L'ADN en tant que texte : style et syntaxe

Une syntaxe commune aux espèces

Sylvain Lespinats*, Patrick Deschavanne*,
Alain Giron*, Bernard Fertil*

* INSERM U.494 91 boulevard de l'hôpital 75634 Paris
lespinats@imed.jussieu.fr
http://genstyle.imed.jussieu.fr

Résumé. L'ADN peut être vu comme un texte dont la signification précise reste encore assez mystérieuse. On sait cependant que les fréquences d'utilisation des mots est spécifique à chacune des espèces (signature génomique). Nous avons montré que la signature génomique résulte d'un style d'écriture que l'on retrouve le long du génome. Bien que les signatures des espèces soient différentes, on observe qu'il existe cependant un consensus entre les espèces sur l'utilisation des mots. En effet, ce sont les mêmes mots qui sont les plus ou les moins variables le long du génome chez les différentes espèces. Certains de ces mots, comme les palindromes par exemple, ont des propriétés fréquentielles originales.

MOTS-CLÉS : style, signature génomique, mots, fréquences, variation le long du génome, palindromes.

1 Introduction

La molécule d'ADN est le support de l'information génétique d'une espèce. Elle contient les informations nécessaires au fonctionnement des cellules. On peut l'assimiler à un texte rédigé avec un alphabet de 4 lettres (les 4 bases : C (cytosine), G (guanine), A (adénine), et T (thymine)). La longueur des génomes (ensemble du matériel génétique d'une espèce) varie de 500 Kbp (bp = paires de bases) à 140 Gbp. En règle générale, le génome des procaryotes est en grande partie codant (plus de 70 %), alors que celui des eucaryotes comprend de larges portions qui ne codent aucun gène (certaines espèces peuvent avoir moins de 5 % de parties codantes).

Le premier génome séquencé a été celui de Haemophilus influenzae (équipe de Craig Venter, TIGR en 1995). Depuis, la quantité de séquences disponibles n'a cessé d'augmenter, le taux de croissance actuel est exponentiel.

L’organisation des génomes est encore mal connue, même si on a déjà situé un bon nombre des gènes des espèces séquencées. L’ADN peut être considéré comme un texte, à ceci près qu’il n’y a pas de séparateur entre les mots. Une des solutions possibles pour l’analyser est d’étudier les propriétés fréquentielles de tous les mots (suites de quelques lettres ou oligonucléotides) qu’il contient (Crochemore, Hancart et al. 2001).

Les fréquences d’utilisation des suites de nucléotides sont caractéristiques de chaque espèce. On appelle signature génomique d’une espèce l’ensemble de ces fréquences. Dans ce travail, nous recherchons ce qui est commun aux signatures des espèces, et ce qui les différencie.

2 Fréquences des mots et signature génomique

2.1 Signature génomique

On définira un mot de n lettres comme une suite de n nucléotides consécutifs. Il existe 4^n mots différents de longueur n. A chacun d’entre eux correspond une fréquence d’utilisation dans la séquence. L’ensemble de ces fréquences peut être visualisé, grâce à la CGR (Chaos Game Representation), sous la forme d’une image où chaque pixel est associé à un mot, et son intensité est proportionnelle à sa fréquence (plus un pixel est foncé, plus le mot correspondant est fréquent (fig. 1)) (Jeffrey 1990). A chaque espèce correspond une image CGR spécifique. Nous appellerons signature génomique l’ensemble des fréquences d’utilisation des mots par une espèce (Deschavanne, Giron et al. 1999). Elle est matérialisée par l’image CGR qui lui correspond.

