

# MATHÉMATIQUES ET BIOLOGIE

---

## Codages de séquences

G. Didier, I. Laprevotte et M. Pupin

---

### 1 Introduction

L'information génétique – le génome – d'un organisme vivant est portée par une ou plusieurs chaînes chimiques orientées : l'ADN (ou des polymères proches appelés ARN pour certains virus). Ces chaînes sont formées d'une succession de quatre types de molécules appelées nucléotides. Depuis quelques années déjà, il est possible de « lire » tout ou partie de ces chaînes, c'est-à-dire d'obtenir la séquence de leurs éléments sous la forme de longs textes ne contenant que quatre lettres A, C, G et T représentant les nucléotides. Le séquençage de divers organismes, représentatifs de l'ensemble du vivant, se poursuit à un rythme soutenu. Les ordres de tailles des séquences ainsi obtenues vont couramment jusqu'à quelques centaines de milliers voire à plusieurs millions et au-delà pour les génomes entiers. Toutefois les problématiques biologiques concernent assez souvent des fragments de taille plus réduite, typiquement de l'ordre de quelques centaines à quelques milliers de nucléotides.

Lorsqu'on travaille sur les séquences d'ADN, une des difficultés principales est due au fait que l'alphabet ne contienne que quatre lettres. Cet alphabet réduit et d'autres considérations sont à l'origine de l'intuition biologique qu'un élément d'une séquence ne peut prendre sens qu'au sein d'un contexte donné. Une première approche pour prendre en compte les nucléotides non plus isolément mais au sein de leurs contextes consiste à faire glisser une fenêtre d'une longueur donnée  $N$  le long de la séquence et à recoder chaque configuration observée dans la fenêtre par une lettre différente. On obtient ainsi une nouvelle séquence dont les « lettres » sont les sous-mots ou blocs de longueur  $N$  de la séquence initiale. De fait de nombreuses méthodes d'analyse de séquences travaillent sur les blocs de nucléotides. C'est par exemple le cas dans certaines méthodes d'alignement qui s'ancrent sur des blocs d'homologies ou encore dans les modélisations de type Markov d'ordre  $N$ . Cette première approche soulève néanmoins plusieurs objections. En particulier, l'écriture par blocs manque de souplesse : il suffit que deux blocs de longueur  $N$  diffèrent sur seulement une position pour que les « lettres » qui leur correspondent soient différentes. Notamment dans le cas de l'alignement, ceci rend difficile la prise en compte d'éventuelles mutations sur les séquences. De plus on doit faire face à une explosion combinatoire dès que l'on s'intéresse à des contextes de longueur un peu importante : le nombre de contextes à considérer croît exponentiellement avec leur longueur.

Pour répondre à ces objections, nous avons étudié cette transformation qui est classique en dynamique symbolique [6], tout d'abord en caractérisant les

mots qui en sont issus, ensuite en montrant comment elle pouvait, dans une certaine mesure, être inversée. Plus précisément, étant donné une suite de contextes de longueur  $N$  d'une séquence, il est possible de construire un antécédent (*i.e.* une séquence qui a le même enchaînement de contextes de longueur  $N$  que la séquence initiale) particulier, maximal dans la mesure où tous les autres antécédents possibles peuvent en être déduits en identifiant certaines de ses lettres. En pratique on obtient à partir d'une séquence d'ADN, une nouvelle séquence dans laquelle on a distingué différents types de nucléotides selon leurs environnements. La section suivante contient les résultats principaux et des éléments de preuves (nous renvoyons à [2] pour les preuves complètes) ainsi qu'un algorithme de faible complexité permettant la construction effective de l'antécédent maximal.

Enfin la dernière section précise les modalités d'utilisation dans l'analyse des séquences génétiques et présente deux applications pratiques : l'identification de régularités et l'alignement de séquences.

## 2 Codages par blocs – $N$ -écritures

### 2.1 Notations et définitions

On appelle **alphabet** tout ensemble fini non vide d'éléments appelés lettres. Un **mot** sur un alphabet  $X$  est une suite de lettres de  $X$  indexée sur  $\{0, 1, \dots, n\}$  pour un entier naturel  $n$ , si le mot est fini, et sur  $\mathbb{N}$  si le mot est infini. Dans le cas fini, on parlera également de **séquence**. La **longueur** d'un mot fini  $w$ , notée  $|w|$ , est le nombre de lettres le composant.

Que  $u$  soit un mot fini ou infini, on appellera **blocs** de  $u$  les sous-mots de  $u$ , c'est-à-dire des mots de la forme  $u_i u_{i+1} \dots u_{i+n-1}$ . Les blocs d'un ensemble de mots se définissent comme l'union des blocs des mots composant l'ensemble. On note  $\mathcal{B}_k(u)$  l'ensemble des blocs de longueur  $k$  apparaissant dans  $u$ . On définit de façon similaire l'ensemble des blocs de longueur  $k$  apparaissant dans un mot infini ou encore un ensemble de mots finis ou infinis.

L'**alphabet** d'un mot  $u$ , noté  $\mathcal{A}_u$ , est l'ensemble des blocs de longueur 1 de  $u$ .

Un codage par blocs naturel consiste à fixer un entier  $N$  et à associer une lettre différente à chaque bloc de longueur  $N$ . L'image d'un mot  $u$  par cette opération est appelée  **$N$ -écriture** de  $u$  et noté  $u(N)$ . Cette opération peut être appliquée sur un ensemble de mots finis, un mot infini ou un ensemble de mots infinis.

