Détection de courts segments inversés dans les génomes - méthodes et applications

David Robelin

UNIVERSITÉ DE PARIS XI
FACULTÉ DE MÉDECINE DE PARIS-SUD

2005

THÈSE

pour obtenir le grade de

DOCTEUR DE L'UNIVERSITÉ PARIS XI

Spécialité : Biostatistique

présentée et soutenue publiquement

par

M. David ROBELIN

le 27 septembre 2005

Titre :

Détection de courts segments inversés dans les génomes : méthodes et applications

Directeur de Thèse : M. Bernard PRUM

JURY

M. Jean MACCARIO	Président
M. Dominique CELLIER	Examinateur
M. Christian GAUTIER	Rapporteur
Mme Chantal GUILHEUNNEUC	Examinateur
M. Bernard PRUM	Directeur
M. Stéphane ROBIN	Rapporteur

Publications et conférences sur le contenu de cette thèse.

David Robelin, Hugues Richard, et Bernard Prum. SIC : a tool to detect short inverted segments in a biological sequence. *Nucl. Acids. Res.*, 31(13) :3669–3671, 2003.

Vincent Miele, Pierre-Yves Bourguignon, David Robelin, Gregory Nuel, et Hugues Richard. seq++ : analyzing biological sequences with a range of Markov-related models. *Bioinformatics*, 21(11) :2783–2784, 2005.

David Robelin et Bernard Prum. Detecting short inverted segments in a biological sequence. In *European Conference on Computationnal Biology*, pages 41–43, Paris, 2003.

Vincent Miele, David Robelin, Yves Bourguignon, Pierre, Gregory Nuel, et Richard Hugues. Seq++ : a c++ library for sequence storage, markov modelization and scoring analysis. International Congress of Bioinformatic. Cuba, 2004.

David Robelin et Pierre Etienne, Marie. Discriminer les inversions de courts segments dans une séquence biologique. In *XI Rencontres de la Société Francophone de Classification*, pages 291–294, Bordeaux, 2004.

Pierre-Yves Bourguignon et David Robelin. Modèle de markov parcimonieux. In *Journées Ouvertes en Biologie, Informatique et Mathématiques*. Montréal, 2004.

Remerciements

Je tiens à remercier les deux rapporteurs de cette thèse, Christian Gautier et Stéphane Robin, pour avoir accepté de lire ce travail et suggéré quelques améliorations.

Merci également aux autres membres du jury (Jean Maccario, Dominique Cellier et Chantal Guihenneuc) pour s'être rendu disponible et apporter leurs contributions aux prolongements de ce travail.

Je salue également Daniel Goldstein, qui a proposé cette problématique dans son versant biologique.

Un remerciement particulier à Bernard Prum, pour de multiples raisons qui ne se résument que grossièrement en quelques mots. Outre le fait qu'il m'a permis d'effectuer cette thèse dans son laboratoire, et mis à disposition toutes les ressources nécessaires (presque sans compter), il s'est montré disponible à tout moment. Merci également pour l'ambiance plus qu'agréable du laboratoire.

Merci aux membres du groupe "Statistique des Séquences Biologiques" (animé par Sophie Schbath) pour leurs exposés enrichissants.

Merci à Vincent Miele et Mark Hoebeke pour m'avoir aidé à concrétiser une partie de cette thèse, et pour leurs apports modérés mais réguliers d'ethanol et autres cochonneries sucrées.

Merci à Adeline pour avoir rédigé cette thèse.

Merci à Maurice, véritable héros discret des temps modernes, dernier rempart contre toutes les agressions binaires, qui sont légions proche du centre commercial d'Evry.

Merci à Hugues pour le temps qu'il n'hésite pas à prendre pour expliquer, même si on ne comprend pas tout ce qu'il dit.

N'oublions pas de remercier Catherine Matias, car elle a relu une partie de ce document, et car c'est plus prudent.

Adeline, Ana, Sophie, Vincent, Florence : merci pour votre soutien en fin de rédaction ; ça a beaucoup compté.

Merci à Simona, aussi pour m'avoir aidé à en rédiger une partie.

Sans oublier Marie-Pierre sans qui je ne serais jamais devenu fumeur passif, mais un simple citoyen du bureau Nord.

Merci à l'ensemble des membres du laboratoire "Statistique et Génome", passés et présents, pour l'ensemble, très agréable.

Table des matières

Chapitre 1

Introduction générale

1.1 Introduction de la problématique biologique

1.1.1 Introduction à la génomique

Le patrimoine génétique des organismes se trouve au coeur de leur(s) cellule(s) sous la forme d'une longue molécule appelée ADN (*Acide Désoxyribo-Nucléique*). Cette molécule est composée de deux brins, chacun étant une séquence linéaire et orientée de quatre **nucléotides** (ou **bases**) : adénine (a), cytosine (c), guanine (g) et thymine (t). L'ADN peut se représenter comme une longue "phrase" composée de lettres issues d'un alphabet à quatre lettres $\mathcal{A} = \{a, c, g, t\}$. Pour des raisons chimiques, l'extrémité initiale est désignée par 5' et la finale par 3'.

Les nucléotides possèdent des affinités chimiques entre eux ; on distingue les liaisons fortes (a-t) des liaisons faibles (c-g). Cette propriété fait qu'une molécule d'ADN s'associe presque toujours avec sa molécule complémentaire dont le sens est inversé (voir la figure suivante 1.1.1). Cette molécule prend ensuite la forme d'une double hélice et se replie de multiples fois.

$$\begin{array}{cccccccccc} 5' & a & c & c & t & g & a & c & 3' \\ & | & | & | & | & | & | & | & \\ 3' & t & g & g & a & c & t & g & 5' \end{array}$$

L'ADN donne naissance par une série de processus aux protéines qui constituent la base de tout organisme. Sa "lecture" et son interprétation représentent donc un enjeu majeur en biologie.

En laboratoire, on sait **séquencer** des organismes, c'est à dire trouver la séquence de nucléotides qui constitue son ADN. Le tableau 1.1.1 donne quelques exemples d'organismes séquencés.

La longueur importante des séquences rend l'étude expérimentale d'un génome entier

organisme	date	taille (Mb)	description
Haemophilus influenzae	05/95	1.8	bacille infectieux
Mycoplasma genitalium	10/95	0.6	parasite des voies génitales
Escherichia coli	09/97	4.6	bacille modèle
Homo sapiens	06/00	3100	l'homme
Arabidopsis thaliana	12/00	120	plante modèle

TAB. 1.1 – Exemples d'organismes séquencés - les tailles sont données en millions de bases (Mb)

difficile. En utilisant des modèles probabilistes, la statistique permet de dégager des caractéristiques de la succession des lettres de la séquence, pouvant être d'intérêt biologique.

1.1.2 Évolution

L'information génétique se transmet de génération en génération par recopie de l'ADN dans les nouveaux organismes. Pour les organismes sexués, le patrimoine génétique du nouvel individu provient pour moitié de la mère et pour moitié du père. Pour les organismes asexués, l'ensemble du génome de l'organisme parent est transmis au nouvel individu.

Le génome ne reste cependant pas stable au cours du temps. Il subit des transformations au cours de la transmission du patrimoine génétique, mais aussi au cours de la vie de la cellule. Ces modifications de l'ADN sont susceptibles d'entraîner la désactivation ou l'activation de la production d'une protéine, de modifier une protéine synthétisée... Elles peuvent donc impliquer des modifications importante du métabolisme de la cellule concernée. Ces transformations peuvent être très diverses, mais nous nous intéressons classiquement à trois opérations élémentaires :

- Les **substitutions** : qui se produisent lorsqu'un nucléotide est remplacé par un autre.
- Les **insertions** de nouveaux nucléotides dans la séquence d'ADN.
- Les **délétions** de nucléotides de la chaîne

La séquence ancestrale n'étant pas connue, il est impossible de distinguer une insertion dans un des brins, d'une délétion dans l'autre. On regroupe généralement ces deux opérations sous le terme **indel**.

Comme Ochman et al. (2000); Jain et al. (2002) l'ont remarqué, ces trois simples opérations ne suffisent pas à décrire toutes les modifications de l'ADN.

A la fin des années 1980, J. Palmer et coll. ont comparé les génomes mitochondriaux de *Brassica oleracea* (chou) et *Brassica campestris* (navet). Ces deux espèces sont très proches (99% de leurs gènes sont identiques), mais l'ordre des gènes sur leurs séquences respectives diffère. D'autres études ont montré que ce type d'évolution moléculaire, appelé **remaniement global** est fréquent.

Le remaniement global se produit lors de la reproduction. A cette occasion, les deux brins complémentaires d'ADN se séparent pour former chacun un nouveau brin complémentaire. Par ce processus, le matériel génétique est doublé et deux nouvelles identités (cellules...) vont pouvoir être créées. Il arrive au brin esseulé d'ADN de se "casser" et de subir l'insertion d'un nouveau segment d'ADN. Ce segment peut provenir du matériel génétique du même individu ou d'un autre organisme tel qu'un virus par exemple. Dans le premier cas, le segment peut être issu d'un autre endroit sur le même brin, ou sur le brin complémentaire, un **retournement** accompagnant alors le déplacement. En effet, les deux brins d'ADN étant de sens opposés, il est nécessaire au segment issu du brin complémentaire de se retourner pour pouvoir être inséré. On sait prendre en compte ces remaniements globaux en supposant connus les segments élémentaires, les gènes par exemple (voir Pevzner (2000)).

Ces phénomènes ne se limitent pas aux segments d'ADN de relativement grande taille tels que les gènes, mais se produisent également sur des segments de quelques bases. La phase de réplication* de l'ADN, par exemple, peut provoquer l'apparition d'inverses complémentaires (Gordon and Halliday, 1995). La figure 1.1, extraite d'un rapport technique illustre ce processus (Richard, 2002). La phase de **réplication** peut se schématiser de la façon suivante. Les deux brins d'ADN se séparent et forment la **fourche de réplication**. Chaque brin est ensuite "copié" indépendamment. Un brin d'ADN ne peut se synthétiser que dans le sens $5' \rightarrow 3'$. Par conséquent, un des deux brins sera synthétisé en continu par la **polymérase III**, alors que le deuxième brin est construit par morceau (d'environ 100 bases chez les eukaryotes, et 1000 à 2000 bases chez les prokaryotes), les morceaux étant ensuite assemblés. Il peut arriver que la polymérase synthétisant le brin $5' - 3'$ "change" de brin (car la synthèse est plus rapide par exemple) et recopie un morceau de l'autre brin avant de revenir au brin initial. Ce brin contiendra ainsi un segment retourné.

Goldstein et al. (2000) souligne le fait que ces modifications du génomes à petite échelle pourraient être un mécanisme majeur de l'évolution. Il est connu, par exemple, qu'une répétition de deux motifs, inverse complémentaire l'un de l'autre, peut initier la formation d'un grand palindrome, lui même associé à l'amplification génique (Tanaka et al., 2002). A une plus grande échelle, Lefebvre et al. (2003b); Hannenhalli et al. (1995) ont étudié l'ordre respectif des gènes présents dans deux génomes proches ; ils ont conclu à une sur-représentation significative de l'inversion d'un seul gène, par rapport à l'inversion de plusieurs gènes. Cela peut être la conséquence d'un phénomène plus général d'inversion sur l'ADN, qui n'est pas nécessairement limité au gènes.

Dans Goldstein et al. (2000, 2003), les auteurs s'intéressent à la conséquence sur les protéines de ces retournements de courtes séquences d'ADN. Dans le cas particulier où l'on suppose que la "phase" est conservée, c'est à dire qu'un codon, après retournement, est encore lu comme un codon, ils définissent des acides "aminés complémentaires" (par

*création d'un nouveau double brin d'ADN identique, sauf erreur, au double brin initial

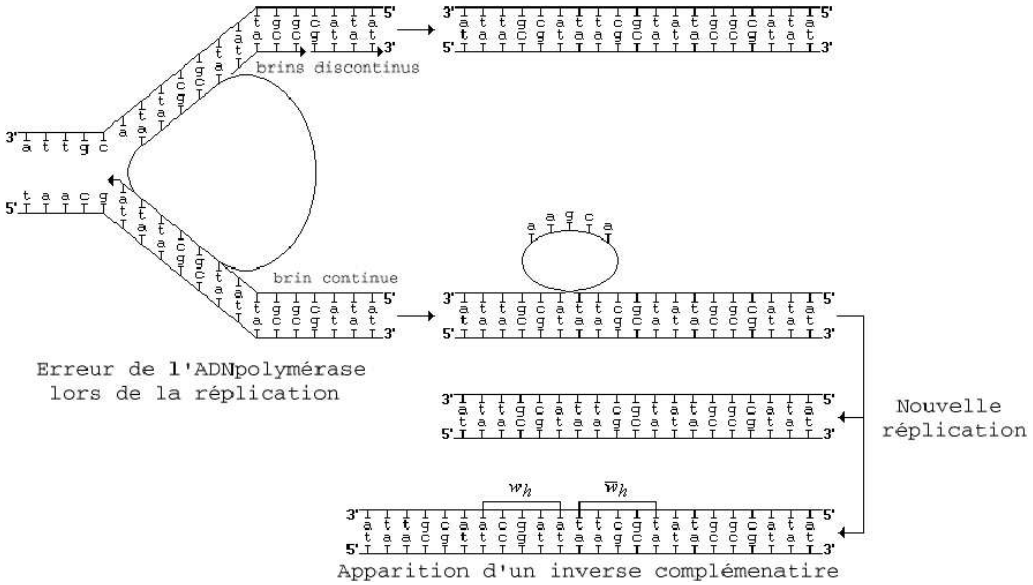


FIG. 1.1 – Erreur de réplication à l'origine de l'apparition d'un inverse complémentaire de petite taille.

exemple, la cystéine, codé tgt ou tgc a ses inverses complémentaires codés par aca, la threonine et gca, l'alanine). Il est montré que dans certaines classes de protéine, il y a davantage de mots (3 à 7 acides aminés) qui sont inverses complémentaires l'un de l'autre que ne le voudrait le hasard. Golstein et coll. y voient la trace de retournements de mots courts dans les génomes au cours de l'évolution. Nous reprenons la terminologie employée par Goldstein et al. pour désigner par le terme **Dincom** (pour **D**NA **in**verse **com**plementary) un court segment d'ADN inversé.

On doit insister ici pour souligner d'un Dincom est différent d'une répétition inversée. En effet, le motif inverse du Dincom n'existe pas nécessairement dans la séquence considérée. La détection des répetitions inversées est largement étudiée dans la littérature et des outils sont déjà disponibles (RepeatFinder de Volfovsky et al. (2001), Tandem repeats finder de Benson (1999), FORRepeats de Lefebvre et al. (2003a), REPuter de Kurtz et al. (2001) ou PatternHunter de Ma et al. (2002) par exemple).

L'objet de cette thèse est de mettre au point et d'éprouver une méthodologie permettant détecter la présence de Dincom dans une séquence d'ADN donnée. Une étude de la présence de dincom dans quelques génomes viraux et bactériens est présentée dans un deuxième temps.

1.2 Presentation de l'exposé

Ce document constitue le corps d'une thèse de mathématiques appliquées à la génomique (dans le domaine de l'analyse de séquence). Ce travail s'inscrit dans la problématique du laboratoire "Statistique et Génome" (département mathématiques de Université d'Evry Val d'Essonne), dont la structure s'appuie sur trois disciplines : les Mathématiques, la Biologie et l'Informatique. Les mathématiques, champ principal du laboratoire, permettent de proposer et d'ajuster des modèles pertinents à la biologie, de mettre en évidence des résultats asymptotique, de calculer des vitesses de convergence, par exemple. Les dérivées de ses résultats trouveront des applications concrètes. L'informatique permet d'évaluer et d'améliorer la mise en oeuvre concrête de méthodes et modèles. Les simulations sont également utiles pour prendre en compte une complexité suffisament "réaliste" dans l'évaluation des différents modèles. La biologie enfin, permet d'éviter les modèlisations trop grotesques (et par conséquent non informative), d'interpréter les modèles mathématiques et leur qualité d'ajustement, mais également de définir le champs d'applications de ces modèles et de choisir l'information pertinente à une problématique donnée.

Ma formation étant principalement mathématique, cette thèse est principalement orientée sur une explicitation détaillé de la méthodologie statistique employée. Néanmoins, une intention particulière a été portée pour rendre son exposé aussi didactique que possible. Sa lecture se veut à la portée de tous les bioinformaticiens, qu'ils soient plutôt biologistes, informaticiens ou mathématiciens ; quelques détails techniques pourront être laissés de coté selon la nature du lecteur. Une attention particulière a été apportée à la faisabilité des méthodes, et un algorithme accompagne généralement les résultats mathématiques. Les résultats biologiques sont principalement présentées à titre d'exemple. En effet, l'aspect biologique de la problématique des inversions de courts segments d'ADN, présentée dans cette thèse, tirerait profit d'un regard de biologiste (spécialiste), notamment pour la finesse de l'analyse et le choix des séquences à traiter. Conscient des limites de mes compétences biologiques, j'ai dévellopé le logiciel SIC, qui met à disposition les méthodes présentées dans cette thèse. Ce logiciel est disponible sous licence GPL, ou via une interface web Robelin et al. (2003).

1.2.1 Plan de la thèse :

Voici, ci-dessous les différents points abordés dans la suite de l'exposé.

- Modèlisation markovienne d'une séquence biologique : (espace d'état discret) :
  - Revue des modèles markoviens utiles en analyse de séquence (avec leurs principales propriétés)
  - Introduction des modèles récents (développés au sein du laboratoire "Statistique et Génome" : modèle de Markov parcimonieux, modèle de mélange de transition,

et chaîne de Markov dérivante)
  - Généralisation au cas d'un ordre supérieur à 1 d'un résultat connu concernant la vitesse de convergence d'une chaîne de Markov vers sa distribution stationnaire (Rosenthal, 1995).
  - Exposé d'un résultat central dans cette thèse sur la chaîne de Markov "inversée".
- Le score local (cas d'une seule séquence, ou d'alignement sans gap)
  - Revue des principaux résultats probabilistes et méthodes de détermination du plus grand score local
  - Etude des r plus grandes valeurs de score local : algorithme de détermination, résultats probabilistes, démarche de tests multiples, sans ou avec détermination de r.
- Méthodologie de recherche de segments inversés dans une séquence markovienne - cas de segments de longueurs connues, et inconnues. Attention : il ne s'agit pas de trouver un couple de segments tel que ces deux segments soient l'inverse l'un de l'autre, mais bien de trouver, à l'aide d'un critère statistique, un segment qui se soit "retourné" dans une chaine de markov.
- Application à quelques séquences de génomes viraux.

Bibliographie

Benson, G. (1999). Tandem repeats finder : a program to analyze DNA sequences. *Nucl. Acids. Res.*, 27(2) :573–580.

Goldstein, D., Muri, F., Saragueta, P., and Prum, B. (2000). Inverse complementary homologues of short cysteine signatures. *C R Acad Sci III*, 323 :167–172.

Goldstein, D. J., Fondrat, C., Muri, F., Nuel, G., Saragueta, P., Tocquet, A.-S., and Prum, B. (2003). Short inverse complementary amino acid sequences generate protein complexity. *Comptes Rendus Biologies*, 326(3) :339–348.

Gordon, A. and Halliday, J. (1995). Inversions with deletions and duplications. *Genetics J*, 140 :411–4.

Hannenhalli, S., Chappey, C., Koonin, E. V., and Pevzner, P. A. (1995). Genome Sequence Comparison and Scenarios for Gene Rearrangements : A Test Case. *Genomics*, 30(2) :299–311.

Jain, R., Rivera, M. C., Moore, J. E., and Lake, J. A. (2002). Horizontal Gene Transfer in Microbial Genome Evolution. *Theoretical Population Biology*, 61(4) :489–495.

Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C. h., Stoye, J., and Giegerich, R. (2001). REPuter : the manifold applications of repeat analysis on a genomic scale. *Nucl. Acids. Res.*, 29(22) :4633–4642.

Lefebvre, A., Lecroq, T., Dauchel, H., and Alexandre, J. (2003a). FORRepeats : detects repeats on entire chromosomes and between genomes. *Bioinformatics*, 19(3) :319–326.

Lefebvre, J., El-Mabrouk, N., Tillier, E., and Sankoff, D. (2003b). Detection and validation of single gene inversions. *Bioinformatics*, 19(90001) :190i–196.

Ma, B., Tromp, J., and Li, M. (2002). PatternHunter : faster and more sensitive homology search. *Bioinformatics*, 18(3) :440–445.

Ochman, H., Lawrence, J. G., and A, G. E. (2000). Lateral gene transfer and the nature of bacterial innovation. *Nature*, 405 :299–304.

Pevzner, P. (2000). *Computational Molecular Biology : An Algorithmic Approach*. MIT Press.

Richard, A. (2002). Etude de la corrélation entre le nombre d'occurences de mots et le nombre d'occurences de leurs inverses complémentaires dans les séquences d'adn. Technical report, IUP Génie Biologique et Informatique, Université d'Evry Val d'Essonne.

Robelin, D., Richard, H., and Prum, B. (2003). SIC : a tool to detect short inverted segments in a biological sequence. *Nucl. Acids. Res.*, 31(13) :3669–3671.

Rosenthal, Jeffrey, S. (1995). Convergence rates of markov chains. *SIAM review*, 37 :387–405.

Tanaka, H., Tapscott, S. J., Trask, B. J., and Yao, M.-C. (2002). Short inverted repeats initiate gene amplification through the formation of a large DNA palindrome in mammalian cells. *PNAS*, 99(13) :8772–8777.

Volfovsky, N., Haas, B., and Salzberg, S. (2001). A clustering method for repeat analysis in dna sequences. *Genome Biology*, 2(8) :research0027.1–research0027.11.

Chapitre 2

Modélisation Markovienne d'une séquence biologique

2.1 Introduction

Ce chapitre répertorie les différentes modélisations de séquences biologiques effectuées à l'aide des chaînes de Markov que l'on trouve classiquement dans la littérature. Elles sont utilisées avec divers objectifs en analyse de séquence ; citons, par exemple, la détection de motifs dont la fréquence est anormalement élevée dans une séquence d'ADN, ou la recherche d'une segmentation en zone homogène afin d'identifier des éléments fonctionnels tels que les gènes par exemple. En outre, la plupart des notations utilisées dans ce document est définie à cette occasion. Les principales propriétés des chaînes de Markov que nous utiliserons dans le développement de cette thèse sont également rappelées ici. Un intérêt particulier est montré pour la distribution du nombre d'apparition d'un certain "mot" dans une chaîne de Markov.

La début de la première partie, rappel des différentes modélisations markoviennes, est inspiré de la thèse de Grégory Nuel (2001).

2.2 Notation

La terminologie employée est inspirée de la biologie, mais aussi de l'analyse textuelle.

$\mathcal{A}$ désigne un **alphabet** de cardinal $|\mathcal{A}| = k$ dont on peut supposer, sans perte de généralité, qu'il est égal à $0, ..., k-1$.

Une séquence observée de longueur n sera notée $x = x_1, ..., x_n$ avec $\forall i \in 1, ..., n, x_i \in \mathcal{A}$. La séquence aléatoire à valeur dans $\mathcal{A}^n$ qui lui est associée est désignée par $X = X_1, ..., X_n$.

$W = w_1...w_h$ avec $\forall i \in \{1, ..., h\}, w_i \in \mathcal{A}$ désigne un **mot** de longueur h ou un **h-mot**.

La variable aléatoire associée au nombre d'occurrences du mot W dans la séquence X est noté N^W :

$$N^W = \sum_{i=1}^{n-h+1} I_i^W$$

où $I_i^W = \mathrm{I}_{\{X_i=w_i\}} \times ... \times \mathrm{I}_{\{X_{i+h}=w_{i+h}\}}$ est l'indicatrice de la présence du mot W (commençant) à la position i dans X.

On s'intéressera également au nombre d'apparitions du mot W dans la sous-séquence $X_i...X_{i+l-1}$ de longueur $l \geqslant h$ que l'on notera $N_{i,l}^W$.

2.3 Les différents modèles Markoviens

Nous présentons ici les modèles markoviens classiquement utilisés en analyse de séquences. Pour chacun d'eux, on mettra en évidence sa relation au modèle markovien d'ordre 1 (défini ci-dessous). Le nombre de paramètres linéairement indépendants de chacun de ces modèles est également précisé.

2.3.1 Modèles "simples"

La séquence d'ADN peut se représenter comme une longue phrase composée de lettres issues d'un alphabet à quatre lettres. Un modèle simple ne consiste qu'à s'intéresser à la probabilité d'apparition de chacune des lettres ; dans ce modèle, cette distribution est la même tout le long de la séquence, et la valeur prise à une position est indépendante des autres lettres de la séquence. Chaque lettre de la séquence est tirée indépendamment des autres selon une loi identique.

Définition 2.1 (Modèle $M0$)
$(X_i)_{i\in\{1,...,n\}}$ *est un échantillon de taille* n *de loi* μ *(**i.e.** les* n *variables aléatoires sont indépendantes et identiquement distribuées - **i.i.d.** - selon* μ*).*

Nombre de paramètres linéairement indépendants : $|\mathcal{A}| - 1$ (car la somme des probabilités de toutes les lettres est égale à 1).

Probabilité d'apparition d'un h-mot W à une position donnée

$$\mathbb{P}(I_i^W = 1) = \mu(w_1) \times ... \times \mu(w_h)$$

Ce modèle, très simple et néanmoins efficace face à certains problèmes, n'est évidemment pas réaliste. Une amélioration possible de ce modèle consiste à supposer que les lettres voisines ont une influence sur la loi d'apparition de la lettre considérée. C'est le

cas dans une chaîne de Markov. La dépendance, mesurée par exemple par un critére de mélange (de type α ou β mélange), des lettres situées à deux positions espacées de d lettres décroît alors exponentiellement avec d. Il s'agît donc d'un modèle où la dépendance est locale, et qui ne prendra pas en compte d'éventuelles corrélations à longue portée. Dans ce modèle, la distribution de la lettre observée à une certaine position dépend directement de la lettre précédente et indirectement des autres lettres.

Définition 2.2 (Modèle $M1$)
*Si $(X_i)_{i\in\{1,\dots,n\}}$ est une chaîne de **Markov** d'ordre 1 : $\forall i \in \{1,\dots,n\}$ et $\forall x_1,\dots,x_i \in \mathcal{A}$, on a*

$$\begin{aligned}\mathbb{P}(X_i = x_i | X_1 = x_1, \dots, X_{i-1} = x_{i-1}) &= \mathbb{P}(X_i = x_i | X_{i-1} = x_{i-1}) \\ &= Q(x_{i-1}, x_i).\end{aligned}$$

*Q est une matrice stochastique que l'on nommera **matrice de transition** de la chaîne de **Markov**.*
*La distribution μ unique solution de l'équation $\mu^T Q = \mu^T$ si elle existe est appelée **distribution stationnaire**.*

Nombre de paramètres linéairement indépendants : $|\mathcal{A}|\times(|\mathcal{A}|-1)$ (car la somme de chaque ligne de la matrice est égale à 1.)

Probabilité d'apparition d'un h-mot W à une position donnée En régime stationnaire*, la probabilité que le mot W apparaisse à la position i pour $i = 1, ..., n-h+1$ vaut sous ce modèle :

$$\mathbb{P}(I_i^W = 1) = \mu(w_1) \times \prod_{j=2}^{h} \mathrm{Q}(w_{j-1}, w_j).$$

On généralise ce modèle en supposant que la distribution de la lettre observée à une certaine position dépend directement des m lettres précédentes et indirectement des autres lettres.

Définition 2.3 (Modèle Mm)
*$(X_i)_{i\in\{m,m+1,\dots,n\}}$ constitue une chaîne de **Markov** d'ordre m si, et seulement si, $\forall i \in \{1,\dots,n\}$ et $\forall x_1,\dots,x_i \in \mathcal{A}$, on a*

$$\begin{aligned}\mathbb{P}(X_i = x_i | X_1 = x_1, \dots, X_{i-1} = x_{i-1}) &= \mathbb{P}(X_i = x_i | X_{i-m} = x_{i-m}, \dots X_{i-1} = x_{i-1}) \\ &= Q(x_{i-m} \dots x_{i-1}, x_i).\end{aligned}$$

*La chaîne de Markov est dite en régime stationnaire lorsque la probabilité d'apparition de n'importe quel mot ne dépend pas de sa position dans la chaîne. Autrement dit, la chaîne a "oublié" son point de départ, ou X_0 est de distribution μ.

Si Q gère les transitions de la chaîne, si l'équation ci-dessous a une solution, on note μ la distribution sur $\mathcal{A}^m$ vérifiant :

$$\forall x_1, \ldots, x_m, x_{m+1} \in \mathcal{A}, \sum_{x_1 \in \mathcal{A}} \mu(x_1, \ldots, x_m) Q(x_1 \ldots x_{m+1}) = \mu(x_2, \ldots, x_{m+1}).$$

Nombre de paramètres linéairement indépendants : $|\mathcal{A}|^m \times (|\mathcal{A}| - 1)$

Probabilité d'apparition d'un h-mot W à une position donnée En régime stationnaire, la probabilité que le h-mot W apparaisse à la position i pour $i = 1, ..., n-h+1$ vaut sous le modèle Mm :

$$\mathbb{P}(I_i^W = 1) = \mu(w_1, ..., w_m) \times \prod_{j=m+1}^{h} \mathrm{Q}(w_{j-m}, ..., w_j).$$

Équivalence à une chaîne de Markov d'ordre 1 Une chaîne de Markov d'ordre m est équivalente à une chaîne de Markov d'ordre 1 définie sur un alphabet correctement choisi. Si on définit le m-uplet $Z_i = (X_i, ..., X_{i+m-1})$, alors la séquence $Z = Z_1, ..., Z_{n-m}$ est une chaîne de Markov d'ordre 1 dont la matrice de transition s'exprime en fonction de celle de (X). Cette propriété est illustrée sur l'exemple suivant. On considère une réalisation d'une chaîne de Markov d'ordre 2 sur l'alphabet $\mathcal{A} = \{a, c, g, t\}$:

$$x = agtgatgccccgt.$$

Alors, la séquence z d'ordre 1 lui correspondant s'écrit dans l'alphabet $\mathcal{A}_z = \mathcal{A}^2$ de la façon suivante :

$$z = (ag)(gt)(tg)(ga)(at)(tg)(gc)(cc)(cc)(cc)(cg)(gt).$$

Si on note Q_x la matrice de transition de X et Q_z celle de Z, on a alors :

$$\mathrm{Q}_z\left((u, v), (u', v')\right) = \begin{cases} \mathrm{Q}_x(uv, v') & \text{si } v = u' \\ 0 & \text{sinon} \end{cases}$$

Emboîtement de ces modèles Soient M et M' deux modèles, on pose alors la relation suivante : $M \subset M'$ qui signifie que M peut être vu comme un cas particulier de M'. On dit alors que M est **emboîté** dans M'.
On a alors :

$$M0 \subset M1 \subset \ldots \subset Mm \subset Mm+1 \subset \ldots$$

2.3.2 Modèles périodiques

Les parties de l'ADN qui sont utilisées pour coder les protéines sont à considérer par triplet de lettres (cf. table 2.1 ci-dessous). Il n'est pas réaliste de considérer que la loi d'apparition des lettres en première position d'un triplet est la même que celle en deuxième ou en troisième position. La table des acides aminés montre, par exemple, que le nombre d'acides aminés qu'il est possible de coder en connaissant la première lettre du codon, est beaucoup plus faible que ce même nombre si on connaît seulement la dernière lettre. En d'autres termes, la première lettre du codon est beaucoup plus déterminante quant au choix de l'acide aminé qui sera codé, que la dernière. Dans les parties codantes de l'ADN, on souhaitera par conséquent tenir compte de la **phase** de lecture, c'est à dire de la position modulo trois des lettres dans la séquence.

Première position	Deuxième position								Troisième position
	A		C		G		T		
A	AAA	Lys	ACA	Thr	AGA	Arg	ATA	Ile	A
	AAC	Asn	ACC	Thr	AGC	Ser	ATC	Ile	C
	AAG	Lys	ACG	Thr	AGG	Arg	ATG	Met	G
	AAT	Asn	ACT	Thr	AGT	Ser	ATT	Ile	T
C	CAA	Gln	CCA	Pro	CGA	Arg	CTA	Leu	A
	CAC	His	CCC	Pro	CGC	Arg	CTC	Leu	C
	CAG	Gln	CCG	Pro	CGG	Arg	CTG	Leu	G
	CAT	His	CCT	Pro	CGT	Arg	CTT	Leu	T
G	GAA	Glu	GCA	Ala	GGA	Gly	GTA	Val	A
	GAC	Asp	GCC	Ala	GGC	Gly	GTC	Val	C
	GAG	Glu	GCG	Ala	GGG	Gly	GTG	Val	G
	GAT	Asp	GCT	Ala	GGT	Gly	GTT	Val	T
T	TAA	Stop	TCA	Ser	TGA	Stop	TTA	Leu	A
	TAC	Tyr	TCC	Ser	TGC	Cys	TTC	Phe	C
	TAG	Stop	TCG	Ser	TGG	Trp	TTG	Leu	G
	TAT	Tyr	TCT	Ser	TGT	Cys	TTT	Phe	T

TAB. 2.1 – Tables des acides aminés

Définition 2.4 (Modèle Mm_3)
Le modèle Mm_3, selon la notation introduite par Schbath (1995), appelé modèle périodique de période 3, est défini ainsi :

$$\mathbb{P}(X_i = x_i | X_1 = x_1, ..., X_{i-1} = x_{i-1}) = Q_{\phi(i)}(x_{i-m}, ..., x_{i-1}, x_i)$$

avec

$$\phi(i) = \begin{cases} 1 & si \quad i = 1[3] \\ 2 & si \quad i = 2[3] \\ 3 & si \quad i = 0[3] \end{cases}$$

$\phi(i)$ *désigne la* **phase** *dans laquelle se trouve la* i^{me} *lettre de la séquence. Si elle existe, on note* μ_i *l'unique distribution sur* $\mathcal{A}^m$ *vérifiant* $\forall i \in \{1,2,3\}$,

$$\mu_i^T Q_{\phi(i+1)} Q_{\phi(i+2)} Q_{\phi(i)} = \mu_i^T$$

Remarque 2.5 *Les modèles périodiques n'admettent pas de distribution stationnaire, car les hypothèses nécessaires ne sont pas vérifiées. Néanmoins, les distributions* $\mu_i, i = 1,2,3$ *jouent le rôle de ces distributions pour chaque phase* i.

Lemme 2.6 *Dans un modèle périodique* Mm_3, *en conservant les notations introduites précédemment, si les distributions* μ_1, μ_2 *et* μ_3 *existent, alors les égalités suivantes sont vérifiées :*

$$\begin{aligned} \mu_1^T Q_2 &= \mu_2^T \\ \mu_2^T Q_3 &= \mu_3^T \\ \mu_3^T Q_1 &= \mu_1^T \end{aligned}$$

Preuve. On montre que ces trois égalités vérifient la définition de $\mu_i, i = 1,2,3$. On a $\mu_1^T \mathrm{Q}_2 \mathrm{Q}_3 \mathrm{Q}_1 = \mu_1^T$. D'où $\mu_2^T \mathrm{Q}_3 \mathrm{Q}_1 = \mu_1^T$, puis $\mu_3^T \mathrm{Q}_1 = \mu_1^T$. Il en est de même pour les deux autres distribution μ_2 et μ_3.

La distribution μ_i est l'unique solution de

$$\mu_i^T \mathrm{Q}_{\phi(i+1)} \mathrm{Q}_{\phi(i+2)} \mathrm{Q}_{\phi(i)} = \mu_i^T.$$

■

Remarque 2.7 *Il n'est pas toujours aisé de bien distinguer les notions de phase, de période ou d'ordre. La convention utilisée dans ce document est la suivante : lorsque l'on parle de mot en phase* i, *cela signifie que la* **dernière** *lettre du mot se trouve en phase* i.

Exemple : Dans l'exemple suivant, nous avons un modèle de période 3, et le mot formé par les lettres TAAGT (en gras) se trouve en phase 1.

Phase	1	2	3	1	2	3	1	2	3	1	2	3	1	2	3	1
Séquence	*A*	*T*	*G*	*A*	*G*	*G*	*T*	*A*	**T**	**A**	**A**	**G**	**T**	*A*	*A*	*A*

Nombre de paramètres linéairement indépendants : Un tel modèle possède trois fois plus de paramètres qu'un modèle Mm puisqu'une matrice de transition particulière est utilisée pour chaque phase. Soit : $3 \times |\mathcal{A}|^m \times (|\mathcal{A}| - 1)$

Probabilité d'apparition d'un h-mot W à une position donnée Cette probabilité dépend de la phase à laquelle commence le mot. En régime stationnaire, la probabilité que le h-mot W ($h > m$) apparaisse à la position i pour $i = 1, ..., n - h + 1$ vaut sous le modèle Mm_3 :

$$\mathbb{P}(I_i^W = 1) = \mu_{\phi(i+m-1)}(w_1, ..., w_m) \times \prod_{j=m+1}^{h} \mathrm{Q}_{\phi(i+j-1)}(w_{j-m}, ..., w_j)$$

Équivalence à une chaîne de Markov d'ordre 1 Notons qu'il est possible de ramener ce modèle au cas markovien simple en modifiant l'alphabet utilisé de la façon suivante. L'alphabet est "étendu" en supposant que chacune des lettres initiales sera distinguée en trois exemplaires selon la phase à laquelle elle apparaît. A titre d'exemple, considérons la séquence écrite dans l'alphabet $\mathcal{A} = \{a, c, g, t\}$:

$$x = atgctcgatctcggtcgagcgcaa.$$

Dans le nouvel alphabet $\mathcal{A}_3 = \{a, c, g, t\} \times \{1, 2, 3\}$, cette séquence se réécrit :

$$x = a_1t_2g_3c_1t_2c_3g_1a_2t_3c_1t_2c_3g_1g_2t_3c_1g_2a_3g_1c_2g_3c_1a_2a_3.$$

Le nombre de paramètres de cette chaîne de Markov n'est pas pour autant $|A_3|^m \times (|A_3| - 1)$ (il vaut $3 \times |\mathcal{A}|^m \times (|\mathcal{A}| - 1)$), car il y a beaucoup de probabilités de transition nulles comme par exemple (à l'ordre 1) : $\mathrm{Q}(a_2, a_1)$.

Finalement, la chaîne de Markov équivalente écrite dans le nouvel alphabet est de matrice de transition :

$$\mathrm{Q} = \begin{pmatrix} 0 & \mathrm{Q}_2 & 0 \\ 0 & 0 & \mathrm{Q}_3 \\ \mathrm{Q}_1 & 0 & 0 \end{pmatrix}$$

2.3.3 Chaînes de Markov cachées

Les longues séquences biologiques présentent généralement des segmentations fonctionnelles en différents types (codant, non codant, intron, exon, intergénique...). Les modèles de Markov cachés permettent de prendre en compte cette hétérogénéité en utilisant différents modèles de Markov pour chaque type de segment. Les transitions entre un type et un autre sont, elles aussi, gérées par un modèle de Markov.

Formellement, on considère la séquence $X = X_1, ..., X_n$ avec $X_i \in \mathcal{A}$ et la suite des états cachés associés $S = S_1, ..., S_n$ avec $S_i \in \{1, ..., s\}$. S_i qui désigne le type de segments, est dite "cachée" (ou variable latente), car elle n'est pas directement observée.

Définition 2.8 (Modèle $M1 - Mm$)
La séquence S est générée par une chaîne de Markov d'ordre 1 sur l'espace des états cachés $\{1, ..., s\}$ et X_i conditionnellement à S_i est généré selon une chaîne de Markov d'ordre m :

$$\begin{aligned}\mathbb{P}(X_i = x_i| \quad & X_1 = x_1, ..., X_{i-1} = x_{i-1}, S_1 = s_1, ..., S_{i-1} = s_{i-1}) \\ & = \mathbb{P}(X_i = x_i | X_{i-1} = x_{i-1}, ..., X_{i-m} = x_{i-m}, S_i = s_i) \\ & = Q_{s_i}(X_{i-1} = x_{i-m}, ..., x_{i-1}, x_i)\end{aligned}$$

Le graphe acyclique dirigé représenté sur la figure 2.1 illustre les dépendances d'une chaîne Markov cachée de type $M1 - Mm$.

$$\begin{array}{lccccccccc} \text{États cachés} & s_{i-2} & \rightarrow & s_{i-1} & \rightarrow & s_i & \rightarrow & s_{i+1} & \rightarrow & s_{i+2} \\ & \downarrow & & \downarrow & & \downarrow & & \downarrow & & \downarrow \\ \text{séquence observée} & x_{i-2} & \rightarrow & x_{i-1} & \rightarrow & x_i & \rightarrow & x_{i+1} & \rightarrow & x_{i+2} \end{array}$$

FIG. 2.1 – Graphe Acyclique Dirigé représentant un modèle de Markov caché de type $M1 - M1$

Le terme "chaîne de Markov cachée" provient de la séquence markovienne des états cachés. Il est important de noter que la séquence observée $X_1, ..., X_n$ (non conditionnellement à S) n'est pas markovienne (cf. le graphe acyclique dirigé, figure 2.1).

L'estimation d'un tel modèle est complexifiée par le fait que la séquence d'états cachés S n'est pas connue. Plusieurs méthodes d'estimation existent. On pourra se reporter à Baum and Petrie (1966), Rabiner (1989) et Muri (1997) pour plus de détails sur les méthodes d'estimation de tels modèles et leurs applications.

Nombre de paramètres linéairement indépendants : Un tel modèle possède une matrice de Markov d'ordre m pour chaque état caché, ce qui représente $s \times |\mathcal{A}|^m \times (|\mathcal{A}| - 1)$ paramètres. De plus, les états cachés sont régis par une chaîne de Markov d'ordre 1 ce qui ajoute $s \times (s - 1)$ paramètres. Finalement, le nombre total de paramètres s'élève à $s \times (s - 1) + s \times |\mathcal{A}|^m \times (|\mathcal{A}| - 1)$.

Probabilité d'apparition d'un h-mot W à une position donnée Le calcul de cette probabilité ne présente pas de difficulté théorique, mais n'est néanmoins pas immédiat. Il utilise une formulation récursive qui nécessiterait l'introduction de notations supplémentaires pour être présenté ici. La suite de ce document ne nécessite pas la connaissance de ce calcul ; les lecteurs intéressés sont invités à consulter les ouvrages de référence cités précédemment. Le calcul de cette probabilité se trouve généralement sous une rubrique intitulé "étape forward".

Équivalence à une chaîne de Markov d'ordre 1 Si les états cachés sont connus, on peut se ramener une fois encore au modèle de Markov simple par une transformation simple. La chaîne de Markov considérée aura comme espace d'état $\mathcal{A} \times \{1, ..., s\}$, c'est à dire que chacune des lettres sera distinguée selon l'état qui lui est associé. Par exemple, on considère la séquence suivante :

$$x = aatttgtgatgccgatgt$$

à laquelle est associée la suite d'états suivante :

$$s = 121431112314432214$$

Cette séquence se réécrit dans le nouvel alphabet de la façon suivante :

$$x = a_1a_2t_1t_4t_3g_1t_1g_1a_2t_3g_1c_4c_4g_3a_2t_2g_1t_4$$

2.3.4 Modèles de Markov à dépendance variable

Contrairement au modèles Mm, les modèles de Markov à dépendance variable ne supposent pas que le nombre de positions précédentes dont dépend directement la loi d'une position donnée est fixé à m. Dans ces modèles, ce nombre de positions dépend du **contexte** c'est à dire du mot qui précède la position considérée.

On représente de tels modèles à l'aide d'arbres de contexte (voir Willems et al. (1995) et Rissanen (1983) pour plus de détails). Dans l'exemple de la figure 2.2, on considère un modèle dans lequel les occurrences des lettres suivant un `a` ou `g` sont réglées par un modèle $M2$ tandis que celles qui suivent les lettres `c` ou `t` le sont par un modèle $M1$ pour un total final de 30 paramètres (au lieu des 48 paramètres du modèle $M2$ complet). Dans le cas de l'arbre représenté en figure 2.3, il n'est même plus possible de décrire le modèle résultant avec les modèles précédents. Ici, il ne reste plus que les 21 paramètres correspondant aux probabilités suivantes :

$$\begin{cases} \mathbb{P}(X_i = x | X_{i-2} = \mathtt{c}, X_{i-1} = \mathtt{a}) \\ \mathbb{P}(X_i = x | X_{i-2} \neq \mathtt{c}, X_{i-1} = \mathtt{a}) \\ \mathbb{P}(X_i = x | X_{i-1} = \mathtt{c}) \\ \mathbb{P}(X_i = x | X_{i-2} = \mathtt{a}, X_{i-1} = \mathtt{g}) \\ \mathbb{P}(X_i = x | X_{i-2} = \mathtt{g}, X_{i-1} = \mathtt{g}) \\ \mathbb{P}(X_i = x | X_{i-2} \neq \mathtt{a}, X_{i-2} \neq \mathtt{g}, X_{i-1} = \mathtt{g}\ ) \\ \mathbb{P}(X_i = x | X_{i-1} = \mathtt{t}) \end{cases}$$

avec $x \in \mathcal{A}$.

En fait, on utilise ces arbres pour connaître le contexte pertinent en une position dans la séquence. Pour cela, on regarde d'abord la lettre précédant la position étudiée : si cette lettre ne possède pas de branche dans l'arbre menant à un nouveau nœud depuis

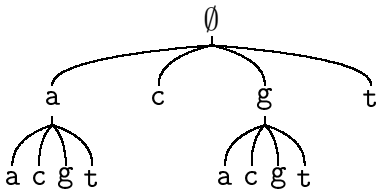


FIG. 2.2 – Exemple d'arbre de contexte complet dans l'alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$.

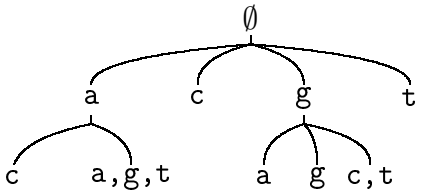


FIG. 2.3 – Exemple d'arbre de contexte incomplet dans l'alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$.

la racine $\emptyset$ alors on s'arrête ; sinon, on réitère le procédé en traitant le nœud comme une nouvelle racine. Le dernier nœud désigné par cet algorithme constitue le contexte pertinent recherché ; c'est à dire les éléments du "passé" qu'il faut considérer pour connaître la distribution d'émission du caractère à la nouvelle position.

Nombre de paramètres linéairement indépendants : Le nombre de paramètres dépend du nombre de contextes qui conduiront à différentes probabilités d'émission de la lettre suivante. Ce nombre peut donc facilement se déduire de l'arbre des contextes, puisqu'il s'agit du nombre de feuilles de l'arbre que multiplie la taille de l'alphabet-1.

2.3.5 Modèles de Markov Parcimonieux

L'utilisation des modèles de Markov à dépendance variable a été essentiellement motivée par un problème de parcimonie dans les modèles de Markov. En effet, on a vu que le nombre de paramètres d'un modèle de Markov augmente exponentiellement avec son ordre m. Néanmoins, la modélisation d'une séquence réelle nécessite généralement d'utiliser un ordre supérieur à 1 afin d'obtenir une adéquation suffisante. L'utilisation des modèles de Markov à dépendance variable permet d'adapter la taille de la mémoire au contexte considéré, et ainsi d'économiser des paramètres par rapport au modèle dit complet.

Le modèle de Markov parcimonieux généralise ces modèles en autorisant que le contexte considéré soit un motif. Un exemple de motif est donné par $m = a[gc]tc$ qui correspond à l'ensemble des deux mots suivants $\{\mathtt{agtc},\ \mathtt{actc}\}$. Dans la figure 2.2, un modèle de Markov parcimonieux autoriserait que la loi d'émission de la lettre à la position t est la même si la lettre à la position $t-1$ est un a ou un c. Ceci se résume dans la figure 2.4. Ce

modèle contient 18 paramètres, au lieu de 48 dans le modèle $M2$ "complet", correspondant aux probabilités suivantes :

$$\begin{cases} \mathbb{P}(X_i = x | X_{i-1} = \texttt{a ou g}, X_{i-2} = \texttt{a}) \\ \mathbb{P}(X_i = x | X_{i-1} = \texttt{a ou g}, X_{i-2} = \texttt{g}) \\ \mathbb{P}(X_i = x | X_{i-1} = \texttt{a ou g}, X_{i-2} = \texttt{c}) \\ \mathbb{P}(X_i = x | X_{i-1} = \texttt{a ou g}, X_{i-2} = \texttt{t}) \\ \mathbb{P}(X_i = x | X_{i-1} = \texttt{c}) \\ \mathbb{P}(X_i = x | X_{i-1} = \texttt{t}) \end{cases}$$

avec $x \in \mathcal{A}$.

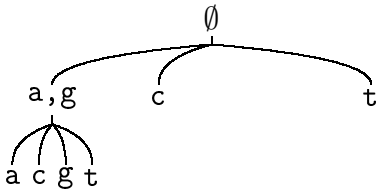


FIG. 2.4 – Exemple de représentation arborescente d'un modèle de Markov parcimonieux dans l'alphabet $\mathcal{A} = \{\texttt{a}, \texttt{c}, \texttt{g}, \texttt{t}\}$.

Ces modèles sont dit parcimonieux car ils permettent de réduire de façon importante le nombre de paramètres d'un modèle de Markov. Remarquons que cette diminution est d'autant plus importante que le "regroupement" a lieu en haut de l'arbre.

Dans ce type de problématique, le choix du modèle revêt une importance particulière, car le modèle choisi devra permettre un bon compromis entre parcimonie et qualité d'ajustement. L'inférence doit donc considérée le problème du choix du modèle parmi la famille des modèle parcimonieux. La taille extrêmement importante du nombre d'arbres de contexte possibles complexifie cette tâche. L'inférence bayésienne de ces modèles et leur application en génomique est le sujet de la thèse de Pierre-Yves Bourguignon*. Une méthode bayésienne d'estimation a été implémentée à l'aide d'un algorithme de programmation dynamique, et permet d'estimer ces modèles en un temps raisonnable dans le cas de l'alphabet nucléotidique (Bourguignon and Robelin, 2004).

Nombre de paramètres linéairement indépendants : Comme dans le cas des modèles de Markov à dépendance variable, le nombre de paramètres est égal au nombre de feuilles de l'arbre que multiplie la taille de l'alphabet-1.

Équivalence à une chaîne de Markov d'ordre m Les chaînes de Markov parcimonieuses d'ordre m sont un cas particulier des chaînes de Markov d'ordre m. En effet, il

*bourguignon@genopole.cnrs.fr

s'agit d'un modèle de Markov d'ordre m dont la matrice contient certaines lignes égales. Dans l'exemple ci-dessus, les lignes correspondant au contexte *aa* et *ga* seront égales, ainsi que les lignes correspondant au contexte *ca*,*cg*,*cc* et *ct* par exemple.

2.3.6 Modèles de mélange de transition Markovienne

Les modèles de Markov à dépendance variable, et les modèles de Markov parcimonieux, se fondent tout deux sur le contexte pour économiser des paramètres. De tels modèles seront particulièrement adaptés à des séquences fondées sur des sortes de syllabes. La taille de la mémoire correspond alors à la taille de la syllabe considérée. Les modèles de mélange de transition* ont également été motivés par un souci de parcimonie (Berchtold, 2001). Néanmoins, cette approche ne se fonde pas sur la notion de contexte, mais plutôt de position, comme la définition ci-dessous le précise.

Définition 2.9 (Modèle de mélange de distribution d'ordre m)
$(X_i)_{i=1,...,n}$ *constitue un modèle de mélange de distribution d'ordre* m *si, et seulement si,* $\forall i \in \{m, m+1, ..., n\}$ *et* $\forall x_1, ..., x_i \in \mathcal{A}$, *on a*

$$\begin{aligned}\mathbb{P}(X_i = x_i | X_1 = x_1, \ldots, X_{i-1} = x_{i-1}) &= \mathbb{P}(X_i = x_i | X_{i-m} = x_{i-m}, \ldots X_{i-1} = x_{i-1}) \\ &= \sum_{g=1}^{m} \phi_g \mathbb{P}(X_i = x_i | X_{i-g} = x_{i-g}) \\ &= \sum_{g=1}^{m} \phi_g \, Q_g(x_{i-g}, x_i)\end{aligned}$$

où $\forall g \in \{1, ..., m\}, \phi_g \geqslant 0$ *et* $\sum_{g=1}^{m} \phi_g = 1$

La loi d'apparition d'une lettre conditionnellement aux m lettres précédentes est un mélange de chaînes de Markov d'ordre 1 de poids $\phi = (\phi_1, ..., \phi_m)$, c'est à dire qu'elle dépend de la *ième* position précédente avec la probabilité ϕ_i et la loi de succession est donnée par la matrice de transition Q_i. Le paramètre ϕ_i reflète l'importance de cette position pour prédire la lettre successive.

Une méthode de maximisation de la vraisemblance par descente du gradient a été proposée pour estimer ce modèle dans le cas où les matrices de transition $\mathrm{Q}_g, g = 1, ..., m$ sont égales (Berchtold, 2001). Sophie Lèbre a mis au point un algorithme EM permettant d'estimer de tel modèle pour des matrices de transition quelconques (Lèbre, 2004). Elle a aussi montré que la paramètrisation en (ϕ_g, Q_g) de ces modèles n'est pas identifiable et a proposé une reparamétrisation. Des travaux effectués sont en cours pour estimer ces modèles de manière bayésienne, dans le cas où les poids peuvent prendre des valeurs nulles, où les matrices de transitions peuvent être égale et où l'ordre m n'est pas nécessairement connu†.

*"Mixture Transition Distribution"(MTD) in english

†me contacter si vous êtes intéressés

Nombre de paramètres linéairement indépendants Le modèle comporte m matrices de Markov d'ordre 1, ainsi que $m-1$ paramètres pour le vecteur ϕ, soit un total de $m \times |\mathcal{A}| \times (|\mathcal{A}|-1) + m - 1$

Équivalence à une chaîne de Markov d'ordre m La définition des modèles de mélanges de distribution d'ordre m montre leur inclusion dans le modèle de Markov d'ordre m.

2.3.7 Modèles de chaîne de Markov dérivantes

Ce modèle s'emploie à lever l'hypothèse de stationnarité le long de la séquence en faisant varier la matrice de transition en fonction de la position dans la séquence. A chaque position dans la séquence, nous aurons une nouvelle matrice de transition : $\mathrm{Q}_{\phi(t)}$ où ϕ est une fonction linéaire ou polynomiale de t. Nous nous donnons une matrice de départ Q_0 et une matrice d'arrivée Q_1 puis nous passons de l'une l'autre en fonction de la position t dans la séquence. Dans le cas où ϕ est linéaire, le modèle est donné par :

$$\mathrm{Q}_{\frac{t}{n}} = \left(1 - \frac{t}{n}\right) \mathrm{Q}_0 + \frac{t}{n} \mathrm{Q}_1$$

(où n est ici la taille de la séquence).

L'étude et l'inférence de tel modèle est l'objet de la thèse en cours de Nicolas Vergne*.

2.4 Quelques propriétés des chaînes de Markov

2.4.1 Estimation des paramètres

Un modèle de Markov Mm possède $(k-1)k^m$ paramètres. Ces paramètres ne sont généralement pas connus. On se propose de les estimer par maximum de vraisemblance.

