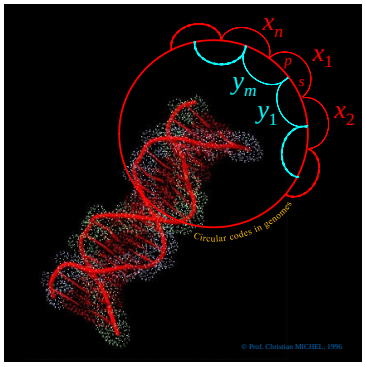
Codes circulaires dans les gènes



Christian J. Michel
Theoretical bioinformatics, ICube
University of Strasbourg, CNRS
Email: c.michel@unistra.fr
URL: http://dpt-info.u-strasbg.fr/~c.michel/

1 Contexte scientifique

Les codes de trinucléotides (mots de trois lettres sur un alphabet à quatre lettres), comme le code génétique, sont à l'origine d'une théorie fascinante qui combine la recherche de solutions à des problèmes anciens et récents en utilisant des méthodes modernes issues de plusieurs domaines scientifiques.[1]

Il y a 60 ans environ, avant la découverte du code génétique, une classe de codes de trinucléotides appelés codes comma-free a été proposée par Crick et al. (1957) pour expliquer comment une lecture d'une séquence (suite) de trinucléotides pouvait coder les acides aminés (mots d'une lettre sur un alphabet à 20 lettres). En particulier, comment la phase de lecture des gènes pouvait être retrouvée, maintenue et synchronisée. Les quatre nucléotides $\{A, C, G, T\}$ comme les 16 dinucléotides $\{AA, ..., TT\}$ sont des codes inappropriés pour coder les 20 acides aminés. Les 64 trinucléotides $\{AAA, ..., TTT\}$ induisent une redondance dans leur codage. Ainsi, Crick et al. (1957) ont proposé que seulement

[1] Le lecteur non spécialiste du domaine peut omettre cette section.

20 trinucléotides parmi 64 codent les 20 acides aminés. Un tel code bijectif implique que les trinucléotides codants se trouvent tous dans une même phase - la propriété de comma-free. La détermination d'un ensemble de 20 trinucléotides constituant un code comma-free repose sur deux conditions nécessaires:
(i) Un trinucléotide périodique de l'ensemble $\{AAA,CCC,GGG,TTT\}$ doit être exclu d'un tel code. En effet, la concaténation de AAA avec lui-même, par exemple, ne permet pas de retrouver la phase de lecture (phase originale) puisqu'il existe trois décompositions possibles: $\ldots AAA,AAA,AAA\ldots$ (phase originale), $\ldots A,AAA,AAA,AA\ldots$ et $\ldots AA,AAA,AAA,A\ldots$, les virgules montrant la décomposition adoptée.
(ii) Deux trinucléotides permutés non-périodiques, i.e. deux trinucléotides liés par permutation circulaire, par exemple ACG et CGA, doivent être également exclus d'un tel code. En effet, la concaténation de ACG avec lui-même, par exemple, ne permet pas de retrouver la phase de lecture puisqu'il existe deux décompositions possibles: $\ldots ACG,ACG,ACG\ldots$ (original frame) and $\ldots A,CGA,CGA,CG\ldots$ Ainsi, en excluant les quatre trinucléotides périodiques et en regroupant les 60 trinucléotides restants en 20 classes de trois trinucléotides tel que, dans chaque classe, les trois trinucléotides se déduisent les uns des autres par permutation circulaire, par exemple ACG, CGA et GAC, on remarque qu'un code comma-free ne peut contenir qu'un seul trinucléotide de chaque classe contient ainsi au plus 20 trinucléotides. Ce nombre de 20 trinucléotides est identique au nombre de 20 acides aminés, conduisant ainsi à un code associant un trinucléotide par acide aminé sans ambiguïté.
Quelques résultats combinatoires sur les codes comma-free de trinucléotides ont été obtenus par Golomb et al. (1958a,b). Cependant, aucun code comma-free de trinucléotides n'a été identifié statistiquement dans les gènes. De plus, au début des années 1960, un résultat biologique a montré que le trinucléotide TTT, i.e. un trinucléotide périodique exclu dans un code comma-free, code l'acide aminé phénylalanine (Nirenberg and Matthaei, 1961). Ce résultat a conduit à l'abandon du concept de code comma-free sur l'alphabet $\{A,C,G,T\}$ à quatre lettres. Pour plusieurs raisons biologiques, en particulier l'interaction entre ARN messager (ARNm) et ARN de transfert (ARNt), ce concept de code comma-free a été repris plus tard sur l'alphabet purine/pyrimidine $\{R,Y\}$ ($R=\{A,G\}$, $Y=\{C,T\}$) à deux lettres avec deux codes comma-free de trinucléotides: RRY (Crick et al., 1976) et $RNY=\{RRY,RYY\}$ (N étant une lettre quelconque sur $\{R,Y\}$) (Eigen and Schuster, 1978). Des résultats statistiques ont confirmé ces deux codes comma-free dans les gènes, au niveau du gène par Shepherd (1981) et de la population de gènes par Michel (1989). En 1986, il est montré que les introns, contrairement aux exons (régions codantes des gènes des eukaryotes), ne possèdent pas de périodicité nucléotidique modulo 3 (Figure 2

dans Michel, 1986, avec une analyse statistiques de 90 introns). Un an plus tard, avec la croissance des données de séquences, une périodicité nucléotidique modulo 2 est identifiée dans les introns avec deux différentes méthodes statistiques (Konopka and Smythers, 1987; Arquès and Michel, 1987). A ce jour, aucun code comma-free ou circulaire n'a été identifié dans les introns.

En 1996, une analyse statistique des fréquences d'occurrence des 64 trinucleotides $\{AAA, ..., TTT\}$ dans les trois phases des gènes ausis bien des procaryotes que des eucaryotes a montré que les trinucléotides ne sont pas uniformément distribués dans ces trois phases (Arquès et Michel, 1996). En excluant, les quatre trinucléotides périodiques et en affectant à chaque trinucléotide une phase préférentielle (phase de fréquence d'occurrence maximale), trois sous-ensembles $X = X_0$, X_1 et X_2 de 20 trinucléotides sont identifiés dans les phases 0 (phase de lecture), 1 (phase 0 décalée d'un nucleotide dans la direction $5' \rightarrow 3'$, i.e. vers la droite) et 2 (phase 0 décalée de deux nucleotides dans la direction $5' \rightarrow 3'$) dans les gènes à la fois des procaryotes et des eucaryotes. Cet ensemble X contient les 20 trinucléotides suivants:

$$X = \{AAC,AAT,ACC,ATC,ATT,CAG,CTC,CTG,GAA,GAC, \\ GAG,GAT,GCC,GGC,GGT,GTA,GTC,GTT,TAC,TTC\}. \quad (1)$$

Les deux ensembles X_1 et X_2 de 20 trinucléotides chacun, dans les phases décalées 1 et 2, respectivement, des gènes se déduisent de X par permutation circulaire. Ces trois ensembles de trinucléotides présentent des propriétés mathématiques remarquables, en particulier X est un code circulaire de trinucléotides qui est en plus maximal, C^3 et autocomplémentaire (Arquès et Michel, 1996). Le sous-ensemble $\{CAG,CTC,CTG,GAG\}$ de X est un code comma-free de trinucléotides et également C^3 et autocomplémentaire (Michel, 2012). Les codes comma-free représentent une classe limite des codes circulaires avec trois propriétés: (i) aucun mot d'un code comma-free ne se trouve en phase décalée; (ii) la longueur (nombre de lettres) pour retrouver la phase de lecture (phase originale ou phase de décomposition) est la plus courte; et (iii) la probabilité de codage de la phase de lecture est maximale et égale à 1 (Michel, 2014). En 2015, en quantifiant l'approche par inspection utilisée en 1996 pour identifier une phase préférentielle à chaque trinucléotide et en appliquant une analyse statistique massive de groupes taxonomiques de gènes, le résultat du code circulaire X est confirmé et renforcé dans les gènes des procaryotes (7,851,762 gènes, 2,481,566,882 trinucléotides) et des eucaryotes (1,662,579 gènes, 824,825,761 trinucléotides), et également identifié dans les gènes des plasmides (237,486 gènes, 68,244,356 trinucléotides) et virus (184,344 gènes, 45,688,798 trinucléotides) (Michel, 2015).

