Distributions de longueurs des codes circulaires

Frédérique Bassino
I.G.M., Université de Marne La Vallée
2, rue de la butte verte 93166 Noisy le Grand Cedex
e-mail: bassino@univ-mlv.fr

Résumé

On décrit, en fonction de différents paramètres, la série génératrice de l'étoile d'un code circulaire. On étend à un alphabet "pondéré" la caractérisation des distributions de longueurs des codes circulaires établie pour un alphabet fini par Schützenberger. On donne également, dans ce cadre, une nouvelle caractérisation de ces distributions. Ce résultat met en évidence la décidabilité dans le cas d'une distribution finie.

Astract

We describe, in terms of different parameters, the generating series of the star of a circular code. We extend the characterization of length distributions of circular codes established for a finite alphabet by Schützenberger to a "weighted" alphabet. We also give another characterization of these distributions. This result underscores the decidability in the case of a finite distribution.

Introduction

Les codes sont les bases des sous-monoïdes libres, ce sont les ensembles de mots qui conduisent à un décodage unique. On étudie ici une famille particulière de codes appelés codes circulaires. Leur spécificité tient au fait qu'ils définissent une unique factorisation des mots écrits sur un cercle.

Ils ont été introduits par Golomb et Gordon ([GG65]) en raison de leur forte propriété de synchronisation. Ainsi, tout code circulaire coupant est synchronisant et tout code uniformément synchronisant est circulaire. Pour cette raison, ils jouent un rôle important dans la correction d'erreurs.

Un type particulier de codes circulaires, les codes comma-free ([GGW58]), ont été l'objet de nombreuses investigations. Ils ont de nombreuses applications, en particulier, à l'analyse des séquences d'ADN. Récemment, par une étude statistique de la répartition des trinucléotides dans les gènes codant les protéines, D. Arquès et C. Michel ([AM95]) ont identifié un ensemble de codons constituant un code circulaire.

Les codes circulaires apparaissent dans de multiples problèmes de combinatoire des mots. Par exemple, pour calculer le nombre de mots circulaires sur l'alphabet $A = \{a, b\}$ ne comportant pas deux occurrences consécutives de la lettre b, on peut utiliser le codage

$$\begin{aligned} \alpha : \quad & A^* \to B^* = \{a, ba\}^* \\ & a \to a \\ & b \to ba. \end{aligned}$$

Le coefficient f_n d'ordre n de la série génératrice du sous-monoïde engendré par le code circulaire $\{a, ba\}$

$$\sum_{n \geq 0} f_n z^n = \frac{1}{1 - z - z^2}$$

est alors égal au nombre de mots de longueur n ($n \geq 2$) n'ayant pas bb comme facteur. D'autres exemples de problèmes combinatoires résolus à l'aide des codes circulaires sont présentés dans [Sta86] (Section 4.7). Enfin, les codes circulaires sont liés aux factorisations finies du monoïde libre, tous les facteurs intervenant dans une telle factorisation sont circulaires ([Sch65a], [Vie78], [Kro87]).

Dans ce qui suit, on prouve trois nouveaux résultats concernant les codes circulaires. On décrit, en fonction de différents paramètres, la série génératrice de l'étoile d'un code circulaire. On généralise, dans plusieurs directions, la caractérisation des distributions de longueurs des codes circulaires établie dans le cas d'un alphabet fini par Schützenberger ([Sch65b]). D'une part, on remplace l'alphabet fini par un alphabet quelconque dont les éléments ont des poids, ce qui permet d'étendre le résultat à deux distributions de longueurs. D'autre part, on restreint les conditions, ce qui permet d'établir la décidabilité dans le cas d'une distribution finie. On donne une nouvelle formulation de cette caractérisation. Ce résultat, établi par des méthodes combinatoires, donne un algorithme de décision dans le cas d'une distribution finie.

La première partie de cet article est consacrée au rappel de définitions relatifs aux codes circulaires et à la relation de conjugaison ([BP85]). On introduit également la notion d'alphabet pondéré, qui n'est autre qu'un ensemble de symboles auxquels sont associés des poids entiers. Shannon fut l'un des premiers à s'intéresser aux alphabets dont les lettres ont des coûts quelconques, ces derniers représentent dans ce cas la durée des symboles de transmission. Ces alphabets furent utilisés par la suite en théorie de l'information afin de généraliser la notion d'entropie qu'il avait proposée. Selon les applications, les coûts sont des paramètres liés à l'amplitude, la durée ou l'énergie d'un signal.

Dans la deuxième partie, on étudie la série génératrice du sous-monoïde engendré par un code circulaire et on établit des variantes de la formule de Witt et de l'identité cyclotomique. Dans la partie suivante, on étend (Théorème 2) la caractérisation des distributions de longueurs des codes circulaires établie par Schützenberger. On en donne enfin une nouvelle formulation, qui fait l'objet du Théorème 3, mettant ainsi en évidence la décidabilité de ce problème dans le cas d'une suite finie.

