Etude des parties codantes de l'ADN en fonction du sens lecture

Guillaume Sicot[1], Ramesh Pyndiah[1]

[1]Dept Signal and Communication - ENST-Bretagne - TAMCIC (CNRS 2658)
Technopôle Brest-Iroise - CS 83818 - 29238 Brest Cedex 3
{guillaume.sicot} {ramesh.pyndiah}@enst-bretagne.fr

Résumé – Dans les séquences ADN, certaines portions dites codantes contiennent l'information nécessaire à l'élaboration des protéines. Le code génétique utilise des règles de correspondance entre les séquences de nucléotides et les acides aminés, constituants élémentaires de protéines. Ce code génétique est dit dégénéré, c'est-à-dire qu'un même acide aminé peut-être représenté par un ensemble de groupe de nucléotides ou codons. Une manière d'étudier les séquences ADN est de s'intéresser aux fréquences d'apparition des mots. Dans cet article nous étudions les fréquences d'apparition des mots suivant les différents sens de lecture de l'ADN et suivant le brin. Nous montrons en particulier que ces fréquences d'apparition des mots dans la séquence ADN présentent des propriétés qui ne peuvent être reproduite à l'aide de modèles statistiques simple.

Abstract – The DNA is the support of the information neccessary to elaborate proteins. Rules that associate nucleotids sequences to amino-acids sequences, constitute the genetic code. This code is said to be degenerative, that means that several nucleotids sets or codons can represent the same amino-acid. A way to study nucleotids sequences consists in the study of the set of frequency of appearence of words in nucleotids sequences. In this document we will study frequencies of appearence of words in relation to the reading direction and the strand of the DNA. We show that some features appear by considering the reading direction that cannot be reproduced with simple statistical model.

1 Introduction

L'acide désoxyribonucléique (ADN), support de l'information génétique, se présente sous la forme d'une double hélice. Chaque brin de la double hélice est une suite de bases nucléotiques au nombre de quatre : *adénine* (A), *thymine* (T), *cytosine* (C), *guanine* (G). Ces deux brins d'ADN sont dits complémentaires, c'est-à-dire qu'un A (resp. un C) sur un brin est associé à un T (resp. G) sur l'autre brin. Chaque brin est orienté, cette orientation est définie à partir de la composition chimique de l'ADN : un côté du brin est noté $3'$, l'autre côté est noté $5'$. De plus si un brin est orienté dans le sens $3' - 5'$ alors l'autre brin sera orienté dans le sens inverse, *i.e.* $5' - 3'$. Ces rappels sur la molécule d'ADN sont représentés sur le figure 1.

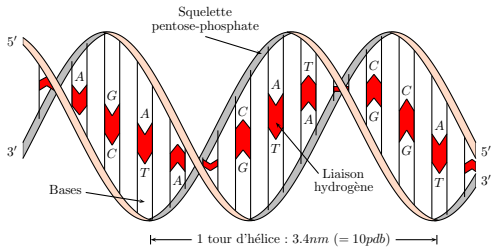


Fig. 1 – Schéma décrivant la double hélice d'ADN

L'information génétique contient l'ensemble de l'information nécessaire à la cellule pour créer les protéines, en d'autres termes elle est utilisée pour construire les séquences d'acides aminés (constituant élémentaire de la protéine). On distingue deux types de séquences, les séquences dites codantes, c'est-à-dire que ces séquences seront traduites en protéines et les séquences non-codantes. Le passage de la séquence ADN codante en protéine se réalise en deux étapes : la transcription qui copie en ARN messager (ARNm) la portion codante de l'ADN puis le traduction qui élabore la protéine à partir de l'ARNm. Les règles de passage entre la séquence de bases nucléotidiques formant l'ARNm et la séquence d'acides aminés formant la protéine constituent le code génétique.

Nous présenterons dans un premier temps le code génétique. Nous montrerons que le code génétique permet de transmettre plus d'information que celle requise pour la transmission de l'information génétique. Ensuite nous présenterons les notions utilisées dans notre étude qui se focalise sur la comparaison des séquences ADN suivant le sens de lecture et le brin. Enfin nous comparerons les résultats obtenus sur des séquences réelles avec des séquences théoriques puis nous terminerons par tirer des conclusions sur ces comparaisons.

