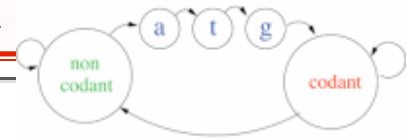


# Des mathématiques dans nos cellules ? Partie 2

To print higher-resolution math symbols, click the [PDF icon](#) in the printing area, or the [list icon](#) in the control panel.



Le 28 octobre 2009, par **Bernard Prum**

Professeur à l'Université d'Evry. Directeur du Laboratoire CNRS "Statistique et génome" ([page web](#))

*Nous proposons ici deux articles décrivant quelques aspects de l'emploi des mathématiques en génétique et génomique — évoquant donc des nouveaux problèmes que ces disciplines posent aux mathématiciens, appliqués comme fondamentaux.*

**A** PRES un survol succinct de la génétique mathématique que l'on peut faire sans avoir accès aux gènes (voir [premier article](#)), ce second article présente l'un des problèmes qui se posent en génomique : la recherche par des algorithmes *ad hoc* des gènes (ou autres objets biologiques) [1].

D'innombrables autres problèmes sont soumis à la sagacité du mathématicien, telles l'analyse des liens entre les maladies et les centaines de millions de variants génomiques *observés* ou la reconstruction des réseaux d'interactions au sein de la cellule. Nous n'en parlerons pas ici.

---

## Où sont les gènes ?

---

Nous avons laissé Gregor Mendel concluant que les caractères phénotypiques étaient hérités au moyen de *gènes*, dont chacun peut prendre plusieurs formes, les *allèles* ; chaque individu porte deux allèles (égaux ou différents), l'un hérité de son père, l'autre de sa mère.

Avant même que la conception de Mendel ressurgisse (de Vries, 1900), les cytologistes comme August Weismann (1834-1914) avaient découvert la structure de la cellule et de son noyau. En particulier ils découvraient les *chromosomes* [2] et surtout la *meïose* qui prend la moitié des chromosomes d'une cellule pour la mettre dans les cellules *germinales*, à savoir ovules ou spermatozoïdes. Ils étaient donc le support rêvé des gènes conçus par Mendel.

Sans entrer dans les détails de l'histoire, citons Thomas Morgan qui a montré (1910) que les chromosomes étaient bien le support des gènes, l'expérience de Frederick Griffith (1928) qui a transféré des caractères à des bactéries en transférant des chromosomes, Oswald Avery (1944) et James Watson et Francis Crick (1953) qui ont établi le rôle de l'ADN.

La question est alors : « Où se trouvent exactement les gènes dans la longue séquence (la suite de **t, c, a, g**) qui constitue les chromosomes ? »

Répondre à cette question, c'est être capable, nous le verrons, de connaître la formule exacte des protéines et surtout de « regarder » sur chaque individu s'il y a des modifications des gènes. On peut ainsi établir un lien entre ces modifications et le phénotype de l'individu, par exemple retrouver la cause exacte de telle ou telle maladie génétique.

Un mot de vocabulaire : on appelle *génom*e l'ensemble des chromosomes d'un individu ; la

*génomique* sera donc l'étude des génomes, depuis leur contenu, leur structure ou leur disposition spatiale, jusqu'à la comparaison des génomes d'organismes différents.

### Signaux locaux

Dans le génome, il y a ... des gènes, mais pas seulement. Un certain nombre des motifs jouent divers rôles biologiques. Le plus souvent il s'agit de *mots* — nous appellerons ainsi des suites de, disons, 5 à 20 lettres consécutives.

Pour nous, qui sommes habitués à la lecture, il n'est pas surprenant qu'une très grande partie de l'information contenue dans les génomes soit portée par de tels mots. La lecture d'un texte est grandement simplifiée par le fait que les six lettres du mot **maison** sont écrites les unes à côté des autres, la lecture par les mécanismes cellulaires est plus aisée si un signal, tel « ici se trouve une position où le mécanisme qui va lire les gènes doit se fixer pour commencer son travail », est un mot.

Rappelons qu'un texte (écrit en français, par exemple) est une intrication de mots porteurs de sens et de dispositifs les articulant. Pour savoir de quoi parle un texte, le plus simple est de chercher quels sont les mots qui y sont fréquents. Si le texte que vous avez sous les yeux (sa version électronique) est envoyé par radio dans l'espace et reçu par les savants de Véga de la Lyre, les ordinateurs de ceux-ci auront tôt fait de s'apercevoir que l'on y rencontre très fréquemment le mot **gén**, par exemple dans des sur-mots tels que **génom**, **génique**, **génomique**, voire **général**. Avant même d'apprendre assez de français pour lire cet article, les Végaiens (?) pourront suspecter qu'il traite de **gén**-quelque chose.

Mais la compréhension d'un texte — surtout mathématique — passe essentiellement par l'articulation des mots ; ne rentrons pas (il le faudrait) dans la grammaire *sujet-verbe-complément*, contentons nous des « petits mots » sans lesquels on ne comprend rien [3] : « **Si** il fait beau, **alors** Pierre **et** Paul sortent, **mais pas** Jacques ».

Il en est de même en génomique. Certains mots sont très utiles au langage du génome, on les trouve fréquemment. À titre d'exemples citons :

- les START placés au début de chaque gène, les STOP placés en fin de gène (voir la **section 2**) ;
- les « sites de fixation », pistes d'atterrissage des mécanismes de lecture des gènes, évoquées il y a un instant ;
- le *Chi* de *E. coli* et autres bactéries, qui est décrit en détail **ici**.

Le statisticien « retournera » cette implication « utile  $\Rightarrow$  très fréquent » et cherchera les mots significativement trop fréquents, pour interpeller les biologistes en leur demandant pourquoi.

Notons que, contrairement aux textes en français (encore que...), certains mots sont évités par un génome. Par exemple *E. coli*, toujours lui, a, à côté du dispositif des *Chi* un autre moyen de se protéger des attaques virales : un *restrieteur* (comme l'aurait appelé Luc Besson) cherche sur les séquences d'ADN le mot **ctag** et, quand il le trouve, il coupe le brin. Nombre de virus n'ont pas survécu à ce traitement. Pour ne pas risquer de « se tirer une balle dans le pied » *E. coli* évite soigneusement d'utiliser ce mot **ctag** dans son propre génome. Un statisticien — fut-il de Véga — saura s'apercevoir de ce déficit et alerter son collègue biologiste, qui pourra rechercher *pourquoi* ce mot manque et, peut-être, découvrir le rôle des restricteurs.

La première idée est naturellement : « comptons les mots d'une longueur fixée, bien sûr, et décrétons « exceptionnellement fréquents » ceux que l'on rencontre le plus, « exceptionnellement rares » ceux que l'on rencontre le moins — et allons les dénoncer pour étude ultérieure à nos collègues biologistes ».

