \infty}{\longrightarrow} 0$ et

$$||\mathcal{D}\left(\sum_{i \in I} V_i\right) - \mathcal{P}(\lambda_n)||_{VT} \leqslant \beta_{1,n} + 2(\beta_{2,n} + \beta_{3,n} + \beta_{4,n}) \overset{n \to \infty}{\longrightarrow} 0 \ .$$

Les paramètres θ^* et K^* dans (8.7) et (8.8) dans le cas une chaîne de Markov doivent être explicités. Ils doivent en fait correspondent à leur analogues dans [31] et [23]. Pour cela, les hypothèses sur ℓ_n devront être mises en évidence afin de s'assurer du résultat.

Dans un document de Travail [22] S. Grusea et S. Mercier vérifient qu'il est facile de s'assurer que $\beta_{2,n} \overset{n \to \infty}{\longrightarrow} 0$ en utilisant une borne supérieure pour $\mathbb{E}[V_i(t)]$ similaire au cas de deux chaînes de Markov ($\mathbb{E}[V_i(t)] \leqslant K \exp(-\theta^* t)$). Cela amène une première condition pour ℓ_n:

$$\ell_n / n \overset{n \to \infty}{\longrightarrow} 0 \ . \tag{8.9}$$

Nous montrons également que les propriétés de α et β mélanges peuvent être utilisées et permettent de s'assurer du contrôle de $\beta_{4,n}$: $\beta_{4,n} \overset{n\to\infty}{\longrightarrow} 0$ quand $\ell_n \overset{n\to\infty}{\longrightarrow} \infty$. Contrôler $\beta_{3,n}$ peut également se faire à l'aide des propriétés de stationnarité et la condition supplémentaire que $\forall\alpha > 0$, $\ell_n \cdot n^\alpha \overset{n\to\infty}{\longrightarrow} 0$. Du travail reste à faire pour le contrôle de $\beta_{1,n}$. L'idée principale est de montrer que

$$\sum_i \mathbb{E}[V_i(t_n, \ell_n, \eta)] \sim \mathbb{E}[C_n(t_n)] \sim \exp(\lambda_n)$$

en utilisant une borne supérieure et inférieure possédant toutes deux le comportement asymptotique désiré. Pour cela définissons de manière similaire à Hansen

$$\tau^-(1) = \inf\{k > 0, S_k \leq 0\}$$

$$q_i(x) = \mathbb{P}(\exists k : i = T_k, X_i = x)$$

$$p(n, x) = \mathbb{P}_x(\max_{k \leqslant \tau^-(1)} S_k > t)$$

$$\tilde{p}(n, x) = \mathbb{P}_x(\max_{k \leqslant \tau^-(1) \wedge \ell_n ; d(\epsilon_\delta, \hat{\pi}) \leqslant \eta}, \ S_k > t)$$

Remarque 7 *$p(n, x)$ correspond à la probabilité que la hauteur de la première excursion est inférieure à t conditionnellement à $X_0 = x$ et $\tilde{p}(n, x)$ à une probabilité analogue en prenant les contraintes sur la longueur de l'excursion et que la composition du segment ne soit pas trop éloignée de $\hat{\pi}$, $q_i(x)$ la probabilité qu'une excursion commence en i avec un score x.*

Rappelons $I = \{1, \ldots, n\}$ et notons $\tilde{I} = \{i \in I : i \leqslant n - \ell_n\}$ on a

$$\sum_{x\in\mathcal{A}} \tilde{p}(n, x) \sum_{i\in\tilde{I}} q_i(x) \leqslant \sum_{i\in I} \mathbb{E}[V_i(t)] \leqslant \mathbb{E}[C_n(t_n)] \leqslant \sum_{x\in E} p(n, x) \sum_{i\in I} q_i(x)$$

Il suffit de s'assurer que les bornes supérieure et inférieure ont un comportement en $\exp(x - x_n)$ ce qui assurera $\beta_{1,n} \to 0$.

La borne supérieure n'est pas un problème en adaptant le travail d'Hansen [24]. Du travail reste à faire pour la borne inférieure. Une idée serait d'utiliser le principe des grandes déviations pour une chaîne de Markov. La clé revient à contrôler la longueur de la première excursion, ainsi que le score de cette première excursion, dans le cas où la distribution des couples $A_i A_{i+1}$ n'est pas forcément très proche de celle déduite à partir de la distribution stationnaire $d(\delta_i^\delta; \hat{\pi}) \geqslant \eta$.

Chapter 9

Projet “Scores de régions et tests multiples”

Collaborateurs : Simona Grusea et Pierre Neuvial.

