Rubrique

À la recherche de mots de fréquence exceptionnelle dans les génomes

INRA, Unité Mathématique, Informatique & Génome, 78352 Jouy-en-Josas

La compréhension de l'information génétique portée par les génomes est un défi pour les biologistes, les biophysiciens, les informaticiens et les mathématiciens. Un des problèmes classiques en bioinformatique est l'identification de "mots" qui apparaissent avec une fréquence inattendue dans ces longues suites de lettres à valeur dans l'alphabet $\{a, c, g, t\}$ que sont les séquences d'ADN. Ces mots exceptionnels peuvent en effet être liés à des mécanismes biologiques cruciaux pour la cellule. Nous présentons ici la démarche du statisticien face à cet enjeu.

L'information génétique de chaque être vivant est portée par son ADN. Celui-ci est une longue succession de nucléotides. Il existe quatre nucléotides différents selon qu'ils portent l'une des bases *adénine*, *cytosine*, *guanine* ou *thymine*. Ainsi, une manière d'appréhender l'ADN est de le considérer comme un texte écrit sur un alphabet à quatre lettres $\{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$. Nous allons nous intéresser ici aux mots, courtes suites de nucléotides, contenus dans une séquence d'ADN donnée, et plus particulièrement à leur fréquence d'apparition.

MOTS EXCEPTIONNELS

Chaque occurrence d'un mot peut être reconnue par une enzyme et ainsi participer à un mécanisme biologique. C'est le cas par exemple des sites de restriction chez les bactéries, généralement constitués de 6 nucléotides (ou lettres), qui constituent des points de cassure de l'ADN dès qu'ils sont reconnus par une enzyme de restriction spécifique. Il n'est donc pas surprenant que ces sites soient spécialement évités dans les génomes bactériens car ils fragilisent l'ADN. À l'inverse, certains mots sont primordiaux pour garantir une certaine stabilité du génome et sont donc présents en grand nombre le long de l'ADN. C'est le cas du mot `gctggtgg` qui est très fréquent dans le génome de la bactérie *E. coli*. On y reviendra à la fin de ce texte. Rechercher des mots significativement rares ou fréquents, ou détecter si un mot donné est exceptionnel dans une séquence est donc devenu un problème classique en génomique. Voyons comment poser le problème de façon statistique.

PROBLÈME STATISTIQUE

On souhaite savoir si un mot W donné, sur l'alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$, a une fréquence d'apparition inattendue dans une séquence d'ADN donnée de longueur n, c'est à dire soit trop forte, soit trop faible. Pour cela, il faut définir ce qu'est une valeur attendue ou espérée du nombre d'occurrences du mot W dans une séquence de longueur n et la variabilité autour de cette valeur. De façon plus précise, il faut déterminer la loi probabiliste du comptage ; ceci nécessite de définir un modèle. Un modèle est d'une part l'ensemble de toutes les séquences possibles dont la séquence d'ADN observée n'est qu'une réalisation, et d'autre part une loi de probabilité sur cet ensemble. On verra que le modèle choisi a une influence sur le caractère exceptionnel d'un mot. La significativité de la fréquence observée n'a en effet de sens que par rapport à une fréquence attendue, celle déterminée par le modèle. En changeant de modèle, on modifie la fréquence attendue, et l'écart entre la fréquence observée et celle attendue peut devenir significatif, ou au contraire ne plus l'être. Si on fait l'analogie avec un texte en français, la combinaison de lettres "tra" est beaucoup plus fréquente que "tsa" car "tr" est déjà plus fréquente que "ts". De même, il est tout à fait possible que sur un génome `tgg` soit beaucoup plus fréquent que `tcg` car `tg` serait plus fréquent que `tc`. Lorsqu'on souhaite étudier de courtes séquences d'ADN comme des mots il est donc probablement pertinent d'utiliser des modèles qui prennent en compte la fréquence des sous-mots qui les composent. On verra qu'une des réponses est d'utiliser des modèles de chaînes de Markov.