Un code circulaire de trinucléotides possède la propriété fondamentale de pouvoir toujours retrouver la phase de lecture de toute séquence construite avec un code circulaire, i.e. quelque soit la position dans la séquence pour retrouver la pahse de lecture. En particulier, les trinucléotides d'initiation et stop ainsi que tout signal de phase sont inutiles pour definir la phase de lecture. En effet, une fenêtre de quelques nucléotides, dont la longueur dépend de la classe des codes circulaires, positionnée n'importe où dans une séquence construite par le code circulaire permet toujours de retrouver la phase de lecture. Prenons comme exemple le mot $w = ...AGGTAATTACCAG...$ du code circulaire X. Le premier nucléotide de w, i.e. A, est-il le 1er, 2ème ou le 3ème nucléotide d'un trinucléotide de X? En essayant les trois factorisations (phases) possibles w_0, w_1 et w_2 (w_1 et w_2 étant w_0 décalé d'un et de deux nucléotides, respectivement) en trinucléotides de X, une seule factorisation, dasn cet exemple w_1, est possible. Ainsi, le 1er nucléotide A de w est le 3ème nucléotide d'un trinucléotide de X. En effet, la factorisation w_1 donne la suite des trinucléotides NNA, GGT, AAT, TAC et CAG (N étant une lettre quelconque appropriée de X) qui appartiennent à X. Les factorisations w_0 et w_2 sont impossibles car aucun trinucléotide de X ne commence avec le préfixe AG. Ce cas apparaît immédiatement pour w_0 et après 11 lettres pour w_2. Ainsi, l'unique factorisation de w est $w_1 = ...A,GGT,AAT,TAC,CAG,....$ Le mot $w' = AGGTAATTACCA$ (w sans la dernière lettre G) de longueur 12 nucléotides est ambigü car il possède deux factorisations w_1 et w_2 en trinucléotides de X. Le mot w' est dit mot ambigü de X. Par définition d'un code circulaire, tous les mots ambigüs sont des mots finis. Le mot w', pris ici comme exemple d'illustration, est un des quatre plus longs mots ambigus de X. Ainsi, la longueur l de fenêtre pour retrouver la phase de lecture pour tout mot d'un code circulaire Y est la longueur des plus long mots ambigus plus une lettre. Avec le code circulaire X, $l = 12 + 1 = 13$ nucléotides (Michel, 2008). Les longueurs l de fenêtre pour les codes circulaires X_1 et X_2 sont également égales à $l = 13$ nucléotides (Michel, 2008). En conclusion, la phase de lecture avec le code circulaire X, la phase 1 avec le code circulaire X_1 et la phase 1 avec le code circulaire X_2 nécessitent une longueur de fenêtre l de 13 nucléotides ($l \geq 13$).

Récemment en 2012, en plus de l'existence du code circulaire X dans les gènes (ADN et ARN messager noté ARNm), une deuxième étape majeure de cette théorie du code circulaire est obtenue avec l'identification de motifs du code circulaire X, en abrégé motifs X, dans les régions $5'$ et/ou $3'$ des ARNs de transfert (ARNt) des procaryotes et des eucaryotes (Michel, 2012, 2013) et des ARNs ribosomiques 16S (ARNr 16S) des procaryotes et des eucaryotes, en particulier dans le centre de décodage du ribosome dans

lequel les nucléotides A1492 and A1493 universellement conservés dans les procaryotes et eucaryotes, et le nucléotide G530 conservé dans les procaryotes appartiennent à des motifs X (Michel, 2012; El Soufi et Michel, 2014). Une visualisation 3D des motifs X dans le ribosome montre plusieurs configurations spatiales impliquant des motifs X de l'ARNm, des motifs X de l'ARNt et des motifs X de l'ARNr 16S (Michel, 2012; El Soufi and Michel, 2014). Tous ces résultats suggèrent un code de translation des gènes basé sur le code circulaire (Michel, 2012).

2 Statistiques des codes circulaires

2.1 Méthodes statistiques identifiant un code circulaire "universel" dans les gènes

Les classes de méthodes statistiques qui ont permis l'identification du code circulaire X (1) dans les gènes sont:

- Trinucleotide frequencies per frame (Arquès et Michel, 1996)
- Correlation functions per frame (Arquès et Michel, 1997)
- Frame permuted trinucleotide frequencies (Frey et Michel, 2003, 2006)
- Covering function (Gonzalez, Giannerini et Rosa, 2011)
- Occurrence probability of a complementary/permutation trinucleotide set in a gene kingdom (Michel, 2015)

Le principe de toutes ces méthodes est simple. Il repose sur l'étude des 64 trinucléotides dans les trois phases des gènes. La méthode la plus simple proposée en 1996 (Arquès et Michel, 1996) est basée sur l'inspection des trinucléotides ayant la fréquence d'occurence maximale parmi les trois phases des gènes. Chaque trinucléotide est affecté à la phase des gènes de fréquence d'occurence maximale. Cette approche permet de classer presque tous les 64 trinucléotides. Pour les quelques trinucléotides ayant des fréquences d'occurrence similaires dans les trois phases ou proches de l'aléatoire ($1/64 = 1.5625\%$), des méthodes statistiques sont nécessaires (cf. ci-dessus).

Occurrence frequencies $P(w^p)$ *of the* 64 *trinucleotide* w *in each frame* p *in the prokaryotic protein coding genes* (13 686 *sequences*, 4 708 758 *trinucleotides*)

w in frame $p=0$	Frequency (%)	w in frame $p=1$	Frequency (%)	w in frame $p=2$	Frequency (%)
AAA	3.38	AAA	2.75	AAA	2.44
AAC	2.18	AAC	1.59	AAC	1.38
AAG	1.98	**AAG**	3.21	AAG	0.81
AAT	2.17	AAT	1.37	AAT	1.69
ACA	1.22	**ACA**	1.91	ACA	1.11
ACC	2.09	ACC	1.60	ACC	0.79
ACG	1.30	**ACG**	2.49	ACG	0.68
ACT	1.13	**ACT**	1.17	ACT	1.09
AGA	0.61	AGA	1.59	**AGA**	2.47
AGC	1.42	**AGC**	1.83	AGC	1.71
AGG	0.31	**AGG**	2.21	AGG	1.45
AGT	0.87	AGT	0.97	**AGT**	1.26
ATA	0.83	**ATA**	2.15	ATA	0.66
ATC	2.61	ATC	1.66	ATC	0.82
ATG	2.38	**ATG**	2.82	ATG	0.41
ATT	2.50	ATT	1.38	ATT	1.50
	Frame 0		**Frame 1**		**Frame 2**

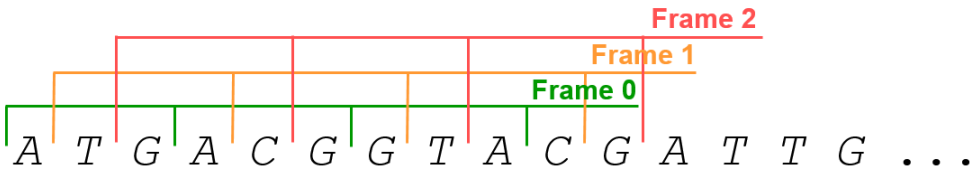


Figure 1: Identification du circular code X (1) dans les gènes (Arquès et Michel, 1996).

2.2 Signaux statistiques des codes circulaires dans les gènes

Nous renvoyons le lecteur aux travaux de Frey et Michel (2006), Ahmed, Frey et Michel (2007, 2010) et de Ahmed et Michel (2008, 2011).

3 Définitions

Les définitions suivantes sont classiques pour un ensemble fini de mots sur un alphabet fini. Soit $\mathcal{A}_4 = \{A, C, G, T\}$ l'alphabet génétique (nucléotides ou lettres) ordonné lexicographiquement par $A < C < G < T$. L'ensemble des mots non vides (mots resp.) sur $\mathcal{A}_4$ est noté $\mathcal{A}_4^+$ ($\mathcal{A}_4^*$ resp.). L'ensemble des 4^n mots de longueur n sur $\mathcal{A}_4$ est noté par $\mathcal{A}_4^n = \{A^n, \ldots, T^n\}$. Les dinucléotides et trinucléotides sont les mots de longueur $n = 2$ (dilettres) et $n = 3$ (trilettres) sur $\mathcal{A}_4$. Ainsi, l'ensemble des 16 mots de longueur $n = 2$ sur $\mathcal{A}_4$ est noté $\mathcal{A}_4^2 = \{AA, \ldots, TT\}$. L'ensemble des 64 mots de longueur $n = 3$ sur $\mathcal{A}_4$ est noté $\mathcal{A}_4^3 = \{AAA, \ldots, TTT\}$. Soit $x_1 \cdots x_n$ la concaténation des mots x_i pour $i = 1, \ldots, n$.

Il existe deux applications biologiques importantes en relation avec les codes dans les gènes sur $\mathcal{A}_4$: l'application de complémentarité qui se base sur la structure de la double hélice d'ADN et l'application de permutation circulaire dont la définition est associée à l'identification du code circulaire X (1) dans les gènes (Arquès et Michel, 1996).