Pour cela, on utilise le lemme suivant :

Lemme 2.10 *On considère la fonction*

$$\begin{array}{rccc} f: & \mathbb{R}^d & \to & \mathbb{R} \\ & (x_1, \ldots, x_d) & \mapsto & \sum_{i=1}^d c_i \log x_i \end{array}$$

où $(c_1, \ldots, c_d) \in (\mathbb{R}^+)^d$ *avec* $C = \sum_{i=1}^d c_i \neq 0$, *alors*

$$\inf_{x \in \Gamma} f(x) = f\left(\frac{c_1}{C}, \ldots, \frac{c_d}{C}\right).$$

où Γ *est le simplexe unité.*

*vergne@genopole.cnrs.fr

Preuve. La fonction f est continue sur le compact Γ et atteint donc son minimum sur Γ en un certain $a \in \Gamma$. On pose

$$\begin{array}{rccc} g: & \mathbb{R}^d & \to & \mathbb{R} \\ & (x_1, \ldots, x_d) & \mapsto & \sum_{i=1}^d x_i - 1 \end{array}$$

ce qui permet d'écrire $\Gamma = \{x \in (\mathbb{R}^+)^d, g(x) = 0\}$ et on sait alors, grâce au théorème des extrema liés, que $\exists \lambda \in \mathbb{R}$ tel que $df_a = \lambda dg_a$ c'est à dire tel que

$$\frac{\partial f}{\partial x_i}(a) = \lambda \frac{\partial g}{\partial x_i}(a) \quad \forall i \iff \frac{c_i}{a_i} = \lambda \quad \forall i$$

or comme $a = (a_1, \ldots, a_d) \in \Gamma$, il est clair que $\lambda = C$ ce qui donne le résultat recherché. ■

Sous le modèle Mm, la vraisemblance des observations s'écrit :

$$\begin{aligned} L &= \log(\mu_0(x_1, ..., x_m)) + \sum_{i=1}^{n-m} \log \mathrm{Q}(x_i, ..., x_{i+m+1}) \\ &= \log(\mu_0(x_1, ..., x_m)) + \sum_{x_1 ... x_{m+1} \in \mathcal{A}} n(x_1 ... x_{m+1}) \log \mathrm{Q}(x_1, ..., x_{m+1}) \end{aligned}$$

où μ_0 est la distribution des m premières observations
et $N(x_1...x_{m+1})$ le nombre d'occurence du mot $x_1...x_{m+1}$.

En utilisant le lemme 2.10, on obtient l'estimateur du maximum de vraisemblance suivant :

$$\hat{\mathrm{Q}}(x_1...x_{m+1}) = \frac{n(x_1...x_{m+1})}{\sum_{y \in \mathcal{A}} n(x_1...x_m y)}, \forall (x_1...x_{m+1}) \in \mathcal{A}^{m+1}$$

Finalement, l'estimation d'un modèle de Markov d'ordre m utilise les comptages des mots de tailles $m+1$, dont on déduit également ceux de taille m.
Noter que l'on suppose ici que tous les mots de tailles m apparaissent au moins une fois. Dans le cas contraire, on peut montrer que l'estimateur du maximum de vraisemblance donne la valeur 0 au paramètre correspondant.

Remarque :

- L'estimateur du maximum de vraisemblance obtenu est valable même si la chaîne de Markov n'est pas à l'état stationnaire.
- Il est courant de supposer que la chaîne de Markov est à l'état stationnaire, c'est à dire que la distribution des m premières lettres μ_0 est la distribution stationnaire de la chaîne. Dans ce cas, l'estimateur du maximum de vraisemblance doit prendre en compte le fait que le terme $\mu(x_1, ..., x_m)$ est une fonction du paramètre Q. La contrainte sur μ est difficile à prendre en compte lors de la maximisation

de la vraisemblance. Néanmoins, lorsque la séquence est suffisamment longue, le terme $\log(\mu(x_1, ..., x_m))$ devient négligeable devant la deuxième partie de la log-vraisemblance.

2.4.2 Détermination de la distribution stationnaire d'une chaîne de Markov d'ordre m

Lorsqu'elle existe, la distribution stationnaire μ d'une chaîne de Markov est définie par :
$\forall x_1, \ldots, x_m, x_{m+1} \in \mathcal{A}$,

$$\sum_{x_1 \in \mathcal{A}} \mu(x_1, \ldots, x_m) \mathrm{Q}(x_1 \ldots x_{m+1}) = \mu(x_2, \ldots, x_{m+1})$$

ou, de manière matricielle en utilisant le vecteur colonne μ contenant cette distribution :

$$\mu^T \mathrm{Q} = \mu^T$$

Cette distribution existe si, et seulement si, la chaîne est ergodique ce qui sera généralement le cas dans les applications génomiques.

La notation matricielle indique que le vecteur μ est le vecteur propre à droite de la matrice Q associé à la valeur propre 1. Il pourra donc être déterminé à l'aide d'une diagonalisation de matrice, ou à l'aide de la procédure d'Arnoldi qui permet d'extraire itérativement les valeurs et vecteurs propres d'une matrice (Lehoucq R et al., 1996). Cette dernière solution présente l'avantage d'être beaucoup moins coûteuse en temps de calcul.

La proposition suivante permet d'éviter cette étape de détermination de vecteur propre, dans le cas où la matrice de transition est estimée par maximum de vraisemblance.

Proposition 2.11 *Soit $X_1, ..., X_n$ une chaîne de Markov d'ordre m, et soit $\hat{Q}$ l'estimateur du maximum de vraisemblance de sa matrice de transition. Alors, la distribution stationnaire associée à la matrice $\hat{Q}$ est donnée par :*

$$\forall (x_1, ..., x_m) \in \mathcal{A}^m, \hat{\mu}(x_1, ..., x_m) = \frac{N(x_1 ... x_m)}{n - m + 1}$$

Preuve. Vérifions que $\hat{\mu}(x_1, ..., x_m) = \frac{N(x_1 ... x_m)}{n-m+1}$ et $\hat{\mu}(x_2, ..., x_{m+1}) = \frac{N(x_2 ... x_{m+1})}{n-m+1}$ sont solutions de l'équation pour tout $x_1 ... x_{m+1} \in \mathcal{A}^{m+1}$:

$$\sum_{x_1 \in \mathcal{A}} \hat{\mu}(x_1, \ldots, x_m) \hat{\mathrm{Q}}(x_1 \ldots x_{m+1}) = \hat{\mu}(x_2, \ldots, x_{m+1})$$

sous la contrainte $\sum_{x_1,...,x_m \in \mathcal{A}^m} \hat{\mu}(x_1,...,x_m) = 1$.
On a : $\hat{\mathrm{Q}}(x_1 \dots x_{m+1}) = \frac{N(x_1...x_{m+1})}{N(x_1...x_m)}$ donc

$$\begin{aligned}\sum_{x_1 \in \mathcal{A}} \hat{\mu}(x_1,\dots,x_m)\hat{\mathrm{Q}}(x_1 \dots x_{m+1}) &= \sum_{x_1 \in \mathcal{A}} \frac{N(x_1...x_m)}{n-m+1}\frac{N(x_1...x_{m+1})}{N(x_1...x_m)} \\ &= \frac{N(x_2...x_{m+1})}{n-m+1} \\ &= \hat{\mu}(x_2...x_{m+1})\end{aligned}$$

et on a bien

$$\begin{aligned}\sum_{x_1,...,x_m \in \mathcal{A}^m} \hat{\mu}(x_1,...,x_m) &= \sum_{x_1,...,x_m \in \mathcal{A}^m} N(x_1,...,x_m)/(n-m+1) \\ &= 1\end{aligned}$$

■

2.4.3 Vitesse de convergence vers la distribution stationnaire

Dans la plupart des problèmes traités, la complexité de la solution est généralement au moins linéaire avec la taille n de la chaîne. C'est le cas, par exemple, du calcul de la variance du nombre d'occurences d'un mot donné, dans une chaîne de Markov (cf. paragraphe 2.4.4 page 33). Lors de la modélisation génomique, les séquences considérées peuvent être de longueurs importantes (plusieurs dizaines de milliers à plusieurs centaines de milliers de lettres), ce qui entraîne des temps de calculs importants. Il est donc utile de chercher à diminuer cette complexité. Pour cela, on s'intéresse à la vitesse de convergence de la chaîne Markov vers sa distribution stationnaire. En effet, à partir d'un certain rang k, la matrice de transition Q^k pourra être approchée par la matrice stationnaire : $\mathrm{Q}^\infty = \lim_{n\to\infty} \mathrm{Q}^n$. Dans le cas d'une chaîne de Markov ergodique, la matrice Q^∞ est la matrice comportant la distribution stationnaire sur chacune de ses lignes. Le problème est donc de déterminer ce rang k en fonction d'une précision ϵ choisie a priori.

Cette question est traitée de manière très didactique et agréable dans Rosenthal (1995) ; la borne proposée a été améliorée dans différents travaux ; on verra par exemple Rosenthal (2002) ou Meyn and Tweedie (1994). Dans une chaîne de Markov irréductible et apériodique, on a la propriété suivante :

Proposition 2.12 *Si une chaîne de Markov $X_1, ..., X_n$ d'ordre 1 à espace d'états discret $\mathcal{A}$ fini est irréductible et apériodique et de matrice de transition Q, pour tout $n \in \mathbb{N}$:*

$$||\,Q^n - Q^\infty||_{vt} \leqslant C|\beta_1|^n$$

où β_1 *est la plus grande valeur propre en module distincte de 1 de la matrice de transition* Q
et $||\bullet||_{vt}$ *désigne la distance en variation totale*[*]
et $Q^\infty = \lim_{n\to\infty} Q^n$
et la constante positive $C = \frac{1}{2}\sum_{x\in\mathcal{A}} C_x$ *où*

$$C_x = \sum_{m=2}^{|\mathcal{A}|} |a_m v_m(x)|$$

avec $v_1, ..., v_{|\mathcal{A}|}$ *est une base de vecteurs propres à droite de la matrice* Q*, correspondant respectivement aux valeurs propres*
$beta_1 = 1$,
$beta_2, ..., \beta_{\mathcal{A}}$
et $(a_1, ..., a_{|\mathcal{A}|})$ *est la solution complexe unique de l'équation*

$$\mu_1 = a_1 v_1 + ... + a_{|\mathcal{A}|} v_{|\mathcal{A}|}$$

lorsque μ_1 *est la distribution de* X_1,

Ce résultat indique que la vitesse de convergence est exponentielle par rapport au module de la deuxième valeur propre de la matrice de transition.

En pratique, il présente deux inconvénients. Le premier consiste à devoir calculer la deuxième plus grande valeur propre en module. Même s'il existe des algorithmes efficaces (Lehoucq R et al., 1996), cette opération reste néanmoins coûteuse en calcul. Le deuxième problème est la détermination de la constante C ; ce point revêt moins d'importance car cette constante n'est influente que pour des valeurs de n faible.

Le résultat suivant, dû à Doob (1953), évite ces problèmes. Il se démontre de manière probabiliste en utilisant des techniques de couplages de variables aléatoires.

Proposition 2.13 *Soit une chaîne de Markov* $X = X_1, ..., X_n$ *d'ordre 1 à espace d'états* $\mathcal{A}$ *de matrice de transition* Q*.* X *est telle qu'il existe un réel* β *telle que* $Q(x, A) \geqslant \beta\zeta(A)$ *pour tout* $x \in \mathcal{A}$ *et pour tout ensemble mesurable de* $A \subseteq \mathcal{A}$ *et une distribution* ζ *sur* $\mathcal{A}$*. On a alors pour tout* $n \in \mathbb{N}$ *:*

$$||Q^n - Q^\infty||_{vt} \leqslant (1-\beta)^n$$

où $||\bullet||_{vt}$ *désigne la distance en variation totale*
et $Q^\infty = \lim_{n\to\infty} Q^n$ *la distribution stationnaire.*

[*]La distance en variation totale entre deux matrices de transition P et Q est donnée par $D(P,Q) = \frac{1}{2}\sum_{u,v\in\{a,c,g,t\}} |P(u,v) - Q(u,v)|$

La plus grande valeur de β satisfaisant les conditions de la proposition précédente est donnée par

$$\beta = \sum_{y \in \mathcal{A}} \min_{x \in \mathcal{A}} \mathrm{Q}(x, y)$$

C'est à dire que β est simplement la somme des minima de chaque colonne de la matrice de transition.

Un des problèmes de cette approche est que β est facilement nul en pratique, car il suffit qu'il y ait une probabilité de transition nulle dans chaque colonne de la matrice. Doob (1953) suggère dans ce cas de considérer la convergence de Q^{n_0}, avec n_0 choisi de façon à ce que tous les éléments de Q^{n_0} soit strictement positif, et de remplacer la valeur de n de la proposition par n/n_0.

Exemple : On considère la matrice de transition suivante :

$$\mathrm{Q} = \begin{pmatrix} 0.2 & 0.2 & 0.3 & 0.3 & 0 \\ 0.4 & 0 & 0.3 & 0.3 & 0 \\ 0.2 & 0.2 & 0.4 & 0.1 & 0.1 \\ 0.2 & 0.1 & 0.3 & 0.1 & 0.3 \\ 0.2 & 0 & 0.5 & 0.3 & 0 \end{pmatrix}$$

On a $1 - \beta = 1 - (0.2 + 0 + 0.3 + 0.1 + 0) = 0.4$ et la deuxième plus grande valeur propre en module vaut $|\beta_1| = 0.31$

Nous nous proposons de généraliser ce résultat au cas de chaîne de Markov d'ordre supérieur à 1. Lorsque l'on s'intéresse à une chaîne X d'ordre $m > 1$, la première réaction consiste généralement à définir la chaîne de Markov Y d'ordre 1 équivalente (cf. paragraphe 2.3.1 page 16). Cette opération, dans notre cas, conduit à une valeur de β nulle, car chaque colonne de la matrice de transition de Y possède nécessairement une valeur de transition nulle. Le résultat se trouve dans la proposition suivante.

Proposition 2.14 *Soit une chaîne de Markov $X = X_1, ..., X_n$ d'ordre m à espace d'états $\mathcal{A}$ de matrice de transition Q. X est telle qu'il existe un réel β telle que $Q(x, A) \geqslant \beta\zeta(A)$ pour tout $x \in \mathcal{A}^m$ et pour tout ensemble mesurable de $A \subseteq \mathcal{A}$ et une distribution ζ sur $\mathcal{A}$. On a alors pour tout $n \in \mathbb{N}$:*

$$||Q^n - Q^\infty||_{vt} \leqslant q_n^m(\beta)$$

où $||\bullet||_{vt}$ désigne la distance en variation totale
et $Q^\infty = \lim_{n\to\infty} Q^n$ la distribution stationnaire
et $q_n^m(\beta)$ désigne la probabilité d'observer aucune suite de m succès consécutifs dans une suite de n variables de Bernoulli indépendantes de paramètre β.

Preuve. La preuve s'inspire de la démonstration fournie dans Rosenthal (1995) dans le cas d'une chaîne de Markov d'ordre 1. Elle consiste à définir deux chaînes de Markov d'ordre $m > 1$ de même matrice de transition Q ; la seule différence entre ces deux chaînes

de Markov réside dans la loi du point de départ. En effet, le point de départ de la première chaîne est issue d'une loi quelconque, alors que le point de départ de la deuxième chaîne est issue de la distribution stationnaire commune à ces deux chaînes. La deuxième chaîne se trouve donc à l'état stationnaire, c'est à dire que $\mathbb{P}\left[(Z_i, Z_{i+1}, ..., Z_{i+k}) = (z_i, z_{i+1}, ..., z_{i+k})\right]$ ne dépend pas de i.

Soit le couple $(X_k, Z_k), k \in \mathbb{N}$ défini de la façon suivante :

- $(X_0, ..., X_{m-1})$ suit une distribution quelconque μ_0
- et $(Z_0, ..., Z_{m-1})$ suit la distribution stationnaire de la chaîne X notée μ_∞.

On définit également la suite de variables aléatoires, indépendantes et identiquement distribuées :
$\forall i \in \mathbb{N}, U_i \sim \text{Bernoulli}(\beta)$.
Conditionnellement à $(X_{k-m}, ..., X_{k-1})$ et $(Z_{k-m}, ..., Z_{k-1})$, la distribution de (X_k, Z_k) est la suivante :

a) si $U_k = 1$ alors z est tiré selon la distribution $\zeta(.)$ et $X_k = Z_k = z$

b) si $U_k = 0$ alors X_k et Z_k sont tirées indépendamment avec :

$$\mathbb{P}(X_k \in A | X_{k-m}...X_{k-1}, U_k = 0) = \frac{\mathrm{Q}(X_{k-m}...X_{k-1}, A) - \beta\zeta(A)}{1-\beta}$$

$$\mathbb{P}(Z_k \in A | Z_{k-m}...Z_{k-1}, U_k = 0) = \frac{\mathrm{Q}(Z_{k-m}...Z_{k-1}, A) - \beta\zeta(A)}{1-\beta}$$

Un simple calcul montre que l'on a bien $\mathbb{P}(X_k \in A | X_{k-m}...X_{k-1}) = \mathrm{Q}(X_{k-m}...X_{k-1}, A)$ (de même pour Z_k). De cette façon, les suites $X_0, ..., X_k$ et $Z_0, ..., Z_k$ constituent bien deux chaînes de Markov d'ordre m de même matrice de transition Q. De plus, il y a une probabilité β à chaque temps k pour que les deux chaînes prennent la même valeur. Si cela se produit m fois de suite, alors on couple les deux chaînes.
Soit T le premier instant où $U_{T-m+1} = ... = U_T = 1$. On définit

$$Y_k = \begin{cases} Z_k & \text{si } k \leqslant T \\ X_k & \text{si } k > T \end{cases}$$

La chaîne (X_k, Y_k) définit un couplage avec T comme temps de couplage. On peut réécrire la distance en variation totale concernée, de la manière suivante :

$$||\mathrm{Q}^n - \mathrm{Q}^\infty||_{vt} = ||\mathcal{L}(X_k) - \mathcal{L}(Y_k)||_{vt}$$

La théorie sur les temps de couplage nous permet de majorer la distance en variation totale entre deux lois :

$$||\mathcal{L}(X_k) - \mathcal{L}(Y_k)||_{vt} \leqslant \mathbb{P}(X_k \neq Y_k)$$

et

$$\mathbb{P}(X_k \neq Y_k) \leqslant \mathbb{P}(T > k)$$

Or, T est associé au premier temps où est observé une suite de r 1 successifs dans une suite de Bernoulli indépendantes de paramètre β, on a donc $\mathbb{P}(T > k) = q_n^m(\beta)$. ■

Ce résultat n'est utile en pratique que si la probabilité $q_n^r(\beta)$ de n'observer aucune série de r succès successifs dans une suite de variables de Bernoulli de paramètres β est connue. Ce problème, dont la solution n'est pas immédiate, est largement étudié dans la littérature. On se contente ici de rappeler deux résultats connus sur cette probabilité. Le premier est un simple calcul récursif permettant de connaître cette probabilité de façon exacte. Le deuxième est une approximation extraite de Feller (1968).

Cette probabilité peut se calculer récursivement de la façon suivante :

$$q_n^r(\beta) = q_{n-1}^r(\beta) - \beta^r(1-\beta)q_{n-r-1}^r(\beta)$$

avec comme valeurs initiales : $\begin{cases} 1 & \text{si} \quad n < r \\ 1-\beta^r & \text{si} \quad n = r \end{cases}$

Feller (1968) donne le comportement asymptotique en n de cette probabilité :

$$q_n^r(\beta) \sim \frac{1-\beta x}{(r+1-rx)(1-\beta)} \frac{1}{x^{n+1}} \tag{2.1}$$

où x est la seule racine positive du polynome de degré r suivant :

$$p(x) = (1-\beta)\sum_{k=0}^{r-1} \beta^k x^k - 1$$

Soulignons que les résultats obtenus à l'aide de cette approximation sont étonnamment précis, même pour n petit (cf. la table 2.2 ci-dessous). Les approximations sont obtenues à l'aide de l'équation 2.1.

	$r=2$			$r=3$			$r=4$		
n	$q_n^2(0.5)$	Approx.	erreur	$q_n^3(0.5)$	Approx.	erreur	$q_n^4(0.5)$	Approx.	erreur
2	0.75	0.7663	0.0163						
3	0.625	0.6200	0.0080	0.875	0.8846	0.0097			
4	0.500	0.5016	0.0016	0.8125	0.8135	0.0011	0.9375	0.9419	0.0044
5	0.40625	0.4058	0.0005	0.75	0.7482	-0.0018	0.90625	0.9078	0.0015

TAB. 2.2 – Probabilité exacte et approchée selon Feller (1968) de n'observer aucune suite de r 1 consécutifs dans une séquence de taille n de Bernoulli indépendantes de paramètre $\beta = 0.5$

Remarque 2.15 *Cette remarque a pour but de gagner en temps de calcul (mais de perdre en précision) en évitant de chercher les racines d'un polynôme de degré r. En effet, la proposition 2.12 garantit que la distance en variation totale entre Q^n et la distribution stationnaire est de la forme : $a\gamma^n$. Dans le cas où $m = 1$, la proposition 2.13 donne $a = 1$ et $\gamma = 1 - \beta$. Pour $r > 1$, on propose ici de déterminer ces valeurs a et γ afin qu'elles coïncident avec les premiers termes de la série $q_n^m(\beta)$.*
On considère les trois premiers termes :

$$\begin{aligned} a\gamma &= q_m^m(\beta) \\ a\gamma^2 &= q_{m+1}^m(\beta) \\ a\gamma^3 &= q_{m+2}^m(\beta) \end{aligned}$$

Comme $m > 1$, un simple calcul basé sur la forme récursive de q_n^r donne $q_{m+1}^m(\beta) = 1 - \beta^m - \beta^m(1-\beta)$ et $q_{m+2}^m(\beta) = 1 - \beta^m - 2\beta^m(1-\beta)$. La valeur de γ est donnée par le rapport de ces deux quantités ; on en déduit ensuite la valeur de a. Les résultats sont les suivants :

$$\gamma = 1 - \frac{\beta^m(1-\beta)}{1 - \beta^m - \beta^m(1-\beta)}$$

$$a = \frac{[1 - \beta^m - \beta^m(1-\beta)]^3}{[1 - \beta^m - \beta^m(1-\beta)]^2}$$

Corollaire 2.16 *Soit une chaîne de Markov $X = X_1, ..., X_n$ d'ordre m à espace d'états $\mathcal{A}$ de matrice de transition Q. X est telle qu'il existe un réel β telle que $Q(x, A) \geqslant \beta\zeta(A)$ pour tout $x \in \mathcal{A}$ et pour tout ensemble mesurable de $A \subseteq \mathcal{A}$ et une distribution ζ sur $\mathcal{A}$. On a alors pour tout $n \in \mathbb{N}$:*

$$||Q^n - Q^\infty||_{vt} \leqslant C(1 - \beta^m)^n$$

où $||\bullet||_{vt}$ désigne la distance en variation totale
et $Q^\infty = \lim_{n\to\infty} Q^n$ la distribution stationnaire,
et C est une constante.

Preuve. La preuve est immédiate en utilisant les propositions 2.14 et 2.12. En effet, $q_m^m(\beta) = 1 - \beta^m$. ■

2.4.4 Distribution des nombres d'occurrence de mots dans une chaîne de Markov

On s'intéresse ici à la distribution jointe de $(N^{W_1}, N^{W_2}, ...)$ où W_i est un mot de h_i lettres et N^{W_i} est la variable aléatoire égale au nombre d'occurrences de ce mot dans une séquence de longueur n. A cette occasion, le calcul de l'espérance de ces nombres, ainsi que de leurs covariances est détaillé.

Le but de ce paragraphe n'est pas de présenter une revue de toutes les approximations existantes de cette distribution, mais plutôt d'introduire quelques résultats qui nous seront utiles pour la suite. Pour avoir plus de détails sur les résultats présentés ou sur d'autres se référer, par exemple, à Reinert et al. (2000), Robin and Schbath (2001) et Robin and Daudin (1999). Nuel (2004) s'intéresse à la distribution de ces comptages afin de découvrir des mots dont la fréquence est exceptionnellement élevée ou faible dans les séquences biologiques ; une comparaison de différentes approximations est présentée à cette occasion.

Les résultats présentés ici sont principalement issus de l'ouvrage de Waterman (1995), notamment le théorème suivant :

Théorème 2.17 *(**Loi jointe des comptages**) Soit X une chaîne de Markov stationnaire, irréductible et apériodique d'ordre 1. Soit $\mathcal{W} = \{w_1, ..., w_m\}$ un ensemble de mots et $N = (N^{w_1}, ..., N^{w_m})$ leur comptage dans une séquence de longueur n. Alors, $n^{-1}N$ est asymptotiquement gaussien de moyenne ν et de matrice de covariance $n^{-1}\Sigma$. Si $det(\Sigma) \neq 0$, alors*

$$n^{1/2}\Sigma^{-1/2}(N/n - \nu) \rightarrow^d \mathcal{N}(0, 1)$$

où
$\nu = \lim_{n\to\infty} n^{-1} \left(\mathbb{E}(N^{w_1}), ..., \mathbb{E}(N^{w_m})\right)$
$\Sigma = (\sigma_{ij})$
avec $\sigma_{ij} = \lim_{n\to\infty} n^{-1} Cov(N^{w_i}, N^{w_j})$

Calcul de l'espérance

On considère deux mots U et V de longueurs respectives k et l, et une chaîne de Markov X stationnaire d'ordre m de longueur n. On suppose $l \geqslant k > m$. Les calculs des moments reposent sur le fait que : $N^U = \sum_{i=1}^{n-k+1} Y_i^U$ où Y_i^U est l'indicatrice de la présence du mot U à la position i dans X.

$\forall i = 1, ..., n-k+1,\ \mathbb{E}(Y_i^U)$ est la probabilité d'apparition du mot U. La chaîne étant stationnaire, on a :

$$\mathbb{E}(Y_i^U) = \mathbb{E}(Y_0^U) = \mu(u_1...u_m) \prod_{j=m}^{k-1} \mathrm{Q}(u_{j-m+1}...u_j, u_{j+1})$$

On en déduit :

$$\mathbb{E}(N^U) = (n-k+1)\mu(u_1...u_m) \prod_{i=m}^{k-1} \mathrm{Q}(u_{i-m+1}...u_i, u_{i+1})$$

Pour le calcul de la covariance, nous avons besoin de connaître $\forall i = 1, ..., n-k+1,\ \forall j = 1, ..., n-l+1,\ \mathbb{E}(Y_i^U Y_j^V)$.

Recouvrement de deux mots Ce calcul est rendu plus difficile par le fait que les mots U et V peuvent se **recouvrir**.

Considérons l'exemple suivant : U="abcd", V="cde" et W="hih". Une séquence de longueur 5 peut contenir une fois le mot U ET une fois le mot V ; il suffit d'ajouter la lettre "e" au mot U. Par contre, il n'est pas possible sur une séquence de cette longueur d'obtenir 1 fois le mot U et 1 fois le mot W.

Ceci est dû au fait que U et V se recouvrent, c'est à dire que la fin du mot U est égale au début du mot V. Remarquons que cette notion n'est pas symétrique. On doit également préciser sur combien de positions les mots se recouvrent. Formellement, on définit le **bit de recouvrement** de deux mots U et V à la position j, $\beta_{U,V}(j)$ tel que :

$$\beta_{U,V}(j) = \begin{cases} 1 \text{ si } U_{1+j} = V_1, ..., U_k = V_{k-j} \\ 0 \text{ sinon} \end{cases}$$

La table 2.3 page 35 donne un exemple de calcul des bits de recouvrement pour les mots U et V. Les bits de recouvrement de W avec W sont : $\beta_{W,W} = (1, 0, 1)$.

TAB. 2.3 – Exemple de calcul de bits de recouvrement avec les mots U="abcd" et V="cde"

j	0	1	2	3
alignement	$a\ b\ c\ d$ $c\ d\ e$	$a\ b\ c\ d$ $c\ d\ e$	$a\ b\ c\ d$ $\|\ \ \|$ $c\ d\ e$	$a\ b\ c\ d$ $c\ d\ e$
$\beta(j)$	0	0	1	0

Calcul des moments d'ordre 2

Nous allons ici calculer la covariance des nombres d'apparitions de deux mots différents, la variance du nombre d'apparition d'un mot en est un cas particulier. L'expression de cette variance n'étant pas plus simple que l'expression de la covariance, seule la formule de la covariance est donnée. On obtient facilement $\mathbb{V}(N^U) = \mathrm{Cov}(N^U, N^U)$.

On utilise le bit de recouvrement pour calculer l'espérance de $Y_i^U Y_j^V$. On distingue deux cas : soit les deux mots apparaissent à des positions telles qu'ils sont totalement séparés ($k \leqslant j - i$), soit ils se recouvrent ($j - i < k$) et on doit prendre en compte les lettres restantes du deuxième mot :

Pour $k \leqslant j - i$,

$$\mathbb{E}(Y_i^U Y_j^V) = \mathbb{E}(Y_i^U) p^{(j-i-k+1)}_{u_{k-m+1}...u_k, v_1...v_m} \prod_{s=j+1}^{j+l-1} \mathrm{Q}(v_{s-m}...v_{s-1}, v_s)$$

Pour $i \leqslant j$ et $j - i < k$,

$$\mathbb{E}(Y_i^U Y_j^V) = \beta_{U,V}(j-i)\mathbb{E}(Y_i^U) \prod_{s=i+k}^{i+k-j+l-1} \mathrm{Q}(v_{s-m}...v_{s-1}, v_s)$$

où $p^n_{w_1...w_m,w'_1...w'_m}$ est la probabilité d'aller du mot de taille m $w_1...w_m$ au mot $w'_1...,w'_m$ en n pas. Formellement,

$$p^n_{w_1...w_m,w'_1...w'_m} = \mathbb{P}(X_{n+m} = w'_1, ..., X_{n+2m-1} = w'_m | X_1 = w_1, ..., X_m = w_m)$$

Si X est d'ordre 1, on a $p^n_{w_1,w_2} = \mathrm{Q}^n(w_1, w2)$

On en déduit la valeur de

$$\mathrm{Cov}(Y_i^U, Y_j^V) = \mathbb{E}(Y_i^U, Y_j^V) - \mathbb{E}(Y_i^U)\mathbb{E}(Y_j^V)$$

$$= \begin{cases} \mathbb{E}(Y_0^U)\left\{\beta_{U,V}(j-i)\prod_{s=i+k}^{i+k-j+l}\mathrm{Q}(v_{s-m}...v_{s-1}, v_s) - \mathbb{E}(Y_0^V)\right\} & \text{si } 0 \leqslant j-i < k \\ \mathbb{E}(Y_0^V)\left\{\beta_{V,U}(i-j)\prod_{s=j+l}^{j+l-i+k}\mathrm{Q}(u_{s-m}...u_{s-1}, u_s) - \mathbb{E}(Y_0^U)\right\} & \text{si } 0 \leqslant i-j < l \\ \mathbb{E}(Y_0^U)\left\{p^{(j-i-k+1)}_{u_k,v_1}\prod_{s=j+1}^{j+l-1}\mathrm{Q}(v_{s-m}...v_{s-1}, v_s) - \mathbb{E}(Y_0^V)\right\} & \text{si } k \leqslant j-i \\ \mathbb{E}(Y_0^V)\left\{p^{(i-j-l+1)}_{v_k,u_1}\prod_{s=i+1}^{i+k-1}\mathrm{Q}(u_{s-m}...u_{s-1}, u_s) - \mathbb{E}(Y_0^U)\right\} & \text{si } l \leqslant i-j \end{cases}$$

On calcule ensuite la covariance des nombres d'apparitions de U et de V de la façon suivante :

$$\begin{aligned} \mathrm{Cov}(N^U, N^V) &= \mathrm{Cov}\left(\sum_{i=1}^{n-k+1} Y_i^U, \sum_{j=1}^{n-l+1} Y_j^V\right) \\ &= \sum_{i=1}^{n-k+1}\sum_{j=1}^{n-l+1}\mathrm{Cov}(Y_i^U, Y_j^V) \end{aligned}$$

Beaucoup de termes ont la même valeur dans cette double somme. En effet, la chaîne étant en régime stationnaire, à δ fixé, $\delta = 0, ..., n-l$, toutes les $\mathrm{Cov}(Y_i^U, Y_{i+\delta}^V)$ pour $i = 1, ..., n-k-\delta+1$ sont égales, le nombre de termes égaux étant $n-k+1-\delta$. En faisant la même remarque pour $\mathrm{Cov}(Y_i^V, Y_{i+\delta}^U)$, on réécrit :

$$\begin{aligned} \mathrm{Cov}(N^U, N^V) &= (n-l+1)\times\mathrm{Cov}(Y_1^U, Y_1^V) + \\ &\quad \sum_{\delta=1}^{n-l}(n-l+1-\delta)\times\mathrm{Cov}(Y_1^U, Y_{1+\delta}^V) + \\ &\quad \sum_{\delta=1}^{n-k}(n-k+1-\delta)\times\mathrm{Cov}(Y_1^V, Y_{1+\delta}^U) \end{aligned}$$

En pratique...

Dans les applications génomiques, la longueur n de la chaîne peut être très élevée. Par conséquent le calcul de la covariance entre les nombres d'occurrences de deux mots est relativement long (en $o(n)$). En tenant compte du fait que la précision des nombres représentés par l'ordinateur est limitée, on peut faire le calcul en $o(1)$.

En effet, le calcul de la covariance entre deux nombres d'apparitions fait intervenir $\mathrm{Cov}(Y_1^U, Y_{1+\delta}^V)$. Cette covariance tend vers 0 quand δ tend vers $+\infty$. L'ordinateur considérera comme nulle ces covariances pour $\delta > \delta_0$ pour un certain δ_0. Il n'est donc pas nécessaire de faire varier la somme sur δ de 1 à $n-l$, mais de 1 à $\min(\delta_0, n-l)$.

Le problème est donc le suivant : Si l'on note ε la précision de la machine utilisée (généralement, ce nombre est de l'ordre de 10^{-16}), quelle est la valeur seuil δ_0 telle que pour tout $\delta > \delta_0$, la représentation en machine de $\mathrm{Cov}(Y_1^U, Y_{1+\delta}^V)$ peut être considérée comme nulle ?

$$\mathrm{Cov}(Y_1^U, Y_{1+\delta}^V) = \mathbb{E}(Y_1^U Y_{1+\delta}^V) - \mathbb{E}(Y_1^U)\mathbb{E}(Y_{1+\delta}^V)$$

On considère $\delta \geqslant k$, donc

$$\begin{aligned}\mathbb{E}(Y_1^U Y_{1+\delta}^V) &= \mathbb{E}(Y_i^U) p^{(\delta-k+1)}_{u_{k-m+1}\ldots u_k, v_1\ldots v_m} \prod_{s=j+1}^{j+l-1} \mathrm{Q}(v_{s-m}\ldots v_{s-1}, v_s) \\ &= \mathbb{E}(Y_i^U) p^{(\delta-k+1)}_{u_{k-m+1}\ldots u_k, v_1\ldots v_m} \mathbb{E}(Y_{1+\delta}^V)/\mu(v_1, \ldots, v_m)\end{aligned}$$

On a donc l'équivalence suivante :

$$\mathrm{Cov}(Y_1^U, Y_{1+\delta}^V) = 0 \Longleftrightarrow p^{(\delta-k+1)}_{u_{k-m+1}\ldots u_k, v_1\ldots v_m} = \mu(v_1, \ldots, v_m)$$

En ce qui nous concerne, nous préférerons ici, l'approche utilisant la seconde valeur propre. Ce choix est motivé par le fait qu'il existe des algorithmes efficaces de recherche de valeurs propres, et qu'une solution plus précise est fournie. Finalement, δ_0 est la solution de l'équation suivante en δ :

$$\varepsilon = C|\beta_1|^{\delta-k+1}$$

donc δ_0 est égale à :

$$\delta_0 = \frac{\log \epsilon - \log C}{\log(|\beta_1|)} + k - 1$$

k désigne dans cette formule la longueur du mot le plus long.

Le tableau 2.4 présente des valeurs de δ_0 pour des matrices de transition estimées sur les génomes complets de quelques organismes (avec $k = 1$). Nous avons considéré ici que la constante C valait 1. Cette valeur est de l'ordre de ce qui est observée en pratique sur des séquences génomiques.

Utilisation de la série des puissances de Q La solution présentée précédemment permet d'approximer de manière suffisamment précise les variances des comptages dans les cas réels, c'est à dire que la quantité négligée reste généralement très faible. Néanmoins, l'erreur commise sur une séquence de taille n tend vers l'infini lorsque n tend vers l'infini.

Séquence \ Ordre	1	2	3	4	5
Hiv	16	29	45	63	84
Saccharomyces Cerevisiae	15	28	50	52	60
Bacillus Subtilis	18	32	48	62	62
Escherichia Coli	17	29	55	59	67

TAB. 2.4 – Quelques exemples de valeurs de δ_0 pour des chaînes de Markov estimée sur des séquences d'ADN et une précision machine de $\varepsilon = 10^{-16}$ (avec k=1).

Une autre manière d'effectuer ce calcul consiste à s'appuyer sur les résultats connus de séries matricielles. En effet, ces variances et covariances font intervenir les séries suivantes : $\sum_{k=1}^{n} \mathrm{Q}^k$ et $\sum_{k=1}^{n} k\mathrm{Q}^k$.

Proposition 2.18 *Soit X une chaîne de Markov irréductible d'ordre 1 de matrice de transition Q. On note Q^i la ième puissance de Q et $Q^\infty = \lim_{n\to\infty} Q^n$. On a les résultats suivants :*

$$\sum_{k=0}^{n} Q^k = (I - Q^{n+1} + Q^\infty)(I - Q + Q^\infty)^{-1} + nQ^\infty \tag{2.2}$$

$$\sum_{k=1}^{n} Q^k = (Q - Q^{n+1})(I - Q + Q^\infty)^{-1} + nQ^\infty \tag{2.3}$$

$$\sum_{k=0}^{n} kQ^k = (nQ^{n+2} - (n+1)Q^{n+1} + Q)(I - Q + Q^\infty)^{-2} + \frac{n(n+1)}{2}Q^\infty \tag{2.4}$$

Preuve. On rappelle les résultats suivants dans le cas où $I - A$ est une matrice inversible. I désigne la matrice identité.

$$\begin{aligned}
\textstyle\sum_{k=0}^{n} A^k &= (I - A^{n+1})(I - A)^{-1} \\
\textstyle\sum_{k=1}^{n} A^k &= A(I - A^n)(I - A)^{-1} \\
\textstyle\sum_{k=0}^{n} kA^k &= (A - (n+1)A^{n+1} + nA^{n+2})(I - A)^{-2}
\end{aligned}$$

Dans notre cas, $I - \mathrm{Q}$ n'est pas inversible, car il s'agit d'une matrice stochastique, et qui admet, par conséquent, 1 comme valeur propre.

La preuve est fondée sur la propriété suivante : $\forall k \geqslant 1$:

$$(\mathrm{Q} - \mathrm{Q}^\infty)^k = \mathrm{Q}^k - \mathrm{Q}^\infty$$

En effet,

$$\begin{aligned}
(\mathrm{Q} - \mathrm{Q}^\infty)^2 &= \mathrm{Q}^2 - \mathrm{Q}\mathrm{Q}^\infty - \mathrm{Q}^\infty\mathrm{Q} + \mathrm{Q}^{\infty 2} \\
&= \mathrm{Q}^2 - 2\mathrm{Q}^\infty + \mathrm{Q}^\infty \\
&= \mathrm{Q}^2 - \mathrm{Q}^\infty
\end{aligned}$$

On suppose que la propriété $(\mathrm{Q} - \mathrm{Q}^\infty)^k = \mathrm{Q}^k - \mathrm{Q}^\infty$ est vraie pour un certain $k \geqslant 1$. On montre qu'elle est vraie en $k + 1$.

$$\begin{aligned}(\mathrm{Q}-\mathrm{Q}^{\infty})^{k+1} &= \mathrm{Q}^{k+1}-\mathrm{Q}^{k}\mathrm{Q}^{\infty}-\mathrm{Q}^{\infty}\mathrm{Q}+\mathrm{Q}^{\infty 2}\\ &= \mathrm{Q}^{k+1}-2\mathrm{Q}^{\infty}+\mathrm{Q}^{\infty}\\ &= \mathrm{Q}^{k+1}-\mathrm{Q}^{\infty}\end{aligned}$$

Les deux premiers résultats de la proposition proviennent du fait que :
$\sum_{k=1}^{n}\mathrm{Q}^{k}=\sum_{k=1}^{n}(\mathrm{Q}-\mathrm{Q}^{\infty})^{k}+n\mathrm{Q}^{\infty}$
d'où,
$\sum_{k=0}^{n}\mathrm{Q}^{k}=\sum_{k=0}^{n}(\mathrm{Q}-\mathrm{Q}^{\infty})^{k}+n\mathrm{Q}^{\infty}$
La matrice $I-(\mathrm{Q}-\mathrm{Q}^{\infty})$ est inversible. En effet, Q^{∞} est une matrice transition de rang 1, et ses valeurs propres sont toutes nulles exceptées la plus grande qui vaut 1. Donc les valeurs propres de $(\mathrm{Q}-\mathrm{Q}^{\infty})$ sont différentes de 1 et $I-(\mathrm{Q}-\mathrm{Q}^{\infty})$ ne possède aucune valeur propre nulle.
On applique le résultat sur les séries de matrices énoncé précédemment pour trouver le résultat.

Pour le deuxième résultat, cela donne :

$$\begin{aligned}\textstyle\sum_{k=1}^{n}\mathrm{Q}^{k} &= (\mathrm{Q}-\mathrm{Q}^{\infty})(I-(\mathrm{Q}-\mathrm{Q}^{\infty})^{n})(I-\mathrm{Q}+\mathrm{Q}^{\infty})^{-1}+n\mathrm{Q}^{\infty}\\ &= (\mathrm{Q}-\mathrm{Q}^{n+1}+\mathrm{Q}^{\infty}-\mathrm{Q}^{\infty}+\mathrm{Q}^{\infty}-\mathrm{Q}^{\infty})(I-\mathrm{Q}+\mathrm{Q}^{\infty})^{-1}+n\mathrm{Q}^{\infty}\\ &= (\mathrm{Q}-\mathrm{Q}^{n+1})(I-\mathrm{Q}+\mathrm{Q}^{\infty})^{-1}+n\mathrm{Q}^{\infty}\end{aligned}$$

Pour montrer le troisième résultat, on utilise le fait que : $\sum_{k=0}^{n}k\mathrm{Q}^{k}=\sum_{k=0}^{n}k(\mathrm{Q}-\mathrm{Q}^{\infty})^{k}+\mathrm{Q}^{\infty}\sum_{k=0}^{n}k$. Là encore, on s'appuie sur la série de $\sum_{k=0}^{n}k(\mathrm{Q}-\mathrm{Q}^{\infty})^{k}$ pour trouver le résultat.
■

Remarque 2.19 *Ce résultat peut s'étendre aux chaînes de Markov d'ordre supérieur à 1. La matrice de transition Q sera alors la matrice carrée obtenue en considérant la chaîne de Markov d'ordre 1 équivalente (c.f. page 16).*

Du point de vue de la complexité, cette méthode nécessite trois types d'opération. Le premier consiste en le calcul de Q^{∞}, c'est à dire du vecteur propre (à gauche) associée à la valeur propre 1. La librairie ARPACK implémente la méthode d'Arnoldi permettant d'extraire ce vecteur propre sans nécessairement extraire les autres (Lehoucq R et al., 1996). La complexité de cet algorithme est une fonction environ linéaire du nombre de valeurs non nulles de la matrice considérée, c'est à dire $|\mathcal{A}|^{m+1}$ dans un modèle Mm. Le deuxième est l'inversion de $(I-\mathrm{Q}+\mathrm{Q}^{\infty})$ qui est une opération dont la complexité est de l'ordre de k^2 pour une matrice carré de taille k avec, dans notre cas, $k=|\mathcal{A}|^{m}$. Le troisième est le calcul de Q^{n+1} dont la complexité est plus élevée ; en effet, un produit de deux matrices de taille k possède une complexité de l'ordre de k^3 avec dans notre cas, $k=|\mathcal{A}|^{m}$. L'opération pourra être optimisée de telle façon qu'il ne soit pas nécessaire d'effectuer n produits de matrices*, néanmoins la complexité reste élevée. Il est à noter

*en effectuant une décomposition en base 2 de n par exemple

qu'à partir de $n = \delta_0$ calculé ci-dessus, les valeurs de Q^k et de Q^∞ sont assimilées. Pour une séquence suffisamment longue, le calcul de Q^{n+1} peut être évité au prix d'une approximation. Il est important de noter ici que l'erreur due à cette approximation tend vers 0 quand n vers l'infini ce qui n'était pas le cas précédemment.

2.5 Chaîne de Markov inversée

Dans ce paragraphe, nous verrons que si $X = X_1...X_n$ est une chaîne de Markov, alors $X^- = X_n...X_1$ est également une chaîne de Markov dont la matrice de transition et la distribution stationnaire sont connues. Cette notion est le coeur du travail présenté dans le reste cette thèse. Nous introduisons également des notations qui seront utiles dans la suite du document.

2.5.1 Notations

Soit un mot $W = w_1...w_h$. On note le mot inverse $W^- = w_h...w_1$.

On note également $\bar{w} \in \mathcal{A}$ le complémentaire d'une lettre $w \in \mathcal{A}$. On a la propriété suivante :

$$\bar{\bar{w}} = w$$

Pour les nucléotides, on a $\bar{a} = t$ et $\bar{c} = g$

On définit également le mot inverse complémentaire de W : $\bar{W} = \bar{w}_h...\bar{w}_1$. On omet l'exposant "$^-$" pour ne pas alourdir les notations, mais $\bar{W}$ est néanmoins inversé.

Dans la suite du document, les formules seront établies en supposant que l'on passe à l'alphabet complémentaire. Pour retrouver les formules sans passer au complémentaire, il suffit de poser $\bar{w} = w$.

Proposition 2.20 *Soit $X = X_1...X_n$ une chaîne de Markov d'ordre m de matrice de transition Q et de distribution stationnaire μ. Alors $\bar{X} = \bar{X}_n...\bar{X}_1$ est également une chaîne de Markov d'ordre m de matrice de transition $\bar{Q}$ et de distribution stationnaire $\bar{\mu}$. On a*

$$\bar{\mu}(x_1...x_m) = \mu(\bar{x}_m...\bar{x}_1)$$

$$\bar{Q}(x_1...x_{m+1}) = Q(\bar{x}_{m+1}...\bar{x}_1) \times \frac{\mu(\bar{x}_{m+1}...\bar{x}_2)}{\bar{\mu}(x_1...x_m)}$$

Preuve. On prouve que $\bar{X}$ est une chaîne de Markov. Pour $i = 1, ..., n$:

$$\mathbb{P}(\bar{X}_i = \bar{x}_i | \bar{X}_{i+1} = \bar{x}_{i+1}, ..., \bar{X}_n = \bar{x}_n) = \frac{\mathbb{P}(\bar{X}_i = \bar{x}_i, \bar{X}_{i+1} = \bar{x}_{i+1}, ..., \bar{X}_n = \bar{x}_n)}{\mathbb{P}(\bar{X}_{i+1} = \bar{x}_{i+1}, ..., \bar{X}_n = \bar{x}_n)}$$

Or, le numérateur vaut :

$$\begin{aligned}\mathbb{P}(\bar{X}_i = \bar{x}_i, \bar{X}_{i+1} = \bar{x}_{i+1}, ..., \bar{X}_n = \bar{x}_n) &= \mathbb{P}(\bar{X}_i = \bar{x}_i, ..., \bar{X}_{i+m-1} = x_{i+m-1}) \times \\ &\quad \mathbb{P}(\bar{X}_{i+m} = \bar{x}_{i+m} | \bar{X}_i = \bar{x}_i, ..., \bar{X}_{i+m-1} = x_{i+m-1}) \times \\ &\quad \mathbb{P}(\bar{X}_{i+m+1} = \bar{x}_{i+m+1} | \bar{X}_i = \bar{x}_i, ..., \bar{X}_{i+m} = x_{i+m})\end{aligned}$$

Comme X est une chaîne de Markov d'ordre m,

$$\begin{aligned}\mathbb{P}(\bar{X}_i = \bar{x}_i, \bar{X}_{i+1} = \bar{x}_{i+1}, ..., \bar{X}_n = \bar{x}_n) &= \mathbb{P}(\bar{X}_i = \bar{x}_i, ..., \bar{X}_{i+m-1} = x_{i+m-1}) \times \\ &\quad \mathbb{P}(\bar{X}_{i+m} = \bar{x}_{i+m} | \bar{X}_i = \bar{x}_i, ..., \bar{X}_{i+m-1} = x_{i+m-1}) \times \\ &\quad \mathbb{P}(\bar{X}_{i+m+1} = \bar{x}_{i+m+1} | \bar{X}_{i+1} = \bar{x}_{i+1}, ..., \bar{X}_{i+m} = x_{i+m})\end{aligned}$$

De même, on a pour le dénominateur :

$$\begin{aligned}\mathbb{P}(\bar{X}_{i+1} = \bar{x}_{i+1}, ..., \bar{X}_n = \bar{x}_n) &= \mathbb{P}(\bar{X}_{i+1} = \bar{x}_{i+1}, ..., \bar{X}_{i+m} = x_{i+m}) \times \\ &\quad \mathbb{P}(\bar{X}_{i+m} = \bar{x}_{i+m} | \bar{X}_{i+1} = \bar{x}_{i+1}, ..., \bar{X}_{i+m} = x_{i+m}) \times \\ &\quad \mathbb{P}(\bar{X}_{i+m+1} = \bar{x}_{i+m+1} | \bar{X}_{i+1} = \bar{x}_{i+1}, ..., \bar{X}_{i+m} = x_{i+m})\end{aligned}$$

En simplifiant, on obtient finalement :

$$\begin{aligned}\mathbb{P}(\bar{X}_i = \bar{x}_i | \bar{X}_{i+1} = \bar{x}_{i+1}, ..., \bar{X}_n = \bar{x}_n) &= \mathbb{P}(\bar{X}_i = \bar{x}_i, ..., \bar{X}_{i+m-1} = x_{i+m-1}) \times \\ &\quad \mathbb{P}(\bar{X}_{i+m} = \bar{x}_{i+m} | \bar{X}_i = \bar{x}_i, ..., \bar{X}_{i+m-1} = x_{\bar{i+m-1}}) / \\ &\quad \mathbb{P}(\bar{X}_{i+1} = \bar{x}_{i+1}, ..., \bar{X}_{i+m} = x_{\bar{i+m}}) \\ &= \mathbb{P}(\bar{X}_i = \bar{x}_i | \bar{X}_{i+1} = \bar{x}_{i+1}, ..., \bar{X}_{i+m} = \bar{x}_{i+m})\end{aligned}$$

Il est par ailleurs évident que $\bar{\mu}(x_1...x_m) = \mu(\bar{x}_m...\bar{x}_1)$ puisque le sens du temps n'intervient pas dans la distribution stationnaire.

On calcule la matrice de transition. On a pour $x_1, ..., x_{m+1}$ quelconque de $\mathcal{A}$:

$$\bar{\mathbb{P}}(x_1, ..., x_{m+1}) = \mathbb{P}(\bar{x}_{m+1}, ..., \bar{x}_1)$$

où $\bar{\mathbb{P}}$ représente la probabilité sous le modèle inversé.

d'où

$$\bar{\mathrm{Q}}^-(x_1...x_{m+1}) = \mathrm{Q}(\bar{x}_{m+1}...\bar{x}_1) \times \frac{\mu(\bar{x}_{m+1}...\bar{x}_2)}{\bar{\mu}^-(x_1...x_m)})$$

■

Bibliographie

Baum, L. and Petrie, T. (1966). Statistical inference for probabilistic functions of finite state markov chains. *Ann. Math. Stat.*, 37 :1554–1563.

Berchtold, A. (2001). Estimation in the mixture transition distribution model. *Journal of Time Series Analysis*, 22(4) :379–397.

Bourguignon, P.-Y. and Robelin, D. (2004). Modèle de markov parcimonieux. In *Journées Ouvertes en Biologie, Informatique et Mathématiques*. Montréal.

Doob, J, I. (1953). *Stochastic Processes*. Wiley.

Feller, W. (1968). *An introduction to probability theory and its applications*. Wiley series in Probability and Mathematical Statistics. Wiley, 3 edition.

Lehoucq R, B., Sorensen, D, C., and Yang, C. (1996). ARPACK Users' Guide : Solution of Large Scale Eigenvalue Problems with Implicitly Restarted Arnoldi Methods. Technical report, Rice University.

Lèbre, S. (2004). Estimation du modèle mtd "mixture transition distribution". Technical report, DEA Mathématiques Fondamentales et Applications, Université de Rennes 1.

Meyn, S, P. and Tweedie, R, L. (1994). Computable bounds for geometric convergence rates of markov chains. *Ann. Appl. Probab*, 4 :981–1011.

Muri, F. (1997). *Comparaison d'algorithmes d'identification de chaînes de Markov cachées et applications à la détection de régions homogènes dans les séquences d'ADN.* PhD thesis, Université René Descartes, Paris V.

Nuel, G. (2001). *Grandes déviations et chaînes de Markov pour l'étude des occurences de mots dans les séquences biologiques*. PhD thesis, Université d'Evry Val d'Essonne.

Nuel, G. (2004). Ld-spatt : Large deviations statistics for patterns on markov chains. *Journal of computational Biology*, 11(6) :1023–1033.

Rabiner, L. (1989). A tutorial on hidden markov models and selected applications in speech recognition. In *Proceedings of the IEEE*, volume 77, pages 257–286.

Reinert, G., Schbath, S., and Waterman, Michael, S. (2000). Probabilistic and statistical properties of words : An overview. *Journal of computationnal Biology*, 7(1-2) :1–46.

Rissanen, J. (1983). A universal data compression system. *IEEE Trans. Inf. Theory*, 29 :656–664.

Robin, S. and Daudin, J.-J. (1999). Exact distribution of word occurrences in a random sequence of letters. *J. Appl. Probab.*, 36(1) :179–193.

Robin, S. and Schbath, S. (2001). Numerical comparison of several approximations of the word count distribution in random sequences. *J. Comp. Biol.*, 8 :349–359.

Rosenthal, S., J. (2002). Quantitative convergence rates of markov chains : a simple account. *Elect. Comm in Probab.*, 7 :123–128.

Rosenthal, Jeffrey, S. (1995). Convergence rates of markov chains. *SIAM review*, 37 :387–405.

Schbath, S. (1995). *Etude asymptotique du nombre d'occurences d'un mot dans une chaîne de Markov et application à la recherche de mots de fréquences exceptionnelle dans les séquence d'ADN*. PhD thesis, Université de Paris V.

Waterman, M. (1995). *Introduction to Computational Biology : Maps, sequences and genomes*, chapter Probability and Statistics for Sequence Patterns, pages 305–326. Chapman & Hall.

Willems, Frans, M., Shtarkov, Yuri, M., and Tjalkens, Tjalling, J. (1995). The context-tree weighting method : Basic properties. *IEEE Trans. Inf. Theory*, 41(3) :653–664.

Chapitre 3

Le score local : principaux résultats probabilistes et méthodes.

3.1 Introduction

La méthode dite de "score local" a pour but de détecter une accumulation de valeurs élevées dans une série. Cette série de valeurs est généralement issue de l'application d'une certaine fonction, appelé fonction de score, à une séquence qui n'est généralement pas numérique ; cette fonction est définie a priori de manière à prendre des valeurs élevées lorsque l'élément considéré de la séquence présente la particularité nous intéressant.

Initialement, cette méthode a été développée pour détecter des similitudes entre deux séquences biologiques ; à chaque couple de lettres est associé un score d'autant plus élevé que ces deux lettres se ressemblent. Ainsi, une accumulation de scores élevées désignera deux segments très ressemblants. Dans ce cas, la fonction de score reflète la "ressemblance" entre deux lettres. Dans le cadre de l'alignement de séquences, on peut citer, par exemple, des matrices BLOSUM (Henikoff and Henikoff, 1992) ou PAM (Dayhoff et al., 1978) qui définissent le score de similarité entre deux acides aminés quelconques.

