Publ. I.R.M.A. Strasbourg, 1984, 229/S–08
Actes 8e Séminaire Lotharingien, p. 79–86

POLYNÔMES ORTHOGONAUX ET PROBLÈMES D'ÉNUMÉRATION EN BIOLOGIE MOLÉCULAIRE

PAR

M. VAUCHASSADE DE CHAUMONT ET GÉRARD VIENNOT

RÉSUMÉ. — Le but de cet exposé est de montrer comment certains polynômes de Techebycheff généralisés à un paramètre permettent de dénombrer les structures secondaires d'acides nucléiques simple-brin ayant une complexité donnée.

I. Introduction. — La stucture primaire des acides nucléiques simple-brin (acide ribonucléique ou ARN, ARN messager, ARN de transfert, etc.) est constituée par l'enchaînement linéaire de nucléotides reliés par des liaisons phosphodiesters. Chaque nucléotide est caractérisé par la base qu'il contient. Pour les ARN, il y a quatre bases possibles : Adénine, Cytosine, Guanine, Uracycle. Cette structure primaire est codée par un mot sur un alphabet à quatre lettres A, C, G, U. Les liaisons hydrogène replient la molécule sur elle-même pour former une figure planaire : la *structure secondaire*.

Les liaisons hydrogène ne peuvent se croiser. WATERMAN [11] a défini mathématiquement la notion de structure secondaire comme une certaine famille de *graphes* (en fait *cartes*) *planaires*. Cette famille contient toutes les structures secondaires connues. Notons également que l'on ne s'intéresse qu'au dessin formé par les liaisons, c'est-à-dire, que l'on oublie l'étiquetage des sommets du graphe par les bases A, C, G, U.

Les biologistes s'intéressent particulièrement à la prédiction des structures secondaires, la structure primaire étant connue (voir, par exemple, TINOCO et al. [10]). Les liaisons hydrogène étant de faible énergie, la stabilité de la molécule dépend beaucoup de ses "épingles à cheveux", c'est-à-dire de régions contenant des liaisons hydrogène parallèles. Un paramètre de *complexité* (appelé aussi *ordre*) d'une structure secondaire a été défini par MITIKO GÔ [4] et WATERMAN [11] pour le calcul de l'énergie libre des épingles à cheveux entre elles est plus "compliquée."

Le problème combinatoire sous-jacent est de dénombrer toutes les structures secondaires possibles de complexité k. Soit donc $a_{n,k}$ le nombre de structures secondaires ayant n bases (c'est-à-dire les sommets du graphe) et de complexité k. WATERMAN a montré que $a_{n,1}$ est asymptotiquement équivalent à λ^n, en désignant par λ la plus grande racine de l'équation $x^2 - 2x^3 - 1 = 0$ (soit $\lambda = 2,2055\ldots$).

Notons $s_k(t) = \sum_{n\geq 0} a_{n,k} t^n$ la série génératrice des structures d'ordre k. Le résultat fondamental de l'exposé est le suivant :

THÉORÈME 1. — *La série génératrice des structures secondaires d'ordre k est*

$$s_k(t) = \frac{t^{(5.2^{k-1}-2)}}{(1-t)Z_1\ldots Z_k},$$

dans lequel $Z_1, \ldots, Z_k$ désigne la suite de polynômes définis par la récurrence

$$Z_1 = 1 - 2t - t^2, \quad Z_{k+1} = Z_k^2 - 2t^{5.2^{k-1}} \quad (k \geq 1).$$

La méthode employée, chère à M.-P. SCHÜTZENBERGER, consiste à coder les objets combinatoires ayant une série algébrique (ou rationnelle) par les mots d'un *langage algébrique* (ou "context-free," au sens de la théorie des langages en Informatique Théorique). La série génératrice (ordinaire) s'obtient par passage à l'image commutative de la *série génératrice non-commutative* formée par la somme formelle des mots du langage (plus précisément en envoyant par morphisme toutes les variables sur la même variable t). La série génératrice des mots du langage est solution d'un système algébrique (en séries formelles à variables non-commutatives). Chaque équation traduit une propriété combinatoire des objets. Ce passage au non-commutatif est ainsi un intermédiaire commode entre l'objet combinatoire et la série génératrice classique.

La résolution du système provenant des structures secondaires nécessite la méthodologie des interprétations combinatoires des *polynômes orthogonaux*. En particulier, apparaissent des "polynômes de Tchebycheff" $F_n(x,b)$ à un paramètre b, qui sont orthogonaux pour la forme linéaire $P \mapsto \int_\alpha^\beta P\,d\psi$, définie par ses *moments* :

$$\mu_n(b) = \int_\alpha^\beta x^n\,d\psi = \sum_{k\geq 1} \frac{1}{n}\binom{n}{k}\binom{n}{k-1} b^k. \tag{1}$$

Pour $b=1$, le polynôme $F_n(x,1)$ est le polynôme $U_n(x/2)$, dans lequel U_n est le *polynôme de Tchebycheff* de deuxième espèce défini par

$$\sin(n+1)\theta = \sin\theta\, U_n(\cos\theta). \tag{2}$$

Cet exposé est un résumé d'un article qui sera publié ultérieurement. Il a fait l'objet d'une communication au congrès *Mathematics in Biology and Medicine* [Bari, Italie. Juillet 1983].