**Exemple.** —

$$u = abacaabcaabc$$

Sa 2-écriture  $u(2)$  est :

$$\begin{aligned} u(2) &= (ab)(ba)(ac)(ca)(aa)(ab)(bc)(ca)(aa)(ab)(bc) \\ &= 0 \quad 1 \quad 2 \quad 3 \quad 4 \quad 0 \quad 5 \quad 3 \quad 4 \quad 0 \quad 5 \end{aligned}$$

### 2.2 Propriétés des $N$ -écritures

1. Si  $u$  est de longueur  $|u|$  alors  $u(N)$  est de longueur  $|u| - N + 1$ .
2. L'opération peut s'itérer et l'on a :  $(u(N))(K) = u(N + K - 1)$ .
3. Le nombre de blocs différents de longueur  $k$  apparaissant dans  $u(N)$  est égal au nombre de blocs différents de longueur  $k + N - 1$  apparaissant dans  $u$ .
4. Un changement sur une position d'un mot entraîne  $N$  changements sur sa  $N$ -écriture : si deux mots  $u$  et  $v$  diffèrent sur une position, alors  $u(N)$  et  $v(N)$  diffèrent sur les  $N$  positions incluant celle considérée.
5. Une  $N$ -écriture peut avoir plusieurs antécédents. On peut avoir  $u(N) = v(N)$  avec  $u \neq v$  (même à une bijection d'alphabet près).

### 2.3 Caractérisation

Pour caractériser les mots qui sont des  $N$ -écritures, l'idée principale est d'utiliser le fait que les blocs correspondant à deux lettres consécutives d'une  $N$ -écriture se chevauchent. Ceci a pour conséquence que si deux lettres  $a$  et  $b$  d'une  $N$ -écriture voient chacune une de leurs occurrences précédée (resp. suivie) par une même lettre  $c$  alors les préfixes (resp. les suffixes) de longueur  $N - 1$  des blocs correspondant aux lettres  $a$  et  $b$  coïncident. Ce type d'argument a amené la définition d'une relation d'équivalence sur les lettres de l'alphabet qui est construite à partir de l'ensemble des blocs de longueur 2 du mot (ou de l'ensemble de mots considéré).

**Définition 1.** — Pour  $k \in \{0, \dots, N - 1\}$ , on définit  $\sim_k^*$  la relation sur les lettres de  $\mathcal{A}_v$  par  $a \sim_k^* b$  si l'une des conditions suivantes est vérifiée :

1.  $a = b$ ,
2. Pour un entier  $i \in [1, N - 1 - k]$ , il existe des lettres  $\alpha_1, \dots, \alpha_i, \beta_1, \dots, \beta_i$  de  $\mathcal{A}_v$  avec  $\alpha_i = \beta_i$  telles que  $\alpha_i \alpha_{i-1}, \dots, \alpha_2 \alpha_1, \alpha_1 a, \beta_i \beta_{i-1}, \dots, \beta_2 \beta_1, \beta_1 b$  appartiennent à  $\mathcal{B}_2(v)$ ,
3. Pour un entier  $i \in [1, k]$ , il existe des lettres  $\alpha_1, \dots, \alpha_i, \beta_1, \dots, \beta_i$  de  $\mathcal{A}_v$  avec  $\alpha_i = \beta_i$  telles que  $a \alpha_1, \alpha_1 \alpha_2, \dots, \alpha_{i-1} \alpha_i, b \beta_1, \beta_1 \beta_2, \dots, \beta_{i-1} \beta_i$  appartiennent à  $\mathcal{B}_2(v)$ .

On note  $\sim_k$  la fermeture transitive de  $\sim_k^*$  et  $\mathcal{P}_k$  la partition de  $\mathcal{A}_v$  correspondante.

**Théorème 1.** — *Un mot  $v$  (resp. un ensemble de mots) est une  $N$ -écriture si et seulement si il n'existe pas deux lettres  $a$  et  $b$  dans  $\mathcal{A}_v$  telles que  $a \sim_k b$  pour tout entier  $k \in \{0, 1, \dots, N - 1\}$ .*

Preuve :

( $\Rightarrow$ ) S'il existe un mot  $u$  tel que  $u(N) = v$ , alors à chaque lettre  $a$  de  $v$  correspond un unique bloc  $a_0 a_1 \dots a_{N-1}$  de  $u$  et inversement. L'argument de « chevauchement » implique que, pour tout couple de lettres  $a$  et  $b$  de  $v$ , on a  $a \sim_k b \Rightarrow a_k = b_k$ . Si  $a$  et  $b$  sont deux lettres de  $v$ , il existe donc un entier  $k$  tel que  $a_k \neq b_k$  et donc  $a \not\sim_k b$

( $\Leftarrow$ ) La condition du théorème permet de construire un antécédent particulier, noté  $A_N(v)$ . Cette construction est présentée dans la sous-section suivante.

## 2.4 Construction

L'exposé a été ici légèrement adapté. Pour ce faire, on ajoute l'hypothèse que les  $N$  dernières lettres du mot considéré admettent plus d'une occurrence. Une présentation plus générale peut être trouvée dans [2], mais le traitement des cas, « dégénérés » alourdit un peu l'exposé.