Le score local est un objet mathématique déjà bien défini. On se propose ici d'en rappeler la définition, ainsi que les principaux résultats sur les lois des statistiques mises en jeu. Il ne s'agit pas d'une revue exhaustive, mais de fournir au lecteur non familier avec ces problématiques un point sur les connaissances actuelles et un bagage permettant de comprendre la suite de ce document. Nous nous bornerons au cas du score local appliqué à une seule séquence ou, de manière équivalente, au cas de comparaison de séquences lorsqu'il n'y a ni insertion, ni délétion. En effet, la possibilité d'introduire des insertions/délétions lors de l'alignement de deux séquences induit la présence de valeurs manquantes, ce qui complique considérablement la distribution du score local. Une large littérature a été développée sur ce sujet. Le cadre dans lequel nous appliquerons cette méthode dans le chapitre 5 ne subit pas ces problèmes. Nous incitons le lecteur intéressé à

la problématique de l'alignement de séquences à consulter des ouvrages plus spécifiques.

Ce chapitre est composé de deux grandes parties. La première concerne le score local maximal, tel qu'il est présenté dans la littérature. La deuxième partie s'intéresse aux valeurs successives du meilleur score local. Une définition de ces $k$ième scores locaux est présentée. Des résultats sur la distribution de ces scores sont ensuite établis dans le cas où chaque élément de la séquence sous-jacente est distribué indépendamment et identiquement aux autres éléments de la séquence (modèle M_0). Un algorithme de recherche des k meilleurs scores locaux, de complexité linéaire avec la taille de la séquence, est présenté. Enfin, une procédure séquentielle de test permettant de déterminer quels sont les scores locaux significatifs pour un risque de première espèce α donné est proposée et évaluée.

Un accent important est porté sur la faisabilité pratique des méthodes présentées. Nous prenons donc soin d'illustrer les résultats théoriques rencontrés avec des algorithmes de complexité raisonnable afin qu'ils puissent être utilisés.

3.2 Le score local maximal

Cette partie est largement inspirée de la thèse de Sabine Mercier (1999). La lecture de cette thèse fournit, selon moi, un excellent point de départ pour un mathématicien intéressé au problème du score local.

3.2.1 Définition et exemples

Définition 3.1 (Score local)
Soit $X = X_1X_2...X_n$ une suite de variables aléatoires à valeurs dans un alphabet fini $\mathcal{A}$. Soit s une fonction de score définie sur cet alphabet :

$$\begin{aligned} s: \quad & \mathcal{A} \to \mathbb{R} \\ & X_i \to s(X_i) \end{aligned}$$

On définit le score local H_n de la manière suivante :

$$H_n = \max_{1 \leq i \leq j \leq n} \left(\sum_{k=i}^{j} s(X_k) \right)$$

Une valeur élevée du score local est due à une série de X_i dont les scores sont élevés.

Si la fonction de score s est correctement choisie, un score élevé peut refléter des propriétés biophysiques ou biochimiques propres à la séquence. A ce jour, on recense des études concernant la charge, le volume, le caractère hydrophobe ou une structure tridimensionnelle potentielle s'appuyant sur des techniques de score local. Quelques fonctions de score pertinentes sont données dans la littérature ; on peut citer par exemple Kyte

and Doolittle (1982) pour un score simple d'hydrophobicité lié aux acides aminés (voir table 3.1), ou Muegge and Martin (1999) pour un score d'affinité entre une protéine et un ligand* en fonction de leurs structures tridimensionnelles.

Acide aminé	Code	Score d'hydrophobicité
Isoleucine	I	4.5
Valine	V	4.2
Leucine	L	3.8
Phenylalanine	F	2.8
Cysteine	C	2.5
Methionine	M	1.9
Alanine	A	1.8
Glycine	G	-0.4
Threonine	T	-0.7
Tryptophan	W	-0.9
Serine	S	-0.8
Tyrosine	Y	-1.3
Proline	P	-1.6
Histidine	H	-3.2
Acid Glutamic	E	-3.5
Glutamine	Q	-3.5
Acid Aspartic	D	-3.5
Asparagine	N	-3.5
Lysine	K	-3.9
Arginine	R	-4.5

TAB. 3.1 – Fonction de score d'hydrophobicité proposée par Kyte et Doolittle (1982)

Exemple On considère la séquence d'acides aminés suivantes dans laquelle on cherche la zone la plus hydrophobe. On utilise pour cela la fonction de score de la table 3.1 page 46. Cette séquence est arbitrairement extraite de la protéine transmembranaire *Rhodopsin* :

Séquence :	F	C	Y	G	Q	L	V	F	T	V
Scores associées :	2.8	2.5	-1.3	-0.4	-3.5	3.8	4.2	2.8	-0.7	4.2

On pourra vérifier que le maximum des sommes obtenues pour tous les segments possibles est obtenu pour la séquence complète, et qu'il vaut 14.4. On verra au paragraphe 3.2.3 un algorithme linéaire avec la longueur de la séquence permettant d'obtenir ce score ainsi que le segment associé.

*Molécule capable de s'attacher à un récepteur cellulaire

Remarquons qu'il existe deux phénomènes distincts qui conduiront à un score élevé mais qui ne présentent pas nécessairement les mêmes causes sous-jacentes. En effet, un score élevé peut être dû à une série courte de X_i de scores très élevés, ou à une série longue de X_i de scores moyennement élevés. Il semble important de noter ici que l'étude du score local seule ne permettra pas de distinguer ces deux phénomènes qui sont pourtant de nature biologique très différente.

3.2.2 Distribution asymptotique du score local H_n

Le but d'une étude de score local est de détecter une accumulation de scores élevés. La seule donnée de la valeur du score locale n'est généralement pas très informative. Il est plus intéressant de quantifier son exceptionnalité par rapport à un modèle probabiliste sur la séquence initiale des X_i.

Nous rappelons dans ce paragraphe les résultats asymptotiques connus sur la distribution de H_n quand n tend vers l'infini.

$\mathbf{s(X_1)s(X_2)...s(X_n)}$ sont indépendants et identiquement distribués (i.i.d.) Le signe de l'espérance de $s(X_i)$ détermine le comportement asymptotique de H_n. En effet, si l'espérance du score est strictement positive, on comprend intuitivement que plus on considérera de termes dans le calcul de $s(X_i) + s(X_{i+1}) + ... + s(X_{i+l})$, plus la valeur de cette somme sera élevée. Réciproquement, pour obtenir une telle somme élevée, on aura intérêt à prendre un grand nombre de termes. Finalement, la valeur asymptotique de H_n sera proportionnelle à la longueur de la séquence.

Si l'espérance du score $s(X_i)$ est négative, le comportement asymptotique est plus délicat à déterminer. Le résultat principal, dans ce cas, est dû à Karlin and Altschul (1990), Dembo and Karlin (1991b), Karlin and Dembo (1992) et Dembo et al. (1994) sous l'hypothèse, généralement vérifiée, que $\mathbb{P}(s(X_i) > 0) > 0$. Les auteurs ont mis en évidence la convergence vers une loi de Gumbel, dont ils donnent une forme analytique des paramètres dans le cas du score local. Ces paramètres sont rarement facilement calculable en pratique, et on verra une manière de les déterminer (voir le paragraphe 3.6). Ce résultat est utilisé par le programme de comparaison de séquence BLAST (Altschul et al., 1990), qui est un des programmes les plus utilisés dans le domaine de la bioinformatique actuellement.

Des résultats concernant la distribution asymptotique du score dans le cas où l'espérance de $s(X_i)$ est nulle ont été formulés par Daudin et al. (2003) et Etienne (2002). Ils expriment la convergence en loi de H_n correctement normalisée vers la loi du supremum d'un mouvement Brownien. Ils obtiennent également un résultat dans le cas où l'espérance est "petite".

Voici un résumé de ces différents résultats sur le comportement de H_n quand n tend vers l'infini (cas X_i i.i.d.).

– Si $\mathbb{E}[\mathbf{X_i}] < \mathbf{0}$ et si $\mathbb{P}(s(X_i) > 0) > 0$ alors

$$\lim_{n\to\infty} \mathbb{P}\left(H_n \leqslant \frac{\ln n}{\lambda} + x\right) = \exp(-K\exp(-\lambda x)) \tag{3.1}$$

où λ et K sont deux constantes qui ne dépendent que de la distribution de X_i. λ est l'unique solution positive de l'équation en x, $\mathbb{E}[e^{xs(X_i)}] = 1$ et K est une constante dont la formule est donnée ci-dessous.

$$K = \frac{\exp\left\{-2\sum_{k=1}^{\infty}\frac{1}{k}\left(\mathbb{E}[e^{\lambda S_k}; S_k < 0] + \mathbb{P}(S_k \geqslant 0)\right)\right\}}{\lambda\mathbb{E}(s(X)e^{\lambda s(X)})}$$

où S_k désigne la somme cumulée des scores jusqu'à l'indice k : $S_k = \sum_{i=1}^{k} s(X_k)$
Ces deux constantes peuvent être interprétées : K prend en compte la dépendance qui existe entre les différents objets dont on considère le maximum, et λ est un paramètre de normalisation du score qui vaut 1 si le score est construit selon un log-rapport de vraisemblance.

– Si $\mathbb{E}[\mathbf{X_i}] = \mathbf{0}$ alors

$$\lim_{n\to\infty} \frac{H_n}{\sqrt{n}} = \sigma B_1^* \tag{3.2}$$

où $B_1^* = \max_{0\leqslant u\leqslant 1} |B_u|$, et B_u est un mouvement brownien standard.
La proposition 2 produite par Daudin et al. (2003) permet le calcul de la fonction de répartition de B_1^* :

$$\mathbb{P}(B_1^* \leqslant x) = \frac{2}{\pi}\sum_{k\in\mathbb{Z}} \frac{(-1)^k}{2k+1}\exp\left(-\frac{(2k+1)^2\pi^2}{8x^2}\right), x \geqslant 0 \tag{3.3}$$

– Si $\mathbb{E}[\mathbf{X_i}] > \mathbf{0}$ alors H_n/n tend presque sûrement vers $\mathbb{E}(s(X_1))$

Daudin et al. (2003) et Etienne (2002) ont également montré la convergence suivante de la distribution de H_n quand **l'espérance du score est "petite"** et que sa variance est finie. Plus précisément, on se place dans le cas d'une famille de score $\left\{\left(s^{(N)}(X_i)\right)_{i\geqslant 1}; N \geqslant 1\right\}$ indépendante et indicée par N. On suppose que $\lim_{N\to\infty}\sqrt{N}\mathbb{E}[s^{(N)}(X_i)] = \delta \in \mathbb{R}$ et $\lim_{N\to\infty}\mathbb{V}[s^{(N)}(X_i)] = \sigma^2 > 0$. Alors,

$$\lim_{N\to\infty} \frac{H_N^{(N)}}{\sqrt{N}} = \sigma\xi_{\delta/\sigma}$$

où $\xi_\gamma = \max_{0\leqslant u\leqslant 1}\{B(u) + \gamma u - \min_{0\leqslant s\leqslant u}(B(s) + \gamma s)\}$. Les auteurs montrent que la queue de la distribution de ξ_γ est approchée par

$$\mathbb{P}(\xi_\gamma \geqslant a) \underset{a\to\infty}{\sim} 2\sqrt{\frac{2}{\pi}}\frac{1}{a}e^{-(\gamma-a)^2/2}$$

Pour terminer cette revue, citons le travail de Mercier et Daudin (Daudin and Mercier, 1999; Mercier and Daudin, 2001) qui calcule la **distribution exacte du score local** dans le cas où la fonction de score est à valeur dans $\mathbb{Z}$. Leur approche s'appuie sur l'égalité suivante :

$$H_n = \max_{0 \leqslant j \leqslant n} U_j$$

avec

$$U_j = (U_{j-1} + s(X_j))^+ \text{ et } U_0 = 0$$

(U_n) est le processus de Lindley associée à celui des $(s(X_n))$. La valeur de U_j est égale à $S_j - \min_{1 \leqslant i \leqslant j} S_i$ où $S_j = \sum_{i=1}^{j} s(X_i)$ est la somme cumulée des scores. En effet, $(S_j - \min_{1 \leqslant i \leqslant j} S_i) = \max_{1 \leqslant i \leqslant j} s(X_{i+1}) + ... + s(X_j)$ et on s'aperçoit facilement qu'il suffit de déterminer la valeur j qui maximise cette quantité pour déterminer H_n. La valeur de H_n s'illustre graphiquement par la plus grande augmentation entre deux points du graphe des S_j. Dans l'algorithme 3.3 de détermination du score local présenté page 52, la valeur de U_j correspond à la variable notée s_j.

Pour calculer la probabilité $\mathbb{P}(H_n < x)$, on s'intéresse au processus U_i tant qu'il n'a pas pris la valeur x. On définit par conséquent le processus arrêté $\tilde{U}_n$ qui est égal à U_n tant que U_n n'a pas atteint la valeur x et qui vaut x sinon.

$\tilde{U}_n$ est une chaîne de Markov à valeur dans $\{0, 1, ..., x\}$. Sa matrice de transition Q de taille $(x+1) \times (x+1)$ s'exprime en fonction de la distribution de $s(X_i)$ de la façon suivante :

$$\begin{aligned}
\mathbb{P}(\tilde{U}_{i+1} = 0 \mid \tilde{U}_i = y) &= \mathbb{P}(s(X_1) \leqslant -y) \qquad \forall y \in \{0, ..., x-1\} \\
\mathbb{P}(\tilde{U}_{i+1} = z \mid \tilde{U}_i = y) &= \mathbb{P}(s(X_1) = y - z) \qquad \forall y \in \{0, ..., x-1\}, \forall z \in \{1, ..., x-1\} \\
\mathbb{P}(\tilde{U}_{i+1} = x \mid \tilde{U}_i = y) &= 1 - \sum_{y=0}^{x-1} \mathbb{P}(\tilde{U}_{i+1} = z \mid \tilde{U}_i = y) \qquad \forall y \in \{0, ..., x-1\} \\
\mathbb{P}(\tilde{U}_{i+1} = x \mid \tilde{U}_i = x) &= 1 \\
\mathbb{P}(\tilde{U}_{i+1} = z \mid \tilde{U}_i = x) &= 0 \qquad \forall z \in \{0, ..., x-1\}
\end{aligned}$$

Comme le processus $\tilde{U}_n$ conserve la valeur x lorsqu'il l'a atteinte, la probabilité $\mathbb{P}(H_n < x)$ est égale à la probabilité $1 - \mathbb{P}(\tilde{U}_n \neq x \mid \tilde{U}_0 = 0)$. Elle s'exprime à l'aide du vecteur P_n de taille $x + 1$:

$$P_n = P_0 Q^n$$

où $P_0 = (1, 0, ..., 0)$ est de longueur x et Q est la matrice de transition de la chaîne de Markov $\tilde{U}$ définie ci-dessus. On a alors

$$\mathbb{P}(H_n \geqslant x) = P_n(x)$$

Remarquons que ce résultat est valable quel que soit le signe de l'espérance du score. Néanmoins, son utilisation pratique peut poser quelques problèmes. En effet, la taille de la

matrice Q sera, en général, très importante étant donné qu'elle est de la taille de la valeur du score local. On a vu par exemple que le score local est presque sûrement proportionnel à la taille de la séquence lorsque l'espérance du score est strictement positive. De plus, les séquences que l'on étudie sont en général de taille n très élevée. Ceci ajoute encore de la complexité pour calculer Q^n. Enfin, il est nécessaire pour utiliser cette approche que la fonction de score soit à valeur dans $\mathbb{Z}$; très souvent, la fonction de score est issue d'un rapport de vraisemblance et est donc à valeurs dans $\mathbb{R}$. Les valeurs doivent dans ce cas être arrondies ce qui n'est pas nécessairement souhaitable. Ce résultat sera donc très utile dans les cas où les résultats asymptotiques ne peuvent être appliqués, mais son utilisation reste pour l'instant limitée en pratique par une grande complexité algorithmique et une occupation mémoire importante.

$\mathbf{s(X_1)s(X_2)...s(X_n)}$ ne sont pas indépendants Il n'y a pas de résultats généraux dans le cas où $s(X_1)s(X_2)...s(X_n)$ ne sont pas indépendants. Certains des résultats précédents sont également valable si $s(X_1)s(X_2)...s(X_n)$ forment une chaîne de Markov. Remarquons ici, que si la séquence $X_1...X_n$ est une chaîne de Markov, alors la séquence de scores associée n'est généralement pas une chaîne de Markov, à moins que la fonction s soit bijective, ou que s vérifie le critère de Dynkin, rappelé ci-dessous.

Proposition 3.2 Critère de Dynkin
Soit (X_n) une chaîne de Markov sur E dénombrable, de matrice de transition P. Soit ϕ une application surjective de E dans F.

$Y_n = \phi(X_n)$ est une chaîne de Markov de matrice de transition Q, si :

$$\forall j \in F, \forall x \in E, \phi(x) = \phi(y) \Rightarrow P(x, \phi^{-1}(j)) = P(y, \phi^{-1}(j)))$$

où $\phi^{-1}(j)$ dénote l'ensemble des antécédants de j. On a alors :

$$\forall i, j \in F, Q(i, j) = P(x, \phi^{-1}(j))$$

où x est un état quelconque de E tel que $\phi(x) = i$.

Dans le cas où l'espérance du score est nulle, le résultat énoncé précédemment sur la distribution asymptotique du score reste valable. La démonstration se trouve dans le thèse de Etienne (2002). Le résultat de Karlin, qui est le plus utilisé, reste également valable (Dembo and Karlin, 1991a) si les scores forment une chaîne de Markov.

Remarquons que des travaux sont en cours pour essayer de montrer que cette limite est encore valable lorsque la séquence initiale $X_1...X_n$ est une chaîne de Markov, et, par conséquent, lorsque $s(X_1)...s(X_n)$ est une fonction d'une chaîne de Markov (communication personnelle de Sabine Mercier).

3.2.3 En pratique : Algorithme de recherche et détermination de la significativité statistique.

Le but de cette partie est de présenter comment déterminer efficacement la valeur du score local H_n ainsi que le segment $X_i...X_j$ qui lui est associé. On traitera également la détermination pratique des constantes K et λ de Karlin pour associer une mesure de significativité à ce score. On considère que l'on observe une séquence de score $c_1...c_n$.

Algorithme de recherche On cherche dans cette séquence le plus grand score local ainsi que le segment* qui lui est associé. Remarquons qu'il n'y a pas une solution unique pour la détermination de ce segment. En effet, considérons par exemple la séquence de score suivante : 4,-5,3,-3,1,2,-2,2,1,5. Après un rapide calcul de toutes les sommes possibles, on s'aperçoit que le score local vaut alors 7 et il y a deux segments qui le réalisent : les segments $[1, 2, -2, 2, 1, 5]$ et $[3, -3, 1, 2, -2, 2, 1, 5]$. Afin, de limiter le nombre de solutions, on choisit arbitrairement de considérer le segment le plus court parmi tous les segments de même score, et qui ne sont pas d'intersection vide. Remarquons qu'il n'y a pas nécessairement unicité de la solution. En effet, il se peut que deux segments disjoints réalisent le meilleur score local ; c'est le cas, par exemple, lorsque ces deux segments sont identiques. Dans cette partie, un segment sera choisi arbitrairement parmi tous les segments qui réalise le score local maximal (s'il y en a plusieurs). On verra dans la suite, le cas de plusieurs segments de scores élevés (voir le paragraphe 3.3), qui permet de contourner ce choix arbitraire.

Trouver le segment de score maximal ainsi que la valeur de score local qui lui est associé est un problème qui se résout en temps linéaire avec la taille de la séquence n (Bates and Constable, 1985). Pour cela, la somme cumulée des scores jusqu'en position i est utilisée : $S_i = \sum_{j=1}^{i} s(X_i)$.

L'algorithme repose sur le fait que l'on peut réécrire le score local de la façon suivante :

$$H_n = \max\{S_i - \min_{1 \leqslant j \leqslant i} S_j\}$$

Il suffit donc de construire itérativement la somme cumulée des scores, tout en conservant à chaque itération le minimum obtenu jusqu'ici et la valeur maximale de la somme cumulée à laquelle est soustraite ce minimum. Les positions respectives des minima et maxima doivent également être conservées.

La valeur de la somme cumulée peut devenir très grande ce qui posera des problèmes lors de l'implémentation de l'algorithme. En effet, la valeur maximale que peut stocker un ordinateur est fixée (par le nombre de bits utilisés pour coder le nombre). Pour contourner ce problème, on utilisera le schéma équivalent de programmation dynamique suivant :

*Une sous-séquence contigue pourra être dénommée par les termes *segment, sous-chaîne, intervalle ou région.*

Algorithme 3.3 (Recherche du score local)

1. *Initialisation :*

$$s_1 = \begin{cases} 0 \ si \ c_1 \leqslant 0 \\ c_1 \ sinon \end{cases}$$

2. *Pour i allant de 2 à n*

$$s_i = \begin{cases} 0 \ si \ s_{i-1} + c_i \leqslant 0 \\ s_{i-1} + c_i \ sinon \end{cases}$$

3. *Conlusion :*
   Le maximum des s_i correspond à la valeur du score local recherché, et le segment lui correspondant se termine à la position du maximum des s_i et commence à la position correspondant à la dernière valeur nulle des s_i avant ce maximum.

Dans cet algorithme, la valeur maximale des s_i est largement inférieure à celle de la somme cumulée des scores, car elle est égale à la valeur de $S_i - \min_{1 \leqslant j \leqslant i} S_j$.

Détermination de la significativité du score local On cherche, dans ce paragraphe à déterminer en pratique la probabilité critique* suivante $p = \mathbb{P}(H_n \geqslant h_n)$ sous un modèle de séquence déterminé a priori (par exemple $M0$, $M1$, etc.).

Nous avons vu dans le pararaphe précédent, un ensemble de résultats asymptotiques sur la loi de H_n quand la taille de la séquence n tend vers l'infini. Nous présentons plus bas une procédure pour déterminer les paramètres de ces lois et ainsi déterminer la loi asymptotique du score local H_n.

Evidemment, ces résultats asymptotiques ne seront pas suffisamment précis si la séquence n'est pas suffisamment longue. Dans ce cas, la faible complexité de l'algorithme permettant de trouver la valeur de H_n (présenté dans le paragraphe précédent) permet d'estimer la loi du score local par Monte-Carlo dans un temps acceptable et avec une précision contrôlable. La procédure est détaillée ci-dessous.

Algorithme 3.4 (Détermination de la loi du score local par Monte-Carlo quand n est petit)
Pour i allant de 1 à N :

1. *Simuler une séquence $x_1x_2...x_n$ selon le modèle désiré*
2. *Déterminer le score local $h_n^{(i)}$ associé à cette séquence*

On peut montrer que la fonction de répartition empirique du N-échantillon $h_n^{(1)}...h_n^{(N)}$ tend vers la fonction de répartition de H_n quand N tend vers l'infini. La probabilité $P(H_n \geqslant h_{obs})$ est alors estimée par la proportion de $h_n^{(i)}$ supérieurs à h_{obs}.

*appelé aussi "degré de signification" ou "p-valeur"

La puissance de calcul devenant de plus en plus grande, le principal inconvénient de la procédure de Monte Carlo, qui consiste en des temps de calcul trop longs, devient de moins en moins rédhibitoire. Cette remarque est d'autant plus valable que cette méthode présente des avantages sur les méthodes paramétriques. Premièrement, elle ne repose pas sur un résultat asymptotique (avec la taille de la séquence) et permet donc d'obtenir la loi de H_n pour la taille de séquence qui nous intéresse effectivement. Le deuxième et principal avantage à mon goût est le fait que la loi de la séquence n'est pas restreinte à un modèle de séquence où les X_i sont indépendants et identiquement distribués, ou bien Markoviens. Il est tout a fait possible d'obtenir la loi du score local sous d'autres modèles de séquences (en considérant un déséquilibre de liaison pour des données de SNPs, ou un modèle de Markov non stationnaire par exemple). Il suffit pour cela de simuler les séquences sous le modèle désiré, la seule limite résidant dans la complexité à générer les données sous ce modèle.

Dans le cas où l'espérance du score est nulle, et les scores $s(X_1), s(X_2), ..., s(X_n)$ sont indépendants et identiquement distribués ou forment une chaîne de Markov, la distribution de H_n ne dépend que de la variance $\mathbb{V}(X_1) = \sigma^2$ (cf. formule 3.2). L'espérance μ et la variance σ^2 du score s'obtiennent sans difficulté, l'alphabet étant généralement de taille réduite :

$$\mu = \sum_{x \in \mathcal{A}} \mathbb{P}(X_1 = x) s(x) \qquad \text{et} \qquad \sigma^2 = \sum_{x \in \mathcal{A}} \mathbb{P}(X_1 = x) s(x)^2 - \mu^2$$

La série 3.3 converge rapidement et il sera alors possible de l'approcher avec une bonne précision en omettant les termes à partir d'un certain rang L^* :

$$\mathbb{P}(B_1^* \leqslant x) \simeq \frac{2}{\pi} \sum_{k=-L}^{L} \frac{(-1)^k}{2k+1} \exp\left(-\frac{(2k+1)^2 \pi^2}{8x^2}\right), x \geqslant 0$$

Dans le cas le plus courant où l'espérance du score est négative, l'approximation de Karlin pourra être appliquée dès que la séquence considérée sera suffisamment longue. Il reste néanmoins à déterminer en pratique la valeur des deux constantes K et λ.

Proposition 3.5 *On considère une suite de variables aléatoires $X_1 ... X_n$ indépendantes et identiquement distribuées à valeurs dans un alphabet $\mathcal{A}$ de taille finie, et une fonction de score s associée d'espérance négative. Si la fonction de score est issue du log-rapport de vraisemblance :*

$$\forall u \in \mathcal{A}, s(u) = \log\left(\frac{\mathbb{P}_1(X_1 = u)}{\mathbb{P}_0(X_1 = u)}\right)$$

où $\mathbb{P}_0$ (resp. $\mathbb{P}_1$) correspond à la loi de probabilité de X_i de l'hypothèse nulle (resp. alternative), alors $\lambda = 1$ est solution de l'équation $\mathbb{E}\left[e^{xs(X_1)}\right] = 1$

*$L \simeq 100$ donne des résultats convenables

Preuve. La preuve est immédiate par simple calcul :

$$\begin{aligned}\mathbb{E}[e^{\lambda s(X_1)}] &= \sum_{u\in\mathcal{A}} \mathbb{P}_0(X_1=u)e^{\lambda s(u)} \\ &= \sum_{u\in\mathcal{A}} \mathbb{P}_0(X_1=u)\left(\frac{\mathbb{P}_1(X_1=u)}{\mathbb{P}_0(X_1=u)}\right)^{\lambda}\end{aligned}$$

$\lambda = 1$ est alors solution de l'équation $\mathbb{E}[e^{\lambda s(X_1)}] = 1$, car

$$\begin{aligned}\mathbb{E}[e^{s(X_1)}] &= \sum_{u\in\mathcal{A}} \mathbb{P}_0(X_1=u)\frac{\mathbb{P}_1(X_1=u)}{\mathbb{P}_0(X_1=u)} \\ &= \sum_{u\in\mathcal{A}} \mathbb{P}_1(X_1=u) \\ &= 1\end{aligned}$$

■

Le paramètre K, quant à lui, n'est pas aussi facile à déterminer. La procédure suivante pourra être appliquée pour estimer K, et λ dans le cas où le score n'est pas un log-rapport de vraisemblance. Cette technique est initialement proposée par Lawless (1982) pour les données temporelles. Elle a ensuite été reprise par Waterman (1995) dans le cadre de la problématique du score local.

Cette technique repose sur la linéarisation de la loi de Gumbel. En effet, d'après la formule 3.1 page 48, on a l'approximation suivante dès lors que n est "grand" :

$$\ln(-\ln(\mathbb{P}(H_n \leqslant y))) \simeq \ln K - \lambda y + \ln n$$

Cette formulation permet d'estimer les paramètres λ et K par Monte-Carlo en un temps fini, quelle que soit la valeur de n, grâce à l'algorithme ci-dessous.

Algorithme 3.6 (Détermination de la loi du score local par Monte-Carlo quand n est grand)

- *Choisir une taille de séquence n_0. On prendra soin de choisir n_0 suffisamment élevé pour que le résultat asymptotique de la loi du score puisse être appliqué sans perdre trop de précision. D'autre part, plus n_0 sera élevé, plus le déroulement de l'algorithme sera coûteux en terme de calcul et d'utilisation mémoire.*
- *Calculer la fonction de répartion empirique $\hat{F}_{n_0}$ de H_{n_0}. Cette étape s'effectue à l'aide de l'algorithme 3.4.*
- *A l'aide d'une régression linéaire, estimer les paramètres a et b (par $\hat{a}$ et $\hat{b}$) de :*

$$\ln(-\ln(\hat{F_{n_0}}(y))) = a + b \times y$$

- *On estime λ et K par $\hat{\lambda} = -\hat{b}$ et $\hat{K} = e^{\hat{a}}/n_0$*

La complexité de cette algorithme est de l'ordre de n_0. Le choix de n_0 est très important, car ce paramètre détermine le rapport entre la vitesse et la précision désirée. En effet, si n_0 est trop faible, la loi de H_{n_0} sera encore trop éloignée de la loi asymptotique

de Karlin, et si n_0 est trop élevé, l'algorithme de détermination du score local, de complexité linéaire avec la taille de la séquence, utilisera des ressources inutilement. La valeur à donner à n_0 dépend de la fonction de score choisie et de la distribution de la séquence $X_1...X_n$. Il est nécessaire de procéder à une détermination empirique de n_0.

Remarque 3.7 *Afin de gagner de la précision sur l'estimation des paramètres, il est important de fixer la valeur de λ à 1 lors de la régression linéaire lorsque la fonction de score est construite par un log-rapport de vraisemblance.*

Néanmoins, la précision sur le paramètre λ n'est pas nécessairement la plus importante, contrairement à ce qu'il est parfois affirmé. En effet, un calcul de dérivée fondé sur

$$F_n(x, K, \lambda) = \exp\left(-Ke^{-\lambda y + \ln n}\right)$$

donne :

$$\frac{dF_n(x, K, \lambda)}{d\lambda} = -yK\frac{dF_n(x, K, \lambda)}{dK}$$

Dès lors que Ky est supérieur à 1, la précision sur le paramètre K prend le pas sur la précision du paramètre λ.

Remarquons également que la précision sur les paramètres dépend plus du choix d'une valeur n_0 suffisante que du nombre d'itérations de Monte-Carlo.

Remarque 3.8 *La distribution du score local maximal est issue de la théorie des valeurs extrêmes. Cette théorie propose une paramétrisation plus générale de la loi du maximum de n variables aléatoires (généralement dénommée la distribution de valeur extrême généralisée*) ; ce modèle possède un paramètre supplémentaire qui donne plus de souplesse à la forme de la densité. Karlin and Dembo (1992) ont montré que la distribution du score local converge vers la loi de Gumbel, utilisée ci-dessus. Néanmoins, la distribution de valeur extrême généralisée pourrait être utilisée lorsque les hypothèses nécessaires ne sont pas respectées, comme ce serait le cas d'une séquence de scores non indépendants, ni Markoviens, par exemple. Une étude de simulation permettrait d'ajuster les paramètres de cette distribution et de vérifier qu'elle s'ajuste correctement au modèle.*

Remarquons également que d'autres méthodes d'estimation des paramètres K et λ permettent d'utiliser mieux l'information contenue dans les données. Ces méthodes ont un coût en terme de complexité algorithmique et de robustesse que nous avons choisi d'éviter.

Remarque 3.9 *Cette méthode de linéarisation est fréquemment utilisée dans le cadre de la comparaison de deux séquences. Récemment, Olsen et al. (1999) ont proposé une autre méthode (Island-method) pour estimer les paramètres K et λ dans ce cadre, de complexité légérement plus élevée. Altschul et al. (2001) ont montré à l'aide de simulation*

*GEV : "Generalized Extreme Value distribution"

qu'elle améliorait de façon significative la précision des estimations de K et λ. Pour une précision fixée, cette méthode sera finalement plus rapide que la méthode de Monte-Carlo classique.

3.3 Cas des deuxième, troisième,..., r^{ième} plus grandes valeurs de score local

3.3.1 Introduction

Dans la plupart des problèmes biologiques, les phénomènes étudiés à l'aide du score local peuvent se répéter plusieurs fois le long d'une séquence. Il peut exister plusieurs régions hydrophobes, ou y avoir plusieurs zones de fortes homologies, par exemple. Finalement, les deuxième, troisième,..., $r^{ième}$ plus grandes valeurs de score local présentent également un intérêt.

Il est nécessaire, ici, de caractériser ce qui sera désigné par le terme "deuxième plus grand score local". En effet, la sous-séquence qui donnera la deuxième plus grande valeur à la somme $s(X_i)+...+s(X_j)$ sera probablement la même que la sous-séquence qui conduit au plus grand score, excepté quelques termes en début ou fin de segment. Ce segment ci n'est pas très intéressant dans la recherche d'une seconde zone de score élevé ayant une pertinence biologique. On définit donc la $k^{ième}$ meilleure sous-séquence comme étant celle qui maximise $s(X_i)+...+s(X_j)$ parmi toute les sous-séquences **disjointes** des $k-1$ meilleures sous-séquences. De plus, pour éviter la multiplicité des résultats, on considère la sous-séquence la plus courte parmi celles qui ont le même score, comme cela était déja le cas pour la meilleure sous-séquence.

Formellement, Ruzzo and Tompa (1999) donnent la définition suivante :

Définition 3.10 Segment de score local maximal
Soit C une séquence non vide de score. Une sous-séquence I est de score maximal dans C si, et seulement si :

1. *Toutes les sous-séquences de I ont un score plus faible,*
   ET
2. *Aucune sur-séquence de I contenu dans C ne satisfait la condition 1.*

On s'intéresse ici à **toutes les sous-séquences de score local maximal** d'une séquence donnée.

Exemple : On considère la séquence de score à valeur dans $\{-1,1\}$ suivante :

position :	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
score :	1	-1	-1	1	1	-1	1	1	-1	-1	1	1	-1	-1	-1	1
score cumulée :	1	0	-1	0	1	0	1	2	1	0	1	2	1	0	-1	0

L'ensemble des scores locaux maximaux I_k est donné par :

	I_1	I_2	I_3	I_4
Position :	$\{1\}$	$\{4,5,6,7,8\}$	$\{11,12\}$	$\{16\}$
Valeur :	1	3	2	1

On peut vérifier que tous ces segments sont en accord avec la définition 3.10 précédente ; Le point 1. est facilement vérifiable :

- Les sous-séquences de $\{11,12\}$ sont de scores 1 et ont par conséquent un score plus faible que celui du segment $\{11,12\}$ qui vaut 2.
- Les sous-séquences de $\{4,5,6,7,8\}$ sont plus nombreuses. La séquence de score associée est $\{1,1,-1,1,1\}$; On vérifie facilement qu'aucune sous-séquence n'a un score plus élevé que 3.

Le point 2 de la définition est plus difficile à vérifier. En ce qui concerne I_2, il s'agit du score local maximal sur toute la séquence, il n'y a donc aucun segment ayant un score plus élevé, et par conséquent, toute sur-séquence de I_2 possédera nécessairement une sous-séquence ayant un score plus élevé. En ce qui concerne les autres segments, ils sont tous bordés par des nombres négatifs. Si on inclut seulement ces nombres négatifs dans le segment, alors la condition 1 ne sera plus respectée. Si on inclut ces nombres négatifs et les nombres positifs suivants (ou précédents, selon le sens de l'élargissement), on s'aperçoit sur notre exemple que la condition 1 ne sera pas respectée.

Inversement, le segment $\{4,5,6\}$, de score 3, n'est pas un segment de score local maximal. En effet, la condition 2 n'est pas respectée, car la sur-séquence $\{4,5,6,7,8\}$ satisfait la condition 1.

3.3.2 Distribution asymptotique des plus grandes valeurs de score local (cas i.i.d.)

Distribution asymptotique du plus grand score local Avant de discuter de la distribution des plus grandes valeurs du score local, nous détaillons la démarche employée par Karlin and Dembo (1992) pour trouver la distribution du score local maximal H_n dans le cas où la séquence sous-jacente suit un modèle M_0 et où l'espérance du score est négative. Ce rappel est l'occasion d'introduire des notations et des notions qui serviront à trouver la distribution des autres plus grandes valeurs du score local.

La principale idée de leur approche est de découper la séquence en blocs déterminés de manière aléatoire, et dont les distributions sont indépendantes et identiquement distribuées. La manière dont ces blocs sont déterminés est présentée dans la suite. La méthode consiste à réécrire le score local sur la séquence entière à l'aide d'une fonction définie sur chacun de ces blocs. De cette façon, le score local s'exprime comme une fonction de variables aléatoires réelles indépendantes et identiquement distribuées ; la loi de ces variables aléatoires est étudiée sur le premier bloc.

On définit la variable aléatoire $C_i = s(X_i)$, où s est la fonction de score choisie, et X_i la variable aléatoire associée à la *ième* lettre de la séquence étudiée.

On note $\forall k \in \mathbb{N}, S_k = \sum_{i=1}^{k} C_i$, les sommes cumulées partielles, et on pose $S_0 = 0$.

On note M le maximum des sommes cumulées partielles $M = \sup_{k \geqslant 0} S_k$

La démarche de Karlin and Dembo (1992) consiste à étudier la distribution de M, pour ensuite en déduire la distribution du maximum des sommes cumulées partielles sur le premier bloc. Cette distribution permet ensuite de caractériser la distribution du plus grand score local.

Les blocs sont définis à partir des **temps de records négatifs** de la marche aléatoire $(S_k)_{k \geqslant 0}$. Les temps de records négatifs sont définis comme étant les temps d'arrêts suivants :

$$T_0 = 0 \text{ et } T_{k+1} = \inf\{i : i > T_k \text{ et } S_i \leqslant S_{T_k}\}$$

Un temps de record correspond à un nouveau minimum dans la marche aléatoire formée par $(S_k)_{k \geqslant 0}$. Dans l'algorithme 3.3 page 52 un temps de record correspond à un s_i nul. Comme l'espérance du score est négative, on a $\mathbb{P}(T_k < \infty) = 1$. Et on note

$$\mu = \mathbb{E}[T_1]$$

Entre deux temps de records successif, on s'intéresse aux variables aléatoires valant le maximum des sommes cumulées partielles sur un bloc :

$$Q_i = \max_{T_i \leqslant k < T_{i+1}} (S_k - S_{T_i})$$

D'autre part, les trajectoires des sommes cumulées partielles entre deux temps de records successifs $\Sigma_k = (S_i, T_k \leqslant i \leqslant T_{k+1})_k$ sont indépendantes et identiquement distribuées. Ceci est dû à l'indépendance des variables aléatoires C_i et au fait qu'elles sont identiquement distribuées. Par conséquent, les maxima de ces trajectoires, Q_i, sont également des variables aléatoires indépendantes et identiquement distribuées.

Le score local maximal obtenu à un certain temps de record s'exprime comme le maximum de variables aléatoires i.i.d :

$$\forall m \geqslant 0, H_{T_m} = \max_{0 \leqslant i \leqslant N_n} Q_i$$

où N_n est le nombre (aléatoire) de temps de records dans la séquence de taille n.

Cette propriété a inspiré l'algorithme 3.3 du score local maximal, car elle entraîne que le début du segment de score local maximal est un temps de record, et que la fin de ce segment correspond à la réalisation du maximum entre ce temps de record et le suivant.

Karlin and Dembo (1992) ont établi une approximation de la queue de la distribution du maximum des sommes cumulées partielles entre deux temps de records :

$$\mathbb{P}(Q_1 > y) \overset{y \to +\infty}{\sim} Ce^{-\lambda y}$$

où λ est l'unique solution de l'équation en x, $\mathbb{E}[e^{-xC_1}] = 1$, et C est une constante dépendant de la loi de C_i. Ils ont également établi que le domaine d'attraction de la loi de Q_i est la loi de Gumbel. Autrement dit, la distribution du maximum des Q_i, c'est à dire H_n, correctement normalisé, converge vers une loi de Gumbel. On note :

$$M_{N_n} = \max_{1 \leqslant i \leqslant N_n} Q_i$$

D'après la théorie des valeurs extrêmes et le résultat de Karlin and Dembo (1992), il existe deux séquences a_k et b_k telles que

$$\lim_{k \to \infty} \mathbb{P}((M_k - b_k)/a_k < x) = e^{-e^{-x}}$$

Proposition 3.11 *En utilisant les notations précédentes, les séquences a_k et b_k suivantes sont telles que :*

$$\lim_{k \to \infty} \mathbb{P}((M_k - b_k)/a_k < x) = e^{-e^{-x}}$$

avec

$$\begin{aligned} a_k &= \frac{1}{\lambda} \\ b_k &= \frac{\ln k + \ln C}{\lambda} \end{aligned}$$

Preuve. La preuve s'effectue par un simple calcul.

$$\begin{aligned} \mathbb{P}((M_k - b_k)/a_k < (y - b_k)/a_k)) &= \mathbb{P}(M_k < y) \\ &= (1 - \mathbb{P}(Q_1 > y))^k \end{aligned}$$

On pose $x = (y - b_k)/a_k$, c'est à dire $x = \lambda y - \ln k - \ln C$, d'où $y = (x + \ln k + \ln C)/\lambda$. On a donc

$$\begin{aligned} \mathbb{P}((M_k - b_k)/a_k < x) &= \left(1 - \mathbb{P}\left(Q_1 > \frac{x + \ln k + \ln C}{\lambda}\right)\right)^k \\ &\sim \left(1 - C \exp\left(-\lambda \frac{x + \ln k + \ln C}{\lambda}\right)\right)^k \\ &= \left(1 - \frac{\exp(-x)}{k}\right)^k \\ &\overset{k \to \infty}{\to} \exp(-e^{-x}) \end{aligned}$$

■

Cette proposition donne les constantes de normalisation en fonction du nombre de temps de records. Il reste à trouver ces constantes en fonction de la longueur de la

séquence. On utilise pour cela la convergence presque sûre de N_n/n vers $1/\mu$. On peut donc affirmer que N_n se comporte presque sûrement comme n/μ.

Cette observation conduit aux constantes de normalisations suivantes pour H_n :

Proposition 3.12 *En utilisant les notations précédentes, les séquences a_n et b_n suivantes sont telles que :*

$$\lim_{n\to\infty} \mathbb{P}((H_n - b_n)/a_n < x) = e^{-e^{-x}}$$

avec

$$\begin{aligned} a_n &= \frac{1}{\lambda} \\ b_n &= \frac{\ln n + \ln(C/\mu)}{\lambda} \end{aligned}$$

Preuve. La preuve, présentée dans Mercier (1999), s'appuie sur deux résultats. Le premier est la convergence presque sûre de N_n/n.

$$\lim_{n\to\infty} \frac{N_n}{n} \overset{p.s.}{=} \frac{1}{\mu}$$

Le deuxième est le fait que H_n est croissante est en n donc :

$$\mathbb{P}(H_{T_{N_{n+1}}} \leqslant x) \leqslant \mathbb{P}(H_n < x) \leqslant \mathbb{P}(H_{T_{N_n}} \leqslant x)$$

On considère $\varepsilon > 0$, et la majoration suivante :

$$\begin{aligned} \mathbb{P}(H_{T_{N_n}} \leqslant x) &= \mathbb{P}(H_{T_{N_n}} \leqslant x; n(1/\mu - \varepsilon) \leqslant N_n \leqslant n(1/\mu + \varepsilon)) \\ &\quad + \mathbb{P}(H_{T_{N_n}} \leqslant x; |N_n/n - 1/\mu| > \varepsilon) \\ &\leqslant \mathbb{P}(H_{T_{\lceil n(1/\mu-\varepsilon)\rceil}} \leqslant x) + \mathbb{P}(H_{T_{N_n}} \leqslant x; |N_n/n - 1/\mu| > \varepsilon) \\ &= (\mathbb{P}(Q_1 \leqslant x))^{\lceil n(1/\mu-\varepsilon)\rceil} + o(1) \end{aligned}$$

car pour tout $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|N_n/n - 1/\mu| > \varepsilon) = 0$$

De manière similaire, on obtient la minoration suivante :

$$\mathbb{P}(H_{N_{n+1}} \leqslant x) \geqslant (\mathbb{P}(Q_1 \leqslant x))^{\lfloor n(1/\mu+\varepsilon)\rfloor+1} - o(1)$$

Finalement,

$$G_n(\varepsilon) \leqslant \mathbb{P}(H_n \leqslant x) \leqslant D_n(\varepsilon)$$

avec

$$G_n(\varepsilon) = (\mathbb{P}(Q_1 \leqslant x))^{\lfloor n(1/\mu+\varepsilon)\rfloor+1} - o(1)$$

et

$$D_n(\varepsilon) = (\mathbb{P}(Q_1 \leqslant x))^{\lceil n(1/\mu-\varepsilon)\rceil} + o(1)$$

On pose

$$\begin{aligned} x &= a_n z + b_n \\ &= \frac{1}{\lambda}(z + \ln n + \ln(C/\mu)) \end{aligned}$$

Et on a, à l'aide du résultat sur la queue de distribution de Q_1 :

$$\begin{aligned} G_n(\varepsilon) &\overset{n\to\infty}{\sim} \left(1 - C\exp\left(-\lambda\frac{1}{\lambda}(z + \ln n + \ln(C/\mu))\right)\right)^{\lfloor n(1/\mu+\varepsilon)\rfloor+1} \\ &= \left(1 - \frac{\mu e^{-z}}{n}\right)^{\lfloor n(1/\mu+\varepsilon)\rfloor+1} \\ &\overset{n\to\infty}{\sim} \left(1 - \frac{\mu e^{-z}}{n}\right)^{n(1/\mu+\varepsilon)} \\ &\overset{n\to\infty}{\sim} \exp\left(-(1+\varepsilon\mu)e^{-z}\right) \end{aligned}$$

Donc

$$\lim_{n\to\infty} G_n(\varepsilon) = \exp\left(-(1+\varepsilon\mu)e^{-z}\right)$$

De la même façon :

$$\lim_{n\to\infty} D_n(\varepsilon) = \exp\left(-(1-\varepsilon\mu)e^{-z}\right)$$

On a donc :

$$\lim_{n\to\infty} G_n(\varepsilon) \leqslant \mathbb{P}\left(H_n \leqslant \frac{1}{\lambda}(z + \ln n + \ln(C/\mu))\right) \leqslant \lim_{n\to\infty} D_n(\varepsilon)$$

Or

$$\lim_{n\to\infty,\varepsilon\to 0} G_n(\varepsilon) = \lim_{n\to\infty,\varepsilon\to 0} D_n(\varepsilon) = \exp\left(-e^{-z}\right)$$

Nous avons donc montré que les constantes de normalisation, définies dans la proposition, conduisent à la convergence de la loi score local maximal vers la loi de Gumbel. ■

Ce résultat est lié à la convergence du maximum des Q_i vers la loi de Gumbel. La qualité de l'approximation de la loi de H_n par sa loi limite, dans le cas usuel de séquences de tailles finies, dépendra par conséquent du nombre de temps de records. Un indicateur intéressant est la valeur de l'espérance de ce nombre : n/μ, dont un estimateur non biaisé est N_n.

Distribution asymptotique conjointe des plus grandes valeurs de score local
Nous proposons dans ce paragraphe de déterminer la distribution asymptotique conjointe des r plus grandes valeurs de score local, telles qu'elles sont définies page 56 (définition 3.10). On se place dans le cas où les lettres de la séquence sous-jacente sont indépendantes et identiquement distribuées, et dans le cas le plus fréquent où :

$$\mathbb{E}[C_1] < 0 \text{ et } \mathbb{P}(C_1 > 0) > 0$$

avec $C_i = s(X_i)$, où s est la fonction de score choisie, et X_i la variable aléatoire associée à la *ième* lettre de la séquence étudiée.

Notation : Si l'on considère une suite de variables aléatoires, $Z_1, ..., Z_n$, le maximum de cette suite sera noté $Z^{(1)} = \max_{i=1,...,n} Z_i$, la deuxième plus grande valeur sera notée $Z^{(2)}$, etc*. Dans la suite, nous considérerons :

- $H_n^{(1)} \geqslant ... \geqslant H_n^{(r)}$: les r plus grandes valeurs de score local.
- $Q_n^{(1)} \geqslant ... \geqslant Q_n^{(r)}$: les r plus grandes valeurs des maxima des sommes cumulées partielles entre deux temps de records.

Nous avons vu, dans le paragraphe précédent, que le score local maximal est le maximum de quantités indépendantes et identiquement distribuées $(Q_1, ..., Q_n)$. La loi de ce maximum, normalisée par les constantes a_n et b_n définies dans la proposition 3.12 page 60, converge vers la loi de Gumbel. Pour déterminer la distribution asymptotique des plus grandes valeurs de score local, on utilisera un résultat connu sur la distribution conjointe asymptotique des r plus grandes valeurs d'un échantillon (Weissman, 1978; Smith, 1986; Coles, 2001) :

Proposition 3.13 *Soit $Y_1 ... Y_n$ une suite de variables aléatoires indépendantes et identiquement distribuées. On note $M_n^{(k)}$ la $k^{ième}$ plus grande valeur. On se place dans le cas où $M_n^{(1)}$ converge vers une loi de Gumbel : $\lim_{n\to\infty} \mathbb{P}(M_n^{(1)} < a_n x + b_n) = \exp(-\exp(-(x-\mu)/\sigma))$ où a_n et b_n sont deux suites de constantes dépendantes de la loi de Y_i. La densité jointe des r plus grandes valeurs, correctement normalisées par a_n et b_n*

$$\left(\frac{M_n^{(1)} - b_n}{a_n}, ..., \frac{M_n^{(r)} - b_n}{a_n} \right)$$

est donnée par :

$$L(\bullet) = \exp\left\{ -\exp\left(-\frac{M_n^{(r)} - \mu}{\sigma} \right) \right\} \prod_{i=1}^{r} \sigma^{-1} \exp\left(-\frac{M_n^{(i)} - \mu}{\sigma} \right)$$

Une remarque importante concernant cette densité est que les constantes de normalisation a_n et b_n sont les mêmes pour toutes les statistiques (maximum, deuxième plus grande valeur, etc.). Cette observation justifie la remarque ci-dessous.

*Attention : cette notation n'est pas celle utilisée généralement dans les documents traitant des statistiques d'ordre, dans lesquels $Z^{(1)}$ désigne usuellement le minimum de l'échantillon.

Remarque 3.14 *La distribution conjointe des r plus grandes valeurs étant connue ,il est possible d'améliorer l'estimation des constantes de Karlin K et λ (voir l'algorithme 3.6 page 54) en ne se fondant pas uniquement sur un échantillon des valeurs maximales, mais en considérant également l'information contenue dans les r plus grandes valeurs des Q_i. Smith (1986) a développé une méthode d'estimation par maximum de vraisemblance utilisant cette remarque. Dupuis (1997) a montré que cet estimateur manquait de robustesse et propose un estimateur biais-robuste optimal* (Hampel et al., 1986).*

Le problème de la distribution conjointe des plus grandes valeurs de score local serait résolu à ce point, si la $r^{ième}$ plus grande valeur de score local correspondait à la $r^{ième}$ plus grande valeur de l'échantillon i.i.d. $Q_1, ..., Q_{N_n}$. Ce n'est pas le cas, car deux plus grandes valeurs de score local peuvent se trouver dans la même excursion entre deux temps de records successifs.

Exemple : On considère la série de scores élémentaires suivante : 1, -5, 8, -3, 6, -8, 7, -9, 3, -10. Cette série est évidemment un court exemple permettant d'illuster que le premier et le deuxième plus grand score local peuvent se trouver dans la même excursion (c'est à dire entre deux temps de records successifs). On peut imaginer que ce segment soit extrait d'une séquence plus longue.
Le tableau ci-dessous présente les scores, les scores cumulés permettant de trouver facilement (et visuellement) les plus grandes excursions, ainsi que les variables $s_i = S_i - \min_{j \leqslant i} S_j$ utilisées dans l'algorithme de détermination du score local maximal (algorithme 3.3 page 52).

i	1	2	3	4	5	6	7	8	9	10
Scores initiaux c_i	1	-5	8	-3	6	-8	7	-9	3	-10
Scores cumulés S_i	1	-4	4	1	7	-1	6	-3	0	-10
s_i	1	0	8	5	11*	3	10	1	4	0

On rappelle qu'une valeur nulle de s_i correspond à un temps de record. Il y a donc deux temps de records ici (différents $T_0 = 1$) : $T_1 = 2$ et $T_2 = 10$. L'algorithme de recherche du plus grand score local donne le segment $\{3, 4, 5\}$ de score 11. La deuxième plus grande valeur de score local vaut 7 et est obtenue par le segment constitué d'un seul élément : $\{7\}$. On peut vérifier que ce deuxième segment est bien en accord avec la définition 3.10 présentée page 56. Un algorithme de détection systématique des plus grandes valeurs de score local ainsi que des segments associés est donné dans la suite du document (algorithme 3.10 page 56).

La figure 3.1 page 64 représente les scores cumulés ainsi que les temps de records et les plus grandes valeurs de score local. Sur cette courbe, les temps de records correspondent à un nouveau minimum. Les segments, correspondant aux trois plus grandes valeurs de score local, se situent tous entre les mêmes deux temps de records consécutifs.

*"Optimal Bias-Robust Estimator" pour Shakespeare.

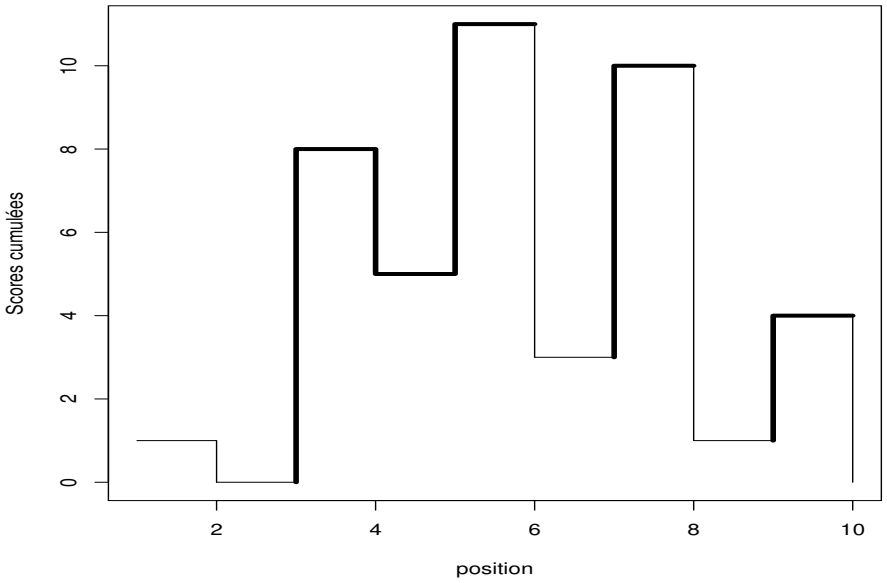


FIG. 3.1 – Représentation des scores cumulés de l'exemple donné page 63 ; les lignes dessinées en gras correspondent au plus grandes valeurs de score local. Les temps de records se trouvent en position 2 et 10.

Les r plus grandes valeurs de score local ne sont donc pas les r plus grandes valeurs d'un échantillon i.i.d.. On s'intéresse ici à la distribution conjointe de ces valeurs lorsque la taille de la séquence (et par conséquent le nombre de temps de records) tend vers l'infini. Dans ce cas, la probabilité qu'un nombre fini de plus hautes valeurs de score local soient issues de segments contenus dans la même excursion (entre deux temps de records successif) tend vers 0. Cette remarque nous permettra de conclure que la distribution asymptotique conjointe des r (avec r fini) plus grandes valeurs de score local se comporte comme la distribution des r plus grandes valeurs d'un échantillon i.i.d.