Definition 1 *L'application de complémentarité sur les nucléotides* $\mathcal{C} : \mathcal{A}_4 \to \mathcal{A}_4$ *est définie par* $\mathcal{C}(A) = T$, $\mathcal{C}(C) = G$, $\mathcal{C}(G) = C$ *et* $\mathcal{C}(T) = A$.

D'après les propriétés de complémentarité et antiparallèle de la double hélice d'ADN, l'application de complémentarité sur les mots est définie de la façon suivante:

Definition 2 *L'application de complémentarité sur les mots* $\mathcal{C} : \mathcal{A}_4^n \to \mathcal{A}_4^n$ *est définie par* $\mathcal{C}(uv) = \mathcal{C}(v)\mathcal{C}(u)$ *pour tous* $u \in \mathcal{A}_4, v \in \mathcal{A}_4^{n-1}$.

Definition 3 *L'application de complémentarité sur les ensembles de mots* $\mathcal{C} : \mathbb{P}(\mathcal{A}_4^n) \to \mathbb{P}(\mathcal{A}_4^n)$ ($\mathbb{P}$ *étant l'ensemble des sous-ensembles*) *est définie par* $\mathcal{C}(\mathcal{S}) = \{v \mid u, v \in \mathcal{A}_4^n, u \in \mathcal{S}, v = \mathcal{C}(u)\}$.

Example 4 *Sur* $\mathcal{A}_4^2$, *nous avons* $\mathcal{C}(\{AC,AG\}) = \{CT,GT\}$ *et sur* $\mathcal{A}_4^3$, *nous avons* $\mathcal{C}(\{ACG,AGT\}) = \{ACT,CGT\}$.

Remark 5 *L'application de complémentarité* $\mathcal{C}$ *est involutionelle, i.e. pour tout ensemble de mots* $\mathcal{S}$, $\mathcal{C}(\mathcal{C}(\mathcal{S})) = \mathcal{S}$.

Definition 6 *L'application de permutation circulaire sur les trinucléotides* $\mathcal{P} : \mathcal{A}_4^3 \to \mathcal{A}_4^3$ *est définie par* $\mathcal{P}(l_0l_1l_2) = l_1l_2l_0$ *pour tous* $l_0, l_1, l_2 \in \mathcal{A}_4$. *Le 2ème itéré de* $\mathcal{P}$ *est noté* $\mathcal{P}^2$.

Definition 7 *L'application de permutation circulaire sur les ensembles de trinucléotides* $\mathcal{P} : \mathbb{P}(\mathcal{A}_4^3) \to \mathbb{P}(\mathcal{A}_4^3)$ *est définie par* $\mathcal{P}(\mathcal{S}) = \{v \mid u, v \in \mathcal{A}_4^3, u \in \mathcal{S}, v = \mathcal{P}(u)\}$.

Example 8 $\mathcal{P}(\{ACG,AGT\}) = \{CGA,GTA\}$ *et* $\mathcal{P}^2(\{ACG,AGT\}) = \{GAC,TAG\}$.

Definition 9 (Code) *Un ensemble de mots* $\mathcal{S} \subset \mathcal{A}_4^n$ *est un code si pour tous* $x_1, \ldots, x_n$, $y_1, \ldots, y_m \in \mathcal{S}$, $n, m \geq 1$, *la condition* $x_1 \cdots x_n = y_1 \cdots y_m$ *implique* $n = m$ *et* $x_i = y_i$ *pour* $i = 1, \ldots, n$.

Les codes sont lus sur une droite.

Remark 10 *L'ensemble $\mathcal{A}_4^n$ est un code. En particulier, le code génétique $\mathcal{A}_4^3$ est un code.*

Remark 11 *L'ensemble $\{A,GC,AGC\}$ n'est pas un code parce qu'il existe la décomposition $A{\cdot}GC{=}AGC$.*

Remark 12 *Les sous-ensembles non vides de $\mathcal{A}_4^n$ sont des codes.*

Definition 13 *Les sous-ensembles de $\mathcal{A}_4^2$ et $\mathcal{A}_4^3$ sont appelés codes de dinucléotides et codes de trinucléotides, respectivement.*

Definition 14 (Code comma-free) *Un code $\mathcal{S} \subset \mathcal{A}_4^n$ est comma-free si pour tous $y \in S$ et $u, v \in \mathcal{A}_4^*$, la condition $uyv = x_1 \cdots x_n$ avec $x_1, \ldots, x_n \in S$, $n \geq 1$, implique $u, v \in S^*$.*

Example 15 *Le code à 20 trinucleotides $\{AAC,AAG,AAT,CAC,CAG,CAT,CCG,CCT,GAC,GAG,GAT,GCG,GCT,GGT,TAC,TAG,TAT,TCG,TCT,TGT\}$ est comma-free.*

Example 16 *Le code génétique $\mathcal{A}_4^3$ n'est pas comma-free parce qu'il existe la décomposition $A{\cdot}CGA{\cdot}CG{=}ACG{\cdot}ACG$.*

Definition 17 (Code circulaire) *Un code $\mathcal{S} \subset \mathcal{A}_4^n$ est circulaire si pour $x_1, \ldots, x_n$, $y_1, \ldots, y_m \in \mathcal{S}$, $n, m \geq 1$, $r \in \mathcal{A}_4^*$, $s \in \mathcal{A}_4^+$, les conditions $sx_2 \cdots x_n r = y_1 \cdots y_m$ et $x_1 = rs$ impliquent $n = m$, $r = \varepsilon$ (mot vide) et $x_i = y_i$ pour $i = 1, \ldots, n$.*

Les codes circulaires sont lus sur un cercle.

4 Théorie combinatoire des codes circulaires

Quelques théorèmes combinatoires sur les codes circulaires de trinucléotides et dinucléotides sont présentés dans ce chapitre. Deux preuves de deux théorèmes permettent également d'illustrer l'approche mathématique. Le lecteur est renvoyé aux articles publiés pour les preuves non décrites dans ce chapitre.

4.1 Codes circulaires de trinucléotides

Deux nouvelles classes de codes circulaires de trinucléotides, appelés code circulaires forts, ont été récemment définies.

Definition 18 (Code circulaire fort *DLD*, Michel et Pirillo, 2011) *Un code circulaire fort de trinucléotides $\mathcal{S} \subset \mathcal{A}_4^3$ est DLD si pour tous $l_1, l_2, l_3, l'_1, l'_2, l'_3 \in \mathcal{A}_4$, les conditions $l_1 l_2 l_3 \in S$ et $l'_1 l'_2 l'_3 \in S$ impliquent $l_1 \neq l'_3$.*

Aucune lettre de $\mathcal{A}_4$ ne peut être simultanément en 1er position d'un trinucléotide de $\mathcal{S}$ et en dernière position (3ème) d'un autre trinucléotide of $\mathcal{S}$.

Definition 19 (Code circulaire fort *LDL*, Michel et Pirillo, 2011) *Un code circulaire fort de trinucléotides $\mathcal{S} \subset \mathcal{A}_4^3$ est LDL si pour tous $l_1, l'_1 \in \mathcal{A}_4$, $d_1, d'_1 \in \mathcal{A}_4^2$, les conditions $l_1 d_1 \in S$, $d'_1 l'_1 \in S$ impliquent $d_1 \neq d'_1$.*

Aucune dilettre de $\mathcal{A}_4^2$ ne peut être simultanément prefixe d'un trinucléotide de $\mathcal{S}$ et suffixe d'un autre trinucléotide of $\mathcal{S}$.

Example 20 *Le code circulaire fort $\{ACG, GTA\}$ est LDL mais pas DLD. Le code circulaire fort $\{ACG, CGT\}$ est DLD mais pas LDL.*

Definition 21 (Code circulaire autocomplémentaire) *Un code circulaire $\mathcal{S} \subset \mathcal{A}_4^n$ est autocomplémentaire si pour tout $x \in \mathcal{S}$, $\mathcal{C}(x) \in \mathcal{S}$, i.e. $\mathcal{C}(\mathcal{S}) = \mathcal{S}$.*

Definition 22 (Code circulaire C^3) *Un code circulaire $\mathcal{S} \subset \mathcal{A}_4^n$ de trinucléotides est C^3 si $S_1 = \mathcal{P}(S)$ et $S_2 = \mathcal{P}^2(S)$ sont des codes circulaires.*

Definition 23 (Code circulaire C^3 et autocomplémentaire) *Un code circulaire $\mathcal{S} \subset \mathcal{A}_4^n$ de trinucléotides est C^3 et autocomplémentaire si $\mathcal{S}$, $S_1 = \mathcal{P}(S)$ et $S_2 = \mathcal{P}^2(S)$ sont des codes circulaires vérifiant la relation $\mathcal{S} = \mathcal{C}(\mathcal{S})$ (autocomplémentaire) et $\mathcal{C}(S_1) = S_2$ et $\mathcal{C}(S_2) = S_1$ (S_1 et S_2 sont complémentaires).*

