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.