Proposition 3.15 *Soit $C_1, ..., C_n$ une séquence de scores i.i.d.*
Soit $H_n^{(1)} \geqslant ... \geqslant H_n^{(r)}$ les r plus grandes valeurs de score local telles qu'elles sont définies à la page 56 (Définition 3.10) correspondant à cette séquence. On note

$$\left[\frac{H_n^{(1)} - b_n}{a_n}, ..., \frac{H_n^{(r)} - b_n}{a_n}\right]$$

la densité jointe de $H_n^{(1)}, ..., H_n^{(r)}$. On a :

$$\lim_{n\to\infty}\left[\frac{H_n^{(1)} - b_n}{a_n}, ..., \frac{H_n^{(r)} - b_n}{a_n}\right] = \exp\left\{-\exp\left(-H_n^{(r)}\right)\right\}\prod_{i=1}^{r}\exp\left(-H_n^{(i)}\right)$$

avec

$$a_n = \frac{1}{\lambda} \text{ et } b_n = \frac{\ln n + \ln(C/\mu)}{\lambda}$$

Preuve. On rappelle les notations utilisées précédemment, on introduit H' et $\Sigma^{(k)}$:
- $H_n^{'(k)} = (H_n^{(k)} - b_n)/a_n$ le score local normalisé,
- $S_i = \sum_{j=1}^{i} S_i, i = 1, ..., n$, les sommes cumulées partielles,
- $T_k, k = 1, ..., N_n$ les temps de records négatifs de la série $\{S_i\}_{i=1,...,n}$,
- $\Sigma_k = (S_i, T_k \leqslant i \leqslant T_{k+1})_k, k = 1, ..., N_n - 1$ les excursions entre deux temps de records successifs.
- $\Sigma^{(k)}$ l'excursion contenant le segment qui réalise $H_n^{(k)}$.

On décompose la densité en deux termes selon que les r plus grandes valeurs de score local sont toutes issues d'excursions différentes ou non :

$$\begin{aligned}\left[H_n^{'(1)}, ..., H_n^{'(r)}\right] &= \left[H_n^{'(1)}, ..., H_n^{'(r)} \mid \exists 1 \leqslant i < j \leqslant N_n, \Sigma^{(i)} = \Sigma^{(j)}\right] \\ &\quad \left[\exists 1 \leqslant i < j \leqslant N_n, \Sigma^{(i)} = \Sigma^{(j)}\right] \\ &\quad + \left[H_n^{'(1)}, ..., H_n^{'(r)} \mid \forall 1 \leqslant i < j \leqslant N_n, \Sigma^{(i)} \neq \Sigma^{(j)}\right] \\ &\quad \left[\forall 1 \leqslant i < j \leqslant N_n, \Sigma^{(i)} \neq \Sigma^{(j)}\right]\end{aligned}$$

Or, r est fini, et $C_1, ..., C_n$ sont indépendantes, donc :

$$\lim_{n\to\infty} \left[\exists 1 \leqslant i < j \leqslant N_n, \Sigma^{(i)} = \Sigma^{(j)}\right] = 0.$$

Autrement dit :

$$\lim_{n\to\infty} \left[\forall 1 \leqslant i < j \leqslant N_n, \Sigma^{(i)} \neq \Sigma^{(j)}\right] = 1$$

et

$$\lim_{n\to\infty} \left[H_n'^{(1)}, ..., H_n'^{(r)} \mid \exists 1 \leqslant i < j \leqslant N_n, \Sigma^{(i)} = \Sigma^{(j)}\right] < \infty.$$

Finalement :

$$\lim_{n\to\infty} \left[H_n'^{(1)}, ..., H_n'^{(r)}\right] = \lim_{n\to\infty} \left[H_n'^{(1)}, ..., H_n'^{(r)} \mid \forall 1 \leqslant i < j \leqslant N_n, \Sigma^{(i)} \neq \Sigma^{(j)}\right].$$

Or,

$$\forall 1 \leqslant i < j \leqslant N_n, \Sigma^{(i)} \neq \Sigma^{(j)} \Rightarrow \forall 1 \leqslant i \leqslant r, H_n^{(i)} = Q_n^{(i)}.$$

Donc,

$$\lim_{n\to\infty} \left[H_n^{(1)}, ..., H_n^{(r)}\right] = \lim_{n\to\infty} \left[Q_n^{(1)}, ..., Q_n^{(r)}\right].$$

Comme $Q_1, ..., Q_{N_n}$ est un échantillon i.i.d., la proposition ci-dessus est prouvée. ■

3.3.3 Algorithme de recherche

La méthode la plus couramment utilisée pour trouver le $k^{ième}$ score local le plus élevé consiste à appliquer successivement l'algorithme de recherche du meilleur score, en supprimant les segments correspondant aux meilleurs scores déjà détectés. Partant de la séquence initiale, le segment de plus haut score est recherché et la valeur de score associée est $h_1^{(n)}$. Ensuite, le deuxième meilleur score est recherché parmi les deux séquences obtenues après la suppression du meilleur score. Cette opération est itérée jusqu'à l'obtention du nombre de segments désiré ou jusqu'à ce qu'il ne reste que des scores négatifs.

Dans le pire des cas, cet algorithme possède une complexité de l'ordre de n^2 où n est la taille de la séquence. Ce cas correspond à la situation où les segments détectés sont à chaque fois de taille 1. On montre ci-dessous que cet algorithme possède une complexité moyenne de l'ordre de $n \log n$ dans le cas où l'espérance est négative, et les éléments de la séquence sous-jacente sont distribués de manière indépendante et identique.

Le fait que l'espérance soit négative nous permet de négliger la longueur du segment détecté par rapport à la longueur de la séquence. Une itération de l'algorithme va donc couper la séquence en 2 parties. La question de la complexité de l'algorithme se ramène en ces termes : combien va-t-il falloir d'itérations pour n'avoir plus que des séquences de taille 1 ? La réponse est k tel que $n = 2^k$, en supposant que la séquence soit d'une taille qui soit une puissance de 2. On en déduit que $k = \ln_2(n)$, où $\ln_2$ est le logarithme en base 2. Etant donné d'autre part que la recherche du segment de score maximal est de

complexité $O(n_i)$ pour une sous-séquence de taille n_i, la complexité de la recherche à une étape donnée vaut $\sum_i O(n_i) = O(n)$ avec $n = \sum_i n_i$. Finalement, la complexité globale moyenne vaut : $kO(n)$ soit $O(n \ln n)$.

Ruzzo and Tompa (1999) proposent un algorithme de recherche de tous les scores locaux positifs présents dans une séquence dans un temps linéaire*. Nous présentons cet algorithme ci-dessous. Les scores sont lus de gauche à droite. A chaque itération (nouveau score lu), une liste de segments disjoints notés $I_1, ..., I_k$ évolue. A chacun de ces segments I_j est associée la somme cumulé L_j de tous les scores précédants ce segment ; est également associée la somme cumulée R_j de tous les scores du début de la séquence jusqu'à la fin du segment incluse.

Algorithme 3.16 *La liste est initialement vide. Seules les scores positifs sont inclus dans un nouvel I_k, qui est ensuite traité par la procédure suivante :*

1. *La liste est parcourue de droite à gauche jusqu'à ce que $L_j < L_k$.*
2. *S'il n'y a pas de tel j, alors le segment I_k est ajouté en fin de liste.*
3. *S'il y a un tel j, et $R_j \geqslant R_k$, alors le segment I_k est ajouté en fin de liste.*
4. *Sinon :*
   - *Le segment I_k est étendu sur la gauche jusqu'à ce qu'il contienne le segment I_j.*
   - *Les segments numérotés de j à $k-1$ sont éliminés de liste, et le segment I_k est renommé I_j.*
   - *Ce segment I_j subit à nouveau le processus à l'étape 1.*

Exemple :
On considère la série de scores élémentaires suivante :

i	1	2	3	4	5	6	7	8	9
Scores initiaux c_i	1	-5	8	-3	6	-8	7	-9	3
Scores cumulés S_i	1	-4	4	1	7	-1	6	-3	0
s_i	1	0	8	5	11*	3	10	1	4

La deuxième ligne du tableau correspond aux sommes cumulées $S_k, k = 1, ..., 9$. La troisième ligne du tableau correspond aux s_i tel qu'ils sont calculés dans l'algorithme proposé pour trouver la plus grande valeur de score local (voir algorithme 3.3 page 52). Le symbole "*" désigne la valeur maximale, et le segment de score local maximal sera le segment $(8, -3, 6)$ dont le score local vaut 11. Dans cet exemple simple, on s'aperçoit, "à l'oeil", que le deuxième score local positif sera donné par le segment (7), ensuite (3) et enfin (1).

Déroulons maintenant l'algorithme 3.16. A chaque étape, on présente les segments contenus dans la liste ; chaque segment I_i est décrit par sa position dans la séquence ([

*Cet algorithme a été implémenté en C et est disponible sur demande auprès des auteurs

début ; fin]), la somme cumulée avant le début du segment L_i, et la somme cumulée jusqu'à la fin du segment incluse R_i.

On lit le premier nombre $c_1 = 1$

$$I_1 \quad \{1\} \quad L_1 = 0 \quad R_1 = 1$$

c_2 est négatif, il n'est pas traité.
On traite ensuite $c_3 = 8$

$$\begin{array}{llll} I_1 & \{1\} & L_1 = 0 & R_1 = 1 \\ I_2 & \{3\} & L_2 = -4 & R_2 = 4 \end{array}$$

Il n'existe pas dans la liste L_i d'élément inférieur à L_2. On passe donc au nombre positif suivant : $c_5 = 6$

$$\begin{array}{llll} I_1 & \{1\} & L_1 = 0 & R_1 = 1 \\ I_2 & \{3\} & L_2 = -4 & R_2 = 4 \\ I_3 & \{5\} & L_3 = 1 & R_3 = 7 \end{array}$$

L_2 est inférieur strictement à L_3, et R_2 est supérieur à R_3, les segments I_2 et I_3 sont par conséquent fusionnés, ce qui donne :

$$\begin{array}{llll} I_1 & \{1\} & L_1 = 0 & R_1 = 1 \\ I_2 & [3;5] & L_2 = -4 & R_2 = 7 \end{array}$$

Comme L_2 n'est pas inférieur à L_1, on passe à l'élément positif suivant, $c_7 = 7$:

$$\begin{array}{llll} I_1 & \{1\} & L_1 = 0 & R_1 = 1 \\ I_2 & [3;5] & L_2 = -4 & R_2 = 7 \\ I_3 & \{7\} & L_3 = -1 & R_3 = 6 \end{array}$$

L_2 est inférieur strictement à L_3, mais R_2 est supérieur à R_3. Par conséquent, I_3 reste en fin de liste.
On traite maintenant le dernier nombre positif $c_9 = 3$.

$$\begin{array}{llll} I_1 & \{1\} & L_1 = 0 & R_1 = 1 \\ I_2 & [3;5] & L_2 = -4 & R_2 = 7 \\ I_3 & \{7\} & L_3 = -1 & R_3 = 6 \\ I_4 & \{9\} & L_4 = -3 & R_4 = 0 \end{array}$$

Le même cas de figure est obtenu, c'est à dire que L_2 est inférieur strictement à L_4, mais R_2 est supérieur à R_4. Donc I_4 reste inchangé.

L'algorithme se termine, car plus aucun score positif n'est disponible. La liste de tous les scores locaux positifs de la séquence initiale est donnée par les segments I_1 à I_4. On retrouve bien les segments que l'on avait prévu "à l'oeil".

Cet algorithme, qui est l'algorithme initial proposé par Ruzzo and Tompa (1999), peut être optimisé de manière conséquente. En effet, si l'étape 2 est appliquée alors tous les segments $I_1, ..., I_{k-1}$ sont maximaux. Ils peuvent être retirés de la liste de recherche pour être placés dans une autre liste contenant les segments définitivement déclarés de score maximal. On remplace donc l'étape 2 par l'étape 2′ :

2' S'il n'y a pas de tel j, alors tous les segments $I_1, ..., I_{k-1}$ sont maximum. Ces segments sont alors déplacés dans la liste M des segments de score maximal, et la liste de "travail" est réinitialisée pour ne contenir que I_k (renommé maintenant I_1).

La recherche de l'étape 1 peut, elle aussi, être accélérée. Pour cela, les segments I_k ajoutés à l'étape 3 doivent être liés (par un pointeur) à la sous-séquence I_j résultant de l'étape 1. Cette opération fournira une liste de segments dont les valeurs L_j sont décroissantes. La recherche de l'étape 1 sera donc effectuée sur cette liste-ci au lieu de la liste complète.

Remarque 3.17 *La liste des scores locaux donnée par l'algorithme n'est pas triée par valeur de score local (croissante ou décroissante).*

Remarque 3.18 *La valeur du score local d'un segment I_i donné est $R_i - L_i$.*

3.3.4 Détermination de la significativité des plus grands scores locaux

La seule donnée des scores locaux et des segments qui leur sont associés n'est pas suffisante. En effet, on cherche, ici encore, à associer à ces valeurs un critère d'exceptionnalité par rapport au modèle supposé. Pour cela, la significativité statistique du résultat observé doit être éprouvée.

A notre connaissance, seul un article (Karlin and Altschul Stephen, 1993) dans la littérature considère le problème des r plus grandes valeurs de score local. Les auteurs considèrent la somme des r plus grands score locaux, dont ils explicitent la loi asymptotique, ainsi qu'une approximation de cette loi. Ils considèrent ensuite comme indice de significativité la probabilité d'observer une valeur de cette somme au moins aussi élevée que celle observée sur l'échantillon. Outre le fait que cette méthode nécessite de fixer le nombre r de segments considérés a priori, elle présente également l'inconvénient de "noyer" l'effet d'un segment exceptionnel parmi les $r-1$ segments sous H_0 dans le cas où r est mal choisi. En effet, la statistique considérée est une somme (ou de façon équivalente, une moyenne) des scores locaux, où tous les scores locaux ont le même poids. Par conséquent, imaginons qu'il n'y ait en réalité qu'une seule zone sous H_1 et que les $r-1$ autres zones soient sous H_0, alors la distribution de la statistique de test sera beaucoup plus proche de sa loi sous H_0 (avec un poids $r - 1/r$) que de sa loi sous H_1 (avec un poids $1/r$).

La méthode habituellement pratiquée consiste à considérer la loi du maximum des scores H_n, tel que cela a été effectué dans le paragraphe précédent. Ensuite, un seuil de significativité s est déterminé en fonction du risque de première espèce α désiré, de la façon suivante : $\mathbb{P}(H_n \geqslant s) \leqslant \alpha$. Finalement, tous les scores locaux $H_n^{(r)}$ qui dépassent ce seuil seront considérés comme significatifs au risque α.

Cette méthode présente l'avantage de préserver le risque de première espèce du test :
H_0 : La séquence est i.i.d. (*il n'y a pas "d'accumulation"*)
contre

H_1 : La séquence n'est pas i.i.d. (*il y a au moins une "accumulation"*)
Ce résultat est inhabituel, car un problème de test multiple semble apparaître. En effet, plusieurs tests sont effectués au risque α, ce qui généralement entraîne un risque de première espèce global supérieur à α. Le risque global se définit comme étant la probabilité de rejeter H_0 alors que H_0 est vraie, ou, autrement dit 1 moins la probabilité d'accepter H_0 lorsque H_0 est vraie. Dans notre cas, cette dernière probabilité vaut :

$$\begin{aligned}\mathbb{P}(\text{accepter } H_0 \mid H_0) &= \mathbb{P}(H_n^{(1)} \leqslant s, H_n^{(2)} \leqslant s, ... \mid H_0) \\ &= \mathbb{P}(H_n^{(1)} \leqslant s \mid H_0) \\ &= 1-\alpha\end{aligned}$$

Ce cas particulier est dû au fait que les scores locaux sont ordonnés, donc

$$\left\{H_n^{(1)} \leqslant s\right\} \Rightarrow \left\{H_n^{(i)} \leqslant s\right\}_{i>1}$$

Plus formellement, l'hypothèse nulle ci-dessus peut-être écrite comme l'intersection de plusieurs hypothèses nulles intermédiaires correspondant chacune à la statistique de test $H_n^{(i)}$. En effet, on désire savoir si $H_n^{(i)}$ est "trop" élevé ou non, par rapport à ce qui est attendu dans le cas où la séquence est indépendante et identiquement distribuée, et cela, indépendamment du fait que $H_n^{(i-1)}$ le soit ou non. Autrement dit, pour l'hypothèse nulle $H_0^{(i)}$, on considérera la séquence privée des segments qui ont réalisé les scores locaux $H_n^{(j)}$, pour tout $j < i$.
On note $\Sigma'^{(k)}$ le **segment** qui réalise $H_n^{(k)}$ (à ne pas confondre avec $\Sigma^{(k)}$ qui désignait **l'excursion** contenant ce segment).
On définit les hypothèses nulles et alternatives suivantes, pour tout $i > 2$:

$H_0^{(i)}$: la séquence privée de $\Sigma'^{(k)}, k < i$ est indépendante et identiquement distribuée.
$H_1^{(i)}$: la séquence privée de $\Sigma'^{(k)}, k < i$ n'est pas indépendante et identiquement distribuée.
et
$H_0^{(1)}$: la séquence est indépendante et identiquement distribuée.
$H_1^{(1)}$: la séquence n'est pas indépendante et identiquement distribuée.

Ces différentes hypothèses nulles ont une structure particulière dans le sens où elles sont emboitées de la façon suivante :

$$H_0^{(1)} \Rightarrow H_0^{(2)} \Rightarrow H_0^{(3)} \Rightarrow \dots$$

En effet, toute séquence vérifiant $H_0^{(i)}$ vérifie nécessairement $H_0^{(j)}, j > i$, mais l'inverse n'est pas vrai.

Cette structure impose une démarche particulière de test afin de ne pas conduire à des conclusions incohérentes. Par exemple, il serait tout à fait génant de ne pas rejeter $H_0^{(1)}$ et de rejeter $H_0^{(2)}$.

La démarche consistant à calculer le seuil de significativité s à partir de la loi du score local maximal, et à considérer les scores dépassant ce seuil peut se formuler de la manière suivante :

Algorithme 3.19 (Démarche habituelle de détermination des scores locaux significatifs) *On note s^α le seuil obtenu en se fondant sur la distribution du score local maximal, lorsque la séquence est i.i.d.*

- $i \leftarrow 1$
- *Tant que* $H_n^{(i)} > s^\alpha$,
  - *on rejette* $H_0^{(i)}$
  - $i \leftarrow i + 1$
- *On ne rejette pas les hypothèses :* $\forall j \geqslant i, H_0^{(j)}$

Rappelons les notions de **cohérence** et **consonance** d'une démarche hiérarchique dans le cadre des tests multiples. Ces notions sont dues à Gabriel (1969) et reprise dans Hochberg and Tamhane (1987).

Définition 3.20 (Cohérence et Consonance d'une démarche hiérarchique dans le cadre des tests multiples)

Une démarche est dite **cohérente** *si l'acceptation d'une hypothèse nulle entraîne l'acceptation de toutes les sous-hypothèses qu'elle implique.*

Une démarche est dite **consonante** *si le rejet d'une hypothèse nulle entraîne le rejet d'au moins une des sous-hypothèses qu'elle implique.*

Proposition 3.21 *La démarche proposée est cohérente, mais pas consonante.*

Preuve. Les sous-hypothèses impliquées par l'hypothèse $H_0^{(i)}$ sont toutes les hypothèses $\{H_0^{(j)}, j > i\}$.

Consonance

Dans notre cas, le rejet de l'hypothèse $H_0^{(i)}$ ne garantit pas qu'au moins une des suivantes le sera également. La démarche n'est donc pas consonante*.

Cohérence

Par contre, le non rejet de l'hypothèse nulle $H_0^{(i)}$ entraîne effectivement l'acceptation de toutes les hypothèses nulles impliquées par $H_0^{(i)}$, par construction de la démarche ■

Cette démarche ne serait pas cohérente, s'il était possible de ne pas rejeter $H_0^{(3)}$ par exemple, mais de rejeter $H_0^{(4)}$, ce qui n'est pas le cas ici.

*La notion de consonance n'est pas vraiment adaptée à ce type de problème. Elle convient mieux lorsque le rejet de l'hypothèse nulle induit plusieurs nouvelles hypothèses nulles possibles. C'est le cas dans un test d'égalité d'espérances de plusieurs variables comme en analyse de variance, par exemple. Dans ce cas, après avoir rejeté l'égalité de k espérances, on se demandera ce qu'il en est des différentes façons de comparer $k-1$ parmi les k

Néanmoins, elle possède l'inconvénient de ne pas être optimum en terme de puissance. En effet, on comprend intuitivement que le seuil associé à la deuxième plus grande valeur devrait être plus faible que le seuil associé au maximum. Idem pour la troisième, quatrième,..., $r^{ième}$ plus grande valeur de score local. Utiliser la méthode habituelle pourra donc conduire à sous-estimer la significativité de certains segments, et sous-estimer ainsi un éventuel phénomène biologique sous-jacent.

Deux types de méthodes améliorant l'approche classique sont présentés ci-dessous. Le premier consiste à fixer a priori le nombre de maxima considérés ; deux procédures de test sont alors déduites de cette approche. Le deuxième type de méthode ne fixe pas a priori le nombre de maxima. Il s'agit alors d'une procédure séquentielle.

Détermination de la significativité des plus grands scores locaux - Nombre de scores locaux fixé L'ensemble de ces méthodes repose sur la distribution conjointe des valeurs des k plus grandes valeurs d'un échantillon rappelée précédemment (voir proposition 3.13 page 62 et les travaux de Weissman (1978); Smith (1986); Coles (2001)). Citons également, les travaux de Oncel et al. (2005) qui établissent des récurrences sur la distributions des valeurs prises au différents temps de records, et sur celle des statistiques d'ordre, de manière élégante, à l'aide de techniques de contractions de variables aléatoires.

En notant $M_n^{(i)}$ la $ième$ plus grande valeurs d'un n-échantillon, la fonction de répartition de chaque valeur $M_n^{(i)}$ normalisée par a_n et b_n se déduit de cette densité jointe. Le calcul est détaillé en annexe A. Trois manières de l'expliciter sont présentées ici, en notant $M_n^{\prime(i)} = (M_n^{(i)} - b_n)/a_n$:

$$\mathbb{P}\left(M_n^{\prime(i)} < x\right) = \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\} \sum_{j=0}^{i-1} \frac{\exp\left(-\frac{x-\mu}{\sigma}\right)^j}{j!}$$

$$\mathbb{P}\left(M_n^{\prime(i)} < x\right) = \mathbb{P}\left(M_n^{(i-1)} < x\right) + \frac{1}{(i-1)!}\left\{\exp\left(-\frac{x-\mu}{\sigma}\right)\right\}^{i-1} \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\}$$

$$\mathbb{P}\left(M_n^{\prime(i)} < x\right) = \mathbb{P}\left(Y < i\right) \text{ où } Y \sim Poisson\left\{\exp\left(-\frac{x-\mu}{\sigma}\right)\right\}$$

Néanmoins, la démarche qui consiste à tester successivement les valeurs les plus élevées des scores ne peut se faire sans observer une augmentation du risque de première espèce. En effet, si chaque test est effectué au risque α, la probabilité de rejeter l'hypothèse nulle alors qu'elle est vraie est dans ce cas $1-(1-\alpha)^r$ lorsque r tests indépendants sont effectués. Dans notre cas, les tests ne sont pas indépendants (car les statistiques d'ordre ne sont pas indépendantes) et l'augmentation du risque de première espèce sera par conséquent plus faible.

On note $s_1^\alpha, s_2^\alpha, ..., s_r^\alpha$ les valeurs telles que

$$\mathbb{P}\left(M_n^{'(i)} < s_i^\alpha\right) = 1 - \alpha$$

On définit la règle de décision suivante : "On rejette H_0 si au moins une des réalisations de $M_n^{'(i)}$ dépasse s_i^α". Dans ce cas, le risque de première espèce global vaut $1 - \mathbb{P}\left(M_n^{'(1)} < s_1^\alpha, ..., M_n^{'(r)} < s_r^\alpha\right)$ lorsque r tests sont effectués.

La table 3.2 donne les valeurs des risques de première espèce effectifs en fonction du risque alpha classiquement utilisé pour chaque test, et du nombre r de tests effectués.

α \ r	2	5	10	20
1%	1.9%	3.5%	5.1%	6.7%
	(2%)	(5%)	(10%)	(18%)
5%	8.6%	14.5%	19.2%	24.8%
	(10%)	(23%)	(40%)	(64%)
10%	16.2%	26.4%	32.1%	38.4%
	(19%)	(41%)	(65%)	(87%)

TAB. 3.2 – Risque de première espèce globale lorsque r tests sont effectués au risque α en utilisant les lois marginales asymptotiques des statistiques d'ordre pour déterminer les seuils. Les nombres entre parenthèses valent $1 - (1 - \alpha)^r$: risque de première espèce si les tests sont indépendants.

On s'intéresse ici au cas où le nombre de plus grandes valeurs considérées est fixé. On note r ce nombre. L'hypothèse nulle du test considéré est la suivante :
H_0 : les données $Y_1...Y_n$ sont indépendantes et identiquement distribuées.

Dans la problématique de la recherche d'accumulations de valeurs élevées, deux hypothèses alternatives sont considérées ici :

- H_1 : les données $Y_1...Y_n$ ne sont pas indépendantes et identiquement distribuées. Il y a au moins **1** accumulation de valeurs élevées.

- H_1' : les données $Y_1...Y_n$ ne sont pas indépendantes et identiquement distribuées. Il y a au moins **r** accumulations de valeurs élevées.

Ces deux hypothèses alternatives conduisent à deux formes différentes de régions de rejet de H_0. En effet, on rejettera H_1 si au moins une valeur exceptionnelle est observée parmi les r plus grandes valeurs ; on rejettera H_1' si toutes les r plus grandes valeurs sont exceptionnelles. La première alternative permet simplement d'affirmer que les données

ne sont pas en accord avec l'hypothèse nulle, sans en préciser la teneur. La deuxième alternative conduit à une conclusion beaucoup plus certaine puisqu'en un sens elle "désigne les coupables". Par contre, elle ne rejettera pas H_0 lorsque le nombre de points d'accumulation n'est pas suffisant.

On note α le risque de première espèce du test. Ces deux nuances conduisent alors à deux manières différentes de déterminer les seuils notés respectivement $s_n^{(1)}...s_n^{(r)}$ et $s_n'^{(1)}...s_n'^{(r)}$:

$$\begin{aligned} H_1 : &\quad 1 - \mathbb{P}(M_n^{(1)} < s_n^{(1)}, ..., M_n^{(r)} < s_n^{(r)} \mid H_0) = \alpha \\ H_1' : &\quad \mathbb{P}(M_n^{(1)} > s_n'^{(1)}, ..., M_n^{(r)} > s_n'^{(r)} \mid H_0) = \alpha \end{aligned}$$

De manière duale, on s'intéresse au degré de signification des valeurs $m_n^{(1)}...m_n^{(r)}$ observées. Dans le premier cas, ce degré de signification vaut $p = 1 - \mathbb{P}(M_n^{(1)} < m_n^{(1)}, ..., M_n^{(r)} < m_n^{(r)} \mid H_0)$, et dans le deuxième cas, il vaut $p' = \mathbb{P}(M_n^{(1)} > m_n^{(1)}, ..., M_n^{(r)} > m_n^{(r)} \mid H_0)$.

La théorie des tests nous permet d'affirmer qu'il est équivalent de comparer les valeurs observées au seuil déterminé au risque α, et de comparer le degré de signification à α. Cette remarque prend son importance pour effectuer le test concrètement. En effet, il sera dans certains cas, plus facile ou plus rapide d'obtenir numériquement le degré de signification que les valeurs des seuils ; on verra un exemple ci-dessous.

Une autre remarque, qui prend son importance en pratique, est le fait que toutes les valeurs les plus grandes sont normalisées par les mêmes constantes a_n et b_n (cf. lemme 3.13) lors de la convergence vers la loi jointe. Il suffit de les déterminer pour la valeur maximale et de les utiliser ensuite pour les autres valeurs.

Revenons en au cas des plus grands scores locaux lorsque les scores sont indépendants et identiquement distribués. On cherche alors à déterminer les degrés de signification des tests présentés plus haut. On note les variables aléatoires associées aux plus grands scores observés rangés dans l'ordre décroissant $(H_n^{(1)}, ..., H_n^{(r)})$. D'après le résultat de Karlin et al., les constantes de normalisation sont $a_n = 1$ et $b_n = \ln(n)/\lambda$. Il suffit alors d'utiliser le lemme 3.13 pour en déduire les probabilités recherchées. La forme analytique de la fonction de répartition multidimentionnelle de $\left(M_n^{(1)} - b_n)/a_n, ..., (M_n^{(r)} - b_n)/a_n\right)$ est donnée pour une valeur quelconque de r par la forme récursive suivante (les calculs sont détaillés dans l'annexe A) :

$$\begin{aligned} &\mathbb{P}\left(M_n^{(i)} < x_i, M_n^{(i+1)} < x_{i+1}, ..., M_n^{(r)} < x_r\right) = \\ &\qquad \mathbb{P}\left(M_n^{(r)} < x_r\right) - \sum_{j=i}^{r-1} \frac{\left(e^{-\frac{x_j - \mu}{\sigma}}\right)^j}{j!} \mathbb{P}\left(M_n^{(1)} < x_{j+1}, ..., M_n^{(r-j)} < x_r\right) \end{aligned}$$

Cette formulation récursive peut-être coûteuse en temps de calcul et en mémoire lorsque r devient grand. Une alternative consiste à utiliser des méthodes d'intégration numérique pour calculer ces fonctions de répartition. Les simulations de Monte-Carlo,

effectuées dans un premier temps pour déterminer K et λ peuvent notamment être réutilisées dans ce but. Cette procédure permet une économie substancielle de temps de calcul, même s'il elle est coûteuse en espace mémoire. La procédure est détaillée ci-dessous.

Algorithme 3.22 (Détermination de la significativité des r plus grands scores locaux par Monte-Carlo quand n est grand)

- *Choisir une taille de séquence n_0. On prendra soin de choisir n_0 suffisamment élevé pour que le résultat asymptotique de la loi du score puisse être appliqué sans perdre trop de précision. Néanmoins, plus n_0 sera élevé, plus le déroulement de l'algorithme prendra du temps.*
- *Simuler N séquences, calculer et conserver les valeurs des plus grands scores locaux* $\left\{(h_{n_0,i}^{(1)},...,h_{n_0,i}^{(r)}), i=1,...,N\right\}$.
- *Déduire K et λ de la fonction de répartition de $H_{n_0}^{(1)}$ de la manière indiquée dans l'algorithme 3.6.*
- $\mathbb{P}\left(H_n^{(1)} < h_n^{(1)},...,H_n^{(r)} < h_n^{(r)}\right)$ *est estimé par*

$$\hat{p} = \frac{1}{N}\sum_{i=1}^{N} I\left\{h_{n_0,i}^{(1)} < h_n^{(1)} + \frac{\ln n_0 - \ln n}{\lambda},...,h_{n_0,i}^{(r)} < h_n^{(r)} + \frac{\ln n_0 - \ln n}{\lambda}\right\}$$

- $\mathbb{P}\left(H_n^{(1)} > h_n^{(1)},...,H_n^{(r)} > h_n^{(r)}\right)$ *est estimé par*

$$\hat{p}' = \frac{1}{N}\sum_{i=1}^{N} I\left\{h_{n_0,i}^{(1)} > h_n^{(1)} + \frac{\ln n_0 - \ln n}{\lambda},...,h_{n_0,i}^{(r)} > h_n^{(r)} + \frac{\ln n_0 - \ln n}{\lambda}\right\}$$

Cette méthode présente l'avantage de pouvoir relaxer l'hypothèse d'indépendance de la séquence, en simulant les données sous le modèle qui nous intéresse. En effet, cette procédure n'utilise pas la densité jointe des r plus grandes valeurs, mais seulement le fait que les constantes de normalisation a_n et b_n restent les mêmes pour toutes les variables $H_n^{(1)},...,H_n^{(r)}$.

Par contre, elle présente l'inconvénient d'être "gourmande" en mémoire et en complexité. En ce qui concerne la mémoire, il sera nécessaire de conserver la totalité des résultats des simulations de Monte-Carlo pour pouvoir calculer les probabilités désirées. L'espace mémoire occupé par ces résultats sera de taille $N \times r$, où N est le nombre d'itérations de Monte-Carlo et r est le nombre de plus grandes valeurs considérées. En général, N sera de l'ordre de la dizaine de milliers et r de l'ordre de la dizaine.

La tableau 3.3 présente la taille mémoire occupée par l'algorithme pour différentes valeurs de N et de r. La taille mémoire occupée par chaque valeur est supposée être de 8 octets comme c'est généralement le cas pour le type double pour les compilateurs du langage C/C++.

r \ n	1000	10000	50000	100000
5	0.04	0.40	2.00	4.00
10	0.08	0.80	4.00	8.00
15	0.12	1.20	6.00	12.00
50	0.4	4.0	20.0	40.0
100	0.8	8.0	40.0	80.0

TAB. 3.3 – Taille mémoire utilisée (en Mo) par l'algorithme de détermination de la significativité des r plus grands scores locaux par Monte-Carlo. Chaque nombre stocké est supposé ici utiliser 8 octets.

La détermination du nombre de fois où $\left\{h_{n_0,i}^{(1)} < h_n^{(1)} + \frac{\ln n_0 - \ln n}{\lambda}, ..., h_{n_0,i}^{(r)} < h_n^{(r)}, i = 1...N\right\}$ s'effectue simplement de la manière suivante :

Algorithme 3.23 (Calcul du cardinal de $\left\{h_{n_0,i}^{(1)} < h_n^{(1)} + \frac{\ln n_0 - \ln n}{\lambda}, ..., h_{n_0,i}^{(r)} < h_n^{(r)}, i = 1...N\right\}$)

- *comptage* $\leftarrow 0$
- *pour i allant de 1 à N*
  - *si* $h_{n,i}^{(1)} < h_n^{(1)}$ *alors*
    - $j \leftarrow 2$
    - $Trouve = FAUX$
    - *tant que* $j <= r$ *et* $Trouve = FAUX$
      - $Trouve \leftarrow \left(h_{n,i}^{(j)} < h_n^{(j)}\right)$
      - $j \leftarrow j + 1$
    - *si* $Trouve = VRAI$, *alors comptage* $\leftarrow$ *comptage* $+ 1$

On procède de même pour déterminer le cardinal de

$$\left\{h_{n_0,i}^{(1)} > h_n^{(1)} + \frac{\ln n_0 - \ln n}{\lambda}, ..., h_{n_0,i}^{(r)} > h_n^{(r)}, i = 1...N\right\}.$$

Puisque ces probabilités seront amenées à être calculées plusieurs fois, il est intéressant de trier les résultats selon la variable $h_n^{(1)*}$. La complexité de cet algorithme se calcule assez facilement. Le nombre d'itérations à effectuer pour trouver le nombre de $h_{n,i}^{(1)} < h_n^{(1)}$, dans le cas où les données sont triées, est en moyenne $N \times \mathbb{P}\left(H_n^{(1)} < h_n^{(1)}\right)$. Ensuite, on s'intéresse au cas où $h_{n,i}^{(2)} < h_n^{(2)}$ uniquement si $h_{n,i}^{(1)} < h_n^{(1)}$, ce qui conduit à un nombre d'itérations moyen de $N \times \mathbb{P}\left(H_n^{(2)} < h_n^{(2)} \mid H_n^{(1)} < h_n^{(1)}\right)$. En poursuivant ce raisonnement,

*les résultats ne seront pas pour autant triés pour les autres variables

on déduit la complexité moyenne de l'algorithme :

$$N^r \times \mathbb{P}\left(H_n^{(1)} < h_n^{(1)}\right) \times \prod_{j=2}^{r} \mathbb{P}\left(H_n^{(j)} < h_n^{(j)} \mid \left\{H_n^{(j-1)} < h_n^{(j-1)}\right\}_{k=1,\ldots,j-1}\right)$$

ce qui donne finalement :

$$N^r \times \mathbb{P}\left(H_n^{(1)} < h_n^{(1)}, ..., H_n^{(r)} < h_n^{(r)}\right).$$

En pratique, les valeurs observées sur des séquences seront en général élevées. Ceci va conduire à une valeur élevée de $\mathbb{P}(M_n^{(1)} < m_n^{(1)}, ..., M_n^{(r)} < m_n^{(r)} \mid H_0)$ et faible valeur de $\mathbb{P}(M_n^{(1)} > m_n^{(1)}, ..., M_n^{(r)} > m_n^{(r)} \mid H_0)$. Par conséquent, la complexité moyenne pour déterminer la première probabilité sera beaucoup plus élevée que celle pour déterminer la deuxième.

Cette méthode permet de déterminer le degré de signification d'un résultat observé en un temps certes important mais concevable. On voit bien ici que la détermination d'un **seuil de significativité** en fonction d'un risque de première espèce choisi a priori sera beaucoup plus délicate. De plus, il n'y aura pas une solution unique.

Détermination des plus grands scores locaux significatifs - Nombre de scores locaux non fixé Les méthodes précédentes présentent l'avantage d'exprimer le résultat du test à l'aide d'un degré de signification. Cette approche est souvent préférée à une approche de type détermination d'un seuil de significativité. Néanmoins, le nombre de scores locaux à considérer est fixé à r. Cette limite est gênante pour plusieurs raisons. Qu'en est-t-il du $(r+1)-$ième score local ? Comment choisir r ? Si r est trop grand, on n'acceptera probablement pas l'hypothèse que tous les scores locaux sont trop élevés, et si l'on rejette l'hypothèse que la séquence initiale est indépendante et identiquement distribuée, on ne saura pas préciser quel(s) segment(s) en est (sont) responsable(s).

On reprend chacune des deux hypothèses précédentes, mais en ne supposant plus que le nombre de scores locaux à considérer est fixé a priori. On cherche à détecter le plus grand nombre de scores locaux tout en maîtrisant le risque de première espèce du test. Pour cela, on formule la conjecture suivante :

D'un point de vue biologique, plus la valeur d'un score local est élevée, plus le segment qui l'a engendré est caractéristique du phénomène étudié.

En effet, les fonctions de scores sont construites pour refléter un phénomène physico-chimique tel que l'hydrophobicité par exemple. Le score observé d'un certain segment reflète donc cette propriété. Statistiquement, il n'est pas difficile d'expliquer pourquoi le troisième score, par exemple, est significatif alors que les deux premiers ne le sont pas. Exemple : alors que les deux premiers scores observés seront compatibles avec le modèle H_0, le troisième score serait trop élevé, c'est à dire que sa valeur serait trop "proche" de la deuxième plus grande valeur. Il semble difficile dans ce cas, de conclure

biologiquement que le segment correspondant au troisième score présente une "anomalie", mais pas le segment qui a conduit à la deuxième valeur. Ce serait, de plus, accorder trop de confiance au modèle supposé générer les données. Affirmer que les séquences biologiques ne sont que rarement générées selon un modèle de variables aléatoires indépendantes et identiquement distribuées, ou même une chaîne de Markov tient plus de l'euphémisme que de la profession de foi.

Cette conjecture permet de s'intéresser à la significativité des scores successivement, dans une démarche du type : "Si un score local est significatif, alors on s'intéresse au suivant". Cette démarche s'appuie sur la formulation d'hypothèses suivantes :

$$H_0^{'(i)} : \text{Il y a (au plus) } i-1 \text{ accumulations de valeurs élevées}$$
$$\text{contre}$$
$$H_1^{'(i)} : \text{Il y a au moins } i \text{ accumulations de valeurs élevées}$$

Le cas $i = 1$ est celui où le maximum des valeurs est testé, c'est à dire celui correspondant aux hypothèses précédentes H_0 et H_1.

La démarche de test proposée est la suivante :

Algorithme 3.24 (Démarche de test pour déterminer la significativité des plus grands scores)

1. *Effectuer le test de l'hypothèse $H_0^{'(i)}$ contre $H_1^{'(i)}$.*
2. *Si le test est significatif au risque α_i choisi, alors*
   $i \leftarrow i+1$ et continuer au point 1.
3. *Sinon*
   Conclure :
   - *si $i > 1$ alors "Les $i-1$ plus grands scores sont significatifs*."*
   - *sinon "Il n'y a pas de score significatif dans la séquence".*

Il nous reste à préciser comment effectuer le test à l'étape i. Précisons tout d'abord le niveau de risque auquel chaque test sera effectué. On désire contrôler le niveau de risque global α, c'est à dire que la probabilité de rejeter l'hypothèse H_0 (Les données sont indépendantes et identiquement distribuées) alors que H_0 est vraie vaut α.

Proposition 3.25 *Soit une séquence ordonnée et linéaire de tests, dans laquelle le ième test est effectué seulement si le test précédent est significatif au risque α (algorithme 3.24). Alors, le niveau de risque global de cette démarche vaut α, quel que soit le nombre de tests effectués.*

*Le risque est à préciser en fonction des statistiques de test utilisées

Preuve. $H_0^{(i)}$ désigne l'hypothèse nulle numéro i. On note $s^{(i)}$ le seuil de significativité associé au test de l'hypothèse $H_0^{(i)}$ au risque α. Le risque global est défini comme étant la probabilité de rejeter au moins une hypothèse $H_0^{(i)}$ alors que toutes ces hypothèses sont vérifiées. L'événement complémentaire est de ne rejeter aucune hypothèse $H_0^{(i)}$ alors qu'elles sont vérifiées. La probabilité de cet événement complémentaire se calcule aisément dans la démarche de l'algorithme 3.24. On a :

$$\begin{aligned}
\mathbb{P}(\text{rejet de } H_0 \mid H_0) &= 1 - \mathbb{P}(\text{accepter } H_0 \mid H_0) \\
&= 1 - \mathbb{P}\left(\text{accepter } H_0^{(1)} \mid H_0\right) \\
&= 1 - (1-\alpha) \\
&= \alpha
\end{aligned}$$

■

Cette preuve s'appuie sur le fait que le niveau global ne dépend que du niveau du premier test. Ceci peut sembler gênant, car une règle de décision quelconque pour les autres plus grandes valeurs garantit néanmoins le risque de première espèce global, tout en fournissant des résultats qui peuvent n'avoir aucun sens. Cela s'explique par le fait que nous avons implicitement supposé que l'alternative est H_1 : "Il y a au moins **un** score trop élevé". Dans ce cas, et en supposant la conjecture que les scores les plus élevés sont aussi les plus intéressants, il semble évident qu'il suffit de tester la valeur maximale pour conclure.

Cette remarque nous conduit à introduire la notion de risque de première espèce partiel, dont la définition est la suivante.

Définition 3.26 (Risque de première espèce partiel) *On considère une séquence ordonnée et linéaire de tests, dans laquelle le ième test est effectué seulement si le test précédent conclut au rejet de l'hypothèse nulle précédente. H_0^i désigne l'hypothèse nulle numéro i. On définit le risque de première espèce partiel de rang i noté α_i comme la probabilité de rejeter à tort l'hypothèse nulle numéro i conditionnellement au fait que les hypothèses nulles précédentes ont été rejetées.*

$$\alpha_i = \mathbb{P}\left(\bigcup_{j \geqslant i} \left\{ Rejeter\ H_0^j \right\} \mid \bigcup_{j<i} \left\{ Rejeter\ H_0^j \right\}\right)$$

Le risque de première espèce global (ou risque de première espèce) est alors α_1.

Il semble une propriété intéressante pour une démarche séquentielle de test de garantir que le risque de première espèce partiel de chaque rang est majoré par le risque de première espèce global.

La proposition 3.25 ne garantit pas cette propriété.

Proposition 3.27 *Soit une séquence ordonnée et linéaire de tests* **indépendants**, *dans laquelle le ième test est effectué seulement si le test précédent est significatif au risque* α'_{i-1}. *Alors, si on applique un risque* α'_i *à chaque test, les niveaux de risque partiels* α_i *de cette démarche sont tous égaux à* α'_i. *Si* k *tests sont effectués, le risque de première espèce global vaut alors :* α.

Preuve. La preuve est similaire à la preuve précédente ; En effet :

$$\begin{aligned}
\alpha_i &= \mathbb{P}\left(\bigcup_{j\geqslant i}\left\{\text{Rejeter } H_0^j\right\} \mid \bigcup_{j<i}\left\{\text{Rejeter } H_0^j\right\}\right) \\
&= 1-\mathbb{P}\left(\bigcap_{j\geqslant i}\left\{\text{Accepter } H_0^j\right\} \mid \bigcup_{j<i}\left\{\text{Rejeter } H_0^j\right\}\right) \\
&= 1-\mathbb{P}\left(\left\{\text{Accepter } H_0^i\right\} \mid \bigcup_{j<i}\left\{\text{Rejeter } H_0^j\right\}\right) \\
&= 1-(1-\alpha'_i) \\
&= \alpha'_i
\end{aligned}$$

■

Dans notre cadre, l'hypothèse alternative adéquate est une hypothèse du type de H'_1 : "Il y a au moins **r** scores trop élevés", r devant être le plus grand possible. Dans ce cas, le risque de première espèce s'exprime de la façon suivante :

$$\alpha = \mathbb{P}(H_n^{(1)} > s'^{(1)}, ..., H_n^{(r)} > s'^{(r)})$$

où $s'^{(1)}, ..., s'^{(r)}$ sont les seuils de significativité.

La première idée repose sur la factorisation de la probabilité suivante :

$$\begin{aligned}
&\mathbb{P}(H_n^{(1)} > s^{(1)}, ..., H_n^{(r)} > s^{(r)}) = \\
&\quad \mathbb{P}\left(H_n^{(1)} > s^{(1)}\right) \times \prod_{j=2}^{r} \mathbb{P}\left(H_n^{(j)} > s^{(j)} \mid \left\{H_n^{(j-1)} > s^{(j-1)}\right\}_{k=1,...,j-1}\right)
\end{aligned}$$

Cette factorisation suggère de déterminer le seuil de significativité $s^{(1)}$ pour le test de la valeur maximale à l'aide de $\mathbb{P}\left(H_n^{(1)} > s^{(1)}\right)$, puis le seuil $s^{(2)}$ pour le test de la deuxième plus grande valeur à l'aide de $\mathbb{P}\left(H_n^{(2)} > s^{(2)} \mid H_n^{(1)} > s^{(1)}\right)$, puis le seuil $s^{(3)}$ pour le test de la troisième plus grande valeur à l'aide de $\mathbb{P}\left(H_n^{(3)} > s^{(3)} \mid H_n^{(2)} > s^{(2)}, H_n^{(1)} > s^{(1)}\right)$, etc.

Dans ce cas, cette série de tests a la propriété intéressante de garantir que la probabilité $\mathbb{P}(H_n^{(1)} > h^{(1)}, ..., H_n^{(r)} > h^{(r)})$ est inférieure à α. Cette démarche constitue donc une généralisation du cas où le nombre r de scores à considérer est fixé a priori. Remarquons néanmoins que la conjecture concernant la validité biologique d'un score est nécessaire pour effectuer cette généralisation.

Remarque 3.28 *Dans ce cas, la valeur du seuil $s'^{(1)}$ pour le maximum est plus faible que dans le cas où ce seuil $s^{(1)}$ est fixé sur la base de la distribution de $H_n^{(1)}$ seule. La factorisation de la probabilité suivante permet de s'en rendre compte :*

$$\alpha = \mathbb{P}\left(H_n^{(1)} > s'^{(1)}\right) \times \prod_{j=2}^{r} \mathbb{P}\left(H_n^{(j)} > s'^{(j)} \mid \left\{H_n^{(j-1)} > s'^{(j-1)}\right\}_{k=1,\ldots,j-1}\right)$$

En effet,

$$\begin{aligned}\left\{\alpha = \mathbb{P}\left(H_n^{(1)} > s^{(1)}\right) \ ET \ s'^{(1)} > s^{(1)}\right\} \ &\Rightarrow \ \mathbb{P}\left(H_n^{(1)} > s'^{(1)}\right) < \alpha \\ &\Rightarrow \ \mathbb{P}\left(H_n^{(1)} > s'^{(1)}, ..., H_n^{(r)} > s'^{(r)}\right) < \alpha\end{aligned}$$

Cette démarche itérative aura donc une puissance plus faible que dans le cas où r est connu pour détecter r scores trop élevés.

En fait, la probabilité $\mathbb{P}(H_n^{(1)} > h^{(1)}, ..., H_n^{(r)} > h^{(r)})$ est inférieure à $\alpha^{\mathbf{r}}$ si on prend un risque α pour calculer les valeurs de chaque seuil. On rappelle que r n'est pas déterminé à l'avance dans cette démarche. Le tableau 3.4 donne en illustration quelques valeurs de α^r en fonction de r. Ces valeurs sont beaucoup plus faibles que la valeur du risque α pris dans le cas où r est fixé, ceci est un indicateur de la baisse de puissance qui se produira losrque r n'est pas fixé. Néanmoins, cette baisse de puissance est progressive, dans le sens où la puissance pour détecter une seule valeur élevé sera évidemment la même, ensuite le risque pour en détecter deux correspond à α^2, etc.. Cette procédure est, de toute manière, plus puissante que la méthode usuellement utilisée qui consiste à considérer $s^{(1)}$ comme valeur seuil pour toutes les variables.

α \ r	2	5	10	15	50	100
1	0.01	10^{-8}	10^{-18}	10^{-28}	10^{-98}	10^{-198}
5	0.25	3×10^{-5}	10^{-13}	3×10^{-19}	9×10^{-65}	8×10^{-130}
10	1	0.001	10^{-8}	10^{-13}	10^{-48}	10^{-98}

TAB. 3.4 – Quelques valeurs de α^r pour des risques de première espèce α et des nombres de segments r arbitraires (exprimées **en %**).

Néanmoins, la détermination concrète de ces seuils n'est pas aisée car les fonctions de répartition sont certes calculables, mais leur forme est complexe (cf. annexe A). On pourrait ici aussi réutiliser les simulations de Monte-Carlo effectuées pour déterminer les constantes K et λ. Cependant, on s'intéresse ici aux queues de distribution, et il faudrait un grand nombre d'itérations pour avoir une précision suffisante pour la détermination des seuils (la probabilité recherchée vaut α^r).

Pour contourner cette difficulté, une autre solution consiste à conditionner les distributions des événements $H_n^{(j)} > s^{(j)}$ par $\left\{H_n^{(j-1)} = h^{(j-1)}, k = 1, ..., j-1\right\}$ ou par $\left\{H_n^{(j-1)} = s^{(j-1)}, k = 1, ..., j-1\right\}$. En effet, les distributions obtenues sont explicites ce qui permet de simplifier de façon importante les calculs à effectuer :

Lemme 3.29 *La distribution de la kième plus grande valeur conditionnellement aux $k-1$ valeurs plus élevées d'un n-échantillon ne dépend que de la valeur immédiatement plus élevée :*

$$[M^{(i)} \mid M^{(i-1)}, ..., M^{(1)}] = [M^{(i)} \mid M^{(i-1)}]$$

où la notation $[X]$ désigne la distribution de la variable aléatoire X.

La densité asymptotique de $(M^{(i)} - b_n)/a_n$ conditionnellement à $(M^{(i-1)} - b_n)/a_n$, où a_n et b_n sont les constantes de normalisation adéquates, est donnée par :

$$f_{M^{(i)}|M^{(i-1)}}(m^{(i)} \mid m^{(i-1)}) = \frac{f_{M^{(1)}}(m^{(i)})}{F_{M^{(1)}}(m^{(i-1)})}$$

et

$$\mathbb{P}\left(M^{(i)} \leqslant m^{(i)} | M^{(i-1)} = m^{(i-1)}\right) = \frac{F_{M^{(1)}}(m^{(i)})}{F_{M^{(1)}}(m^{(i-1)})}$$

c'est à dire, dans le cas où la distribution du maximum $M^{(1)}$ est une loi de Gumbel :

$$f_{M^{(i)}|M^{(i-1)}}(m^{(i)} \mid m^{(i-1)}) = \frac{1}{\sigma} \frac{\exp\left\{-\exp\left(-\frac{m^{(i)}-\mu}{\sigma}\right)\right\} \exp\left(-\frac{m^{(i)}-\mu}{\sigma}\right)}{\exp\left\{-\exp\left(-\frac{m^{(i-1)}-\mu}{\sigma}\right)\right\}} I_{\left\{m^{(i)} \leqslant m^{(i-1)}\right\}}$$

et

$$\mathbb{P}\left(M^{(i)} \leqslant m^{(i)} \mid M^{(i-1)} = m^{(i-1)}\right) = \exp\left(-e^{-\frac{m^{(i)}-\mu}{\sigma}} + e^{-\frac{m^{(i-1)}-\mu}{\sigma}}\right)$$

où $F_{M^{(1)}}$ désigne la fonction de répartition du maximum de l'échantillon.

Preuve. La preuve s'effectue par simple calcul d'intégration. Ces calculs sont présentés en annexe A. ■

Finalement, nous proposons les deux démarches de tests suivantes :

1. Calcul des constantes de normalisations K et λ
2. Standardisation des plus grandes valeurs de score local :

$$\forall j, H_n'^{(j)} = \lambda H_n^{(j)} - \log(K \times n)$$

3. On teste au risque α la plus grande valeur à l'aide du seuil $s^{(1)}$ tel que $\exp(-\exp(-s^{(1)})) = 1-\alpha$.
   Ou, de manière équivalente, on compare le degré de signification $p = \exp(-\exp(-h_n'^{(1)}))$ à α.
4. Tant que l'hypothèse $H_0^{(i)}$ est rejetée ,
   on effectue le test de la valeur $H_n'^{(i+1)}$ en utilisant une des deux règles de décision suivantes : on calcule le seuil $s^{(i+1)}$ tel que
   (a) **Règle 1** : $\mathbb{P}(H_n'^{(i+1)} > s^{(i+1)} \mid H_n'^{(i)} = s^{(i)}) = \alpha$
   Ou, de manière équivalente, on compare le degré de signification $p = \exp\left(-\exp(-h_n'^{(i+1)}) + \exp(-s^{(i)})\right)$ à α.
   (b) **Règle 2** : $\mathbb{P}(H_n'^{(i+1)} > s^{(i+1)} \mid H_n'^{(i)} = h'^{(i)}) = \alpha$
   Ou, de manière équivalente, on compare le degré de signification $p = \exp\left(-\exp(-h_n'^{(i+1)}) + \exp(-h_n'^{(i)})\right)$ à α.

La deuxième règle de décision semble a priori plus naturelle, car on conditionne par les valeurs déjà observées. Néanmoins, ce conditionnement n'intervient que lorsque la valeur observée à l'étape précédente (et par laquelle on conditionne) est "anormalement" élevée, et sans doute le modèle i.i.d. n'est pas valable. Ce constat sous-tend la proposition de la première règle de décision.
La première règle de décision présente l'avantage d'être efficace d'un point de vue calculatoire. En effet, en utilisant les scores renormalisés $H_n'^{(i)}$, ces seuils peuvent être déterminés une fois pour toutes et tabulés. Le tableau 3.5 donne les cinquantes premières valeurs de ces seuils $s^{(i)}$.

1-5	2.970	2.277	1.872	1.584	1.361
6-10	1.178	1.024	0.891	0.773	0.668
11-15	0.572	0.485	0.405	0.331	0.262
16-20	0.198	0.137	0.080	0.026	-0.026
21-25	-0.074	-0.121	-0.165	-0.208	-0.249
26-30	-0.288	-0.326	-0.362	-0.397	-0.431
31-35	-0.464	-0.496	-0.526	-0.556	-0.585
36-40	-0.613	-0.641	-0.667	-0.693	-0.719
41-45	-0.743	-0.767	-0.791	-0.814	-0.836
46-50	-0.858	-0.880	-0.901	-0.922	-0.942

TAB. 3.5 – 50 premières valeurs des seuils $s^{(i)}$ pour la règle de décision numéro 1 et des valeurs de score local renormalisées.