1 Préliminaires

On définit un *alphabet pondéré* comme un couple (A, λ) formé d'un alphabet A et d'une application $\lambda : A \to \mathbb{N}^*$. Celle-ci peut être interprétée comme un coût aussi bien que comme une distance ou une longueur. On étend la fonction λ aux mots de A^* de la façon suivante

$$\begin{array}{rccl} \lambda: & A^* & \to & \mathbb{N}^* \\ & x & \to & \lambda(x) = \sum_{a \in A} \lambda(a)|x|_a \end{array}$$

où $|x|_a$ désigne le nombre d'occurrences de la lettre a dans le mot x. L'entier $\lambda(x)$ est alors appelé la *longueur* du mot x de (A^*, λ) .

Un alphabet ordinaire A peut être vu comme un alphabet pondéré pour lequel la fonction λ vaut identiquement 1 sur A, on notera dans ce cas $|w|$ au lieu de $\lambda(w)$ la longueur d'un mot w. Les définitions et résultats établis sur un alphabet ordinaire pour lesquels la notion de longueur n'intervient pas restent inchangés sur un alphabet pondéré.

Un sous-ensemble X du monoïde libre A^* est un *code* sur A si, et seulement si, pour tous entiers $n, m \geq 1$ et $x_1, x_2, \ldots, x_n, x'_1, x'_2, \ldots, x'_m \in X$

$$x_1 x_2 \cdots x_n = x'_1 x'_2 \cdots x'_m \quad \Rightarrow \quad n = m \quad \text{et} \quad x_i = x'_i \quad \text{pour} \quad i = 1, \ldots, n.$$

Une partie X de A^* est donc un code si tout mot de X^+ se factorise de manière unique en mots de X.

Une partie $X \subset A^+$ est un *code circulaire* si, pour tous entiers $n, m \geq 1$ et $x_1, x_2, \ldots, x_n \in X$, $y_1, y_2, \ldots, y_m \in X$ et $p \in A^*$ et $s \in A^+$, les égalités

$$s x_2 x_3 \cdots x_n p = y_1 y_2 \cdots y_m, \qquad x_1 = ps$$

impliquent

$$n = m,\ p = 1 \text{ et } \forall i,\ x_i = y_i.$$

En d'autres termes, tout mot de X^+ écrit sur un cercle (Figure 1) admet une unique décomposition en mots de X.

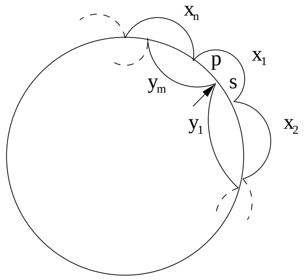

FIG. 1 – *Deux factorisations circulaires*

Un code circulaire est un code et toute partie d'un code circulaire est encore un code circulaire.

Dans ce qui suit, on introduit la notion de conjugaison dans un monoïde. Cette relation d'équivalence permet de caractériser les codes circulaires. Deux mots x et y de A^* sont *conjugués* s'il existe des mots u et v tels que $x = uv$ et $y = vu$. Les classes de cette relation d'équivalence sont appelée les *classes de conjugaison* (et parfois *mots circulaires* ou *colliers* cf. [Reu88] et [MR83]).

Un mot $x \in A^*$ est *primitif* s'il n'est puissance d'aucun autre mot : x est primitif si, et seulement si, $(x = y^n \quad \text{et} \quad n \geq 0) \quad \Rightarrow \quad x = y$. Le mot vide n'est pas primitif.

Remarque 1 Un code circulaire X ne peut contenir deux mots conjugués distincts. De plus, tout mot de X est primitif.

Soit X un code. Deux mots w et w' de X^* sont dits *X-conjugués* s'il existe $x, y \in X^*$ tels que $w = xy$ et $w' = yx$. On appelle *classes de X-conjugaison* les classes de cette relation d'équivalence. Un mot $x \in X^*$ est dit *X-primitif* s'il n'est puissance d'aucun mot de X^*.

Remarque 2 L'alphabet étant lui-même un code circulaire, la relation de conjugaison peut être vue comme celle de la A-conjugaison et la notion de primitivité comme celle de A-primitivité.

Deux mots de X^* qui sont X-conjugués sont conjugués. De même, un mot de X^* qui est primitif est également X-primitif. Quand X est un code circulaire, deux mots de X^* sont X-conjugués si, et seulement si, ils sont conjugués. De même, un mot de X^* est primitif si, et seulement si, il est également X-primitif.

On peut maintenant donner une caractérisation des codes circulaires basée sur la notion de conjugaison (cf. [BP85] Proposition 1.1 p.323). Soit $X \subset A^+$ un code, X est un code circulaire si, et seulement si, X^* est un sous-monoïde *très pur* de A^*, *i.e.*,

$$\forall u, v \in A^*, \quad uv, vu \in X^* \qquad \Rightarrow \qquad u, v \in X^*.$$

Exemple 1 Soient $A = \{a, b\}$ et $X = a^*b$, alors $X^* = A^*b \cup \{1\}$. Par suite, si uv et vu sont dans X^*, chacun des mots u et v est le mot vide ou se termine par un b, il appartient donc à X^*. On en déduit que X^* est très pur, et que X est un code circulaire.

Soient $A = \{a, b, c, d\}$ et $X = \{ab, cd, bc, da\}$, les mots $abcd$ et $bcda$ sont dans X^*, alors que ni a, ni bcd ne sont des mots du code ; par suite, X n'est pas circulaire.