Pour toute classe  $C$  de  $\mathcal{P}_k$ , on note  $P(C) = \{a \in \mathcal{A}_v \mid \exists b \in C \text{ tel que } ab \in \mathcal{B}_2(v)\}$ , l'ensemble des lettres précédant une occurrence d'une lettre de  $C$  dans  $v$ .

**Lemme 1.** — *Pour tout  $k < N - 1$  et toute classe  $C$  de  $\mathcal{P}_k$ , si  $P(C)$  est non vide alors  $P(C)$  est inclus dans une unique classe de  $\mathcal{P}_{k+1}$  que l'on notera  $P_\star(C)$ .*

A chaque classe  $C$  de  $\mathcal{P}_0$  correspond donc un unique  $N$ -uplet dont le premier élément est  $C$  et les éléments suivants, les itérés  $P_\star^i(C)$ . L'ensemble de ces  $N$ -uplets va constituer notre nouvel alphabet :

$$\mathcal{A}' = \{(C, P_\star(C), P_\star^2(C), \dots, P_\star^{N-1}(C)); C \in \mathcal{P}_0\}$$

On définit alors  $N$  projections (*i.e.* applications lettre à lettre se prolongeant par concaténation sur les mots) de  $\mathcal{A}_v$  vers  $\mathcal{A}'$  par :

$$\varphi_k : a \rightarrow (C, P_\star(C), P_\star^2(C), \dots, P_\star^{N-1}(C)) \text{ tel que } a \in P_\star^k(C)$$

La projection  $\varphi_k$  associe à toute lettre  $a$  de  $\mathcal{A}_v$ , le  $N$ -uplet de  $\mathcal{A}'$  dont le  $k^{\text{ième}}$  élément contient  $a$  (l'hypothèse supplémentaire faite dans cette sous-section assure de l'existence d'un tel  $N$ -uplet de  $\mathcal{A}'$  pour toute lettre de  $\mathcal{A}_v$ ).

En appliquant ces projections sur le mot  $v$ , on obtient un nouveau mot, que nous appellerons **antécédent maximal à l'ordre  $N$**  de  $v$  et noterons  $A_N(v)$ . A cause de la propriété 1, concernant la longueur des  $N$ -écritures, il nous faut distinguer le cas fini du cas infini :

- si  $v$  est fini, on pose  $A_N(v) = \varphi_0(v)\varphi_1(v_{|v|-1}) \dots \varphi_{N-1}(v_{|v|-1})$ ,
- si  $u$  est infini, on pose  $A_N(v) = \varphi_0(v)$ .

Remarquons que la relation  $\sim_k$ , et donc les projections qui en découlent, peuvent être définies y compris dans le cas où  $v$  n'est pas une  $N$ -écriture. Cependant la propriété de caractérisation implique que deux lettres différentes  $a$  et  $b$  de  $\mathcal{A}_v$  sont telles que  $\varphi_k(a) \neq \varphi_k(b)$  pour au moins un entier  $k$ . Ceci permet de conclure que la  $N$ -écriture de l'antécédent maximal à l'ordre  $N$  de  $v$  est bien  $v$ .

**Corollaire 1.** — *Si  $u$  est un mot tel que  $u(N) = v$  alors il existe une application lettre à lettre  $\delta$  de  $\mathcal{A}'$  vers  $\mathcal{A}_u$  telle que  $u = \delta(A_N(v))$ .*

La démonstration utilise le même argument que pour le sens ( $\Rightarrow$ ) du théorème. En notant  $a_0a_1 \dots a_{N-1}$  et  $b_0b_1 \dots b_{N-1}$  les blocs de  $u$  correspondant à deux lettres  $a$  et  $b$  de  $v$ , on a  $a \sim_k b$ , autrement dit  $\varphi_k(a) = \varphi_k(b)$ , implique  $a_k = b_k$ .

**Exemple.** — Reprenons l'exemple précédent en posant  $v = u(2)$ .

On a :  $v = 01234053405$

L'alphabet  $\mathcal{A}'$  se déduit des blocs de longueur 2 de  $v$  :

$$\begin{array}{l} a_0 = ( \begin{array}{cc} \mathcal{P}_0 & \mathcal{P}_1 \\ \{0\} & , \quad \{4\} \end{array} ) \\ b_0 = ( \begin{array}{cc} \{1,5\} & , \quad \{0\} \end{array} ) \\ a_1 = ( \begin{array}{cc} \{2\} & , \quad \{1\} \end{array} ) \\ c_0 = ( \begin{array}{cc} \{3\} & , \quad \{2,5\} \end{array} ) \\ a_2 = ( \begin{array}{cc} \{4\} & , \quad \{3\} \end{array} ) \end{array}$$

Pour une raison qui apparaîtra par la suite, on a baptisé chaque couple de la forme  $(C, P(C))$  par une lettre de  $\mathcal{A}_u$  suivie d'un indice (cette possibilité nous est assurée par le corollaire).

Les projections  $\varphi_0$  et  $\varphi_1$  correspondantes nous donnent un antécédent  $A_2(v)$  :

$$A_2(v) = a_0b_0a_1c_0a_2a_0b_0c_0a_2a_0b_0c_0$$

On remarque que  $A_2(v)$  est différent de  $u$  (son alphabet contient 5 lettres contre 3 pour  $u$ ) mais  $u$  peut bien être obtenu à partir de  $A_2(v)$  par la projection « supprimant les indices ».