Definition 24 (Code circulaire maximal) *Un code circulaire $\mathcal{S} \subset \mathcal{A}_4^n$ est maximal si pour tout $x \in \mathcal{A}_4^n$, $x \notin \mathcal{S}$, $\mathcal{S} \cup \{x\}$ n'est pas un code circulaire.*

Definition 25 *Un code circulaire de dinucléotides (trinucléotides resp.) de longueur l, i.e. contenant exactement l éléments, est appelé code circulaire l-dinucléotides (l-trinucléotides resp.).*

Remark 26 *Un code circulaire* 6*-dinucléotides est toujours maximal. Un code circulaire* 20*-trinucleotides est toujours maximal.*

Theorem 27 (Michel et Pirillo, 2011) *Un code circulaire fort DLD de trinucléotides est comma-free.*

Theorem 28 (Michel et Pirillo, 2011) *Un code circulaire fort LDL de trinucléotides est comma-free.*

Remark 29 *Il existe des codes comma-free qui ne sont pas des codes circulaires forts DLD. Example: $\{ACA\}$.*

Remark 30 *Il existe des codes comma-free qui ne sont pas des codes circulaires forts LDL. Example: $\{ACG, CGT\}$.*

Proposition 31 (Michel et Pirillo, 2011) *Pour toutes lettres $x, y, z \in \mathcal{A}_4$, un trinucléotide singleton $xyz \in \mathcal{A}_4^3$ est un code circulaire fort DLD sur $\mathcal{A}_4$ SSI $x \neq z$.*

Proposition 32 (Michel et Pirillo, 2011) *Pour toutes lettres $x, y, z \in \mathcal{A}_4$, un trinucléotide singleton $xyz \in \mathcal{A}_4^3$ est un code circulaire fort LDL sur $\mathcal{A}_4$ SSI au moins deux de ses lettres sont différentes.*

Theorem 33 (Michel et Pirillo, 2011) *Aucun code circulaire* 20*-trinucleotides n'est un code circulaire fort DLD de trinucléotides.*

Theorem 34 (Michel et Pirillo, 2011) *Un code circulaire* 20*-trinucleotides est comma-free SSI il est un code circulaire fort LDL de trinucléotides.*

Notation 35 *$l_1, l_2, \ldots, l_n$ sont des lettres dans $\mathcal{A}_4$, $d_1, d_2, \ldots, d_n$ sont des dilettres dans $\mathcal{A}_4^2$ et n est un entier satisfaisant $n \geq 2$.*

Definition 36 *Letter Diletter Necklaces (LDN): La séquence ordonnée* $l_1, d_1, l_2, d_2, \ldots, d_{n-1}, l_n, d_n$ *est dite* $nLDN$ *pour un sous-ensemble* $S \subset A_4^3$ *si* $l_1d_1, l_2d_2, \ldots, l_nd_n \in S$ *et* $d_1l_2, d_2l_3, \ldots, d_{n-1}l_n \in S$.

Definition 37 *Letter Diletter Continued Necklaces (LDCN): La séquence ordonnée* $l_1, d_1, l_2, d_2, \ldots, d_{n-1}, l_n, d_n, l_{n+1}$ *est dite* $(n+1)LDCN$ *pour un sous-ensemble* $S \subset A_4^3$ *si* $l_1d_1, l_2d_2, \ldots, l_nd_n \in S$ *et* $d_1l_2, d_2l_3, \ldots, d_{n-1}l_n, d_nl_{n+1} \in S$.

Definition 38 *Diletter Letter Necklaces (DLN): La séquence ordonnée* $d_1, l_1, d_2, l_2, \ldots, l_{n-1}, d_n, l_n$ *est dite* $nDLN$ *pour un sous-ensemble* $S \subset A_4^3$ *si* $d_1l_1, d_2l_2, \ldots, d_nl_n \in S$ *et* $l_1d_2, l_2d_3, \ldots, l_{n-1}d_n \in S$.

Definition 39 *Diletter Letter Continued Necklaces (DLCN): La séquence ordonnée* $d_1, l_1, d_2, l_2, \ldots, l_{n-1}, d_n, l_n, d_{n+1}$ *est dite* $(n+1)DLCN$ *pour un sous-ensemble* $S \subset A_4^3$ *si* $d_1l_1, d_2l_2, \ldots, d_nl_n \in S$ *et* $l_1d_2, l_2d_3, \ldots, l_{n-1}d_n, l_nd_{n+1} \in S$.

Definition 40 *Letter Diletter Continued Closed Necklaces (LDCCN): La séquence ordonnée* $l_1, d_1, l_2, d_2, \ldots, d_{n-1}, l_n, d_n, l_{n+1}$ *est dite* $(n+1)LDCCN$ *pour un sous-ensemble* $S \subset \mathcal{A}_4^3$ *si* $l_1d_1, l_2d_2, \ldots, l_nd_n \in S$, $d_1l_2, d_2l_3, \ldots, d_{n-1}l_n, d_nl_{n+1} \in S$ *et* $l_1 = l_{n+1}$.

Theorem 41 (Pirillo, 2003) *Soit* S *un code de trinucleotides. Les conditions suivantes sont équivalentes:*
(i) S *est un code circulaire.*
(ii) S *n'a pas de collier* $5LDCN$.

Definition 42 *Letter Diletter Continued Closed Necklaces (LDCCN): We say that the ordered sequence* $l_1, d_1, l_2, d_2, \ldots, d_{n-1}, l_n, d_n, l_{n+1}$ *is an* $(n+1)LDCCN$ *for a subset* $X \subset \mathcal{A}_4^3$ *if* $l_1d_1, l_2d_2, \ldots, l_nd_n \in X$ *and* $d_1l_2, d_2l_3, \ldots, d_{n-1}l_n, d_nl_{n+1} \in X$ *and* $l_1 = l_{n+1}$.

Theorem 43 (Michel et Pirillo, 2010) *Soit* S *un code de trinucleotides. Les conditions suivantes sont équivalentes:*
(i) S *est un code circulaire.*
(ii) S *n'a pas de collier* $nLDCCN$ *pour tout entier* $n \in \{2, 3, 4, 5\}$.

Proof. $(i) \Rightarrow (ii)$. Par contradiction, suppose que S a un $nLDCCN$ pour un entier $n \in \{2, 3, 4, 5\}$.
Si le collier est un $2LDCCN$ alors $l_1, d_1, l_1, d_1, l_1, d_1, l_1, d_1, l_1$ est un $5LDCN$ pour S.
Si le collier est un $3LDCCN$ alors $l_1, d_1, l_2, d_2, l_1, d_1, l_2, d_2, l_1$ est un $5LDCN$ pour S.
Si le collier est un $4LDCCN$ alors $l_1, d_1, l_2, d_2, l_3, d_3, l_1, d_1, l_2$ est un $5LDCN$ pour S.
Si le collier est un $5LDCCN$ alors $l_1, d_1, l_2, d_2, l_3, d_3, l_4, d_4, l_1$ est un $5LDCN$ pour S.
Dans chacun des quatre cas, d'après le Théorème (41), S n'est pas un code circulaire de trinucléotides. Contradiction.
$(ii) \Rightarrow (i)$. Par contradiction, suppose que S n'est pas un code circulaire de trinucléotides. D'après le Théorème (41), S a un $5LDCN$, soit $l_1, d_1, l_2, d_2, l_3, d_3, l_4, d_4, l_5$.
Comme l'alphabet $\mathcal{A}_4$ possède 4 lettres, alors $l_i = l_j$ pour i, j, $1 \leq i \leq j \leq 5$.
Si $j - i = 4$ alors $l_1 = l_5$ et $[l_1, d_1, l_2, d_2, l_3, d_3, l_4, d_4]$ est un $5LDCCN$ pour S.
Si $j - i = 3$ alors $[l_i, d_i, l_{i+1}, d_{i+1}, l_{i+2}, d_{i+2}]$ est un $4LDCCN$ pour S.
Si $j - i = 2$ alors $[l_i, d_i, l_{i+1}, d_{i+1}]$ est un $3LDCCN$ pour S.
Si $j - i = 1$ alors $[l_i, d_i]$ est un $2LDCCN$ pour S.
Dans chacun des quatre cas, d'après le Théorème (41), il existe une contradiction avec (ii).

Proposition 44 (Michel et Pirillo, 2010) *La fonction de croissance des codes circulaires de trinucléotides est*

l	1	2	3	4	5	6	7	8	9	10
Nb(l)	60	1,704	30,432	382,164	3,568,212	25,507,512	141,639,780	614,568,102	2,086,742,208	5,542,646,244

l	11	12	13	14	15	16	17
Nb(l)	11,503,061,124	18,615,667,124	23,403,485,556	22,700,634,924	16,787,523,072	9,279,022,320	3,708,717,048

18	19	20
1,012,099,740	168,726,792	12,964,440

Le nombre de codes circulaires (de trinucléotides) de longueur $l = 1$ est égale à 60. Le nombre de codes circulaires maximum de longueur $l = 20$ est égale à 12,964,440 (Arquès et Michel, 1996). Le nombre maximum de codes circulaires est 23,403,485,556 pour la longueur $l = 13$.

