Modèles statistiques pour l'analyse des séquences biologiques

Franck Picard*

*UMR CNRS-5558, Laboratoire de Biométrie et Biologie Evolutive

franck.picard@univ-lyon1.fr

Outline

1 Introduction

2 Préliminaires & Notations

3 Caractérisation statistique du modèle M0

4 Mise au point du Modèle Markovien d'ordre 1 (M1) et généralisation

5 Quel modèle choisir ?

6 Définition mathématique d'un motif dans une séquence

Rapide historique de l'analyse automatique de séquences

- Avant la détermination de la structure de l'ADN, il n'existait pas de bases moléculaires à la génétique
- Une fois la structure élucidée (succession de 4 monomères) se pose la question de l'analyse de l'information contenue dans la molécule
- L'ancrage dans l'algorithmique du texte est "immédiat" : l'ADN est un texte composé de 4 lettres
- Comment décoder ce langage ?
- C'est aussi la "grande époque" de l'analyse automatique du langage
- Comment extraire l'information contenue dans les séquences ?
- À la main ?

Un ancrage dans les "Computer Sciences"

- C'est la méthode de Sanger (1975) qui permet la détermination des séquences bases après bases
- A la fin des années 70, se lancent les grands projets de séquencage
- Tout de suite se pose la question du stockage, de l'accès et de l'organisation des données
- C'est aussi l'ère de la micro-informatique et de la popularisation des méthodes automatiques
- Mais une fois les séquences organisées, comment extraire de l'information pertinente de toute cette masse d'information ?

Premiers développements méthodologiques

- Une des idées fondatrices de l'analyse de séquences est de supposer que la comparaison de deux séquences peut se faire par alignement, étant donné le mécanisme d'évolution des séquences
- Les premiers développements mathématiques majeurs concernent la résolution algorithmique du problème d'alignement
- Une question qui se pose également est l'étude de la composition des génomes en bases, et en motifs
- Trouver les motifs d'une taille donnée est un problème d'algorithmique dont les repercusions pratiques sont considérables

Un petit problème de significativité

- Après avoir aligné deux séquences, que dire du score d'alignement ? Il est grand ? Petit ? **significativement** grand/petit ?
- On sait noter cet alignement, mais que faire de cette note ?
- Comment définir la significativité statistique des informations contenues dans les séquences ?
- Deux séquences s'alignent bien, mais par rapport à quoi ?
- Travaux de Karlin proposent une p-value pour le score d'alignement, utilisée dans BLAST

Les motifs

- On peut s'intéresser aux caractéristiques globales des compositions en base des génomes
- La motivation principale est que les structures observées ont un sens biologique
- Une question présente : y a-t-il des structures plus présentes que d'autres ?
- Plutôt que de s'intéresser aux structures globales, on peut se demander si certains mots sont évités dans un langage, ou au contraire utilisés de manière très fréquentes.
- Plusieurs contextes : motifs exceptionnels dans une séquence, motifs consensus dans plusieurs séquences

Les motifs dans les séquences d'ADN

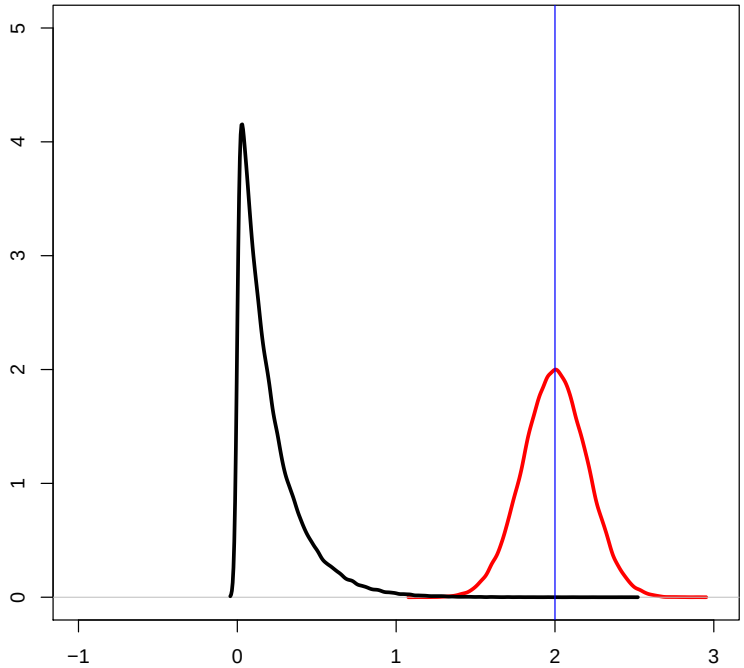
- Un exemple historique dans l'étude des génomes sont les sites de restrictions chez les bactéries
- Ce sont des motifs de 6 lettres (nucléotides) qui constituent un point de cassure de l'ADN dès qu'ils sont reconnus par une enzyme
- Ils sont "peu" présents dans les génomes bactériens
- D'autres motifs sont primordiaux et garantissent une stabilité du génome
- Exemple du motif chi : `GCTGGTGG` très présent chez E.Coli
- Très présent ? Mais par rapport à quoi ?

Vers une démarche de tests statistiques

- Pour dire si un mot est exceptionnel, on doit se donner une référence
- Un mot sera exceptionnel par rapport à un attendu qui sera un modèle de référence
- La significativité du motif n'aura de sens que par rapport au modèle de référence
- Par exemple: il est possible que `TGG` soit beaucoup plus fréquent que `TCG` parce que `TG` est plus fréquent que `TC`
- Il sera donc important que le modèle de référence prenne en compte la fréquence des sous mots qui composent un mot.