Introduction

Il est apparu clairement que le travail de nombreux biologistes consistant à mettre en évidence des régions d'une séquence $\mathbb{A} = A_1, \ldots, A_n$ significativement intéressantes ne se posait pas uniquement sur la région potentiellement la plus intéressante dans une séquence, comme celle réalisant le score local pour une fonction de scores donnée s, mais sur toutes les régions potentielles réalisant un score positif (voir par exemple [38]). Cela se traduit par l'étude de la significativité de chacune des régions. Ces régions correspondent aux excursions du processus de Lindley $\mathbb{U}$ associé à la séquence des scores étudiée $(X_i) = (s(A_i))_{1 \leq i \leq n}$, avec $\mathbb{U}$ défini récursivement par : $U_0 = 0$ et $U_{i+1} = \max(0, U_i + X_{i+1})$. On peut également définir les excursions de la manière suivante. Soit $(T_i)_i$ le processus des temps échelles descendant : $T_0 = 0$ et $T_{i+1} = \min\{k \geq T_i,\ X_{T_i} + \cdots + X_k < 0\}$. La i-ème région-excursion, également appelée montagne, correspond à la portion de séquence $X_{T_i}, \ldots, X_{T_{i+1}-1}$. Notons Q_i le score, ou hauteur, de la i-ème excursion : $Q_i = \max_{T_i \leq k \leq T_{i+1}-1} \sum_{j=T_i}^{k} X_j$. Sous l'hypothèse d'un score moyen négatif $\mathbb{E}[s(A)] < 0$, nous avons $\lim_{n\to\infty} S_n = -\infty$ et $\mathbb{E}[T_i] < \infty$. Soit $m(n)$ le nombre d'excursions complètes dans la séquence $\mathbb{A}$ et soit Q^* la hauteur de la dernière montagne incomplète : $Q^* = \max_{T_{m(n)} \leq k \leq n} \sum_{j=T_i}^{k} X_j$. Le score local qui peut être défini mathématiquement de différentes manières correspond à la hauteur de la plus grande excursion : $M_n = \max\left(\max_{1 \leq i \leq m(n)} Q_i, Q^*\right)$.

La question statistique pour répondre au besoin des biologistes, revient à mettre en place une méthode pour choisir un score seuil, noté t tel que toute région ayant un score dépassant cette valeur à déterminer est “intéressante” : $Q_i > t \Rightarrow i$-ème région significativement intéressante. Que l'on peut également

reformuler par trouver un seuil α approprié tel que $\mathbb{P}(Q_i \geq q_i) < \alpha \Rightarrow i$-ème région significativement intéressante, pour q_i le score observé de la i-ème région. Nous sommes donc ici dans un cas de tests multiples où le nombre $m(n)$ de tests effectués, correspondant au nombre d'excursions dans la séquence, est observé mais aléatoire.

Les séquences sont généralement modélisées par une suite de variables aléatoires i.i.d. ou bien une chaîne de Markov. Nous considérons ici les deux modèles.

Nous proposons ici deux approches autour des tests multiples. La première consiste à étendre le travail effectué dans le cas où les séquences sont modélisées par une suite de variables i.i.d. au cas markovien et qui propose une correction de type Bonferroni [12] pour les tests multiples pour le nombre aléatoire $m(n)$ de régions (cf. [38], Supplementary materials) . Cela nécessite d'établir dans le modèle markovien la loi de la hauteur d'une excursion quelconque, Q_i, puis dans une seconde étape à prendre en compte le nombre aléatoire de régions. Nous commençons par rappeler les résultats existants sur la loi des hauteurs d'excursions dans les modèles de séquence i.i.d. et markovien. Nous proposons des conjectures et exposons les grandes lignes de démonstration de celles-ci auxquelles nous pensons.

La seconde approche est une ouverture plus large qui se place à la suite du projet "Chen-Stein : Approximation de Poisson pour la distribution du nombre d'excursions dans un modèle markovien" (cf. chapitre précédent). Ce travail correspond à une étude globale d'une séquence en déterminant si cette dernière, pour un seuil t donné, contient un nombre significatif d'excursions dépassant t. Dans le cas où t n'est pas déterminé par le contexte biologique, il existe une multitude de valeurs de t possibles (de 0 à la hauteur maximale de toutes les régions-excursions). Il s'agit donc d'un test multiple que l'on pourrait appelé "continu" à la différence des tests multiples plus usuels pour un nombre discret de tests effectués. Ce cadre assez nouveau de travail proposant un ensemble continu de tests a été assez récemment abordé par Roquain *et al.* (cf. [10, 44]).

Remarque 8 *Notez que cette seconde approche ne fait pas intervenir les mêmes compétences mathématiques que la première approche pour une correction dans le cas des tests multiples pour un nombre aléatoire de régions. Dans la première approche, tests multiples de régions significatives en nombre aléatoire et procédure de Bonferroni, les compétences mathématiques sont proches de celles nécessaires au projet "Chen-Stein" (théorie des chaînes de Markov, du renouvellement, grandes déviations). Dans la seconde, test multiple "continu", les compétences mathématiques correspondent au domaine des test multiples. Comme le dit Etienne Roquain dans l'introduction de son manuscrit de HdR [46], cela correspond à une grande variété de concepts théoriques tels que combinatoire, réechantillonnage, processus empirique, inégalité de concentration, dépendance positive, ... parmi d'autres.*

Quelques éléments essentiels sur les tests multiples sont présentés à la sous-section 9.1.2.

9.1 Tests multiples en nombre aléatoire et procédure de Bonferroni

9.1.1 Resultats actuels et conjectures

Dans le modèle de séquence i.i.d. nous avons les résultats suivants

Résultat 17 (Modèle i.i.d.)