Remarquons que

$$\mathbb{P}\left(H_n'^{(i)} < h_n'^{(i)} \mid H_n'^{(i-1)} = h_n'^{(i-1)}\right) > \mathbb{P}\left(H_n'^{(i)} < h_n'^{(i)} \mid H_n'^{(i-1)} = s_n^{(i-1)}\right)$$

car $h_n'^{(i-1)} > s_n^{(i-1)}$. La deuxième règle de décision sera par conséquent plus conservative que la première.

Remarquons enfin que

$$\mathbb{P}\left(H_n'^{(i)} < h_n'^{(i)} \mid H_n'^{(i-1)} > h_n'^{(i-1)}\right) < \mathbb{P}\left(H_n'^{(i)} < h_n'^{(i)} \mid H_n'^{(i-1)} = h_n'^{(i-1)}\right)$$

Il suffit de poser $h_n'^{(i)} = h_n'^{(i-1)}$ pour s'en convaincre.

Par conséquent, le produit de degré de signification de chaque test sous-estime le degré de signification du test global, si on considère le nombre r de tests connu à l'avance. C'est à dire que si on note $p^{(i)} = \mathbb{P}\left(H_n'^{(i)} > h_n'^{(i)} \mid H_n'^{(i-1)} = h_n'^{(i-1)}\right)$ le degrè de signification du test numéro i, on a :

$$\prod_{i=1}^{r} p^{(i)} < \mathbb{P}\left(H_n'^{(1)} > h_n'^{(1)}, ..., H_n'^{(r)} > h_n'^{(r)}\right)$$

Ce résultat peut laisser penser que la démarche de test proposée ici aura tendance à sous-estimer la probabilité réelle, et conduire à des rejets d'hypothèses nulles non fondés. En fait, cette impression est inexacte, car la démarche ne conduit pas à comparer $\prod_{i=1}^{r} p^{(i)}$ au risque de première espèce global α.

Etude de Simulation

Afin de comparer les puissances respectives de ces deux approches, nous avons effectué une étude simple de simulation.

On se place pour cela dans le cadre d'une séquence i.i.d. à valeurs dans $\{0, 1, 2, 3\}$ avec les probabilités d'occurence arbitrairement choisies suivantes : $\mathbb{P}(X = 0) = 0.5$, $\mathbb{P}(X = 1) = 0.1$, $\mathbb{P}(X = 2) = 0.2$, $\mathbb{P}(X = 3) = 0.2$

Nous étudions le comportement de la méthode en fonction des quatre paramètres suivants :

1. La taille de la séquence : $n = 1000, 10000, 100000$.
2. Le système de score :
   (a) $S_1(0) = -4, S_1(1) = -3, S_1(2) = 1, S_1(3) = 3$, ce qui donne $E[S_1(X)] = 1.5$
   (b) $S_2(0) = -8, S_2(1) = -6, S_2(2) = 2, S_2(3) = 6$ ce qui donne $E[S_2(X)] = 3$
   (c) $S_3(0) = -16, S_3(1) = -12, S_3(2) = 4, S_3(3) = 12$ ce qui donne $E[S_3(X)] = 6$
3. Le nombre de plages sous H_1 : $r = 0, 5, 10, 20$
4. La longueur des plages sous H_1 : $l = 10, 20$

Pour simuler les scores sous H_1, on utilise l'opposé des scores sous H_0, et deux plages consécutives sous H_1 sont séparées par 30 positions.

Pour chaque combinaison des paramètres, 1000 échantillons sont tirés aléatoirement. Pour résumer ces simulations, on s'intéresse à la proportion de fois où il n'y a pas eu d'hypothèses nulles rejetées, et où il y a eu k hypothèses rejetées pour toutes les valeurs de k observée. Le risque de première espèce α est fixé à 5%.

Comportement sous H_0

Ceci nous permet essentiellement de contrôler que la programmation de l'algorithme est correcte. C'est bien le cas ici, où dans environ 95% des échantillons, l'hypothèse nulle globale ("la séquence est i.i.d.") n'est pas rejetée. Les résultats ne sont pas détaillés ici. Ce résultat indique également que l'approximation de la loi du score local par sa loi limite (Gumbel) est justifié dans ce cas, même pour une séquence de taille 1000, et une espérance du score relativement petite. Ce résultat ne peut néanmoins pas se généraliser à un système de score quelconque et une distribution de la séquence quelconque. Comme pour l'ensemble des méthodes se fondant sur la loi asymptotique du score local présentées dans ce chapitre, il est absoluement nécessaire de s'assurer que cette approximation est valable pour le système de score, la distribution de séquence et la taille de séquence utilisée avant d'employer ces méthodes. Une simple étude de simulation permet de répondre à cette question.

On s'aperçoit également que le nombre r de tests $H_0^{(i)}, i = 1, ..., r$ pour une séquence ne dépasse jamais 3 sur l'ensemble des cas de figures simulés sous H_0, avec une probabilité élevée de ne rejeter à tort qu'une seule hypothèse nulle, car quasiment la totalité des 5% des séquences issues du rejet de $H_0^{(1)}$ ne conduisent pas au rejet de $H_0^{(2)}$. Comme espéré, un rejet erronné de $H_0^{(1)}$ n'aura pas de conséquence gênante sur le rejet des autres hypothèses nulles.

Comportement sous H_1

Les graphiques suivants présentent le résultat de ces simulations pour différentes combinaisons de ces paramètres et pour chacune des deux règles de décision. Un point de ces graphiques s'interprète comme le nombre de fois (sur les 1000 simulations) où l'on a rejeté r hypothèses nulles $H_0^{(i)}, i = 1, ..., r$, r étant placé en abscisse. Le graphique 3.2 page 88 est obtenu pour la fonction de score S_1, le deuxième (3.3 page 89) pour la fonction de score S_2, et le troisième pour la fonction de score S_3 (3.4 page 90).
Sur chacun de ces trois graphiques, La première règle est symbolisée par des cercles, et la deuxième par des triangles. La valeur de l'ordonnée y représente le nombre de fois (sur 1000) où l'on a rejeté la valeur de l'abscisse x. Hypothèses nulles : $H_0^{(i)}, i = 1, ..., x$. Chaque colonne représente respectivement les cas de nombre de retournements : 5, 10 et 20. Les deux premières lignes sont obtenues pour une séquence de taille 1000, et les deux dernières pour une séquence de taille 100000. Pour chacune de ces longueurs de séquence, la première ligne correspond à des plages sous H_1 de taille 10, et de taille 20 pour la deuxième.

La première conclusion globale est que le nombre de segments déclarés à la fin de la procédure comme n'étant pas sous H_0 ne dépasse quasiment jamais le vrai nombre de segments sous H_1. Cette conclusion est valable quelle que soit la taille de la séquence, la fonction de score utilisée, le nombre de segments sous H_1 et la taille de ces segments.

On s'aperçoit également que la règle de décision fondée sur le conditionnement par les seuils précédants, au lieu de la valeur observée précédemment, conduit à rejeter plus de segments.

Le tableau 3.6 page 86 résume ces distributions en présentant les nombres moyens d'hypothèses nulles $H_0^{(i)}$ rejetées en fonction des différentes configurations de paramètres. Comme prévu, les deux démarches de tests sont plutôt conservatives, même si elles permettent de détecter plus de segments que la méthode classique consistant à employer le seuil obtenu à l'aide de la distribution de la plus grande valeur de score local. Ceci est partibulièrement vrai pour le cas de figure où les segments sont de taille 10 dans une longue séquence de taille 10^5. Dans ce cas, la puissance du test de l'hypothèse globale ("la séquence est i.i.d.") n'excède pas 50%. La présence de 20 segments sous H_1 au lieu de 5, modifie assez peu cette puissance.

On remarque également la faible influence de l'espérance de la fonction de score utilisée.

		S_1					
n	l \ r	5		10		20	
1000	10	2.20	1.83	5.48	4.00	12.94	8.49
	20	4.15	3.71	8.36	7.38	9.04	7.79
100000	10	0.19	0.18	0.42	0.34	0.96	0.72
	20	2.50	2.05	5.26	4.03	11.59	8.23
		S_3					
n	l \ r	5		10		20	
1000	10	2.54	2.02	5.69	4.13	13.17	8.67
	20	4.29	3.83	8.47	7.51	16.49	14.70
100000	10	0.23	0.22	0.45	0.40	1.10	0.85
	20	2.52	2.08	5.26	4.09	11.62	8.24

TAB. 3.6 – Nombre moyen de d'hypothèses $H_0^{(i)}$ rejetées. n : taille de la séquence ; r : nombre de segments sous H_1 ; l : taille des segments sous H_1. Pour chaque cas, la colonne de gauche représente la règle de décision numéro 1 (conditionnement par le seuil précédant), et celle de droite la régle de décision numéro 2 (conditionnement par la valeur observée précédente)

Conclusion de cette étude.

L'étude de simulation montre que les deux démarches de tests employées sont conservatives. On préférera donc la démarche utilisant la règle de décision numéro 1 (conditionnement par les seuils précédants) qui est moins conservative que la règle numéro 2 (conditionnement par les valeurs observées précédantes). De plus, elle est plus efficace en pratique, car les valeurs des seuils peuvent être déterminées indépendamment des données, donc être tabulées.

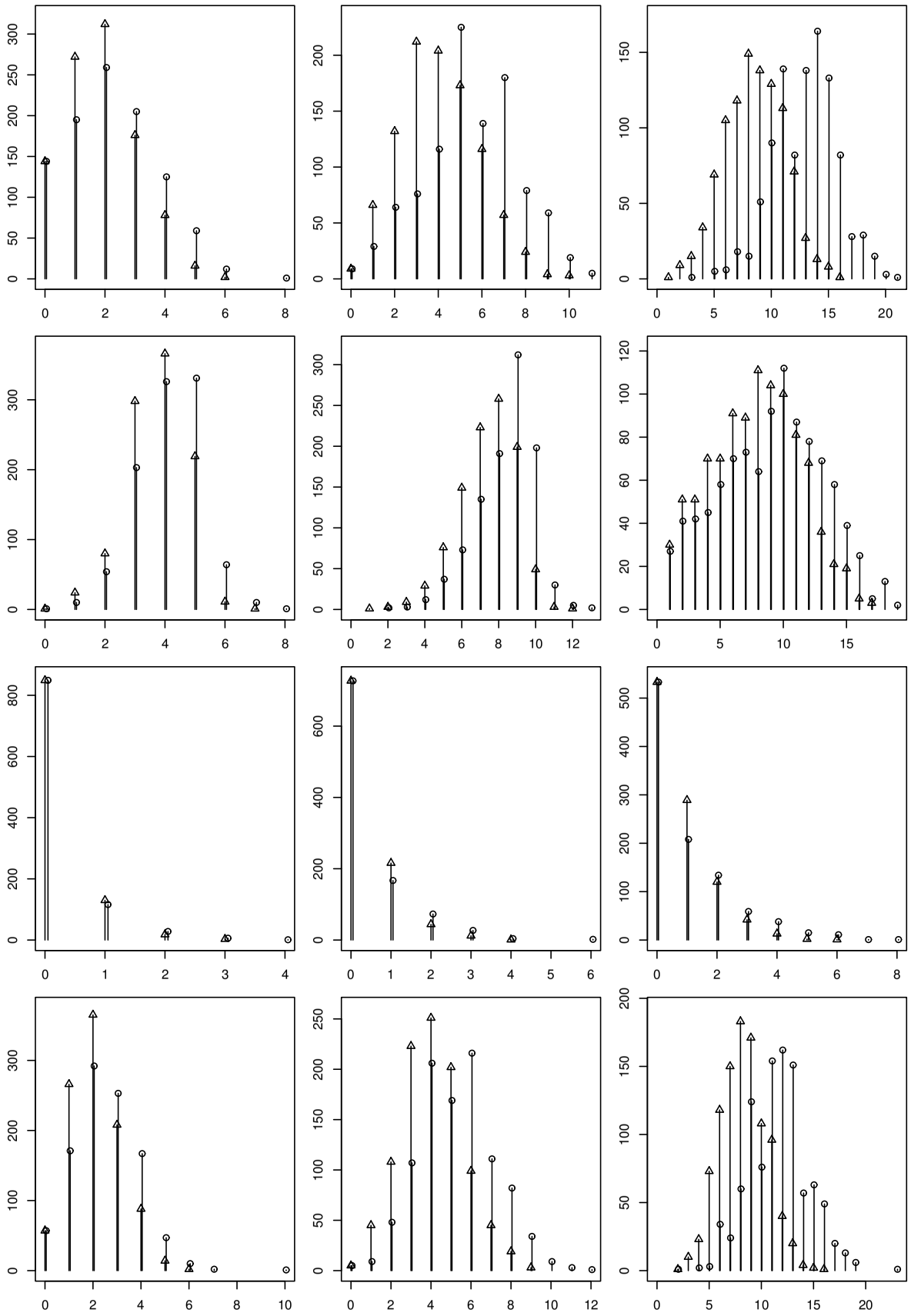


FIG. 3.2 – Taux de rejet des hypothèses nulles obtenues pour chacune des deux règles de décision et la fonction de score S_1 ($E[S_1] = -1.5$). Légende : voir texte page 85

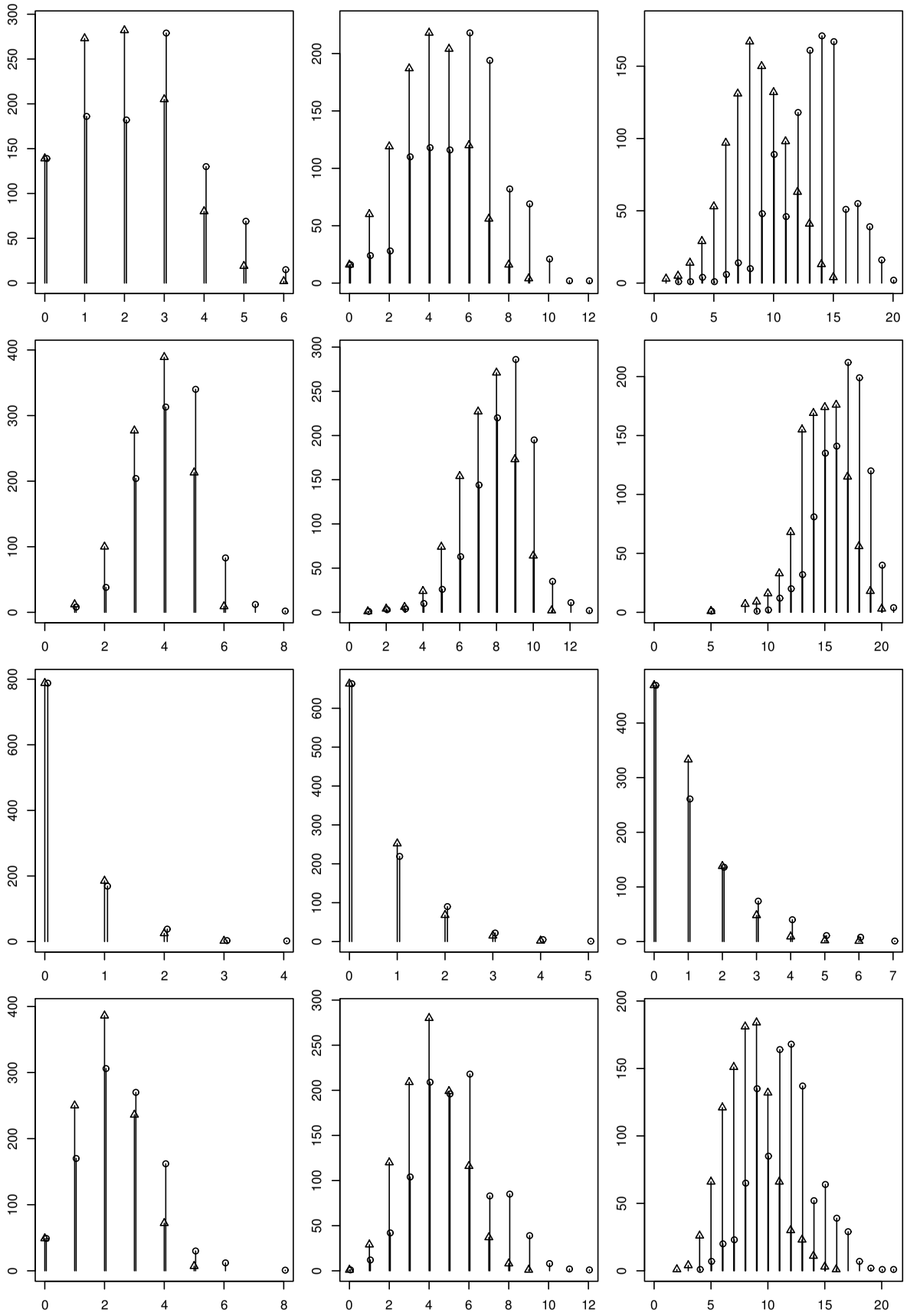


FIG. 3.3 – Taux de rejet des hypothèses nulles obtenus pour chacune des deux règles de décision et la fonction de score S_2 ($E[S_2] = -3$). Légende : voir texte page 85

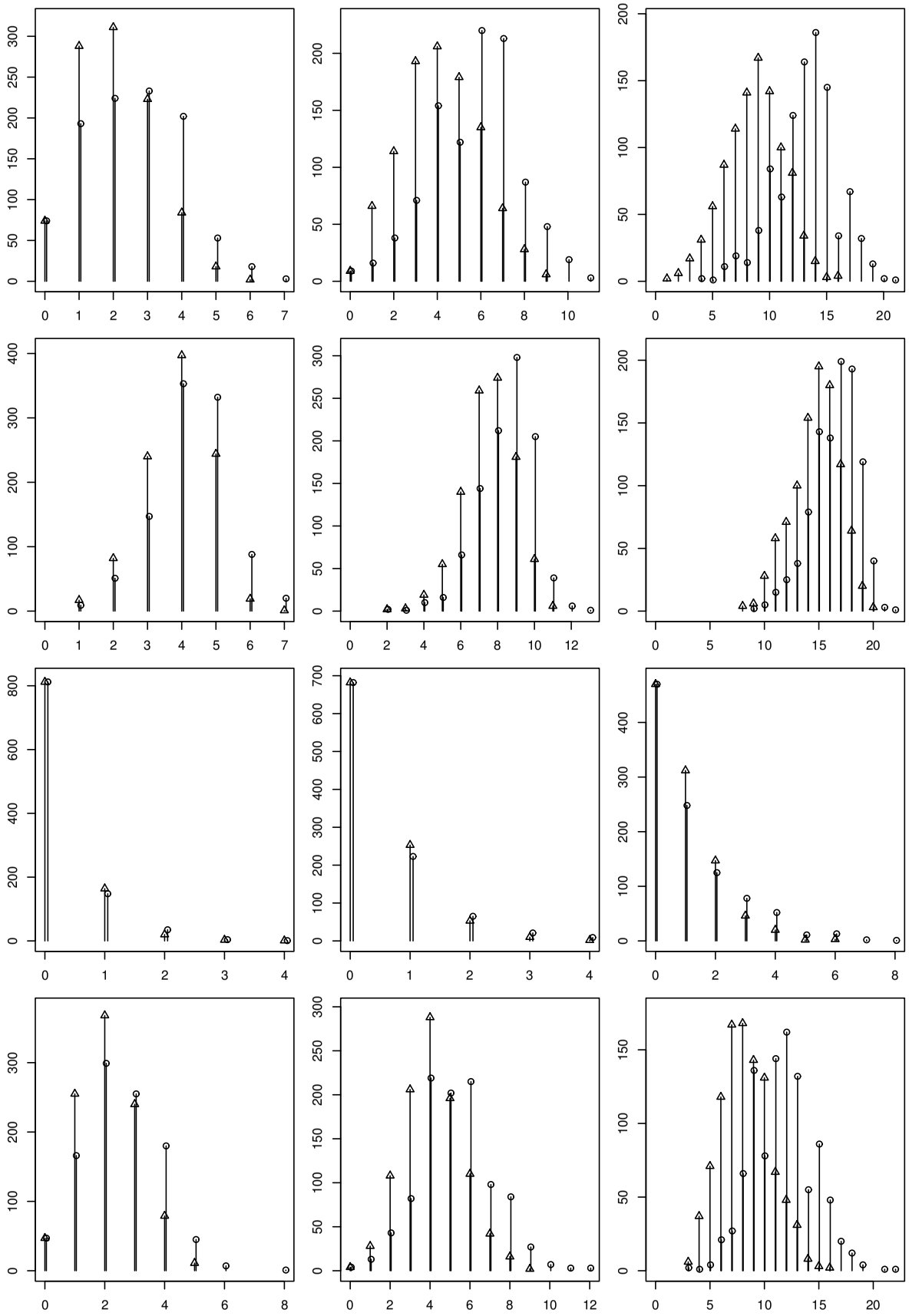


FIG. 3.4 – Taux de rejet des hypothèses nulles obtenus pour chacune des deux règles de décision. et la fonction de score S_3 ($E[S_3] = -6$). Légende : voir texte page 85

Bibliographie

Altschul, S, F., Bundschuh, R., Olsen, R., and Hwa, T. (2001). The estimation of statistical parameters for local alignment score distributions. *Nucleic Acids Res J*, 29 :351–361.

Altschul, S, F., Gish, W., Miller, W., Myers, E, W., and Lipman, D, J. (1990). Basic local alignment search tool. *J Mol Biol J*, 215 :403–410.

Bates, Joseph, L. and Constable, Robert, L. (1985). Proofs as programs. *ACM Trans. Program. Lang. Syst.*, 7 :113–136.

Coles, S. (2001). *An introduction to statistical modeling of extreme values.* Series in Statistics. Springer, London.

Daudin, J., Etienne, M., and Vallois, P. (2003). Asymptotic behaviour of the local score of independant and identically distributed random sequence. *Stochastic Processes and their Applications*, 107 :1–28.

Daudin, J. and Mercier, S. (1999). Distribution exacte du score local d'une suite de variables indépendantes et identiquement distribuées. *C. R. Acad. Sciences*, 9 :815–820. Série I, Math.

Dayhoff, M., O., Schwartz, R., M., and Orcutt, B., C. (1978). *Atlas of protein sequece and structure, supplement 3*, chapter A model of evolutionary change in proteins, pages 345–352. National Biomedical Research Foundation.

Dembo, A. and Karlin, S. (1991a). Strong limit theorems of empirical distributions for large segmental exceedances of partial sums of Markov variables. *Ann. Probab.*, 19(4) :1756–1767.

Dembo, A. and Karlin, S. (1991b). Strong limit theorems of empirical functionals for large exceedances of partial sums of i.i.d. variables. *Ann. Probab.*, 19(4) :1737–1755.

Dembo, A., Karlin, S., and Zeitouni, O. (1994). Limit distribution of maximal non-aligned two-sequence segmental score. *Ann. Probab.*, 22(4) :2022–2039.

Dupuis, D. J. (1997). Extreme value theory based on the r largest annual events : a robust approach. *Journal of Hydrology*, 200(1-4) :295–306.

Etienne, Marie, P. (2002). *Le score local : un outil pour l'analyse de séquences biologiques.* PhD thesis, Université de Nancy I.

Gabriel, K, R. (1969). Simultaneous test procedures - some theory of multiple comparisons. *Ann. Math. Stat.*, 40 :224–250.

Hampel, Frank, R., Ronchetti, Elvezio, M., Rousseeuw, Peter, J., and Stahel, Werner, A. (1986). *Robust statistics. The approach based on influence functions.* Wiley Series in Probability and Mathematical Statistics. Probability and Mathematical Statistics. John Wiley & Son, New York.

Henikoff, S. and Henikoff, J, G. (1992). Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A J*, 89 :10915–9.

Hochberg, Y. and Tamhane, Ajit, C. (1987). *Multiple comparison procedures.* Wiley Series in Probability and Mathematical Statistics. Applied Probability and Statistics. John Wiley & Sons.

Karlin, S. and Altschul, S. F. (1990). Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc. Natl. Acad. Sci. USA*, 87(6) :2264–2268.

Karlin, S. and Altschul Stephen, F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci.*, 90 :5873–5877.

Karlin, S. and Dembo, A. (1992). Limit distributions of maximal segmental score among Markov-dependent partial sums. *Adv. Appl. Probab.*, 24(1) :113–140.

Kyte, J. and Doolittle, R, F. (1982). A simple method for displaying the hydropathic character of a protein. *J Mol Biol J*, 157 :105–132.

Lawless, J, F. (1982). *Statistical models and methods for lifetime data.* Probability and Mathematical Statistics. John Wiley & Sons, New York.

Mercier, S. (1999). *Statistiques des scores pour l'analyse et la comparaison de séquences biologiques.* PhD thesis, Université de Rouen.

Mercier, S. and Daudin, J. (2001). Exact distribution for the local score of one i.i.d random sequence. *Journal of computational biology*, 8 :373–380.

Muegge, I. and Martin, Y, C. (1999). A general and fast scoring function for protein-ligand interactions : a simplified potential approach. *J Med Chem J*, 42 :791–804.

Olsen, R., Bundschuh, R., and Hwa, T. (1999). Rapid assessment of extremal statistics for gapped local alignment. *Proc Int Conf Intell Syst Mol Biol J*, pages 211–222.

Oncel, S. Y., Ahsanullah, M., Aliev, F. A., and Aygun, F. (2005). Switching record and order statistics via random contractions. *Statistics & Probability Letters*, 73(3) :207–217.

Ruzzo, W, L. and Tompa, M. (1999). A linear time algorithm for finding all maximal scoring subsequences. In *Seventh International Conference on Intelligent Systems for Molecular Biology*, pages 234–241, Heidelberg, Germany.

Smith, R. L. (1986). Extreme value theory based on the r largest annual events. *Journal of Hydrology*, 86(1-2) :27–43.

Waterman, M. (1995). *Introduction to Computational Biology : Maps, sequences and genomes*. Chapman & Hall.

Weissman, I. (1978). Estimation of parameters and large quantiles based on the k largest observations. *J. Am. Stat. Assoc.*, 73 :812–815.

Chapitre 4

Étude des retournements - Cas de retournements de longueur connues

4.1 Introduction

Dans ce chapitre, nous allons essayer d'identifier des segments retournés de faible longueur dans une chaîne donnée. Ce retournement peut éventuellement être accompagné par un passage au complémentaire pour les séquences d'ADN. La séquence est alors supposée être parsemée de petits segments du brin complémentaire.

La principale idée de la démarche consiste à utiliser le fait qu'une chaîne de Markov est orientée. La séquence $X = X_1...X_n$ est modélisée par une chaîne de Markov d'ordre m fixé a priori. Dans ce cas, comme on l'a montré dans le premier chapitre (2.5 page 40), la séquence du brin complémentaire notée $X^- = X_1^-, ..., X_n^-$ avec $X_1^- = \bar{X}_n$, $X_2^- = \bar{X}_{n-1}...$ est également une chaîne de Markov d'ordre m, mais dont la distribution sera différente de celle X, excepté dans le cas particulier où X est une chaîne de Markov réversible. La matrice de transition Q^- de X^-, ainsi que sa distribution stationnaire μ^- sont données en fonction des paramètres Q et μ de X, ci-dessous :

$$\mu^-(x_1...x_m) = \mu(\bar{x}_m...\bar{x}_1) \tag{4.1}$$

$$\mathrm{Q}^-(x_1...x_{m+1}) = \mathrm{Q}(\bar{x}_{m+1}...\bar{x}_1) \times \frac{\mu(\bar{x}_{m+1}...\bar{x}_2)}{\mu^-(x_1...x_m)} \tag{4.2}$$

L'idée est de choisir a priori une taille de segment l, et de regarder si chaque segment de taille l de la chaîne est plutôt issu de la distribution de X ou de $\bar{X}$. Pour cela, une statistique de rapport de vraisemblance est utilisée.

Dans une première partie, la distribution exacte et asymptotique de cette statistique est donnée dans le cas où un seul segment est considéré. Nous étudions ensuite le cas où

tous les segments possibles de taille l sont considérés, par une approche de type "fenêtre glissante". Lors de la détermination du degré de significativité du résultat obtenu, un problème de test multiple apparaît, puisque $n-l+1$ différents segments de tailles l sont présents dans une séquence de taille n. Pour pallier ce problème, et puisque nous nous intéressons aux valeurs élevées du rapport de vraisemblance, la distribution du maximum de cette statistique du rapport de vraisemblance sera étudiée. Cette partie repose sur une adaptation de l'approximation de cette distribution proposée par Glaz and Balakrishnan (1999) dans le cadre de sommes glissantes de données discrètes et i.i.d.

4.2 Test de retournement sur un segment

4.2.1 Hypothèse testée

On considère une séquence observée $x = x_1, ..., x_l$ de longueur l. Il s'agit de savoir si ce segment est issu de X ou de $\bar{X}$. Les paramètres de ces lois sont supposées connus.

Formellement, on se propose d'effectuer le test suivant :

H_0 : la séquence x est issue d'une chaîne de Markov X d'ordre m stationnaire de matrice de transition Q connue

contre

H_1 : la séquence x est issue de la chaîne de Markov $\bar{X}$ de matrice de transition $\bar{\mathrm{Q}}$, où $\bar{X}$ est la chaîne X inversée avec passage au complémentaire.

Dans la suite, on ne considérera que des séquences x de probabilités non nulles sous H_0 comme sous H_1. Dans le cas contraire, il est facile de savoir si la séquence est issue de X ou $\bar{X}$ et la conclusion du test est évidente.

4.2.2 Statistique de test

On utilisera pour cela la statistique du rapport de vraisemblance T

$$T = \log \frac{\bar{\mathbb{P}}(x_1, ..., x_l)}{\mathbb{P}(x_1, ..., x_l)}$$

Écritures équivalentes de T

$$\begin{aligned} T &= \log \frac{\bar{\mu}(X_1, ..., X_m)}{\mu(X_1, ..., X_m)} + \sum_{k=1}^{l-m} \log \frac{\bar{\mathrm{Q}}(X_k, ..., X_{k+m})}{\mathrm{Q}(X_k, ..., X_{k+m})} \\ T &= \sum_{w_1 \in \mathcal{A}, ..., w_m \in \mathcal{A}} \mathrm{I}_{\{X_1 = w_1, ... X_m = w_m\}} \log \frac{\bar{\mu}(w_1, ..., w_m)}{\mu(w_1, ..., w_m)} + \\ & \qquad \sum_{w_1 \in \mathcal{A}, ..., w_{m+1} \in \mathcal{A}} N^{w_1 ... w_{m+1}} \log \frac{\bar{\mathrm{Q}}(w_1, ..., w_{m+1})}{\mathrm{Q}(w_1, ..., w_{m+1})} \end{aligned}$$

Dans cette dernière écriture de la statistique T, la somme s'effectue sur tous les mots de probabilités non nulles sous H_0.

4.2.3 Distribution de la statistique de test

Le nombre de mots de taille l étant fini, la statistique T a un support discret. Calculer sa loi ne présente en théorie aucune difficulté. Il suffit de :

1. Générer tous les segments de taille l possible,
2. Pour chaque segment, calculer la valeur prise par la statistique T,
3. Ainsi que la probabilité d'observer ce segment.

Néanmoins, le nombre de points de support augmente exponentiellement avec l. On dénombre $|\mathcal{A}|^l$ mots différents. Deux mots différents peuvent conduire au même rapport de vraisemblance, comme c'est le cas, entre autres, de tous les mots symétriques de la forme $w_1w_2w_3...\bar{w}_3\bar{w}_2\bar{w}_1$. Finalement, le nombre de points de support est inférieur au nombre de mots différents. Si l est pair, il y a $|\mathcal{A}|^{l/2}$ mots symétriques, et il y en a $|\mathcal{A}|^{(l+1)/2}$ sinon. La table 4.1 suivante donne le nombre approximatif de mots différents pour un alphabet de taille 4 en fonction de la taille l du mot.

TAB. 4.1 – Nombre de mots non symétriques différents pour un alphabet de taille 4 en fonction de la taille du mot.

Taille du mot	5	10	15	20	30
Nombre de mots non symétriques	960	1.047×10^6	1.074×10^9	1.100×10^{12}	1.153×10^{18}

La loi exacte de la statistique devient rapidement inutilisable en pratique. Pour se donner une idée, la limite observée sur l'ordinateur que j'utilise actuellement, un PC muni de 512Mo de mémoire, est de $l = 11$. On utilisera donc une approximation de cette loi.

Proposition 4.1 *Soit $x = x_1, ..., x_l$ une réalisation d'une chaîne de Markov X. On note $\mathbb{P}()$ la vraisemblance de x selon la chaîne X et $\bar{\mathbb{P}}()$ la vraisemblance de x selon la chaîne $\bar{X}$. Et on définit :*

$$T = \log \frac{\bar{\mathbb{P}}(x_1, ..., x_l)}{\mathbb{P}(x_1, ..., x_l)}$$

Alors T tend en loi vers une gaussienne d'espérance θ et de variance τ donnée ci-dessous.

Preuve. T s'écrit comme une combinaison linéaire des comptages des mots de taille $m+1$. Or, le vecteur de ces comptages tend en distribution vers une gaussienne multivariée d'espérance et de variance connues, d'après le théorème 2.17 énoncé page 34. Finalement, T tend également en loi vers une gaussienne. ■

Calcul de l'espérance et de la variance de T.

$$\begin{aligned}\theta = \mathbb{E}(T) &= \mathbb{E}\left(\log\frac{\bar{\mu}(X_1,...,X_m)}{\mu(X_1,...,X_m)}\right) + \\ &\quad \sum_{w_1\in\mathcal{A},...,w_{m+1}\in\mathcal{A}} \mathbb{E}(N^{w_1...w_{m+1}})\log\frac{\bar{\mathrm{Q}}(w_1,...,w_{m+1})}{\mathrm{Q}(w_1,...,w_{m+1})} \\ &= \sum_{w_1\in\mathcal{A},...,w_m\in\mathcal{A}} \mathbb{E}(\mathrm{I}_{\{X_1=w_1,...X_m=w_m\}})\log\frac{\bar{\mu}(w_1,...,w_m)}{\mu(w_1,...,w_m)} + \\ &\quad \sum_{w_1\in\mathcal{A},...,w_{m+1}\in\mathcal{A}} \mathbb{E}(N^{w_1...w_{m+1}})\log\frac{\bar{\mathrm{Q}}(w_1,...,w_{m+1})}{\mathrm{Q}(w_1,...,w_{m+1})}\end{aligned}$$

$$\begin{aligned}\tau = \mathbb{V}(T) &= \mathbb{V}\left(\sum_{w_1\in\mathcal{A},...,w_m\in\mathcal{A}} \mathrm{I}_{\{X_1=w_1,...X_m=w_m\}}\log\frac{\bar{\mu}(w_1,...,w_m)}{\mu(w_1,...,w_m)} + \right. \\ &\quad \left. \sum_{w_1\in\mathcal{A},...,w_{m+1}\in\mathcal{A}} N^{w_1...w_{m+1}}\log\frac{\bar{\mathrm{Q}}(w_1,...,w_{m+1})}{\mathrm{Q}(w_1,...,w_{m+1})}\right) \\ &= \sum_{w'_1\in\mathcal{A},...,w'_m\in\mathcal{A}}\sum_{w_1\in\mathcal{A},...,w_{m+1}\in\mathcal{A}} \mathrm{Cov}(\mathrm{I}_{\left\{X_1=w'_1,...X_m=w'_m\right\}}, N^{w_1...w_{m+1}}) \times \\ &\quad \log\frac{\bar{\mu}(w'_1,...,w'_m)}{\mu(w'_1,...,w'_m)} \times \log\frac{\bar{\mathrm{Q}}(w_1,...,w_{m+1})}{\mathrm{Q}(w_1,...,w_{m+1})} + \\ &\quad \sum_{w'_1\in\mathcal{A},...,w'_m\in\mathcal{A}}\sum_{w_1\in\mathcal{A},...,w_m\in\mathcal{A}} \mathrm{Cov}(\mathrm{I}_{\left\{X_1=w'_1,...X_m=w'_m\right\}}, \mathrm{I}_{\{X_1=w_1...X_m=w_{m+1}\}}) \times \\ &\quad \log\frac{\bar{\mu}(w'_1,...,w'_m)}{\mu(w'_1,...,w'_m)} \times \log\frac{\bar{\mu}(w_1,...,w_m)}{\mu(w_1,...,w_m)} + \\ &\quad \sum_{w'_1\in\mathcal{A},...,w'_{m+1}\in\mathcal{A}}\sum_{w_1\in\mathcal{A},...,w_{m+1}\in\mathcal{A}} \mathrm{Cov}(N^{w'_1...w'_{m+1}}, N^{w_1...w_{m+1}}) \times \\ &\quad \log\frac{\bar{\mathrm{Q}}(w'_1,...,w'_{m+1})}{\mathrm{Q}(w'_1,...,w'_{m+1})} \times \log\frac{\bar{\mathrm{Q}}(w_1,...,w_{m+1})}{\mathrm{Q}(w_1,...,w_{m+1})}\end{aligned}$$

On sait calculer toutes les espérances et covariances nécessaires (voir 2.4.4 page 33), on a donc déterminé complètement la loi asymptotique de T lorsque la taille de la fenêtre l tend vers l'infini.

Remarque 4.2 *On est capable de calculer la loi exacte et approchée de T sous H_0, mais également sous H_1 ("le segment est issu du brin complémentaire") ce qui nous permettra de calculer la puissance du test.*

4.3 Test de retournement dans une séquence par fenêtre glissante

4.3.1 Démarche

Nous considérons une séquence de longueur n dans laquelle nous cherchons à localiser des retournements de longueur l.

Pour cela, on définit la chaîne :

$$\forall i \in 1,..,n-l+1,\ T_i = \log \frac{\bar{\mathbb{P}}(x_i,...,x_{i+l-1})}{\mathbb{P}(x_i,...,x_{i+l-1})}$$

T_i est la statistique du rapport de vraisemblance étudiée précédemment, pour le segment de longueur l, $X_i...X_{i+l-1}$

On s'intéresse aux valeurs élevées de T_i. Le graphique de la "trajectoire" de T_i est un bon indicateur pour détecter un retournement. On désire savoir si un "pic" dans la trajectoire est trop improbable pour qu'il n'y ait pas eu de retournement.

On formule les hypothèses suivantes :
H_0 : la séquence est issue d'une chaîne de Markov X d'ordre m stationnaire de matrice de transition Q connue

contre

H_1 : la séquence est issue de la chaîne de Markov X excepté un ou plusieurs segments de longueur l issue de $\bar{X}$ où $\bar{X}$ est la chaîne X inversée avec passage au complémentaire.

Nous cherchons des valeurs anormalement élevées de T_i. La statistique de test suivante :

$$S_l^n = \max_{i=1,n-l+1} T_i$$

sera étudiée dans une première partie. Une autre statistique intéressante, est le nombre de dépassements d'une certaine valeur seuil u :

$$U_u^n = \sum_{i=1}^{n-l+1} \mathrm{I}_{\{T_i \geqslant u\}}$$

Dans un premier temps, nous étudierons la chaîne des $T_i, i = 1, n-l+1$. Puis nous étudierons le comportement des statistiques de test S_l^n et U_u^n.

4.3.2 Étude de la chaîne des $T_i,\ i = 1,...,n-l+1$

L'ensemble des $T_i,\ i = 1,...,n-l+1$ définit une chaîne dont nous nous proposons d'étudier la loi ici.

Stationnarité

La séquence X_i étant supposée stationnaire, la série des T_i l'est également.

Fonction d'autocovariance des T_i

La variance de $T_i,\ i = 1, ..., n$ a déjà été calculée auparavant. En suivant le même principe, on peut calculer l'autocovariance au rang j, noté $\rho(j)$ et définie par

$$\forall i = 1, ..., n - l - j + 1\ \rho(j) = \mathrm{Cov}(T_i, T_{i+j})$$

On a,

$$\begin{aligned}
\mathrm{Cov}(T_1, T_{1+j}) \;=\; & \mathrm{Cov}\left(\sum_{w_1 \in \mathcal{A}, ..., w_m \in \mathcal{A}} \mathrm{I}_{\{X_1 = w_1, ... X_m = w_m\}} \log \frac{\bar{\mu}(w_1, ..., w_m)}{\mu(w_1, ..., w_m)} + \right. \\
& \sum_{w_1 \in \mathcal{A}, ..., w_{m+1} \in \mathcal{A}} N_{1,l}^{w_1 ... w_{m+1}} \log \frac{\bar{\mathrm{Q}}(w_1, ..., w_{m+1})}{\mathrm{Q}(w_1, ..., w_{m+1})}, \\
& \sum_{w'_1 \in \mathcal{A}, ..., w'_m \in \mathcal{A}} \mathrm{I}_{\left\{X_{1+j} = w'_1, ... X_{j+m} = w'_m\right\}} \log \frac{\bar{\mu}(w'_1, ..., w'_m)}{\mu(w'_1, ..., w'_m)} + \\
& \left. \sum_{w'_1 \in \mathcal{A}, ..., w'_{m+1} \in \mathcal{A}} N_{1+j,l}^{w_1 ... w_{m+1}} \log \frac{\bar{\mathrm{Q}}(w'_1, ..., w'_{m+1})}{\mathrm{Q}(w'_1, ..., w'_{m+1})} \right)
\end{aligned}$$

En développant cette expression autour des signes "+", et en extrayant les signes "Σ", on obtient les quatre covariances élémentaires suivantes :

$$\begin{aligned}
(1) \quad A &= \mathrm{Cov}\left(\mathrm{I}_{\{X_1 = w_1, ... X_m = w_m\}}, \mathrm{I}_{\left\{X_{1+j} = w'_1, ... X_{j+m} = w'_m\right\}} \right) \\
(2) \quad B &= \mathrm{Cov}\left(N_{1,l}^{w_1 ... w_{m+1}}, N_{1+j,l}^{w'_1 ... w'_{m+1}} \right) \\
(3) \quad C &= \mathrm{Cov}\left(\mathrm{I}_{\{X_1 = w_1, ... X_m = w_m\}}, N_{1+j,l}^{w'_1 ... w'_{m+1}} \right) \\
(4) \quad D &= \mathrm{Cov}\left(N_{1,l}^{X_1 = w_1, ... X_{m+1} = w_{m+1}}, \mathrm{I}_{\left\{X_{j+1} = w'_1, ... X_{j+m} = w'_m\right\}} \right)
\end{aligned}$$

que l'on sait calculer.

4.3.3 Étude de $S_l^n = \max(T_i,\ i = 1, ..., n - l + 1)$

Les retournements seront suspectés lorsque T_i sera anormalement élevé. Nous allons donc nous intéresser à la loi de $S_l^n = \max(T_i,\ i = 1, ..., n-l+1)$ sous l'hypothèse H_0 : "il n'y a pas de retournement". Pour cela, nous utilisons deux approches différentes. La première

s'appuiera sur l'approximation gaussienne de la loi de T_i, et utilisera les résultats sur la distribution du maximum d'un processus gaussien. La deuxième approche est inspirée des travaux sur les "statistiques par fenêtre glissante" de Glaz and Balakrishnan (1999) et utilise le principe des "fenêtres" de taille l pour approcher la loi de S_l^n.

Maximum d'un processus gaussien

Dans le cas où la taille l de la fenêtre est élevée, la distribution de T_i peut être approchée par une loi gaussienne. Les propriétés du maximum d'un processus gaussien sont connues, et on rappelle dans les paragraphes ci-dessous les résultats sur la distribution de ce maximum.

Formules de Rice On considère un processus stochastique $(Y(t))$ à trajectoires presque sûrement de classe $\mathcal{C}^1$, où t varie dans l'intervalle $[a, b]$ de $\mathbb{R}$.

On note r sa fonction de covariance, et λ_0 et λ_2 les moments spectraux d'ordre 0 et 2 de ce processus. Rappelons que si r est deux fois dérivable en 0, on a :
$\lambda_0 = r(0)$ et $\lambda_2 = r''(0)$.

On définit H_u le nombre de franchissements de u vers le haut de la trajectoire du processus Y :

$$H_u = \#\{t : t \in [a, b], Y(t) = u, Y'(t) > 0\}$$

Une approximation de la distribution de $\sup_{[0,L]} Y_t$ provient de la décomposition suivante de l'événement $\{\sup_{[0,L]} Y_t \geqslant u\}$: soit le point de départ de la trajectoire est supérieur à u, soit il est inférieur et la trajectoire "franchit" nécessairement u :

$$\{\sup_{[0,L]} Y_t \geqslant u\} = \{Y_0 \geqslant u\} \bigcup \{Y_0 < u, H_u \geqslant 1\}.$$

On en déduit que :

$$\mathbb{P}(\sup_{[0,L]} Y_t \geqslant u) \leqslant \mathbb{P}(Y_0 \geqslant u) + \mathbb{P}(H_u \geqslant 1).$$

L'inégalité de Markov nous permet de majorer cette quantité :

$$\mathbb{P}(\sup_{[0,L]} Y_t \geqslant u) \leqslant \mathbb{P}(Y_0 \geqslant u) + \mathbb{E}(H_u).$$

C'est la borne de Davis dans le cas d'un processus gaussien stationnaire centré à trajectoires presque sûrement de classe C^1

Théorème 4.3 *(**Formule de Rice, cas gaussien stationnaire**) Soit $\{Y_t, t \in \mathbb{R}\}$ un processus gaussien, stationnaire et centré à trajectoires presque sûrement de classe C^1. L'espérance de la variable aléatoire H_u définie par*

$$H_u = \#\{t : t \in [a,b], Y(t) = u, Y'(t) > 0\}$$

est donnée par

$$\mathbb{E}\left(H_u\right) = \frac{b-a}{2\pi}\sqrt{\frac{\lambda_2}{\lambda_0}}\exp\left(-\frac{u^2}{2\lambda_0}\right).$$

Une majoration de $\mathbb{P}(\sup_{[0,L]} Y_t > u)$ se déduit directement des formules de Rice :

$$\mathbb{P}(\sup_{[0,L]} Y_t > u) \leqslant 1 - \Phi\left(\frac{u}{\sqrt{\lambda_0}}\right) + \frac{L}{\sqrt{2\pi}}\sqrt{\frac{\lambda_2}{\lambda_0}}\phi\left(\frac{u}{\sqrt{\lambda_0}}\right)$$

où Φ et ϕ désignent respectivement la fonction de répartition et la densité de la loi gaussienne centrée réduite.

Application à notre problème Le processus des T_i est à temps discret. On pose $Z_t = T_i - \mathbb{E}(T_i)$ avec $t = i/n$.
Z_t est alors défini sur l'intervalle $[0,1]$, et devient un processus continu quand n tend vers l'infini.
En utilisant l'approximation de T_i par une loi gaussienne (c'est à dire que la largeur de la fenêtre l est grande), on peut appliquer le résultat précédent sur la distribution de $\sup_{[0,L]} Z_t$. Pour cela, on doit estimer la fonction de covariance r_Z de Z, ainsi que ses deux premières dérivées au point 0.

$$\begin{aligned} r_Z(h) &= \mathrm{Cov}(Z_t, Z_{t+h}) \\ &= \mathrm{Cov}(T_{nt} - \mathbb{E}(T_{nt}), T_{nt+nh} - \mathbb{E}(T_{nt+nh}) \\ &= r_T(nh) \\ \lambda_0 = r_Z(0) &= r_T(0) \end{aligned}$$

On estime maintenant les dérivées de $r_Z(h)$ au point 0. On a

$$\begin{aligned} r'_z(0) &= \lim_{h\to 0} \frac{r_Z(0) - r_Z(h)}{h} \\ &\approx -\frac{r_T(0) - r_T(1)}{1/n} \\ r'_z(1/n) &= \lim_{h\to 1/n} \frac{r_Z(1/n) - r_Z(h)}{1/n - h} \\ &\approx -\frac{r_T(1) - r_T(2)}{1/n} \\ \lambda_2 = r''_z(0) &= \lim_{h\to 0} \frac{r'_Z(0) - r'_Z(h)}{h} \\ &\approx -\frac{r'_Z(0) - r'_Z(1/n)}{1/n} \end{aligned}$$

Approche par "statistique par fenêtre glissante"

Dans le livre "*Scan Statistics and Application*" (page 30-40) Glaz and Balakrishnan (1999) s'intéressent à la scan statistique valant la somme glissante $\{Y_i + ... + Y_{i+l-1}\}$ où Y_i est une variable aléatoire prenant ses valeurs dans $\mathbb{N}$. Une revue des approximations possibles pour la fonction de répartition du maximum de cette somme glissante est donnée. Les résultats présentés dans la suite appliquent ces approximations à notre problème.

On reformule l'événement $A = \{S_l^n < s\}$ à l'aide de $B_i = \bigcap_{j=1}^{l+1}\{T_{(i-1)l+j} < s\}, i = 1, .., k-1$. On suppose qu'il existe k tel que $n = k \times l$.
On a
$B_1 = \{T_1 < s\} \bigcap \{T_2 < s\} \bigcap ... \bigcap \{T_{l+1} < s\}$
$B_2 = \{T_{l+1} < s\} \bigcap \{T_{l+2} < s\} \bigcap ... \bigcap \{T_{2l+1} < s\}$
$B_3 = \{T_{2l+1} < s\} \bigcap \{T_{2l+2} < s\} \bigcap ... \bigcap \{T_{3l+1} < s\}$
etc.

On réécrit :

$$\{S_l^n < s\} = \bigcap_{i=1}^{k-1} \{B_i\}$$

d'où

$$\begin{aligned} P(S_l^n \leqslant s) &= P(\textstyle\bigcap_{i=1}^{k-1}\{B_i\}) \\ &= P(B_1 \bigcap B_2) \prod_{i=3}^{k-1} P(B_i | \bigcap_{j=1}^{i-1} B_j). \end{aligned}$$

Pour simplifier cette expression, on regarde le lien qui existe entre les B_i et les X_i initiaux. On s'aperçoit que B_i est une fonction de $2l$ termes : $X_{(i-1)l+1}, ..., X_{(i+1)l}$, que B_i et B_{i+1} ont en commun l termes : $X_{il+1}, ..., X_{(i+1)l}$, et que B_i et B_{i+2} n'ont aucun terme en

commun. La dépendance qui existe entre B_i et B_{i+2} n'est due qu'au caractère markovien de X et devient négligeable quand l tend vers ∞. On obtient alors :

$$\begin{aligned} P(S_l^n \leqslant s) &\approx P(B_1 \bigcap B_2) \prod_{i=3}^{k-1} P(B_i|B_{i-1}) \\ &\approx P(B_1 \bigcap B_2) \prod_{i=3}^{k-1} P(B_i \bigcap B_{i-1})/P(B_i) \end{aligned}$$

De plus, $P(B_i) = \mathbb{P}(S_l^{2l} < s)$ et $P(B_i \bigcap B_{i-1}) = \mathbb{P}(S_l^{3l} < s)$.

Ce qui donne l'approximation "Product-type" suivante :

$$\mathbb{P}(S_l^n \leqslant s) \approx \mathbb{P}(S_l^{3l} < s) \left(\frac{\mathbb{P}(S_l^{3l} < s)}{\mathbb{P}(S_l^{2l} < s)} \right)^{k-3}$$

On exprime ainsi la fonction de répartition du maximum de la statistique sur une séquence de longueur n à l'aide des fonctions de répartition du maximum de cette statistique sur une séquence de longueur $2l$ et $3l$ avec $l << n$.

Les fonctions de répartition de S_l^{2l} et S_l^{3l} ne sont pas connues. Étant donné que l est relativement petit, on propose de les estimer par Monte Carlo.

Cependant, nous sommes intéressés aux valeurs s_0 telles que $\mathbb{P}(S_l^n > s_0) < \alpha$ avec α petit (de l'ordre de 5%, par exemple). Étant données, les inégalités stochastiques $S_l^{2l} \leqslant S_l^{3l} \leqslant S_l^n$ et les tailles relatives de l et n ($l << n$), les probabilités $\mathbb{P}(S_l^{2l} > s_0)$ et $\mathbb{P}(S_l^{3l} > s_0)$ seront très faibles. Par conséquent, beaucoup d'itérations de Monte Carlo seront nécessaires pour estimer correctement les queues de ces distributions. Pour contourner ce problème, on utilisera la technique d'**échantillonage pondéré** suivante.

Échantillonage Pondéré (Importance Sampling) On veut estimer par Monte Carlo $\varepsilon = \int_{\mathbb{R}} \mathrm{I}_{\{y \in \mathcal{D}\}} dP(y)$
où Y est une variable aléatoire de loi $P(y)$ et $\mathcal{D}$ est un certain domaine de $\mathbb{R}$.

Si $P(\mathcal{D})$ est petit, beaucoup d'itérations de Monte Carlo sont nécessaires pour obtenir une précision relative satisfaisante de cette quantité. Pour pallier ce problème, on cherche un loi $\tilde{P}$ telle que $\tilde{P}(\mathcal{D})$ soit suffisamment grand (0.5 par exemple) et on définit $g(y) = \frac{d\tilde{P}}{dP}$.

On a alors $\varepsilon = \int_{\mathbb{R}} \frac{\mathrm{I}_{\{y \in \mathcal{D}\}}}{g(y)} d\tilde{P}(y)$

Pour estimer ε :

1. On simule un échantillon $\{y_i\}_{i=1,\ldots,p}$ selon $\tilde{P}$
2. $\hat{\varepsilon} = p^{-1} \sum_{i=1}^{p} \frac{\mathrm{I}_{\{y_i \in \mathcal{D}\}}}{g(y_i)}$

Pour avoir plus de détails sur la technique *d'échantillonage pondéré*, on pourra consulter les références Ripley (1987); Green (1992).

Dans notre cas, on veut "favoriser" les tirages des grandes valeurs de la statistique. Pour cela, on tirera la chaîne de Markov selon $\bar{\mathrm{Q}}$.

On a donc $g(x) = \frac{\bar{\mathbb{P}}(x)}{\mathbb{P}(x)}$.
Pour estimer $\mathbb{P}(S_l^{2l} > s_0)$:

1. On simule une réalisation d'une chaîne de Markov de taille $2l$ ou $3l$ $\{x_i\}_{i=1...}$ selon $\bar{\mathrm{Q}}$
2. On calcule la série des T_i dont on conserve le maximum s_j.
3. On calcule la valeur de $g_j = g(x) = \frac{\bar{\mathbb{P}}(x)}{\mathbb{P}(x)}$
4. On répète les trois points précédents jusqu'à obtenir un nombre p suffisant de couples (s_j, g_j)
5. On estime $\mathbb{P}(S_l^{2l} > s_0)$ (ou $\mathbb{P}(S_l^{3l} > s_0)$) par : $\hat{\varepsilon} = \frac{1}{p}\sum_{j=1}^{p} \frac{1}{g_j} \mathrm{I}_{\{s_j > s_0\}}$

Illustration de l'approximation Product-type L'approximation product-type proposée ci-dessus a d'abord été proposée pour des variables aléatoires à valeurs dans $\mathbb{N}$ d'une part, et indépendantee et identiquement distribuées d'autre part. Dans notre cas, les variables élémentaires auxquelles nous nous intéressons dans la somme glissante, ne sont ni entières, ni positives, ni indépendantes, car la séquence est markovienne.

Afin d'évaluer l'impact du non-respect de ces conditions, nous proposons une courte étude fondée sur des simulations.

Nous simulons pour cela des séquences markoviennes d'ordre 1. La matrice de transition choisie est celle estimée sur le génome complet du virus HIV1 et présentée ci dessous :

$$\mathrm{Q} = \begin{pmatrix} 0.326 & 0.300 & 0.165 & 0.209 \\ 0.346 & 0.279 & 0.193 & 0.182 \\ 0.448 & 0.054 & 0.233 & 0.265 \\ 0.316 & 0.273 & 0.158 & 0.253 \end{pmatrix}$$

La distribution stationnaire associée est la suivante : $\mu = (0.351, 0.244, 0.182, 0.223)$

Nous considérons :

- Deux tailles de séquences : $n = 1000$ et 10000
- Trois tailles de fenêtres : $l = 5$, 10 et 50

Les graphiques 4.1 page 105 présentent une comparaison des fonctions de répartition obtenues par Monte-Carlo sur 10000 itérations aux fonctions de répartition obtenues à l'aide de l'approximation Product-type pour les différentes combinaisons de ces paramètres.