Exemple (simple) de mauvaise spécification du modèle

- On observe un phénomène distribué selon une certaine loi (distribution en noir)
- On veut savoir si $\{\texttt{Observer la valeur 2}\}$ est un événement exceptionnel
- On construit un modèle (rouge)
- Au vu du modèle rouge, la valeur 2 n'est pas du tout exceptionnelle !



$\mathbb{P}\{\text{la distribution du modèle rouge dépasse 2}\} \leq \alpha$

Notion de P-valeur et exceptionnalité

- La p-value est l'outil de base pour prendre une décision à l'issue d'un test
- Elle quantifie l'exceptionnalité de l'observation au vu du modèle de référence
- Dans le cas des motifs :

$$\mathbb{P}\{\text{Comptage observé d'un motif} \geq \text{Comptage attendu}\}$$

- Elle s'interprète comme la probabilité d'observer les données si le modèle de référence était vrai

Pourquoi un modèle aléatoire de séquences

- Un modèle ici sera l'ensemble de toutes les séquences possibles dont la séquence d'ADN observée ne constitue qu'une réalisation
- On cherche un modèle qui décrit globalement les caractéristique de la séquence observée (même composition en moyenne par exemple)
- L'objectif n'est pas forcément de modéliser au mieux les séquences, mais de construire un modèle aléatoire qui prenne en compte certaines informations
- On souhaite ensuite détecter des écarts au modèle, c'est à dire des événements exceptionnels compte tenu des contraintes déjà prises en compte

Pourquoi des résultats mathématiques ?

- Une pratique courante consiste à simuler le modèle de référence, pour calculer les p-values empiriquement
- On simule un modèle de référence et on compte le nombre de fois que le modèle est au dessus de la valeur observée par exemple
- Mais il faut aussi bien définir ce modèle ! Pour respecter la composition des séquences en bases par exemple
- Les contraintes combinatoires peuvent rendre cette stratégie impossible en pratique
- Les modèles de Markov sont naturels pour décrire une suite de variables aléatoires dépendantes
- Ils offrent un cadre probabiliste pour l'analyse de séquences
- Les résultats théoriques peuvent permettent d'éviter les stratégies combinatoires

Outline

1 Introduction

2 Préliminaires & Notations

3 Caractérisation statistique du modèle M0

4 Mise au point du Modèle Markovien d'ordre 1 (M1) et généralisation

5 Quel modèle choisir ?

6 Définition mathématique d'un motif dans une séquence

Notations pour les séquences

- On dispose d'une séquence de taille n, $s_n = (x_1, \ldots, x_n)$,
- On fait l'hypothèse que s_n est une réalisation d'une séquence aléatoire $S_n = (X_1, \ldots, X_n)$
- Chaque (X_i) modélise une lettre de la séquence et S_n est une succession de variables aléatoires
- On note $\mathcal{A}$ l'espace des possibles pour chaque lettre:

$$\forall i \in \{1, \ldots, n\},\ X_i \in \mathcal{A}$$

- $\mathcal{A} = \{\mathtt{A}, \mathtt{T}, \mathtt{G}, \mathtt{C}\}$ par exemple $S_6 = \mathtt{ACCTAG}$, $n = 6$
- On note également $|\mathcal{A}|$ la taille de l'alphabet (le cardinal de $\mathcal{A}$)

Loi d'une séquence

- La loi de la séquence S_n se note

$$\begin{aligned} \mathbb{P}\{S_n = s_n\} &= \mathbb{P}\{X_1 = x_1, \ldots, X_n = x_n\} \\ \mathbb{P}\{S_3 = \texttt{ATG}\} &= \mathbb{P}\{X_1 = \texttt{A}, X_2 = \texttt{T}, X_3 = \texttt{G}\} \end{aligned}$$

- C'est la loi jointe de toutes les lettres de la séquence
- Si les X_i sont indépendantes alors

$$\mathbb{P}\{S_n = s_n\} = \prod_{i=1}^{n} \mathbb{P}\{X_i = x_i\}$$

- Si les X_i ne sont pas indépendantes alors la loi de la séquence est déterminée par la loi jointe.
- Mais quel modèle considérer pour la loi d'apparition des lettres ?

Comment définir un modèle statistique ?

- On utilise un modèle statistique pour obtenir une approximation de ce que l'on observe
- En général tous les modèles sont faux, mais certains permettent de bien synthétiser le phénomène observé
- Un modèle statistique est consituté d'une famille de lois de probabilités sur un même espace
- En général ces lois de probabilités dépendent d'un paramètre θ qui appartient à un ensemble Θ
- On note alors:

$$\mathcal{M}_\theta = \{\mathbb{P}_\theta,\ \theta \in \Theta\}$$

Exemples de modèles statistiques

- le modèle binomial de paramètre θ:

$$\mathcal{M}_\theta = \{\theta \in \Theta = [0,1],\ \mathbb{P}\{X = 1\} = \theta\}$$

- le modèle gaussien de paramètres (μ, σ)

$$\begin{aligned} \mathcal{M}_\theta \quad = \quad & \left\{\theta = (\mu,\sigma) \in \Theta = \mathbb{R} \times \mathbb{R}^+, \right. \\ & \left. f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\right\} \end{aligned}$$