2 Le code génétique

Le code génétique permet de représenter les 20 acides aminés plus le signal *Stop* ; soit 21 éléments par un ensemble de mots de code. Ces mots de code sont constitués de trois bases nucléotidiques A, T, C, G encore appelés codons. Le code génétique est dit universel, dégénéré et non-

chevauchant. Il est dit dégénéré car le nombre de codons ou mots de code (64) est supérieur aux 21 éléments à coder et donc un même acide aminé peut être associé à plusieurs codons. Ainsi les codons permettent de transmettre potentiellement plus d'information par rapport aux 21 éléments de la source. Il est intéressant de constater que l'alphabet quaternaire maximise le nombre de mots de codes disponibles pour des alphabets de taille $1 < S < 7$ sous la contrainte de minimiser la longueur n des mots de code (voir Tableau 1).

Taille de l'alphabet (S)	2	3	**4**	5	6	7
n tel que $S^{n-1} < 21 \leq S^n$	5	3	**3**	2	2	2
Nombre de mots de code	32	27	**64**	25	36	49

TAB. 1 – Redondance pour différentes tailles d'alphabet.

Ainsi le code génétique permet de transmettre plus d'information que nécessaire pour les séquences d'acides aminés. En effet plusieurs codons, encore appelés codons synonymes, peuvent représenter un même acide aminé et l'usage des codons synonymes n'est pas uniforme et varie suivant les organismes ([2]). La distribution des fréquences d'apparition des codons synonymes est d'ailleurs relativement caractéristiques de chaque organisme ([3]).

Un moyen de caractériser le codage associé à une espèce est de s'intéresser à sa signature génomique. Cette notion peut être définie de différentes manières ([4], [5], [6]), néanmoins elles reposent toutes sur les fréquences d'apparition de mots de taille K dans la séquence ADN. Cette représentation est d'ailleurs plus à même de caractériser chaque organisme ([3]). La définition de la signature génomique ne distingue pas *a priori* les parties codantes et non-codantes de l'ADN. Les résultats obtenues sont similaires si l'on ne tient compte que des parties codantes ([3]).

C'est cette notion de signature génomique que nous allons utiliser par la suite. En effet nous allons étudier les fréquences d'apparition des mots de taille K suivant les différents sens de lecture des deux brins constituant la double hélice de l'ADN.

3 Signature génomique & sens de lecture

Comme nous l'avons indiqué précédemment, nous allons nous intéresser plus particulièrement aux fréquences d'apparition des mots. Nous allons les comparer suivant les quatre sens de lecture possibles. La figure 2 définit les différents sens de lecture utilisées ainsi que leur notation. Il est important de noter que le sens 1 est le sens de lecture biologique d'une partie codante (de $5'$ vers $3'$). Dans la suite de ce document, nous noterons S_i, les séquences codantes prises dans le sens i, avec $i \in \{1, \ldots, 4\}$.

Nous allons étudier la différence entre les fréquences d'apparition des mots suivant les quatre sens de lecture. Pour cela, considérons deux séquences S_i et S_j de bases

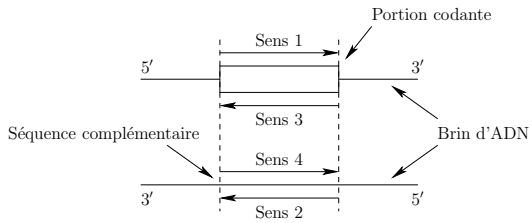


FIG. 2 – Définition des notations des sens de lecture utilisés

nucléotidiques correspondant à deux sens de lecture et soit $\mathcal{M}_K$, l'ensemble des mots de taille K, construit sur l'alphabet, $\mathcal{A}$ des bases nucléotidiques, donc $\mathcal{A} = \{A, T, C, G\}$. Enfin pour étudier la différence de distribution des fréquences d'apparition des mots entre les séquences S_1 et S_2, nous définissons trois fonctions $d_1(S_1, S_2)$, $d_2(S_1, S_2)$ et $d_3(S_1, S_2)$ comme ci-dessous :

$$d_1^K(S_i, S_j) = \frac{1}{\text{card}(\mathcal{M}_K)} \sum_{m \in \mathcal{M}_K} \log^2 \left(\frac{f_{S_i}(m)}{f_{S_j}(m)} \right) \quad (1)$$

$$d_2^K(S_i, S_j) = \frac{1}{\text{card}(\mathcal{M}_K)} \sum_{m \in \mathcal{M}_K} \frac{|f_{S_i}(m) - f_{S_j}(m)|}{f_{S_i}(m)} \quad (2)$$

$$d_3^K(S_i, S_j) = \sum_{m \in \mathcal{M}_K} f_{S_i}(m) \log \left(\frac{f_{S_i}(m)}{f_{S_j}(m)} \right) \quad (3)$$

où $f_S(m)$ représente la fréquence d'apparition du mot m, de taille K dans la séquence S. La fonctions d_3 est la divergence de Kullback-Leibler, permettant de quantifier l'écart entre deux mesures de probabilités.