- *La loi exacte [22] et des approximations asymptotiques [14, 31] pour le score d'une excursion quelconque Q_i.*
- *La loi exacte [40] et approchée du score local M_n [14, 31].*
- *Faire un test multiple sur l'ensemble des $m(n)$ excursions $Q_1, \ldots, Q_{m(n)}$ avec une correction de type Bonferroni [12] revient à utiliser le quantile fourni par la loi du score local M_n (cf. Supplementary materials de [38]).*

Dans le modèle markovien des résultats similaires sont aussi disponibles

Résultat 18 (Cas markovien)

- *Loi exacte de la première montagne Q_1.*
- *Loi approchée de la première excursion Q_1 [23] (HAL Id : hal-01726031) et [31].*
- *La loi exacte [25] et approchée du score local M_n [23, 31].*

Le premier point du résultat précédent s'établit facilement en utilisant les techniques de [25] appliqué à Q_1 au lieu de M_n [22]. Les résultats suivants sont tout à fait envisageables :

Conjecture 2 (Cas markovien)

1. *Approximations asymptotiques de la loi du score d'une excursion quelconque Q_i.*
2. *Faire un test multiple sur l'ensemble des $m(n)$ excursions $Q_1, \ldots, Q_{m(n)}$ avec une correction de type Bonferroni revient à utiliser le quantile fourni par la loi du score local M_n.*

1. Dans [31] (cas markovien) et dans [23], la démarche principale permettant d'établir les résultats sur l'approximation de la loi du score local M_n est la suivante :

   1- Avoir un résultat sur la loi de la hauteur de la première excursion Q_1.

   2- En déduire un résultat sur la loi de n'importe quelle excursion Q_i.

3- Porter les résultats au score local M_n par l'intermédiaire de la relation

$$M_n = \max\left(\max_{1\leq i\leq m(n)} Q_i, Q^*\right) . \tag{9.1}$$

Dans le cas i.i.d. le passage de 1- à 2- n'est pas un problème car les Q_i sont i.i.d. On utilise ensuite des propriétés de la théorie du renouvellement et des chaînes de Markov, ainsi que des propriétés intrinsèques à la définition du score local, pour porter le résultat de 1- à 2-. Dans le cas markovien, il faut considérer que ce sont les Q_i conditionnellement aux valeurs prises par la séquence en début des excursions qui sont i.i.d. Nous proposons ensuite d'utiliser la même démarche pour porter le résultat sur la loi de la première excursion Q_1 dans le cas markovien, aux autres excursions Q_i. (Sujet de stage de M2).

2. Par ailleurs, le passage du point 2- à 3- dans [23,31] repose d'une part sur le fait que les Q_i conditionnellement aux valeurs prises par la séquence en début des excursions qui sont i.i.d. et d'autre part sur l'écriture du score local M_n comme étant de manière approchée le maximum d'un nombre $m(n)$ aléatoire de Q_i (cf. (9.1)). La gestion du nombre aléatoire des Q_i a pu être effectée dans le cas markovien [23,31]. La méthode correspondante dans le cas du modèle i.i.d. a permis d'obtenir une correction de type Bonferroni pour le test multiple d'excursions. Nous proposons de l'écrire d'une manière analogue dans le cas markovien.

9.1.2 Questions ouvertes sur les tests multiples

Commençons par poser quelques notions relatives aux tests multiples. Posons les hypothèses relatives au test d'une excursion j. Soit $H_{0,j}$: "Les composants de la j-ème excursion sont issus d'une distribution $\mathcal{D}_0$ donnée", et $H_{1,j}$: "Les composants de la j-ème excursion sont issus d'une distribution $\mathcal{D}_1 \neq \mathcal{D}_0$". On dira qu'une excursion j est "intéressante" (pour le biologiste) si $H_{1,j}$ est vraie, ce qui est équivalent à $H_{0,j}$ est fausse. On peut poser 3 types de problématique de niveau de contraintes différent.

1. On souhaite savoir s'il existe au moins une excursion intéressante et l'on définit $H_0(totale) = \cap_j H_{0,j}$.

2. On s'intéresse au pourcentage de montagnes intéressantes *ie* le pourcentage de j tel que $H_{0,j}$ est fausse.

3. Une question de sélection : On veut identifier tous les j tels que $H_{0,j}$ est fausse *ie* détection toutes les excursions réellement intéressantes.

On définit les critères de contrôle suivants. On appelle FP le nombre de faux positifs, c'est-à-dire le nombre de tests rejetés alors que l'hypothèse nulle est vraie (erreur de type I). Soit FDP la False Discovery Proportion correspondant au nombre de faux positifs FP divisé par le nombre total de rejet. On appelle

le False Discovery Rate $FDR = \mathbb{E}(FDP)$. Le $FWER$ est la probabilité de faire un faux positif ou plus parmi dans l'ensemble des tests effectués et en ce sens est en lien avec la problématique 3.

$FWER$ et FDR sont des critères de qualité d'un test multiple. Pour contrôler ces critères différentes procédures sont établies ou proposées.

Family-Wise-Error-Rate ($FWER$) : contrôler le $FWER$, comme le propose la procédure de Bonferroni [12], est assez exigeant, car on garantit alors avec une forte probabilité que toutes les montagnes dont la hauteur dépasse le seuil corrigé sont assurées d'être des montagnes "intéressantes" ($H_{0,j}$ rejetée à raison). Et en même temps, cette correction peut s'avérer assez sévère dans le cas de dépendance entre les différents tests.