2. Codage par des mots. — On peut coder les structures secondaires ayant n bases par des mots de longueur n sur l'alphabet $X = \{a, x, \overline{x}\}$. On parcourt les sommets de la structure secondaire selon l'ordre de la séquence primaire, en écrivant x (resp. $\overline{x}$) lorsque l'on rencontre une liaison hydrogène pour la première fois (resp. deuxième fois). Lorsque la base n'est pas reliée par une liaison hydrogène, on écrit a. La définition des structures secondaires donnée par WATERMAN [11] permet de caractériser l'ensemble $\mathcal{S}$ des mots ainsi obtenus. Ce sont les mots w vérifiant les trois conditions suivantes :

(3) pour toute factorisation $w = uv$, $|u|_x \geq |u|_{\overline{x}}$ (dans lequel $|u|_z$ désigne le nombre d'occurrences de la lettre z dans le mot u) ;

(4) $|w|_x = |w|_{\overline{x}}$;

(5) w n'a pas de facteur $x\overline{x}$ (c'est-à-dire, qu'il n'y a pas de factorisation $w = ux\overline{x}v$.

Remarque. — Classiquement, les mots sur l'alphabet $\{a, x, \overline{x}\}$ vérifiant (3) et (4) sont appelés *mots de Motzkin*. Le nombre de tels mots de longueur n est le nombre de Motzkin M_n.

Les mots sur l'alphabet $\{x, \overline{x}\}$ vérifiant (3) et (4.) sont appelés *mots de Dyck*. Le nombre de tels mots de longueur $2n$ est le classique *nombre de Catalan* $C_n = \frac{1}{n}\binom{2n}{n}$. La condition (5) traduit le fait que deux bases ne peuvent être reliées à la fois par des liaisons phosphodiesters et hydrogène.

Notons $S = \sum_{w \in \mathcal{S}} w$ la série génératrice formelle en variables non commutatives $x \in X$ (à coefficients dans $\mathbf{Z}$). Elle est solution de l'équation algébrique

$$S = 1 + aS + x(S-1)\overline{x}\,S. \tag{6}$$

Par le morphisme φ envoyant toutes les lettres a, x, $\overline{x}$ sur la lettre t, on obtient la série génératrice des structures secondaires, dénombrées selon le nombre de bases (voir aussi STEIN, WATERMAN [5])

$$s(t) = \frac{1}{2t^2}\Big(t^2 - t + 1 - (1 + t(t^3 - 2t^2 - t - 2))^{1/2}\Big). \tag{7}$$

3. Complexité d'une structure secondaire. — Nous définissons l'ordre d'un mot, traduisant la définition de la complexité d'une structure secondaire de [11]. Soit w un mot de Motzkin (vérifiant (3) et (4)) et $\alpha(w)$ le mot (de Dyck) obtenu en enlevant les lettres a. Une *pyramide* d'un mot de Dyck est un facteur de la forme $x^p\overline{x}^p$. Une pyramide est dite *maximale* si elle n'est pas facteur d'une autre pyramide. Tout mot de Dyck w admet une décomposition unique en pyramides maximales de la forme $w = u_1v_2 \ldots u_qv_qu_{q+1}$ avec $q \geq 1$ et $v_1, \ldots, v_q$ pyramides maximales.

On définit alors $\pi(w) = u_1 \ldots u_{q+1}$ (opérateur suppression des pyramides maximales). L'*ordre* d'un mot $w \in S$ codant une structure secondaire est le minimum des entiers k tels que

$$\pi^k(\alpha(w)) = e \quad \text{(mot vide)}. \tag{8}$$

Exemple. — Soit $w = x\,x\,x\,\overline{x}\,x\,x\,\overline{x}\,\overline{x}\,\overline{x}\,x\,\overline{x}\,\overline{x}\,x\,x\,\overline{x}\,x\,\overline{x}\,\overline{x}$, alors
$\pi(w) = x\,x\,\overline{x}\,\overline{x}\,x\,\overline{x}, \quad \pi^2(w) = e.$
Le mot w est d'ordre 2 (on a indiqué les décompositions en pyramides maximales).

Remarque. — Les pyramides maximales du mot de Dyck $\alpha(w)$ correspondent aux épingles à cheveux de la structure secondaire codée par w.