2 Sous-monoïdes très pur

Soient (A, λ) un alphabet pondéré et X une partie de A^*. Pour tout entier positif n, on note A_n l'ensemble des mots de longueur n de A^*, soit

$$A_n = \{x \in A^* \mid \lambda(x) = n\},$$

on note α_n le nombre de mots de longueur n de X, soit

$$\alpha_n = \text{Card}\{x \in X \mid \lambda(x) = n\} = \text{Card}\,(X \cap A_n)$$

et f_X la série génératrice associée à X et définie par : $f_X(z) = \sum_{n \geq 0} \alpha_n z^n$.

Si A est un alphabet ordinaire, l'entier α_n est alors le nombre de mots de X qui ont n lettres. De manière générale, la suite $\alpha = (\alpha_n)_{n \geq 1}$ est appelée la *distribution de longueurs* de X.

Etant donné un code X sur un alphabet pondéré (A, λ), on désigne par $L_n(X)$ le nombre de classes de X-conjugaison de mots X-primitifs dans $X^* \cap A_n$.

Proposition 1 *Soit A un alphabet pondéré et soit $X \subset A^+$ un code circulaire sur A. Alors, pour tout entier $n \geq 1$,*

$$L_n(X) \quad \leq \quad L_n(A).$$

et l'égalité est atteinte si, et seulement si, toute classe de conjugaison de mots primitifs de A_n a une intersection non vide avec X^*.

Ce résultat est une extension d'un résultat connu dans le cas des codes circulaires sur un alphabet ordinaire (cf. [BP85] Proposition 3.2 p.338). De plus, d'après un résultat dû à Golomb et Gordon ([GG65]), quand X est un code sur un alphabet à k lettres, le nombre $L_n(X)$ ne dépend que de la distribution de longueurs du code. Cette propriété est encore vraie quand X est un code sur un alphabet pondéré quelconque.

Soit $(\alpha_n)_{n\geq 1}$ la distribution de longueurs de X. On note $\alpha_n^{(i)}$ le nombre de mots de X^i de longueur n ; comme X est un code, $\alpha_n^{(i)}$ est le coefficient d'ordre n de la série $(\sum_{j\geq 1} \alpha_j z^j)^i$. On pose

$$\forall n \geq 1, \qquad s_n(\alpha) = \sum_{i=1}^{n} \frac{n}{i} \alpha_n^{(i)},$$

l'entier $s_n(\alpha)$ peut être interprété comme le nombre de mots de longueur n dont un conjugué est dans X^*. On a alors

$$\forall n \geq 1, \qquad L_n(X) = l_n(\alpha) \quad \text{avec} \quad l_n(\alpha) = \frac{1}{n} \sum_{d/n} \mu(\frac{n}{d}) s_d(\alpha),$$

où μ désigne la fonction de Möbius.

On décrit maintenant, en fonction de différents paramètres, la série génératrice d'un sous-monoïde très pur de A^*.

Théorème 1 *Soient X un code circulaire sur un alphabet pondéré et $f_X(z) = \sum \alpha_n z^n$ sa série génératrice. Dans ces conditions, la série génératrice du sous-monoïde très pur X^* engendré par X peut être décrite par les différents membres de l'égalité suivante*

$$f_{X^*}(z) = \exp\left(\sum_{n\geq 1} \frac{s_n(\alpha)}{n} z^n\right) = \frac{1}{\prod_{n\geq 1}(1-z^n)^{l_n(\alpha)}} = \frac{1}{\prod_{n\geq 1}(1-a_n(\alpha) z^n)}$$

où - $s_n(\alpha)$ est le nombre de mots de longueur n dont un conjugué appartient à X^*,

- $l_n(\alpha)$ est le nombre de classes de conjugaison de mots X-primitifs de $X^* \cap A_n$ et

- $(a_n(\alpha))_{n\geq 1}$ est la distribution de longueurs d'un alphabet pondéré A' engendrant un monoïde commutatif ayant pour série génératrice $f_{X^*}(z)$.

3 Extension d'un théorème de Schützenberger

Les distributions de longueurs $(\alpha_n)_{n\geq 1}$ des codes sur un alphabet à k lettres sont caractérisées par l'inégalité de Kraft-McMillan : $\sum_{n\geq 1} \alpha_n k^{-n} \leq 1$. Schützenberger a identifié par des conditions plus fortes celles de ces suites qui sont distributions de longueurs d'un code circulaire ([Sch65b]). Dans ce qui suit, on étend cette caractérisation dans deux directions. D'une part, on remplace l'alphabet fini par un alphabet pondéré quelconque, ce qui permet également d'étendre le résultat au cas de deux distributions de longueurs (Corollaire 2) ; d'autre part, on réduit l'ensemble des valeurs pour lesquelles les inégalités considérées doivent être vérifiées, ce qui permet d'établir la décidabilité dans le cas d'une distribution de longueurs finie (Corollaire 1).

L'idée est de substituer, dans l'énoncé du théorème de Schützenberger, à l'alphabet à k lettres un alphabet pondéré. L'analogie provient du fait que les propriétés de la relation de conjugaison sont conservées ; on étend ainsi sans difficulté la preuve initiale. De plus, on réduit l'ensemble des valeurs pour lesquelles les inégalités sur les classes de mots primitifs doivent être vérifiées.