False Discovery Rate (FDR) : Benjamini et Hochberg [7] propose une procédure pour contrôler le False Discovery Rate (FDR). Il s'agit d'une garantie plus souple que de contrôler le $FWER$. Le contrôle s'effectue sur le taux de faux positifs qui ne doit pas être trop grand : $\mathbb{E}[FDP]$ <borne, une borne donnée par l'utilisateur. Certes, on se rend compte que ce n'est pas forcément très utile d'avoir un contrôle sur l'espérance seule. Un contrôle sur la variabilité du FDP est aussi nécessaire. Actuellement, Pierre Neuvial et Etienne Roquain travaillent sur les quantiles de FDP [11] (HAL Id : hal-01483585).

Questions :

1. Peut-on faire mieux que Bonferroni en terme de procédure pour le critère $FWER$ pour la problématique du nombre aléatoire de tests ?

2. L'objectif de la méthode de Bonferroni est-il le bon ? N'est-il pas plus pertinent de raisonner en terme de faux positifs (FP) ? Est-ce intéressant pour le biologiste d'avoir une garantie sur le taux de FP de ce type là ?

Remarque 9 *Etant donnée la complexité des articles tels que [31], de l'ensemble de la bibliographie qu'il est nécessaire de s'approprier, et des techniques à utiliser, l'ensemble du travail présenté précédemment pourrait donner lieu à un travail de thèse.*

9.2 Approche : nombre d'excursions dépassant un seuil

Cette approche est beaucoup plus ouverte que la première proposée dans la section précédente. Il s'agit d'une problématique non standard dont la formulation seule est déjà un travail à effectuer.

Il s'agit ici de rebondir sur la conjecture du projet Chen-Stein (voir chapitre précédent), de voir comment ce résultat peut être appliqué et créer une ouverture pour, et avec, le groupe de travail organisé par Pierre Neuvial de l'IMT sur les tests multiples.

Rappelons la notation $C_n(t)$, correspondant au nombre d'excursions de hauteur dépassant t, et donc au nombre de rejets de $(H_{0,j})_j$:

$$C_n(t) = \sum_{i=1}^{m(n)} \mathbb{1}_{\{Q_i > t\}} + \mathbb{1}_{\{Q^* > t\}}.$$

D'après Hansen [24], la loi de $C_n(t)$ dans le cas de la comparaison de deux séquences markoviennes sans insertions-délétions, peut être approchée par une loi de Poisson. On a

$$||\mathcal{D}\left(C_n(t_n)\right) - \mathcal{P}(\lambda_n)||_{VT} \underset{n\to\infty}{\to} 0$$

avec $||.||_{VT}$ la distance en variation totale et

$$t_n = \frac{\log K^* + \log(n) + x}{\theta^*}, \tag{9.2}$$

$$\lambda_n = \exp\left(-x + x_n\right) \text{ et } x_n = \theta^*(t_n - \lfloor t_n \rfloor), \tag{9.3}$$

pour x réel. Les quantités K^* et θ^* dépendent des paramètres de la chaîne de Markov et de la fonction de scores choisie. Nous proposons dans le projet "Chen-Stein" de porter ce résultat au cas des scores d'excursions pour l'analyse d'une chaîne de Markov (alors que le travail de Hansen s'intéresse au cas de la comparaison de deux chaînes de Markov).

Soit alors un seuil $\alpha \in]0,1[$ donné et une longueur de séquence donnée n suffisamment grande pour que l'approximation de Chen-Stein soit de qualité. Ainsi, $\forall t$ donné, on déduit via (9.2) une valeur de $x = x_n(t)$ et une valeur correspondante de λ_n. On peut donc ensuite obtenir le quantile $q_\alpha(n,t)$ de la loi de Poisson de paramètre λ_n tel que de manière approchée $\mathbb{P}(C_n(t_n) > q_\alpha(n,x)) \leq \alpha$. On a donc ainsi la possibilité de déterminer pour tout t si une séquence possède un nombre significatif au seuil α, de montagnes de hauteur dépassant le seuil t. Pour une séquence donnée de longueur n, notons $c_n(t)$ la fonction associant pour chaque t le nombre observé de montagnes dépassant t : réalisation de la variable aléatoire $C_n(t)$ pour cette séquence observée. Cette fonction est décroissante par morceaux et atteint 0 pour t grand. Pour α donné, on peut comparer pour chaque t, $c_n(t)$ à $q_\alpha(n,t)$ et déterminer l'ensemble des valeurs de t tel que $c_n(t) \geq q_\alpha(n,t)$. Notons $\mathcal{T}_\alpha = \{t : c_n(t) \geq q_\alpha(n,t)\}$ cet ensemble.

La Figure 9.2 représente d'une part les quantiles $q_\alpha(n,t)$ obtenus à partir des formules ci-dessus (9.2,9.3) pour une chaîne de Markov de paramètres donnés. Les valeurs de K^* et θ^* ont été calculées à partir des formules de [31] pour le cas markovien. En effet, nous avons $\{C_n(t) = 0\} = \{M_n \leqslant t\}$, d'où $\mathbb{P}(M_n \leqslant t_n) - \exp(-\lambda_n) \overset{n\to\infty}{\longrightarrow} 0$, et nous retrouvons bien par l'approximation de Chen-Stein le fait que la fonction de répartition du score local peut être approchée par une double exponentielle correspondant à une loi de Gumbel. Les paramètres K^* et θ^* doivent donc correspondre entre ceux attendus dans la conjecture du projet "Chen-Stein" sur la loi de $C_n(t)$ et ceux de [31] dans le cas markovien.