La qualité de l'approximation est très statisfaisante, y compris pour des faibles valeurs de la longueur de la fenêtre. De plus l'approximation "product-type" reste valable quelle que soit la distribution des variables élémentaires T_i, contrairement à l'approche précédente reposant sur la normalité de ces statistiques. Ce résultat nous encourage à conserver cette approximation.

4.3.4 Étude de $U_u^n = \sum_{i=1}^{n-l+1} \mathbf{I}_{\{T_i \geqslant u\}}$

La présence de retournements sera suspectée si la chaîne des T_i dépasse trop souvent un seuil donné u. Nous nous intéressons donc ici à la loi du nombre de dépassements

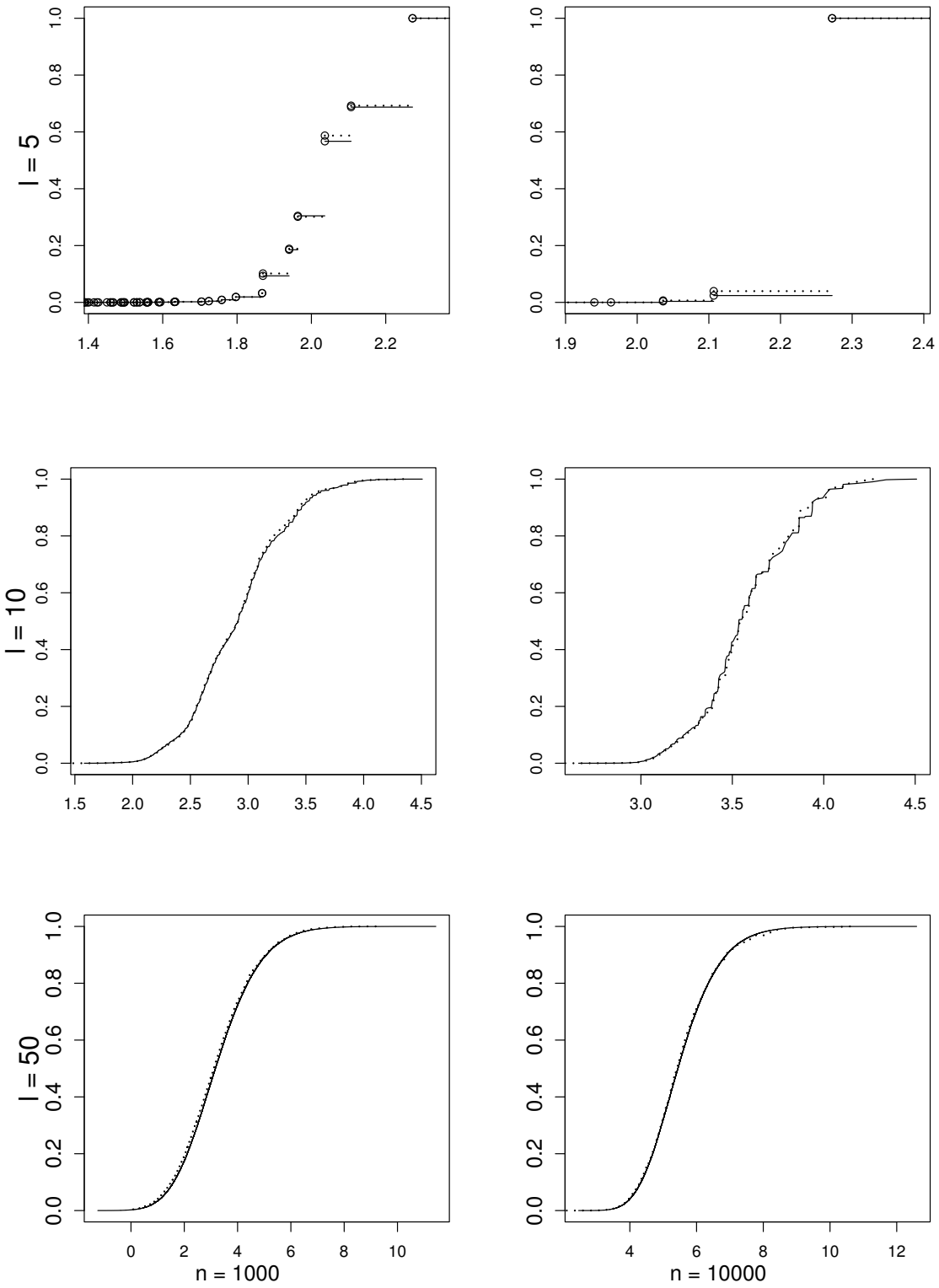


FIG. 4.1 – Qualité de l'approximation "Product-type" de la fonction de répartition de S_l^n pour une séquence markovienne de taille $n = 1000$ et 10000, et pour des fenêtres de longueurs $l = 5$, 10 et 50. La courbe pointillée désigne la vraie fonction de répartition obtenue par Monte-Carlo, et la courbe pleine désigne l'approximation "product-type".

du seuil u sous H_0. Remarquons au passage que les événements $\{U_u^n = 0\}$ et $\{S_l^n < u\}$ sont équivalents. Là encore, deux approches seront développées. La première s'appuiera sur l'approximation gaussienne de la loi de T_i, et utilisera les résultats sur la distribution du nombre de franchissements d'un seuil pour un processus gaussien. La deuxième approche est inspirée des travaux sur les "statistiques par fenêtre glissante" de Glaz and Balakrishnan (1999) et utilise le principe des "fenêtres" de taille l pour approcher la loi de U_u^n.

Approche par "scan statistique"

Une première approche On définit $I_i(s) = \begin{cases} 1 & \text{si } T_i \geqslant s \\ 0 & \text{sinon} \end{cases}$ pour i allant de 1 à $n-l+1$

On a alors $U_u^n = \sum_{i=1}^{n-l+1} I_i(u)$.

$I_i(s)$ suit une loi de Bernoulli d'espérance $\mathbb{P}(T_i \geqslant s)$. L'espérance de U_u^n vaut

$$\begin{aligned} \mathbb{E}(U_u^n) &= \mathbb{E}\left(\sum_{i=1}^{n-l+1} I_i(s)\right) \\ &= (n-l+1)\mathbb{P}(T_1 \geqslant s) \end{aligned}$$

En considérant que l'événement $I_i(s) = 1$ est de probabilité faible, et en première approximation que ces événements sont indépendants pour tout $i = 1, ..., n-l+1$, la loi de U_u^n est approchée par une loi de Poisson d'espérance $\lambda_1 = (n-l+1)\mathbb{P}(T_1 \geqslant s)$.

"Declumping" On peut améliorer l'approximation précédente en tenant compte du fait que les événements $\{I_j(s) = 1\}$ sont positivement autocorrélés. On définit $I_i^\star(s) = \begin{cases} 1 & \text{si } (I_i = 1) \bigcap \{\bigcap_{t=i-l+1}^{i-1} (I_t = 0)\} \\ 0 & \text{sinon} \end{cases}$, pour i allant de 1 à $n-l+1$. Si la chaîne dépasse plusieurs fois la valeur seuil s dans un intervalle de longueur l, la nouvelle variable $I_i^\star$ ne prend qu'une seule fois la valeur 1.

On note $U_u^{\star n} = \sum_{i=1}^{n-l+1} I_i^\star(s)$, le nombre de dépassements "ajusté". On approche cette loi par une loi de Poisson d'espérance $\lambda^\star$.

On peut calculer l'espérance de $U_u^{\star n}$, et on trouve :

$$\begin{aligned} \lambda^\star &= \mathbb{E}(U_u^{\star n}) = \mathbb{E}\left(\sum_{i=1}^{n-l+1} I_i^\star(s)\right) \\ \lambda^\star &= \mathbb{P}(S_l^{2l-2} > s) + (n-2l+2)\left(\mathbb{P}(S_l^{2l-2} < s) - \mathbb{P}(S_l^{2l-1} < s)\right). \end{aligned}$$

Le calcul de cette expression est détaillé ci-dessous.

On définit l'événement : $A_j = \{I_j = 0\}$ et on distingue les $l-1$ premiers termes des

autres :

$$\begin{aligned}\mathbb{E}(U_u^{\star n}) &= \sum_{i=1}^{n-l+1} \mathbb{E}(I_i^\star(s)) \\ &= \mathbb{P}(A_1^c) + \sum_{j=2}^{l-1} \mathbb{P}\left(A_j^c \bigcap \left(\bigcap_{i=1}^{j-1} A_i\right)\right) + (n-2l+2)\mathbb{P}\left(A_l^c \bigcap \left(\bigcap_{j=1}^{l-1} A_j\right)\right).\end{aligned}$$

Or, $\mathbb{P}(A_1^c) = \mathbb{P}(S_l^l > s)$ et,

$$\begin{aligned}\mathbb{P}\left(A_l^c \bigcap \left(\bigcap_{j=1}^{l-1} A_j\right)\right) &= \mathbb{P}\left(\bigcap_{j=1}^{l-1} A_j\right) - \mathbb{P}\left(A_l \bigcap \left(\bigcap_{j=1}^{l-1} A_j\right)\right). \\ &= \mathbb{P}(S_l^{2l-2} < s) - \mathbb{P}(S_l^{2l-1} < s)\end{aligned}$$

Selon le même principe, $\sum_{j=2}^{l-1} \mathbb{P}(A_j^c \bigcap (\bigcap_{i=1}^{j-1} A_i)) = \mathbb{P}(S_l^l < s) - \mathbb{P}(S_l^{2l-2} < s)$.
En sommant ces trois expressions, on obtient le résultat.

Remarquons que les fonctions de répartitions de S^{2l-2} et S^{2l-1} sont très proches, et que leurs estimations doivent par conséquent être précises pour utiliser cette approche.

Remarquons également l'équivalence :

$$\sum_{i=1}^{n-l+1} I_i(s) = 0 \Longleftrightarrow \sum_{i=1}^{n-l+1} I_i^\star(s) = 0.$$

Il sera donc équivalent de s'intéresser à la loi de S_n^l par le biais de U_u^n et de $U_u^{\star n}$.

En utilisant la loi de S_l^n Les événements $\{U_u^n = 0\}$ et $\{S_l^n < u\}$ étant équivalents, on a $\mathbb{P}(U_u^n = 0) = \mathbb{P}(S_l^n < u)$. Si on approche la loi de U_u^n par une loi de Poisson de paramètre λ_2, on peut écrire : $\mathbb{P}(S_l^n < u) = \exp(-\lambda_2)$.

En utilisant l'approximation de la loi de S_l^n trouvée au paragraphe 4.3.3, on obtient :

$$\begin{aligned}\exp(-\lambda_2) &= \mathbb{P}(S_l^{3l} < u)\left(\frac{\mathbb{P}(S_l^{3l} < u)}{\mathbb{P}(S_l^{2l} < u)}\right)^{k-3} \\ \text{d'où } \lambda_2 &= (k-3)\left(\log \mathbb{P}(S_l^{2l} < u) - \log \mathbb{P}(S_l^{3l} < u)\right) - \log \mathbb{P}(S_l^{3l} < u)\end{aligned}$$

où k est tel que $n = kl$.

Qualité des approximations Si l'on considère les variables T_i gaussiennes, Raab (1997) donne une majoration de la distance en variation totale entre la loi de Poisson

d'espérance λ_1 et la loi de U_u^n :

$$|| \sum_{i=1}^{n-l+1} I_i(s) \quad , \quad Po(\lambda_1(s))||_{vt} \leqslant$$
$$\frac{1-e^{-\lambda_1(s)}}{\lambda_1(s)} \left(\lambda_1(s)\mathbb{P}(T_1 > s) + \sum_{k=1}^{n-l+1} \sum_{j=1, j\neq k}^{n-l+1} |\mathrm{Cov}(I_k(s), I_j(s))| \right)$$

Pour la calculer, il suffit d'estimer les covariances des indicatrices. On rappelle également le théorème 3.4 énoncé dans sa thèse page 48 (Raab, 1997) :

Théorème 4.4 *Soit Y_k une séquence gaussienne stationnaire et standardisée de fonction de corrélation satisfaisant $\rho_k \leqslant A \ln k$, $k \geqslant 2$, pour une constante A. Soit $\rho = \max(0, \rho_1, \rho_2, ...)$. Soit a_n une suite de nombres réelles tels que $n(1-\Phi(a_n)) = \lambda$. On définit $U^n = \sum_{k=1}^n I_{\{Y_k < a_n\}}$. Alors*

$$||U^n, Po(\lambda)||_{vt} = O\left(n^{-(1-\rho)/(1+\rho)} (\ln n)^{-\rho/(1+\rho)} + \frac{1}{n} \sum_{k=1}^{n} |\rho_k| \right).$$

Ce théorème s'applique dans notre cas, la suite des ρ_k étant strictement décroissante.

Bibliographie

Glaz, J. and Balakrishnan, N. (1999). *Scan Statistics and Application*. Birkhauser.

Green, P. (1992). Discussion on "Constrained Monte Carlo maximum likelihood for dependent data" by C. J. Geyer and E. A. Thompson. *J.R. Statist. Soc B*, 54 :683–684.

Raab, M. (1997). *Number of exceedances in gaussian and related sequences*. PhD thesis, Royal Institute of Technology, Stockolm.

Ripley, B. D. (1987). *Stochastic Simulation*. Wiley series in probability and mathematical statistics. John Wiley and Sons, New York, NY, USA ; London, UK ; Sydney, Australia.

Chapitre 5

Étude des retournements - Cas de retournements de longueur inconnues

5.1 Introduction

La plupart du temps, la taille des retournements dans une séquence n'est pas connue. On pourra alors utiliser la méthode précédente en essayant plusieurs tailles de retournement. Cette approche présente plusieurs inconvénients. Le premier est que l'interprétation des résultats obtenus avec différentes tailles ne sera pas nécessairement immédiate. Le second est que l'on se confronte à nouveau à un problème de test multiple.

Par conséquent, il semble plus rigoureux de prendre en compte cette incertitude directement dans la méthode de détection. Pour cela, deux méthodes de type "score local" sont présentées dans ce chapitre. On cherchera donc des "zones" de la séquence comportant une densité importante de scores élémentaires élevés.

Dans notre cas, nous cherchons des segments retournés ; le score considéré devra être d'autant plus élevé que le nucléotide considéré à de chance d'être retourné. Dans la suite de ce chapitre, deux méthodes de scores sont mis au point. La première découle directement du chapitre précédant et peut être considérée comme une généralisation de la méthode à fenêtre glissante dans le cas où la taille de la fenêtre n'est pas connue.

Après que quelques propriétés de cette fonction de score sont mises en évidence, on verra que les hypothèses nécessaires à l'application des résultats de Karlin ne sont pas vérifiées dans notre cas. Une validation empirique des résultats est présentée.

La deuxième méthode de score a pour but de pallier cette approximation. Elle repose sur une martingale ce qui permettra d'obtenir des résultats plus rigoureux théoriquement.

5.2 Généralisation de l'approche par fenêtre glissante

5.2.1 Introduction et détermination de la fonction de score

L'approche par fenêtre glissante présentée dans le chapitre précédent repose sur la statistique T_i du rapport de vraisemblance d'un segment donnée $X_i...X_{i+l-1}$ par rapport aux deux modèles X et X^-.

$$\forall i \in \{1,..,n-l+1\},\ T_i = \ln \frac{\bar{\mathbb{P}}(x_i,...,x_{i+l-1})}{\mathbb{P}(x_i,...,x_{i+l-1})}$$

où $\mathbb{P}(x)$ est la vraisemblance de x selon la chaîne X et $\bar{\mathbb{P}}(x)$ la vraisemblance de x selon la chaîne X^-.

Or, chacune de ces deux probabilités peut se décomposer en une somme de termes élémentaires. On définit :

$$\forall i \in \{1,...,n-m\}, Y_i = \sum_{(u_1,...,u_{m+1})\in\mathcal{A}^{m+1}} \mathrm{I}_{\{(X_{i-m+1}=u_1,...,X_i=u_{m+1})\}} \ln\left(\frac{\mathrm{Q}^-(u_{m+1}|u_1,...,u_m)}{\mathrm{Q}(u_{m+1}|u_1,...,u_m)}\right).$$

De manière plus condensée, Y_i peut se réécrire :

$$Y_i = \ln\left(\frac{\mathrm{Q}^-(X_{i+1}|X_i...X_{i-m+1})}{\mathrm{Q}(X_{i+1}|X_i...X_{i-m+1})}\right).$$

La statistique du rapport de vraisemblance se réécrit alors

$$\forall i \in \{1,..,n-l+1\}, T_i = \ln\left(\frac{\mu^-(X_i)}{\mu^+(X_i)}\right) + \sum_{j=1}^{i+l-1} Y_j.$$

Cette écriture suggère de considérer la **fonction de score s** suivante dans un modèle Mm :

$$\begin{array}{rcl} \mathbf{s}: & \mathcal{A}^{\mathbf{m+1}} & \to \mathbf{R} \\ & (\mathbf{u_1},...,\mathbf{u_{m+1}}) & \to \ln\left(\frac{\mathrm{Q}^-(\mathbf{u_{m+1}}|\mathbf{u_1},...,\mathbf{u_m})}{\mathrm{Q}(\mathbf{u_{m+1}}|\mathbf{u_1},...,\mathbf{u_m})}\right). \end{array}$$

Dans le cas particulier d'un modèle de Markov d'ordre 1, la fonction de score devient :

$$\begin{array}{rcl} \mathbf{s}: & \mathcal{A}^{\mathbf{2}} & \to \mathbf{R} \\ & (\mathbf{u_1},\mathbf{u_2}) & \to \ln\left(\frac{\mathrm{Q}^-(\mathbf{u_2}|\mathbf{u_1})}{\mathrm{Q}(\mathbf{u_2}|\mathbf{u_1})}\right). \end{array}$$

Cette fonction de score associe à chaque couple de lettres successives (u_1, u_2) un score d'autant plus élevé que la probabilité d'observer u_2 conditionnellement à u_1 est grande sous le modèle du brin complémentaire, par rapport au brin "principal". Il s'agit d'un **score de retournement**.

5.2.2 Quelques propriétés de la fonction de score

Nous présentons ici quelques résultats sur la distribution de la fonction de score. Nous verrons notamment que la séquence des scores $Y_1...Y_{n-m}$ forme généralement une chaîne de Markov dans le cas où la séquence initiale $X_1...X_n$ est une chaine de Markov d'ordre m. Il faut pour cela que le retournement soit accompagné d'un passage à l'alphabet complémentaire. Dans tous les cas, la chaîne $Y_1...Y_{n-m}$ est α-mélangeante.

Le choix de cette fonction de score permet de donner un sens à la somme des scores élémentaires du segment considéré. En effet, $\forall(i,j)$ tel que $j > i+m$,

$$s(X_i,...,X_{i+m}) + ... + s(X_{j-m},...,X_j) = \ln \frac{\bar{\mathbb{P}}(X_i,...,X_j|X_i,...,X_{i+m-1})}{\mathbb{P}(X_i,...,X_j|X_i,...,X_{i+m-1})}$$

Dans le cas particulier où **l'ordre m vaut 1**, on a : $\forall(i,j)$ tel que $j > i+1$,

$$s(X_i,X_{i+1}) + ... + s(X_{j-1},X_j) = \ln \frac{\bar{\mathbb{P}}(X_i,...,X_j|X_i)}{\mathbb{P}(X_i,...,X_j|X_i)}$$

Lemme 5.1 *Si la séquence initiale $X_1...X_n$ suit une chaîne de Markov d'ordre m stationnaire, alors l'espérance du score est strictement négative :*

$$\mathbb{E}_\mu(Y_i) \leqslant 0.$$

Preuve.

$$\begin{aligned}
\mathbb{E}_\mu(Y_i) &= \mathbb{E}\left(\ln\left(\frac{\mathrm{Q}^-(X_i,X_{i+1})}{\mathrm{Q}(X_i,X_{i+1})}\right)\right) \\
&\leqslant \ln\left(\mathbb{E}\left(\frac{\mathrm{Q}^-(X_i,X_{i+1})}{\mathrm{Q}(X_i,X_{i+1})}\right)\right) \\
&= \ln\left(\sum_{u,v\in\mathcal{A}} \mu(u)\mathrm{Q}(u,v)\left(\frac{\mathrm{Q}^-(u,v)}{\mathrm{Q}(u,v)}\right)\right) \\
&\leqslant \ln\left(\sum_{u,v\in\mathcal{A}} \mu(u)\mathrm{Q}(u,v)\left(\frac{\mathrm{Q}^-(u,v)\mu(u)}{\mathrm{Q}(u,v)\mu(u)}\right)\right) \\
&\leqslant \ln\left(\sum_{u\in\mathcal{A}}\mu(u)\sum_{v\in\mathcal{A}}\mathrm{Q}^-(u,v)\right) \\
&\leqslant \ln\left(\sum_{u\in\mathcal{A}}\mu(u)\right) \\
&\leqslant 0.
\end{aligned}$$

■

Le fait que l'espérance du score soit négative est nécessaire pour appliquer les résultats de Karlin sur la loi du score local.

On s'intéresse maintenant à la dépendance de la série des scores $Y_1, \ldots, Y_n$.

Lemme 5.2 *La suite $(Y_i)_{i\geqslant 1}$ est une suite α-mélangeante.*

Preuve. Calculons tout d'abord $a_{n,k}(s,t) = \mathbb{P}(T_n = s, T_{n+k} = t)$. Par hypothèse de stationnarité de la chaîne de Markov, on a

$$\begin{aligned} a_{n,k}(s,t) &= \mathbb{P}(T_1 = s, T_{1+k} = t), \\ &= \sum_{\substack{u_1,u_2,u_3,u_4 \\ g(u_1,u_2)=s \\ g(u_3,u_4)=t}} \mu(u_1)\mathrm{Q}(u_1,u_2)\mathrm{Q}^{k-1}(u_2,u_3)\mathrm{Q}(u_3,u_4). \end{aligned} \tag{5.1}$$

D'autre part

$$\mathbb{P}(T_n = s) = \sum_{\substack{u_1,u_2 \\ g(u_1,u_2)=s}} \mu(u_1)\mathrm{Q}(u_1,u_2) \tag{5.2}$$

Donc

$$\begin{aligned} a_{n,k}(s,t) &- \mathbb{P}(T_n = s)\,\mathbb{P}(T_{n+k} = t) \\ &= \sum_{\substack{u_1,u_2,u_3,u_4 \\ g(u_1,u_2)=s \\ g(u_3,u_4)=t}} \mu(u_1)\mathrm{Q}(u_1,u_2)\mathrm{Q}^{k-1}(u_2,u_3)\mathrm{Q}(u_3,u_4) \\ &\quad - \left(\sum_{\substack{u_1,u_2 \\ g(u_1,u_2)=s}} \mu(u_1)\mathrm{Q}(u_1,u_2) \right) \left(\sum_{\substack{u_3,u_4 \\ g(u_3,u_4)=t}} \mu(u_3)\mathrm{Q}(u_3,u_4) \right) \\ &= \sum_{\substack{u_1,u_2,u_3,u_4 \\ g(u_1,u_2)=s \\ g(u_3,u_4)=t}} \mu(u_1)\mathrm{Q}(u_1,u_2)\mathrm{Q}(u_3,u_4)\left(\mathrm{Q}^{k-1}(u_2,u_3) - \mu(u_3)\right) \end{aligned} \tag{5.3}$$

Ainsi pour tout $n \geqslant 1$ et $k \geqslant 1$

$$\sup_{(s,t)\in\mathbb{R}^2} \mid a_{n,k}(s,t) - \mathbb{P}(T_n = s)\,\mathbb{P}(T_{n+k} = t) \mid \leqslant \lambda_1^k$$

où λ_1 désigne la seconde plus grande valeur propre ($\lambda_1 < 1$) de la matrice Q. ■

Proposition 5.3 *Si la fonction de score s, définie comme ci dessus en 5.2.1, est bijective alors les images par cette fonction des éléments d'une chaîne de Markov d'ordre m $X_1 \dots X_n$ forme une chaîne de Markov d'ordre m.*

Preuve. On pose $Z_i = (X_i, \dots, X_{i+m})$ qui est à valeur dans $\mathcal{A}^{m+1}$.
La séquence $Z_1, \dots, Z_{n-m}$ forment une chaîne de Markov d'ordre 1. Or, la séquence produite par une fonction f bijective appliquée aux éléments d'une chaîne de Markov est elle-aussi une chaîne de Markov. ■

Si s n'est pas bijective, la séquence des scores ne sera markovienne que si le critère de Dynkin est satisfait. Ce critère est rappelé dans la proposition 3.2 page 50.

Lemme 5.4 *Soit $X = X_1 ... X_n$ une chaîne de Markov à valeur dans l'alphabet $\mathcal{A}$ de matrice de transition Q, et $X^- = X_n ... X_1$ la chaîne de Markov dite "inversée" de matrice de transition Q^-. Alors, dans ce cas où l'inversion n'est pas accompagnée du passage à l'alphabet complémentaire, la fonction de score s telle que définie ci-dessus n'est pas bijective.*

Preuve. Pour effectuer cette démonstration, on utilise la définition "génomique" d'un palindrome.

Définition 5.5 (Palindrome) *Soit un alphabet noté $\mathcal{A}$, muni de la structure de complémentarité suivante :*

$$\forall u \in \mathcal{A}, \begin{cases} \bar{u} \in alphabet \\ \bar{\bar{u}} = u. \end{cases}$$

Soit un mot $w = w_1 ... w_h$ de taille h composé des lettres de l'alphabet $\mathcal{A}$. On note $\bar{w}$ l'inverse complémentaire du mot w :

$$\bar{w} = \bar{w}_h ... \bar{w}_1.$$

w est un palindrome si, et seulement si, il vérifie la condition suivante :

$$w = \bar{w}.$$

Par exemple, le mot "aggcct" écrit dans l'alphabet nucléique ($\bar{A} = T$ et $\bar{G} = C$) est un palindrome*. Si l'on considère un alphabet où il n'y a pas de passage au complémentaire (autrement dit, $\bar{w}_i = w_i$), comme par exemple l'alphabet des acides-aminés, un exemple de palindrome est donné par "Leu-Leu-Pro-Ala-Pro-Leu-Leu".

*Attention : un palindrome génomique n'a pas exactement la définition employée dans la langue courante.

Lemme 5.6 *Soit la fonction de score de retournement s telle que définie ci dessus en 5.2.1 dans le cas d'un modèle de Markov d'ordre m où* **l'inversion n'est pas accompagnée par un passage à l'alphabet complémentaire**. *Soit w un palindrome de taille $m+1$ défini sur le même alphabet que s. On a alors :*

$$s(w) = 0.$$

Preuve.

$$s(w) = \ln\left(\frac{\mathrm{Q}^-(w_{m+1}|w_1...w_m)}{\mathrm{Q}(w_{m+1}|w_1...w_m)}\right).$$

D'après la proposition 2.20 page 40 donnant Q^- en fonction de Q, on a

$$s(w) = \ln\left(\frac{\mu(w_{m+1}...w_2)\mathrm{Q}(w_1|w_{m+1}...w_2)}{\mu^-(w_1...w_m)\mathrm{Q}(w_{m+1}|w_1...w_m)}\right).$$

Or, $\mu^-(w_1...w_m) = \mu(w_m...w_1)$ car il n'y pas de passage à l'alphabet complémentaire. d'où

$$s(w) = \ln\left(\frac{\mu(w_{m+1}...w_2)\mathrm{Q}(w_1|w_{m+1}...w_2)}{\mu(w_m...w_1)\mathrm{Q}(w_{m+1}|w_1...w_m)}\right).$$

Or w est un palindrome, donc $(w_m...w_1) = (w_{m+1}...w_2)$ et $(w_{m+1}...w_1) = (w_1...w_{m+1})$ d'où

$$s(w) = 0.$$

■

Dès lors que la taille de l'alphabet est supérieure strictement à un (ce qui est toujours le cas), il existe au moins deux palindromes différents. Donc, s n'est pas bijective dans le cas où il n'y a pas de passage à l'alphabet complémentaire. ■

Le critère de Dynkin impose aux différentes probabilités des contraintes très fortes pour que la séquence image de la fonction reste Markovienne. Ces contraintes n'ont aucune raison d'être vérifiées dans des matrices de Markov modélisant des "vraies" séquences biologiques. Il en résulte que la séquence des scores ne sera probablement pas Markovienne dans le cas où il n'y a pas de passage au complémentaire.

Lorsqu'au contraire, le retournement se produit avec un passage à l'alphabet complémentaire, comme c'est le cas dans une molécule d'ADN, la fonction de score sera très probablement bijective pour la même raison qu'au paragraphe précédent ; il est en effet, tout à fait improbable que le rapport $\mathrm{Q}(v|u)/\mathrm{Q}^-(v|u)$ conduise à la même valeur pour deux couples de lettres (u, v) différents. On peut se demander ce qu'il advient des palindromes dans ce cas. Un exemple de score de palindrome est donné ci-dessous. On se place dans un modèle $M1$ pour plus de clarté.

Si $w = w_1w_2$ est un palidrome, on a :

$$\begin{aligned} s(w) &= \ln\left(\frac{\mathrm{Q}^-(w_2|w_1)}{\mathrm{Q}(w_2|w_1)}\right) \\ &= \ln\left(\frac{\mu(\bar{w}_2)\mathrm{Q}(\bar{w}_1|\bar{w}_2)}{\mu^-(w_1)\mathrm{Q}(w_2|w_1)}\right). \end{aligned}$$

Or, $\bar{w}_1 = w_2$ et $\mu^-(w_1) = \mu(\bar{w}_1)$ donc :

$$s(w) = \ln\left(\frac{\mu(\bar{w}_2)}{\mu(w_2)}\right).$$

La valeur du score du palindrome w ne sera nulle que si $\mu(\bar{w}_2) = \mu(w_2)$. Ce sera notamment le cas lorsque la distribution de X et celle du brin complémentaire seront les mêmes. Remarquons que dans ce cas, les méthodes de détection proposées dans cette thèse seront totalement inutiles. Nous rediscuterons ce point dans le chapitre 6 consacré aux applications. De manière plus générale, ce sera le cas lorsque la distribution stationnaire de X est la même que la distribution stationnaire du brin complémentaire X^- (au passage au complémentaire près). Dans les applications aux séquence d'ADN, ce point aura son importance, car il est lié à la deuxième règle de parité (dite de Chargaff (Chargaff and Shapiro, 1968; Bell and Forsdyke, 1999b,a) observée sur ces séquences. Cette règle établit que, sur un même brin d'ADN, le nombre d'adenine(a) tend à être le même que le nombre de thymine (t), et que le nombre de cytosine (c) est proche du nombre de guanine (g). Dans ce cas, la fonction de score ne sera pas bijective, et la séquence des scores qui en découle ne sera pas Markovienne.

5.3 Méthode fondée sur l'utilisation de martingale

Cette approche est fondée sur les travaux de Prum et al. (1995) concernant la distribution d'un mot donné dans une chaîne de markov.

On définit la filtration $\mathcal{F} = (\mathcal{F}_k)$ où $\mathcal{F}_k$ est la σ-algèbre engendrée par $X_1, ..., X_k$.
On définit également la fonction suivante* :

$$\begin{array}{rccc} \mathbf{G}: & \mathcal{A}^{\mathbf{m+1}} & \to \mathbf{R} \\ & (X_{k-m}, ..., X_{k-1}, X_k) & \to \log \frac{\mathrm{Q}^-(\mathbf{X_{k-m}...X_{k-1}, X_k})}{\mathrm{Q}^+(\mathbf{X_{k-m}...X_{k-1}, X_k})}. \end{array}$$

Comme auparavant, pour un $l > 0$ quelconque :

$$\sum_{k=i}^{i+l} G(X_k, ..., X_{k+m-1}, X_{k+m}) = \log \frac{\bar{\mathbb{P}}(X_{i+m}, ..., X_{i+l+m} | X_i, ..., X_{i+m-1})}{\mathbb{P}(X_{i+m}, ..., X_{i+l+m} | X_i, ..., X_{i+m-1})}.$$

Proposition 5.7 *On considère une chaîne de Markov stationnaire et ergodique. En utilisant les définitions et notations ci-dessus, et en définissant :*

$$M_n = \frac{1}{n-m} \sum_{(u_1,...,u_m,v) \in A^{m+1}} (N_{1n}(u_1...u_m, v) - N_{1n}(u_1...u_m)\, Q^+(u_1...u_m, v)) \log \frac{Q^-(u_1...u_m, v)}{Q^+(u_1...u_m, v)}$$

*$G(X_{k-m}, ..., X_{k-1}, X_k) = Y_k$ en utilisant les notations précédentes

avec $m \leqslant n$, M_n converge en loi vers une distribution gaussienne centrée, de variance σ_G^2, quand $(j-i)$ tend vers l'infini, avec

$$\begin{aligned}\sigma_G^2 &= \sum_{(u_1...u_m,v)\in A^{m+1}} \mu(u_1,...,u_m)\, Q^+(u_1...u_m,v) \log \frac{Q^-(u_1...u_m,v)}{Q^+(u_1...u_m,v)} \\ &\left(\log \frac{Q^-(u_1...u_m,v)}{Q^+(u_1...u_m,v)} - \sum_{w\in A} Q^+(u_1...u_m,w) \log \frac{Q^-(u_1...u_m,w)}{Q^+(u_1...u_m,w)} \right).\end{aligned}$$

Preuve.

Pour alléger les écritures, on se place dans le cas où $m = 1$. La généralisation est évidente.

La statistique M_n se réécrit de la façon suivante :

$$M_n = (n-1)^{-\frac{1}{2}} \sum_{k=2}^{n} G(X_{k-1}, X_k) - E[G(X_{k-1}, X_k)|\mathcal{F}_{k-1}]$$

En effet,

$$\sum_{k=2}^{n} G(X_{k-1}, X_k) = \sum_{(u,v)\in A^2} N_{1n}(u,v) \log \frac{\mathrm{Q}^-(u,v)}{\mathrm{Q}^+(u,v)}$$

et

$$\sum_{k=2}^{n} E[G(X_{k-1}, X_k)|\mathcal{F}_{k-1}] = \sum_{(u,v)\in A^2} N_{1n}(u)\mathrm{Q}^+(u,v) \log \frac{\mathrm{Q}^-(u,v)}{\mathrm{Q}^+(u,v)}$$

On définit le tableau triangulaire suivant :

$$\xi_{n,k} = \frac{1}{\sqrt{n-1}} Z_k$$

avec

$$\begin{aligned} Z_k &= G(X_{k-1}, X_k) - E[G(X_{k-1}, X_k)|\mathcal{F}_{k-1}] \\ &= \log \frac{\mathrm{Q}^-(X_{k-1}, X_k)}{\mathrm{Q}^+(X_{k-1}, X_k)} - \sum_{v\in\mathcal{A}} \mathrm{Q}^+(X_{k-1}, v) \log \frac{\mathrm{Q}^-(X_{k-1}, v)}{\mathrm{Q}^+(X_{k-1}, v)} \end{aligned}$$

On alors

$$\begin{aligned} M_n &= \sum_{k=1}^{n} \xi_{n,k} \\ &= \frac{1}{\sqrt{n-1}} \sum_{k=1}^{n} Z_k \end{aligned}$$

Dans un premier temps, on montre la convergence vers la distribution gaussienne à l'aide du théorème de la limite centrale sur les martingales. Pour appliquer ce théorème, il suffit que $\{Z_k, k \geqslant 1\}$ forme une séquence ergodique stationnaire telle que $\mathbb{E}[Z_1^2] = \sigma^2 < \infty$, et $\mathbb{E}[Z_{n+1}|\mathcal{F}_n] = 0$.

La stationnarité et l'ergodicité de la chaîne des Z_k est assurée par celle de la chaîne de Markov sous-jacente.

Pour l'espérance conditionnelle, on a :

$$\begin{aligned}\mathbb{E}[Z_{n+1}|\mathcal{F}_n] &= \mathbb{E}[G(X_n, X_{n+1})|\mathcal{F}_n] - E[G(X_n, X_{n+1})|\mathcal{F}_n] \\ &= 0\end{aligned}$$

De plus, Z_1^2 possède une distribution discrète à support fini. Son espérance est donc fini.

Calcul de σ_G^2

$\sigma_G^2 = E[G^2] - E[E[G|F_{k-1}]^2]$

$E[G^2] = \sum_{(u,v)\in A^2} \left(\log \frac{\mathrm{Q}^-(u,v)}{\mathrm{Q}^+(u,v)}\right)^2 \mu(u)\mathrm{Q}^+(u,v)$

et $E[E[G(X_k, X_{k-1})|F_{k-1}]^2] = \sum_{u\in A} \mu(u) \left(\sum_{v\in A} \mathrm{Q}^+(u,v) \log \frac{\mathrm{Q}^-(u,v)}{\mathrm{Q}^+(u,v)}\right)^2$

d'où le résultat trouvé. ■

5.3.1 Equivalence avec l'approche par score local

Cette approche par martingale correspond à une fonction de score particulière. En effet, on peut formaliser le problème de la même façon que pour le score local en utilisant la fonction de score suivante :

$$\begin{array}{rccl} \mathbf{s_2}: & \mathcal{A}^{\mathbf{m+1}} & \to \mathbb{R} & \\ & (X_{k-m}, ..., X_{k-1}, X_k) & \to G(X_{k-m}...X_{k-1}, X_k) - E[G(X_{k-m}...X_{k-1}, X_k)|X_{k-m}...X_{k-1}]. \end{array}$$

On peut expliciter cette fonction :

$$s_2(u_1, ..., u_{m+1}) = \log\left(\frac{\mathrm{Q}^-(u_1, ..., u_{m+1})}{\mathrm{Q}^+(u_1, ..., u_{m+1})}\right) - \sum_{w\in\mathcal{A}} \mathrm{Q}^+(u_1, ...u_m, w) \log\left(\frac{\mathrm{Q}^-(u_1, ..., u_{m+1})}{\mathrm{Q}^+(u_1, ..., u_{m+1})}\right).$$

Dans la problèmatique du score local, on cherche le segment qui maximise la somme de ses scores élémentaires, parmi tous les segments possibles. Ici, la statistique M_{ij} est égale au score du segment $X_i, X_{i+1}, ..., X_j$ renormalisé par la racine carré de sa longueur (à l'ordre de la chaîne de Markov près).

Ici, l'espérance de la fonction de score s_2 est nulle. Les résultats formulés par Daudin et al. (2003) et Etienne (2002), et rappelés page 48, donnent la loi du score local dans ce

cas, et montrent eux aussi que la normalisation dans ce cas doit être effectuée en fonction de la racine de la longueur de la séquence, et non en fonction du logarithme de cette longueur comme lorsque l'espérance du score est négative.

5.3.2 Quelques simulations

Le but des simulations est :

1. Etude de **spécificité** : Eprouver la qualité de l'approximation normale de M_{1n} et du paramètre de variance σ_G^2 sous H_0 en fonction de
   (a) la longueur de la chaine n
   (b) la matrice de transition P^+ (et la distance en variation totale d entre P^+ et P^-)
2. Etude de **sensibilité** : Comportement de M_{ij} sous H_1 en fonction de
   (a) la longueur de la chaine n
   (b) la matrice de transition P^+ (et la distance en variation totale d^* entre P^+ et P^-)
   (c) la longueur (absolue l et relative $r = l/n$ à la longueur totale) du segment retourné

Les simulations peuvent également permettre de détecter une erreur de programmation des algorithmes.

Paramètres Longueur de la séquence n : $100, 1000$
Longueur du segment retourné l : $20, 200$ (positions respectives 10-30 et 100-300)

---

*$d(P, P') = \frac{1}{2}\sum_{(u,v)\in A^2} |\mu(u)P(u,v) - \mu'(u)P'(u,v)|$,
d'où $d(P^+, P^-) = \frac{1}{2}\sum_{(u,v)\in A^2} |\mu(u)P^+(u,v) - \mu(v)P^-(v,u)|$

Les différentes matrices de transition sont classées en ordre croissant de distance $d(P^+, P^-)$:

$$P_1^+ = \begin{pmatrix} .3 & .2 & .2 & .3 \\ .2 & .3 & .3 & .2 \\ .2 & .3 & .3 & .2 \\ .2 & .2 & .3 & .3 \end{pmatrix} \qquad P_2^+ = \begin{pmatrix} .3 & .2 & .2 & .3 \\ .2 & .3 & .3 & .2 \\ .1 & .4 & .4 & .1 \\ .2 & .2 & .3 & .3 \end{pmatrix}$$

$$d(P_1^+, P_1^-) = .123 \qquad d(P_2^+, P_2^-) = .254$$

$$P_3^+ = \begin{pmatrix} .5 & .1 & .1 & .3 \\ .3 & .2 & .1 & .4 \\ .1 & .2 & .05 & .65 \\ .7 & .1 & .1 & .1 \end{pmatrix} \quad P_4^+ = \begin{pmatrix} .6 & .1 & .1 & .2 \\ .2 & .5 & .2 & .1 \\ .05 & .05 & .4 & .5 \\ .1 & .4 & .2 & .3 \end{pmatrix}$$

$$d(P_3^+, P_3^-) = .332 \qquad d(P_4^+, P_4^-) = .491$$

$$P_5^+ = \begin{pmatrix} .3 & .2 & .2 & .3 \\ .3 & .1 & .3 & .3 \\ .1 & .2 & .05 & .65 \\ .7 & .1 & .1 & .1 \end{pmatrix} \quad P_6^+ = \begin{pmatrix} .1 & .2 & .2 & .4 \\ .5 & .2 & .2 & .4 \\ .8 & .1 & .05 & .05 \\ .2 & .4 & .3 & .1 \end{pmatrix}$$

$$d(P_5^+, P_5^-) = .580 \qquad d(P_6^+, P_6^-) = .682$$

Résultats des simulations

Les traces $M_{1j}, j = 1, n$. Les graphiques des pages 125 et 126 montrent 20 trajectoires de la martingale pour des chaines simulées selon P_1^+, P_3^+ et P_6^+, et les différents paramètres choisis. La statistique M_{1n} que nous considérons est l'ordonnée du point final de cette trajectoire.

Le terme "position x-y" dans les titres des graphiques indique le segment retourné. Les lignes pointillées sont définies par $\pm 2\sqrt{\sigma_G^2 j/n}$ pour j allant de 1 à n. Les trois premiers graphiques sont donc sous H_0, alors que les trois graphiques de la page suivante sont sous H_1 avec un retournement de taille 200.

Sous H_0, les trajectoires restent à l'intérieur des bornes de confiance. C'est un élément de validation des programmes utilisés d'une part, et de l'approximation de la variance d'autre part.

Losrqu'on retourne un segment, on attend que la trajectoire soit déviée positivement aux alentours du retournement. Cette déviation est visible pour des segments retournés de longueur 200 (sur 1000) pour toutes les matrices et est marquée à partir de la matrice P_2 ($d(P_2^+, P_2^-) = .254$). Les segments de longueur 20 sont plus difficilement mis en évidence.

Pour les chaînes de longueur 100 ou 1000, les "déviations" sont visibles à partir de P_4 ($d(P_4^+, P_4^-) = .491$) Dans tous les cas, la fin de la trajectoire (qui représente la statistique que nous avons considérée) se rapproche ou se situe dans la zone de confiance.

La statistique M_{1n}. Le tableau suivant présente les caratéristiques de M_{1n} selon les différents paramètres. 500 séquences ont été simulées dans chaque cas.
m est la moyenne empirique de M_{1n}.
v est la variance empirique de M_{1n}. Le nombre entre parenthèses est la variance asymptotique σ_G^2.
p est la proportion de simulations où H_0 a été rejetée au risque 2.5% unilatéral. p est une estimation de la puissance.

TAB. 5.1 – Estimation sur 500 simulations de la moyenne m, de la variance v de M_{1n} et de la puissance p du test fondé sur M_{1n}

		P_1 ($d = .123$)			P_2 ($d = .254$)			P_3 ($d = .332$)		
n	l	m	$v(.022)$	p	m	$v(.084)$	p	m	$v(.215)$	p
100	0	.009	.030	.06	.007	.103	.06	.015	.283	.05
	20	.068	.030	.10	.234	.109	.15	.488	.290	.21
1000	0	.011	.030	.06	.010	.113	.05	-.031	.256	.03
	20	.018	.029	.05	.044	.109	.05	.160	.251	.06
	200	.188	.027	.26	.687	.109	.66	1.556	.277	.89
		P_4 ($d = .491$)			P_5 ($d = .580$)			P_6 ($d = .682$)		
n	l	m	$v(.444)$	p	m	$v(.381)$	p	m	$v(.562)$	p
100	0	-.090	.494	.02	-.020	.481	.05	0.011	.596	.04
	20	1.073	.507	.36	1.059	.525	.42	1.565	.724	.53
1000	0	-.005	.502	.04	.003	.519	.05	-0.003	.738	.05
	20	.328	.605	.11	.278	.496	.10	.427	.738	.11
	200	3.400	.489	1	3.470	.494	1	4.876	.796	1

Sous H_0, la plupart des moyennes ne sont pas significativement différentes de 0 comme attendu. Les variances empiriques sont toujours légèrement supérieures aux variances asymptotiques. Généralement, le risque de 1ère espèce empirique est supérieur au risque attendu (2.5%). Lorsqu'un segment de longueur 20 est retourné, la moyenne augmente, mais la puissance reste faible (<50%), y compris pour une séquence de longueur 100. Pour une séquence de longueur 1000, elle ne dépasse pas 15% dans ce cas. Un retournement de longueur 200 est détecté dans plus de 60% des cas dès P_2.

Les graphiques de la page 125 et de la suivante représentent la densité estimée de $M_{1n}/\sqrt{\sigma_G^2}$.

La normalité asympotique Sur quelques simulations, on essaie de voir à partir de quelle longueur n de séquence, la statistique M_{1n} peut être considérée comme gaussienne sous H_0. Les longueurs allant de 2 à 20 ont été considérées. 500 chaînes de chaque longueur ont été simulées sous H_0 pour chaque matrice de transition. Les graphiques de la page 127 représentent les densités de M_{1n} estimées par un estimateur à noyau gaussien en fonction de la matrice de transition utilisée pour une séquence de longueur 20. La courbe pointillée représente la densité de la loi $N(0, \sigma_G^2)$.
L'approximation gaussienne est déjà satisfaisante pour une séquence de longueur 20 (ce qui est faible pour des données génomiques). La matrice de transition semble néanmoins avoir une influence sur la vitesse de convergence pour des longueurs plus faibles (résultat non présenté).

5.3.3 Conclusion

L'approche par martingale est séduisante, car elle s'appuie sur des résultats théoriques solides. Elle présente aussi l'avantage de présenter le problème sous un angle différent. Néanmoins, l'étude de simulation montre qu'une distance en variation totale importante entre les deux chaînes de Markov X^- et X^+ est nécessaire pour obtenir une puissance acceptable.

Un autre point que nous n'avons pas abordé jusqu'ici est la localisation des segments retournés. Il semble assez évident que cette méthode ne permettra pas une localisation précise de la fin du segment retourné (cf les graphiques page 126). Il en sera d'ailleurs toujours ainsi lorsque la fonction de score est d'espérance nulle. En effet, on comprend intuitivement que l'espérance nulle du score ne modifiera pas en moyenne le score local obtenu jusqu'à un certain point de la séquence ; par conséquent, la fin du segment effectivement sous H_1 sera largement dépassée.

Ces deux observations nous conduisent à ne pas utiliser cette approche dans la recherche de segments inversés dans les génomes.

5.4 Comparaison des méthodes de score local et de fenêtre glissante.

On se propose ici de comparer les puissances relatives des deux méthodes retenues pour détecter un retournement dans une chaîne de Markov. Il s'agit de la méthode dite par "fenêtre glissante" présentée dans le chapitre 4 page 94, et de la méthode de score local présentée dans ce chapitre. Pour la méthode utilisant les fenêtres glissantes, nous nous intéressons à la statistique S_l^n et nous utilisons la meilleure approximation de sa distribution : l'approximation "product-type" présentée page 103. Pour la méthode de score local, nous considérons la première fonction de score s, présentée page 111, car nous avons vu que l'approche martingale présentait quelques défauts.

Simulations

Le calcul de la puissance n'est pas trivial à obtenir analytiquement, on choisit donc de l'évaluer à l'aide de simulations. Avant de présenter les paramètres utilisés, remarquons que cette puissance sera nulle si la chaîne de Markov est réversible. Dans ce cas, la distribution de la chaîne dans les deux sens est la même. Par extension, plus la chaîne X est différente de X^-, plus les deux méthodes seront efficaces. Pour quantifier cette

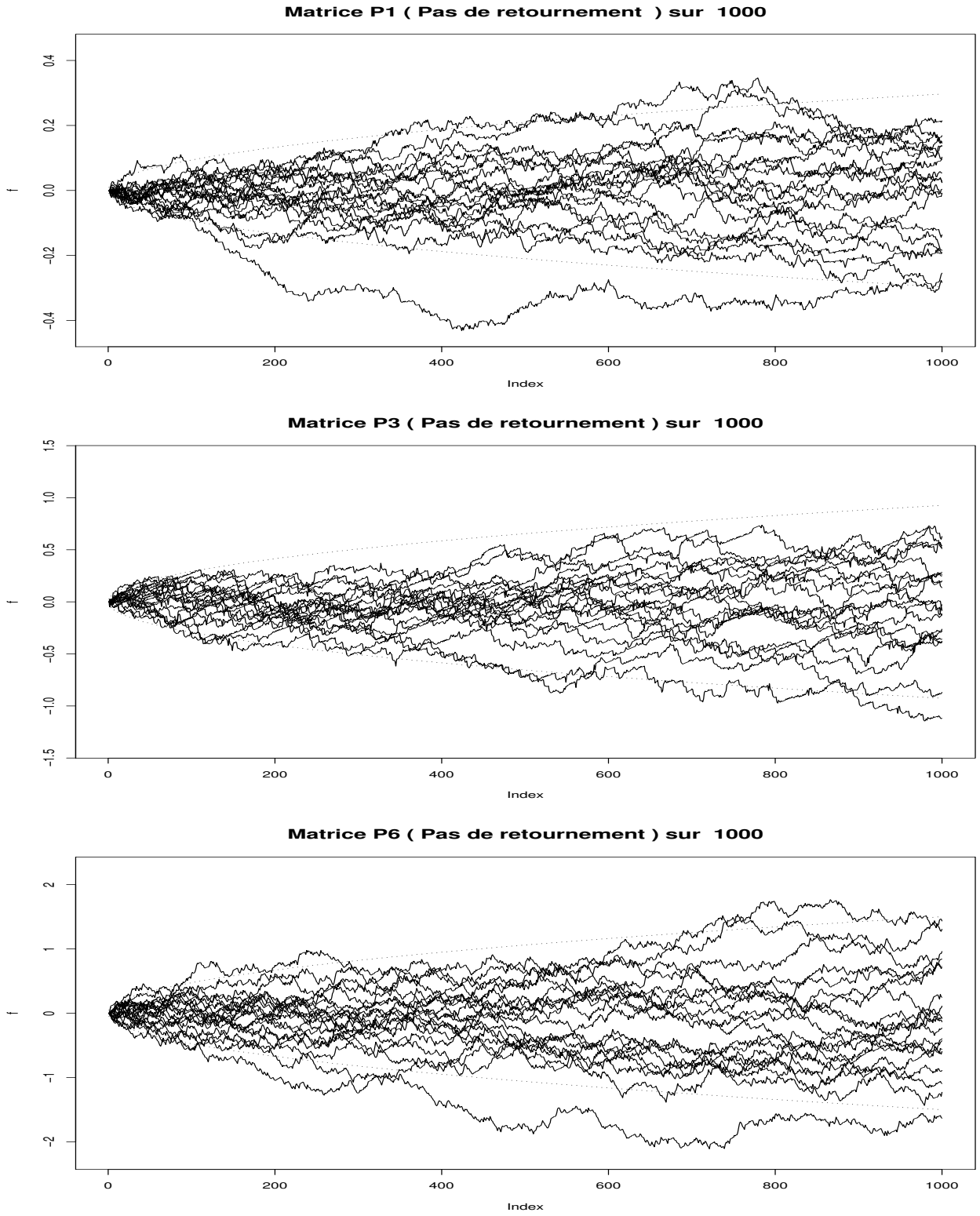


FIG. 5.1 – Vingt trajectoires de la martingales $M_{1j}, j = 1, ..., n$ lorsqu'il n'y a pas de retournement pour des séquences générées selon les matrices de transition P_1^+, P_3^+ et P_6^+.

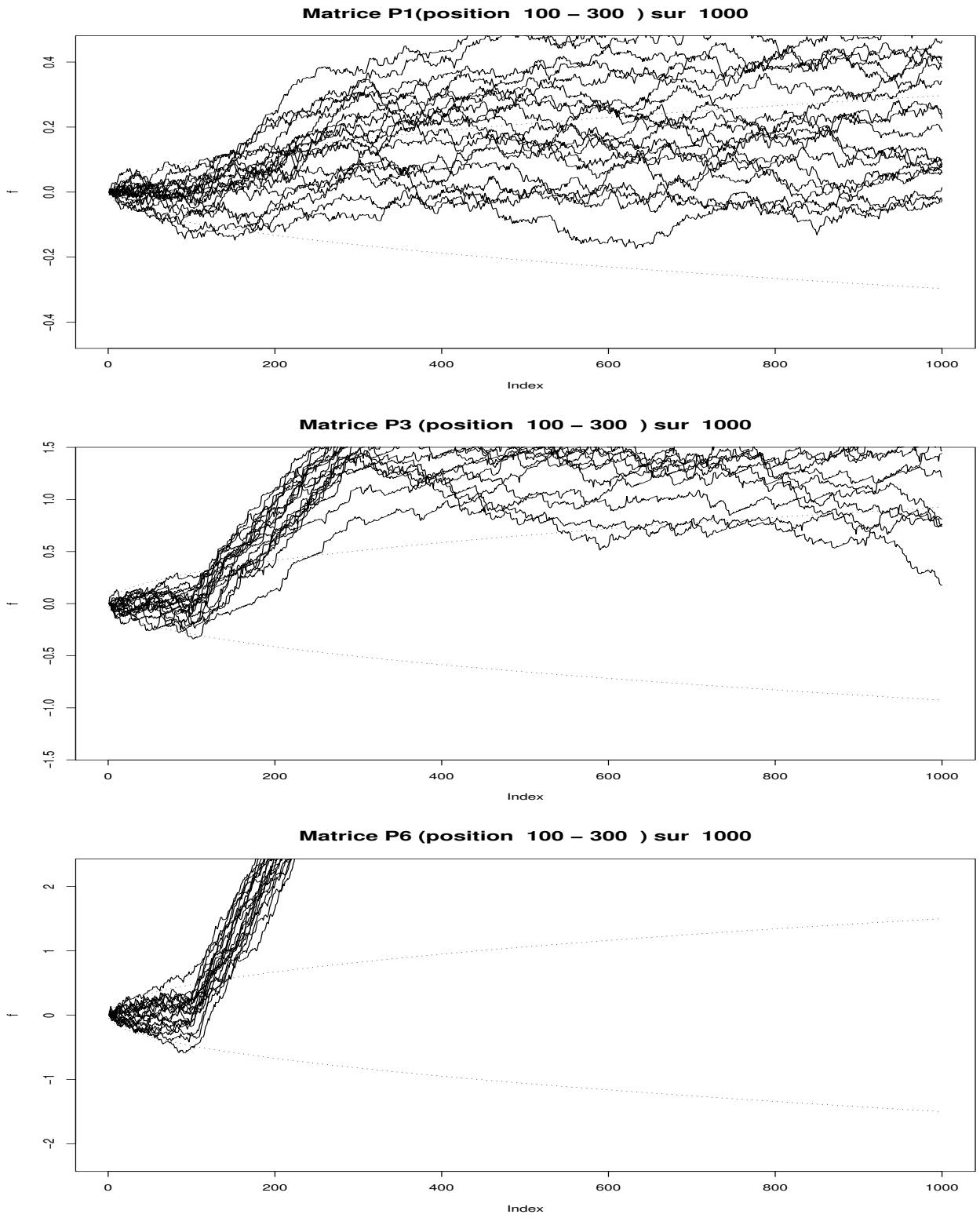


FIG. 5.2 – Vingt trajectoires de la martingales $M_{1j}, j = 1, ..., n$ lorsqu'il y a un retournement de taille 200 pour des séquences générées selon les matrices de transition P_1^+, P_3^+ et P_6^+.