Proposition 45 (Arquès et Michel, 1996) *Le nombre de codes circulaires de trinucléotides qui sont C^3, autocomplémentaires et maximum de longueur $l = 20$ est égale à 216.*

Theorem 46 (Lacan et Michel, 2001) *Le modèle probabiliste de Koch et Lehmann (1997) ne peut pas générer le code circulaire de trinucléotides C^3 et autocomplémentaire X (1) observé dans les gènes.*

Proposition 47 (Lacan et Michel, 2001) *L'extension de la méthode de Koch et Lehmann (1997) permet de générer la sous-classe des 88 codes circulaires de trinucléotides qui sont C^3, autocomplémentaires et maximum de longueur $l = 20$ parmi les 216.*

Theorem 48 (Michel, Pirillo et Pirillo, 2008) *Soit S un code de trinucleotides. Les conditions suivantes sont équivalentes:*
(i) S est un code comma-free.
(ii) S n'a pas de collier $2LDN$ ni de collier $2DLN$.

Proposition 49 (Michel, Pirillo et Pirillo, 2008) *La fonction de croissance des codes comma-free de trinucléotides est*

l	1	2	3	4	5	6	7	8	9	10
Nb(l)	60	1,656	25,608	244,008	1,530,060	6,638,340	20,708,460	47,742,654	82,816,632	109,358,220

l	11	12	13	14	15	16	17	18	19	20
Nb(l)	110,895,036	87,031,844	53,227,980	25,473,732	9,519,912	2,743,080	591,864	90,420	8,760	408

Le nombre de codes comma-free (de trinucléotides) de longueur $l = 1$ est égale à 60. Le nombre de codes comma-free maximum de longueur $l = 20$ est égale à 408. Le nombre maximum de codes comma-free est 110,895,036 pour la longueur $l = 11$.

Proposition 50 (Michel, Pirillo et Pirillo, 2008) *La fonction de croissance des codes comma-free de trinucléotides qui sont C^3 et autocomplémentaires est*

l	2	4	6	8	10	12	14	16	18	20
Nb(l)	28	182	424	498	340	144	36	4	0	0

Le nombre de codes comma-free (de trinucléotides) C^3 et autocomplémentaires de longueur $l = 2$ est égale à 28. Le nombre de codes comma-free C^3 et autocomplémentaires maximum de longueur $l = 16$ est égale à 4. Il n'existe pas de code comma-free C^3 et autocomplémentaire pour les longueurs $l = 18, 20$. Le nombre maximum de codes comma-free C^3 et autocomplémentaires est 498 pour la longueur $l = 8$.

Definition 51 *Soit S un code de trinucleotides. Pour tout entier $n \in \{2,3,4,5\}$, S appartient à la classe C^{nLDN} si S n'a pas de collier $nLDN$ et S appartient à la classe C^{nDLN} si S n'a pas de collier $nDLN$. Similairement, pour tout entier $n \in \{3,4,5\}$, S appartient à la classe C^{nLDCN} si S n'a pas de collier $nLDCN$ et S appartient à la classe C^{nDLCN} si S n'a pas de collier $nDLCN$.*

Notation 52 *Pour tout entier $n \in \{2,3,4,5\}$, $I^n = C^{nLDN} \cap C^{nDLN}$ et $U^n = C^{nLDN} \cup C^{nDLN}$. Similairement, pour tout entier $n \in \{3,4,5\}$, $I^nC = C^{nLDCN} \cap C^{nDLCN}$ et $U^nC = C^{nLDCN} \cup C^{nDLCN}$.*

Theorem 53 (Michel, Pirillo et Pirillo, 2008) *Les suites d'inclusion suivantes sont vérifiées*
(i) $C^{2LDN} \subset C^{3LDCN} \subset C^{3LDN} \subset C^{4LDCN} \subset C^{4LDN} \subset C^{5LDCN} \subset C^{5LDN}$.
(ii) $C^{2DLN} \subset C^{3DLCN} \subset C^{3DLN} \subset C^{4DLCN} \subset C^{4DLN} \subset C^{5DLCN} \subset C^{5DLN}$.
(iii) $C^{2LDN} \subset C^{3DLCN} \subset C^{3LDN} \subset C^{4DLCN} \subset C^{4LDN} \subset C^{5DLCN} \subset C^{5LDN}$.
(iv) $C^{2DLN} \subset C^{3LDCN} \subset C^{3DLN} \subset C^{4LDCN} \subset C^{4DLN} \subset C^{5LDCN} \subset C^{5DLN}$.
(v) $I^2 \subset I^3C \subset I^3 \subset I^4C \subset I^4 \subset I^5C \subset I^5$.
(vi) $U^2 \subset U^3C \subset U^3 \subset U^4C \subset U^4 \subset U^5C \subset U^5$.

Theorem 54 (Michel et Pirillo, 2011) *Les codes circulaires 20-trinucleotides vérifie la suite d'inclusion et d'égalité suivante:*
$\emptyset = LDL \cap DLD \subset LDL \cup DLD = LDL = I^2 \subset U^2 = I^3C \subset U^3C = I^3 \subset U^3 = I^4C \subset U^4C = I^4 \subset U^4 = I^5C \subset U^5C = I^5 = U^5$.

Remark 55 *La chaîne d'inclusion des codes circulaires commence avec les codes circulaires forts LDL et DLD et les codes comma-free I^2.*

4.2 Codes circulaires de dinucléotides

Definition 56 *Soit les lettres $l_1, l_2, l_3, \ldots, l_n, l_{n+1}$ sur $\mathcal{A}_4$. La séquence ordonnée $l_1, l_2, l_3, \ldots, l_n, l_{n+1}$ est dite $(n+1)$-collier pour un sous-ensemble $S \subset \mathcal{A}_4^2$ si chaque dinucleotide $l_1l_2, l_2l_3, \ldots, l_nl_{n+1}$ appartient à S.*

Theorem 57 (Michel et Pirillo, 2013) *Soit S un sous-ensemble de $\mathcal{A}_4^2$. Les conditions suivantes sont équivalentes:*

(i) S est un code circulaire.
(ii) S n'a pas de 5-collier.

Theorem 58 (Michel et Pirillo, 2013) *Soit (i, j, h, k) une permutation de (A, C, G, T). Si $S = \{ij, ih, ik, jh, jk, hk\}$ alors S est un code circulaire de dinucléotides.*

Proof. Par contradiction, suppose que S n'est pas un code circulaire de dinucléotides et soit l_1, l_2, l_3, l_4, l_5 un 5-collier de S. Note que, sauf pour la lettre l_1, les autres lettres l_2, l_3, l_4, l_5 composant le collier doivent être un suffixe d'un dinucléotide de S.

Claim 59 *Pour $\alpha \in \{2, 3, 4, 5\}$, $l_\alpha \neq i$.*

Proof of Claim 1. Par inspection, i n'est jamais un suffixe d'un dinucléotide de S.

Claim 60 *Pour $\alpha \in \{3, 4, 5\}$, $l_\alpha \neq j$.*

Proof of Claim 2. Par inspection, j est uniquement suffixe pour ij. Pour $\alpha \in \{3, 4, 5\}$, si $l_\alpha = j$ alors $l_{\alpha-1} = i$ qui est impossible avec Claim 1.

Claim 61 *Pour $\alpha \in \{4, 5\}$, $l_\alpha \neq h$.*

Proof of Claim 3. Par inspection, h est uniquement suffixe pour ih et jh. Par contradiction, suppose que $l_5 = h$. Alors, $l_4 = i$ ou $l_4 = j$. Si $l_4 = i$ alors il existe une contradiction avec Claim 1 et si $l_4 = j$ il existe une contradiction avec Claim 2. Par contradiction, suppose que $l_4 = h$. Alors, $l_3 = i$ ou $l_3 = j$. Si $l_3 = i$ alors il existe une contradiction avec Claim 1 et si $l_3 = j$ alors il existe une contradiction avec Claim 2.

Claim 62 *$l_5 \neq k$.*

Proof of Claim 4. Par inspection, k est uniquement suffixe pour ik, jk et hk. Par contradiction, suppose que $l_5 = k$. Alors, $l_4 = i$ ou $l_4 = j$ ou $l_4 = h$. Si $l_4 = i$ alors il existe une contradiction avec Claim 1, si $l_4 = j$ il existe une contradiction avec Claim 2 et si $l_4 = h$ alors il existe une contradiction avec Claim 3.

