Recherche de signaux de codes circulaires dans les gènes

Nestor Demeure, Christian J. Michel

- 2017 -

1. ACQUISITION DES GENES

Gènes de bactéries (bacteria), archaea (archaea), eucaryotes (eukaryotes), plasmides (plasmids), mitochondries (mitochondria), chloroplastes (chloroplasts) et virus (viruses).

2. FONCTION DE CORRELATION

Un langage F (génome) est constitué de $n(F)$ mots (gènes) sur l'alphabet $\mathcal{A} = \{A, C, G, T\}$. Soit x un mot de F de longueur $|x|$ lettres (nucléotides). Soient 2 motifs w et w' de longueurs respectives $|w|$ et $|w'|$ sur $\mathcal{A}$. Soit m_i, appelé i-motif, 2 motifs w et w' séparés par i, $i \in \{0, \dots, imax\}$, lettres quelconques N et noté $m_i = wN^iw'$. Pour chaque mot x de F, le compteur $c_i(x)$ compte les occurrences de m_i dans x. Pour compter les occurrences de m_i dans les mêmes conditions pour tout $i \in \{0, \dots, imax\}$, uniquement les $l(x) = |x| - imax - |w'|$ premières lettres de x sont considérées.

Remarque: $l(x)$ se termine sur la dernière lettre de w ($l(x) = m(x) + 2$ avec $m(x)$ la longueur pour la fonction de corrélation moyenne)

Remarque: $imax$ doit être un multiple de 3.

Alors la probabilité d'occurrence $o_i(x)$ de m_i dans x est égale au ratio du compteur par le nombre de lettres étudiées

$$o_i(x) = \frac{3 \times c_i(x)}{l(x)}.$$

La probabilité d'occurrence $A_{w,w'}(i, F)$ de m_i dans F est donc égale à

$$A_{w,w'}(i, F) = \frac{1}{n(F)} \sum_{x \in F} o_i(x).$$

La fonction $i \to A_{w,w'}(i, F)$ donnant la probabilité d'occurrence que w' apparaisse i lettres quelconques N après w dans le langage F, est dite fonction de corrélation wN^iw' (associée au i-motif wN^iw').

Remarque importante:

Les mots w sont analysés en phase 0 modulo 3 (en phase de lecture), c'est-à-dire par pas de 3 lettres à partir des trinucléotides d'initiation.

Les mots w' sont analysés après chaque mot w en position i, $i \in \{0, \dots, imax\}$, c'est-à-dire par pas de 1 lettre.

Remarque importante:

$\sum_{w,w'} A_{w,w'}(i, F) = 1$ pour tout i, $i \in \{0, \dots, imax\}$, lettres quelconques N^i.

Cette fonction de corrélation wN^iw' est représentée par une courbe avec:

- en abscisse, le nombre i de lettres N entre w et w', i variant de 0 à $imax$
- en ordonnée, la probabilité $A_{w,w'}(i,F)$ d'occurrence de wN^iw' dans F.

3. Application

- gènes de bactéries, archaea, eucaryotes, plasmides, mitochondries, chloroplastes et virus avec $|x| \geq 200$ nucléotides.
- $imax = 99$.
- $w, w' \in \{X, X_1, X_2, X_p\}$ où le code circulaire X (maximal, C^3 et autocomplémentaire) est

$$X = \{\mathrm{AAC}, \mathrm{AAT}, \mathrm{ACC}, \mathrm{ATC}, \mathrm{ATT}, \mathrm{CAG}, \mathrm{CTC}, \mathrm{CTG}, \mathrm{GAA}, \mathrm{GAC},$$
$$\mathrm{GAG}, \mathrm{GAT}, \mathrm{GCC}, \mathrm{GGC}, \mathrm{GGT}, \mathrm{GTA}, \mathrm{GTC}, \mathrm{GTT}, \mathrm{TAC}, \mathrm{TTC}\},$$

le code circulaire $X_1 = \mathcal{P}(X)$ ($\mathcal{P}$ étant l'application de permutation) est

$$X_1 = \{\mathrm{AAG}, \mathrm{ACA}, \mathrm{ACG}, \mathrm{ACT}, \mathrm{AGC}, \mathrm{AGG}, \mathrm{ATA}, \mathrm{ATG}, \mathrm{CCA}, \mathrm{CCG},$$
$$\mathrm{GCG}, \mathrm{GTG}, \mathrm{TAG}, \mathrm{TCA}, \mathrm{TCC}, \mathrm{TCG}, \mathrm{TCT}, \mathrm{TGC}, \mathrm{TTA}, \mathrm{TTG}\},$$

le code circulaire $X_2 = \mathcal{P}^2(X)$ est

$$X_2 = \{\mathrm{AGA}, \mathrm{AGT}, \mathrm{CAA}, \mathrm{CAC}, \mathrm{CAT}, \mathrm{CCT}, \mathrm{CGA}, \mathrm{CGC}, \mathrm{CGG}, \mathrm{CGT},$$
$$\mathrm{CTA}, \mathrm{CTT}, \mathrm{GCA}, \mathrm{GCT}, \mathrm{GGA}, \mathrm{TAA}, \mathrm{TAT}, \mathrm{TGA}, \mathrm{TGG}, \mathrm{TGT}\}$$

et le code

$$X_p = \{\mathrm{AAA}, \mathrm{CCC}, \mathrm{GGG}, \mathrm{TTT}\}.$$

4. Programmation

- Paramétrer $|x|$ et $imax$ dans le programme
- Dans une feuille Excel, les valeurs numériques des 16 fonctions de corrélation: XN^iX, XN^iX_1, …, $X_pN^iX_p$:

i	$A_{X,X}$	i	A_{X,X_1}	…
0	0.123	0	0.223	
1	0.145	1	0.245	
…	…	…	…	
$imax$	0.167	100	0.267	

- Dans une autre feuille Excel, les courbes associées aux 16 fonctions de corrélation.

Remarque importante:

Pour chaque fonction de corrélation, on dessine 3 courbes modulo 3: une courbe reliant les points 0 modulo 3, une courbe reliant les points 1 modulo 3 et une courbe reliant les points 2 modulo 3.