- Dans la suite, on va caractériser les modèles de Markov d'ordre m

Taille de modèle et qualité d'ajustement

- On définit la taille d'un modèle $|\mathcal{M}_\theta|$ par le nombre de paramètres (libres) qui le caractérisent
- Modèle de Bernoulli : $|\mathcal{M}_\theta| = 1$, modèle Gaussien: $|\mathcal{M}_\theta| = 2$.
- Plus le modèle sera "riche", plus il décrira les observations de manière précise
- Le nombre d'observations étant limité, un modèle riche aura comparativement peu d'observations pour estimer tous ses paramètres comparé à un modèle "plus simple"
- Il faudra prendre en compte cet élément quand on voudra comparer des modèles

Outline

1 Introduction

2 Préliminaires & Notations

3 Caractérisation statistique du modèle M0

4 Mise au point du Modèle Markovien d'ordre 1 (M1) et généralisation

5 Quel modèle choisir ?

6 Définition mathématique d'un motif dans une séquence

Présentation du modèle M0

- On suppose dans un premier temps l'indépendance des lettres dans la séquence

$$\mathbb{P}\{S_n = s_n\} = \prod_{i=1}^{n} \mathbb{P}\{X_i = x_i\}$$

- La loi de la séquence est déterminée par la probabilité d'apparition des 4 lettres:

$$\mathbb{P}\{X_i = \mathtt{A}\}, \mathbb{P}\{X_i = \mathtt{T}\}, \mathbb{P}\{X_i = \mathtt{G}\}, \mathbb{P}\{X_i = \mathtt{C}\}$$

- On utilise la notation:

$$\forall x \in \mathcal{A},\ \mu(x) = \mathbb{P}\{X_i = x\},\ \text{avec}\ \sum_{x \in \mathcal{A}} \mu(x) = 1$$

Caractérisation formelle du modèle M0

- Dans la suite, on notera:

$$\boldsymbol{\mu} = (\mu(x))_{x\in\mathcal{A}} = (\mu(\mathtt{A}), \mu(\mathtt{T}), \mu(\mathtt{G}), \mu(\mathtt{C}))$$

- Le modèle $\mathcal{M}_{\theta_0}$ est caractérisé par:

$$\mathcal{M}_{\theta_0} = \left\{ \theta_0 = \boldsymbol{\mu} \in \Theta_0 = [0,1]^{|\mathcal{A}|},\ \sum_{x\in\mathcal{A}} \mu(x) = 1,\ \mathbb{P}_\theta\{X = x\} = \mu(x) \right\}$$

- La taille du modèle est $|\mathcal{M}_{\theta_0}| = |\mathcal{A}| - 1$ à cause de la contrainte

Définition des variables indicatrices

- Dans la suite, on aura besoin de ces variables aléatoires
- Si ω est un événement, alors $\mathbb{I}\{\omega\} = 1$ si ω est vrai, et 0 sinon
- L'indicatrice est donc une variable aléatoire: on note $Y = \mathbb{I}\{\omega\}$

$$\mathbb{P}\{Y = 1\} = \mathbb{P}\{\mathbb{I}\{\omega\}\} = p$$

- Y est une variable de Bernoulli: $Y \sim \mathcal{B}(p)$
- Son espérance et sa variance sont donc:

$$\mathbb{E}(Y) = p, \ \mathbb{V}(Y) = p(1-p)$$

Loi d'une séquence sous le modèle M0

- $\mathbb{I}\{X_i = \mathtt{A}\} = 1$ si la $i$ème lettre de la séquence est un $\mathtt{A}$
- Le nombre de $\mathtt{A}$ dans la séquence est donc donné par:

$$N(\mathtt{A}) = \sum_{i=1}^{n} \mathbb{I}\{X_i = \mathtt{A}\}$$

- La loi d'une séquence sous le modèle M0 est donc:

$$\begin{aligned} \mathbb{P}\{S_n = s_n\} &= \prod_{i=1}^{n} \prod_{x \in \mathcal{A}} \mu(x)^{\mathbb{I}\{X_i = x\}} = \prod_{x \in \mathcal{A}} \mu(x)^{N(x)} \\ \mathbb{P}\{S_6 = \mathtt{ACCTAG}\} &= \mu(\mathtt{A})^2 \times \mu(\mathtt{T})^1 \times \mu(\mathtt{G})^1 \times \mu(\mathtt{C})^2 \end{aligned}$$

Notion de vraisemblance

- Le modèle $\mathcal{M}_\theta$ sert de lien entre les observations s_n et le paramètre θ
- Une fois observée, s_n donnera de l'information sur θ: c'est la démarche de l'inférence statistique
- On appelle vraisemblance du modèle $\mathcal{M}_\theta$ au vu de l'observation s la fonction de densité ayant servi à définir le modèle, mais du point de vue de $\mathcal{M}_\theta$

$$\mathcal{L}_s(\mathcal{M}_\theta) = \mathbb{P}_{\mathcal{M}_\theta}(s)$$

- Quand le modèle $\mathcal{M}_\theta$ est caratérisé par un paramètre θ on note aussi:

$$\mathcal{L}_s(\theta) = \mathbb{P}_\theta(s)$$

Pourquoi la log-vraisemblance ?