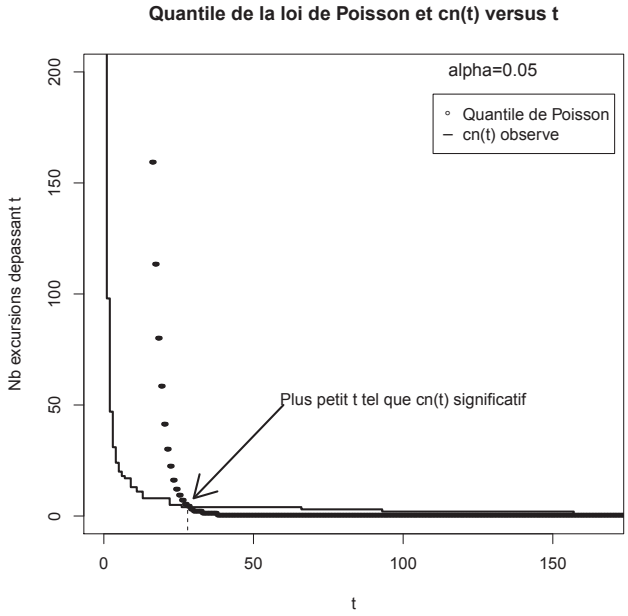


Figure 9.1: Allure de $q_\alpha(n,t)$ pour $\alpha = 5\%$ et comparaison avec le nombre observé $c_n(t)$ d'excursions dépassant t sur une chaîne de Markov simulée.

Nous vérifions bien dans la figure ci-dessous le fait que pour α fixé, $q_\alpha(n,t)$ est décroissante en t et que $c_n(t)$ est constante par morceaux.

Questions : Pour un seuil non déterministe, la statistique choisie dépendra quoiqu'il en soit de l'ensemble continu $\mathcal{T}_\alpha$. Nous pouvons proposer par exemple $\hat{t}_\alpha = \min \mathcal{T}_\alpha$. Mais dans ce cas quelle garantie veut-on fournir ?

Est-il possible de prendre en compte une correction sur le choix de α afin d'obtenir une valeur $\hat{t}$ adéquat, ou ayant un sens, du fait de cette multiplicité de tests ? Suivant le contexte d'application, doit-on choisir le seuil t ou bien α ?

Par ailleurs, deux approches semblent possibles :

1. Choisir un t et tester si la séquence étudiée est exceptionnelle avec l'idée d'étudier plus localement les montagnes significatives.
2. Avoir une approche plus globale de la séquence, la prise en compte de l'allure globale du processus de Lindley définissant les excursions et pour un ensemble de t : $((t_k)_{1\leq k\leq K}$ ou bien $\{t \in I\}$ pour I un intervalle.

La problématique n'est pas familière et il reste à bien définir et formuler les choses.

Conclusion

Le score local d'analyse d'une séquence a été défini en 1990 par Karlin et Altschul. La motivation première était alors de proposer un outil pour la comparaison de séquences : la mise en évidence des segments de chaque séquence les plus ressemblants permettant ainsi d'associer la fonction biologique du segment d'une séquence à l'autre. De nombreux travaux ont été entrepris sur le score local de comparaison qui a longtemps constitué le centre d'intérêt principal. Par ailleurs et de manière assez surprenante, les biologistes continuaient et continuent encore très souvent à analyser ou à scanner leurs séquences en utilisant des fenêtres glissantes de taille fixée, taille dont le choix n'est a priori pas toujours évident et parfois fluctuant. J'ai commencé à m'intéresser au score local à partir de ma thèse. J'ai continué d'en étudier la distribution car j'ai identifié de réels besoins dans le domaine. J'ai tout d'abord identifié de manière évidente le besoin de prendre en compte la non connaissance de la taille des segments ; puis la dépendance des composants ; le fait qu'un score local qui se réalise sur un petit segment ou bien un grand segment ne comporte pas la même densité information ; que les régions d'intérêt ne se limitent pas à la région la plus exceptionnelle...

D'autre part, cette thématique de recherche m'a permis de saisir et parfois susciter des opportunités de collaboration qui se sont présentées à moi : Claudie Chabriac (modèle markovien) ; Louis Ferré (co encadrement de thèse en biostatistique) ; recrutement d'Agnès Lagnoux (événements rares) ; Pierre Vallois (mouvement brownien) ; Grégory Nuel (HMM), ... Le score local est par ailleurs un outil dont le champ d'application est loin de se résumer à l'analyse des séquences biologiques et peut être utilisé dans de multiples domaines comme la gestion de production et le contrôle qualité. Ce point contribue également à renforcer mon intérêt pour le score local étant donné mes différentes activités dans le domaine du contrôle de la qualité, comme mes enseignements, les multiples suivis de stagiaires, ou encore les liens divers avec le monde de l'entreprise que j'ai largement développés.