Une fois le modèle choisi (chaîne de Markov d'ordre m par exemple), l'objectif est d'évaluer la probabilité $\mathbb{P}(N(W) \geq c(W))$, communément appelée p-valeur, où $c(W)$ est le comptage observé du mot W dans la séquence d'ADN de longueur n et $N(W)$ est la variable aléatoire qui représente le nombre d'occurrences de W dans une chaîne de Markov d'ordre m et de longueur n. En effet, si la p-valeur est très proche de 0 alors le mot W sera considéré comme exceptionnellement fréquent — ou significativement sur-représenté — (il y a une probabilité quasi nulle de l'avoir observé autant de fois). Au contraire si la p-valeur est très proche de 1 alors la probabilité $\mathbb{P}(N(W) < c(W))$ est très proche de 0 et le mot est exceptionnellement rare — ou significativement sous-représenté — (il y a une probabilité quasi nulle de l'avoir observé si peu de fois).

ÉTAT DE L'ART

Plusieurs méthodes ont été développées pour évaluer cette p-valeur. L'une des plus récentes repose sur le calcul de la loi exacte du comptage $N(W)$, c'est à dire des probabilités $\mathbb{P}(N(W) = x)$ pour tout entier x. Cette loi est déterminée par sa fonction génératrice et les probabilités ponctuelles peuvent s'obtenir par une récurrence. Cependant, cette méthode est pour le moment coûteuse en temps de calcul pour des longues séquences (et des mots fréquents) et n'est pas disponible en pratique pour des chaînes de Markov d'ordre $m \geq 2$, ce qui est limitant pour l'analyse des mots exceptionnels comme on le verra plus tard.
Auparavant, deux approximations de la loi du comptage ont été proposées pour des séquences suffisamment longues ($n \to \infty$) : une approximation gaussienne valable pour des mots relativement fréquents (des mots pas trop longs par rapport à la longueur de la séquence), et une approximation par une loi de Poisson composée valable pour des mots rares (mots relativement longs par rapport à la longueur de la séquence). Sans rentrer dans les détails, on peut avoir une idée intuitive de ces deux résultats asymptotiques. On verra que le comptage $N(W)$ du mot W n'est rien d'autre qu'une somme de variables aléatoires de Bernoulli Y_i non indépendantes, de même moyenne $\mu(W)$, qui valent 1 si le mot apparaît en position i dans la séquence ou 0 sinon. Si ces variables aléatoires étaient indépendantes, le comptage suivrait alors une loi Binomiale $\mathcal{B}(n, \mu(W))$ qui s'approche selon des résultats classiques soit par une loi normale si $n\mu(W) \to \infty$, soit par une loi de Poisson si $n\mu(W)$ tend vers une constante quand $n \to \infty$ (notons que $n\mu(W)$ est précisément le comptage moyen). La difficulté ici réside dans le fait que les variables aléatoires Y_i ne sont pas indépendantes. Toutefois, il est démontré qu'on peut approcher la loi du comptage par une loi gaussienne ou par une loi de Poisson composée selon les cas (quelques détails vont suivre) ; le prix à payer sera de tenir compte de la structure auto-recouvrante des mots (`tttt` est par exemple très recouvrant tandis que `ttac` ne l'est pas).

MODÈLES POUR LES SÉQUENCES

Le choix du modèle correspond ici à définir une séquence S de n variables aléatoires $X_1, X_2, \ldots, X_n$, à valeur dans l'alphabet $\mathcal{A} = \{\mathtt{a}, \mathtt{c}, \mathtt{g}, \mathtt{t}\}$ qui “ressemblerait” en un certain sens à la séquence d'ADN donnée. Ceci nous permettra de comparer ce qui est observé dans la séquence étudiée à ce à quoi l'on s'attendrait dans une séquence aléatoire générée selon le modèle. Rappelons ici que l'objectif n'est pas de modéliser au mieux une séquence d'ADN mais précisément de construire un modèle aléatoire qui prenne en compte certaines contraintes, certaines informations sur la séquence étudiée, pour détecter des écarts au modèle c'est à dire des événements exceptionnels compte tenu des contraintes déjà prises en compte.