- $\mathcal{L}_s(\theta) = \mathbb{P}_\theta(s)$ est une probabilité donc dans $[0, 1]$
- Si on considère un $n-$échantillon (indépendance) alors la vraisemblance sera très "petite" (numériquement)

$$\mathcal{L}_s(\theta) = \prod_{i=1} \mathbb{P}_\theta(x_i)$$

- La transformation log est une fonction croissante: la maximisation de $\mathcal{L}_s(\theta)$ et de $\log \mathcal{L}_s(\theta)$ donnera la même solution

$$\frac{\partial \log \mathcal{L}_s(\theta)}{\partial \theta} = \frac{1}{\mathcal{L}_s(\theta)} \times \frac{\partial \mathcal{L}_s(\theta)}{\partial \theta}$$

- La transformation log permet de manipuler des sommes au lieu de produits. Pour un $n-$échantillon:

$$\log \mathcal{L}_s(\theta) = \sum_{i=1}^{n} \log \mathbb{P}_\theta(x_i)$$

L'estimateur du maximum de vraisemblance

- Si on considère le modèle $\mathcal{M}_\theta$, plusieurs valeurs de θ sont possibles
- Lorsqu'on dispose d'observations, on peut alors chercher le "meilleur modèle", celui dont la vraisemblance est la meilleure:

$$\widehat{\theta}(s) = \arg\max_{\theta \in \Theta} \{\log \mathcal{L}_s(\theta)\}$$

- La maximisation de la vraisemblance nécessite la résolution de l'équation:

$$\frac{\partial \log \mathcal{L}_s(\theta)}{\partial \theta} = 0$$

- L'estimateur du maximum de vraisemblance $\hat{\theta}(s)$ est la solution de cette équation

Retour au modèle M0

- La log-vraisemblance du modèle M0 est:

$$\log \mathcal{L}_s(\theta) = \log \mathbb{P}_\theta\{S_n = s_n\} = \sum_{x \in \mathcal{A}} N(x) \log \mu(x)$$

- On cherche à maximiser la vraisemblance par rapport aux paramètres $\mu(x)$

$$\frac{\partial \log \mathcal{L}_\theta(S)}{\partial \mu(x)} = 0, \ \sum_{x \in \mathcal{A}} \mu(x) = 1$$

- C'est une **maximisation sous contraintes** qui se résout à l'aide des multiplicateurs de Lagrange.

L'estimateur du MV pour le modèle M0

- La solution de la maximisation sous contrainte donne (pour la séquence s):

$$\forall x \in \mathcal{A},\ \widehat{\mu}_s(x) = \frac{N_s(x)}{n}$$

- C'est la **fréquence empirique** de chaque lettre dans la séquence
- Exemple pour $s_6 = \texttt{ACCTAG}$:

$$\widehat{\mu}_{s_6}(\texttt{A}) = 2/6;\ \widehat{\mu}_{s_6}(\texttt{T}) = 1/6;\ \widehat{\mu}_{s_6}(\texttt{G}) = 1/6;\ \widehat{\mu}_{s_6}(\texttt{C}) = 2/6.$$

Propriétés statistiques des estimateurs

- Lorsqu'on estime les paramètres $\mu(x)$, les résultats dépendent des observations

s_6^1	ACCTAA	$\widehat{\mu}_{s_6^1}(x)$
s_6^2	TCCTAG	$\widehat{\mu}_{s_6^2}(x)$
s_6^3	AACTAG	$\widehat{\mu}_{s_6^3}(x)$
$\vdots$	$\vdots$	$\vdots$
s_6^{10000}	ACTAAG	$\widehat{\mu}_{s_6^{10000}}(x)$

- **Un estimateur est une variable aléatoire**
- on distinguera l'estimateur $\widehat{\mu}_S(x)$ de sa réalisation $\widehat{\mu}_s(x)$
- On peut donc s'intéresser à sa loi, et à ses propriétés asymptotiques

Rappels sur l'espérance et la variance

- Si Y prend la valeur réelle y avec probability $p(y)$ alors l'espérance de Y s'écrit:

$$\mathbb{E}(Y) = \sum_y yp(y)$$

- l'espérance est un opérateur **linéaire**: $\mathbb{E}(Y + Z) = \mathbb{E}Y + \mathbb{E}Z$
- Si $Y \perp Z$, $\mathbb{E}(YZ) = \mathbb{E}Y \times \mathbb{E}Z$, sinon $\mathbb{E}(YZ) = \mathbb{E}Y \times \mathbb{E}Z - cov(Y, Z)$
- La variance de Y s'écrit:

$$\mathbb{V}(Y) = \mathbb{E}[Y - \mathbb{E}(Y)]^2 = \mathbb{E}[Y^2] - \mathbb{E}[Y]^2$$

- $\mathbb{V}(Y)$ mesure l'écart de Y à son espérance (sa dispersion)

Espérance du de l'EMV pour le modèle M0 - 1

- l'EMV pour les probabilités d'apparition des lettres dans le modèle M0: $\widehat{\mu}_S(x) = N_S(x)/n$
- La loi de l'EMV dépend de la loi du comptage des lettres

$$N_S(x) = \sum_{i=1}^{n} \mathbb{I}\{X_i = x\}$$

- L'espérance du comptage peut se calculer:

$$\mathbb{E}\left(N_S(x)\right) \quad = \quad \mathbb{E}\left(\sum_{i=1}^{n} \mathbb{I}\{X_i = x\}\right) = n\mu(x)$$