### 2.5 Propriétés des antécédents maximaux

1. Bien que l'on ait  $(u(N))(K) = u(N + K - 1)$ , l'égalité  $(A_{N+K-1}(v))(K) = A_N(v)$  n'est pas vérifiée en général. Autrement dit, le passage à l'antécédent maximal ne peut pas toujours s'itérer. Il est des cas où l'antécédent maximal à l'ordre  $N$  d'une  $(N + K - 1)$ -écriture n'est pas une  $K$ -écriture.
2. Si  $N$  et  $M$  sont deux entiers tels que  $N \geq M$ , on peut relier l'antécédent maximal à l'ordre  $M$  de la  $M$ -écriture de  $u$  à l'antécédent maximal à l'ordre  $N$  de la  $N$ -écriture de  $u$ . Plus précisément, il existe une projection  $\varphi$  de  $\mathcal{A}_{A_N(u(N))}$  vers  $\mathcal{A}_{A_M(u(M))}$  telle que  $A_M(u(M)) = \varphi(A_N(u(N)))$ .
3. Le modèle de Markov à l'ordre 1 que l'on peut déterminer sur l'antécédent  $A_N(u(N))$ , associé à la projection  $\delta$  telle que  $u = \delta(A_N(u(N)))$  définit un modèle de Markov à états cachés particulier, où les états correspondent aux lettres de  $\mathcal{A}_{A_N(u(N))}$  et la fonction d'émission est  $\delta$  (voir [7] pour une présentation des modèles de Markov à états cachés). En effet, ce modèle est celui possédant le plus d'états et vérifiant que  $N$  lettres de  $u$  incluant une position donnée de manière quelconque déterminent complètement l'état, finalement pas si caché, associé à cette position.

### 2.6 Algorithme

Soit  $v$  une  $N$ -écriture. L'algorithme suivant, de complexité  $O(N \times |v|)$ , instancie le tableau des projections  $(\varphi_i)_{i \in \{0, \dots, N-1\}}$  de toutes les lettres de  $v$ .

[1] Initialiser toutes les projections  $(\varphi_i)_{i \in \{0, \dots, N-1\}}$  de toutes les lettres à « non instanciée » *lettre courante*  $\leftarrow 0$  lettre  $l \in \mathcal{A}_v$  projection  $\varphi_i$  non instanciée de  $l$   $\varphi_i(l) \leftarrow$  *lettre courante* (*i.e.* instancier  $\varphi_i(l)$  à *lettre courante*) Empiler le couple  $(l, i)$  la pile n'est pas vide Dépiler le dernier couple  $\rightarrow (a, j)$  position  $p$  d'une occurrence de  $a$  position  $k \in \{p + j - N - 1, \dots, p + j\}$  la projection de  $v_k$  par  $\varphi_{p+j-k}$  n'est pas instanciée  $\varphi_{p+j-k}(v_k) \leftarrow$  *lettre courante* Empiler le couple  $(v_k, p + j - k)$  *lettre courante*  $\leftarrow$  *lettre courante* +1

### 3 Analyse des séquences génétiques

#### 3.1 Présentation

Comme on l'a dit dans l'introduction, une première approche pour pallier aux problèmes posés par l'écriture par blocs de longueur  $N$  consiste à recoder une séquence ou un ensemble de séquences en l'antécédent maximal de rang  $N$  de leur  $N$ -écriture. On parlera par la suite de recodage à l'ordre  $N$  des séquences. Pratiquement, le calcul de la  $N$ -écriture d'une séquence  $v$  peut être réalisé par un algorithme simple avec un temps de calcul  $O(N \times |v|)$ . L'algorithme présenté dans la section précédente permet ensuite de déterminer l'antécédent maximal d'une  $N$ -écriture avec le même ordre de complexité. Le recodage de séquences ou d'ensemble de séquences est donc possible même pour des longueurs conséquentes.

Le recodage n'a qu'un seul paramètre : l'ordre  $N$  qu'il faut fixer. L'expérience montre qu'en général, la plage des ordres pertinents est assez réduite : pour un  $N$  trop petit, le recodage ne modifie rien (on reste sur l'alphabet  $\{A, C, G, T\}$ ); pour un  $N$  trop grand, il y a une lettre différente en chaque position de la ou des séquences recodées. De plus les séquences recodées à l'ordre  $N + K$  se projettent toujours sur les séquences recodées à l'ordre  $N$ .

Les deux sous-sections suivantes présentent deux applications effectives illustrant l'aide que peut apporter le recodage à des analyse brutes de séquences. La première concerne l'alignement de séquences de LTR viraux et la seconde l'étude et l'identification des régularités présentes dans les portions de séquences de la levure.

Il est à noter que, par nature, l'approche nécessite que les blocs dans la ou les séquences aient une certaine redondance. Ce qui est le cas dans les deux problématiques biologiques précédentes : l'identification de régularités a été effectuée sur des séquences sélectionnées car contenant des blocs sur-représentés et il n'est utile d'aligner des séquences que si elles présentent un minimum d'homologies.

#### 3.2 Alignement de séquences

L'alignement de séquences biologiques homologues (c'est à dire pour lesquelles on a de bonnes raisons de penser qu'elles dérivent d'un ancêtre commun dans l'évolution) permet de faire correspondre les zones les mieux conservées entre différents sous-types ou différentes espèces. Cette approche est très utilisée en génétique moléculaire car elle permet notamment d'établir des arbres phylogéniques (*i.e.* « généalogiques ») et de proposer des modèles d'évolution moléculaire ciblés sur des zones particulières des génomes.

