

Représentation par jeu du chaos de séquences d'ADN

Henin Thibaut, ENS Cachan - Symbiose IRISA

19 septembre 2007

1 Introduction

Le jeu du chaos [1, 2] est un algorithme permettant de dessiner des images fractales en faisant décrire une trajectoire à un point dans le plan. À chaque étape de l'algorithme, on fait bouger un point en le rapprochant de moitié d'un point de contrôle, choisi aléatoirement ou de manière déterministe parmi un ensemble de points fixes dans le plan. Suivant ces points et la manière de les choisir, les images présentent ou non diverses structures fractales.

Dans [1], il a été suggéré d'utiliser des suites de symboles pour remplacer le générateur aléatoire. L'idée est d'assigner un symbole (ou plusieurs) à chaque point de contrôle. Ensuite, à chaque étape de l'algorithme, on lit une lettre de la suite, qui déterminera le point cible de la transformation. La figure obtenue après la lecture de la suite caractériserait ainsi la suite elle-même.

On peut alors tracer des figures pour représenter des séquences d'ADN [1, 3]. Nous avons ici représenté les séquences du chromosome 14 de l'homme et de E. Coli [Fig. 1.1]. Chaque génome produit des images plus ou moins structurées. Dans la représentation par jeu du chaos du génome humain [Fig. 1.1], on note, par exemple un trou, en haut à droite, qui se répète un peu partout en plus petit dans l'image. Par contre, sur un génome comme E. Coli [Fig. 1.1], on a le sentiment que l'image contient un certain ordre qu'on ne peut pas caractériser aussi facilement que pour l'homme.

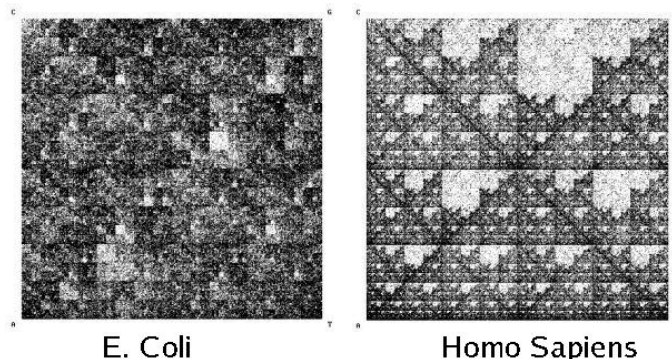


FIG. 1.1 – Le jeu du chaos avec les 500 000 premières bases du chromosome 14 de l'Homme (à droite), et de E. Coli (à gauche) (Chaque point noir correspond à une itération de l'algorithme)

2 Autosimilarité de la représentation de langages réguliers à l'aide d'IFS

Les IFS [4] sont des systèmes de transformations $\{T_i\}_{i \in I}$ appliquées itérativement. Ainsi, on définit la suite E_k des étapes de la transformation telle que pour une étape k , $E_k = \bigcup_{i \in I} T_i(E_{k-1})$. On appelle attracteur de l'IFS l'ensemble E , s'il est point fixe de la transformation $E = T(E) = \bigcup_{i \in I} T_i(E)$. Par définition, cet ensemble, quand il existe, est auto-similaire, c'est à dire un ensemble formé de copies plus ou moins déformées de lui-même.

Il est également possible de construire automatiquement un IFS pour la représentation par jeux du chaos d'un langage régulier qui produira des motifs observés. Cette construction utilise les états et les transitions de l'automate complet déterministe correspondant au langage pour construire les transformations. Ce résultat est important car il montre que les langages réguliers ont une représentation autosimilaire. De plus, les motifs notables des représentations sont produit par de simples langages réguliers.

3 Caractérisation des structures fractales des représentation de séquences d'ADN

La représentation par jeu du chaos permet, entre autres, de regrouper visuellement les images des mots de même suffixe dans des cadrans. On note qu'en ajoutant la même lettre à la fin de différents mots, les points des images de ces nouveaux mots se retrouvent dans le même quartier de la figure. On continue le raisonnement en constatant qu'en ajoutant une autre lettre, les images de ces mots se retrouvent encore une fois regroupées dans des carrés plus petit encore.

$$\forall k, i, j \in \mathbb{N} : 0 \leq i, j < 2^k \quad E_{i,j}^k =]2^{-k} \cdot i, 2^{-k} \cdot (i+1)[\times]2^{-k} \cdot j, 2^{-k} \cdot (j+1)[$$

On définit la distance ($d_c \in \mathbb{R}^+$) entre deux points du carré unitaire x et y dans $]0, 1[^2$ comme la taille du plus petit cadran les contenant tous les deux, et zéro s'ils sont égaux :

$$d_c(x, y) = \begin{cases} 0 & \text{si } x = y \\ \inf_k(2^{-k} : \exists 0 \leq i, j < 2^k \wedge \{x, y\} \subset E_{i,j}^k) & \text{sinon} \end{cases}$$

On définit ensuite la distance entre langages ($d_l(L_1, L_2)$) en utilisant la distance de Hausdorff utilisant entre leurs représentations par jeu du chaos respectives : $d_l(A, B) = \max_{a \in A}(\min_{b \in B} d_c(a, b))$. On interprète alors la valeur de $d_l(L_1, L_2)$ comme une estimation de la taille des mots présents dans les deux langages. Plus d_l est petit, plus la longueur des mots de L_2 contenus dans L_1 est grande.

Comme pour la distance de Hausdorff en général, la version non symétrique (d_l) donne plus d'informations que la distance elle-même (d_L). Par exemple, la distance d_l du le génome humain vers le langage modélisant la diagonale (ce langage contient tous les mots n'utilisant que a et g) est inférieure à $\frac{1}{2}^{10}$ (car dans notre représentation, tous les pixels de la diagonale sont noirs). On sait alors que tous les mots de $\{a, g\}^*$ de taille inférieure à 10 sont présents dans la séquence. Par contre, la distance du langage vers la séquence est de 1 car la séquence utilise les deux autres bases. De même, la distance entre la séquence et le langage interdisant cg est inférieure à $\frac{1}{2}^{10}$ alors que la distance dans l'autre sens est de $1/4$ (puisque la séquence, contient quand même une occurrence de cg).

4 Conclusion

Grâce à cette distance, on caractérise la présence d'un motif autosimilaire en terme d'inclusion de langage régulier dans la séquence. En effet, étant donné un langage régulier, on détermine automatiquement sa représentation. Le fait que les motifs de la représentation du langage soit marqués dans la représentation d'une séquence se traduit par l'inclusion de tous les mots de taille au moins 10 du langage régulier dans la séquence. Ainsi, si un motif est marqué dans une séquence, cela indique que les mots du langage représenté par le motif sont tous présents jusqu'à une certaine taille dans la séquence.

Références

- [1] H. Joel Jeffrey. Chaos game representation of gene structure. 1990.
- [2] Peter Tiño. Spacial representation of symbolic sequences through iterativ function systems. 1999.
- [3] Patrick J. Deshavanne, Alain giron, Joseph vilain, Guillaume Fagot, and Bernard Fertil. Genomic signature : characterization and classification of species assessed by chaos game representation of sequences. 1999.
- [4] Kenneth Falconer. *Fractal Geometry*. John Wiley, 2003.