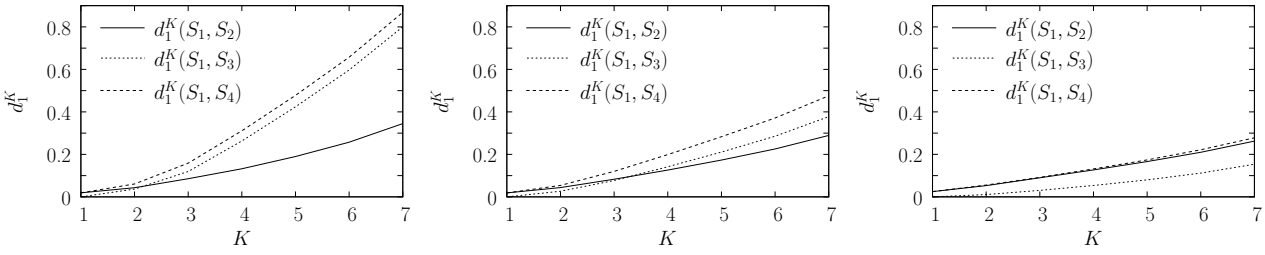
Les résultats présentés sur la figure 3(a) présentent les résultats obtenus avec les séquences codantes[1] de la bactérie *B. Thetaiotaomicron VPI-5482* avec la fonction d_1^K. On observe une forte différence statistique entre les séquences codantes prises dans le sens 1 avec les séquences codantes prises dans les sens 3 et 4 (valeurs élevées de d_1^K). Cette différence est très atténuée si l'on compare le sens 1 à le sens 2, qui est la séquence complémentaire des séquence codantes.

Nous avons ensuite cherché à reproduire ces résultats à partir de séquences codantes définies par des modèles théoriques. A partir de la séquence d'acide aminé, on détermine la séquence de codons associés suivant les hypothèses suivantes :

- la distribution des probabilités des codons synonymes pour représenter un acide aminé est équiprobable, la séquence obtenue par cette hypothèse sera qualifiée "d'équiprobable".
- la distribution des probabilités des codons est définie à partir de la distribution estimée sur la séquence biologique, cette séquence sera appelée séquence "suivant l'usage des codons".

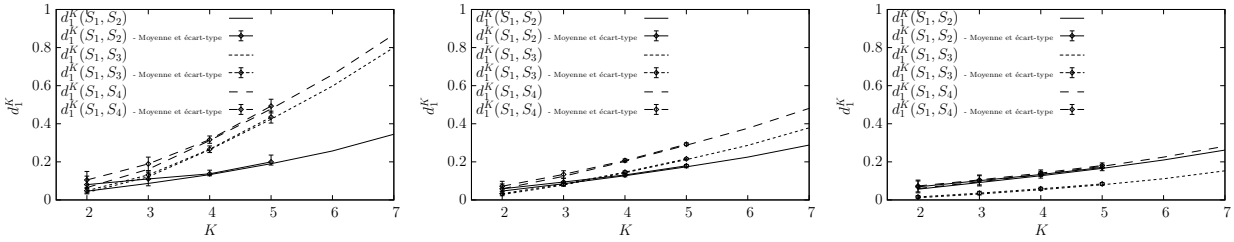
Il est important de remarquer que les séquences obtenues à partir de ces modèles théoriques représentent la même information génétique que la séquence biologique. Les figures 3(b) et 3(c) présentent les valeurs d_1^K pour les séquences dites "équiprobable" et "suivant l'usage des codons" pour la bactérie *B. Thetaiotaomicron VPI-5482*.

[1] Les séquences biologiques proviennent de la banque génomique NCBI (http ://www.ncbi.nlm.nih.gov/)



(a) Séquence biologique (b) Séquence "suivant l'usage des codons" (c) Séquence "équiprobable"

FIG. 3 – Valeur de d_1^K suivant les sens de lecture.



(a) Séquence biologique (b) Séquence "suivant l'usage des codons" (c) Séquence "équiprobable"

FIG. 4 – Comparaison des valeurs de d_1 obtenues sur l'ensemble des parties et sur des sous-ensembles de parties codantes.