Elle a conduit (voir le cas du *Chi*) à des impasses. Tâchons de comprendre pourquoi.

### Le modèle markovien

Pour manipuler des nombres plus petits que les millions de lettres d'un génome de bactérie, considérons une souche de virus du Sida, HIV1. Elle compte 9718 lettres qui se répartissent en :

$$2164 \text{ t} \quad 1773 \text{ c} \quad 3411 \text{ a} \quad 2370 \text{ g}$$

Il y a (presque) deux fois plus de **a** que de **c**. Si l'on fixe la longueur des mots (8 par exemple) et que l'on cherche les mots très fréquents on trouvera *plutôt* des mots riches en **a** et pauvres en **c** [4]. Sous un modèle où l'on tirerait à chaque position une lettre selon sa fréquence d'apparition, le mot **a[8]**, composé de huit **a** successifs, a une probabilité d'apparaître  $2^8$  fois plus grande que le mot **c[8]** composé de huit **c** successifs [5].

Il faut donc *corriger* les nombres attendus de chaque mot pour tenir compte de la fréquence des lettres. Mais les ennuis ne s'arrêtent pas là. La matrice  $N$  suivante donne les décomptes, toujours chez HIV1, des mots de deux lettres : il y a 548 fois **tt**, 342 fois **tc**, etc.

$$\begin{array}{c}
 \mathbf{t} \\
 \mathbf{c} \\
 \mathbf{a} \\
 \mathbf{g}
 \end{array}
 N = \begin{array}{c}
 \mathbf{t} \quad \mathbf{c} \quad \mathbf{a} \quad \mathbf{g} \\
 \left( \begin{array}{cccc}
 548 & 342 & 684 & 590 \\
 470 & 413 & 795 & 95 \\
 713 & 561 & 1112 & 1024 \\
 432 & 457 & 820 & 661
 \end{array} \right)
 \end{array}
 N(u+) = \begin{array}{c}
 \left( \begin{array}{c}
 2164 \\
 1773 \\
 3410 \\
 2370
 \end{array} \right)
 \end{array}$$

On constate une grande disparité dans les 16 nombres affichés, que les fréquences différentes des lettres n'explique pas [6].

Il convient donc d'évaluer si un mot est exceptionnel *conditionnellement* aux 16 décomptes reportés dans la matrice  $N$  ci-dessus [7]. Il se trouve que ces décomptes constituent *la statistique suffisante du modèle de Markov*. Derrière cette expression effrayante se cachent deux concepts assez simples :

#### Le modèle

Andrei Andreyevitch Markov (1956-1922) est censé [8] avoir inventé le modèle suivant pour décrire la loi d'une suite d'observations  $X_k$  prenant leurs valeurs dans un ensemble fini, nous dirons un *alphabet*. Pour introduire notre modèle, quittons un instant l'alphabet à quatre lettres (**t**, **c**, **a**, **g**) du génome pour un exemple plus simple.

Considérons l'ampoule qui éclaire l'entrée de votre immeuble. On regarde chaque semaine – disons, par exemple le lundi à minuit – si elle fonctionne (notée *Fct*) ou si elle est en panne (notée *Pan*). Si une semaine donnée  $k$  elle fonctionne, admettons qu'elle a 90 % de chance de

fonctionner encore la semaine  $k + 1$ , et donc 10 % de tomber en panne (piètre qualité !!). Si, par contre elle est en panne la semaine  $k$ , il y a 70 % de chance que votre gardienne l'ai réparée avant la semaine suivante (et donc 30 % de chance qu'elle soit encore en panne) [9].

Ceci peut se représenter dans un matrice, dite *de transition* [10] :

$$\begin{array}{c} \\ \\ \end{array} \begin{array}{cc} & \begin{array}{c} Fct \quad Pan \end{array} \\ \begin{array}{c} Fct \\ Pan \end{array} & \pi = \begin{pmatrix} 0.90 & 0.10 \\ 0.70 & 0.30 \end{pmatrix} \end{array}$$

et l'on conçoit (et démontre) aisément que ceci définit [11] la loi de la séquence  $(X_0 X_1 X_2 \dots X_n)$  décrivant l'état de votre ampoule aux semaines numérotées 0 à  $n$ . Voici, ci dessous une séquence (il y a 101 points) tirée selon cette loi, où l'on a représenté *Fct* par un point vert et *Pan* par un point rouge :



On voit, par exemple qu'une panne (rouge) dure rarement plus d'une semaine, parfois (ici deux fois) deux semaines. Les neuf périodes de fonctionnement cumulent 91 semaines (et la dernière à droite peut encore durer un moment !). La matrice  $N$  des comptages, analogue à celle vue sur HIV, s'écrit :

$$\begin{array}{c} \\ \\ \end{array} \begin{array}{cc} & \begin{array}{c} Fct \quad Pan \end{array} \\ \begin{array}{c} Fct \\ Pan \end{array} & N = \begin{pmatrix} 82 & 8 \\ 8 & 2 \end{pmatrix} \quad N(u+) = \begin{pmatrix} 90 \\ 10 \end{pmatrix} \end{array}$$

Si l'on n'observe *que* cette séquence, on peut essayer de deviner la matrice  $\pi$  qui a servi à la construire :

1 - Sur les 90 transitions partant de vert, 8 amènent à une panne. On estime naturellement  $\pi(Fct, Pan)$  par  $\frac{8}{90} = 0,089$  alors que nous savons que la vraie valeur est  $\pi(Fct, Pan) = 0,10$ . Nous mettrons désormais, comme le font les statisticiens, un chapeau sur les valeurs *estimées* et noterons  $\hat{\pi}(Fct, Pan) = 0,089$ .

2 - Sur les 10 transitions partant de rouge, 8 conduisent au vert ; donc  $\pi(Pan, Fct)$  est *estimé* par  $\hat{\pi}(Pan, Fct) = 0,80$  (alors qu'il vaut 0,70).

On constate que les estimations ne sont pas très précises (nous avons observé peu de transitions !), mais néanmoins donnent déjà l'ordre de grandeur des vraies valeurs [12].

Avec une simple calculatrice, le lecteur vérifiera que la matrice de transition associée aux comptages du virus HIV donnés en début de paragraphe vaut :

$$\begin{array}{c}
 \mathbf{t} \\
 \mathbf{c} \\
 \mathbf{a} \\
 \mathbf{g}
 \end{array}
 N = \begin{pmatrix}
 0.253 & 0.158 & 0.316 & 0.273 \\
 0.265 & 0.233 & 0.448 & 0,054 \\
 0.209 & 0.165 & 0.326 & 0.300 \\
 0.182 & 0.193 & 0.346 & 0.279
 \end{pmatrix}$$

Le miracle (? ?) est que pour estimer les paramètres (la matrice  $\pi$ ) d'une chaîne de Markov, il suffit d'avoir les comptages des mots de 2 lettres (et d'ailleurs il faut les avoir). Plus précisément, on estimera  $\pi(u, v)$  par  $\hat{\pi}(u, v) = \frac{N(uv)}{N(u+)}$ , où  $N(u+) = \sum_v N(uv)$ .