- L'estimateur du MV des probabilités d'apparition est un estimateur sans biais:

$$\mathbb{E}\left(\widehat{\mu}_S(x)\right) = \mu(x)$$

Variance du de l'EMV pour le modèle M0 - 1

- La variance de l'EMV nécessite le calcul du carré de l'espérance du comptage
- Rappel sur les carré de sommes:

$$\left(\sum_i a_i\right)^2 = \sum_i a_i^2 + \sum_{i \neq j} a_i a_j$$

- Pour le carré du comptage on a:

$$\begin{aligned} N_S^2(x) &= \left(\sum_{i=1}^{n} \mathbb{I}\{X_i = x\}\right)^2 \\ &= \sum_{i=1}^{n} \mathbb{I}\{X_i = x\} + \sum_{i=1}^{n} \sum_{j \neq i} \mathbb{I}\{X_i = x, X_j = x\} \end{aligned}$$

Variance du de l'EMV pour le modèle M0 - 2

- L'espérance du carré du comptage est donc:

$$\begin{aligned}
\mathbb{E}(N_S^2(x)) &= \sum_{i=1}^{n} \mathbb{E}\left(\mathbb{I}\{X_i = x\}\right) + \sum_{i=1}^{n} \sum_{j \neq i} \mathbb{E}\left(\mathbb{I}\{X_i = x, X_j = x\}\right) \\
&= n\mu(x) + n(n-1)\mu^2(x) \\
\mathbb{V}(N_S(x)) &= n\mu(x)(1-\mu(x))
\end{aligned}$$

- La variance de l'EMV pour les probabilités d'apparition des lettres:

$$\mathbb{V}\left(\widehat{\mu}_S(x)\right) = \frac{\mu(x)(1-\mu(x))}{n}$$

Propriétés statistiques de l'EMV pour le modèle M0 - 2

- L'inégalité de Bienaymé-Tchebychev peut s'utiliser pour quantifier la concentration d'une variable aléatoire autour de son espérance

$$\mathbb{P}\{|Y - \mathbb{E}Y| \geq \varepsilon\} \leq \frac{\mathbb{V}Y}{\varepsilon^2}$$

- A l'aide de l'inégalité de Bienaymé-Tchebychev on montre la convergence de l'estimateur vers la vraie valeur du paramètre

$$\mathbb{P}\{|\widehat{\mu}(x) - \mu(x)| \geq \varepsilon\} \quad \leq \quad \frac{\mu(x)(1-\mu(x))}{n\varepsilon^2} \leq \frac{1}{4n\varepsilon^2}$$

- Déterminer la loi du comptage est plus difficile

Outline

1 Introduction

2 Préliminaires & Notations

3 Caractérisation statistique du modèle M0

4 Mise au point du Modèle Markovien d'ordre 1 (M1) et généralisation

5 Quel modèle choisir ?

6 Définition mathématique d'un motif dans une séquence

Passage au modèle de Markov d'ordre 1: M1

- La fréquence d'apparition des dinucléotides suggère que l'hypothèse d'indépendance entre les bases est trop simplificatrice

	A	C	G	T	somme
A	1112	561	1024	713	3410
C	795	413	95	470	1773
G	820	457	661	432	2370
T	684	342	590	548	2164

- $N(\mathtt{AG}) = 561$ désigne le comptage du dinucléotide AG,
- $N(\mathtt{A}+) = 3410$ désigne le comptage des dinucléotide qui commencent par un A.

	A	C	G	T
A	0.33	0.16	0.30	0.21
C	0.45	0.23	0.05	0.27
G	0.35	0.19	0.28	0.18
T	0.32	0.16	0.27	0.25

Présentation du modèle M1 - 1

- Le modèle de Markov d'ordre 1 introduit une dépendance des positions à l'ordre 1 (mémoire à distance 1)
- $(X_1, \ldots, X_n)$ est une chaîne de Markov d'ordre 1 ssi:

$$\mathbb{P}\{X_{k+1} = x_{k+1} | X_k = x_k, \ldots, X_1 = x_1\} = \mathbb{P}\{X_{k+1} = x_{k+1} | X_k = x_k\}$$

- Ce modèle suppose que les variables $(X_{k-1}, \ldots, X_1)$ ne donnent pas d'information sur la loi de X_{k+1}
- La loi d'une séquence de taille n sous le modèle M1 s'écrit:

$$\begin{aligned} \mathbb{P}\{S_n = s_n\} &= \mathbb{P}\{X_1 = x_1, \ldots, X_n = x_n\} \\ &= \mathbb{P}\{X_1 = x_1\} \times \mathbb{P}\{X_2 = x_2 | X_1 = x_1\} \ldots \\ &\times \mathbb{P}\{X_n = x_n | X_{n-1} = x_{n-1}\} \end{aligned}$$

Présentation du modèle M1 - 2

- La loi de la chaîne de Markov est entièrement déterminée par:
  - La loi d'émission de la première lettre: $\mathbb{P}\{X_1 = x\} = \mu(x)$
  - Les probabilités conditionnelles $\mathbb{P}\{X_i = y | X_{i-1} = x\} = \pi(x, y)$
- On fait en général l'hypothèse de stationarité : la loi des X_i ne dépend pas de l'endroit où l'on se trouve dans la séquence
- Exemple $\mathbb{P}\{S = \texttt{ACCACC}\}$

$$
\begin{aligned}
&= \mu_1(\texttt{A}) \times \pi_2(\texttt{A},\texttt{C}) \times \pi_3(\texttt{C},\texttt{C}) \times \pi_4(\texttt{C},\texttt{A}) \times \pi_5(\texttt{A},\texttt{C}) \times \pi_6(\texttt{C},\texttt{C}) \\
&= \mu(\texttt{A}) \times \pi^2(\texttt{A},\texttt{C}) \times \pi^2(\texttt{C},\texttt{C}) \times \pi(\texttt{C},\texttt{A}) \text{ si stationaire}
\end{aligned}
$$