On remarque que les résultats obtenus avec la séquence "suivant l'usage des codons" se rapproche de la séquence biologique bien que les différences statistiques sont moins marquées (figure 3(b)), et ceci alors que l'on tient compte de la statistique de l'utilisation des codons. Enfin les observations faites sur la séquence biologique n'apparaissent pas avec la séquence "équiprobable" (figure 3(c)).

Les résultats obtenus avec les fonctions d_2^K et d_3^K sont tout à fait similaires. Il est intéressant de noter que ces résultats se retrouvent sur les autres organismes testés. En effet les parties codantes de 25 organismes ont été testés, et 23 présentent cette caractéristique.

3.1 Pertinence des résultats

Nous avons ensuite vérifié la stationnarité des résultats ci-dessus en fonction de la position des codons dans la séquence étudiée. Pour cela, nous avons défini des fenêtres d'analyse juxtaposées sur la séquence composée de l'ensemble des parties codantes. L'objectif étant de disposer de suffisamment de fenêtres pour avoir une mesure statistique suffisante sur la dispersion des résultats tout en garantissant une mesure fiable sur chaque fenêtre d'analyse. Pour cela nous avons adopté le compromis suivant : la fenêtre d'analyse doit contenir au moins 100 représentants pour chaque mot de taille K avec une limite sur le nombre de nucléotides dans la fenêtre fixé à 200000.

Les figures 4(a), 4(b) et 4(c) présentent les résultats obtenus avec les séquences de la bactérie *B. Thetaiotaomicron VPI-5482* à titre d'illustration. L'ensemble des séquences codantes est constitué 5606913 bases nucléotidiques et les intervalles de confiance à σ associés à chaque mesure sont également indiqués sur les figures. On observe une très bonne concordance entre les résultats obtenus avec les fenêtres d'analyse et ceux obtenus sur la séquence complète ce qui indique une bonne stationnarité de la caractéristique étudiée en l'occurrence la dispersion de la fréquence d'apparition des mots de longueur K suivant les différents sens de lecture.

.

3.2 Comparaison avec un estimateur de la fréquence d'apparition des mots

Dans cette section nous allons poursuivre notre étude en comparant les résultats obtenus sur les séquences biologiques et sur les séquences obtenues à partir des modèles théoriques avec un estimateur de la fréquence d'apparition des mots de taille K dans une séquence ADN proposé. Cet estimateur s'obtient en utilisant le principe de maximisation de l'entropie ([7]), ou par le maximum de vraisemblance lorsque la séquence ADN est considérée comme un processus markovien ([10]). Soit S un séquence ADN et en notant $m_1^K = m_1 \dots m_K$ un mot de taille K, cet estimateur F_S sur une séquence S s'exprime de la manière suivante :

$$F_S(m_1^K) = \frac{f_S(m_1^{K-1}) f_S(m_2^K)}{f_S(m_2^{K-1})} \tag{4}$$

Comme nous pouvons le voir dans l'expression de cette estimateur, la fréquence d'apparition d'un mot de taille K est calculée à partir des fréquences d'apparition des mots de taille $K-1$ et $K-2$. Cet estimateur est souvent utilisé pour détecter si des mots sont sous-utilisés ou au contraire sur-utilisés dans les séquence ADN ([8], [9], [10]). A l'inverse des autres séquences considérés précédemment, cet estimateur ne tient en aucun cas compte de l'information génétique. Néanmoins ces propriétés (estimateur obtenu à partir du principe de maximisation d'entropie, estimateur du maximum de vraisemblance) en font un modèle théo-

rique intéressant pour étudier plus en détail la propriété des séquences biologiques vis-à-vis des différents sens de lecture.

Considérons les fonctions D_i^K, $i \in \{1,2,3\}$, définies comme les fonction d_1^K, d_2^K et d_3^K où les fréquences d'apparition des mots est donné par l'estimateur définit en (4). Il apparaît que D_1^K et D_3^K possède la propriété[2] suivante :

$$D_i^K(S_1,S_j) - d_i^{K-1}(S_1,S_j) = d_i^{K-1}(S_1,S_j) - d_i^{K-2}(S_1,S_j) \quad (5)$$

avec $i \in \{1,3\}$ et $j \in \{2,3,4\}$. La figure 5 illustre les valeurs prises par D_1^K vis-à-vis de d_1^K. Etant donné la propriété (5), nous considérons un nouveau paramètre $\epsilon_i^K(.,.) = D_i^K(.,.) - d_i^K(.,.)$ et représente l'erreur d'estimation en K. La figure 5 illustre la définition de ce paramètre sur un exemple.