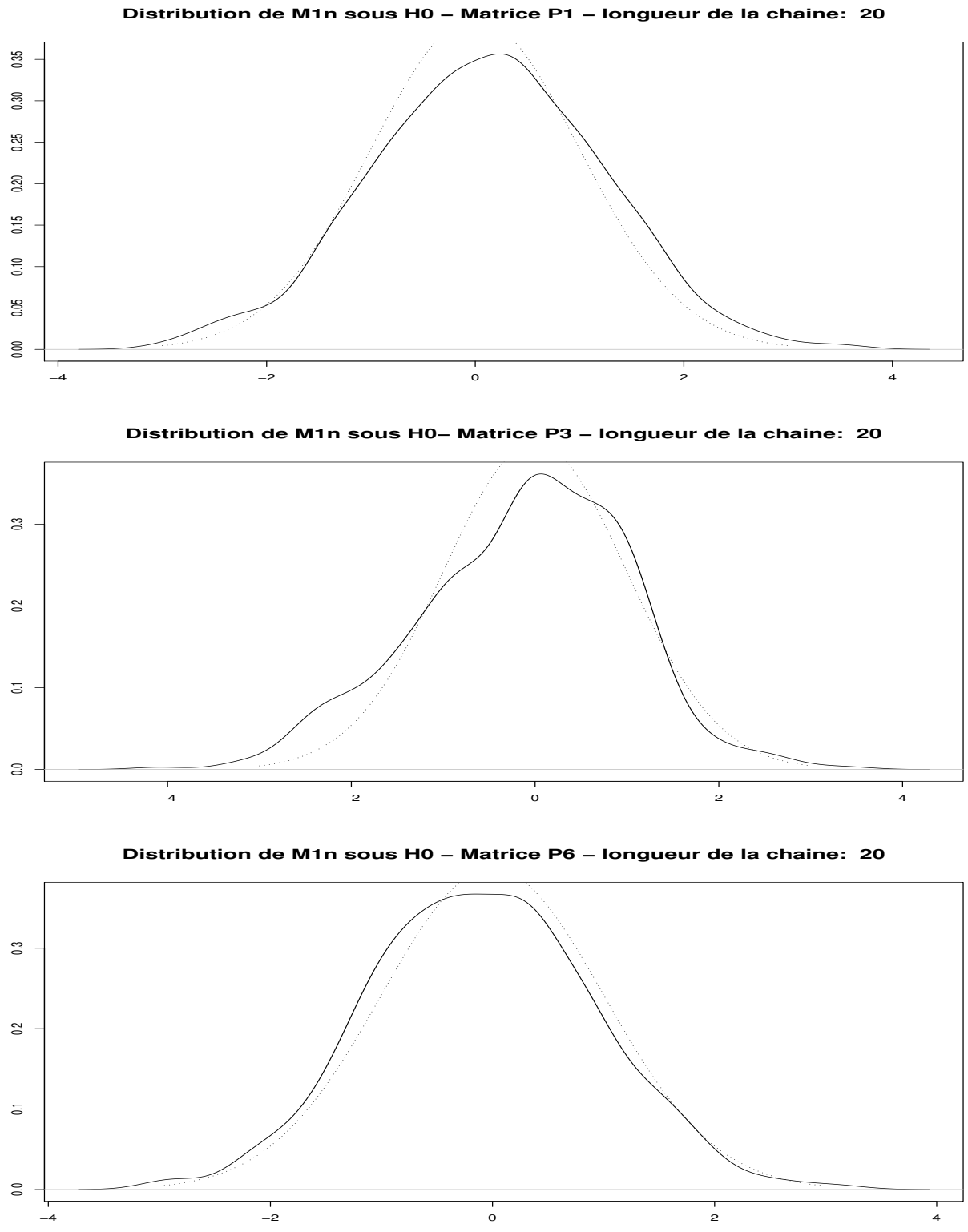


FIG. 5.3 – Estimation des densités de la martingale $M_{1,20}$ lorsqu'il n'y a pas de retournement pour des séquences de taille 20 générées selon les matrices de transition P_1^+, P_3^+ et P_6^+.

différence, on utilise le taux d'entropie Er de la chaîne X^+ relative à la chaîne X^- (Cover and Thomas, 1991) :

$$Er(X^-, X) = \sum_{u,v \in \{a,c,g,t\}} \mathbb{P}^+(u,v) \log(\mathrm{Q}^+(v|u)/\mathrm{Q}^-(v|u))$$

Er est positif et plus cette quantité est élevée, plus la chaîne de Markov est "orientée". Notons qu'il existe également d'autres mesure de distance séparant deux chaînes de Markov. Par exemple, la distance en variation totale $D(X^-, X) = \frac{1}{2} \sum_{u,v \in \{a,c,g,t\}} |\mathbb{P}^+(u,v) - \mathbb{P}^-(u,v)|$, qui varie entre 0 et 1, est une mesure classique. Une autre mesure appropriée dans notre cas serait $M(X^-, X) = \max_{u,v \in \{a,c,g,t\}} \log(\mathrm{Q}^+(v|u)/\mathrm{Q}^-(v|u))$. Ces mesures peuvent diverger lorsque l'on compare le degré d'orientation de deux chaînes de Markov différentes dans certains cas particuliers (voir Li (1991) pour exemple). Ceci explique pourquoi une méthode appliquée à une chaîne de Markov particulière peut être plus puissante que la même méthode appliquée à une autre chaîne de Markov, même si cette dernière semble plus "orientée". Cependant les différentes mesures coïncident ou sont légèrement différentes sur la modélisation de séquences biologiques. Notons que Er est également l'opposé de l'espérance de la fonction de score Y_i lorsqu'il n'y a pas de segment retourné. Nous conserverons cet indicateur comme un résumé du degré d'orientation.

Choix des paramètres Pour observer comment les méthodes se comportent, on simule 10000 chaînes de Markov, dans lesquelles on introduit éventuellement un retournement. On choisit durant la simulation de contrôler trois éléments pertinents :

1. la longueur de la séquence markovienne
2. la taille du segment retourné
3. les paramètres de la chaîne de Markov

Le premier élément, la longueur de la séquence markovienne simulée, n'est pas censé refléter la taille réelle des séquences biologique, mais celle d'une séquence dans laquelle un hypothétique retournement pourrait apparaître. Nous considérons des séquences markoviennes de trois longueurs $n = 1000$, 5000 ou 10000.

Le second élément est la taille du retournement introduit dans la séquence. Puisqu'il existe peu de travaux sur le sujet, les tailles suivantes : $l = 5$, 10, 15, 20, 25, 30, 40, 50, 75, 100 et 150 ont été choisies en accord avec l'intuition de biologiste de Daniel Goldstein.

Le troisième élément concerne le choix des paramètres de la chaîne de Markov simulée et son degré relatif d'orientation. On choisit 6 matrices différentes de transition. La première est la matrice calculée sur la séquence HIV-1. Il se trouve que, comparé aux autres organismes, la séquence d'ADN de HIV-1 est plutôt orientée. Ensuite, pour étudier l'influence du degré d'orientation, on calcule les matrices suivantes

$$\mathrm{Q}_p = p * \mathrm{Q} + (1-p) * \mathrm{Q}^-$$

où $p > 0$, Q est la matrice de la séquence HIV-1 et Q^- est la matrice de la chaîne complémentaire inversée. La valeur $p = 1$ correspond donc à la chaîne de Markov initiale, et la valeur $p = 0.5$ correspond à la chaîne de Markov presque réversible. Plus p est proche de 0.5, moins la chaîne est orientée. Pour p, nous choisissons les valeurs suivantes : $p =$ 0.5, 0.6, 0.7, 0.8, 0.9, 1.0, 1.1, 1.2, 1.3, 1.4 et 1.5. On vérifie que Q_p demeure une matrice de Markov pour ces valeurs de p (i.e. les éléments de Q_p sont positifs et inférieur à 1). Le tableau 5.2 présente différentes mesures d'orientation pour chaque valeur de p.

p	0.5	0.6	0.7	0.8	0.9	**1.0**
Taux d'entropie relatif	7.10^{-4}	4.10^{-3}	0.01	0.03	0.05	**0.09**
Distance en variation totale	0.02	0.06	0.12	0.17	0.23	**0.28**
Maximum de Y_i	0.12	0.13	0.31	0.52	0.74	**0.97**
p	1.1	1.2	1.3	1.4	1.5	
Taux d'entropie relatif	0.12	0.17	0.23	0.30	0.38	
Distance en variation totale	0.34	0.39	0.44	0.49	0.54	
Maximum de Y_i	1.23	1.52	1.87	2.32	3.02	

TAB. 5.2 – Différentes mesures d'orientation pour des chaînes de Markov de matrice de transition $Q_p = p * Q + (1 - p) * Q^-$ où Q est la matrice estimée sur la séquence HIV-1. La valeur $p = 1$ correspond à la séquence HIV-1 (colonne en gras).

L'approche par "fenêtre glissante" nécessite une spécification a priori sur la taille de la fenêtre. On choisit les tailles suivantes : $w = 10$, 15, 20, 25, 30, 40, 50, 75, 100 et 150. La taille 5 n'est pas prise en compte car dans ce cas, l'erreur de type I ne pourrait pas être controlée avec assez de précision, comme illustré sur la figure 5.4 page 131. Dans chaque cas, 10000 simulations ont été effectuées.

On choisit de considérer une erreur de type I fixé a priori, notée α. Ensuite, un seuil correspondant est calculé pour chaque méthode. A partir de celui-ci, trois indicateurs sont construits pour résumer les simulations : la **puissance**, la **sensitivité** et la **spécificité**. La puissance est la proportion empirique de séquences simulées sous H_1 pour lesquelles la valeur maximale de la statistique sur la séquence totale est plus grande que le niveau. Cet indicateur ne prend pas en compte d'information sur la position du retournement mais indique qu'un retournement est présent quelque part dans la séquence. On définit la sensitivité comme le quotient moyen du nombre de bases correctement détectées dans un retournement sur la longueur du retournement. Enfin, la spécificité est définie ici comme le quotient moyen du nombre de bases correctement détectées dans un retournement sur la longueur des retournements détectés. Les deux derniers indicateurs reflètent des propriétés positionnelles des deux méthodes. Notons que la spécificité est calculée ici conditionnellement au fait qu'un retournement est détecté ; ce n'est pas le cas pour la sensitivité. Ainsi, la sensibilité n'est pas nécessairement plus grande que un

moins la spécificité comme c'est généralement le cas dans la courbe caractéristique classique d'opération du récepteur (Receiver Operating Characteristic curve) par exemple. Contrairement à la méthode du score local, celle de la fenêtre glissante peut prédire des retournements qui se chevauchent. Dans ce cas, le résultat n'est pas immédiatement interprétable et une observation plus fine des résultats est nécessaire lorsqu'on analyse une séquence réelle. Dans le cas présent de simulations répétées par ordinateur, on fixe la règle automatique suivante : si deux retournements détectés ou plus se chevauchent, on choisit comme retournement détecté final l'union de tous ces segments.

Résultats

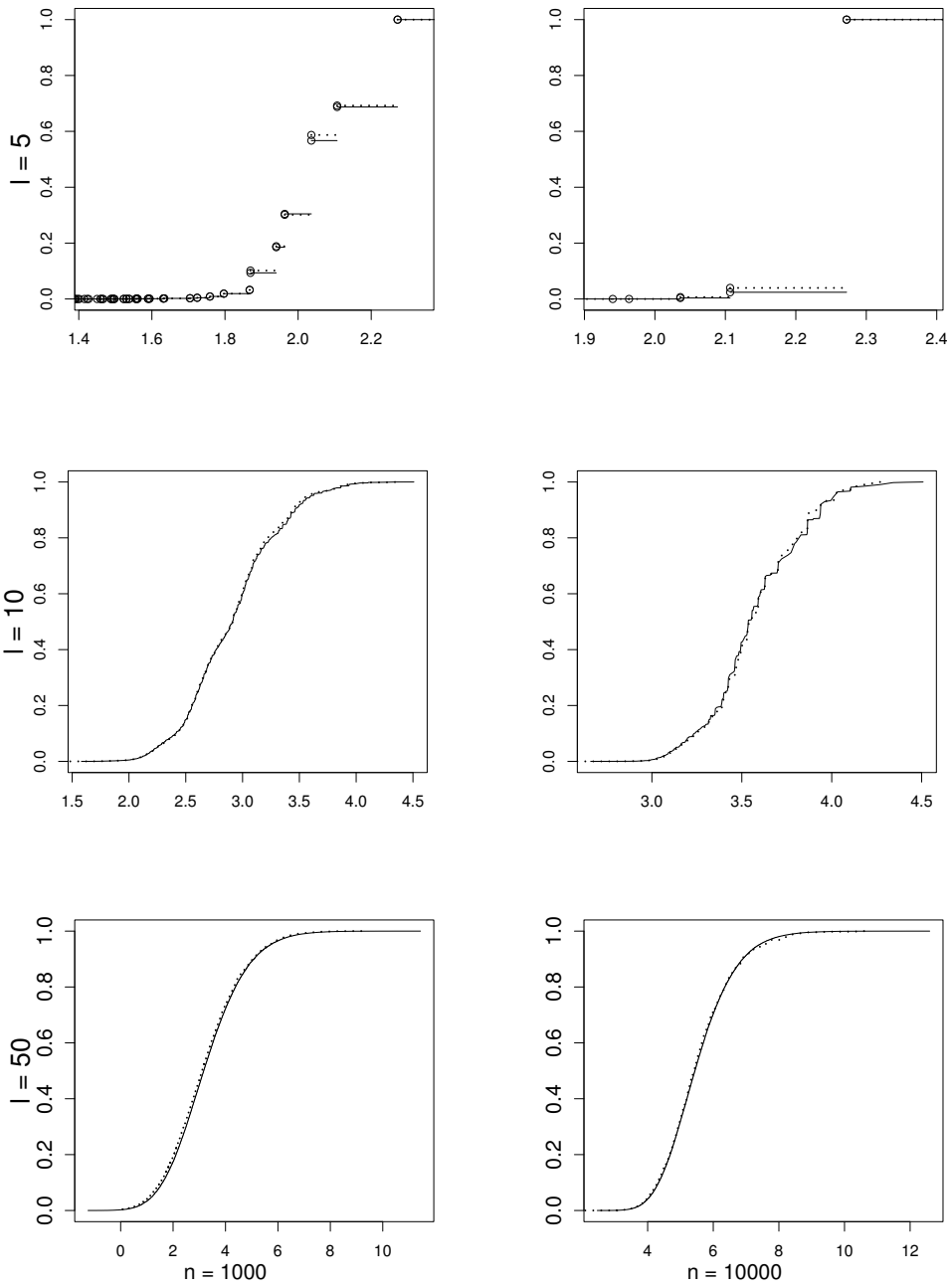
Le graphique 5.5 page 132 montre la puissance de l'entropie de la chaîne positive relativement à la chaîne négative (c'est également l'opposé de l'espérance de Y_i). Cette relation est tracée graphiquement pour différents retournements de longueur l. Pour la méthode par fenêtre glissante, une fenêtre de même taille que le retournement est choisie. Notons que c'est le cas le plus efficace pour cette méthode. Les résultats présentés ici sont obtenus quand un retournement apparaît dans une séquence de 1000 bases.

On remarque qu'un retournement de taille 10 sera mal détecté, même dans une séquence qui présente une orientation relativement bonne. Néanmoins, la puissance est meilleure pour un retournement de taille supérieure à 20 et satisfaisante pour un retournement de taille au moins 30. Comme prévu, la méthode de fenêtre glissante est plus puissante que celle du score, au moins lorsque la taille de la fenêtre est correctement spécifiée. Notons tout de même que cette différence est assez faible.

On considère maintenant la chaîne de Markov estimée sur la **séquence HIV-1**. On s'intéresse à la puissance, à la spécificité et à la sensibilité des deux méthodes pour détecter un retournement dans la séquence HIV-1. Le graphique 5.6 page 133 donne la puissance calculée des deux méthodes en fonction de la longueur du retournement. La taille de la fenêtre glissante a été choisie pour correspondre à la taille du retournement.

Pour les trois longueurs de séquence, on obtient une puissance d'environ 80% pour un retournement de 100 bases. Notons qu'on observe seulement de légères variations dans la puissance lorsqu'un retournement est présent dans une séquence de 10000 bases par rapport à une séquence de 1000 bases. Le graphique 5.7 montre la relation entre spécificité et sensibilité pour chaque méthode en fonction de la taille du retournement.

Pour obtenir ces graphiques, on a fixé 17 valeurs différentes pour l'erreur de type I, variant de $\alpha = 1\%$ à 80%. La spécificité et la sensibilité ont été évaluées pour chaque valeur de α. Ici encore, ce sont des séquences de longueur 1000 qui ont été simulées. Pour des tailles faibles de retournement ($\leqslant 30$), la spécificité de la méthode du score dépend faiblement de la sensibilité. Ce phénomène n'est pas observé pour la méthode de la fenêtre glissante.



$n = 1000$ $n = 10000$

FIG. 5.4 – Qualité de l'approximation "product-type" (voir page 103) pour une longueur de séquence $n = 1000$ et 10000, et une taille de fenêtre $l = 5, 10$ et 50. — : distribution cumulée de S_l^n obtenue par Monte Carlo ; ... : distribution cumulée de S_l^n obtenue par l'approximation "product-type"

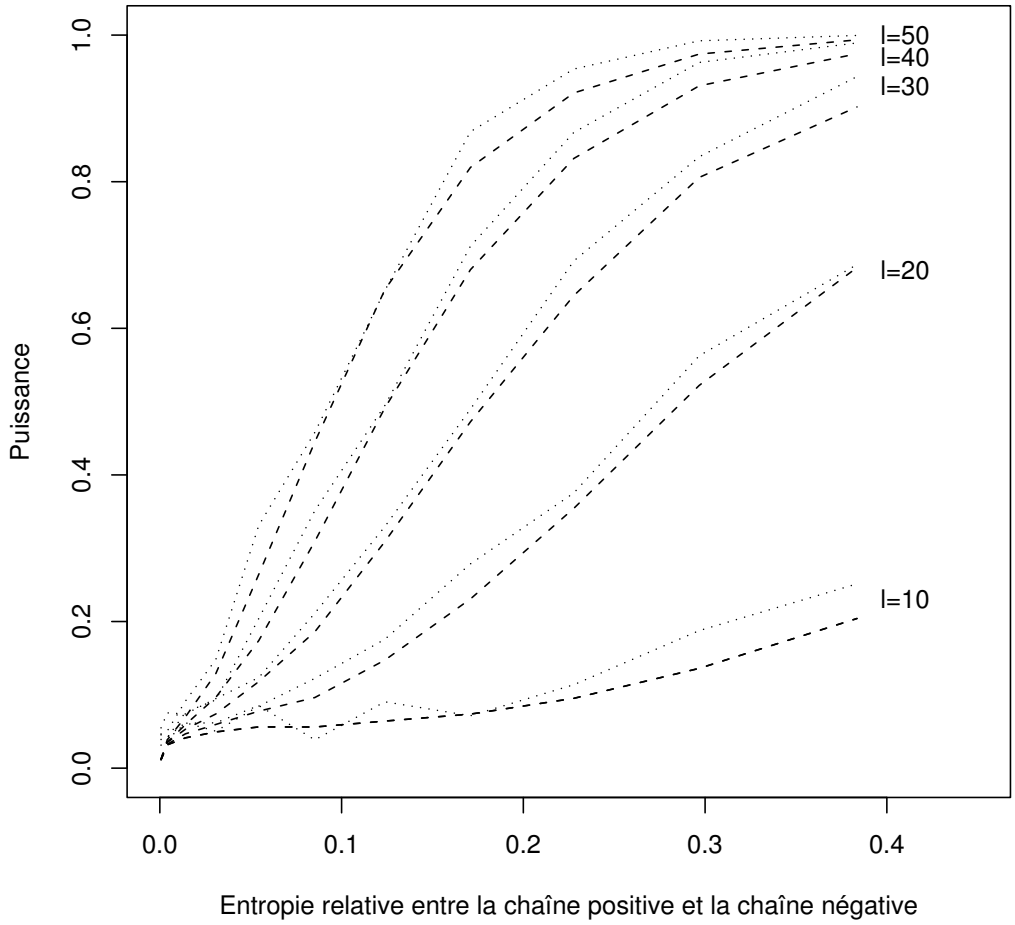


FIG. 5.5 – Puissance de détection d'un retournement par la méthode du score local et la méthode par fenêtre glissante en fonction de l'entropie relative entre la chaîne positive et la chaîne négative, et la longueur l du retournement. . . . . : Méthode par fenêtre glissante ; – – – : Méthode du score local.

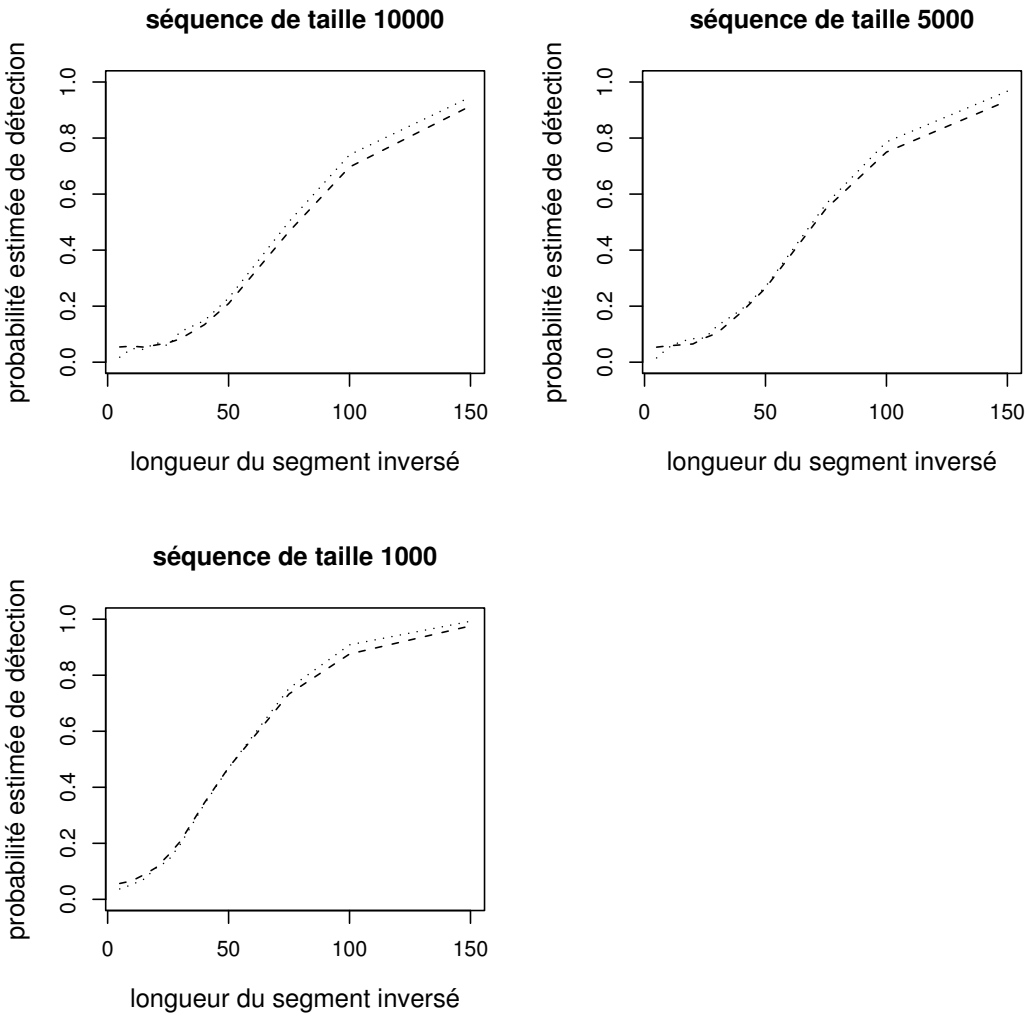


FIG. 5.6 – Puissance de détection d'un retournement par la méthode du score et de la fenêtre glissante en fonction de la longueur du retournement. ..... : Méthode par fenêtre glissante ; – – – : Méthode du score local. Dans le premier graphique, un retournement est présent dans une séquence de longueur 1000 ; dans le second, il est présent dans une séquence de longueur 5000 et dans le dernier, il est présent dans une séquence de longueur 10000.

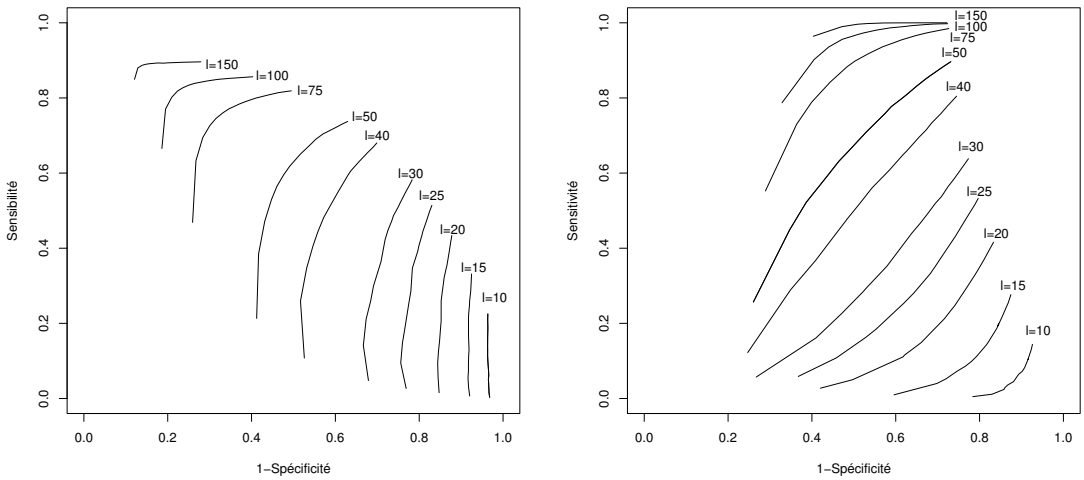


FIG. 5.7 – Sensibilité et spécificité de la méthode par fenêtre glissante (gauche) et de la méthode du score (droite) en fonction de la longueur l du retournement. Pour la méthode par fenêtre glissante, une fenêtre de même taille que le retournement est choisie. Ici un retournement est présent dans une séquence de 1000 bases et la séquence est simulée selon la chaîne de Markov estimée sur la séquence HIV-1.

Evaluation sur une séquence biologique

Pour prendre en compte le fait qu'une séquence réelle n'est pas une chaîne de Markov, on complète l'évaluation précédente par une étude des génomes de *Escherichia Coli K12*. Tout d'abord, on recense toutes les parties du génome qui contiennent cinq gènes tels qu'il existe un gène parmi les cinqs, qui ne soit pas sur le même brin que les autres. On peut considérer que ce gène est "inversé" par rapport aux autres, c'est à dire que le gène n'est pas sur le brin auquel on pourrait s'attendre au vue de son contexte (les deux brins n'ayant pas un rôle symétrique dans la plupart des génomes). On totalise 1443 segments avec de telles caractéristiques, de longueur moyenne 5342 (écart type 1658). La longueur moyenne du gène inversé est d'environ 960 bases, avec un écart type de 600. La méthode du score local est mise en oeuvre sur chaque segment dans le but de détecter ab initio le gène inversé avec une p-value limite de 20%. Résultat : la puissance globale (définie précédemment) est de 68%. Si l'on s'attache plus précisément à la position des segments détectés, la sensibilité est 57% et la spécificité est 60% de bases correctement détectées. Ces chiffres ne sont pas plus élevés en raison de la présence de quelques segments atypiques. Lorsque ces segments sont enlevés, la sensibilité monte à 76% et la spécificité à 79%. Malgré le fait qu'une séquence réelle ne soit pas une chaîne de Markov, ces résultats sont satisfaisants. Notons que l'entropie relative entre la chaîne positive et la chaîne

négative est faible pour quelques segments, ce qui donne au final une faible puissance. De plus, la spécificité donnée est une minoration car rien ne garantit qu'il n'y a pas de retournement hors du gène inversé. Les résultats obtenus sont plutôt bons comparés à ce à quoi on pourrait s'attendre au vu de notre approche très automatisée.

Bibliographie

Bell, S, J. and Forsdyke, D, R. (1999a). Accounting units in DNA. *J Theor Biol J*, 197 :51–61.

Bell, S, J. and Forsdyke, D, R. (1999b). Deviations from Chargaff's second parity rule correlate with direction of transcription. *J Theor Biol J*, 197 :63–76.

Chargaff, E. and Shapiro, H, S. (1968). Remarks on sequence characteristics of the DNA and transfer RNA of yeast. *Proc Natl Acad Sci U S A J*, 59 :161–163.

Cover, Thomas, M. and Thomas, Joy, A. (1991). *Elements of Information Theory*. Wiley & Sons.

Daudin, J., Etienne, M., and Vallois, P. (2003). Asymptotic behaviour of the local score of independant and identically distributed random sequence. *Stochastic Processes and their Applications*, 107 :1–28.

Etienne, Marie, P. (2002). *Le score local : un outil pour l'analyse de séquences biologiques*. PhD thesis, Université de Nancy I.

Li, W. (1991). On the relationship between complexity and entropy for markov chains and regular languages. *Complex Systems*, 5 :381–399.

Prum, B., Rodolphe, F., and de Turckheim, E. (1995). Finding words with unexpected frequencies in deoxyribonucleic acid sequences. *J. R. Statist. Soc. B*, 57(1) :205–220.

Chapitre 6

Recherche de segments inversés dans quelques génomes viraux.

On propose dans ce chapitre d'appliquer les méthodes de détection présentées dans les chapitres précédents à des séquences "réelles" d'ADN de virus. Ces exemples sont principalement présentés dans un but illustratif, une expertise biologique, encore manquante à ce stade de développement, est nécessaire afin de permettre des conclusions plus subtiles. Il semble par exemple utile de ne pas considérer des séquences entières ou même des portions prises au hasard, mais plutôt, de choisir avec un a priori biologique des génomes particuliers, et/ou des parties de génomes pouvant être sensible à l'événement inversion de génome. Une première idée dans ce sens consiste à utiliser le fait que ces zones inversées sont probablement un facteur d'instabilité de la séquence. En effet, une zone dont la composition est proche de la composition du brin inverse complémentaire aura tendance à être "attirée" par une autre partie du même brin ; ces zones peuvent générer de l'instabilité, notamment au moment de la réplication par exemple. On peut notamment utiliser les conclusions de l'étude de Lefebvre et al. (2003); Hannenhalli et al. (1995) qui ont trouvé un excès significatif d'inversion d'un gène unique, par rapport à l'inversion de plusieurs gènes, pour regarder si ces gènes ne seraient pas flanqués de segments inversés. De la même façon, on peut penser à étudier les zones flanquant les transposons*. En effet, ces zones pourraient ne pas être restreintes à être exactement inverses complémentaires l'une de l'autre, mais plutôt être de "styles" inverses complémentaires (en terme de composition nucléotidique).

*gènes dits "sauteurs" qui se déplacent le long du génome et peuvent être responsables de changement phénotypique, comme la couleur du grain dans un épi de maïs par exemple.

6.1 SIC (Scan Inverse Complementary) : logiciel de recherche d'inversions dans une séquence

L'essentiel du travail fourni dans cette thèse est d'ordre méthodologique. Il nous semble un point essentiel de mettre à disposition ce travail à l'ensemble de la communauté biologiste, afin d'en tirer le plus grand profit. Pour cela, nous avons écrit le logiciel SIC (Robelin et al., 2003) disponible sous licence GPL. La disponibilité du code source permet d'en modifier le contenu et de l'adapter éventuellement à un problème particulier*. Ce logiciel est écrit en C++, et s'appuie sur la librairie Seq++ (Miele et al., 2005) développée par les membres du laboratoire "Statistique et Génome". Cette librairie implémente les différents modèles markoviens présentés dans le 1er chapitre. Elle permet de lire facilement une séquence au format FASTA ou Genbank, mais permet également à l'utilisateur de définir lui même, grâce à un fichier de configuration, un alphabet quelconque adapté à sa séquence (alphabet d'hydropobicité, ou alphabet d'angles par exemple).

Le logiciel SIC peut être téléchargé à l'adresse suivante :

http ://stat.genopole.cnrs.fr/SIC/

Il est prévu pour être compilé dans un environnement de type UNIX/Linux disposant de la bibliothèque Gnu Scientific Library (GSL) largement répandue†.

Ce logiciel peut être installé localement pour une utilisation intensive, et plus souple, ou être utilisé via un navigateur web à l'adresse suivante :

http ://stat.genopole.cnrs.fr/websic

Dans le cas d'une installation locale, la spécification des différents paramètres se fait à l'aide d'un fichier de configuration (voir la documentation du logiciel pour la syntaxe). En ce qui concerne le service web, le logiciel est hébergé par le laboratoire "Statistique et Génome" et son utilisation est gérée par un serveur d'application Apache-Tomcat. Il propose un accès simplifié au logiciel SIC. La figure 6.1 page 140 présente la page de saisie des paramètres. L'utilisateur peut choisir l'alphabet nucléotidique ou protéique, ou bien spécifier l'alphabet de son choix à l'aide d'un fichier de configuration. La séquence, sur laquelle la recherche d'inversions est effectuée, est spécifiée soit à l'aide d'un copier-coller, soit en indiquant le nom du fichier si la séquence est longue. La séquence du virus HIV est également disponible et permet d'essayer le logiciel. Il reste ensuite à préciser l'ordre du modèle markovien, la méthode utilisée (score local, ou fenêtre glissante) et le nombre de résultats affichés (soit à l'aide d'un nombre fixé, soit en spécifiant un seuil de probabilité critique). Remarquons que le calcul des probabilités critiques nécessite une étape d'évaluation par Monte-Carlo des fonctions de répartition. Cette étape est relativement lente (1 à 2 minutes actuellement) et ne change pas l'ordre des segments détectés, mais associe une probabilité critique à chaque score. Une option a donc été

*Merci de nous avertir dans ce cas : robelin@genopole.cnrs.fr

†http ://www.gnu.org/software/gsl/

ajoutée afin de supprimer éventuellement le calcul de ces probabilités afin d'accélérer le processus (qui devient quasiment instantané).

La figure 6.2 page 141 présente la page de résultat (obtenu ici sur le génome du VIH1 pour la méthode du score local). Le graphe représente la valeur de $S_k - \min_{i<k} S_k$ pour k variant de 1 à la taille de la séquence, où S_k est la somme cumulée des scores élémentaires. Pour mémoire : le score local maximal représente le maximum de ces quantités ; par conséquent, plus cette valeur est élevée, plus la présence d'un segment inversé est suspectée.

6.2 Etude de génomes viraux

Les virus possèdent généralement des génomes linéaires, avec un degré d'orientation relativement élevé par rapport aux séquences bactériennes. On a vu au paragraphe précédent que le degré d'orientation du génome est fortement lié à la puissance de la méthode de détection. L'étude sur ces génomes fournit donc un cadre favorable pour l'utilisation de SIC.

L'ensemble des séquences ont été téléchargées sur le serveur du NCBI (National Center for Biotechnology Information).

Pour chacune de ces séquences, nous avons estimé les matrices de transition de modèles de Markov d'ordre 1 à 4 sur l'ensemble du gènome. Ces matrices représententla matrice Q^+ dans les développements théoriques présentés dans les chapitres précédents. Nous faisons donc implicitement l'hypothèse que les éventuels retournements qui se produisent dans la séquence représentent une faible partie du génome et, par conséquent, influencent de manière négligeable l'estimation de la "vraie" matrice Q^+. Nous devons souligner ici que les développements théoriques présentés précédemment considèrent que la matrice de transition Q^+ est connue, et ne prennent pas en compte la variabilité liée à son estimation. Néanmoins, il est possible de considérer que la matrice estimée sur la séquence représente les probabilités d'apparition d'une lettre conditionnellement à son contexte (une autre lettre pour un modèle d'ordre 1) dans la séquence considérée, mais que cette séquence n'est pas aléatoire ; Le modèle indique seulement que si l'on considère une position au hasard dans la séquence et que la lettre u apparait, alors la probabilité d'observer ensuite la lettre v est donnée par $Q(u, v)$ (où Q est la matrice de transition de la chaîne de Markov ajustée sur la séquence). Un autre point de vue consiste à considérer que la séquence est le résultat d'un processus d'évolution complexe, dans lequel le hasard tient une place importante ; de ce point de vue, la composition de la séquence peut-être considérée comme aléatoire.

Ces considérations ont des conséquences sur la significativité d'une statistique calculée sur la séquence, telle que nos mesures de retournement par exemple. En effet, dans le premier point de vue, la statistique obtenue sur la séquence doit être comparée aux statistiques obtenues sur toutes les séquences qui présentent les mêmes probabilités de

FIG. 6.1 – Capture d'écran de l'interface Web du logiciel SIC : Ecran de saisie des paramètres.

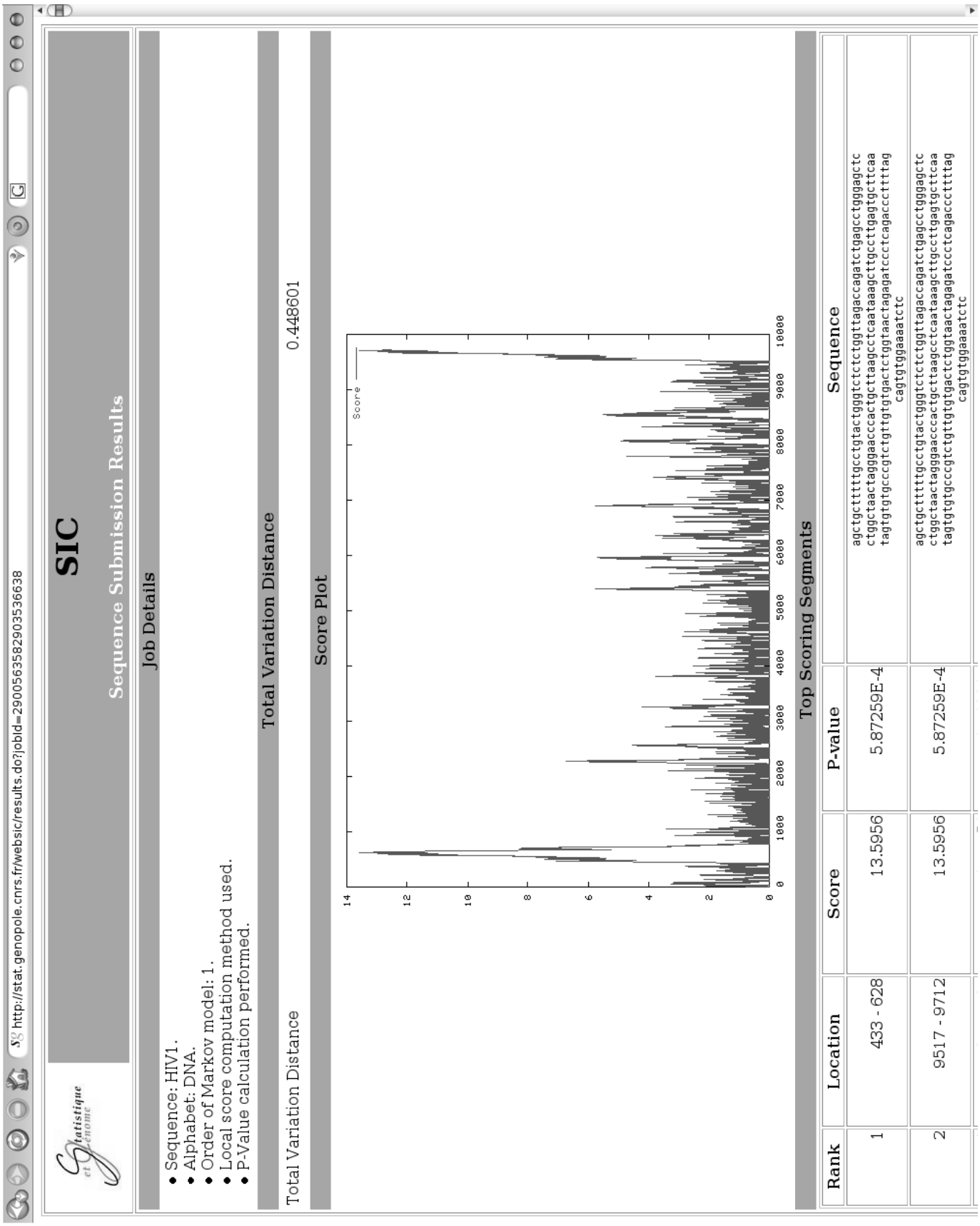


FIG. 6.2 – Capture d'écran de l'interface Web du logiciel SIC : Ecran de présentation des résultats.

succession de lettres **conditionnellement à la séquence**, c'est à dire qui donneront exactement la même matrice estimée Q (en considérant un modèle uniforme sur toutes ces séquences). Pour un modèle de Markov d'ordre m, on s'intéresse donc à toutes les séquences qui conduisent aux mêmes nombres d'occurence de tous les mots de taille $m+1$ que la séquence considérée : on parle de modèle "shuffle". Dans le deuxième point de vue, la séquence est une réalisation d'une chaîne de Markov. Cela a deux implications. La première est que la matrice de transition n'est pas connue, mais est estimée, ce qui induit une certaine variabilité de cette estimation. La deuxième conséquence est que la statistique observée doit, par conséquent, être comparée à la statistique obtenue sur les réalisations de cette chaîne de Markov (dans une approche de Monte-Carlo par exemple). On comprend intuitivement que la variance de la statistique sera plus grande dans le deuxième modèle que dans le modèle "shuffle", qui est conditionnée par le nombre d'occurence des mots de la séquence considérée.

Dans notre cas, nous estimons la matrice de transition sur la séquence, puis nous estimons la distribution des différentes statistiques par Monte-Carlo en simulant des réalisation de la chaîne de Markov selon la matrice de transition estimée. Notre approche donnera une variance de la statistique supérieure à celle du modèle "shuffle", car les chaînes simulées ne sont pas contraintes à avoir les mêmes nombres d'occurences de mots de taille $m+1$ lettres, mais une variance inférieure à l'approche considérant que la séquence est une réalisation d'une chaîne de Markov, car l'incertitude liée à l'estimation de la matrice de transition est négligée.

6.2.1 Virus de l'Immunodéficience Humaine VIH1

La séquence utilisée porte le numéro d'accès NC_001802 sur le serveur du NCBI. C'est un génome linéaire d'un peu moins de 10000 bases, dont les gènes codent 9 protéines.

Les graphiques 6.3 page 143 présentent les valeurs $S_k - \min_{i<k} S_k$ pour k variant le long de la séquence, reflétant la présence d'inversion par la méthode du score local, lorsque la séquence est modélisée par une chaîne de Markov d'ordre 2 et 4. On distingue trois pics sur le premier graphique qui n'apparaissent plus lorsque l'ordre vaut 4. Ceci est probablement dû au fait que le segment détecté comme inversé est composé de motifs de taille 2, 3 ou 4 proche de la séquence complémentaire. Dans le modèle de Markov d'ordre 4, le nombre de paramètres est beaucoup plus élevé, et ces motifs sont intégrés dans le modèle et ne ressortent plus comme exceptionnels. Le tableau 6.1 page 144 donne les trois segments significatifs dans un modèle d'ordre 2. Notons que le pic observé sur le graphique correspondant au modèle d'ordre 4 n'est pas significatif (probabilité critique de 0.51).

La méthode par fenêtre glissante donne les résultats listés dans le tableau 6.2 page 144 pour un modèle de Markov d'ordre 2. Ces résultats sont illustrés par le graphique 6.4 page 145 représentant la valeur de la statistique de rapport de vraisemblance T_k pour k variant le long du génome.

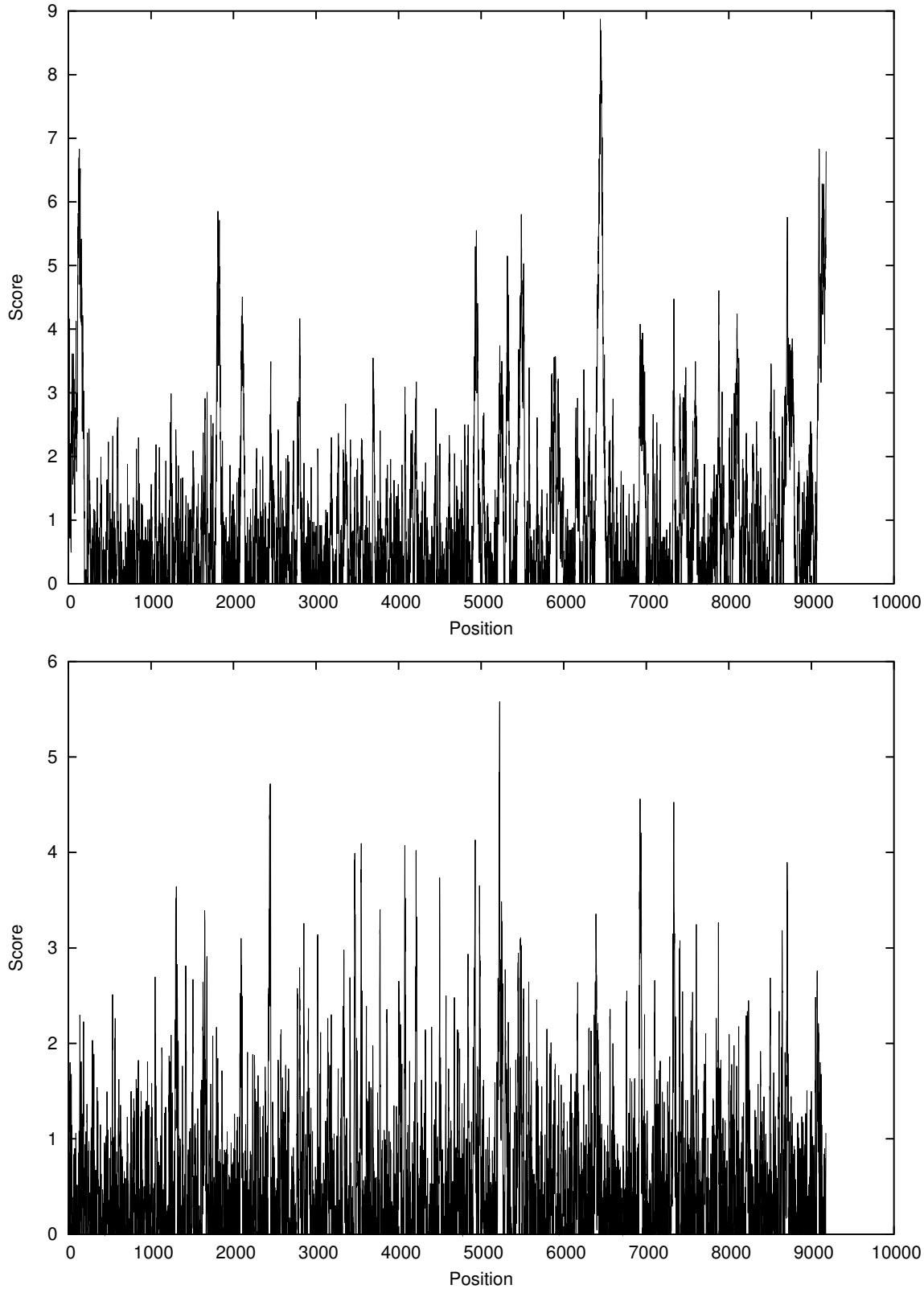


FIG. 6.3 – Score de retournement $S_k - \min_{i<k} S_k$ sur la séquence complète du Virus de l'Immunodéficience Humaine VIH1 pour des modèles de Markov d'ordre 2 et 4

.

Ordre	Position	Value	p-value
1	1 - 174	4.64	0.01
	gtctctctggttagaccagatctgagcctgggagctctctggctaactagggaaccc		
	actgcttaagcctcaataaagcttgccttgagtgcttcaagtagtgtgtgcccgtct		
	gttgtgtgactctggtaactagagatccctcagacccttttagtcagtgtggaaaatctc		
1	9065 - 9179	2.90	0.05
	gtgctttttgcctgtactgggtctctctggttagaccagatctgagcctgggagctc		
	tctggctaactagggaacccactgcttaagcctcaataaagcttgccttgagtgctt		
2	6377 - 6444	2.60	0.07
	aggcctgtccaaaggtatcctttgagccaattcccatacattattgtgccccggctggttttgcgatt		

TAB. 6.1 – Retournements détectés dans le virus VIH-1 en utilisant un modèle de Markov d'ordre 1 et 2 et la méthode du score local.

Window	Position	p-value
25	9069 - 9093	0.28
50	6397 - 6446	0.22
	6396 - 6445	0.25
75	6372 - 6446	0.04
	6379 - 6453	0.04
	6380 - 6454	0.04
	6382 - 6456	0.05
	6381 - 6455	0.05
100	6354 - 6453	0.05
150	6297 - 6446	0.05
	6304 - 6453	0.05
200	6269 - 6468	0.19

TAB. 6.2 – Retournements significativement détectés au risque 5% ou plus grand scores dans le virus VIH-1 avec la méthode par fenêtre glissante pour des tailles de fenêtres allant de 25 à 100 et un modèle de markov d'ordre 2.

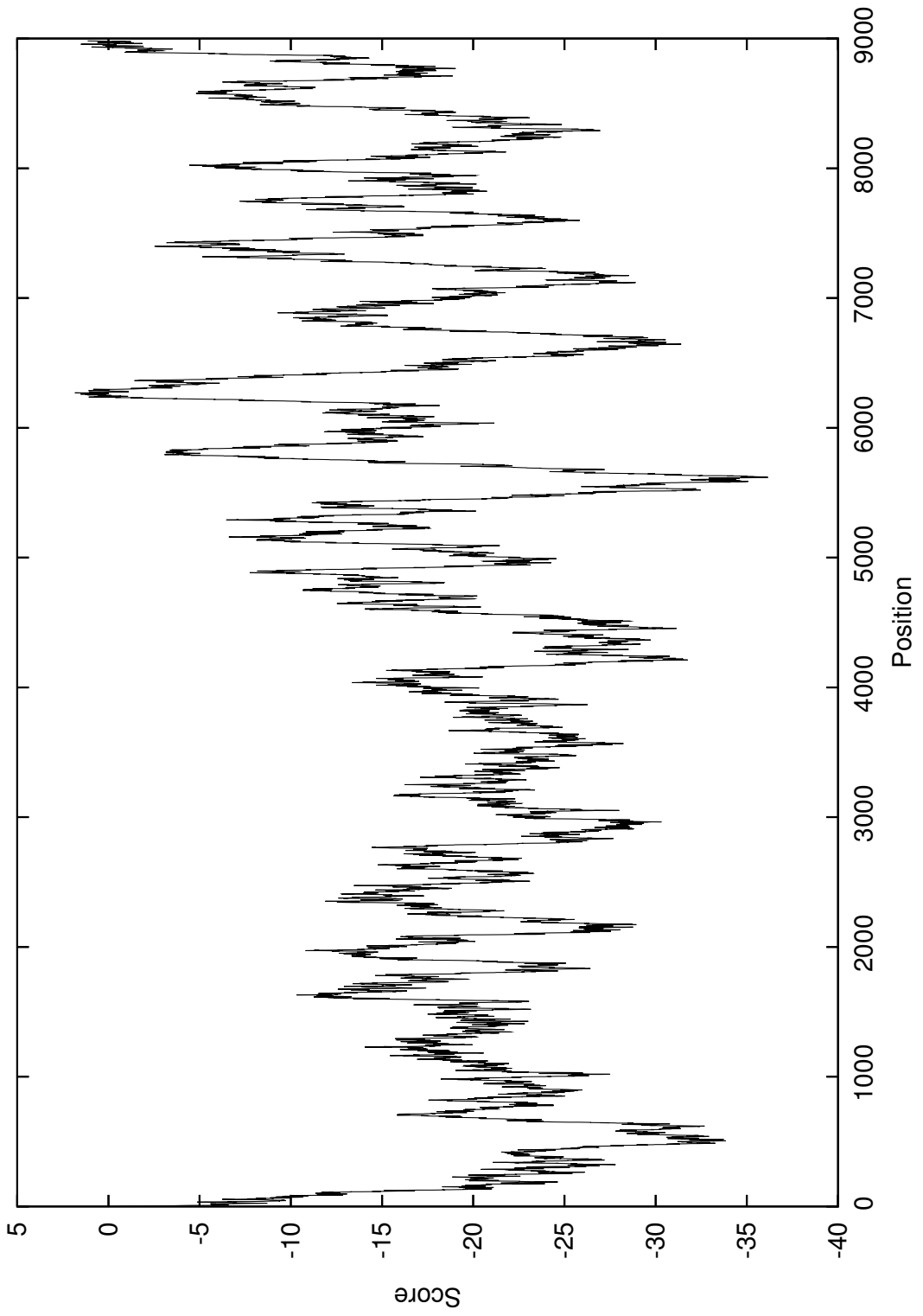


FIG. 6.4 – Valeur de la statistique de retournement du rapport de vraisemblance T_k sur la séquence complète du Virus de l'Immunodéficience Humaine VIH1 pour un modèle de Markov d'ordre 2 et une fenêtre glissante de taille 200.

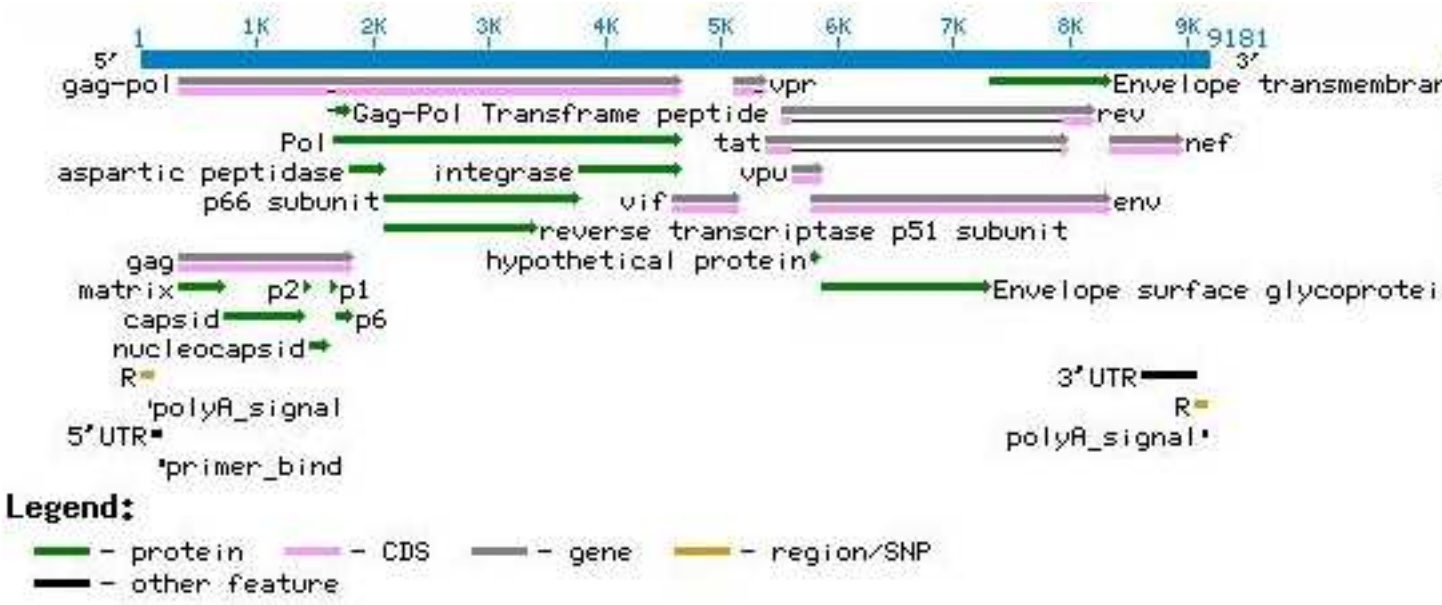


FIG. 6.5 – Représentation de l'annotation du génome VIH1 (source NCBI).