Avec Claims 1, 2, 3, 4, nous avons $l_5 \neq i, l_5 \neq j, l_5 \neq h, l_5 \neq k$ et donc, S n'a pas de 5-collier. En conséquence, S un code circulaire de dinucléotides.

Theorem 63 (Michel et Pirillo, 2013) *Il existe 24 codes circulaires de dinucléotides maximum de longueur* $l = 6$.

Proposition 64 (Michel et Pirillo, 2013) *La liste des 24 codes circulaires de dinucléotides maximum de longueur* $l = 6$ *est*

$\{AC, AG, AT, CG, CT, GT\}$, $\{AC, AG, AT, CG, CT, TG\}$, $\{AC, AG, AT, CG, TC, TG\}$,
$\{AC, AG, AT, CT, GC, GT\}$, $\{AC, AG, AT, GC, GT, TC\}$, $\{AC, AG, AT, GC, TC, TG\}$,
$\{AC, AG, CG, TA, TC, TG\}$, $\{AC, AG, GC, TA, TC, TG\}$, $\{AC, AT, CT, GA, GC, GT\}$,
$\{AC, AT, GA, GC, GT, TC\}$, $\{AC, GA, GC, GT, TA, TC\}$, $\{AC, GA, GC, TA, TC, TG\}$,
$\{AG, AT, CA, CG, CT, GT\}$, $\{AG, AT, CA, CG, CT, TG\}$, $\{AG, CA, CG, CT, TA, TG\}$,
$\{AG, CA, CG, TA, TC, TG\}$, $\{AT, CA, CG, CT, GA, GT\}$, $\{AT, CA, CT, GA, GC, GT\}$,
$\{CA, CG, CT, GA, GT, TA\}$, $\{CA, CG, CT, GA, TA, TG\}$, $\{CA, CG, GA, TA, TC, TG\}$,
$\{CA, CT, GA, GC, GT, TA\}$, $\{CA, GA, GC, GT, TA, TC\}$, $\{CA, GA, GC, TA, TC, TG\}$.

Proposition 65 (Michel et Pirillo, 2013) *Si* S *est un code circulaire de dinucléotides maximum de longueur* $l = 6$ *alors* $\mathcal{C}(S)$ *est également un code circulaire de dinucléotides maximum de longueur* $l = 6$.

Proposition 66 (Michel et Pirillo, 2013) *Si* S *est un code circulaire de dinucléotides maximum de longueur* $l = 6$ *alors* $\mathcal{P}(\mathcal{C}(S)) = \mathcal{C}(\mathcal{P}(S))$.

5 Algorithmique des codes circulaires

Un algorithme basé sur une matrice de colliers et parallélisant l'arbre lexicographique des codes permet de déterminer rapidement (5 heures) les fonctions de croissance des codes circulaires de trinucléotides et des codes circulaires de trinucléotides maximaux pour les longueurs $l = 1, ..., 20$ (Herrmann, Michel et Zugmeyer, 2013).

6 Théorie d'évolution des codes circulaires

Nous renvoyons le lecteur aux travaux de Arquès, Fallot et Michel (1997, 1998), Arquès, Fallot, Marsan et Michel (1999), Bahi et Michel (2004, 2008, 2009), Frey et Michel (2006) et de Michel (2007).

7 Théorie biologique des codes circulaires

Le code circulaire X (1) observé dans les gènes codants des procaryotes, eucaryotes, plasmides et virus (Michel, 2015) ainsi que l'identification de motifs du code circulaire X (motifs construits à partir du code X) dans le centre du centre de décodage du ribosome, autour du centre de décodage du ribosome, ainsi que dans les ARN de transfert ont conduit à l'hypothèse d'un code de translation des gènes basé sur le code circulaire X (Michel, 2012). Il est remarquable de souligner que les nucléotides A1492 et A1493 (dinucléotide AA) du centre de décodage du ribosome qui sont universellement conservés dans les procaryotes et eucaryotes et le nucléotide G530 du centre de décodage du ribosome qui est conservé dans les procaryotes appartiennent à des motifs du code circulaire X. Nous renvoyons le lecteur aux travaux de Michel (2012, 2013, 2014, 2015a), Michel et Seligmann (2014) et de El Soufi et Michel (2014, 2015).

8 Théorie des groupes des codes circulaires

Nous renvoyons le lecteur aux travaux de Fimmel, Giannerini, Gonzalez et Strüngmann (2014, 2015) et de Fimmel et Strüngmann (2015a, 2015b).

9 Théorie des graphes des codes circulaires

Soit $B = \{A, C, G, T\}$. Pour $n \in \mathbb{N}$ avec $n \geq 2$ un *n-nucléotides code* est un sous-ensemble $X \subseteq B^n$. La définition qui suit met en relation un graphe orienté avec un code quelconque à n-nucleotides. On rappelle qu'en théorie des graphes (Clark and Holton, 1991) un *graphe* $\mathcal{G}$ consiste en un ensemble fini de sommets V et en un ensemble fini d'arêtes E. Ici, une

arête est un ensembe $\{v, w\}$ de sommets de V. Le graphe est dit orienté si les arêtes ont une orientation, i.e. les arêtes sont considérées comme des paires ordonnées $[v, w]$.

Definition 67 (Fimmel, Michel, Strüngmann, 2016) *Soit $X \subseteq B^n$ un n-nucleotides code ($n \in \mathbb{N}$). On définit le graphe orienté $\mathcal{G}(X) = (V(X), E(X))$ avec l'ensemble des sommets $V(X)$ et l'ensemble des arêtes $E(X)$ de la façon suivante:*

- $V(X) = \{N_1...N_i, N_{i+1}...N_n : N_1N_2N_3...N_n \in X, 1 \leq i \leq n-1\}$
- $E(X) = \{[N_1...N_i, N_{i+1}...N_n] : N_1N_2N_3...N_n \in X, 1 \leq i \leq n-1\}$

Le graphe $\mathcal{G}(X)$ est dit graphe associé à X.

Fondamentalement, le graphe $\mathcal{G}(X)$ associé au code X interprète les mots à n-nucleotides de X en $(n-1)$ façons par paires de i-nucléotides et $(n-i)$-nucléotides pour $1 \leq i \leq n-1$.

Example 68 *Les trois Figures 1, 2 et 3 donnent des exemples de codes et leurs graphes associés pour $n = 2$ (dinucléotide code), $n = 3$ (trinucléotide code) et $n = 4$ (tétranucléotide code).*

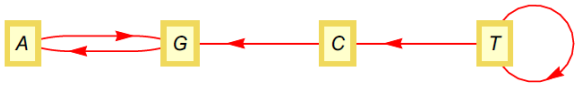


FIGURE 1. Graph representing the dinucleotide code $\{AG, CG, GA, TC, TT\}$

Le graphe du tétranucléotide code comporte quatre parties disjointes. Deux parties sont construites avec des sommets étiquetés par des dinucléotides et deux parties sont construites avec des sommets étiquetés par des nucléotides et trinucléotides. Ces parties sont appelées *composantes de $\mathcal{G}$*. On rappelle qu'un sous-ensemble V' de l'ensemble des sommets V est dit *connecté* si pour deux sommets quelconques $v, w \in V'$ il existe un chemin $[v, v_1][v_1, v_2] \cdots [v_{n-1}, v_n][v_n, w]$ de sommets de V' connectant v et w. Un graphe se décompose de façon unique en composantes connectées qui sont disjointes deux à deux. On rappelle également qu'un graphe est *biparti* si son ensemble de sommets V peut être decomposé en deux sous-ensembles disjoints V' et V'' tels que les arêtes de $\mathcal{G}$ connectent

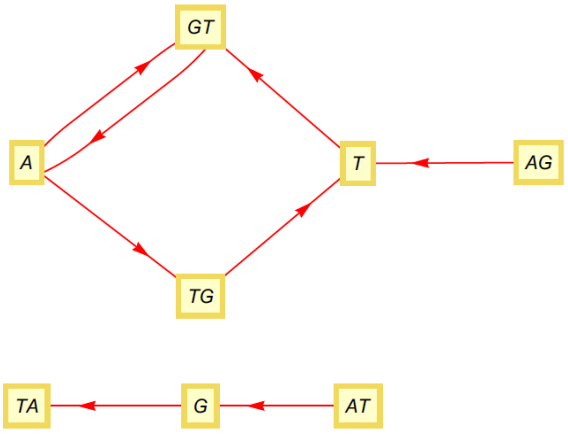


FIGURE 2. Graph representing the trinucleotide code $\{AGT, ATG, GTA, TGT\}$

uniquement les sommets de V' avec les sommets de V'' et vice versa. Evidemment, si X est un n-nucléotides code, alors les composantes de $\mathcal{G}(X)$ sont exactement les graphes

$$\mathcal{G}(X)_j = (V(X)_j, E(X)_j) \text{ pour } 1 \leq j \leq n-1$$

avec

$$V(X)_j = \{N_1...N_j, N_{j+1}...N_n, N_1...N_{n-j}, N_{n-j+1}...N_n : N_1N_2N_3...N_n \in X\}$$

et

$$E(X)_j = \{[N_1...N_j, N_{j+1}...N_n], [N_1...N_{n-j}, N_{n-j+1}...N_n] : N_1N_2N_3...N_n \in X\}.$$

Ces composantes ne sont pas nécessairement connectées comme on le constate avec la Figure 3, cependant, elles le sont souvent. En fait, $\mathcal{G}(X)_j$ comporte exactement les sommets (et leurs arêtes correspondantes) qui traduisent les éléments de X de deux façons: comme paire de j-nucléotide et $(n-j)$-nucléotide et comme paire de $(n-j)$-nucléotide et j-nucléotide. Remarquons que par symmétrie, nous avons $\mathcal{G}(X)_j = \mathcal{G}(X)_{n-j}$ pour tout $j < n-1$. Par exemple, dans la Figure 3, les deux composantes du graphe associé au tétranucléotide code sont $\mathcal{G}(X)_1$ $(= \mathcal{G}(X)_3)$ et $\mathcal{G}(X)_2$. L'observation qui suit est évidente.