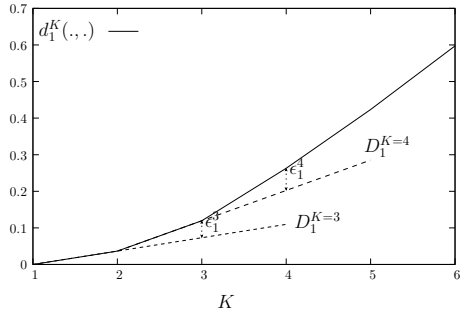


FIG. 5 – Interprétation de ϵ^K

La figure 6 présente les résultats obtenus pour $\epsilon_1^K(S_1,S_3)$. On remarque que quelque soit la taille des mots considérés, K, c'est pour la séquence bioloque qu'$\epsilon_K(S_{dir1},S_{dir3})$ est le plus élevé. De plus le modèle markovien utilisé dans [10] pour calculer cet estimateur pour des mots de taille K est d'ordre $K-2$. Ainsi il apparaît que la propriété markovienne n'est pas suffisante pour expliquer la propriété des séquences codantes mis en avant dans ce document.

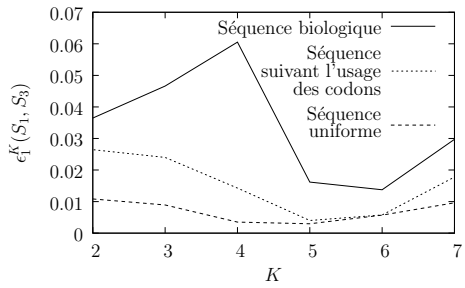


FIG. 6 – $\epsilon_1^K(S_1,S_3)$ pour les trois séquences considérées

4 Conclusion

Comme nous avons pu le constater, le code génétique permet de transmettre plus d'information que nécessaire pour transmettre les séquences des acides aminés (section 2). L'étude des séquences ADN codantes montrent de plus qu'il existe un mécanisme régissant l'association acide aminé-codon qui dépasse les notions statistiques prises en compte dans les séquences théoriques (différences observées entre la séquence biologique et la séquence "suivant l'usage des codons", section 3). De plus il apparaît que la propriété markovienne n'est pas suffisante pour expliquer la propriété des séquences codantes de l'ADN vis-à-vis du sens de lecture. Ainsi cette étude nécessite d'être poursuivie afin de comprendre les mécanismes présents dans les séquences biologiques.

[2] cette propriété est vérifié théoriquement pour la fonction $D_3^K(.,.)$, pour la fonction $D_1^K(.,.)$, cette propriété n'est pas rigoureuse théoriquement mais s'avère vérifié par simulation sur les séquences étudiées

Références

[1] R.G. Gallager, "Information theory and reliable communication", *John Wiley & sons*, 1968.

[2] P. Sharp, T. Tuohy,K. Mosurski, *Nucleic Acids Res.* 14,5125-5143, 1986.

[3] R. Sandberg, C.I. Bränden, I. Ernberg, J. Cöster, "Quantifying the species-specificity in genomic signatures, synonymous codon choice, amino acid usage and G+C content", *Gene* 311,35-42, 2003.

[4] R. Grantham, C. Gautier, M. Gouy, R. Mercier, A. Mercier, *Nucleic Acids Res.* 8, 49-62,1980.

[5] S. Karlin, C. Burge, "Dinucleotide relative abundance extremes : a genomic signature", *Trends Genet*, 11, 283-290, 1995.

[6] P.J. Deschavanne, A. Giron, J. Vilain, G. Fagot, B. Fertil, "Genomic signature : characterization and classification of species assessed by chaos game representation", *Mol. Biol. Evol.* 16, 1391-1399, 1999.

[7] R. Hu, B. Wang, "Statistically significant strings are related to regulatory elements in promoter regions of *Saccharomyces cerevisiae*, *Physica A*, 290, 464-474, 2001.

[8] B. Prum, F. Rodolphe, E. de Turckheim, "Finding words with unexpected frequencies in deoxyribonucleic acid sequences", *Journal of the Royal Society, Series B*, Vol.57, No.1, 205-220, 1995.

[9] M.Y. Leung, G.M. Marsh, T.S. Speed, "Over- and underrepresentation of short DNA words in herpesvirus genomes", *Journal of Computational Biology*, Vol.3,No.3, 345-360, 1996.

[10] G. Reinert, S. Schbath, M.S. Waterman, "Probabilistic an statistical properties of words : an overview", *Journal of Computational Biology*, Vol.7, No.1/2, 1-46, 2000.