Détection de motifs exceptionnels dans l'ADN

Modélisation par des chaînes de Markov

Pierre Boyeau, projet encadré par Jean-François Delmas

École des Ponts ParisTech

Introduction

Aujourd'hui, on dispose de génomes complets d'espèces dont on ne connait presque rien.
Il est donc important de fournir aux biologistes des méthodes **rapides**, **automatiques** et **pertinentes** d'aide à l'analyse de génomes.
Les mots **exceptionnels**, c'est-à-dire très ou au contraire très peu exprimés dans l'ADN, jouent un rôle souvent crucial en biologie, comme le sont par exemple les sites **Chi** ou les **sites de restriction**.

Comment détecter des motifs exceptionnels dans l'ADN ?

1 Modélisation probabiliste du problème

Notations

On considère une **séquence** $S = x_1 x_2 \dots x_N$, $\forall i, x_i \in \{A, T, C, G\}$.
On appelle **mot**, ou **motif**, tout $w = w_1 w_2 \dots w_k \subseteq S, \forall i, w_i \in \{A, T, C, G\}$
$(X_n)_{n\in\{1\dots N\}}$: nième nucléïde de la séquence
$N(w)$ est le nombre d'occurrences de w dans la séquence S

Hypothèses fondamentales

$(X_n, n \in \{1, \dots, N\})$: chaîne de Markov d'ordre m et de matrice de transition P:

$$\mathbb{P}(X_n = a | X_1 = x_1, \dots, X_{n-1} = x_{n-1}) = \mathbb{P}(X_n = a | X_{n-m} = x_{n-m}, \dots, X_{n-1} = x_{n-1}) \quad (1)$$
$$= P((x_{n-m} \dots n_{n-1}), a) \quad (2)$$

Toute information utile pour la prédiction d'un nucléïde ne dépend du passé que par l'enchaînement des m nucléïdes précédentes.
En outre, l'hypothèse (2) fait l'hypothèse que l'ADN est homogène, et qu'il n'existe pas de zones qui portent un sens spécifique.

2 Résolution du problème

Pour un mot w:

- $N_{obs}(w)$ correspond au nombre d'occurrences de w dans notre séquence.
- $N(w)$ est la *variable aléatoire* du nombre d'occurrences de w *prédit par le modèle*

On s'intéresse à l'événement

$$\begin{cases} N_{obs}(w) \geq \mathbb{E}[N(w)] \text{ pour un mot anormalement présent} \\ N_{obs}(w) \leq \mathbb{E}[N(w)] \text{ pour un mot rare} \end{cases}$$

Détermination de la loi de comptage

Estimations gaussiennes

Le Théorème Central Limite appliqué aux chaînes de Markov assure [1] que la loi de comptage vérifie:

$$p_{\text{score}}(w) = \frac{N_{obs}(w) - \mathbb{E}_m(N(w))}{\sigma(w)} \to \mathcal{N}(0,1) \text{ quand } N \to \infty$$

Pour estimer les p_{score} de motifs, il est donc important de pouvoir estimer $\mathbb{E}_m[N(w)]$ et $\sigma(w)$.

Première approche

- théorème ergodique: $\frac{N(w)}{N} \to \pi(w) = \pi(w_1 \dots w_{k-1}) P(w_{k-1}, w_k) = \frac{\pi(w_1 \dots w_{k-1})\pi(w_{k-1} w_k)}{\pi(w_{k-1})}$

On peut trouver un estimateur de $\mathbb{E}_m[N(w)]$:

$$\widehat{\mathbb{E}}_1[N(w)] = \frac{N(w_1 \dots w_{k-1}) N(w_{k-1} w_k)}{N(w_{k-1})}$$

- On montre que sous ce modèle, on peut choisir l'estimateur suivant pour la variance :

$$\widehat{\sigma}_1^2(N(w)) = \frac{N(w)}{N}(1 - \frac{N(w_1 \dots w_{k-1}}{N(w_{k-1})})(1 - \frac{N(w_{k-1} w_k)}{N(w_{k-1}})$$

Approche par Estimateur de Maximum de Vraisemblance

Idée: Trouver des estimateurs dans le cadre plus général d'une chaîne de Markov d'ordre m. Pour cela, on calcule les EMV pour le cas $m = 1$, et on peut facilement généraliser dans le cas $k = m - 2$ par changement d'alphabet.
$\Rightarrow$Obtention d'estimateurs $\tilde{\mathbb{E}}_m[N(w)]$ et $\tilde{\sigma}_m^2(N(w))$