Dans le cas des rétrovirus, des analyses antérieures ont suggéré qu'une bonne part de cette évolution s'est faite par des duplications/délétions mélangées par étapes successives de secteurs courts et de secteurs plus longs. L'existence de motifs consensuels laisse à penser que cette évolution aurait pu se faire à partir d'un ou plusieurs motifs répétés en tandem. Nous avons poursuivi cette étude sur des séquences rétrovirales particulières, celles des virus HIV-1 et HIV-2 qui causent le syndrome de l'immunodéficience acquise humaine et celles de virus simiens apparentés. Comme les autres rétrovirus, les HIV ont un génome ARN monobrin présent sous forme de deux copies identiques dans la particule virale infectieuse. La réplication du virus passe par un ADN double brin (le provirus) qui s'intègre dans le génome de la cellule hôte et qui possède deux extrémités identiques les « long terminal repeats » (LTRs). Nous avons focalisé l'étude sur ces LTRs et les séquences adjacentes. Ces séquences (qui ne dépassent pas 1000 nucléotides) ont l'avantage d'être bien individualisées et de contenir des sites de contrôle de transcription et des zones bien délimitées. Les zones d'intérêt biologique sont souvent bien conservées dans l'évolution et ont permis de construire manuellement un alignement multiple stable qui concerne ici 27 séquences HIV-1, HIV-2 et SIVs et qui peut ainsi servir à évaluer la pertinence biologique des alignements déterminés par des programmes informatiques existants. La figure 1 (voir page 34) illustre les difficultés rencontrées pour construire un alignement – même entre seulement deux séquences. Il s'agit du « dotplot » de deux séquences de LTR  $s_1$  et  $s_2$ , c'est-à-dire une image où le point de coordonnées  $(i, j)$  est noirci si le nucléotide en position  $i$  dans la séquence  $s_1$  est le même que celui en position  $j$  dans  $s_2$ . Un alignement des séquences peut être vu comme un déplacement du coin gauche supérieur vers le coin droit inférieur du dotplot (avec la contrainte que chaque pas du déplacement doit incrémenter au moins une des coordonnées). Un « bon alignement » est vu en général comme un déplacement passant sur un maximum de points noirs (*i.e.* qui met beaucoup de nucléotides en correspondance), et se faisant surtout en diagonale (les cas contraires correspondant à des insertions/délétions).

Nous renvoyons à [8] pour une étude des performances des différents programmes d'alignements. Brièvement, les programmes utilisant le procédé de programmation dynamique (par exemple Clustal-X) sont pris en défaut à partir de la zone où existent les plus importantes délétions/duplications, car ils sont confrontés au problème de l'attribution de pénalités aux insertions/délétions de nucléotides d'une séquence à l'autre. Ce type de programmes est adapté aux évolutions progressives de séquences par mutations ponctuelles et insertions/délétions d'un petit nombre de nucléotides. De meilleurs résultats ont été obtenus avec les programmes calculant des blocs d'homologie (Dialign et Mabios) dont les meilleures combinaisons sont choisies pour sélectionner les points d'ancrage qui permettent de construire le reste de l'alignement. Néanmoins, là encore, tous les résultats doivent être corrigés manuellement.

C'est pour poursuivre le contrôle de l'alignement dont la fiabilité était importante pour des études de modèles d'évolution sur des zones réduites que nous avons réécrit l'alignement nucléotidique multiple sur les séquences recodées à l'ordre 8, puis à l'ordre 12. L'ordre 8 a permis de contrôler des blocs d'homologie plus éloignés et le 12 a permis de confirmer les blocs d'homologie les plus significatifs. Cette bonne correspondance entre un alignement manuel