On dit que ces décomptes sont une *statistique suffisante* — on dit aussi exhaustive — pour le modèle de Markov.

### Le bon sens

On a donc compris que :

- 1 - on ne pouvait rien dire de pertinent sur le texte du génome sans prendre en compte sa richesse en chacune des quatre lettres, mais aussi sa richesse en chacun des 16 mots de deux lettres que l'on peut composer ;
- 2 - que tenir compte de ces comptages, c'est *faire comme si* le génome était produit par la chaîne de Markov.

On se trouve dans la même situation que Monsieur Jourdain, « faire du modèle markovien sans le savoir, ou faire du modèle markovien en le sachant ». Autant se souvenir que depuis Markov, les probabilistes ont écrit des centaines d'ouvrages sur ce modèle, qui permettent — quitte à fabriquer les outils conceptuels ou informatiques manquants — de répondre aux questions :

- 1 - Etant donnée une séquence de longueur  $n$ , générée par le modèle markovien de matrice de transition  $\pi$  et un mot  $W$ , que dire du nombre (aléatoire) de fois,  $N(W)$ , où  $W$  sera observé ? Peut-on calculer son espérance  $\mathbb{E}(N(W))$  ?, sa variance  $\mathbb{V}(N(W))$  ?, sa loi ? la probabilité que l'on avait d'observer autant (ou davantage) de  $W$  que ce qu'il y a sur le génome réellement séquencé ?
- 2 - Et en quoi ces résultats sont-ils altérés parce que l'on ne connaît pas les  $\pi(u, v)$  et que l'on doit les estimer, comme il vient d'être dit, sur la même séquence ?

Plutôt que de nous attarder sur ce problème — une présentation succincte des résultats est présentée **ici** —, revenons à notre fil rouge « comment trouver les gènes ? ». Retenons une seule chose : affirmer quoi que ce soit sur le génome (mots fréquents/rare, périodicité, etc.) ne peut se faire qu'en tenant compte de sa composition en mots de, disons, 2 lettres, c'est à dire en utilisant une modélisation par chaîne de Markov.

---

### L'annotation

---

Le « dogme central » de la biologie affirme que l'information transmise de génération en génération est écrite le long des chromosomes. Une part essentielle de cette information consiste en « la formule » des protéines, qui sont les molécules sur lesquelles se fondent tous les processus du vivant, depuis la reproduction (des individus comme des cellules), jusqu'à l'apoptose (la « mort programmée »), en passant par la nutrition ou la respiration. Nous appellerons ici *gènes* [13] les segments de chromosomes portant cette formule.

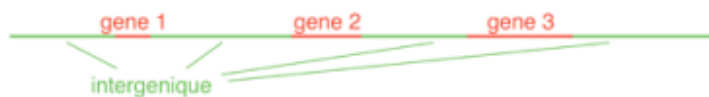
Une protéine est, comme l'ADN, un assemblage linéaire et orienté de molécules de base, les *acides aminés*. Il existe des milliers d'acides aminés différents, mais « la vie » n'en utilise que 20, repérés eux aussi par une lettre [14]. Pour coder un acide aminé, un gène (qui est donc écrit dans l'alphabet à quatre lettres, (**t, c, a, g**) de l'ADN), a besoin (d'au moins) 3 lettres ( $4^2 = 16 < 20$ ). De fait, il convient de regrouper ces lettres 3 par 3 (on obtient des *codons*) et un *code génétique* [15] associe un acide aminé à chaque triplet.

Cette association n'est donc pas injective — on dit qu'il y a *dégénérescence* du code génétique. Voici un exemple très factice de gène (10 acides aminés !!), sur lequel on peut observer cette dégénérescence.

a t g c a a t t g g c g t c g g c t a g t c c c a t a t a t g t g g t g a  
 start Q L A S A S P H M W stop

On y voit aussi qu'un gène commence par un codon START et se termine par un codon STOP. Le codon START est toujours [16] codé **atg** (qui sert aussi à coder la méthionine, voir l'avant dernier codon ci-dessus !) alors que le codon STOP peut être **taa**, **tag** ou **tga**.

Une première description d'un génome est donc : un texte dans lequel alternent des gènes et ce qui sépare des gènes — disons simplement « de l'intergénique ».



Sans être grand biologiste, on soupçonne que la succession des lettres dans les gènes ne suit pas les mêmes règles que dans l'intergénique [17]. C'est naturellement le premier élément que nous allons utiliser pour localiser les gènes.

### *Sprechen Sie Deutsch ?*

Peut-être serons nous plus clairs en parlant de langues que vous connaissez mieux que le génique et l'intergénique.

Supposons disposer sur notre ordinateur d'un texte écrit en français (les *Rougon Macquart* de Zola) et que l'on y « colle » des passages plus ou moins longs d'un texte écrit en allemand ( *Die Leiden des jungen Werthers* de Goethe) :



Nous pouvons admettre, il me semble, que la succession des lettres ne suit pas les mêmes règles en français et en allemand [18] : la lettre la moins utilisée en français, le **w**, est 38 fois plus utilisée en allemand (on passe de 0.04 % à 1.52 %) ; le **k**, 30 fois plus (0.05 % et 1.46 %) et même le **h** passe de moins de 1 % à près de 5 % etc. À l'inverse le français utilise 50 fois plus de **q**, 15 fois plus de **x** et 4 fois plus de **p**. On pourrait en dire autant des mots de deux lettres, par exemple 60 fois plus de **qu** pour les Français, 30 fois plus de **nz** pour les Allemands, etc.

Supposons d'abord connaître ces spécificités des deux langues, résumées dans une matrice de transition  $\pi_{al}$  pour l'allemand et  $\pi_{fr}$  pour le français.

**Premier problème.** Étant donné un texte  $x_m x_{m+1} \dots x_n$  (disons de la millième à la deux-millième lettre de notre fichier), est-il écrit en français ou en allemand ?

La probabilité que l'allemand l'ait produit (on dira « la vraisemblance que ce soit de l'allemand ») est  $L(A) = \prod_i \pi_{al}(x_{i-1}, x_i)$  tandis que la même quantité pour le français vaut  $L(F) = \prod_i \pi_{fr}(x_{i-1}, x_i)$ .

Selon un principe énoncé dans le premier des deux articles présents (Note 17), on sera d'autant plus poussé à pencher pour l'allemand que le quotient  $\frac{L(A)}{L(F)}$  est grand.