3 Validation du modèle

Validité de nos choix d'estimateurs

- Simuler une séquence à partir d'une chaîne de Markov connue
- Calculer l'estimation d'espérance de la séquence
- La comparer à l'espérance réelle de la loi de comptage, que l'on connaît vu qu'on connait la loi de (X_n)

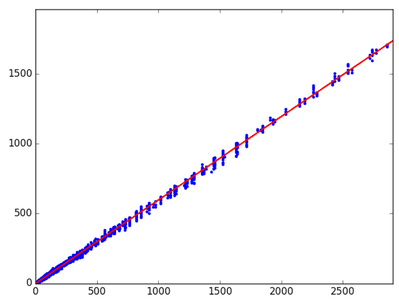
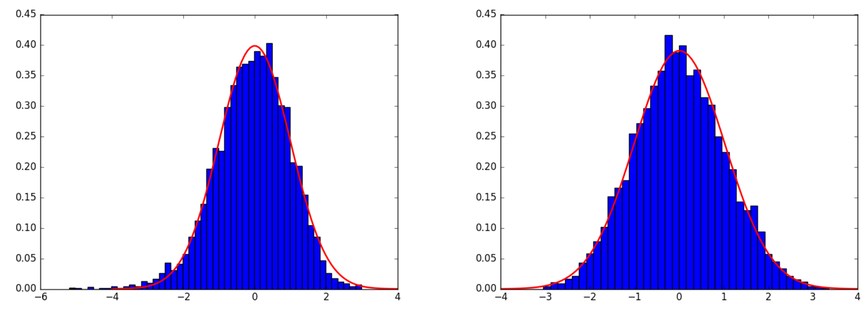


Figure 1: Couples (Espérance Réelle, Espérance estimée) par la 2nde approche pour une séquence simulée de longueur 10^6 et $k = 6, m = 4$

Adéquation à la loi Normale centrée réduite



(a) 1ère méthode **(b)** 2ème méthode

Figure 2: Comparaison de p_{score} pour une séquence simulée de taille 10^6 et des mots de taille $k = 6$

4 Applications

p_{score} pour un génôme réel

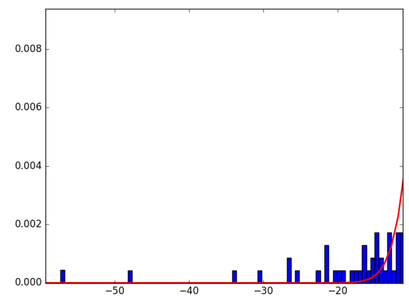


Figure 3: p_{score} **en queue de distribution** pour $k = 6, m = 4$ pour le génome E.Coli

Comment interpréter ce résultat?

Détection de sites de restriction pour E.Coli

Pour le génome d'E.Coli, l'estimation de p_{score} par EMV donne des résultats pertinents.

Mot	Rareté (rang)	Enzymes de restriction
GGCGCC, CCGCGG	1, 7	Eco78I, Eco29kI
GCCGGC, CGGCCG	2, 5	Eco52I, Eco56I
AGCGCT	4	Eco47III
TCCGGA	6	Eco147I
CACGTG	11	Eco72I
GAGCTC	13	Eco53kI, EcoICRI

Mot	Rareté (rang)	Enzymes de restriction
CTATAG, GATATC	1043, 3799	EcoRV

Conclusion

- Apport personnel du projet
- Approfondissements intéressants:
  - 1 mois: approximations plus fines (mots de taille importante, chevauchements de mots)
  - 6 mois: théorie des automates et calculs exacts des p_{score}

Remerciements

Chaleureux remerciements à Jean François Delmas pour son encadrement et son aiguillage, ainsi qu'à Florence Rieu pour son aide pour ma recherche bibliographique.

References

[1] S. Robin, F. Rodolphe, S. Schbath *ADN, mots et modèles* 2003.

[2] J.F. Delmas, B. Jourdain *Modèles aléatoires* Mathematiques et Applications 57, 2007.

[3] G. Nuel *Significance Score of Motifs in Biological Sequences* Bioinformatics: Trends and Methodologies Intech 2011; 978-53. Relations.

[4] Wikipedia.org *List of restriction enzyme cutting sites*