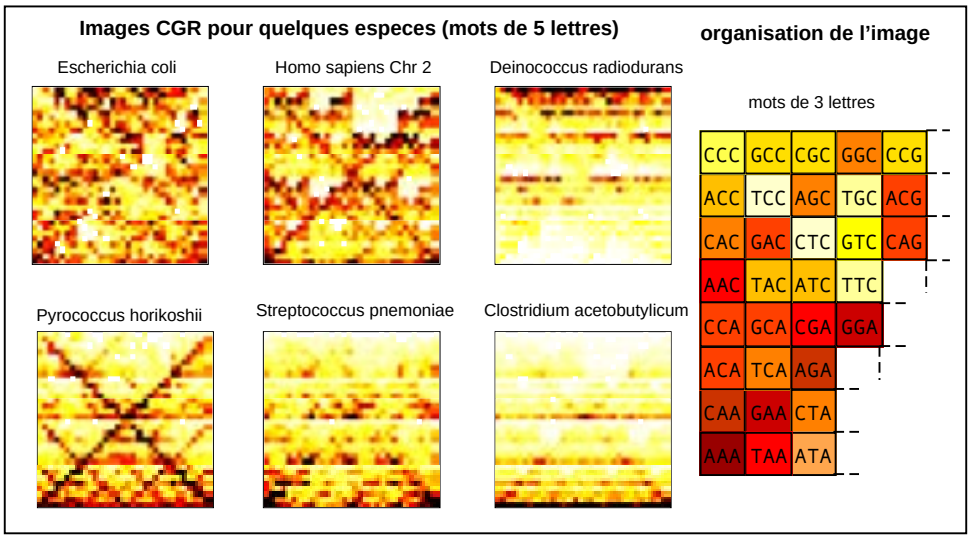

FIG. 1 - Images CGR de l’ensemble du génome de 6 espèces, pour des mots de 5 lettres. Les pixels sombres représentent les mots les plus fréquents, les pixels clairs représentent les mots les moins fréquents.

2.2 Style d'écriture du génome

La majorité des fragments d'ADN ont une signature proche de celle du génome dont ils proviennent (fig. 2). La signature génomique apparaît comme la conséquence d'un « style d'écriture » de l'ADN qui peut être observé presque partout dans le génome (Deschavanne, Giron et al. 2000).

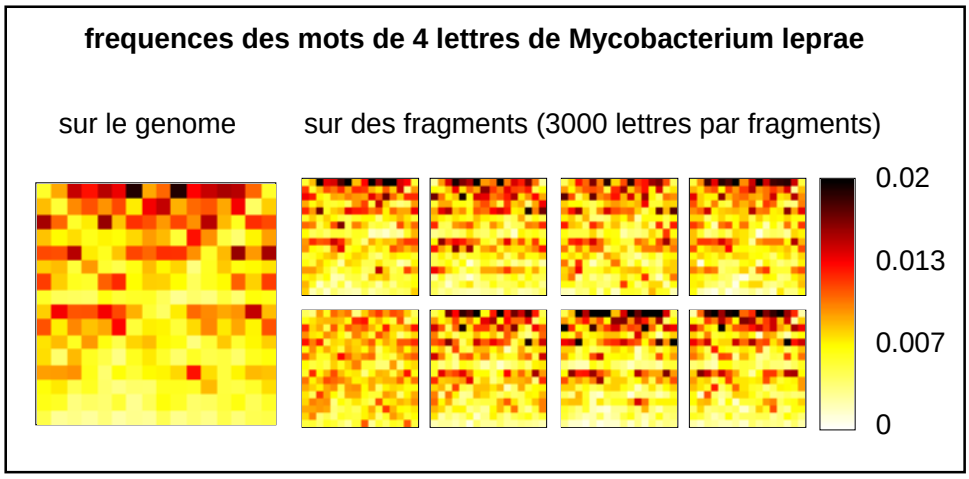

FIG. 2 - *Signature génomique de Mycobacterium leprae.*
Mots de 4 lettres : à gauche, l'image moyenne des fréquences des mots dans l'ensemble du génome, à droite, exemples d'images pour 8 fragments de 3000 lettres consécutives.

D'une manière générale, les variations intra-génome de la signature sont beaucoup plus faibles que les variations inter-génome. Une signature locale permet, dans la majorité des cas, de caractériser le style de l'espèce. En effet, il est possible de retrouver l'origine de plus de 90 % de fragments de 2000 nucléotides extraits d'une base de 34 espèces, grâce à leur signature (mots de 4 lettres) (Deschavanne, Giron et al. 1999).