**Deuxième problème.** L'ennui, c'est que l'on ne sait pas où commencent les morceaux en allemand et où ils finissent. On ne va pas clamer au premier **nz** rencontré : « C'est de l'allemand ! Peut-être que juste avant et juste après, c'est du français, mais ces deux lettres, c'est clairement de l'allemand ! ».

Et rien ne dit que le segment que l'on vient de choisir,  $x_{1000} \dots x_{2000}$  soit entièrement écrit dans une même langue !

Il faut donc modéliser la longueur des segments d'allemand collés dans le texte de Zola, ou ce qui revient au même — nous sommes maintenant bien familiers de cette idée ! — quelle probabilité a-t-on, juste après une lettre venant du français de trouver une lettre venant de l'allemand (et réciproquement).

### Une chaîne de Markov cachée

Décidons de modéliser la suite des *états* (i.e. « allemand » ou « français ») de chaque lettre par ... une chaîne de Markov, autrement dit d'introduire une matrice, notons la  $\pi_0$ , qui gère les transitions entre textes allemands et français, par exemple [19] :

$$\begin{array}{cc} & \begin{array}{cc} Al & Fr \end{array} \\ \begin{array}{c} Al \\ Fr \end{array} & \pi_0 = \begin{pmatrix} 0.9990 & 0.0010 \\ 0.0001 & 0.9999 \end{pmatrix} \end{array}$$

Notant  $s_i$  la langue (inconnue) de la lettre numéro  $i$ , la suite des  $s_i$  forme une chaîne de Markov  $S$  non observée — c'est elle qui donne son nom à cette modélisation « Chaîne de Markov cachée » (en anglais « Hidden Markov Model », souvent réduit à « HMM »).

Les vraisemblances  $L(A)$  et  $L(F)$  introduites dans le « Premier problème » se regroupent en :

$$L(S) = \prod_i (\pi_0(s_{i-1}, s_i) \pi_{s_i}(x_{i-1}, x_i))$$

qu'il ne reste plus qu'à maximiser [20].

**Troisième problème.** Mais, si l'on trouve sur le web les matrices de transition des langues française et allemande (au moins les comptages qui permettent instantanément de les calculer), on ne trouve pas celles du codant et du non-codant, surtout si l'on s'intéresse à la recherche de gènes sur de nouvelles espèces — même chez l'Homme, on ne connaît pas ces transitions pour chaque morceau de chromosome sur lequel on peut se pencher.

En d'autres termes, on nous demande de séparer codant et non-codant, sans savoir à quoi ressemble du codant et à quoi ressemble du non-codant. Autant demander à Charlemagne de séparer les bons élèves des mauvais sans lui fournir les carnets de notes.

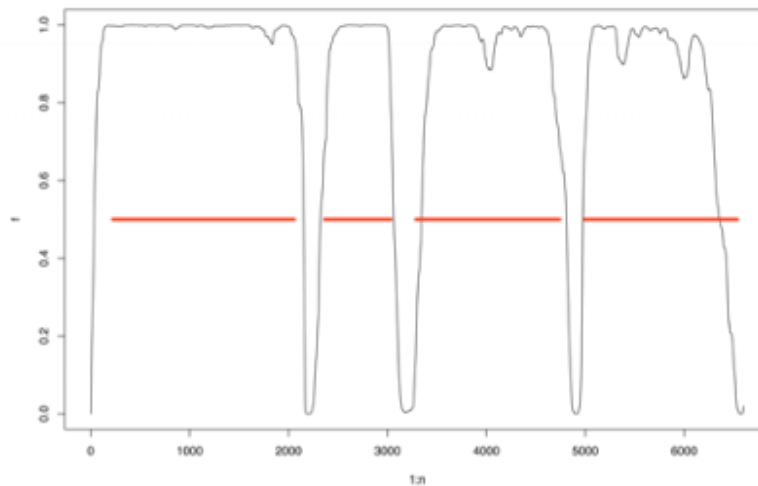
Il faut donc estimer ces transitions. Mais l'on a le même problème : si l'on savait ce qui est codant, on pourrait y compter combien de fois apparaît chaque mot de deux lettres et avec une simple calculatrice (comme *vous* l'avez fait tout à l'heure pour HIV) estimer la matrice de transition.

L'idée qui nous sauve vient de Dempster [21] qui triait deux populations selon un caractère de lois gaussiennes,  $\mathcal{N}(\mu_1, \sigma_1^2)$  pour l'une,  $\mathcal{N}(\mu_2, \sigma_2^2)$  pour l'autre — les paramètres  $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2$  étant inconnus —, après qu'on lui eut *caché* à quelle population appartenait chaque observation, et qu'il a fallu adapter à notre cas de chaîne de Markov cachée :

- On suppose les transitions connues (n'importe quoi si l'on ne sait rien — mieux si possible : celles déjà estimées sur des espèces voisines, par exemple). On sait calculer et même maximiser la vraisemblance  $L(S)$  ;
- son maximum est atteint pour des valeurs  $s_i$ , donc pour une segmentation codant/non-codant ;
- on peut utiliser cette segmentation pour estimer la transition côté codant sur les plages « trouvées » codantes et pour estimer la transition côté non-codant sur les plages « trouvées » non-codantes ;
- on utilise ces transitions estimées pour re-segmenter le chromosome ;
- on utilise les nouveaux segments estimés pour ré-estimer les transitions.
- et on recommence...

Si l'idée est simple (à vrai dire Dempster s'était heurté à un mur d'incompréhension totale : « ça ne marchera jamais ! »), reste à *démontrer* que, sous des hypothèses convenables, les suites de ces estimations convergent vers « ce qu'il faut ». C'est le travail qui a incombé aux mathématiciens [22].

La Figure 1. donne un exemple de résultat. On est, de fait, un petit peu plus subtil que ce qui a été décrit ci-dessus et l'on gère en chaque position  $i$  la probabilité  $p_i$  d'être dans du codant (et donc  $1 - p_i$  d'être dans du non codant). C'est cette quantité  $p_i$  qui est tracée en fonction de  $i$  :



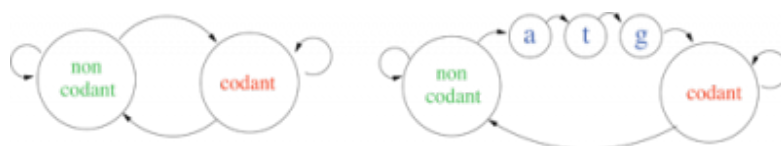
**Figure 1** Sur un morceau de génome de Levure (\*) d'environ 6 600 lettres, un modèle de chaîne de Markov cachée calcule la probabilité pour chaque position d'être dans un gène. Cette probabilité est représentée par la courbe. Comme on connaît ce génome, on sait par ailleurs où sont les gènes : c'est ce qu'indiquent les quatre segments rouges.