Lemma 69 *Soit X un n-nucléotides code pour $n \in \mathbb{N}$. Les conditions suivantes existent:*

1. *Si n est impair, alors $\mathcal{G}(X)$ est un graphe biparti. En particulier, toutes ses composantes $\mathcal{G}(X)_j$ sont biparties.*

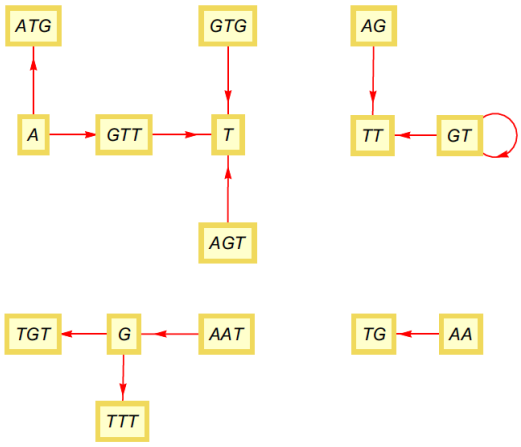


FIGURE 3. Graph representing the tetranucleotide code $\{AATG, AGTT, GTGT, GTTT\}$

2. *Si n est pair, alors toutes les composantes de $\mathcal{G}(X)$ sont biparties à l'exception peut-être pour $\mathcal{G}(X)_{\frac{n}{2}}$.*

Le graphe associé aux codes circulaires est toujours simple. On rappelle qu'en théorie des graphes (Clark and Holton, 1991) un graphe orienté est *simple* s'il ne contient pas de boucles internes, i.e. pas d'arêtes entre un sommet et lui-même et pas d'arêtes multiples avec la même orientation entre deux sommets. Pour un graphe orienté simple, on peut cependant avoir $[x, y] \in E(\mathcal{G})$ et $[y, x] \in E(\mathcal{G})$ ce qui signifie qu'il existe un cycle (cercle) de longueur 2. Cependant, pour les codes circulaires, cette structure est également exclue. On rappelle qu'un *cycle* dans $\mathcal{G}$ est un chemin orienté fermé dans $\mathcal{G}$. Un *cercle* est un cycle qui ne visite aucun sommet deux fois, sauf le sommet de départ (qui est également le sommet de terminaison en même temps). Par exemple, dans la Figure 2, la suite des sommets T, GT, A, TG, T est un cercle alors que la suite des sommets T, GT, A, GT, A, TG, T est un cycle qui n'est pas un cercle.

Lemma 70 (Fimmel, Michel, Strüngmann, 2016) *Soit $X \subseteq B^n$ un code circulaire. Alors son graphe associé est un graphe orienté simple sans cercle de longueur 2.*

Ce Lemme dit simplement que le graphe associé à un code circulaire possède un graphe non-orienté simple sous-jacent.

Example 71 *La Figure 4 donne deux exemples de trinucléotide codes et leurs graphes associés. Le code $\{ATG, CAC, CAT, GTG\}$ (à gauche) est circulaire avec son graphe asso-*

cié qui est simple alors que le code $\{ATG, ATT, TGA, TGT\}$ *(à droite) est non-circulaire avec son graphe associé qui n'est pas simple.*

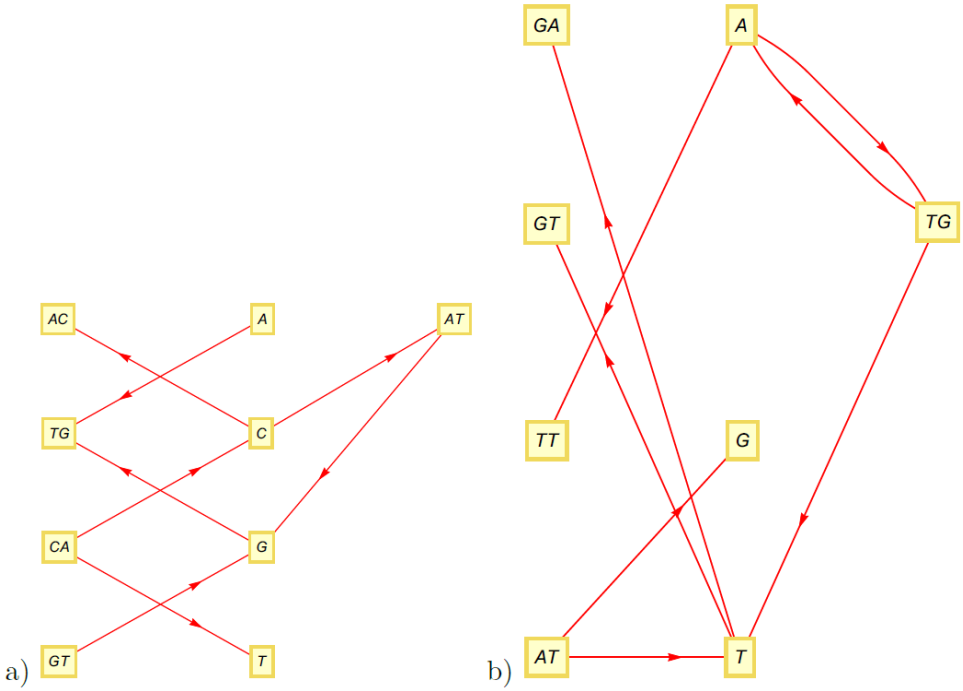


FIGURE 4. a) The trinucleotide code $\{ATG, CAC, CAT, GTG\}$ is circular and has a simple graph, b) The trinucleotide code $\{ATG, ATT, TGA, TGT\}$ is non-circular and its representing graph is not simple.

On rappelle qu'en théorie des graphes (Clark and Holton, 1991) un graphe est *acyclique* s'il ne contient pas de cycle, i.e. pas de chemin orienté fermé.

Theorem 72 (Fimmel, Michel, Strüngmann, 2016) *Soit un code* $X \subseteq B^n$. *Les conditions suivantes sont équivalentes:*
(i) X *est circulaire.*
(ii) $\mathcal{G}(X)$ *est acyclique.*

10 Conclusion

Les gènes seraient constitués de deux codes: (i) le code génétique universel et ses codes génétiques variants qui codent les trinucléotides des gènes en acides aminés des protéines,

et (ii) le code circulaire universel X (1) et ses codes circulaires variants (Michel, 2015) qui permettent de synchroniser et de retrouver automatiquement la phase de lecture des gènes. Les codes circulaires seraient des codes de translation pour les gènes primitifs, i.e. avant l'apparition des protéines, ou pour les gènes actuels sachant qu'à ce jour il n'existe aucune preuve biologique expérimentale d'une telle fontion de codage. Ils sont également des objets mathématiques-informatiques passionnants avec de nombreuses propriétés qui restent à découvrir.

11 Références

Ahmed A., Frey G., Michel C.J. (2007). Frameshift signals in genes associated with the circular code. In Silico Biology 7, 155-168.

Ahmed A., Frey G., Michel C.J. (2010). Essential molecular functions associated with the circular code evolution. Journal of Theoretical Biology 264, 613-622.

Ahmed A., Michel C.J. (2008). Plant microRNA detection using the circular code information. Computational Biology and Chemistry 32, 400-405.

Ahmed A., Michel C.J. (2011). Circular code signal in frameshift genes. Journal of Computer Science and Systems Biology 4, 7-15.

Arquès D.G., Fallot J.-P., Marsan, L., Michel C.J. (1999). An evolutionary analytical model of a complementary circular code. Biosystems 49, 83-103.