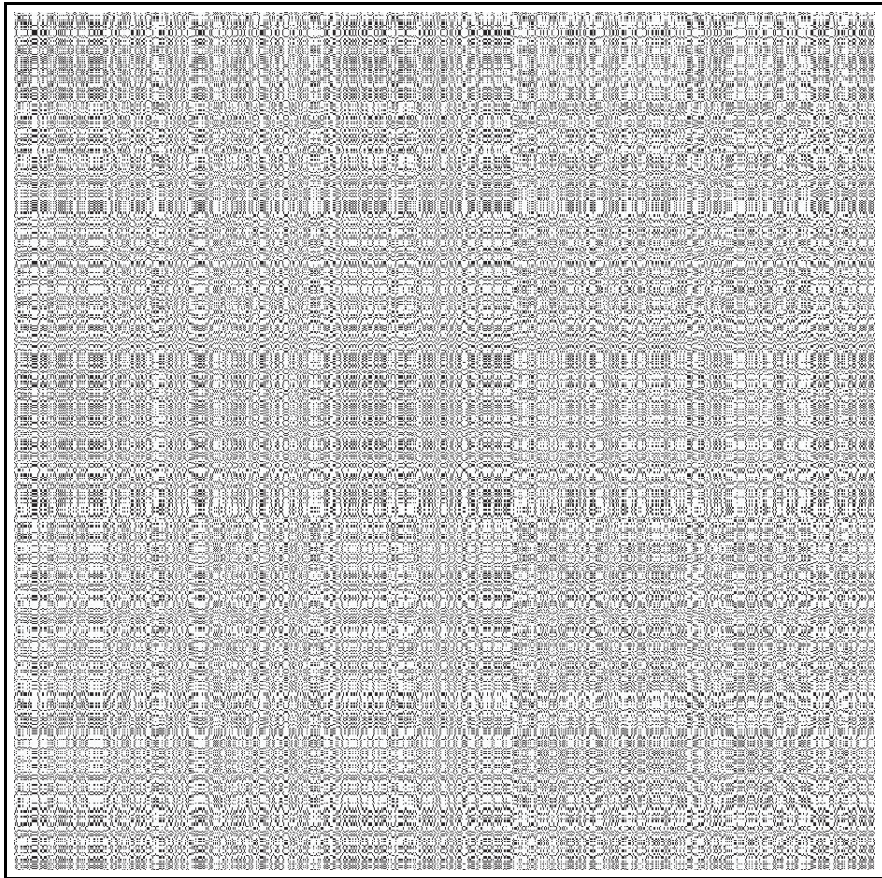


FIG. 1. DotPlot réalisé à partir des séquences brutes des LTR de HIV-1.

fiable (car tenant compte de tout ce que la littérature apprend sur la structure et les fonctions de ces séquences) et les zones conservées entre les séquences recodées est prometteuse dans la perspective de la mise au point de programmes d'alignement utilisant ce recodage. En illustration, la figure 2 est un « dot-plot » correspondant aux mêmes séquences que celui de la figure 1 mais il a été déterminé à partir de leurs recodées à l'ordre 8. Les « points d'ancrages » de l'alignement entre les deux séquences y apparaissent sous forme de diagonales.

Nous développons actuellement un algorithme basé sur le calcul des antécédents maximaux de tous les ordres possibles. Le principe général est, étant données deux séquences  $s_1$  et  $s_2$ , d'associer à chaque couple  $(i, j)$  où  $i$  est une position de  $s_1$  et  $j$  une position de  $s_2$ , un score dépendant du plus grand ordre de codage pour lequel on observe la même lettre sur les deux positions.

### 3.3 Identification de régularités

Les génomes contiennent des régularités au sens large du terme. Les répétitions, qui en sont la manifestation la plus étudiée, possèdent des structures variées. Les unités répétées sont soit contiguës (répétitions en tandem), soit

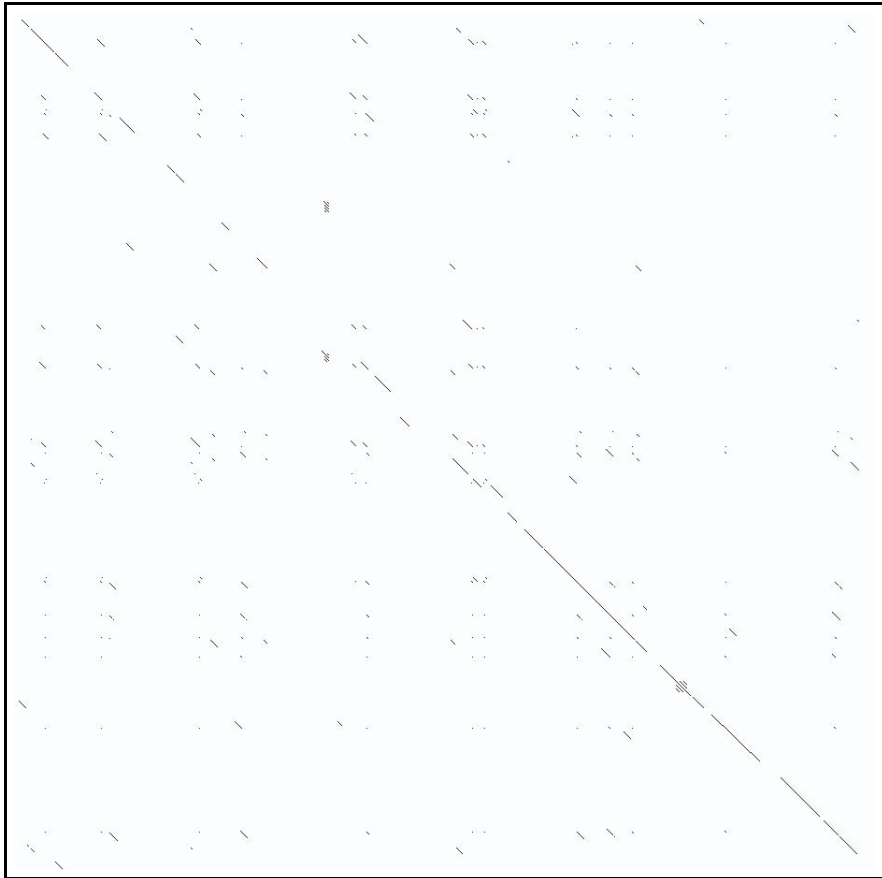


FIG. 2. DotPlot réalisé à partir des recodées à l'ordre 8 des séquences utilisées pour la figure 1.

dispersées dans le génome. La taille d'une unité et son nombre d'occurrence est très variable d'une répétition à une autre (de un à quelques milliers de nucléotides et de deux à plusieurs centaines d'occurrences). Souvent, l'évolution des séquences nucléotidiques introduit des différences entre les unités répétées. Ces erreurs, appelées mutations, sont soit le changement d'un nucléotide en un autre (substitution), soit l'ajout ou la suppression d'un ou plusieurs nucléotides consécutifs (insertion ou délétion). Les répétitions ont des fonctions diverses telles que la protection des extrémités des chromosomes (représentants physiques du génome), la duplication d'un motif important pour la cellule, la protection contre le système immunitaire de l'hôte infecté dans le cas des microbes, l'apparition de maladies génétiques humaines, etc. Certaines répétitions ont une fonction encore inconnue. La recherche exhaustive de répétitions dans les génomes complets aide à mieux comprendre leur mode d'évolution et leurs

fonctions potentielles. L'étude rapportée ici porte sur la Levure de bière, *Saccharomyces Cerevisiae*, un organisme eucaryote<sup>1</sup>.