Dans notre formalisme, on peut énoncer commodément une propriété récemment mise en évidence à Strasbourg au laboratoire du professeur J.-P. Ebel [8] : les ARN de la petite sous-unité ribosomale de diverses expèces ont "en gros" le même mot de Dyck $\alpha(w)$ associé (bien que le nombre de bases puisse varier de 954 pour l'ARN 12-S des mitochondries humaines à 1825 pour l'ARN 18-S du cytomlasme de Xenopus laevis).

Pour $k \geq 1$ la série génératrice non commutative S_k des mots codant les structures secondaires d'ordre k est solution du système algébrique suivant :

$$\begin{cases} S_k = S_{\leq k} - S_{\leq k-1} \quad (k \geq 1); \\ S_{\leq k} = (1 - T_{\leq k})^{-1} \quad (k \geq 0); \\ T_{\leq k} = T_0 + T_1 + \cdots + T_k \quad (k \geq 0); \\ T_k = xS_{\leq k-2}T_{k-1}S_{\leq k-2}T_{k-1}S_{\leq k-1}\overline{x} + xS_{\leq k-1}T_kS_{\leq k-1}\overline{x} \ (k \geq 2); \\ T_0 = a; \\ T_1 = xS_0T_1S_0\overline{x} + xS_0T_0\overline{x}. \end{cases} \tag{9}$$

La série $S_{\leq k}$ est la série génératrice des mots codant les structures secondaires d'ordre $\leq k$. Un mot de Motzkin est dit *premier* lorsqu'il ne peut se factoriser en produit de $p \geq 2$ mots de Motzkin. La notation T_k (resp. $T_{\leq k}$) désigne la série génératrice des mots de Motzkin premiers codant les structures secondaires.

4. Mots de Dyck d'ordre donné. — Soit D_k la série génératrice non-commutative des mots de Dyck d'ordre k. Cette série vérifie un système analogue à (9), mais en remplaçant les deux dernières équations par

$$T_0 = 0, \qquad T_1 = \sum_{n \geq 1} x^n \overline{x}^n. \tag{10}$$

L'image par le morphisme φ envoyant x et $\overline{x}$ sur t donne la série génératrice ordinaire $d_k(t)$ (resp. $d_{\leq k}(t)$) des mots de Dyck d'ordre k (resp. $\leq k$).

PROPOSITION 2. — *On a*

$$d_k(t) = \frac{t^{(2^{k+1}-2)}}{R_{(2^{k+1}-1)}(t)}, \qquad d_{\leq k}(t) = \frac{R_{(2^{k+1}-2)}(t)}{R_{(2^{k+1}-1)}(t)},$$

où $R_n(t)$ désigne le polynôme réciproque du polynôme $F_n(t) = U_n(t/2)$ défini en (2).

La preuve repose sur les deux identités suivantes

$$R_p^2 - t^2 R_{p-1}^2 = R_{2p} \quad (1 \leq p); \tag{11}$$

$$R_{q+1}R_p - R_q R_{p+1} = t^{2q+2} R_{p-q-1} \quad (1 \leq q < p). \tag{12}$$

Nous en donnons des preuves bijectives simples : (11) repose sur une interprétation de R_p par des *pavages (valués) de dominos* deux à deux disjoints du segment $[1,p]$; (12) repose sur l'interprétation de R_p/R_{p+1} comme série génératrice des “chemins de Dyck” associés aux mots de Dyck et de hauteur bornée p (voir KREWERAS [3]).

5. Résolution dy système (9). — On remarque que les polynômes $F_n(x) = U_n(x/2)$ sont orthogonaux pour la suite des moments $\pi_n = C_n$ (nombre de Catalan). La condition (3) nécessite de comptabiliser les “pics” (ou les facteurs $x\overline{x}$) des mots de Dyck, c'est-à-dire, d'introduire une variable formelle non-commutative b dans (9) par la relation suivante

$$T_1 = xS_0T_1S_0\overline{x} + xS_0b\overline{x}. \tag{13}$$

Il est bien classique [3] que la distribution des mots de Dyck selon le nombre de pics est donnée par les *nombres de Runyon* (appelés aussi nombres de Narayana $\frac{1}{n}\binom{n}{k}\binom{n}{k-1}$. On est donc conluit à chercher des polynômes orthogonaux $F_n(x,b)$ (à un paramètre b) pour la suite de moments définis par (1). Ces polynômes se réduisent à $F_n(x)$ pour $b=1$.

Nous donnons une preuve purement bijective du fait que ces polynômes sont donnés par la récurrence suivante :

$$F_{n+1}(x,b) = xF_n(x,b) - \lambda_n F_{n-1}(x,b), \tag{14}$$

avec

$$\lambda_n = \begin{cases} 1, & \text{si } n \text{ impair ;} \\ b, & \text{si } n \text{ pair.} \end{cases}$$

En introduisant le paramètre b dans les bijections prouvant (11) et (12), on prouve l'analogue de ces relations.