Le modèle de Bernoulli

Un premier modèle consiste à supposer que les lettres X_i sont indépendantes les unes des autres et que les bases `a`, `c`, `g` et `t` apparaissent avec les probabilités $\mu(\mathtt{a}), \mu(\mathtt{c}), \mu(\mathtt{g})$ et $\mu(\mathtt{t})$ (la somme étant égale à 1). Notons M0 ce modèle. Dans le modèle M0, la vraisemblance, c'est à dire la probabilité de la suite $S = X_1 X_2 \cdots X_n$, s'écrit

$$\prod_{b \in \mathcal{A}} (\mu(b))^{N(b)}$$

où $N(b)$ est le nombre de b dans la séquence S. La statistique exhaustive du modèle M0 est donc le comptage de chacune des 4 bases. Choisir le modèle M0 consiste à ne retenir de la séquence que sa composition en bases `a`, `c`, `g` et `t`. Les paramètres $\mu(b)$ étant inconnus dans notre cas, on les estime au vu d'une séquence observée par ceux qui maximisent la vraisemblance ci-dessus, c'est à dire par les proportions : $\widehat{\mu}(b) = \frac{N(b)}{n}$. Ainsi en moyenne, le nombre de b dans les séquences aléatoires générées sous ce modèle M0 sera égal au nombre de b dans la séquence d'ADN.

Les chaînes de Markov

Un modèle qui permettrait de plus de prendre en compte la fréquence des 16 mots de longueur 2, aa, ac, …, tc et tt, est le modèle de chaîne de Markov stationnaire d'ordre 1, noté M1 dans la suite. En effet, notons $\mu(\cdot)$ la mesure stationnaire sur $\mathcal{A}$ et $\pi(\cdot,\cdot)$ les probabilités de transition de la chaîne ; la vraisemblance s'écrit alors

$$\mu(X_1)\prod_{a,b\in\mathcal{A}}\Big(\pi(a,b)\Big)^{N(ab)}$$

où $N(ab)$ est le nombre de mots ab dans la séquence S. La statistique exhaustive du modèle M1 est composée de la première lettre X_1 de la séquence et de la collection des comptages des 16 mots de longueur 2 sur l'alphabet $\mathcal{A}$. Les estimateurs du maximum de vraisemblance des paramètres sont $\widehat{\mu}(a)=\frac{N(a)}{n}$ et

$$\widehat{\pi}(a,b)=\frac{N(ab)}{\sum_{b\in\mathcal{A}}N(ab)}\simeq\frac{N(ab)}{N(a)}$$

c'est-à-dire que la probabilité $\pi(a,b)$ qu'un a soit suivi d'un b est estimée par la proportion de a suivis d'un b.

De la même façon, le modèle de chaîne de Markov stationnaire d'ordre m, Mm, permettra de prendre en compte la composition de la séquence d'ADN en mots de longueurs 1 à $(m+1)$. Un mot W de longueur h peut donc être étudié dans les modèles Mm avec $0\leq m\leq h-2$. Le modèle d'ordre $m=h-2$ est dit "maximal" pour étudier les mots de longueur h. C'est celui qui prend en compte la composition de la séquence d'ADN en mots de longueur $h-1$.

Pour tenir compte d'une réalité biologique, à savoir que certaines séquences d'ADN (les gènes) codent pour des protéines et sont naturellement lues de 3 bases en 3 bases (les codons), il est souvent pertinent d'utiliser un modèle où par exemple la probabilité de transition de gtcg vers a (modèle M4) diffère selon que le a est la première, deuxième ou troisième base d'un codon. Les probabilités de transition sont alors périodiques de période 3, ce qui revient en fait à avoir 3 matrices de transition.

Dans la suite nous nous placerons dans le modèle M1 d'ordre 1.

NOMBRE D'OCCURRENCES ATTENDU

Soit $W=w_1w_2\cdots w_h$ un mot de longueur h sur l'alphabet $\mathcal{A}$. On note Y_i la variable aléatoire qui vaut 1 si une occurrence de W commence en position i dans la séquence S et 0 sinon. Le nombre $N(W)$ d'occurrences du mot W dans la séquence S de longueur n est donc défini par

$$N(W)=\sum_{i=1}^{n-h+1}Y_i.$$

Le nombre moyen d'occurrences de W dans S sous le modèle M1 (ou espérance) vaut $\mathbb{E}_1 N(W)=(n-h+1)\mu_1(W)$ où $\mu_1(W)$ est la moyenne des variables de Bernoulli Y_i, c'est à dire la probabilité d'observer le mot W à une position donnée dans une chaîne de Markov d'ordre 1. Cette probabilité s'écrit de la façon suivante en fonction des probabilités de transition et de la loi stationnaire :