La recherche de répétitions au sein du génome de *Saccharomyces Cerevisiae* est décomposée en deux étapes [4]. Dans un premier temps, la séquence du génome est parcourue à l'aide de fenêtres de 500 nucléotides, chevauchantes de moitié. Les fenêtres présentant une régularité exceptionnelle sont sélectionnées à l'aide du programme Excep [3] qui permet d'associer à chaque fenêtre un score reflétant la présence de blocs sur-représentés. Lorsque les fenêtres sélectionnées sont chevauchantes, ces fenêtres sont considérées comme appartenant à une seule et même région. L'étude sur *Saccharomyces Cerevisiae* a nécessité 48 268 fenêtres dont seules 230, soit environ 0,5%, ont été sélectionnées. Ces dernières forment 91 régions.

La deuxième étape consiste en la description des régularités présentes dans les régions détectées. L'analyse manuelle des régions est un travail fastidieux et difficile compte tenu de la diversité des répétitions observées. Certaines régions contiennent même plusieurs répétitions différentes, pouvant être enchevêtrées. La plupart des programmes ont un a priori sur la structure des répétitions qu'ils recherchent et le nombre de mutations qu'elles contiennent. Ces contraintes limitent fortement l'étude des régions qui ne contiennent pas nécessairement des régularités classiques. Le recodage s'est révélé être un outil performant pour caractériser des répétitions de structures variées. Le recodage d'une région à l'aide de l'antécédent maximal de sa  $N$ -écriture met en évidence les nucléotides potentiellement impliqués dans une régularité puisque possédant un environnement redondant. La figure 3 en page 37 présente l'interprétation d'un court fragment de séquence.

Nous avons ainsi pu caractériser les répétitions présentes dans les régions détectées chez *Saccharomyces Cerevisiae* [4]. Deux catégories sont distinguées en fonction de la taille des unités répétées : 22 régions contiennent des unités de 3 ou 6 nucléotides et 69 régions des unités de plus de 9 nucléotides (maximum : 285). Un phénomène encore jamais observé jusqu'à présent a été mis en évidence par les recodages : les échos. Il s'agit de blocs de la répétition principale présents autour de celle-ci (la figure 4 en présente un exemple – page 38).

Pratiquement toutes les régions étudiées possèdent des échos. Ceux-ci semblent plus dégénérés et moins nombreux à mesure que l'on s'éloigne de la répétition principale, jusqu'à leur disparition (environ 2000 nucléotides plus loin). Les échos sont sûrement une trace de l'évolution des répétitions. Cette observation, ainsi que la caractérisation de répétitions très dégénérées (possédant de longues insertions/délétions) est due à la capacité du recodage à associer la même lettre à deux nucléotides présents dans des environnements différents au premier abord (voir la seconde partie de la figure 3). La saturation par transitivité effectuée lors de la recherche de l'antécédent maximal intervient de façon essentielle dans ce phénomène. Le recodage permet ainsi de mettre en

---

<sup>1</sup> Les eucaryotes sont des organismes dont les cellules comportent un noyau qui protège le matériel génétique. Certains ne sont composés que d'une cellule, comme la Levure, les autres sont les organismes pluricellulaires dont fait partie l'Homme. *Saccharomyces Cerevisiae* est le premier eucaryote dont le génome a été complètement déterminé.

Cette petite portion de séquence fait partie d'un segment sélectionné :

caaggatgcagtaccactgctgctactaccactgaaagtaccactgctgcagtaccactgccgac

Son recodage à l'ordre 5 est représenté ci-dessous. Pour aider à l'interprétation, les lettres recodées n'admettant qu'une seule occurrence sont représentées par des lettres minuscules correspondant aux nucléotides sur lesquels elles se projettent. Les autres le sont par les lettres majuscules des nucléotides correspondants suivies de numéros en indice pour les distinguer. Chaque ligne représente une unité de la répétition (la seconde unité contient une insertion de 12 nucléotides composée d'une partie de l'unité initiale).

caagga T <sub>0</sub> G <sub>0</sub> C <sub>0</sub>	A <sub>0</sub> G <sub>0</sub> T <sub>1</sub> A <sub>1</sub> C <sub>1</sub> C <sub>2</sub> A <sub>2</sub> C <sub>0</sub> T <sub>0</sub> G <sub>0</sub> C <sub>0</sub>
T <sub>0</sub> G <sub>0</sub> C <sub>0</sub> T <sub>0</sub> ac T <sub>1</sub> A <sub>1</sub> C <sub>1</sub> C <sub>2</sub> A <sub>2</sub> C <sub>0</sub> T <sub>0</sub> G <sub>0</sub> a A <sub>0</sub> G <sub>0</sub> T <sub>1</sub> A <sub>1</sub> C <sub>1</sub> C <sub>2</sub> A <sub>2</sub> C <sub>0</sub> T <sub>0</sub> G <sub>0</sub> C <sub>0</sub>	
T <sub>0</sub> G <sub>0</sub> C <sub>0</sub>	A <sub>0</sub> G <sub>0</sub> T <sub>1</sub> A <sub>1</sub> C <sub>1</sub> C <sub>2</sub> A <sub>2</sub> C <sub>0</sub> T <sub>0</sub> G <sub>0</sub> C <sub>0</sub> cgac