Théorème 2 *Soit (A, λ) un alphabet pondéré et soit $\alpha = (\alpha_n)_{n\geq 1}$ sa distribution de longueurs. Une suite d'entiers positifs $\beta = (\beta_n)_{n\geq 1}$ est alors la distribution de longueurs d'un code circulaire $Y \subset A^+$ si, et seulement si,*

$$\forall n \geq 1 \quad \textit{tel que} \quad \beta_n \neq 0 \qquad l_n(\beta) \leq l_n(\alpha)$$

On appellera dorénavant *inégalités circulaires* ces inéaglités. On peut tout de suite noter le corollaire suivant qui montre que, lorsque la distribution de longueurs β est une suite finie, on peut décider si les inégalités circulaires sont satisfaites.

Corollaire 1 *Une suite finie $\beta = (\beta_i)_{1\leq i\leq m}$ d'entiers positifs est la distribution de longueurs d'un code circulaire sur l'alphabet (A, λ) de distribution de longueurs $\alpha = (\alpha_n)_{n\geq 1}$ si, et seulement si,*

$$\forall n \leq m \quad \textit{tel que} \quad \beta_n \neq 0 \qquad l_n(\beta) \leq l_n(\alpha).$$

On peut donc décider si une suite finie est la distribution de longueurs d'un code circulaire construit sur un alphabet pondéré donné.

Remarque 3 Le fait de ne vérifier les inégalités circulaires que jusqu'au rang m dans le cas d'une suite $\beta = (\beta)_{1\leq i\leq m}$ n'a rien d'évident *a priori*. En effet, le mot le plus court ayant plus d'une factorisation circulaire sur un code peut être plus long que les mots du code. Par exemple, si $X = \{ab, cd, bc, da\}$, le mot circulaire ambigu le plus court, $abcd$ qui se factorise circulairement en $ab.cd$ et $bc.da$, est de longueur 4.

Construction Pour obtenir le code Y circulaire de distribution de longueurs $(\beta_n)_{n\geq 1}$ du Théorème 2, on construit une suite de codes circulaires préfixes $(Y_i)_{i\geq 0}$ à partir de l'alphabet pondéré (A, λ) de distribution de longueurs $(\alpha_n)_{n\geq 1}$. Le code circulaire que l'on obtient finalement coïncide, pour tout entier i, sur l'ensemble des mots de longueur inférieure ou égale à i, avec Y_i.

La construction des codes $(Y_i)_{i\geq 0}$ est la suivante : $Y_0 = A$ et pour tout i, si $(y_{i,j})_{j\geq 1}$ est la distribution de longueurs du code Y_i, le code Y_{i+1} est obtenu en supprimant successivement

$$\epsilon_i = y_{i-1,i} - \beta_i = l_i(\alpha) - l_i(\beta)$$

mots de longueur i au code Y_{i-1} : $x_{i,1}, x_{i,2}, \ldots, x_{i,\epsilon_i}$ et en construisant les codes auxiliaires

$X_{i,1} = x^*_{i,1}(Y_{i-1} \setminus x_{i,1})$, $X_{i,2} = x^*_{i,2}(X_{i,1} \setminus x_{i,2})$, …, $Y_i = X_{i,\epsilon_i} = x^*_{i,\epsilon_i}(X_{i,\epsilon_i-1} \setminus x_{i,\epsilon_i})$.

Si $\beta_i = 0$ et si $l_n(\beta) \leq l_n(\alpha)$ pour $1 \leq n \leq i-1$, on a alors nécessairement l'inégalité $l_i(\beta) \leq l_i(\alpha)$, ce qui justifie la réduction de l'intervalle pour lequel les inégalités circulaires doivent être vérifiées et assure la décidabilité dans le cas où la suite β est finie.

Exemple 2 Soient $k = 3$ et $A = \{a, b, c\}$. Si $\beta_1 = 1$, $\beta_2 = 1$, $\beta_3 = 7$, $\beta_4 = 10$ et $\beta_i = 0$ pour $i \geq 5$, alors $\epsilon_1 = l_1(3) - l_1(\beta) = 2$, en supprimant a puis b, on obtient le code

$$Y_1 = \{c, \mid ab, ac, bc \mid, a^2b, a^2c, bab, bac, b^2c, a^3b, a^3c, ba^2b, ba^2c, b^3c, \ldots\}.$$

Comme $\epsilon_2 = 2$, en supprimant ab puis ac, on obtient le code

$$\begin{aligned} Y_2 &= \{c, bc, \mid a^2b, a^2c, bab, bac, b^2c, abc, acc \mid, a^3b, a^3c, \\ &\quad ba^2b, ba^2c, b^2ab, b^2ac, b^3c, abac, abbc, acbc, \ldots\} \end{aligned}$$

Enfin, comme $\epsilon_3 = 0$ et $\epsilon_4 = 0$, $Y_2 = Y_3 = Y_4$. On obtient le code circulaire Y :

$$\begin{aligned} Y = Y_4 \cap A^4 &= \{c, bc, a^2b, a^2c, bab, bac, b^2c, abc, acc, \mid a^3b, a^3c, \\ &\quad ba^2b, ba^2c, b^2ab, b^2ac, b^3c, abac, abbc, acbc\} \end{aligned}$$

Le résultat suivant se déduit du précédent par morphisme codant.