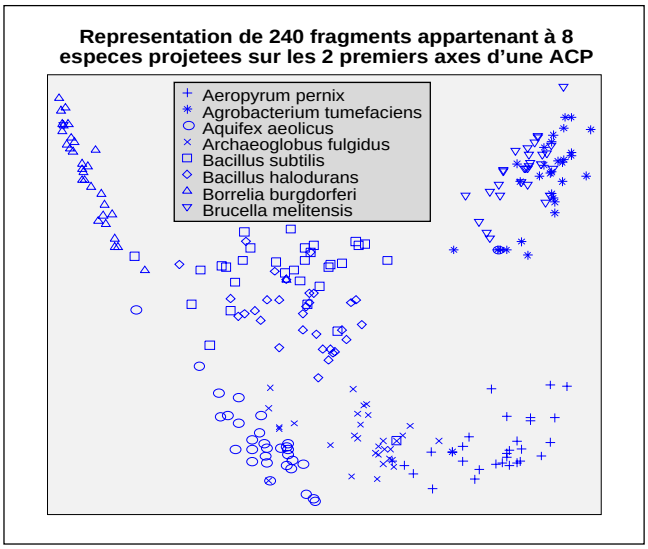

FIG 3 - *Visualisation des distances entre les signatures de fragments d'ADN. 8 espèces, 30 fragments de 3000 lettres par espèces, mots de 4 lettres (256 variables).*

3 Variation intra-génome des fréquences des mots

Les images des fragments d'un même génome sont assez comparables (fig. 2). Pour un génome donné, un ensemble de signatures locales peut être obtenu par échantillonnage à l'aide d'une fenêtre glissante. Si l'image CGR de chacune des fenêtres reste globalement invariante le long du génome, certains mots, cependant, ont une fréquence qui semble varier beaucoup.

Dans la suite de ce travail, nous étudions les fréquences des mots de 4 lettres, dans des fenêtres de 3000 lettres, pour 49 génomes complets de bactéries et archéobactéries.

3.1 Un indice pour quantifier la variation au long du génome

Sous l'hypothèse d'invariance de la fréquence des mots le long du génome, le nombre d'occurrences d'un mot dans une fenêtre suit une loi binomiale B(n,p) où p est la fréquence du mot dans la séquence entière et n le nombre de mots dans la fenêtre. On calcule une p-value à laquelle il est possible d'associer un Z-score. En pratique, on doit choisir n assez grand pour caractériser chaque mot par une fréquence d'utilisation.

On calcule un Z-score par mot et par fenêtre. La variance des Z-scores d'un mot donné dans les différentes fenêtres permet de mesurer la variation de fréquence d'un mot au long d'un génome indépendamment de sa fréquence moyenne. On appellera indice de variation intra-genome d'un mot, la variance du Z-score du mot le long du génome.

Il faut remarquer que l'utilisation d'une loi binomiale nous fait négliger les corrélations entre les fréquences des mots qui se recouvrent. On observe que les mots autorecouvrants (par exemple ATAT) ont des Z-scores comparables à ceux des autres mots, ce qui justifie à posteriori le fait de négliger l'incidence des recouvrements entre les mots.

3.2 Comparaison des variations intra-génome entre les différents génomes

On peut ainsi construire des images où chaque pixel représente la variation intra-génome d'un mot (fig. 4).

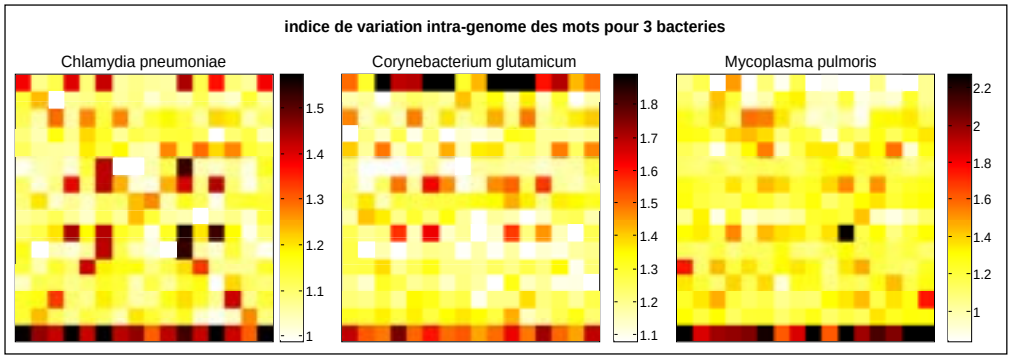

FIG 4 - *indice de variations intra-genomes des mots de 4 lettres pour 3 bactéries (les mots sont disposés dans l'image de la même façon que dans les images CGR).*

La plupart des mots, pour les espèces considérées dans ce travail, ont des indices de variation proches de 1. Cependant, certains mots, (les mots constitués uniquement de A et de T qui composent la bande du bas de l'image CGR, les mots constitués de C et de G (bande du haut de l'image CGR), et 8 mots au centre de l'image (ce sont 8 palindromes)) ont des fréquences d'utilisation plutôt variables.