- On peut aussi écrire $\mathcal{D}(X_i, \ldots, X_{i+h}) = \mathcal{D}(X_{i+\ell}, \ldots, X_{i+h+\ell})$

Propriétés remarquables des matrices de transition

- Rappel sur les probabilités conditionnelles

$$\sum_{y \in \mathcal{A}} \pi(x, y) = 1$$

- La matrice π est une matrice stochastique

$$\boldsymbol{\pi} = \begin{array}{c|cccc} & \mathtt{A} & \mathtt{C} & \mathtt{G} & \mathtt{T} \\ \hline \mathtt{A} & \pi(\mathtt{A},\mathtt{A}) & \pi(\mathtt{A},\mathtt{C}) & \pi(\mathtt{A},\mathtt{G}) & \pi(\mathtt{A},\mathtt{T}) \\ \mathtt{C} & \pi(\mathtt{C},\mathtt{A}) & \pi(\mathtt{C},\mathtt{C}) & \pi(\mathtt{C},\mathtt{G}) & \pi(\mathtt{C},\mathtt{T}) \\ \mathtt{G} & \pi(\mathtt{G},\mathtt{A}) & \pi(\mathtt{G},\mathtt{C}) & \pi(\mathtt{G},\mathtt{G}) & \pi(\mathtt{G},\mathtt{T}) \\ \mathtt{T} & \pi(\mathtt{T},\mathtt{A}) & \pi(\mathtt{T},\mathtt{C}) & \pi(\mathtt{T},\mathtt{G}) & \pi(\mathtt{T},\mathtt{T}) \end{array}$$

Notion de stationarité - 1

- Pour déduire la loi de X_{i+2} à partir de la loi de X_i

$$\begin{aligned}\mathbb{P}\{X_{i+1} = x_{i+1} | X_i = x_i\} &= \pi(x_i, x_{i+1}) \\ \mathbb{P}\{X_{i+2} = x_{i+2} | X_i = x_i\} &= \sum_{x_{i+1} \in \mathcal{A}} \pi(x_i, x_{i+1}) \times \pi(x_{i+1}, x_{i+2})\end{aligned}$$

- On reconnait la formule d'un produit matriciel entre la ligne x_i et la colonne x_{i+2} de $\boldsymbol{\pi}$
- C'est le terme (x_i, x_{i+2}) de la matrice $\boldsymbol{\pi} \times \boldsymbol{\pi} = \boldsymbol{\pi}^2$

$$\mathbb{P}\{X_{i+1} = \bullet\} = \mathbb{P}\{X_i = \bullet\} \times \boldsymbol{\pi}$$

- Par recurrence on peut montrer que la transition en k pas dans les modèles M1 est donnée par l'élément de $\boldsymbol{\pi}^k$

$$\mathbb{P}\{X_{i+1} = \bullet\} = \mathbb{P}\{X_1 = \bullet\} \times \boldsymbol{\pi}^i$$

Notion de stationarité - 2

- Si la loi stationnaire existe, elle doit vérifier la relation suivante:

$$\begin{aligned} \mu(x_{i+1}) &= \sum_{x_i \in \mathcal{A}} \mu(x_i) \times \pi(x_i, x_{i+1}) \\ \boldsymbol{\mu} &= \boldsymbol{\mu} \times \boldsymbol{\pi} \end{aligned}$$

- Donc si on suppose que X_1 est de loi $\boldsymbol{\mu}$ alors tous les (X_i) seront de même loi (sans être indépendants)
- Sous certaines conditions (ergodicité) on sait que cette distribution stationnaire est unique

Caratérisation formelle du modèle M1

- Le modèle $\mathcal{M}_{\theta_1}$ est donc caractérisé par:

$$\theta_1 = \begin{cases} \boldsymbol{\mu} = (\mu(x))_x \in [0,1], x \in \mathcal{A}, , \sum_{x\in\mathcal{A}} \mu(x) = 1 \\ \boldsymbol{\pi} = (\pi(x,y))_{x,y} \in [0,1], x \in \mathcal{A}, y \in \mathcal{A}, \sum_{y\in\mathcal{A}} \pi(x,y) = 1 \end{cases}$$

$$\Theta_1 = [0,1]^{\mathcal{A}} \times [0,1]^{\mathcal{A}\times\mathcal{A}}$$

- La taille du modèle $\mathcal{M}_{\theta_1}$ est donc $|\mathcal{M}_{\theta_1}| = 4 - 1 + 16 - 4$

Log-Vraisemblance et EMV du modèle M1

- La log-vraisemblance des paramètres $\boldsymbol{\mu}, \boldsymbol{\pi}$ pour une séquence S:

$$\begin{aligned}
\log \mathcal{L}_S(\boldsymbol{\mu}, \boldsymbol{\pi}) &= \log \mu(x_1) \\
&+ \sum_{i=2}^{n} \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{A}} \mathbb{I}\{X_i = x, X_{i+1} = y\} \log \pi(x, y) \\
&= \log \mu(x_1) + \sum_{x \in \mathcal{A}} \sum_{y \in \mathcal{A}} N(x, y) \log \pi(x, y)
\end{aligned}$$