J'ai exposé dans la Partie I de ce document les résultats établis avec mes différents collaborateurs permettant de détecter une région atypique au sein d'une séquence et ceci sous des modèles différents, suites de variables aléatoires indépendantes et identiquement distribuées (i.i.d.), chaînes de Markov (CM), ou bien chaînes de Markov cachées (HMM) pour des distributions a posteri-

ori. Nous avons vu que pour chaque cas de longueur de séquences (petites, moyennes, longues ou très longues), une méthode est disponible, dans le modèle i.i.d. comme pour les CM. Certes, il n'existe pas de résultat exploitable pour de très longues séquences et un score moyen positif. Cependant, cela constitue un cas d'application peu intéressant pour une recherche "locale" de région : la région d'intérêt sera dans ce cas plutôt la séquence dans son ensemble. Nos résultats récents (2015-2018) portent également sur la localisation du segment réalisant le score local et/ou la longueur du score local. Ces derniers résultats dans le cas des séquences aléatoires sont applicables à des séquences de très grandes longueurs et un score moyen nul. L'approche par chaînes de Markov cachées, pour une séquence observée donnée, permet des informations sur la longueur et la position pour de petites à très grandes séquences. Par ailleurs, les applications sur les séquences biologiques présentées dans la Partie II montrent la pertinence du score local et nous encourage à développer l'outil pour d'autres domaines d'application. C'est le cas pour la maîtrise statistique des procédés développée au Chapitre 6, projet pour lequel nous avons obtenu, mon collaborateur François Bergeret d'Ippon innovation et moi-même, un financement PEPS de 6,5K euros pour 2 ans.

Nous pouvons dire qu'actuellement, un ensemble de résultats existe pour répondre aux besoins. A ceci, nous pouvons toutefois opposer deux points. Le premier étant qu'un résultat théorique peut longtemps rester sous silence et inutilisé. L'approximation de Karlin et Dembo de 1992 [31] dans le cas markovien n'a pas à ma connaissance été appliqué malgré les besoins dans les modèles de dépendance. Il nous a paru important de fournir un outil pratique et adapté aux utilisateurs. Un package R est en cours de réalisation (juillet 2018) qui a pour objectif de réunir les résultats disponibles, dans un premier temps pour le modèle i.i.d. puis CM (stage financé grâce au projet CIMI "Hightlight", projet dont je suis porteuse pour un montant de 11K euros pour 2 ans). Le second point qu'il me semble important de soulever est que si de nombreux résultats existent actuellement et sont bientôt disponibles pour les utilisateurs, du travail théorique reste toutefois encore à faire. En effet, nous avons mis en évidence l'importance des scores de montagnes et les questions à multiples facettes de la multiplicité des tests qui en découlent. Nous avons proposé pour cela deux projets de recherche qui permettraient d'avancer dans cette direction. Ce document comporte également d'autres ouvertures comme le score local à plusieurs dimensions pour l'analyse d'images par exemple.

Je reprendrai pour finir, les éléments d'une conversation que j'ai eue avec plaisir avec Bernard Prum il y a quelques années, alors que je passais par un moment de découragement et m'interrogeais sur ma spécialité mathématique. Le score local est un outil puissant. Etudier la distribution du score local m'a amené à utiliser et découvrir des outils mathématiques divers provenant de théories mathématiques variées. A la vue des récents résultats théoriques, des travaux d'applications entrepris et des ouvertures possibles présentés dans ce document, il me semble que cela valait la peine de persévérer.

Annexes

9.3 Abréviations

v.a. variable aléatoire
i.i.d. indépendant et identiquement distribué
CM chaîne de Markov
p.s. presque surement

9.4 Principales notations

Notations par ordre alaphabétique (latin puis grec)

$\mathbb{A} = (A_i)_i$: séquence de composants (nucléotides, acides aminées, locis, ...)
M_n : le score local de la séquence, notation de Karlin *et al.* de 1990
$\mathbf{P} = (p_{\alpha\beta})_{\alpha\beta}$: probabilité de transition de la CM des $(A_i)_i$
$\tilde{\mathbf{P}}$: probabilité de transition de la CM (U_i^*, X_{i+1})
$\breve{\mathbf{P}}$: probabilité de transition de la CM (U_i^*, A_i, A_{i+1})
Q_i : hauteur de la i-ème excursion dans le processus des sommes partielles et de Lindley
$\mathbf{Q}$: matrice de probabilité de transition de $(A_{T_i})_i$ dans le cas markovien
s : fonction de scores, appelée aussi échelle de scores
$S_k = \sum_{1 \leq i \leq k} s(A_i)$: somme partielle ($S_0 = 0$)
S^+: maximum des sommes partielles, $S^+ > 0$
T_i : temps échelle descendant des sommes partielles
U : processus de Lindley associé à la séquence des scores (X_i)
U^* : processus de Lindley arrêté en une valeur donnée
$(X_i)_i$: séquence des scores ($X_i = s(A_i)$)

$\mathcal{A} = \{\alpha, \beta, \dots\}$: ensemble fini des lettres codants pour les composants des séquences
λ : un des deux paramètres de la loi de Gumbel dans le cas i.i.d.
π : probabilité stationnaire de $\mathbf{P}$
θ^* : un des deux paramètres de la loi de Gumbel dans le cas markovien.

Bibliography

[1] Fayyaz A., Mercier S., and Ferré L. h-tuple approach to evaluate statistical significance of biological sequence comparison with gaps. *Statistical Applications in Genetics and Molecular Biology*, 6-1, 2007.

[2] Fayyaz A., Mercier S., Ferré L., and Hassenforder C. New approximate p-value of gapped local sequence alignments. *C. R. Acad. Sci. Paris*, 346/1-2:87–92, 2007.

[3] Lagnoux A., Mercier S., and Vallois P. Probability that the maximum of the reflected brownian motion over a finite interval $[0; t]$ is achieved by its last zero before t. *Electronical Communication in Probability*, 20(62), 2015.