Les indices de variation d'un mot dans différents génomes ne sont pas comparables car il existe des génomes plus homogènes que d'autres. En revanche, les indices des mots différents de même longueur dans un génome sont comparables (le Z-score est conçu pour cela). Pour chaque espèce, on peut attribuer un rang à chaque mot, en fonction de son indice de variation le long du génome, et comparer les rangs qu'occupe un même mot dans différents génomes. On associe ainsi un rang élevé aux mots les plus variables. Sous l'hypothèse d'indépendance des rangs, la moyenne des rangs pris par un mot dans les différents génomes doit suivre une loi normale.

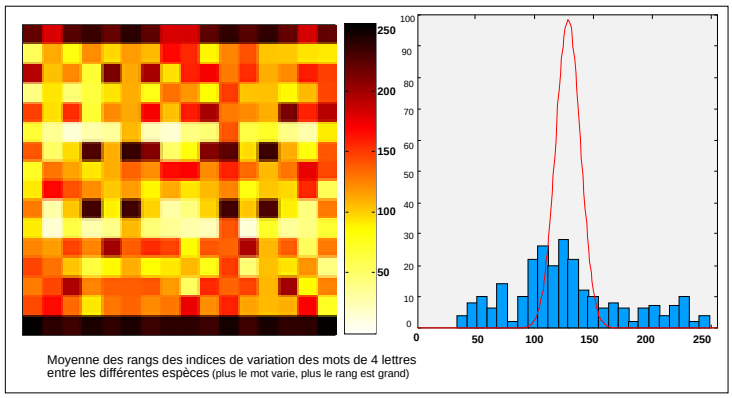

FIG 5 - Consensus entre les espèces sur la variation des mots.
La partie supérieure montre l'image des moyennes des rangs des indices de variation des mots de 4 lettres chez les différentes espèces (plus un mot est variable au long des génomes, plus le pixel qui le représente est foncé). La partie inférieure montre l'histogramme des moyennes des rangs des Z-scores, ainsi que leur distribution théorique sous l'hypothèse d'indépendance des rangs.

Les mots qui varient le plus le long du génome sont souvent les mêmes (fig. 5). On observe 12 mots pour lesquels la moyenne des rangs est inférieure à 50 (sur 256), et 15 mots dont la moyenne des rangs est supérieure à 226. Il existe donc un consensus entre espèces sur l'ordre de variation intra-génome des mots (fig. 5).

On peut penser que, par analogie avec le langage, il existe des mots qui structurent le texte et des mots qui portent le sens. Les mots de structure sont les mots les moins variables (quel que soit le « sens de la phrase », ils doivent être présents avec des fréquences stables). Les mots de contenu sont les mots variables (ils peuvent être très fréquents, ou absents selon le sens du fragment étudié).

4 Contribution des mots aux variations de la signature

On peut penser que les mots qui varient beaucoup le long des génomes ont très peu d'utilité pour définir la signature génomique. Pour connaître l'effet de la variation intra-génome des mots sur la signature, nous allons étudier leur contribution à la reconnaissance de l'espèce d'origine de fragments d'ADN.

4.1 Méthode

On souhaite, retrouver l'espèce d'origine d'une fenêtre extraite d'un génome en utilisant comme seule information les fréquences des mots (fig. 3). Chaque séquence est caractérisée par la fréquence d'apparition des mots de 4 lettres. Elle représente donc un point dans un espace vectoriel à 136 dimensions (correspondant à 136 mots si on élimine les redondances liées à l'analyse conjointe des deux brins de l'ADN). Une procédure de classification

(affectation au plus proche voisin) affecte l'origine d'un fragment d'ADN à l'espèce la plus proche, au sens d'une métrique définissant la distance entre les signatures (Lebart, Morineau et al. 1995). L'expérience montre que la distance euclidienne répond bien au problème posé. Pour cette étude, des fenêtres non recouvrantes de 3 Kb, extraites de 49 espèces, ont fait l'objet d'une classification.