Arquès D.G., Fallot J.-P., Michel C.J. (1997). An evolutionary model of a complementary circular code. Journal of Theoretical Biology 185, 241-253.

Arquès D.G., Fallot J.-P., Michel C.J. (1998). An evolutionary analytical model of a complementary circular code simulating the protein coding genes, the 5' and 3' regions. Bulletin of Mathematical Biology 60, 163-194.

Arquès D.G., Lacan J., Michel C.J. (2002). Identification of protein coding genes in genomes with statistical functions based on the circular code. Biosystems 66, 73-92.

Arquès D.G., Michel C.J. (1987). Periodicities in introns. Nucleic Acids Research 15, 7581-7592.

Arquès D.G., Michel C.J. (1996). A complementary circular code in the protein coding

genes. Journal of Theoretical Biology 182, 45-58.

Arquès D.G., Michel C.J. (1997). A circular code in the protein coding genes of mitochondria. Journal of Theoretical Biology 189, 273-290.

Arquès D.G., Michel C.J. (1997). A code in the protein coding genes. Biosystems 44, 107-134.

Bahi J.M., Michel C.J. (2004). A stochastic gene evolution model with time dependent mutations. Bulletin of Mathematical Biology 66, 763-778.

Bahi J.M., Michel C.J. (2008). A stochastic model of gene evolution with chaotic mutations. Journal of Theoretical Biology 255, 53-63.

Bahi J.M., Michel C.J. (2009). A stochastic model of gene evolution with time dependent pseudochaotic mutations. Bulletin of Mathematical Biology 71, 681-700.

Benard E., Michel C.J. (2013). Transition and transversion on the common trinucleotide circular code. Computational Biology Journal 2013, Article ID 795418, 1-10.

Bussoli L., Michel C.J., Pirillo G. (2011). On some forbidden configurations for self-complementary trinucleotide circular codes. Journal for Algebra and Number Theory Academia 2, 223-232.

Bussoli L., Michel C.J., Pirillo G. (2012). On conjugation partitions of sets of trinucleotides. Applied Mathematics 3, 107-112.

Clark J., Holton D.A. (1991). A first look at graph theory. World Scientific, New Jersey.

Crick F.H.C., Griffith J.S., Orgel L.E. (1957). Codes without commas. Proceedings of the National Academy of Sciences U.S.A. 43, 416-421.

Crick F.H., Brenner S., Klug A., Pieczenik G. (1976). A speculation on the origin of protein synthesis. Origins of Life 7, 389-397.

Eigen M., Schuster P. (1978). The Hypercycle. A principle of natural self-organization. Part C: The realistic hypercycle. Naturwissenschaften 65, 341-369.

El Soufi K., Michel C.J. (2014). Circular code motifs in the ribosome decoding center. Computational Biology and Chemistry 52, 9-17.

El Soufi K., Michel C.J. (2015). Circular code motifs near the ribosome decoding

center. Computational Biology and Chemistry in press.

Fimmel E., Giannerini S., Gonzalez D.,Strüngmann L. (2014). Circular codes, symmetries and transformations. Journal of Mathematical Biology http://10.1007/s00285-014-0806-7.

Fimmel E., Giannerini S., Gonzalez D., Strüngmann L. (2015). Dinucleotide circular codes and bijective transformations. Journal of Theoretical Biology, in press.

Fimmel E., Michel C.J., Strüngmann L. (2016). n-nucleotide circular codes in graph theory. Philosophical Transactions A in press.

Fimmel E., Strüngmann L. (2015a). On the hierarchy of trinucleotide n-circular codes and their corresponding amino acids. Journal of Theoretical Biology 364, 113-120.

Fimmel E., Strüngmann L. (2015b). Maximal dinucleotide comma-free codes, submitted (personal communication).

Frey G., Michel C.J. (2003). Circular codes in archaeal genomes. Journal of Theoretical Biology 223, 413-431.

Frey G., Michel C.J. (2006). An analytical model of gene evolution with 6 mutation parameters: an application to archaeal circular codes. Computational Biology and Chemistry 30, 1-11.

Frey G., Michel C.J. (2006). Identification of circular codes in bacterial genomes and their use in a factorization method for retrieving the reading frames of genes. Computational Biology and Chemistry 30, 87-101.

Golomb S.W., Gordon B., Welch L.R. (1958a). Comma-free codes. Canadian Journal of Mathematics, 10, 202-209.

Golomb S.W., Welch L.R., Delbrück M. (1958b). Construction and properties of comma-free codes. Biologiske Meddelelser, Kongelige Danske Videnskabernes Selskab. 23.

Gonzalez D.L., Giannerini S., Rosa R. (2011). Circular codes revisited: a statistical approach. Journal of Theoretical Biology 275, 21-28.

Herrmann M., Michel C.J., Zugmeyer B. (2013). A necklace algorithm to determine the growth function of trinucleotide circular codes. Journal of Applied Mathematics and Bioinformatics 3, 1-40.

Koch A.J., Lehmann J. (1997). About a symmetry of the genetic code. Journal of Theoretical Biology 189, 171-174.

Konopka A.K., Smythers G.W. (1987). DISTAN - A program which detects significant distances between short oligonucleotides. Bioinformatics 3, 193-201.

Lacan J., Michel C.J. (2001). Analysis of a circular code model. Journal of Theoretical Biology 213, 159-170.

Lassez J.L. (1976). Circular codes and synchronization. International Journal of Computer and Information Sciences 5, 201-208.

Lassez J.L., Rossi R.A., Bernal A.E. Crick's hypothesis revisited: the existence of a universal coding frame," in Proceedings of the 21st International Conference on Advanced Information Networking and Applications Workshops (AINAW'07), pp. 745–751, Niagara Falls, Canada, May 2007.

Michel C.J. (2007). An analytical model of gene evolution with 9 mutation parameters: an application to the amino acids coded by the common circular code. Bulletin of Mathematical Biology 69, 677-698.

Michel C.J. (2008). A 2006 review of circular codes in genes. Computer and Mathematics with Applications 55, 984-988.

Michel C.J. (2012). Circular code motifs in transfer and 16S ribosomal RNAs: a possible translation code in genes. Computational Biology and Chemistry 37, 24-37.

Michel C.J. (2013). Circular code motifs in transfer RNAs. Computational Biology and Chemistry 45, 17-29.

Michel C.J. (2014). A genetic scale of reading frame coding. Journal of Theoretical Biology 355, 83-94.

Michel C.J. (2015). An extended genetic scale of reading frame coding. Journal of Theoretical Biology 365, 164-174.

Michel C.J. (2015). The maximal C^3 self-complementary trinucleotide circular code X in genes of bacteria, eukaryotes, plasmids and viruses. Journal of Theoretical Biology 380, 156-177.

Michel C.J., Pellegrini M., Pirillo G. (2015). Maximal dinucleotide and trinucleotide circular codes. Journal of Theoretical Biology, in press.

Michel C.J., Pirillo G, Pirillo M.A. (2008). A relation between trinucleotide comma-free codes and trinucleotide circular codes. Theoretical Computer Science 401, 17-26.

Michel C.J., Pirillo G, Pirillo M.A. (2008). Varieties of comma free codes. Computer and Mathematics with Applications 55, 989-996.

Michel C.J., Pirillo G. (2010). Identification of all trinucleotide circular codes. Computational Biology and Chemistry 34, 122-125.

Michel C.J., Pirillo G. (2011). Strong trinucleotide circular codes. International Journal of Combinatorics 2011, Article ID 659567, 1-14.

Michel C.J., Pirillo G. (2013). A permuted set of a trinucleotide circular code coding the 20 amino acids in variant nuclear codes. Journal of Theoretical Biology 319, 116-121.

Michel C.J., Pirillo G. (2013). Dinucleotide circular codes. ISRN Biomathematics 2013, Article ID 538631, 1-8.

Michel C.J., Pirillo G., Pirillo M.A. (2012). A classification of 20-trinucleotide circular codes. Information and Computation 212, 55-63.

Michel C.J., Seligmann H. (2014). Bijective transformation circular codes and nucleotide exchanging RNA transcription. Biosystems 118, 39-50.

Nirenberg M.W., Matthaei J.H. (1961). The dependence of cell-free protein synthesis in *E. coli* upon naturally occurring or synthetic polyribonucleotides. Proceedings of the National Academy of Sciences U.S.A. 47, 1588-1602.

Pirillo G. (2003). A characterization for a set of trinucleotides to be a circular code. In: Pellegrini C., Cerrai P., Freguglia P., Benci V., Israel G. (Eds.). Determinism, Holism and Complexity. Kluwer Academic Publisher, NewYork, NY, USA.

Shepherd J.C.W. (1981). Method to determine the reading frame of a protein from the purine/pyrimidine genome sequence and its possible evolutionary justification. Proceedings of the National Academy of Sciences U.S.A. 78, 1596-1600.