Soit maintenant $H_n(x,b)$ le polynôme réciproque (à un paramètre) du polynôme $F_n(x-1,b)$. Posons $V_n(x) = H_n(x,x)$. On démontre par récurrence sur k que la série $s_k = \varphi(S_k)$ est donnée par l'expression suivante :

$$s_k = \frac{t^{(5.2^{k-1}-2)}}{V_{(2^{k+1}-1)}}. \tag{15}$$

En fait V_n est défini par la récurrence

$$V_0 = 1, \quad V_1 = x, \quad V_{n+1} = (1-x)V_n - \lambda_n V_{n-1}, \tag{16}$$

avec

$$\lambda_n = \begin{cases} x^2, & \text{pour } n \text{ impair}; \\ x^3, & \text{pour } n \text{ pair}. \end{cases}$$

Une bijection prouve la relation

$$V_{2^{k+1}-1} = V_{2^k-1} Z_k \qquad (k \geq 1), \tag{17}$$

dans laquelle Z_k est un polynôme obtenu comme somme de "pavages" valués sur un cercle ayant 2^k points, de dominos et monominos deux à deux disjoints. Une autre bijection sur ces pavages prouve la récurrence :

$$Z_{k+1} = Z_k^2 - 2t^{5.2^{k-1}} \qquad (k \geq 1). \tag{18}$$

On obtient ainsi le théorème 1.

Remarque finale. — Il est curieux de constater que la série ordinaire $d_k(t)$ de la Proposition 2 est la même que celle donnant la série génératrice des *arbres binaires* ayant un *nombre de Strahler* égal à k. Ce nombre est un paramètre introduit en Hydrographie Fluviale par STRAHLER [9], apparaissant aussi en Informatique Théorique [1], et aussi en Botanique et Anatomie (voir pages 113 et 225 de la traduction française [6]). On trouvera dans FRANÇON [2] d'intéressantes considérations bijectives montrant que ces deux paramètres ont même distribution qu'un troisième : le plus grand entier inférieur ou égal au logarithme (en base 2) de la hauteur maximale des chemins de Dyck associés.

BIBLIOGRAPHIE

[1] Flajolet (Philippe), Raoult (Jean-Claude) et Vuillemin (Jean). — The number of registers required for evaluating arithmetic expressions, *Theor. Comp. Sc.*, t. **9**, 1979, p. 99–125.

[2] Françon (Jean). — Sur le nombre de registres nécessaires à l'évaluation d'une expression arithmétique, à paraître dans *R.A.I.R.O.*

[3] Kreweras (Germain). — Sur les éventails de segments, *Cahiers du B.U.R.O.*, t. **15**, 1970, p. 3–41.

[4] Mitiko Gô. — Statistical mechanics of biopolymers and its application to the melting transition of polynucleotides, *J. Phys. Soc. Japan*, t. **23**, 1967, p. 597–607.

[5] Stein (P.R.) and Waterman (M.S.). — On some new sequences generalizing the Catalan and Motzkin numbers, *Discrete Math.*, t. **26**, 1979, p. 261–272.

[6] Stevens (P.R.). — *Patterns in Nature.* — Little, Brown and Co., 1974 [Traduction française : *Les formes dans la nature.* — Paris, Seuil, 1978].

[7] Stiegler (P.), Carbon (P.), Zuker (M.), Ebel (J.-P.) and Ehresmann (C.). — *Nucleic Acids Res.*, t. **9**, 1981, p. 2153–2172.

[8] Stiegler (P.), Carbon (P.), Ebel (J.-P.) and Ehresmann (C.). — A general secondary structure model for procaryotic and euraryotic RNA's of the small ribosomal subunits, *Europ. J. Biochem.*, t. **120**, 1981, p. 487–495.

[9] Strahler (A.N.). — Hypsometric (area-altitude) analysis of erosional topology, *Bull. Geological Soc. Amer.*, 1952, p. 1117–1142.

[10] Tinoco (I.), Uhlenbeck (0.C.) and Levine (M.D.). — Estimation of secondary structure in ribonucleic acids, *Nature*, t. **230**, 1971, p. 362–367.

[11] Waterman (Michael S.). — Secondary structure of single-stranded nucleic acids, in *Studies in Foundations and Combinatorics* (Adv. in Math., suppl.) t. **1**, 1978, p. 167–172.

M. Vauchassade de Chaumont,
Université de Bordeaux II,
Bordeaux, France

et

Gérard Viennot,
U.E.R. de mathématiques et d'informatique,
Université de Bordeaux I,
351, cours de la Libération,
33405 Talence Cedex, France.