(\*) Source : Vincent Miele — Biométrie et biologie évolutive — Lyon

### Utiliser des connaissances biologiques

Cette méthode donne ici un assez bon résultat, mais il est très insatisfaisant aux changements de régime : il faut jusqu'à 200 positions pour passer de « presque sûr c'est codant » à « presque sûr ce n'est pas codant ».

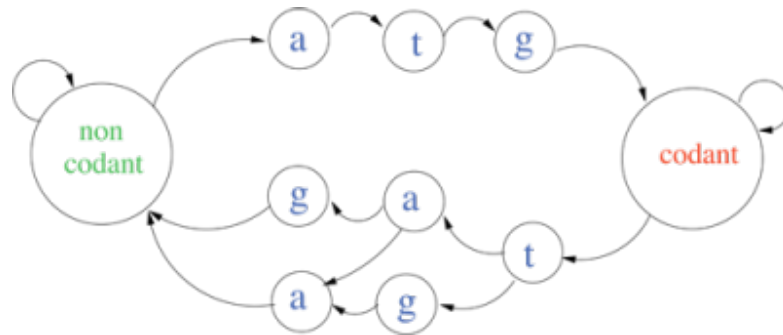
Il est temps de se souvenir du début du paragraphe, où l'on disait qu'un gène commence toujours pas un START codé **atg**. L'une des grandes forces des modèles de Markov cachés est qu'ils sont capables d'intégrer une telle information : il suffit de remplacer le modèle alternant codant/non codant schématisé à gauche de la figure 2. par celui schématisé à droite :



**Figure 2.** Le modèle à deux états est enrichi d'états « dégénérés » : le premier capable seulement de produire un **a**, le deuxième un **t** et le troisième un **g** ; qui plus est, on ne reste dans chacun de ces états que pour *une* position, ensuite on transite avec probabilité 1 vers le suivant.

Ce modèle impose, pour passer de « non codant » à « codant » de traverser les états produisant **a** puis **t** puis **g**. Chaque gène commencera par un **atg** !

Il est à peine plus difficile d'imposer qu'il finisse par un STOP, comme le montre un coup d'œil à la figure 3. :



**Figure 3.** Le modèle est encore enrichi par une ensemble d'états dégénérés qu'il est impossible de traverser sans produire un codon *stop*, **tag**, **taa** ou **tga**.

### *Tout ça pour finir en papillon*

Restent (ici !) deux améliorations à évoquer :

**Modèles périodiques** On a vu que le codant était écrit par groupes de trois lettres les codons. On peut à nouveau soupçonner que les premières lettres des codons ne sont pas générées par les mêmes transitions que leurs deuxièmes ou troisièmes lettres — et tester (avec succès) cette idée. Pour le codant on utilisera donc trois matrices différentes, utilisées périodiquement,  $\pi_1$  pour fabriquer le début,  $\pi_2$  pour le milieu et  $\pi_3$  pour la fin des codons.

**Il y a deux brins d'ADN**, chacun portant des gènes ! Les biologistes nous apprennent que ces deux brins sont lus en sens inverses. La situation est analogue à la suivante : vous marchez dans une rue ; sur la façade à votre gauche vous pouvez trouver écrites les lettres **C A F É** ; sur votre droite vous pouvez rencontrer dans l'ordre les lettres **E I C A M R A H P**.

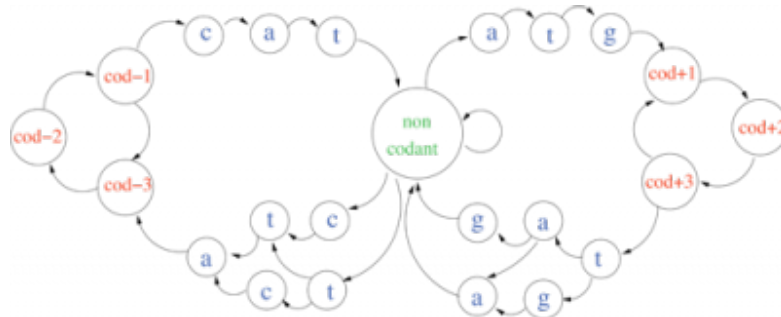
Les biologistes nous disent aussi qu'en face d'un **a**, il y a toujours un **t** et que, en face d'un **c**, il y a toujours un **g**. Il suffit donc de conserver dans les banques de données *un* seul brin d'ADN pour avoir toute l'information, y compris celle portée par le brin opposé. Sur *le* brin lu dans la banque de données, on cherche les gènes qu'il porte et la « photocopie inversée » des gènes situés sur le brin opposé.

Prolongeons notre image de la rue : les enseignes lumineuses se reflètent inversées sur les vitrines d'en face. Face à la pharmacie, à votre droite, vous lirez **E' I C' A M R' A H P'**, où, faute de disposer des caractères nécessaires, on a noté ' le symétrique des lettres. Notons que certaines lettres sont leur propre symétrique (par exemple **A' = A**).

Si une caméra filme la façade de gauche et enregistre **C A F É E' I C' A R' A H P' R' E' H C' U**

**O B'**, peut-on retrouver ce qui était écrit à droite et ce qui était écrit à gauche ? C'est plus facile en français qu'en VOX TUMIHAWY, langue connue pour n'utiliser que les 11 lettres (toutes symétriques !) de son nom... ou en génomique [23].

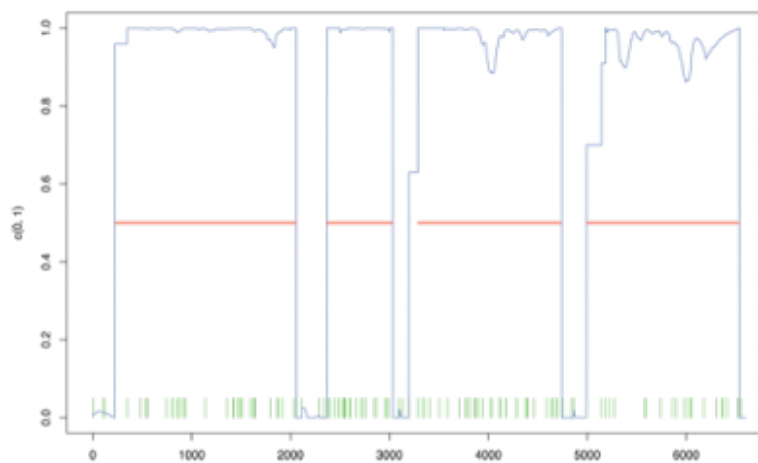
Un peu de réflexion permet néanmoins d'intégrer cette nouvelle connaissance au modèle, qui se complexifie encore et devient le beau papillon de la Figure 4. :



**Figure 4.** Au centre figure le non codant. Les gènes situés sur le brin lu sont décrits par l'aile droite du dessin, qui reprend la Figure 3, à ceci près que l'on a distingué les trois phases du codant.