Corollaire 2 *Soit X un code circulaire sur l'alphabet (ordinaire) A et soit $\alpha = (\alpha_n)_{n\geq 1}$ sa distribution de longueurs. Une suite d'entiers positifs $\beta = (\beta_n)_{n\geq 1}$ est alors la distribution de longueurs d'un code circulaire $Y \subset X^+$ si, et seulement si,*

$$\forall n \geq 1 \quad \textit{tel que} \quad \beta_n \neq 0 \qquad l_n(\beta) \leq l_n(\alpha)$$

4 Nouvelle formulation des inégalités circulaires

Dans cette partie, on montre que les distributions de longueurs $(\beta_n)_{n\geq 1}$ des codes circulaires sur un alphabet pondéré de distribution de longueurs $(\alpha_n)_{n\geq 1}$ peuvent être caractérisées par des inégalité portant directement sur les β_n. Ce résultat met en évidence le fait que si un code circulaire est fini, les inégalités circulaires sont nécessairement vérifiées à partir du rang strictement supérieur au plus long mot du code.

On déduit de la construction précédente une expression à coefficients positifs de $\beta_i + \epsilon_i$ en fonction des $(\alpha_j)_{j\leq i}, (\beta_j)_{j<i}$ et $(\epsilon_j)_{j<i}$, où $\epsilon_i = l_i(\alpha) - l_i(\beta)$.

Proposition 2 *Soit A un alphabet pondéré de distribution de longueurs $\alpha = (\alpha_i)_{i\geq 1}$ et soit $\beta = (\beta_i)_{i\geq 1}$ la distribution de longueurs d'un code circulaire sur A. Alors, pour tout entier positif n, la somme $\epsilon_n + \beta_n$ est un polynôme à coefficients entiers positif en les variables $(\beta_i)_{i<n}$, $(\epsilon_i)_{i<n}$ (où $\epsilon_i = l_i(\alpha) - l_i(\beta)$) et $(\alpha_i)_{i\leq n}$.*

Le Théorème 3 est une conséquence directe de la Proposition 2 et du Théorème 2.

Théorème 3 *Soit (A, λ) un alphabet pondéré de distribution de longueurs $\alpha = (\alpha_i)_{i\geq 1}$ et soit $\beta = (\beta_i)_{i\geq 1}$ une suite d'entiers positifs. Alors, il existe une suite $(Q_n)_{n\geq 1}$ de polynômes à coefficients entiers positifs, en les variables $(\beta_i)_{i<n}$, $(\epsilon_i)_{i<n}$ et $(\alpha_i)_{i\leq n}$, telle que les conditions suivantes soient équivalentes*

1. *la suite $\beta = (\beta_i)_{i\geq 1}$ est la distribution de longueurs d'un code circulaire sur A.*

2. *$\forall n \geq 1, \qquad \beta_n \leq Q_n$.*

Soit $(\alpha_n)_{n\geq 1}$ la distribution de longueurs de l'alphabet pondéré (A,λ). Le calcul des sommes $\beta_i+\epsilon_i$ donne alors, pour les premières valeurs de i, les polynômes suivants

$$Q_1=\beta_1+\epsilon_1=\alpha_1, \qquad Q_2=\epsilon_1\beta_1+\binom{\epsilon_1}{2}+\alpha_2,$$

$$Q_3=\left(\binom{\epsilon_1+1}{2}+\epsilon_2\right)\beta_1+2\binom{\epsilon_1+1}{3}+\epsilon_1\alpha_2+\alpha_3.$$

et

$$Q_4=\left(\binom{\epsilon_1+2}{3}+\epsilon_3\right)\beta_1+\epsilon_2\beta_2+3\binom{\epsilon_1+2}{4}+\binom{\epsilon_2}{2}+\binom{\epsilon_1+1}{2}\alpha_2+\epsilon_1\alpha_3+\alpha_4.$$

On donne maintenant une preuve combinatoire de la Proposition 2 dans le cas d'un alphabet ordinaire fini, puis dans le cas d'un alphabet pondéré. Par construction (p.6), le nombre de mots de longueur n du code Y_{n-1} est égal à $\beta_n+\epsilon_n$. En étudiant la manière dont ces mots sont engendrés, on donne une autre expression de cette somme qui correspond au polynôme Q_n cherché.

Cas d'un alphabet ordinaire à k lettres Pour compter le nombre de mots de longueur n du code Y_{n-1}, on en distingue deux types distincts :

1. les produits de mots qui ont été supprimés à une étape de la construction :

   $$w=w_pw_{p-1}\cdots w_1$$

   où les w_i $(1\leq i\leq p)$ sont des mots qui ont été supprimés précédemment

2. de tels mots concaténés avec un mot qui appartient lui-même au code Y_{n-1} :

   $$w=w_pw_{p-1}\cdots w_1w_0, \qquad w_0\in Y_{n-1}$$

   où les w_i $(1\leq i\leq p)$ sont des mots qui ont été supprimés.