$$\mu_1(W)=\mu(w_1)\prod_{j=1}^{h-1}\pi(w_j,w_{j+1}).$$

En pratique, les paramètres du modèle sont inconnus et estimés, comme on l'a vu plus haut, à partir de la séquence d'ADN observée ; ceci nous permet d'obtenir un estimateur $\widehat{\mathbb{E}}_1(N(W))$ du comptage moyen, noté plus simplement $\widehat{N}_1(W)$, donné par :

$$\widehat{N}_1(W)=\frac{\prod_{j=1}^{h-1}N(w_jw_{j+1})}{\prod_{j=2}^{h-1}N(w_j)}.$$

Par exemple, le comptage moyen estimé sous le modèle M1 du mot atc vaut $N(\mathtt{at})N(\mathtt{tc})/N(\mathtt{t})$.

SIGNIFICATIVITÉ DU COMPTAGE

Voyons maintenant un peu plus en détail comment calculer ou approcher la p-valeur $\mathbb{P}(N(W)\geq c(W))$ sous le modèle M1.

Approximation gaussienne

Dans le cas d'un mot W fixé, l'approximation gaussienne du comptage $N(W)$ permet de déterminer si l'écart entre $N(W)$ et $\widehat{N}_1(W)$ (comptage estimé dans M1) est significativement grand en valeur absolue. En effet, lorsque l'on normalise cet écart par un estimateur de son écart type, on montre que la statistique $U_1(W)$ ainsi formée :

$$U_1(W)=\frac{N(W)-\widehat{N}_1(W)}{\widehat{\sigma}_1(W)}\qquad(1)$$

converge en loi vers une variable aléatoire gaussienne d'espérance 0 et de variance 1. Nous ne donnerons pas ici l'expression de $\widehat{\sigma}_1(W)$ mais faisons simplement remarquer qu'elle fait intervenir la structure d'auto-recouvrement du mot W.

La statistique $U_1(W)$ nous permet ainsi d'évaluer la significativité de l'écart $N(W)-\widehat{N}_1(W)$ compte tenu de la fréquence des mots de longueur 2. En effet, il suffit de calculer la probabilité $\mathbb{P}(X\geq U_1^{\mathrm{obs}}(W))$ pour une variable aléatoire X distribuée suivant la loi gaussienne centrée réduite, et $U_1^{\mathrm{obs}}(W)$ calculée avec la valeur observée de $N(W)$ dans la séquence. Ainsi lorsque $U_1^{\mathrm{obs}}(W)$ est négative et grande en valeur absolue, le mot W est un mot exceptionnellement rare pour le modèle ; inversement, si la statistique $U_1^{\mathrm{obs}}(W)$ est positive et grande en valeur absolue, W est un mot exceptionnellement fréquent pour le modèle.

Encadré 1

NORMALITÉ ASYMPTOTIQUE DU COMPTAGE

La normalité asymptotique de $n^{-1/2}(N(W) - \mathbb{E}N(W))$ s'obtient par un Théorème de Limite Centrale pour chaînes de Markov. Celle de $n^{-1/2}(N(W) - \widehat{N}_1(W))$ en découle par la δ-méthode puisque $\widehat{N}_1(W)$ est une fonction de certains comptages, eux aussi asymptotiquement gaussiens. Cependant, cette méthode est trop fastidieuse pour nous donner la variance asymptotique. La variance de l'écart $N(W) - \widehat{N}_1(W)$ n'est en effet pas égale à la variance de $N(W)$ du fait du caractère aléatoire de l'estimateur $\widehat{N}_1(W)$. La variance asymptotique de $n^{-1/2}(N(W) - \widehat{N}_1(W))$ est en fait égale à la limite de $n^{-1}\, Var(N(W) \mid X_1, N(ab), a, b \in \mathcal{A})$ où $\{X_1, N(ab), a, b \in \mathcal{A})\}$ est la statistique exhaustive du modèle M1.