- Sa maximisation se fait également sous contrainte
- Les EMV sont:

$$\begin{aligned}
\widehat{\mu}(x) &= N(x)/n \\
\widehat{\pi}(x, y) &= N(x, y)/N(x+)
\end{aligned}$$

Généralisation à l'ordre m

- Une séquence S_n est une chaîne de Markov d'ordre $m \geq 1$ avec une distribution initiales $\mu_m(x)$ et matrice de transition $\boldsymbol{\pi}$
- $\forall x_i \in \mathcal{A},\ \mathbb{P}\{X_1 = x_1, \ldots, X_m = x_m\} = \mu(x_1, \ldots, x_m)$
- $\mathbb{P}\{X_i | X_{i-m} = x_{i-m}, \ldots, X_{i-1} = x_{i-1}\} = \pi(x_{i-m}, \ldots, x_{i-1}, x_i)$
- l'EMV de $\boldsymbol{\mu}$ et de $\boldsymbol{\pi}$ sont:

$$\begin{aligned} \widehat{\mu}(x_1, \ldots, x_{m-1}, x_m) &= \frac{N(x_1, \ldots, x_{m-1}, x_m)}{n - m + 1} \\ \widehat{\pi}(x_1, \ldots, x_m, x_{m+1}) &= \frac{N(x_1, \ldots, x_m, x_{m+1})}{N(x_1, \ldots, x_m, +)} \end{aligned}$$

- La taille du modèle $\mathcal{M}_{\theta^m}$ est $(|\mathcal{A}| - 1) \times |\mathcal{A}|^m$

Outline

1 Introduction

2 Préliminaires & Notations

3 Caractérisation statistique du modèle M0

4 Mise au point du Modèle Markovien d'ordre 1 (M1) et généralisation

5 Quel modèle choisir ?

6 Définition mathématique d'un motif dans une séquence

Un modèle, pour quoi faire ?

- Si on cherche le modèle qui explique au mieux les données, alors l'idée est de mettre en compétition plusieurs modèles et de choisir le "meilleur" au sens d'un certain critère
- Si on cherche à détecter des structures exceptionnelles par rapport au modèle de référence, il ne faut pas que ces structures soient prévues par le modèle de référence
- Analogie : on met dans le modèle tout ce qu'on sait et on regarde ce qu'il reste !

Quel ordre pour quel motif ?

- Dans le modèle de Markov d'ordre 1, $\widehat{\pi}(x, y) = N(x, y)/N(x+)$
- C'est la probabilité qu'un x soit suivi d'un y qui est estimée par la proportion de x suivis d'un y
- Un modèle d'ordre m prend en compte la composition ("s'adapte à") des mots de taille 1 à $m + 1$
- Un motif de taille h peut donc être étudié dans les modèles d'ordre maximal $h - 2$.
- Le modèle d'ordre $h - 2$ prend en compte la composition de la séquence en mots de longueur $h - 1$

Principe de la selection de modèles

- Pour comparer des modèles, on cherche à les "noter"
- La vraisemblance d'un modèle permet de quantifier la qualité d'ajustement d'un modèle aux données
- Mais cette qualité dépend du nombre de paramètres du modèle ! La vraisemblance augmente avec la dimension du modèle
- Pour une comparaison "équitable" il faut comparer des modèles en "pénalisant" leur dimension
- On utilise des vraisemblances pénalisées:

$$\log \widetilde{\mathcal{L}}_S(\mathcal{M}_\theta) = \log \mathcal{L}_S(\mathcal{M}_\theta) - \beta \text{pen}\left(|\mathcal{M}_\theta|\right)$$

Différents critères de selection de modèles

- Les critères diffèrent dans leurs objectifs et dans leur définition de la pénalité

$$\begin{aligned} \text{AIC} &= \log \mathcal{L}_S(\mathcal{M}_\theta) - |\mathcal{M}_\theta|/2 \\ \text{BIC} &= \log \mathcal{L}_S(\mathcal{M}_\theta) - \frac{|\mathcal{M}_\theta|}{2} \times \log(n) \end{aligned}$$

	M0	M1	M2	M3	M4	M5	M6
HIV							
AIC	26.37	25.80	**25.68**	25.70	26.10	28.03	40.00
BIC	26.39	**25.89**	26.03	27.08	31.62	50.10	128.26
E. Coli							
AIC	12861	12743	12626	12546	12497	12456	**12435**
BIC	12862	12743	12627	12548	12508	**12497**	12599

Outline

1 Introduction

2 Préliminaires & Notations

3 Caractérisation statistique du modèle M0

4 Mise au point du Modèle Markovien d'ordre 1 (M1) et généralisation

5 Quel modèle choisir ?

6 Définition mathématique d'un motif dans une séquence

Notations

- Un motif est défini comme une sous-séquence **w** d'une séquence **S**
- C'est une séquence connue de longueur h telle que:

$$\mathbf{W} = (W_1, \ldots, W_h) \in \mathcal{A}^h$$

- On définit les occurrences pour savoir combien a-t-on de motifs (et où ils se trouvent)
- La position d'un motif est définie par la position de sa première lettre W_1
- C'est une variable aléatoire !