Comme la construction est faite par suppressions successives de mots de longueur croissante, l'ordre dans lequel ils ont été supprimés induit un ordre total sur l'ensemble de ces mots (le premier mot supprimé étant le plus petit) et conditionne la structure des concaténations. Par suite, les produits, préfixes d'un mot du code ou eux-mêmes éléments du code, se décomposent de manière unique en une suite décroissante (pour cet ordre) de mots qui ont été supprimés :

$$\forall i, \qquad |w_{i+1}|\geq|w_i|.$$

Cependant, quand le produit appartient lui-même au code, *i.e.*, si $w=w_pw_{p-1}\cdots w_1$ alors $|w_1|=|w_2|$ et w_1 a été supprimé après w_2.

1. Plus précisément, les mots de longueur n de la forme $w=w_pw_{p-1}\cdots w_1$, où les w_i $(1\leq i\leq p)$, sont des mots qui ont été supprimés précédemment (et sont, par conséquent, de longueur strictement inférieure à n) vérifient les conditions suivantes sur les longueurs :

- soit $|w_{i+1}|=|w_i|$,

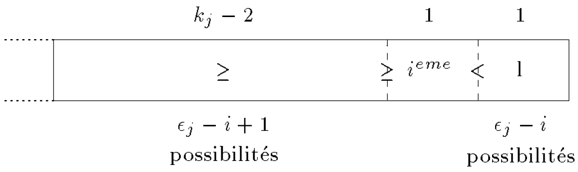

FIG. 2 – *Suffixes formés à partir des ϵ_j mots supprimés de longueur j*

– soit $|w_{i+1}| > |w_i|$ et $|w_{i+1}| < |w_i w_{i-1} \dots w_1|$.
En effet, dans le cas contraire (*i.e.* si $|w_{i+1}| \geq |w_i w_{i-1} \dots w_1|$), $W = w_i w_{i-1} \dots w_1$ est lui-même, soit un mot du code et w est alors un mot du second type distingué, soit un mot que l'on a supprimé et w doit alors nécessairement être factorisé sous la forme $w_p \cdots w_{i+1} W$ où $|w_{i+1}| = |W|$.

On calcule d'abord le nombre de produits de k_j éléments de même longueur qui peuvent apparaître dans un tel mot. On peut noter que les éléments intervenant dans un tel produit peuvent être répétés, mais leur ordre est déterminé de manière unique.

Si le facteur associé n'est pas en position de suffixe (*i.e.*, $|w_1| \neq j$), ce nombre est égal au nombre de combinaisons avec répétitions de k_j éléments pris parmi les ϵ_j éléments supprimés de longueur j soit $\binom{\epsilon_j + k_j - 1}{k_j} = P_{k_j}$.

Quand ce produit est en position de suffixe (*i.e.*, $|w_1| = j$), k_j est nécessairement supérieur ou égal à 2 et le dernier élément du produit est strictement plus grand (pour l'ordre de suppression) que le précédent (Figure 2).

Soit i le rang de l'avant dernier élément du produit, alors le nombre de préfixes associés de longueur $k_j - 2$ est égal au nombre de combinaisons avec répétition de $k_j - 2$ éléments pris parmi $\epsilon_j - i + 1$, soit $\binom{\epsilon_j + k_j - i - 2}{k_j - 2}$ et le nombre de valeurs que peut prendre la dernière lettre est égal à $\epsilon_j - i$. Le nombre de produits de k_j éléments de même longueur en position suffixe est donc égal à

$$T_{k_j} = \sum_{i=1}^{\epsilon_j - 1} \binom{\epsilon_j + k_j - i - 2}{k_j - 2} (\epsilon_j - i) \binom{\epsilon_j + k_j - i - 2}{k_j - 1} = (k_j - 1) \binom{\epsilon_j + k_j - 2}{k_j}.$$

Par suite, le nombre de mots de longueur n de Y_{n-1} obtenus par concaténation de mots supprimés est

$$R_n = \sum_{\mathcal{P}} \prod_{j_1 < j_2 < \dots < j_p} P_{k_{j_p}} \dots P_{k_{j_2}} T_{k_{j_1}} \quad \text{avec} \quad \left\{ \begin{array}{l} T_{k_j} = (k_j - 1)\binom{\epsilon_j + k_j - 2}{k_j} \quad \text{et} \quad k_j > 1 \\ P_{k_j} = \binom{\epsilon_j + k_j - 1}{k_j} \end{array} \right.$$

où la somme est prise sur l'ensemble $\mathcal{P}$ des p-uplets vérifiant les deux conditions suivantes (la deuxième condition exprimant la croissance des longueurs des facteurs) :

$$\sum_l j_l k_{j_l} = n \qquad \text{et} \qquad \forall p, \quad j_p < \sum_{l=1}^{p-1} j_l k_{j_l}.$$