Approximation par une loi de Poisson composée

La méthode de Chen-Stein permet de mesurer l'erreur commise lorsque l'on approche une somme de variables aléatoires de Bernoulli dépendantes par une loi de Poisson. C'est ainsi une généralisation de l'approximation d'une loi binomiale par une loi de Poisson. La loi de Poisson étant la "loi des événements rares", il n'est pas surprenant de devoir ici se placer dans le cas où le mot W a un comptage attendu de l'ordre de 1 quand la longueur n de la séquence croît vers l'infini. Sous cette condition, qui est équivalente à considérer un mot dont la longueur est de l'ordre de $\log n$, on montre que l'approximation de la loi du comptage par la loi de Poisson de même moyenne n'est valable que pour des mots non auto-recouvrants. Pour les mots auto-recouvrants, les indicatrices d'occurrences Y_i ne peuvent donc pas être considérées comme indépendantes même asymptotiquement. Ceci vient simplement du fait que les occurrences d'un mot auto-recouvrant ont tendance à former des paquets, ou "trains", d'occurrences chevauchantes. En fait, la méthode de Chen-Stein nous permet de montrer que le nombre de trains, noté $\widetilde{N}(W)$, peut lui s'approcher par une variable de Poisson. De plus, les "tailles" des trains (nombres de "wagons" W) sont indépendantes et de même loi (une loi géométrique dont le paramètre est connu et bien sûr lié à la structure d'auto-recouvrement du mot). Ainsi, en écrivant le comptage $N(W)$ comme la somme sur les trains de W de la taille des trains :

$$N(W) = \sum_{j=1}^{\widetilde{N}(W)} \{\text{taille du } j^{\text{ième}} \text{ train}\}$$

on obtient par définition une loi de Poisson composée comme loi limite. La p-valeur sera alors approchée par la queue de la loi de Poisson composée ad-hoc.

Encadré 2

MÉTHODE DE CHEN-STEIN

Cette méthode donne une majoration de l'erreur commise lorsque l'on remplace la loi d'une somme de n variables aléatoires de Bernoulli Y_i non indépendantes et de paramètre p_i, par la loi de Poisson de même espérance $\lambda := \sum_{i=1}^n p_i$. Cette erreur est mesurée en terme de distance en variation totale entre les deux lois ; cette distance se définit pour deux variables aléatoires discrètes N et Z par :

$$d_{VT}(\mathcal{L}(N), \mathcal{L}(Z)) := 2 \sup_{A \subset \mathbb{N}} |\mathbb{P}(X \in A) - \mathbb{P}(Z \in A)|.$$

Soient donc $N = \sum_{i=1}^n Y_i$ et Z une variable aléatoire de Poisson de paramètre λ. La borne de la distance en variation totale entre les lois de N et Z va s'exprimer en fonction du niveau de dépendance entre les variables de Bernoulli Y_i dont les indices font partie d'un certain voisinage (à définir) B_i de i.
Le théorème est le suivant :

$$d_{VT}(\mathcal{L}(N), \mathcal{L}(Z)) \leq 2(b_1 + b_2 + b_3)$$

avec

$$\begin{aligned} b_1 &= \sum_{i=1}^n \sum_{j \in B_i} p_i p_j \\ b_2 &= \sum_{i=1}^n \sum_{j \in B_i \setminus \{i\}} \mathbb{E}(Y_i Y_j) \\ b_3 &= \sum_{i=1}^n \mathbb{E}|\mathbb{E}(Y_i - p_i | \sigma(Y_j, j \notin B_i))|. \end{aligned}$$

Comparaison des méthodes

Si la loi exacte du comptage n'est généralement pas accessible en pratique, il est important de pouvoir juger de la qualité des deux approximations. Outre une comparaison théorique des 3 lois, nous avons comparé, à partir de l'analyse d'une séquence d'ADN particulière, les p-valeurs obtenues en utilisant la loi exacte (en négligeant l'estimation des paramètres), la loi gaussienne limite et la loi de Poisson composée limite. Comme on s'y attendait théoriquement, la loi gaussienne est à éviter lorsque les mots sont de plus en plus longs c'est à dire de plus en plus rares. De façon surprenante, la loi de Poisson composée semble donner de très bons résultats même pour des mots a priori non rares (cf. encadré pour plus de détails).