Ces études suggèrent la présence de deux segments de tailles 150 bases en début et fin de séquence, et éventuellement d'un troisième de taille 70 bases en milieu de séquence. Ces segments ne se confondent pas avec des gènes ou d'autres annotations connues de ce génome (voir le graphique 6.5 page 146).

6.2.2 Syndrome Respiratoire Aigu Sévère

La séquence utilisée porte le numéro d'accès NC_004718 sur le serveur du NCBI. Il s'agit d'un génome plus long (environ 30000 bases) dont les gènes codent 14 protéines.

Ce virus étant plus long, il a été arbitrairement découpé en 3 parties d'environ 10000 bases chacune. Ce découpage permettra de réduire l'effet d'un changement de composition en nucléotide le long de la séquence. Un modèle de Markov d'ordre 1 est ajusté sur chaque partie.

Seule l'étude fondée sur le score local est présentée, la taille des segments recherchés n'étant pas connue, et la puissance de cette méthode égalant presque la puissance de la méthode utilisant les fenêtres glissantes. Les graphiques 6.6 page 147 montre la statistique du score local sur chaque partie du génome pour un modèle de Markov d'ordre 1. Le tableau 6.3 page 148 indique les segments les plus significatifs. En comparant avec l'annotation de ce génome, on s'aperçoit que le segment de longueur 600 en position $[28134-28728]$ correspond à la première moitié du gène codant la protéine nucléocapside. L'autre segment détecté de longueur 750 ne semble pas lié à une annotation particulière.

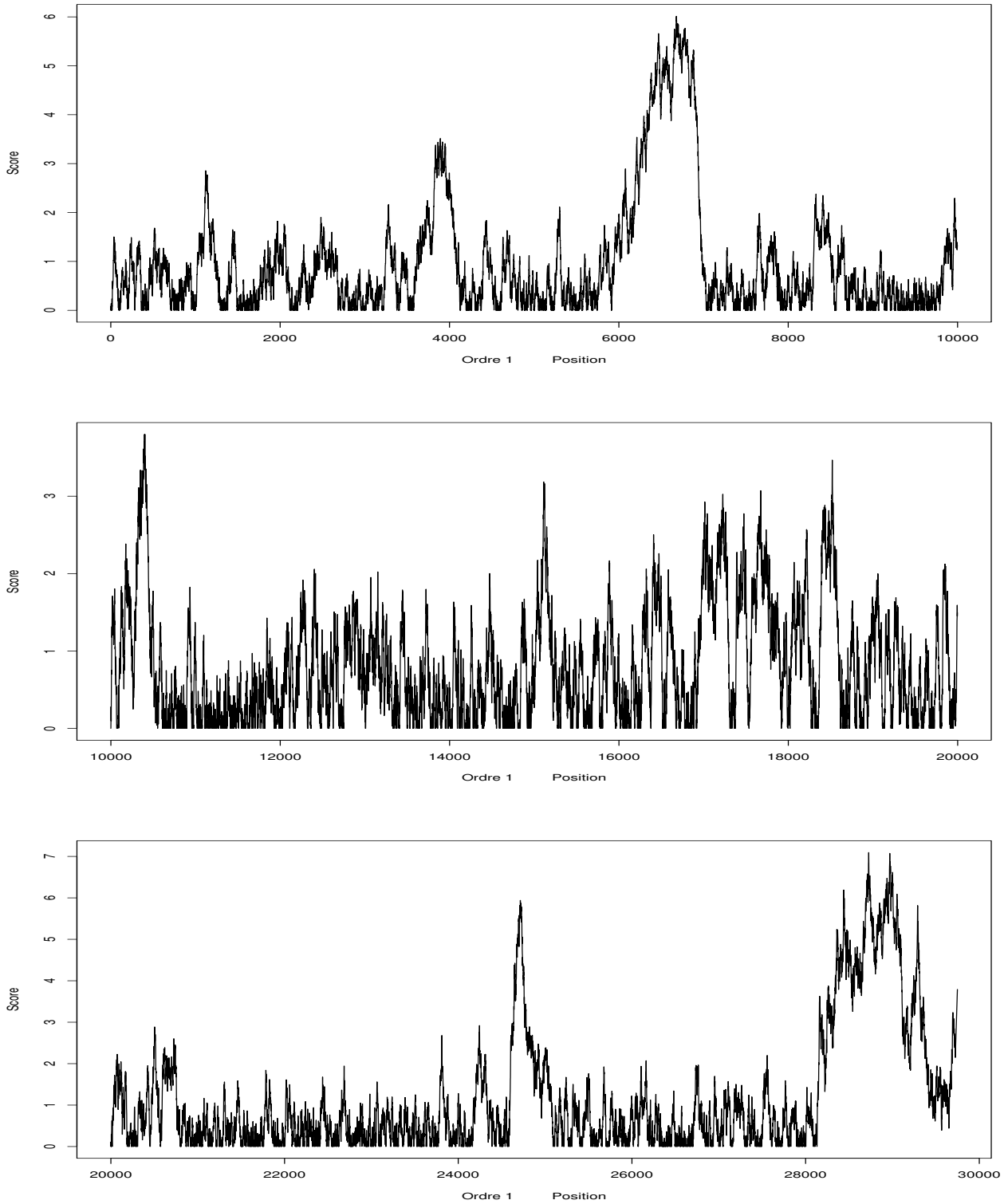


FIG. 6.6 – Score de retournement $S_k - \min_{i<k} S_k$ sur la séquence complète du SRAS pour des modèles de Markov d'ordre 1

.

Position	Value	p-value
5915 - 6679	2.24	0.10
28134 - 28728	2.53	0.08

TAB. 6.3 – Retournements détectés dans le virus du SRAS en utilisant un modèle de Markov d'ordre 2 et la méthode du score local sur 3 parties de taille 10000 chacune.

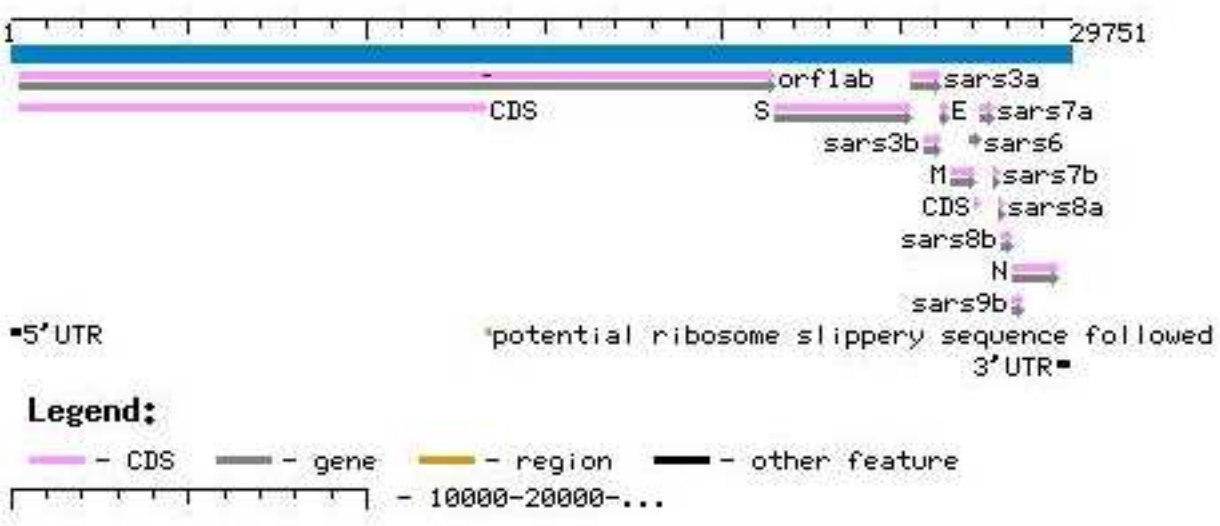


FIG. 6.7 – Représentation de l'annotation du génome SRAS (source NCBI).

6.2.3 Bactériophage Lambda

La séquence utilisée porte le numéro d'accès NC_001416 sur le serveur du NCBI. Il s'agit d'un génome linéaire plus long (environ 50000 bases) et d'un organisme plus complexe : ses gènes codent 71 protéines. Ce phage est un parasite de Escherichia coli.

Le graphique 6.8 page 150 présente la statistique du score local obtenue sur l'ensemble de la séquence. Il apparait une très large zone inversée entre les positions 20000 et 40000. Cette zone semble liée à la présence à la présence de gènes codant les ARNs, dont la plupart sont situés sur le brin complémentaire (voir graphique 6.9 page 151). L'origine de réplication est également située vers la position 40000.

Outre la mise en évidence de ce grand segment inversé, cette observation confirme, si nécessaire qu'il est utile de découper le génome pour en partie tenir compte de son inhomogénéité. Les figures 6.10 et 6.10 page 152 et 152 présentent les courbes sur les 5 parties de taille 10000 chacune. Les segments les plus significativement retournées sont reportés dans le tableau 6.4 page 149. Le premier retournement (3042 - 3130) se situe dans un gène de codant une composante de la capside (position 2836-4437), ainsi que le deuxième retournement (position du retournement : 6061 - 6344, position du gène : 6135-7160). Le troisième retournement est assez long, et aucun rapport n'apparait directement avec l'annotation. Le quatrième segment détecté se trouve près de l'origine de réplication. Le dernier segment (46223 - 46353) se trouve à la fin du gène (position 45966-46427) qui code une protéine concernant la lyse de la membrane de la cellule hôte.

Position	Value	p-value
3042 - 3130	2.10	0.11
6061 - 6344	2.87	0.06
20000 - 21625	55.67	$< 10^{-8}$
39173 - 39958	7.65	5×10^{-4}
46223 - 46353	1.95	0.13

TAB. 6.4 – Retournements détectés dans le bactériophage Lambda en utilisant un modèle de Markov d'ordre 1 et la méthode du score local sur 5 parties de 10000.

6.3 Discussion

Nous avons présenté dans ce chapitre trois exemples d'étude sur la présence d'éventuel segments inversés dans des génomes. Le terme de "segment inversé" un peu abusif, car il n'est pas possible à l'aide d'une simple étude sans comparaison à au moins une autre séquences de savoir s'il s'agit réellement d'une inversion au court de l'évolution, ou simplement d'un segment dont la composition est beaucoup plus proche de la séquence du brin complémentaire, plutôt que du brin principal. Ses études, bien qu'elles puissent sem-

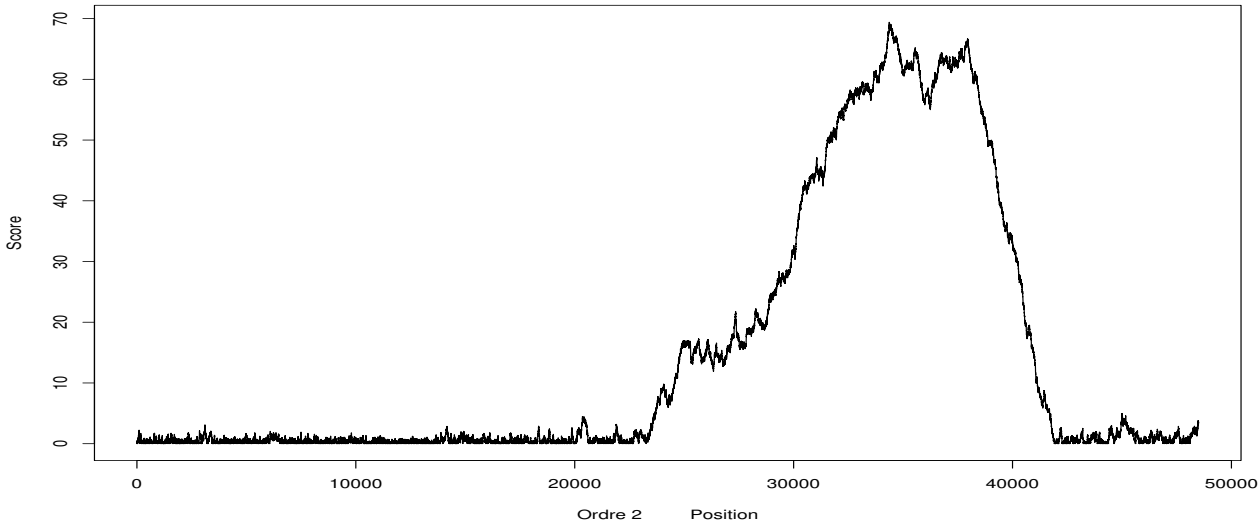


FIG. 6.8 – Score de retournement $S_k - \min_{i<k} S_k$ sur la séquence complète du bactériophage Lambda pour un modèle de Markov d'ordre 2

.

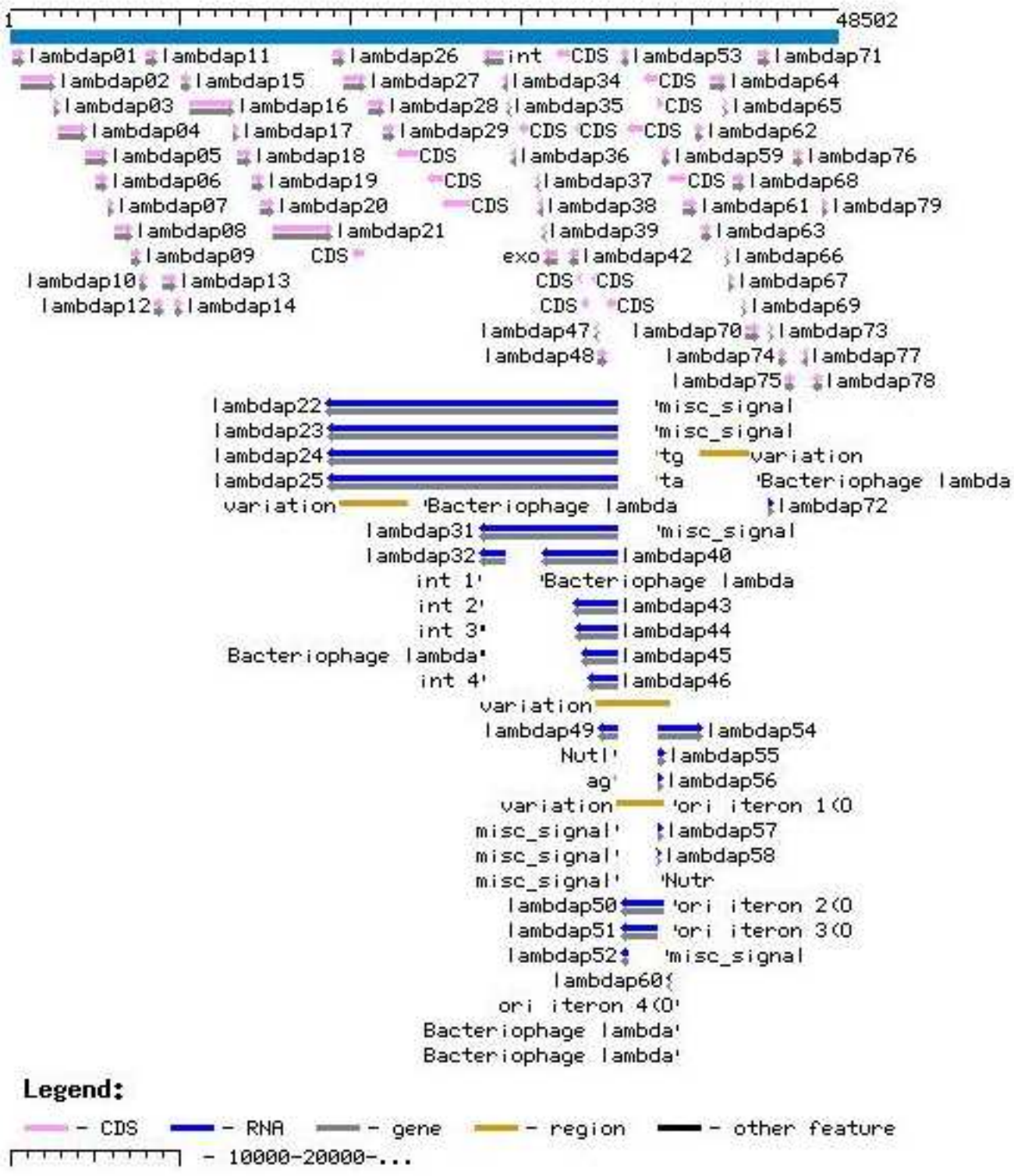


FIG. 6.9 – Représentation de l'annotation du génome bacteriophage lambda (source NCBI).

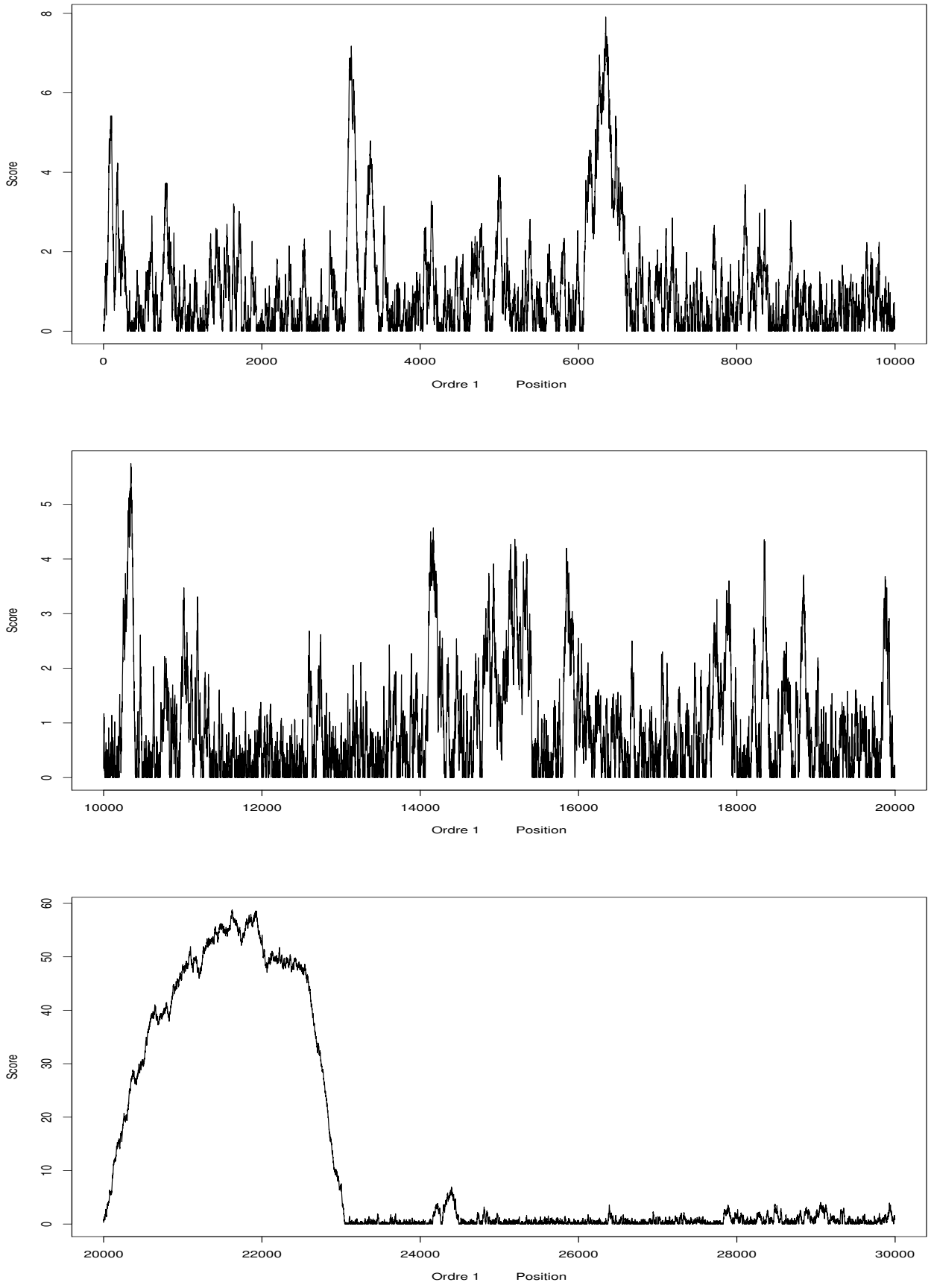


FIG. 6.10 – Score de retournement $S_k - \min_{i<k} S_k$ sur la séquence du bactériophage Lambda pour des modèles de Markov d'ordre 1 (position 0 à 29999)

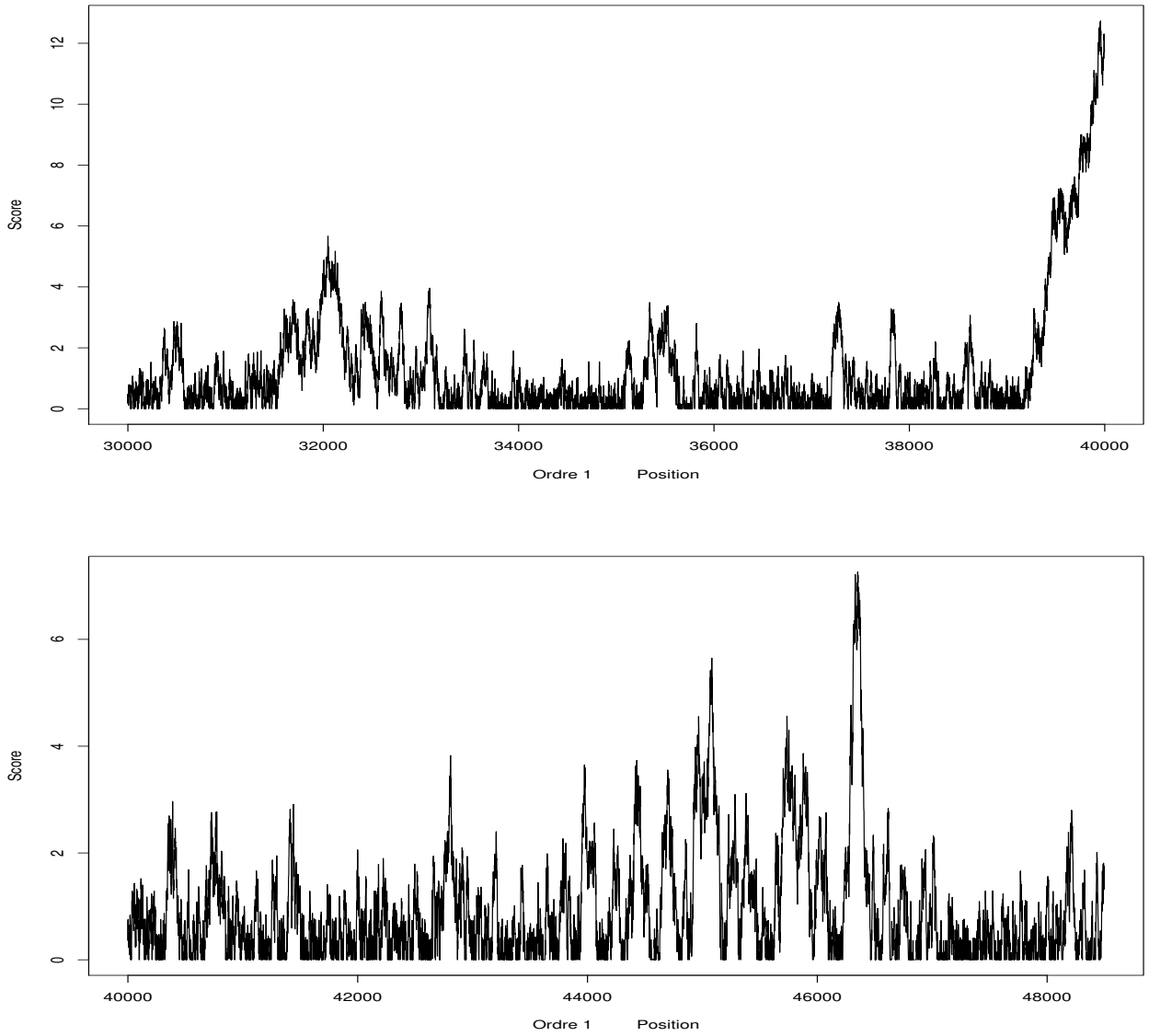


FIG. 6.11 – Score de retournement $S_k - \min_{i<k} S_k$ sur la séquence complète du bactériophage Lambda pour des modèles de Markov d'ordre 1 (position 30000 à 48501)

.

bler un peu grossière, permettent déjà de mettre en évidence la présence de tels segments et de les localiser.

L'existence du logiciel SIC permet aux personnes désireuses de pratiquer des études plus subtiles de les effectuer avec un minimum de travail. Je serais bien évidemment ravi d'apporter éventuellement mon aide dans ce cas.

Nous n'avons pas abordé ici le problème des génomes bactériens ou des eucaryotes. La structure des génomes eukaryotes est beaucoup plus complexe et requiert plus de finesse pour mettre en évidence des segments intéressants ; les éléments répétés requiert notamment un traitement particulier. Concernant les bactéries, nous avons été confrontés à un problème de manque d'orientation des génomes. En effet, la chaîne de Markov estimé sur un génome bactérien est quasiment réversible (au passage à l'alphabet complémentaire près), c'est à dire que la distribution de X est la même que la distribution de X^-. Les méthodes de détection ont une puissance nulle dans ce cas. Ce phénomène est du à une généralisation de la deuxième loi de parité de Chargaff (appelé aussi règle de parité II, voir Forsdyke and Mortimer (2000) pour une revue sur le sujet). Cette loi établi que le nombre de A (respectivement C) est environ le même que le nombre de T (respectivement G) sur un **même** brin d'ADN. Ceci est du au fait que l'ADN est constitué de deux brins **complémentaire**. C'est à dire qu'une mutation sur un brin entraîne également une mutation complémentaire sur l'autre brin. Le nombre de nouveau A sur un brin est égal au nombre de nouveau T sur le brin complémentaire. Imaginons que le taux de mutation d'un nucléotide en un autre soit le même pour les deux brins. Dans ce cas, le nombre de mutations vers un A, par exemple, est le même sur les deux brins, ce qui a pour conséquence, que sur un seul brin, le nombre de mutations vers un A sera environ le même que le nombre de mutations vers T, d'où la seconde règle de parité de Chargaff. On constate dans les séquence un écart à cette règle, car les taux de mutations "observés" sont différents selon les deux brins. Le débat est encore ouvert afin de savoir si cet écart est du à un taux de mutations différents selon l'endroit de la séquence, ou à un mécanisme de sélection a posteriori qui conduit à éliminer les individus ne présentant pas ces écarts (neutraliste vs selectionniste). La littérature est largement fournie à ce sujet.

Cette règle n'est pas seulement valable pour les nucléotides, mais aussi pour les dinucléotides, trinucléotides... (Nussinov, 1984; Alff-Steinberger, 1984; Yomo and Ohno, 1989; Prabhu, 1993). Par conséquent, si la règle est respectée, les chaînes de Markov ajustées sur les séquences sont réversibles. Généralement, on constate une déviation de cette règle selon deux critères. Le premier concerne la position par rapport à l'origine de réplication dans les génomes circulaires (voir Frank et al. (2000) pour une application à la détection de l'origine de réplication). On peut par exemple considérer le brin principale entre l'origine de réplication et la terminaison, puis le brin complémentaire. Le deuxième critère, appelé la règle de direction de transcription de Szybalski, concerne le brin sur lequel les gènes sont codés (uniquement valable pour les séquences codantes). Si un gène est codé sur le brin complémentaire, alors on considère sa séquence inverse complémentaire, sinon on considère sa séquence directe. Malheureusement, ces deux critères classiquement utili-

sés n'ont pas suffit dans notre cas pour obtenir des séquences plus orientées (résultats non présentés). En ce qui concerne le deuxième critère, Tillier and Collins (2000) ont mis en évidence une "adaptation" des gènes qui change de brin (par une inversion par exemple) pour être plus en accord avec le biais de composition de ce brin. Ceci contribue à atténuer le degré d'orientation du génome. Cette discussion nécessite une étude plus approfondie.

Bibliographie

Alff-Steinberger, C. (1984). Evidence for a coding pattern on the non-coding strand of the E. coli genome. *Nucleic Acids Res. J.*, 12 :2235–41.

Forsdyke, D, R. and Mortimer, J, R. (2000). Chargaff's legacy. *Gene J.*, 261 :127–37.

Frank, A, C., Frank, A., Lobry, J, R., and Lobry, J. (2000). Oriloc : prediction of replication boundaries in unannotated bacterial chromosomes. *Bioinformatics J.*, 16 :560–1.

Hannenhalli, S., Chappey, C., Koonin, E. V., and Pevzner, P. A. (1995). Genome Sequence Comparison and Scenarios for Gene Rearrangements : A Test Case. *Genomics*, 30(2) :299–311.

Lefebvre, J., El-Mabrouk, N., Tillier, E., and Sankoff, D. (2003). Detection and validation of single gene inversions. *Bioinformatics*, 19(90001) :190i–196.

Miele, V., Bourguignon, P.-Y., Robelin, D., Nuel, G. o., and Richard, H. (2005). seq++ : analyzing biological sequences with a range of Markov-related models. *Bioinformatics*, 21(11) :2783–2784.

Nussinov, R. (1984). Strong doublet preferences in nucleotide sequences and DNA geometry. *J. Mol. Evol. J.*, 20 :111–9.

Prabhu, V, V. (1993). Symmetry observations in long nucleotide sequences. *Nucleic Acids Res. J.*, 21 :2797–800.

Robelin, D., Richard, H., and Prum, B. (2003). SIC : a tool to detect short inverted segments in a biological sequence. *Nucl. Acids. Res.*, 31(13) :3669–3671.

Tillier, E, R. and Collins, R, A. (2000). Replication orientation affects the rate and direction of bacterial gene evolution. *J Mol Evol*, 51 :459–63.

Yomo, T. and Ohno, S. (1989). Concordant evolution of coding and noncoding regions of DNA made possible by the universal rule of TA/CG deficiency-TG/CT excess. *Proc. Natl. Acad. Sci. USA J.*, 86 :8452–6.

Chapitre 7

Discussion générale

Cette thèse concerne la détection et la localisation de courts segments inversés dans les génomes. Ses principaux apports sont d'ordre méthodologique. Outre une extension d'un résultat sur la vitesse de convergence d'une chaîne de Markov vers sa distribution stationnaire (chapitre 2), le chapitre 3 présente des avancés sur la problématique du score local, outil largement utilisé en bioinformatique. Ces avancés concernent le problème des r plus grandes valeurs de score local. Seul un article aborde ce problème dans la littérature (Karlin and Altschul Stephen, 1993). Les travaux présentés dans cette thèse concerne la significativité des r plus grandes valeurs de score local d'une part, et établit et évalue une démarche de tests pour choisir r. Enfin, une partie directement liée à notre problème initial propose une méthode de détection dans le cas où la longueur du segment retourné est connue, puis est généralisé au cas de recherche de segments retournés dont la longueur est inconnue, c'est à dire sur toutes les tailles de segments possible (chapitres 4 et 5). Le dernier chapitre (chapitre 6) présente une une recherche de segments inversés dans trois génomes viraux (VIH-1, SRAS et bactériophage Lambda) et aborde le problème des génomes bactériens dans la discussion.

Nous devons souligner que très peu de travaux existe sur le problème de l'inversion de courts segments d'ADN durant l'évolution. Un article de Gordon and Halliday (1995) décrit le phénomène biologique pouvant être à l'origine de telles inversions, mais aucune vérification expérimentale n'a été effectué à ce jour. Si un tel phénomène existe avec suffisamment d'ampleur, cela pourrait changer les distances établies entre génomes et avoir des conséquences en phylogénie d'une part, mais surtout au niveau de la recherche de similarité où les programmes d'alignement généralement utilisés comparent les séquences en progressant de manière linéaire le long des séquences (BLAST Altschul et al. (1990), par exemple). Une inversion courte dans une séquence conduirait de manière erronée à l'insertion de gap ou de mismatch dans l'alignement.

Deux articles ont abondé dans le sens de l'existence de retournement de segment d'ADN au niveau des protéines en se fondant sur une étude statistique sur les bases de

données de protéines (Goldstein et al., 2000, 2003). Les trois applications présentés sont un élément de plus apporté à cette hypothèse. Les graphiques, et les degrés de signification associés, mettent clairement en évidence des zones dont la composition est plus proche de la composition du brin complémentaire que du brin principal. Néanmoins, il n'est pas du tout sûr que ces zones proviennent d'une inversion. De plus, il s'agît de zones relativement longues (plus d'une centaine de bases en général), alors que les observations de Goldstein et al. conduisent plutôt à des retournements de l'ordre d'une vingtaine de bases. L'étude de simulation a montré que les méthodes proposées sont assez peu puissantes pour détecter des retournements de cette taille, à moins que la séquence ne soit très orientée.

Une solution pour contourner ce manque de puissance, et également pour mettre en évidence le fait qu'il s'agisse d'un retournement plutôt que d'une ressemblance avec le brin complémentaire, consiste à effectuer des études de génomique comparative explicitement dans ce dessein. Etant donné deux séquences "proches", le problème consiste à aligner ces deux séquences en permettant des retournements et des déplacements de segments. Ce problème est connu sous la dénomination "sous-ensemble non-conflictuel maximal de rectangle*" et il s'agit d'un problème NP-difficile (Bafna et al., 1996; Nagashima and Yamazaki, 2004). Une approximation du problème a été proposée par Berman et al. (2001), mais elle reste à être adaptée à l'alignement de séquences, et à être implémentée à notre connaissance.

Des outils plus spécifiquement orienté vers la bioinformatique sont apparus récemment. Citons les logiciels MUMmer (Kurtz et al., 2004) et Mauve (Darling et al., 2004) dans le cadre de l'alignement multiple, mais ces logiciels sont conçus pour aligner des génomes entiers, en prenant en compte des réarrangements à grande échelle. Citons enfin le travail de Conant and Wagner (2004) qui propose une méthode permettant de réordonner les segments obtenus à l'aide d'alignements locaux pour obtenir un alignement global des deux séquences.

Une dernière approche intérressante est la recherche de point de recombinaison dans un alignement multiple. Ces méthodes utilisent une fenêtre glissante dans laquelle un arbre phylogénétique est ajusté, et compare l'ajustement cet arbre au reste de l'alignement multiple. S'il y a eu recombinaison dans la fenêtre, alors cet ajustement sera de mauvaise qualité. On trouve ce principe dans les travaux de Paraskevis et al. (2005) implémentés dans le logiciel Milne et al. (2004), et les travaux de Husmeier and Wright (2005). Notre idée d'utiliser la chaîne de Markov inverse pourraient être inclue dans une telle approche pour voir si le point de recombinaison serait du à une inversion.

*maximum non-conflict subset of retangles

Bibliographie

Altschul, S, F., Gish, W., Miller, W., Myers, E, W., and Lipman, D, J. (1990). Basic local alignment search tool. *J Mol Biol J*, 215 :403–410.

Bafna, V., Narayanan, B., and Ravi, R. (1996). Nonoverlapping local alignments (weighted independent sets of axis-parallel rectangles). *Discrete Applied Mathematics*, 71.

Berman, P., DasGupta, B., Muthukrishnan, S., and Ramaswami, S. (2001). Efficient Approximation Algorithms for Tiling and Packing Problems with Rectangles. *Journal of Algorithms*, 41(2) :443–470.

Conant, Gavin, C. and Wagner, A. (2004). A fast algorithm for determining the best combination of local alignments to a query sequence. *BMC Bioinformatics*, 5 :62.

Darling, A. C., Mau, B., Blattner, F. R., and Perna, N. T. (2004). Mauve : Multiple Alignment of Conserved Genomic Sequence With Rearrangements. *Genome Res.*, 14(7) :1394–1403.

Goldstein, D., Muri, F., Saragueta, P., and Prum, B. (2000). Inverse complementary homologues of short cysteine signatures. *C R Acad Sci III*, 323 :167–172.

Goldstein, D. J., Fondrat, C., Muri, F., Nuel, G., Saragueta, P., Tocquet, A.-S., and Prum, B. (2003). Short inverse complementary amino acid sequences generate protein complexity. *Comptes Rendus Biologies*, 326(3) :339–348.

Gordon, A. and Halliday, J. (1995). Inversions with deletions and duplications. *Genetics J*, 140 :411–4.

Husmeier, D. and Wright, F. (2005). Detecting interspecific recombination with a pruned probabilistic divergence measure. *Bioinformatics*, 21(9) :1797–806.

Karlin, S. and Altschul Stephen, F. (1993). Applications and statistics for multiple high-scoring segments in molecular sequences. *Proc. Natl. Acad. Sci.*, 90 :5873–5877.

Kurtz, S., Phillippy, A., Delcher, A, L., Smoot, M., S humway, M., Antonescu, C., and Salzberg, Steven, L. (2004). Versatile and open software for comparing large genomes. *Genome Biol J*, 5 :R12.

Milne, I., Wright, F., Rowe, G., Marshall, David, F. H. D., and McGuire, G. (2004). TOPALi : software for automatic identification of recombinant sequences within DNA multiple alignments. *Bioinformatics*, 20 :1806–7.

Nagashima, H. and Yamazaki, K. (2004). Hardness of approximation for non-overlapping local alignments. *Discrete Applied Mathematics*, 137(3) :293–309.

Paraskevis, D., Deforche, K., Lemey, P., Magiorkinis, Gand Hatzakis, A., and Vandamme, A. (2005). SlidingBayes : exploring recombination using a sliding window approach based on Bayesian phylogenetic inference. *Bioinformatics*, 21 :1274–5.

Annexe A

Loi jointes de r plus grandes valeurs d'un n-échantillon

Dans cet annexe, on utilisera une notation qui diffère de celle généralement utilisée lorsque l'on traite des statistiques d'ordre ; ici, la plus grande valeur de l'échantillon de taille n est notée $M_n^{(1)}$, la deuxième plus grande valeur est notée $M_n^{(2)}$, etc. On a donc : $M_n^{(1)} > M_n^{(2)} > ... > M_n^{(n)}$.

On rappelle que le lemma 3.13 page 62 donne la densité jointe des r plus grandes valeurs normalisées d'une suite $Y_1 ... Y_n$ de variables aléatoires indépendantes et identiquement distribuées lorsque le domaine d'attraction de la loi de Y_1 est la loi de Gumbel :

$$L\left(\frac{M_n^{(1)} - b_n}{a_n}, ..., \frac{M_n^{(r)} - b_n}{a_n}\right) = \exp\left\{-\exp\left(-\frac{M_n^{(r)} - \mu}{\sigma}\right)\right\} \prod_{i=1}^{r} \sigma^{-1} \exp\left(-\frac{M_n^{(i)} - \mu}{\sigma}\right)$$

où $M_n^{(1)}$ est supposé converger vers une loi de Gumbel :

$$\lim_{n\to\infty} \mathbb{P}(M_n^{(1)} < a_n x + b_n) = \exp(-\exp(-(x-\mu)/\sigma))$$

et a_n et b_n sont deux suites de constantes dépendantes de la loi de Y_i.
On s'intéresse maintenant à la distribution jointe des $r-i$ plus grandes valeurs normalisées $M_n^{(i)} > M_n^{(i+1)} > ... > M_n^{(r)}$ où $i = 1, 2, ..., r-1$. La densité jointe de ces variables est donnée par

$$f_{M_n^{(i)}, M_n^{(i+1)}, ..., M_n^{(r)}}(x_i, x_{i+1}, ..., x_r) = \frac{e^{-e^{\left(-\frac{x_r-\mu}{\sigma}\right)}}}{\sigma} \frac{\left(e^{-\frac{x_i-\mu}{\sigma}}\right)^i}{(i-1)!} \prod_{j=i+1}^{r} \frac{e^{\left(-\frac{x_j-\mu}{\sigma}\right)}}{\sigma}.$$

Cette expression se vérifie par le calcul récursif suivant. On a vu le cas $i = 1$ ci-dessus,

on va vérifier le cas i en supposant vrai le cas $i-1$:

$$\begin{aligned}
f_{M_n^{(i)},\dots,M_n^{(r)}}(x_i,\dots,x_r) &= \int_{x_i}^{+\infty} f_{M_n^{(i-1)},M_n^{(i)},\dots,M_n^{(r)}}(x_{i-1},x_i\dots,x_r)dx_{i-1} \\
&= \frac{e^{-e^{\left(-\frac{x_r-\mu}{\sigma}\right)}}}{(i-2)!}\int_{x_i}^{+\infty}\frac{\left(e^{-\frac{x_{i-1}-\mu}{\sigma}}\right)^{i-1}}{\sigma}dx_{i-1}\prod_{j=i}^{r}\frac{e^{\left(-\frac{x_j-\mu}{\sigma}\right)}}{\sigma} \\
&= \frac{e^{-e^{\left(-\frac{x_r-\mu}{\sigma}\right)}}}{(i-2)!}\int_0^{e^{-\frac{x_i-\mu}{\sigma}}} y^{i-2}\,dy\prod_{j=i}^{r}\frac{e^{\left(-\frac{x_j-\mu}{\sigma}\right)}}{\sigma} \\
&= \frac{e^{-e^{\left(-\frac{x_r-\mu}{\sigma}\right)}}}{(i-2)!}\frac{\left(e^{-\frac{x_i-\mu}{\sigma}}\right)^{i-1}}{i-1}\prod_{j=i}^{r}\frac{e^{\left(-\frac{x_j-\mu}{\sigma}\right)}}{\sigma} \\
&= \frac{e^{-e^{\left(-\frac{x_r-\mu}{\sigma}\right)}}}{\sigma}\frac{\left(e^{-\frac{x_i-\mu}{\sigma}}\right)^{i}}{(i-1)!}\prod_{j=i+1}^{r}\frac{e^{\left(-\frac{x_j-\mu}{\sigma}\right)}}{\sigma}.
\end{aligned}$$

Il sera utile dans la suite d'avoir l'expression de la fonction de répartition de $M_n^{(i)}, M_n^{(i+1)}, ..., M_n^{(r)}$. On l'exprime ci-dessous de façon récursive :

$$\mathbb{P}\left(M_n^{(i)} < x_i, M_n^{(i+1)} < x_{i+1}, ..., M_n^{(r)} < x_r\right) =$$

$$\begin{aligned}
&= \int_{-\infty}^{x_r}\int_{t_r}^{x_{r-1}} ... \int_{t_{i+1}}^{x_i} \frac{e^{-e^{\left(-\frac{t_r-\mu}{\sigma}\right)}}}{\sigma}\frac{\left(e^{-\frac{t_i-\mu}{\sigma}}\right)^i}{(i-1)!}\prod_{j=i+1}^{r}\frac{e^{\left(-\frac{t_j-\mu}{\sigma}\right)}}{\sigma}\,dt_i...dt_r \\
&= \int_{-\infty}^{x_r} e^{-e^{\left(-\frac{t_r-\mu}{\sigma}\right)}}\frac{e^{\left(-\frac{t_r-\mu}{\sigma}\right)}}{\sigma}\int_{t_r}^{x_{r-1}}\frac{e^{\left(-\frac{t_{r-1}-\mu}{\sigma}\right)}}{\sigma}... \int_{t_{i+1}}^{x_i}\frac{\left(e^{-\frac{t_i-\mu}{\sigma}}\right)^i}{\sigma(i-1)!}\,dt_i...dt_r \\
&= \int_{-\infty}^{x_r} e^{-e^{\left(-\frac{t_r-\mu}{\sigma}\right)}}\frac{e^{\left(-\frac{t_r-\mu}{\sigma}\right)}}{\sigma}\int_{t_r}^{x_{r-1}}\frac{e^{\left(-\frac{t_{r-1}-\mu}{\sigma}\right)}}{\sigma} \\
&\quad ... \int_{t_{i+2}}^{x_{i+1}}\frac{e^{\left(-\frac{t_{i+1}-\mu}{\sigma}\right)}}{\sigma}\left(\int_{e^{-\frac{x_i-\mu}{\sigma}}}^{e^{-\frac{t_{i+1}-\mu}{\sigma}}}\frac{y^{i-1}}{(i-1)!}dy\right)dt_{i+1}...dt_r \\
&= \int_{-\infty}^{x_r} e^{-e^{\left(-\frac{t_r-\mu}{\sigma}\right)}}\frac{e^{\left(-\frac{t_r-\mu}{\sigma}\right)}}{\sigma}\int_{t_r}^{x_{r-1}}\frac{e^{\left(-\frac{t_{r-1}-\mu}{\sigma}\right)}}{\sigma} \\
&\quad ... \int_{t_{i+2}}^{x_{i+1}}\frac{e^{\left(-\frac{t_{i+1}-\mu}{\sigma}\right)}}{\sigma\, i!}\left[\left(e^{-\frac{t_{i+1}-\mu}{\sigma}}\right)^i - \left(e^{-\frac{x_i-\mu}{\sigma}}\right)^i\right]dt_{i+1}...dt_r \\
&= \mathbb{P}\left(M_n^{(i+1)} < x_{i+1}, ..., M_n^{(r)} < x_r\right) - \frac{\left(e^{-\frac{x_i-\mu}{\sigma}}\right)^i}{i!}\mathbb{P}\left(M_n^{(1)} < x_{i+1}, ..., M_n^{(r-i)} < x_r\right) \\
&= \mathbb{P}\left(M_n^{(r)} < x_r\right) - \sum_{j=i}^{r-1}\frac{\left(e^{-\frac{x_j-\mu}{\sigma}}\right)^j}{j!}\mathbb{P}\left(M_n^{(1)} < x_{j+1}, ..., M_n^{(r-j)} < x_r\right)
\end{aligned}$$

(Une version plus explicite se trouve à la page suivante).

La densité de la distribution de chaque valeur $M_n^{(i)}$ normalisée par a_n et b_n se déduit de la densité jointe. On en détaille le calcul ici :

$$
\begin{aligned}
f_{M_n^{(i)}}(t_i) &= \int_{t_i}^{+\infty} ... \int_{t_2}^{+\infty} e^{-e^{\left(-\frac{t_i-\mu}{\sigma}\right)}} \prod_{j=1}^{i} \frac{e^{\left(-\frac{t_j-\mu}{\sigma}\right)}}{\sigma}\, dt_1 ... dt_{i-1} \\
&= e^{-e^{\left(-\frac{t_i-\mu}{\sigma}\right)}} \frac{e^{\left(-\frac{t_i-\mu}{\sigma}\right)}}{\sigma} \int_{t_i}^{+\infty} ... \int_{t_3}^{+\infty} \prod_{j=2}^{i-1} \frac{e^{\left(-\frac{t_j-\mu}{\sigma}\right)}}{\sigma} \left(\int_{t_2}^{+\infty} \frac{e^{\left(-\frac{t_1-\mu}{\sigma}\right)}}{\sigma} dt_1 \right) dt_2 ... dt_{i-1} \\
&= e^{-e^{\left(-\frac{t_i-\mu}{\sigma}\right)}} \frac{e^{\left(-\frac{t_i-\mu}{\sigma}\right)}}{\sigma} \int_{t_i}^{+\infty} ... \int_{t_3}^{+\infty} \prod_{j=2}^{i-1} \frac{e^{\left(-\frac{t_j-\mu}{\sigma}\right)}}{\sigma}\, e^{\left(-\frac{t_2-\mu}{\sigma}\right)} dt_2 ... dt_{i-1} \\
&= e^{-e^{\left(-\frac{t_i-\mu}{\sigma}\right)}} \frac{e^{\left(-\frac{t_i-\mu}{\sigma}\right)}}{\sigma} \int_{t_i}^{+\infty} ... \int_{t_4}^{+\infty} \prod_{j=3}^{i-1} \frac{e^{\left(-\frac{t_2-\mu}{\sigma}\right)}}{\sigma} \left(\int_{0}^{e^{-\frac{t_j-\mu}{\sigma}}} y\, dy \right) dt_3 ... dt_{i-1} \\
&= e^{-e^{\left(-\frac{t_i-\mu}{\sigma}\right)}} \frac{e^{\left(-\frac{t_i-\mu}{\sigma}\right)}}{\sigma} \int_{t_i}^{+\infty} ... \int_{t_4}^{+\infty} \prod_{j=3}^{i-1} \frac{e^{\left(-\frac{t_2-\mu}{\sigma}\right)}}{\sigma} \frac{1}{2} \left(e^{-\frac{t_3-\mu}{\sigma}} \right)^2 dt_3 ... dt_{i-1} \\
&= ... = \exp\left\{ -\exp\left(-\frac{t_i-\mu}{\sigma} \right) \right\} \frac{1}{(i-1)!\,\sigma} \left\{ \exp\left(-\frac{t_i-\mu}{\sigma} \right) \right\}^i
\end{aligned}
$$

La fonction de répartition correspondante est obtenue par intégration de la densité de la façon suivante :

$$
\begin{aligned}
\mathbb{P}\left(M_n^{(i)} < x\right) &= \int_{-\infty}^{x} \exp\left\{ -\exp\left(-\frac{t_i-\mu}{\sigma} \right) \right\} \frac{1}{(i-1)!\,\sigma} \left\{ \exp\left(-\frac{t_i-\mu}{\sigma} \right) \right\}^i dt_i \\
&= \frac{1}{(i-1)!} \int_{e^{-\frac{x-\mu}{\sigma}}}^{\infty} e^{-y}\, y^{i-1} dy \\
&= \frac{1}{(i-1)!} \left\{ \exp\left(-\frac{x-\mu}{\sigma} \right) \right\}^{i-1} \exp\left\{ -\exp\left(-\frac{x-\mu}{\sigma} \right) \right\} \\
&\quad + \frac{1}{(i-2)!} \int_{e^{-\frac{x-\mu}{\sigma}}}^{\infty} e^{-y}\, y^{i-2} dy \\
&= \frac{1}{(i-1)!} \left\{ \exp\left(-\frac{x-\mu}{\sigma} \right) \right\}^{i-1} \exp\left\{ -\exp\left(-\frac{x-\mu}{\sigma} \right) \right\} + \mathbb{P}\left(M_n^{(i-1)} < x\right)
\end{aligned}
$$

On peut la présenter de façon équivalente avec les trois expressions suivantes :

$$
\mathbb{P}\left(M_n^{(i)} < x\right) = \mathbb{P}\left(M_n^{(i-1)} < x\right) + \frac{1}{(i-1)!} \left\{ \exp\left(-\frac{x-\mu}{\sigma} \right) \right\}^{i-1} \exp\left\{ -\exp\left(-\frac{x-\mu}{\sigma} \right) \right\}
$$

$$\mathbb{P}\left(M_n^{(i)} < x\right) = \exp\left\{-\exp\left(-\frac{x-\mu}{\sigma}\right)\right\} \sum_{j=0}^{i-1} \frac{\left(\exp\left(-\frac{x-\mu}{\sigma}\right)\right)^j}{j!}$$

$$\mathbb{P}\left(M_n^{(i)} < x\right) = \mathbb{P}\left(Y < i\right) \text{ où } Y \sim Poisson\left\{\exp\left(-\frac{x-\mu}{\sigma}\right)\right\}$$

Cette fonction de répartition n'est pas loglog-linéaire.

On reprend la fonction de répartition de $M_n^{(i)}, M_n^{(i+1)}, ..., M_n^{(r)}$, car on peut maintenant mieux l'expliciter de la façon suivante :

$$\mathbb{P}\left(M_n^{(i)} < x_i, M_n^{(i+1)} < x_{i+1}, ..., M_n^{(r)} < x_r\right) =$$

$$= \mathbb{P}\left(M_n^{(r)} < x_r\right) - \sum_{j=i}^{r-1} \frac{\left(e^{-\frac{x_j-\mu}{\sigma}}\right)^j}{j!} \mathbb{P}\left(M_n^{(1)} < x_{j+1}, ..., M_n^{(r-j)} < x_r\right)$$

Détection de courts segments inversés dans les génomes : méthodes et applications

L'inversion de courts segments (moins de 1000 bases) est soupçonnée être un mécanisme majeur de l'évolution des génomes. Deux méthodes de détection ab initio de tels segments sont présentées. La séquence est modélisée par une chaîne de Markov X^+. La séquence inverse-complémentaire est alors également modélisée par une chaîne de Markov notée X^-. Le premier chapitre présente de façon didactique les modèles de Markov utilisés en analyse de séquences génomiques. Une généralisation au cas d'un ordre supérieur à 1 d'un résultat sur la vitesse de convergence vers la distribution stationnaire est également établie. Le deuxième chapitre est consacré à l'étude du score local : $H_n = \max_{1 \leqslant i \leqslant j \leqslant n}(Y_i + ... + Y_j)$, pour une séquence $(Y_1, ..., Y_n) \in \mathbb{R}^n$. La loi jointe asymptotique des r plus grandes valeurs de score local est établie à l'aide de la théorie des valeurs extrêmes. Enfin, une démarche de test multiple permettant de choisir r est proposée. Le troisième chapitre propose une statistique de détection fondée sur un rapport de vraisemblance (modèle X^+ contre modèle X^-) lorsque la longueur du segment retourné est connue. Une approche de type "fenêtre glissante" est ensuite appliquée. Une approximation connue de la loi du maximum de ce type de statistique est utilisée pour associer un degré de signification aux segments détectés. Dans le quatrième chapitre, le cas de recherche de segments de longueurs inconnues est traité à l'aide d'une méthode de type score local. Le cinquième chapitre présente l'application de ces méthodes à quelques génomes viraux. Un logiciel développé pour traiter cette problématique est également présenté.

Mots-clés : chaîne de Markov, analyse de sequence génétique, score local, inversion courte, "scan" statistique.

Detection of short inverted segments in genomes : methods and applications

Inversion of short segments (less than 1000 bases) is suspected to be a major mechanism of genome's evolution. Two methods to detect ab initio these segments are presented. The sequence is modelized by a Markov chain X^+. Therefore the inverted complementary sequence is modelized by a Markov chain denoted X^-. The first chapter didactically presents the Markov models used in genomic sequence analysis. A generalization to order greater than 1 of a result on the speed of convergence of a Markov chain to its stationnary distribution is also established. The second chapter deals with the theory of local score $H_n = \max_{1 \leqslant i \leqslant j \leqslant n}(Y_i + ... + Y_j)$, for a sequence $(Y_1, ..., Y_n) \in \mathbb{R}^n$. The asymptotic joint distribution of the r greatest values of the local score is established using the extreme value theory. Finally a multiple test approach is derived to determine r. The third chapter propose a statistic of detection based on a likelihood ratio (model X^+ vs X^-) when the length of the inverted segment is known. A "scan-approach" is then applied. A known approximation of the distribution of the maximum of this type of statistic is then used for obtaining a p-value. In the forth chapter, the method of the local score is applied to deal with segments of unknown length and calculate the corresponding p-value. The fifth chapter presents the application of these methods to viral genomes. A software which implemented both methods is also presented.

Keywords : Markov chain, genetic sequence analysis, local score, short inversion, "scan" statistic.