L'aile gauche correspond aux gènes situés sur l'autre brin d'ADN. Comme les deux brins sont en sens opposés, on y entre par le STOP pour en sortir par le START. On lit les lettres dans l'ordre inverse et en plus chacune a été remplacée par sa jumelle :  $t \leftrightarrow a$  et  $c \leftrightarrow g$ .

Appliqué aux données de la Levure que nous avons commencé à traiter à la Figure 1, ce modèle donne l'annotation :



**Figure 5.** Annotation du brin de levure de la Figure 1 tenant compte de la nécessité des codons START et STOP en début et fin de gène. Les traits rouges indiquent toujours la vraie position des gènes, connue par ailleurs.

Plusieurs codons **atg** peuvent se succéder, même dans le codant (**atg** code pour un acide aminé, **M** — on a indiqué de petites marques vertes les positions de tous les **atg**), et le logiciel ne sait pas trancher.

Pour le premier gène de la Figure, il propose deux START, un auquel il attribue une probabilité de 96 % (ordonnée de la petite marche) et un second, de probabilité 4 %. Pour le quatrième gène, le logiciel propose même trois START.

Par contre, au premier STOP rencontré, le gène s'arrête !

Ces modèles de chaînes de Markov cachées [24] sous-tendent souvent le premier traitement qu'un spécialiste de la génomique applique à une nouvelle séquence, avant de confirmer ou d'infirmer les résultats en se fondant sur des comparaisons avec d'autres espèces ou sur une connaissance biologique *a priori*.

---

## Conclusion

---

Nous avons donc décrit ici une démarche mathématique — nous avons essayé d'en cacher les difficultés techniques — en tâchant de montrer que son cheminement passe par un dialogue constant entre le biologiste et le mathématicien. Le premier apporte le problème et d'innombrables connaissances, le second des outils (les chaînes de Markov, les chaînes cachées, ...) et essaye d'apprendre à son algorithme comment utiliser le savoir du biologiste. Pour garantir au biologiste que la « sortie » des logiciels fondés sur une telle approche a un sens réel, il lui faut *démontrer* les bonnes convergences de ses algorithmes.

Cette interface enrichit donc les mathématiques de nouveaux théorèmes, et la biologie d'outils permettant l'analyse des millions de lettres produites chaque jour par les séquenceurs.

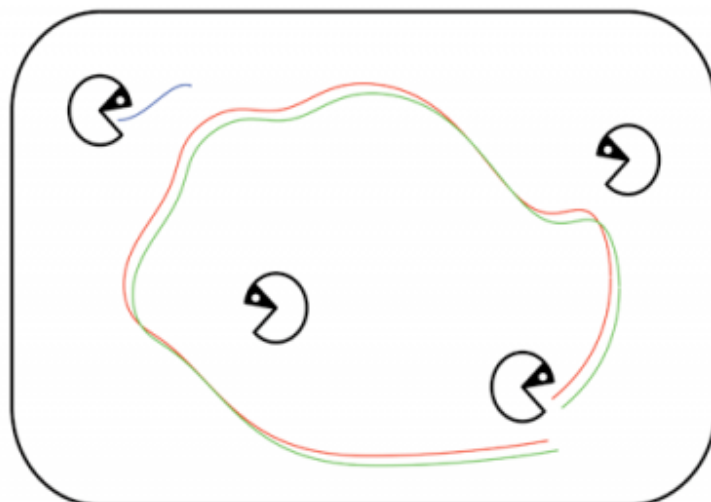
Signalons pour finir une approche alternative à celle de Cowan, fondée sur des automates déterministes — et redoutablement efficace — développée par Grégory Nuel (*J. Applied Prob.* 45 (2008)). On pourra trouver tout ceci (et bien d'autres choses) dans le livre *Analyse statistique des séquences biologiques*, G. Nuel et B. Prum (Hermès 2007).



### Annexe 1 : Le Chi de coli

Décrivons un exemple déjà ancien [25] : *Escherichia coli* est le nom savant [26] actuel de ce que les médecins du début du XX<sup>ième</sup> siècle appelaient le colibacille, voire le bacille colique. Comme son nom l'indique, il vit en particulier dans le colon humain, où il participe sans problème à notre flore intestinale — sauf certaines souches qui sont responsables de gastroentérites et autres désagréments.

La figure schématise une cellule de *E. coli* dans laquelle on remarque tout de suite son génome : un seul chromosome, qui est circulaire. Il est constitué de deux brins d'ADN (représentés ici en vert et en rouge) qui sont d'une façon complexe « photocopie » l'un de l'autre [27].



**Figure 6.** Schéma d'une cellule de *E. coli*. Quelques RecBCD rodent à la recherche de brins d'ADN à détruire.

Nous avons aussi représenté, en bleu en haut à gauche, un virus qui a pénétré la cellule : c'est un brin d'ADN de même nature que le génome cellulaire. Comme pour les Humains, les virus peuvent être extrêmement dangereux pour les bactéries, aussi *E. coli* se défend : un complexe protéique appelé RecBCD (représenté ici comme des gloutons masqués) est présent ; son rôle est de trouver l'extrémité d'un brin d'ADN pour le détruire (il ôte une après l'autre les « lettres » qui le composent).

Le génome de la bactérie est circulaire et n'est donc pas soumis à cette destruction ; mais, nous l'avons dit (Partie 1, note 28), il arrive souvent que ce chromosome se casse et présente des extrémités à la voracité de RecBCD (c'est le cas sur la figure). Si rien ne l'arrête, RecBCD va alors détruire le chromosome de la bactérie, y compris ses gènes et donc tuer la cellule.

Pour se protéger contre ce danger, *E. coli* a disposé [28] le long de son chromosome un certain nombre de fois un « mot de passe », appelé le Chi, à savoir **gctggtgg**. Chaque fois que RecBCD rencontre un Chi, une réaction a lieu qui fait qu'il cesse sa destruction.

Comme l'autre brin du génome n'a (sans doute) pas été atteint, les mécanismes de réparation (dont il a été question dans la même Note 28) peuvent se mettre en oeuvre et reconstituer le brin endommagé [29].

Finalement, notons que pour être efficace, ce dispositif doit placer « beaucoup » de Chi sur le chromosome de la bactérie. S'il n'y en avait — disons — qu'une dizaine, il arriverait que 10 % du génome soit détruit avant que les réparations ne commencent, et ce serait souvent mortel pour la bactérie..

**QUESTIONS :** Existe-t-il des Chi sur d'autres bactéries ? sur toutes les bactéries ? sur les bactéries proches de *E. coli* ? Les Chi sont-ils d'autant plus ressemblants que les bactéries sont voisines ? etc.