Encadré 3

COMPARAISON DES MÉTHODES

En pratique, l'utilisation de la loi exacte du comptage pour calculer les p-valeurs est souvent rédhibitoire compte tenu des longueurs de séquences d'ADN manipulées. A titre d'exemple, il faut 6 heures pour l'étude des mots de longueur 9 dans une séquence de seulement 50 000 bases sous le modèle M0 (le temps passe à 44 heures dans M1), et seulement une vingtaine de secondes en utilisant les lois limites gaussienne ou de Poisson composée. Il est alors important de juger des qualités des p-valeurs approchées calculées comme alternatives. La figure ci-contre montre que l'approche gaussienne (première colonne) est adaptée pour les mots courts mais devient très mauvaise au fur et à mesure que la longueur de mots augmente. La loi de Poisson composée (deuxième colonne) donne quant à elle de bons résultats. Une approximation de la p-valeur par des techniques de Grandes Déviations (non détaillées ici) semble aussi donner de bons résultats pour les mots exceptionnels. Un point important pour les biologistes est que quelque soit la méthode utilisée, les mots sont quasiment classés dans le même ordre d'exceptionnalité. Cela leur permet, grâce à la loi gaussienne ou de Poisson composée, de dégager efficacement un lot de mots exceptionnels pour lesquels un calcul exact de la p-valeur peut être effectué.

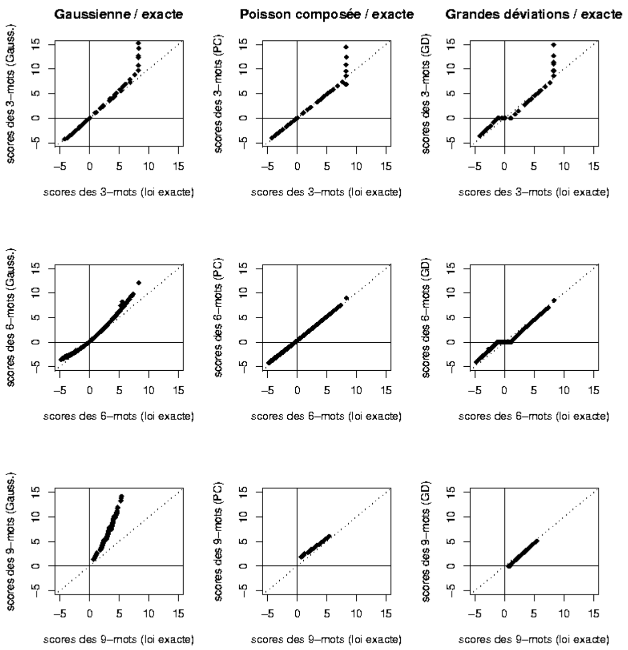

***Figure** 1 – Comparaison des p-valeurs exactes (en abscisse) et des p-valeurs approchées (en ordonnée) obtenues sous le modèle M0 pour les mots de longueur 3 (ligne 1), 6 (ligne 2) et 9 (ligne 3) dans le génome complet du phage Lambda ($n = 48\,502$). Les 3 colonnes correspondent successivement aux approches gaussienne, de Poisson composée et de grandes déviations. Les p-valeurs sont représentées sous la forme d'un score gaussien centré réduit.*

INFLUENCE DU MODÈLE

Le caractère exceptionnel d'un mot dans une séquence est relatif à la quantité d'information prise en compte sur la séquence, c'est à dire au modèle choisi. C'est en effet le modèle qui détermine le comptage attendu, c'est à dire le comptage de référence. Changer de modèle peut alors modifier les résultats.

En guise d'illustration, regardons la fréquence des tri-nucléotides (mots de longueur 3) dans une séquence de longueur 111 402 contenue dans le génome de *E. coli* (>4,6 millions de bases) en se plaçant successivement dans les modèles M0 et M1. On va voir que les mots ne seront pas exceptionnels de la même façon dans les deux modèles, pour la simple raison que le modèle M1 prend en compte le biais éventuel de composition en di-nucléotides (mots de longueur 2) de la séquence, alors que le modèle M0 n'intègre que le biais de composition en bases.