[4] Lagnoux A., Mercier S., and Vallois P. Statistical significance based on length and position of the local score in a model of i.i.d. sequences. *Bioinformatics*, 33(5):654–660, 2017.

[5] Lagnoux A., Mercier S., and Vallois P. Probability density function of the local score position. *Accepted Stochastic Processes and their Applications*, 2018.

[6] A. D. Barbour. Poisson approximation and the Chen-Stein method: Comment. *Statist. Sci.*, 5(4):425–427, 11 1990.

[7] Y. Benjamini and Y. Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, 57(1):289–300, 1995.

[8] F. Bergeret and S. Mercier. *Maîtrise Statistique des Procédés - Principes et cas industriels*. Gestion industrielle, Usine Nouvelle. Technique et Ingénieurie, Dunod, 2011.

[9] P. Billingsley. *Convergence of probability measures*. Wiley Series in Probability and Statistics: Probability and Statistics. John Wiley & Sons Inc., New York, second edition, 1999. A Wiley-Interscience Publication.

[10] G. Blanchard, S. Delattre, and E. Roquain. Testing over a continuum of null hypotheses with False Discovery Rate control. *Bernoulli*, 20(1):304–333, 2014.

[11] Gilles Blanchard, Pierre Neuvial, and Etienne Roquain. Post hoc inference via joint family-wise error rate control. March 2017. Submitted.

[12] C. Bonferroni. Il calcolo delle assicurazioni su gruppi di teste. *Tipografia del Senato*, 1935.

[13] Sonesson C. and Bock D. A review and discussion of prospective statistical surveillance in public health. *J. R. Statist. Soc. A*, 166:5–21, 2003.

[14] F. Cellier, D. Charlot and S. Mercier. An improved approximation for assessing the statistical significance of molecular sequence features. *Jour. Appl. Prob.*, 40:427–441, 2003.

[15] F. Cellier D., Charlot and S. Mercier. An improved approximation for assessing the statistical significance of molecular sequence features. *Jour. Appl. Prob.*, 40:427–441, 2003.

[16] C. Chabriac, A. Lagnoux, S. Mercier, and P. Vallois. Elements related to the largest complete excursion of a reflected Brownian motion stopped at a fixed time. Application to local score. *Stochastic Processes and their Applications*, 124(12), 2014.

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

[18] A. Dembo, S. Karlin, and O. Zeitouni. Critical phenomena for sequence matching with scoring. *Ann. Probab.*, 22(4):1993–2021, 1994.

[19] M.P. Etienne. Le score local : un outil pour l'?analyse de séquences biologiques. *Thèse de doctorat, Université de Henri Poincaré, Mathématiques Appliquées*, 2002.

[20] Food and Drug Administration. Guideline for postmarketing reporting of adverse drug experiences. *Fed. Reg.*, 57(248):61437, 1992.

[21] Food and Drug Administration. Postmarketing expedited adverse experience reporting for human drugs and licensed biological products; increased frequency report final rule. *Fed. Reg.*, 62(122):34166–34168, 1997.

[22] S. Grusea and S. Mercier. Scores d'excursions dans les modèles i.i.d. et markovien. *Document de travail*, 2017.

[23] S. Grusea and S. Mercier. Improvement on the distribution of maximal segmental score in a Markovian sequence. *submitted*, March 2018. Submitted.

[24] N.R. Hansen. Local Alignment in Markov Chains. *Ann. Appl. Proba.*, 16(3):1262–1296, 2006.

[25] C. Hassenforder and S. Mercier. Exact distribution of the local score for markovian sequences. *AISM*, 59(4):741–755, 2007.

[26] S. Mercier J.-N. Bacro, J.-J. Daudin and S. Robin. Back to the local score in the logarithmic case : a direct and simple proof. *AISM*, 54(4):748–757, 2002.

[27] Benneyam J.C. Statistical control charts based on geometric and negative binomial populations. *Master thesis, University of Massachusetts, Amherst*, 1991.

[28] Z. Kabluchko. Extremes of the stabdardized gaussian noise. *Stoch. Proc. and Appl.*, 121:515–533, 2011.

[29] Davis R.B. Kaminsky F.C., Benneyan J.C. and Burke R.J. Statistical control charts based on a geometric distribution. *Journal of Quality Technology*, 24(2):63–69, 1992.

[30] S. Karlin and S.-F. Altschul. Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *PNAS*, 87:2264–2268, 1990.

[31] S. Karlin and A. Dembo. Limit distributions of maximal segmental score among Markov-dependent partial sums. *AdAP*, 24:113–140, 1992.

[32] P. E. Kjelle. Alarm criteria for the fixed gamma radiation monitoring stations. *Report 87-07, Swedish Radiation Protection Institute, Stockholm*, 1987.

[33] Chakraborti S. Kumar N. Phase ii shewart-type control charts for monitoring times between events and effects of parameter estimation. *Quality and Reliability Engineering International*, 2014.

[34] Chakraborti S. Kumar N. Improved phase i control charts for monitoring times between events. *Quality and Reliability Engineering International*, 31:659–668, 2015.

[35] Raubenheimer L. and van der Merwe A.J. Bayesan control chart for nonconformities. *Quality and Reliability Engineering International*, 31:1359–1366, 2015.

[36] Zang M., Nie G., and He Z. Performance of cumulative count of conforming chart of variable sampling intervals with estimated control limits. *Int. J. Production Economics*, 150:114–124, 2014.