4.2 Effet des variations intra-génome sur la classification

Lorsque tous les mots sont utilisés pour calculer la distance entre les signatures, 89 % des fenêtres sont bien classées. Pour tester diverses hypothèses, la classification est reprise avec un nombre variable de mots choisis au hasard (fig. 6). On observe une variabilité importante de la classification en fonction des mots utilisés, le pourcentage de bonne classification augmente toutefois continûment avec le nombre de mots.

Il en est de même quand on utilise les axes principaux de l'analyse en composantes principales. Toutefois, la prise en compte des 30 premiers axes donne déjà des résultats comparables à ceux obtenues avec l'ensemble des mots. Le meilleur résultat est obtenu pour 99 mots, quand on utilise les mots dans l'ordre croissant de leur variance intra-génomique (91.5 %) (encart droit de la figure 6). On peut cependant remarquer que les mots les moins variants des génomes ont un pouvoir discriminant réduit (encart gauche de la figure 6). La détermination, à l'aide d'un algorithme génétique, du meilleur ensemble de mots pour discriminer les espèces, donne les meilleurs résultats de classification (94.5 %). Plusieurs combinaisons de mots (qui associent des mots de faible variance intra-génome et quelques mots de variance forte) conduisent en fait à des résultats comparables, mais le nombre de mots utilisés est relativement constant (70 mots ±3).

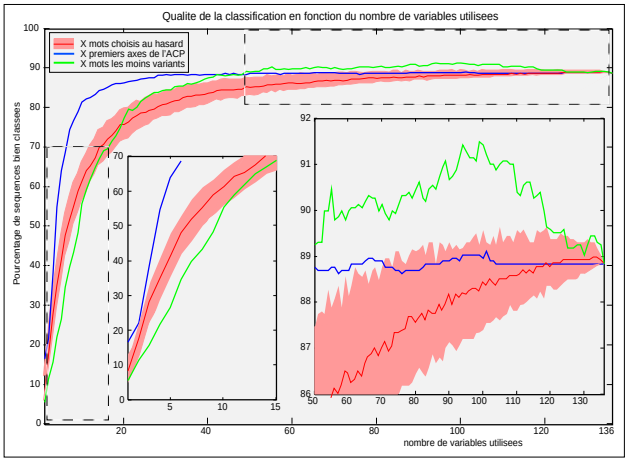

FIG 6 - Qualité de la classification en fonction du nombre de variables utilisées. On classe les séquences avec un nombre croissant de variables sur la base :
- de mots sélectionnés au hasard (partie grisée : intervalle de confiance à 5 %, courbe centrale : médiane des valeurs obtenues)
- des composantes de l'ACP
- des mots ordonnés selon leur variance intra-génome.

4.3 variation intra-génome – variation inter-génomes

On quantifie la variation inter–génome d'un mot par la variance de sa fréquence dans les différents génomes.

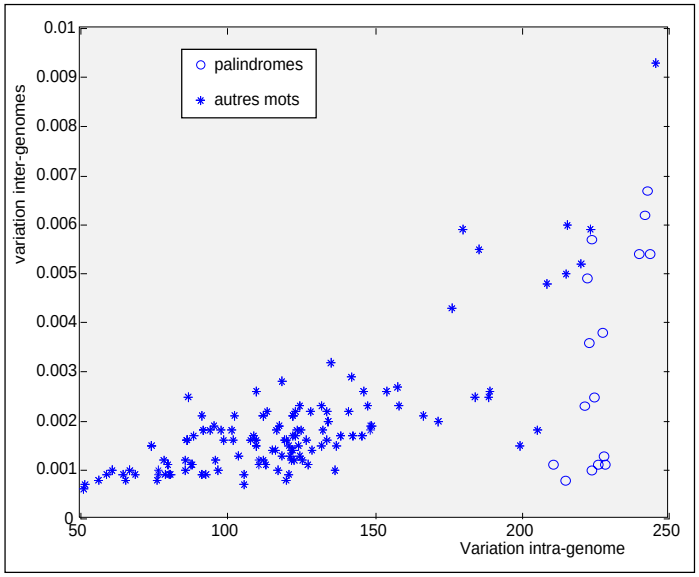

FIG 7 - Comparaison des variations intra-génome et inter-génome des mots de 4 lettres. Chaque point représente un mot de 4 lettres.