Observons les différents contextes du nucléotide C<sub>0</sub> (les indices des autres lettres recodées ont été supprimés pour plus de clarté) :

1. gaTG C<sub>0</sub> AGTA
2. ACCA C<sub>0</sub> TGCT
3. ACTG C<sub>0</sub> TGCT
4. GCTG C<sub>0</sub> TacT
5. ACCA C<sub>0</sub> TGaA
6. ACTG C<sub>0</sub> TGCA
7. GCTG C<sub>0</sub> AGTA
8. ACCA C<sub>0</sub> TGcC
9. ACTG C<sub>0</sub> cgac

En particulier, les blocs incluant les occurrences 1 et 2 de C<sub>0</sub> sont totalement différents. Pourtant, dans toutes ses occurrences, C<sub>0</sub> peut être vu comme appartenant à une portion de répétition d'un bloc du type TGC (dans le premier cas TGC<sub>0</sub> et dans le second C<sub>0</sub>TGCT). D'un point de vue biologique, il est pertinent de considérer ces deux contextes comme équivalents.

FIG. 3. Un exemple d'interprétation d'un fragment de séquence sélectionné.

```

tacagaagtctgctcccatgaggcatgctccttcgcatcgacggtgcaa
ccaccaccttatctgtgactt
                                CC
AAGTTCA
cttcatatatttgcctacttgtcacacaaccgctatcagctcattatcc
gaagtaggaactacaaccgtggatcatccagcgccattgaaccatcaag
tgcctctataatctcacctgtcacctctacactttcgagtacaacatcgt
ccaatccaactactacctccct
AAGTTGACATCTACATCTCC
AAGCTCTACATCTACATCTCC
AAGCTCTACATCTACCTCATC
AAGTTGACATCTACCTCATC
AAGTTGACATCTACCTCATC
AAGTTGACATCTACATCTCC
AAGTTGACATCCACATCTTC
AAGTTTGACATCCACATCTTC
AAGTTCTACATCTACATCCCA
AAGTTCTACATCTACCTCATC
AAGTTGACATCTACATCTCC
AAGCTCTACATCTACCTCATC
AAGTTCAACATCTACATCTCC
AAGTTC
taaactacttctgcaagctccacttccactac
    TTCTTCAT
at
    TCAACATCTACATCCCC
AAGTTTGACTTCTTCAT
ctccaactttggcttccact
                                TCTCC
AAGTTCAACATCTA
ttagctctacttttactgattcaacttcatcccttggctcctctatagca
tcttcatcaacgtctgtgtcattatacagcccatccacacctgtttactc
cgtccc

```

FIG. 4. Un exemple de répétitions perturbées avec échos tiré du génome de *Saccharomyces Cerevisiae*. L'interprétation a été faite à l'aide d'un recodage à l'ordre 9. Les indices distinguant les lettres recodées ont été supprimés et la séquence a été mise en forme (les lettres majuscules correspondent aux nucléotides faisant partie de la répétition et des sauts de lignes ont été ajoutés avant chaque unité répétée, avec des espaces lorsque celle-ci était tronquée).

évidence des répétitions plus complexes que celles vues par une simple étude de blocs avec erreurs.

*Gilles Didier*, Université Évry-Val-Essonne  
*Ivan Laprevotte*, Université Évry-Val-Essonne  
*et Maude Pupin*, Université de Lille

### Références

- [1] G. Didier, *Contribution à l'étude des suites symboliques — Applications aux séquences génétiques*, Thèse, Université de Provence.
- [2] G. Didier, Caractérisation des  $N$ -écritures et application à l'étude des suites de complexité ultimement  $n+cste$ , *Theoret. Comput. Sci.* **215** (1999), no. 1-2, 31–49.
- [3] M. Klaerr-Blanchard, H. Chiapello et E. Coward, Detecting localized repeats in genomics sequences : a new strategy and its application to *B. subtilis* and *A. thaliana*, *Comput. Chem.* **24** (2000), 57–70.
- [4] M. Klaerr-Blanchard, *Etude de répétitions locales et approximatives et élaboration d'une base de données génomiques*, Thèse, Université de Versailles Saint-Quentin.
- [5] I. Laprevotte, M. Pupin, E. Coward, G. Didier, C. Terzian, C. Devauchelle et A. Hénaut, HIV-1 and HIV-2 LTR Nucleotide Sequences : Assessment of the Alignment by N-block presentation ; « Retroviral Signature » of Overrepeated Oligonucleotides, and a Probable Important Role of Scrambled Stepwise Duplications/Deletions in Molecular Evolution, *Molecular Biology and Evolution* **18** (7) (2001), 1231–1245.
- [6] D. Lind et B. Marcus, *An Introduction to Symbolic Dynamics and Coding*, Cambridge University Press (1995).
- [7] F. Muri-Majoube et B. Prum, Une approche statistique de l'analyse des génomes, *Gazette des Mathématiciens* **89** (2001).
- [8] J. D. Thompson, F. Plewniak et O. Pooch, A comprehensive comparison of multiple sequence alignment programs, *Nucleic Acid Res.* **27** (1999), 2682–2690.