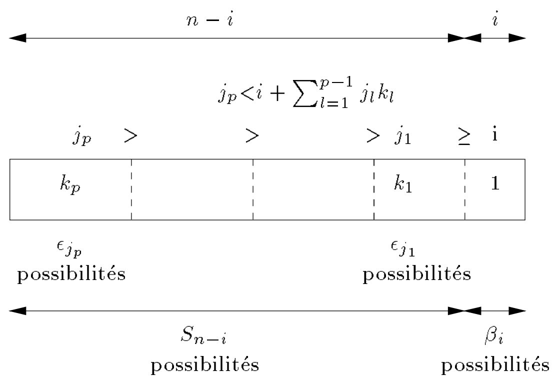

FIG. 3 – *Mots formés à partir des mots supprimés et d'un mot du code*

2. Les autres mots de longueur n du code Y_{n-1} sont de la forme $w = w_p w_{p-1} \cdots w_1 w_0$ où les w_i $(1 \leq i \leq p)$ sont des mots qui ont été supprimés et où w_0 est un mot du code Y_{n-1} qui a été engendré avant la suppression du mot w_1. Les w_i vérifient alors les conditions suivantes sur les longueurs (Figure 3) :

- $\forall i \geq 0, \quad |w_{i+1}| \geq |w_i|$,
- $\forall i \geq 1, \quad |w_{i+1}| < |w_i w_{i-1} \ldots w_1 w_0|$.
  En effet, si $|w_{i+1}| \geq |w_i w_{i-1} \ldots w_1 w_0|$, alors $W = w_i w_{i-1} \ldots w_1 w_0$ est lui-même, soit un mot du code et w doit alors nécessairement être factorisé sous la forme plus courte $w_p \cdots w_{i+1} W$, soit un mot que l'on a supprimé et w est alors un mot du premier type distingué.

Soit S_{n-i} le nombre de produits $w_p w_{p-1} \cdots w_1$ de longueur $n-i$ pouvant être concaténés avec un mot w_0 de Y_{n-1} de longueur i, alors

$$S_{n-i} = \sum_{\mathcal{S}} \prod_{i \leq j_1 < j_2 < \ldots < j_p} P_{k_{j_p}} \ldots P_{k_{j_2}} P_{k_{j_1}} \quad \text{avec} \quad P_{k_j} = \binom{\epsilon_j + k_j - 1}{k_j},$$

(les P_{k_j} représentent le nombre de combinaisons avec répétitions de k_j éléments pris dans un ensemble de cardinal ϵ_j) où la somme est prise sur l'ensemble $\mathcal{S}$ des p-uplets vérifiant les conditions suivantes (la condition sur j_k exprimant les relations entre les longueurs des facteurs)

$$\sum_l j_l k_{j_l} = n - i, \qquad i \leq j_1 \quad \text{et} \qquad \forall p \geq 2, \quad j_p < i + \sum_{l=1}^{p-1} j_l k_{j_l}.$$

Le nombre de mots de longueur n, $\beta_n + \epsilon_n$, du code Y_{n-1} est finalement donné par

$$\boxed{\beta_1 + \epsilon_1 = k, \qquad \forall n \geq 2, \quad \beta_n + \epsilon_n = \sum_{i=1}^{\lfloor n/2 \rfloor} S_{n-i} \beta_i + R_n,}$$

où $\lfloor x \rfloor = \max\{n / n \leq x, n \in \mathbb{N}\}$.

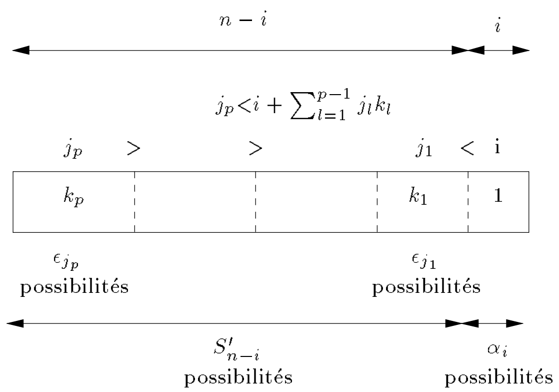

FIG. 4 – *Mots formés à partir des mots supprimés et d'un mot de l'alphabet*

Exemple 2 (suite) Les produits de mots supprimés sont pour $n = 2$: ab, pour $n = 3$: a^2b, bab, pour $n = 4$: $a^3b, ba^2b, b^2ab, abac$. Les autres mots se décomposent en produits de mots supprimés et d'un mot appartenant au code, ce dernier est alors en position suffixe (par exemple, pour $n = 4$: $a^3.c, b.a^2.c, b^2.a.c, ab.bc, ac.bc$). On a alors

$$\beta_2 + \epsilon_2 = P_{1_1}\beta_1 + T_2 = \binom{\epsilon_1}{1}\beta_1 + \binom{\epsilon_1}{2} = 3.$$

$$\beta_3 + \epsilon_3 = (P_{2_1} + P_{1_2})\beta_1 + T_{3_1} = \left(\binom{\epsilon_1 + 1}{2} + \epsilon_2\right)\beta_1 + 2\binom{\epsilon_1 + 1}{3} = 7.$$

$$\beta_4 + \epsilon_4 = \left(\binom{\epsilon_1 + 2}{3} + \epsilon_3\right)\beta_1 + \epsilon_2\beta_2 + 3\binom{\epsilon_1 + 2}{4} + \binom{\epsilon_2}{2} = 10.$$