Indicatrice d'occurrence

- On note $Y_i(\mathbf{w})$ la variable indicatrice qui vaut 1 si $\mathbf{w}$ est à la position i dans la séquence $\mathbf{S}$

$$Y_i(\mathbf{w}) = \begin{cases} 1 \text{ si } (X_i, \ldots, X_{i+h-1}) = (W_1, \ldots, W_h) \\ 0 \text{ sinon} \end{cases}$$

- La loi de $Y_i(\mathbf{w})$ est une loi de Bernoulli

$$\mathbb{P}\{Y_i(\mathbf{w}) = 1\} = \mathbb{P}\{X_i = W_1, \ldots, X_{i+h-1} = W_h\}$$

- On note cette probabilité $\mu(\mathbf{w})$, la probabilité d'occurrence du mot $\mathbf{w}$

Probabilité d'occurence d'un mot

- Le calcul de cette probabilité dépend du modèle de référence
- On choisit souvent le modèle M1

$$\begin{aligned} \mu(\mathbf{w}) &= \mathbb{P}\{X_i = W_1, \ldots, X_{i+h-1} = W_h\} \\ &= \mu(W_1) \times \pi(W_1, W_2) \times \pi(W_{h-1}, W_h) \end{aligned}$$

- Etant donné que le modèle de Markov est stationnaire, cette probabilité ne dépend pas de l'endroit où l'on se place dans la séquence
- L'esperance et la variance de l'indicatrice sont:

$$\begin{aligned} \mathbb{E}Y_i(\mathbf{w}) &= \mu(\mathbf{w}) \\ \mathbb{V}Y_i(\mathbf{w}) &= \mu(\mathbf{w})(1 - \mu(\mathbf{w})) \end{aligned}$$

Comptage d'un motif

- Le nombre d'occurences d'un motif est défini à partir des indicatrices d'occurrence:

$$N(\mathbf{w}) = \sum_{i=1}^{n-h+1} Y_i(\mathbf{w})$$

- L'espérance du comptage se déduit

$$\mathbb{E}(N(\mathbf{w})) = (n - h + 1)\mu(\mathbf{w})$$

- Mais la variance du comptage dépend du recouvrement des occurrences ! les $Y_i(\mathbf{w})$ ne sont pas indépendantes

Loi exacte ou approximation ?

- Déterminer la loi exacte du comptage $N(\mathbf{w})$ signifie calculer pour toutes les valeurs k de $N(\mathbf{w})$

$$\mathbb{P}\{N(\mathbf{w}) = k\}$$

- Des développements existent pour calculer ces probabilités par reccurrence mais leur calcul est coùteux
- Une alternative est de considérer une loi approchée
- Une manière d'approcher une loi est de s'intéresser à la loi **asymptotique**,quand la longueur de la séquence tend vers l'infini

Rappel sur le Théorème central limite

- C'est un théorème à la base de beaucoup de demonstrations / approximations en statistique
- Si les X_k sont des variables aléatoires réelles i.i.d. d'espérance $\mathbb{E}(X)$ et de variance $\mathbb{V}(X)$ alors

$$\sqrt{n}\frac{\bar{X} - \mathbb{E}(X)}{\sqrt{\mathbb{V}(X)}} \underset{n\to\infty}{\sim} \mathcal{N}(0,1)$$

- Dans le cas des motifs, $N(\mathbf{w})$ est bien une somme de variables aléatoires
- Quelle est la qualité d'approximation de la loi asymptotique ?

Approximation Gaussienne pour les comptages de mots

- En pratique, on considère une estimation de l'esperance du comptage:

$$\begin{aligned}\widehat{\mathbb{E}}(N(\mathbf{w})) &= (n-h+1)\widehat{\mu}(\mathbf{w}) \\ &= (n-h+1)\frac{\prod_{j=1}^{h-1} N(W_j, W_{j+1})}{\prod_{j=2}^{h-1} N(W_j)}\end{aligned}$$

- On construit ensuite le score d'exceptionalité:

$$Z(\mathbf{w}) = \frac{N_{\text{obs}}(\mathbf{w}) - \widehat{\mathbb{E}}(N(\mathbf{w}))}{\sqrt{\widehat{\mathbb{V}}(N(\mathbf{w}))}}$$

- $\widehat{\mathbb{V}}(N(\mathbf{w}))$ est difficile à calculer parce qu'elle dépend du recouvrement des motifs

Approximation de Poisson pour les comptages

- D'autres approximations asymptotiques ont été développées
- Si Y_i sont des variables aléatoires iid de loi p_i alors $\lambda = \sum_i p_i$

$$\sum_{i=1}^{n} Y_i \sim \mathcal{P}(\lambda)$$

- Mais dans le cas des motifs, les $Y_i(\mathbf{w})$ ne sont pas indépendantes !
- La méthode de Chen-Stein permet de mesurer l'erreur commise lorsque l'on approche une somme de variables aléatoires de Bernoulli dépendantes par une loi de Poisson

Domaines de validité des approximations

- Il n'existe pas d'approximation meilleure qu'une autre sur tous les critères
- Pour étudier les qualités d'approximation, on peut comparer les p-values obtenues par la loi exacte (quand on peut la calculer) aux p-values calculées à partir des approximations
- L'approximation Gaussienne est valide pour les **mots courts et fréquents**
- L'approximation de Poisson composée donne de bons résultats également même pour les mots rares
- Idée: utiliser plusieurs approximations pour vérifier la concordance des résultats