On associe à chacun des 64 tri-nucléotides abc le couple $(U_0(abc), U_1(abc))$, représentant leurs statistiques asymptotiquement gaussiennes d'espérance 0 et d'écart type 1 (cf. eq. (1)) dans les modèles M0 et M1. Par conséquent, une forte valeur positive de $U(abc)$ identifie un mot abc exceptionnellement fréquent, tandis qu'une forte valeur négative de $U(abc)$ identifie un mot abc exceptionnellement rare.

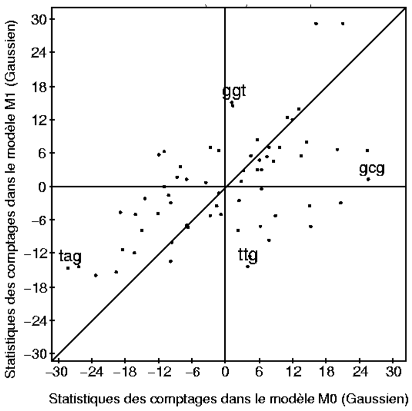

Figure 1 – *Exceptionnalité des tri-nucléotides dans la séquence ECOMORI de E. coli respectivement sous les modèles M0 (en abscisse) et M1 (en ordonnée).*

La figure ci-dessus représente les 64 tri-nucléotides *abc* par un point dont les coordonnées sont $(U_0(abc), U_1(abc))$. Tous les points n'étant pas alignés sur la bissectrice, les statistiques sont donc bien différentes lorsque l'on change de modèle ; le modèle choisi a par conséquent une importance dans l'interprétation des résultats. Certains mots peuvent conserver leur exceptionnalité en augmentant l'ordre du modèle (c'est le cas de `tag` qui est exceptionnellement rare dans M0 et M1) tandis que d'autres peuvent la perdre : c'est la contamination par les sous-mots. Prenons l'exemple de `gcg` ; si l'on tient compte uniquement de la composition en bases (M0), `gcg` est attendu 2105 fois or il est présent 3194 fois et l'écart est très significatif. `gcg` est donc exceptionnellement fréquent sous M0. Si maintenant on tient compte de la composition en di-nucléotides (M1), `gcg` est attendu 3148 fois ce qui n'est pas significativement inférieur au comptage observé. Cette perte d'exceptionnalité en augmentant l'ordre du modèle traduit le fait que la fréquence élevée de `gcg` est simplement due au fait qu'il est composé d'un ou deux mots de longueur 2 fréquents. En effet, `gc` (resp. `cg`) est exceptionnellement fréquent dans M0, mais est "normalement" suivi (resp. précédé) d'un `g`.

Au contraire, l'exceptionnalité de certains mots peuvent être masqués dans des petits modèles (c'est le cas de `ggt` qui n'est pas exceptionnel compte tenu de la composition en bases mais devient significativement sur-représenté dans le modèle M1). C'est en général dû à des phénomènes de compensation entre les sous-mots. Le cas de `ttg` est encore plus surprenant car dans M0 il est sur-représenté (sans être particulièrement exceptionnel) : il est attendu 1758 et apparaît 1904 fois, alors que dans M1 il est finalement attendu 2396 fois, et `ttg` est considéré comme exceptionnellement rare. Ici, `tt` et `tg` sont plutôt sur-représentés dans la séquence mais une contrainte semble qu'ils ne "doivent pas" être juxtaposés. Pourquoi ? C'est une question adressée aux biologistes ...

À travers ces exemples, on s'aperçoit de l'utilité d'étudier la fréquence d'un mot dans plusieurs modèles plutôt que de se restreindre à un seul. On obtient ainsi une information très précise sur la structure de la séquence ou encore de son "vocabulaire". Se placer dans le modèle d'ordre maximal est le meilleur moyen pour détecter si un mot est vraiment exceptionnel de par lui-même (c'est le cas par exemple de `ggt`, `ttg` et `tag`), mais il a l'inconvénient de masquer des mots qui seraient exceptionnels uniquement car ils contiennent un sous-mot exceptionnel (c'est le cas par exemple de `gcg`).