Les biologistes savent déterminer « à la paillasse » si un mot donné de 8 lettres [30] joue ou non le rôle de Chi, mais ça prend des mois, et il n'est pas question d'essayer tous les mots de 8 lettres (il y en a  $4^8 = 65536$ ) !

L'idée « pour répondre à cette question cherchons les mots de 8 lettres les plus fréquents sur le génome de la bactérie étudiée » a mené à un échec : les mots ainsi sélectionnés ne montraient pas de rôle de Chi.

Pourquoi ? On l'a déjà dit : parce que l'on n'a pas pris en compte les fréquences des lettres, voire des mots de 2 lettres dans le génome étudié. Et l'on a dit que tenir compte de ces données, c'est se placer dans un modèle de chaîne de Markov. On a donc maintenant un pur problème de mathématiques :

**Problème :** Etant donnée une séquence écrite dans l'alphabet (**t, c, a, g**), ajuster sur cette séquence un modèle de chaîne de Markov et, pour chaque mot **W** (de 8 lettres par exemple), mesurer combien il est trop fréquent par :

$$d(W) = \mathbb{P}[N(W) \geq n_{obs}(W)]$$

où, bien sûr,  $n_{\text{obs}}(W)$  est le nombre de fois où l'on a réellement observé  $W$  dans la séquence [31].

Ce problème se résout — d'ailleurs de façon assez simple, comme il est expliqué à l'annexe 2, ci dessous.

Appliquée au génome de *E. coli*, cette méthode classe, sans surprise, le Chi, **gctggtgg** comme le plus exceptionnel [32], mais, plus innovateur, elle a « craqué » le mot de passe Chi d'autres génomes, comme *Lactococcus lactis*, celui de votre fromage blanc du matin.

## Annexe 2 : Une formule

Le calcul de  $\mathbb{E}(N(W))$  a été très élégamment mené à bien par Cowan [33]. À un facteur près qui intervient très peu dans le résultat et que nous oublierons ici [34] :

$$\mathbb{E}(N(W)) = \frac{K(X)}{K(X_w)}$$

$X$  représente la séquence observée, tandis que  $K(X)$  est le produit des quatre termes ( $u$  parcourt  $(t,c,a,g)$ ) [35] :

$$K(X) = \prod_u \frac{N(u)!}{N(ut)! N(uc)! N(ua)! N(ug)!}$$

Les quantités  $N(\ )$  désignent bien sûr les comptages dans la séquence  $X$ . Quant à  $X_w$ , c'est la séquence obtenue en remplaçant une occurrence de  $W$  par le mot de deux lettres  $w = w_1 w_h$ , la première et la dernière lettre de  $W$ . Nous appellerons cette opération la transformation de Cowan.

Les factorielles de la formule  $K(X)$  ci dessus portent sur des nombres gigantesques ! Pour les comptages de HIV donnés comme exemple de modèle markovien (voir la **matrice**), on a :

$$K(X) = \frac{2164!}{548! 342! 684! 590!} \times \frac{1773!}{470! 413! 795! 95!} \times \frac{3410!}{713! 561! 1112! 1024!} \times \frac{2370!}{432! 457! 820! 661!}$$

Une vieille formule, due à Stirling, évalue  $K(X)$  à  $10^5$  à la puissance 5627, hors de portée de tout ordinateur raisonnable.

Le miracle résulte du fait que la transformation de Cowan ne modifie les comptages que de quelques unités (de moins qu'il n'y a de lettres dans  $W$ ) ce qui produit un miracle : dans le quotient  $\frac{K(X)}{K(X_w)}$  les termes des factorielles vont presque tous se simplifier, et le calcul va devenir immédiat. Voyons ceci sur un exemple.

Toujours sur HIV1, intéressons nous à  $W = \mathbf{agta}$  [36] qui est présent 96 fois. La transformation de Cowan remplace  $W$  par  $w = \mathbf{aa}$  ; les comptages faisant intervenir  $c$  seront inchangés, il y aura un  $g$  et un  $t$  en moins, un  $ag$ , un  $gt$  et un  $ta$  en moins ; mais aussi un  $aa$  en plus ; on aura donc, par exemple :

$$\frac{N(g)!}{N^-(g)!} = \frac{2370!}{2369!} = 2370 \quad \frac{N(gt)!}{N^-(gt)!} = \frac{432!}{431!} = 432$$

(où  $N^-$  correspond aux décomptes dans la séquence après transformation de Cowan). Et finalement :

$$\mathbb{E}(N(W)) = \frac{1024 \times 432 \times 684}{2370 \times 2164} = 40.35$$

Un calcul similaire donne pour variance  $\mathbb{V}(N(W)) = 33.31$ , soit un écart-type de 5.77. Le véritable comptage, 96, est donc à  $(96 - 40.35)/5.77 = 9.64$  écarts-types du comptage attendu, ce qui — comme on l'enseigne classiquement — est un très gros écart.

Le mot  $W = \mathbf{agta}$  est trop présent chez HIV1. Il ne nous reste qu'à appeler un généticien de HIV pour lui demander s'il sait pourquoi ? Et s'il ne sait pas pour exciter sa curiosité et l'inciter à enquêter à la paillasse.

## Notes

[▲1] Je remercie Gilles Grasseau, Pierre Latouche et Catherine Matias, qui ont relu cet article et formulé des critiques très constructives

[▲2] ainsi nommés parce qu'il est facile de les colorer par diverses teintures.

[▲3] et que les étudiants adorent omettre dans leurs copies. Il écrivent  $x = 2$  et nul ne sait si c'est une hypothèse qu'ils posent, un exemple ou une conclusion péremptoire.

[▲4] De même en français, les mots fréquents sont pleins de **e**, **a**, **s** et comportent peu de **w** ou **k**.

[▲5] Avec le rapport exact  $3411/1773$  on attend 190 fois plus de **a**[8] que de **c**[8].

[▲6] Un test élémentaire de  $\chi^2$  montre qu'il n'y a pas indépendance entre la première lettre et la seconde lettre d'un mot de 2 lettres. [Pour les spécialistes, un  $\chi^2(9)$  valant 471.47, soit un dds de  $10^{-16}$ ].

[▲7] Il n'y a bien sûr aucune raison pour ne conditionner que par les mots de longueur 2. Prendre en compte les mots de longueur 3 conduirait à utiliser des modèles markoviens de mémoire 2, ... Conditionnellement aux décomptes des mots de longueur  $h$ , on travaillerait en modèle de Markov de mémoire  $h - 1$ . Ceci se fait ... pas ici.

La seule question ardue est « Jusqu'à quelle longueur faut-il prendre les mots quand on cherche les mots exceptionnels de longueur, disons, 8 ? »

[▲8] Une légende veut que, en ces temps troubles et staliniens, « on » lui aurait attribué la découverte d'un collègue, excellent en mathématiques, mais admirant trop peu le régime.