Il existe une corrélation positive forte entre la variation intra-génome et la variation inter-génome (Fig. 7). Cela explique que les mots qui ont une faible variation intra-génome ne permettent pas une bonne classification des séquences. En effet, leurs fréquences sont comparables entre les espèces. Ces mots, qui sont candidats potentiels pour le rôle de mots structurants, sont les suivants : AGTC, AGAC, GACA, ACTC, et TGAC, et leurs complémentaires.

Les mots très variables au long des génomes (et qui sont massivement absents des groupes qui permettent les meilleures classifications des séquences), sont les palindromes, les mots uniquement constitués de A et de T et ceux uniquement constitués de C et de G (un palindrome, en génétique, est un mot qui est son propre complémentaire, lu dans l'autre sens, exemple AGCT) (fig. 5).

4.4 Fréquences des palindromes

Nous comparons les fréquences des palindromes avec celles des autres mots de 4 lettres par une méthode d'échantillonnage.

Il a été montré que la moyenne des fréquences des palindromes de 4 lettres est inférieure à celle des non-palindromes de même taille dans plusieurs fragments de génomes de

bactéries (Karlin, Burge et al. 1992). Notre approche par échantillons multiples, appliquée aux séquences aujourd'hui disponibles, confirme ce résultat. Les fréquences du groupe des palindromes sont hétérogènes, il y a en particulier un nombre important de palindromes très fréquents.

Ils sont de plus caractérisés par une grande variance intra-génome, tandis que leurs variances inter-génome sont très diverses (fig. 7).

5 Conclusion

Il existerait une syntaxe commune aux espèces : les « mots de structure » (avec des fréquences assez stables au long des génomes) organisent la « phrase », les « mots de contenu » (dont les fréquences sont variables) portent le « sens ». La signature s'appuie davantage sur les « mots de structure » que sur les « mots de contenu ».

Ces résultats préliminaires ont été obtenues avec 49 espèces. Leur généralité ne pourra être évaluée qu'au fur et à mesure de la disponibilité de nouvelles séquences génomiques de grande taille. En particulier on pourra rechercher l'existence de structures de phrase spécifiques à certaines branches taxonomiques. Le domaine d'expression du style ayant été confirmé, nous utiliserons ces informations pour structurer une base de données de signatures génomiques, qui sera mise à la disposition de la communauté scientifique sur le site http://genstyle.imed.jussieu.fr/.

Références

Brunet, E. (2002). "le lemme comme on l'aime." *JADT 2002, Saint Malo France,13-15 mars 2002* :221-232.

Crochemore, M., C. Hancart, T. Lecoq (2001). "Algorithme du texte." Vuibert.

Deschavanne, P., A. Giron, et al. (2000). "Genomic signature is preserved in short DNA fragments." *BIBE2000 IEEE international Symposium on bio-informatics & biomedical engineering, Washington USA, 8-10 november 2000*: 161-167.

Deschavanne, P. J., A. Giron, et al. (1999). "Genomic signature: characterization and classification of species assessed by chaos game representation of sequences." *Mol Biol Evol* **16**(10): 1391-9.

Jeffrey, H. J. (1990). "Chaos game representation of gene structure." *Nucleic Acids Res* **18**(8): 2163-70.

Lebart, L., A. Morineau, M. Piron (1995). "Statistique exploratoire multidimensionnelle." Dunod.

Lebart L., A. Salem (1994) "Statistique textuelle." Dunod

Karlin, S., C. Burge, et al. (1992). "Statistical analyses of counts and distributions of restriction sites in DNA sequences." *Nucleic Acids Res* **20**(6): 1363-70.

L’ADN en tant que texte : style et syntaxe

Remerciements

Ce travail a été financé en partie par un contrat Bioinformatique Inter-EPST (2001).

Summary

It is known that word usage is species-specific, thus allowing to derive the so-called genomic signature. We have shown that the genomic signature results from a regular usage of words along sequences (writing style). Although genomic signatures are species-specific, there is an agrement between species about the variability of word usage along genomes. The same words are found less or more variable whatever the species. the frequencies of some of these words such as palindroms, for exemple, are particulary original.

KEYWORDS : style, genomic signature, words, frequencies, variation along genome, palindroms.