EXEMPLE DE MOT EXCEPTIONNEL

Chez *E. coli*, la séquence `gctggtgg`, appelée Chi, interagit avec une enzyme, appelée RecBCD, capable de dégrader très efficacement l'ADN double brin (provenant par exemple d'une cassure du chromosome de la bactérie ou provenant d'un virus l'ayant infecté). La séquence Chi module en fait les fonctions de RecBCD : lorsqu'elle n'est pas présente, la fonction "dégradation" de RecBCD est active et la molécule d'ADN est efficacement détruite. Si au contraire la séquence Chi est présente, celle-ci est reconnue par RecBCD et une fonction "réparation" est activée. Ce phénomène conduit les biologistes à proposer que Chi pourrait être particulièrement fréquente sur le génome d'*E. coli* afin d'assurer la protection du génome et la réparation en cas de cassure. Au contraire, l'ADN d'un virus ne comporterait probablement pas de séquence Chi et serait donc dégradé dès son entrée dans la cellule ce qui protégerait *E. coli*. Il s'agit donc de savoir si la séquence Chi est significativement fréquente sur le génome de la bactérie.

Nous avons donc analysé le génome complet de *E. coli*, pris dans le sens de la réplication (biais dans l'activité de Chi), long de 4 638 858 bases. La séquence `gctggtgg` y est présente 762 fois. Le tableau ci-dessous rassemble, pour chaque modèle d'ordre $m \in \{0, \dots, 6\}$, le comptage attendu $\widehat{N}_m$ de ce mot, le carré $\widehat{\sigma}_m^2$ du facteur de normalisation, la statistique asymptotiquement gaussienne U_m et le rang de cette statistique parmi celles des 65 536 mots de longueur 8 classées par ordre décroissant.

m	$\widehat{N}_m$	$\widehat{\sigma}_m^2$	U_m	rang
0	85.9	85.8	72.96	3
1	84.9	84.8	73.54	1
2	206.8	203.9	38.88	1
3	355.5	338.9	22.08	5
4	355.3	314.4	22.94	2
5	420.9	298.0	19.76	1
6	610.1	203.3	10.65	3

Figure 2 – *Statistiques de* `gctggtgg` *dans le génome de E. coli sous différents modèles Mm.*

On s'aperçoit que quelque soit l'ordre du modèle choisi, Chi est significativement sur-représenté et est toujours parmi les 5 mots de longueur 8 les plus exceptionnellement fréquents. La contrainte est donc forte pour que ce mot soit fréquent : la composition en mots de longueur 7 (M6) ne suffit pas à expliquer les 762 occurrences de Chi.

Cela conduit à proposer qu'une telle répartition de la séquence Chi sur le génome a été sélectionnée au cours de l'évolution. Les résultats de l'analyse statistique semblent donc confirmer les résultats de l'analyse génétique qui accorde à la séquence Chi et à l'enzyme RecBCD, son partenaire, un rôle fondamental dans un processus qui permet à la fois de réparer le génome propre de la bactérie et de dégrader l'ADN étranger.

POUR EN SAVOIR PLUS

R. Arratia, L. Goldstein, L. Gordon, « Two moments suffice for Poisson approximations : the Chen-Stein method », *Ann. Prob.*, 17, 9–25, 1989.

G. Nuel, « Grandes déviations et chaînes de Markov pour l'étude des mots exceptionnels dans les séquences biologiques », *Thèse de l'Université d'Evry*, 2001.

B. Prum, F. Rodolphe, E. de Turckheim, « Finding words with unexpected frequencies in DNA sequences », *J. R. Statist. Soc. B*, 57, 205–220, 1995.

S. Robin, J.-J. Daudin, « Exact distribution of word occurrences in a random sequence of letters », *J. Appl. Prob.*, 36, 179–193, 1999.

S. Robin, S. Schbath, « Numerical comparison of several approximations of the word count distribution in random sequences », *J. Comp. Biol.*, 8, 349–359, 2001.

S. Schbath, B. Prum, E. de Turckheim, « Exceptional motifs in different Markov chain models for a statistical analysis of DNA sequences », *J. Comp. Biol.*, 2, 417–437, 1995.

S. Schbath, « Compound Poisson approximation of word counts in DNA sequences », *ESAIM : Prob. Stat.*, 1, 1–16, 1995.

Article proposé par : Sophie Schbath, tel : 01 34 65 28 90, mail `Sophie.Schbath@jouy.inra.fr`
Ont également participé à ce travail V. Brunaud, M. El Karoui, B. Prum, S. Robin, F. Rodolphe, E. de Turckheim.