[▲9] Nous ne prenons pas en compte une éventuelle usure qui augmenterait le risque de panne avec l'âge de l'ampoule. Et nous ne considérons pas les cas où elle tomberait en panne et serait réparée — ou l'inverse — dans la même semaine. Il est facile au mathématicien d'intégrer ces événements doubles, voire triples, etc.

[▲10] La ligne correspond à l'état de départ, la colonne à l'état vers lequel on *transite*.

[▲11] pour peu que l'on se soit donné le point de départ  $X_0$ , voire une loi (de Bernoulli) pour choisir ce point de départ.

[▲12] Les deux autres valeurs s'obtiennent clairement par complément à 1 :  $\hat{\pi}(Fct, Fct) = 0.911$  et  $\hat{\pi}(Pan, Pan) = 0.20$

[▲13] Il existe d'autres gènes tout aussi essentiels, les gènes à ARN par exemple — nous n'en parlerons pas ici.

Tous ces gènes et « ce qui est à leur service » (annonceurs, sites de fixation des mécanismes de lecture, etc.) constituent une toute petite partie des génomes — disons 5 % chez l'Homme. Il semble inconcevable que 95 % du génome ne serve à rien — grand mystère aujourd'hui : à quoi cela sert-il ?

[▲14] souvent l'initiale de leur nom. On utilise des majuscules pour ne pas les confondre avec les constituants de l'ADN : **A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y**.

[▲15] établi principalement par Marshall Nirenberg dans les années 60.

[▲16] Il y a un théorème qui s'applique toujours en Biologie, qui dit qu'à tout théorème il y a des exceptions.

[▲17] En entendant cette affirmation, le Statisticien se précipite pour la *tester*. On a vu qu'à partir d'une séquence on pouvait estimer la matrice de transition. Faisons ça séparément sur les gènes (disons de *E. coli*) et sur les régions intergéniques. Un coup d'oeil aux bons livres de Statistique vous permettra d'y trouver comment tester que les deux estimations sont significativement différentes. Et, comme prévu, elles le sont !

[▲18] si vous ne le croyez pas, ... testez ! — ou bien consultez [ce site](#).

[▲19] après chaque lettre allemande, on a une chance sur 1000 de changer de langue ; on en déduit — mais si ! — que les plages en allemand compteront en moyenne 1000 lettres. De même, avec l'exemple donné, les plages en français dureront en moyenne 10 000 lettres.

[▲20] Ce qui est loin d'être facile : les  $s_i$  étant inconnus, il y a  $2^n$  configurations (si chaque  $s_i$  peut prendre 2 valeurs,  $3^n$  configurations s'il y avait 3 langues, etc.). De très beaux algorithmes récursifs ont dû être mis au point — et ils sont très efficaces. (Baum et al. *Ann. Math. Stat.* 41 (1970) ; Viterbi *IEEE Trans. Inform. Th.* 13 (1967)).

[▲21] Dempster et al. *JRSS B* 39 (1977).

[▲22] C. F. Jeff Wu, *Ann. Statist.* 11 (1983) Cette démarche n'est pas très éloignée de celle décrite dans notre premier article, où nous partions d'une solution approchée (voire très mal approchée) pour l'améliorer pas à pas.

[▲23] La dite génomique s'apparenterait davantage à la langue (assez imprononçable) *bdpq*, qui n'utilise que ces quatre lettres, en minuscules, un **b** se reflétant en forme de **d** et un **p** se lisant **q** dans la vitrine opposée.

[▲24] souvent encore complexifiés pour prendre en compte d'autres connaissances biologiques (par exemple plusieurs types de gènes, les hydrophyles, les hydrophobes, ...) — sans parler des gènes des organismes supérieurs qui s'écrivent ... en pointillé, alternant des segments finalement traduits en protéines (les *exons*) et des segments qui sont éliminés (les *introns*).

[▲25] C'est celui par lequel notre laboratoire est entré dans la problématique génomique : Prum B., Rodolphe F., de Turckheim E. (1995) Finding words with unexpected frequencies in DNA sequences, *J. Royal Statistical Society* 57, p. 205-220.

[▲26] en hommage à Theodor Escherich qui l'a découvert en 1885.

[▲27] Outre le rôle protecteur explicité quelques lignes plus loin, ceci sert aussi lors de la reproduction de la cellule. Quand celle-ci se coupe en deux, un brin va dans chaque « cellule-fille » et le brin manquant est copié sur ce brin présent, pour obtenir de nouveau du « double brin ».

[▲28] le processus de sélection, au cours de l'évolution des espèces, a disposé...

[▲29] Exercice : décrire un scénario catastrophe.

Réponse : un virus très destructeur place le Chi de *coli* au début de son génome. Il échappe à RecBCD. Il détruit tous les *E. coli* du monde, et l'humanité ne survit pas.

Conclusion : ne dévoilez pas le mot de passe aux virus qui vous le demandent.

[▲30] comme chez *E. coli* ; en cas d'échec on essaiera d'autres longueurs.

[▲31] c'est ce que le statisticien appelle avec beaucoup de pertinence le *degré de significativité* :  $d(W)$  évalue le risque que l'on a de dire une bêtise en proclamant qu'il y a plus de  $W$  que ce que l'on attendrait dans le modèle.

[▲32] ceci dans le modèle markovien de mémoire 1, mais aussi de mémoire 2. Il est encore le champion pour la mémoire 5 (conditionnellement aux décomptes des mots de 6 lettres) ; pour les mémoires 3, 4 et 6 il reste dans les tout premiers rangs. Voir l'HdR de Sophie Schbath (Evry 2003).

[▲33] *J. Applied Prob.* 28 (1991) ; pour la variance et les approximations gaussiennes voir Prum-Rodolphe-de Turckheim *JRSS B* 57 (1995), pour les approximations de type Poisson, Schbath *ESAIM-PS* 1 (1995).

[▲34] pardon à F. Rodolphe et E. de Turckheim qui ont beaucoup peiné sur ce « facteur ». Bien sûr les logiciels en ligne à notre laboratoire sont moins désinvoltes et calculent minutieusement la correction.

[▲35] Qui a fait un peu de « combinatoire » ne sera pas surpris par cette formule qui dénombre des combinaisons.

[▲36] le Chi, **gctggtgg** est trop long pour notre propos. Dans une séquence de longueur 9718, on attendrait largement moins que 1 occurrence.

#### ► Crédits images

Pour citer cet article : **Bernard Prum**, **Des mathématiques dans nos cellules ? Partie 2. Images des Mathématiques**, CNRS, 2009. En ligne, URL : <http://images.math.cnrs.fr/Des-mathematiques-dans-nos,406.html>