[37] Zang M., Peng Y., Schuh A., Megahed F.M., and Woodall W.H. Geometric charts with estimated control limits. *Qual. Reliab. Engng. Int.*, 29:209–223, 2013.

[38] Fariello M.-I., Boitard S., Mercier S., Robelin D., Faraut T., Arnould C., Le Bihan-Duval E., Recoquillay J., Salin G., Dahais P., Pitel F., Leterrier C., and Sancristobal M. Accounting for linkage disequilibrium in genome scans for selection without individual genotypes : the local score approach. *Molecular Ecology*, 26(14):3700–3714, 2017.

[39] S. Mercier, H. Chiapello, and G. Nuel. Detecting several atypical segments in sequences. *In preparation*, x.

[40] S. Mercier and J.J. Daudin. Exact distribution for the local score of one i.i.d. random sequence. *Jour. Comp. Biol*, 8(4):373–380, 2001.

[41] S. Mercier and C. Hassenforder. Distribution exacte du score local, cas markovien. *C. R. Math. Acad. Sci. Paris*, 336(10):863–868, 2003.

[42] S. Mercier and G. Nuel. Probabilizing the segmentation space in local score approaches. 2018. Submitted.

[43] G. Nuel and Prum B. *Analyse statistique des séquences biologiques. Modélisation markovienne, alignements et motifs.* Coll. Bioinformatique, 2006.

[44] F. Picard, P. Reynaud-Bouret, and E. Roquain. Continuous testing for poisson process intensities: A new perspective on scanning statistics. 2017. Submitted.

[45] A. Krogh R. Durbin, S. Eddy and G. Mitchison. *Biological sequence analysis. Probabilistic models of proteins and nucleic acids.* Cambridge University Press, Cambridge, seventh edition, 2002.

[46] E. Roquain. *Contributions to multiple testing theory for high-dimensional data., HdR.* Université Pierre et Marie Curie, 2015.

[47] J. Sharpnack and E. Arias-Castro. Exact asymptotics for the scan statistic and fast alternatives. *Electronic Journal of Statistics*, 10:2641–2684, 2016.

[48] James Sharpnack. Learning patterns for detection with multiscale scan statistics. Feb. 2018. Submitted.

[49] A. Sveréus. Detection of successive changes: statistical methods in postmarketing surveillance. *Research Report, Department of Statistics, GÂ¨oteborg University, Goteborg*, 2, 1995.

[50] M. S. Waterman. *Introduction to Computational Biology: Maps, Sequences and Genomes.* Chapman & Hall, 1995.

[51] M.-S. Waterman and M. Vingron. Sequence comparison significance and Poisson approximation. *SS*, 9:367–381, 1994.

Distribution du score local pour la mise en évidence de segments atypiques au sein de séquences

-

Sabine MERCIER

Le score local est un outil classique utilisé pour mettre en évidence des régions d'intérêt dans une séquence. Il quantifie le niveau maximal d'une propriété présente localement. Pour l'établir il est nécessaire de définir un score pour chaque composant. Le score local correspond au maximum obtenu en cumulant les scores de composants consécutifs pour tous les segments possibles de la séquence. La question principale consiste à déterminer si le score local réalisé est significativement élevé ou non.

Les résultats théoriques, exacts ou asymptotiques, sont tout d'abord présentés en considérant tout d'abord les séquences comme une suite de variables aléatoires, indépendantes et identiquement distribuées puis comme une chaîne de Markov. Par ailleurs, j'utilise la fonction de scores pour probabiliser l'ensemble de tous les segments possibles de la séquence. Sous des hypothèses usuelles, une dualité entre cet espace et celui des chaînes de Markov cachées, exempt de scores, est établie qui permet un transfert d'outils et de compétences, avec notamment le calcul de la probabilités conditionnellement à la séquence observée qu'un composant soit dans un segment atypique.

La deuxième partie du document présente des applications aux séquences biologiques mais également dans d'autres domaines tels que la maîtrise statistique des procédés. Des sujets de recherche sont ensuite proposés, portant sur la distribution du score des segments suboptimaux et la question de procédure de tests multiples en nombre aléatoire ou encore continu.

Local score distribution to highlight atypical segments in sequences

-

Sabine MERCIER

The local score is a classical tool used to detect an atypical segment inside a sequence. The local score quantifies the degree of local presence of a studied property in a sequence. One must first choose a scoring function f such as $\forall a \in \mathcal{A},\ x = f(a) \in \mathbb{R}$. The local score is then defined by $M_n = \max_{0 \leq i \leq j \leq n} \sum_{k=i}^{j} f(a_k)$ with $f(a_0) = 0$ and the question is to determine if the local score value is significant.

Theoretical results, exact and asymptotic ones, are gathered in the first part. We first consider sequences as random variable sequences, independent and identically distributed or Markovian sequences. The Hidden Markov Model is also considered and we propose a probabilization of the space of every possible sequence segment. Under usual hypothesis on the scoring function we establish a duality between this space and the one deduced from the HMM exempted from scores. This duality allows many transfers of tools and skills such as the posterior probabilities computation of a component to be in an atypical state.

The second part of the document presents applications on biological sequences but other domains are presented such as quality control and Statistical Process Control. Research projects are proposed in the third part, first dealing with the distribution of the score of suboptimal segments and the question of multiple test procedure, random or continuous number.