Cas d'un alphabet pondéré Soit $(\alpha_n)_{n\geq 1}$ la distribution de longueurs de l'alphabet pondéré (A, λ). Le nombre de mots de longueur n du code Y_{n-1} est comme précédemment égal à $\beta_n + \epsilon_n$. Leur description est identique à la précédente à ceci près que l'alphabet étant pondéré, il contient des mots de longueur supérieure à 1, ce qui amène à distinguer un troisième type de mots. En effet, les mots de longueur n du code Y_{n-1} peuvent, dans ce cas, être de la forme $w_p w_{p-1} \cdots w_1 w_0$ où les w_i $(1 \leq i \leq p)$ sont des mots qui ont été supprimés et où w_0 appartient à l'alphabet. De plus, les α_i mots de longueur i de l'alphabet étant comptés parmi les ϵ_i mots que l'on supprime ou les β_i qui font partie du code, les longueurs des facteurs vérifient (Figure 4)

- $\forall i \geq 1, \quad \lambda(w_{i+1}) \geq \lambda(w_i)$,
- $\forall i \geq 0, \quad \lambda(w_{i+1}) < \lambda(w_i w_{i-1} \ldots w_1 w_0)$.
  En effet, si $\lambda(w_{i+1}) \geq \lambda(w_i w_{i-1} \ldots w_1 w_0)$, $w_i w_{i-1} \ldots w_1 w_0$ est lui-même, soit un mot du code, soit un mot que l'on a supprimé et w est alors un mot d'un des deux types étudiés précédemment.

En utilisant le calcul fait pour les préfixes des mots du codes dont le nombre est donné par le coefficient S_{n-i} et en tenant compte des relations sur les longueurs des facteurs, on obtient

$$S'_{n-i} = \sum_{\mathcal{S}'} \prod_{j_1<j_2<\ldots<j_p} P_{k_{j_p}} \ldots P_{k_{j_2}} P_{k_{j_1}} \quad \text{avec} \quad P_{k_j} = \binom{\epsilon_j + k_j - 1}{k_j} \quad \text{et} \quad S'_0 = 1,$$

où la somme est prise sur l'ensemble $\mathcal{S}'$ des p-uplets tels que

$$\sum_l j_l k_{j_l} = n - i, \qquad i > j_1 \quad \text{et} \qquad \forall p \geq 2, \quad j_p < i + \sum_{l=1}^{p-1} j_l k_{j_l}.$$

Finalement, la somme $\beta_n + \epsilon_n$, égale au nombre de mots de longueur n du code Y_{n-1}, s'écrit

$$\beta_1 + \epsilon_1 = \alpha_1,$$

$$\boxed{\forall n \geq 2, \qquad \beta_n + \epsilon_n = \textstyle\sum_{i=1}^{\lfloor n/2 \rfloor} S_{n-i}\beta_i + R_n + \sum_{i=\lceil n/2 \rceil}^{n} S'_{n-i}\alpha_i.}$$

où $\lfloor x \rfloor = \max\{n \mid n \leq x, n \in \mathbb{N}\}$ et $\lceil x \rceil = \min\{n \mid n \geq x, n \in \mathbb{N}\}$.

La preuve du Théorème 2 ne permet pas d'établir le rationalité du code circulaire que l'on construit dans le cas où la suite d'entiers positifs est supposée $\mathbb{N}$-rationnelle. La caractérisation des distributions de longueurs des codes circulaires rationnels demeure un problème ouvert.

Références

[AM95] D. Arquès and Christian Michel. A possible code in the genetic code. In E. W. Mayr and C. Puech, editors, *STACS 95*, volume 900 of *Lect. Notes in Comput. Sci.*, pages 640–651. Springer, 1995.

[BP85] J. Berstel and D. Perrin. *Theory of codes*. Pure and Applied Mathematics. Academic Press, Inc., 1985.

[GG65] S. W. Golomb and B. Gordon. Codes with bounded synchronisation delay. *Inform. and Control*, 8:355–372, 1965.

[GGW58] S. W. Golomb, B. Gordon, and L. R. Welch. Comma-free codes. *Canadian J. Math.*, 10:202–209, 1958.

[Kro87] D. Krob. Codes limités et factorisations finies du monoïde libre. *RAIRO Inform. Theor. Appl.*, 21:437–467, 1987.

[MR83] N. Metropolis and G.-C. Rota. Witt vectors and the algebra of necklaces. *Adv. Math.*, 50:95–125, 1983.

[Reu88] C. Reutenauer. Mots circulaires et polynômes irréductibles. *Ann. Sci. Math. Québec*, 12:275–285, 1988.

[Sch65a] M. P. Schützenberger. On a factorization of free monoids. *Proc. Amer. Math. Soc.*, 16:21–24, 1965.

[Sch65b] M. P. Schützenberger. Sur une question concernant certains sous-monoïdes libres. *C. R. A. S. Paris, Groupe 1*, 261:2419–2420, 1965.

[Sta86] R. P. Stanley. *Enumerative Combinatorics*, volume I. Wadsworth & Brooks/Cole, 1986.

[Vie78] G. Viennot. *Algèbres de Lie libres et monoïdes libres*. Number 691 in Lect. Notes in Math. Springer-Verlag, 1978.