# Mathématiques et biologie

Bernard Prum(*)

Chacun d'entre nous a appris dès ses premières années d'école que le monde physique obéissait à des lois qui s'exprimaient par des formules mathématiques : $F = m\gamma$, $P V = n R T$, $\sin i = n \sin r$, etc. jusqu'à $E = m c^2$. Personne ne s'en étonne plus(1). Mais cette néanmoins surprenante capacité des mathématiques à modéliser le monde réel s'étend-elle à la sphère du vivant ? Le « libre arbitre » de chaque être vivant – qu'il soit (relativement) grand comme chez un Humain ou fort limité pour une bactérie – ne compromet-il pas définitivement la modélisation inflexible par des Mathématiques ?

On n'a pas trop de souci quand il s'agit de modèles déterministes, qui sont souvent l'application au monde vivant de lois physico-chimiques universelles : liens entre pression artérielle et teneur en sel, équations régissant les flux de sang dans les vaisseaux ou d'air dans les poumons ou même vitesse de croissance d'une population bactérienne en fonction de la température et de la concentration en sucre ; tout ceci rentre naturellement dans nos schémas mentaux. Le problème est plus délicat précisément là où les mathématiques sont les plus efficaces pour aider le médecin et le biologiste, c'est à dire lorsqu'elles utilisent des modélisations fondées sur l'aléatoire : études épidémiologiques, essais cliniques et, aujourd'hui, modélisations des mécanismes intracellulaires.

On peut, de façon un peu simpliste, dire que le problème se pose à deux niveaux : celui concernant un organisme et celui concernant une communauté d'organismes ; la frontière n'est nullement précise, comme le montrent la fourmilière ou la colonie bactérienne qui « optimisent une stratégie collective ». Le comportement statistique d'une collectivité ne remet pas nécessairement en cause le déterminisme de chacun de ses membres ; c'est ainsi que l'on modélise de façon probabiliste une élection ou le choix d'une marque d'essence, sans croire que les électeurs ou les consommateurs tirent au sort leurs préférences.

Plus profonde est la question : peut-on modéliser un être vivant en utilisant des modèles probabilistes. Un rat de laboratoire choisit-il « au hasard » son chemin dans un labyrinthe ? Un accident coronarien est-il dû à l'obstruction aléatoire d'un vaisseau ? L'évolution des espèces s'est-elle faite « au hasard » ? et leurs génomes peuvent-ils donc être considérés comme des successions aléatoires de nucléotides(2) ?

(*) La génopole, Évry. mail : prim@genopole.cnrs.fr.

(1) Rappelons la formule célèbre de E.P. Wigner : « la déraisonnable efficacité – the unreasonable effectiveness – des Mathématiques dans les Sciences de la Nature » *Comm. in Pure and Applied Math*. 13 (1960)

(2) Bien évidemment « au hasard » ne signifie pas « selon une loi uniforme » – confusion fréquente dans la presse … et dans les copies d'étudiants. Entre « seul un événement peut se produire » (déterminisme) et « tout peut arriver avec la même probabilité », il y a une immense marge. L'évolution, par exemple, est soumise à de nombreuses contraintes. Laissent-elles une place à une part de hasard ? Telle est la question.

Bref, un être vivant peut-il être modélisé à l'aide de cette branche des Mathématiques que sont les Probabilités ?

## 1 – De la biométrie aux mécanismes biomoléculaires

La connaissance que l'Homme a du vivant a fait depuis une quinzaine d'années un progrès quantitatif inconnu jusqu'alors, mais aussi un progrès qualitatif certain. Il est l'aboutissement – provisoire – d'un processus où « le vivant » est compris de façon sans cesse plus analytique et mécaniste. Jusque vers le milieu du XX$^{e}$ siècle, la description se faisait d'abord en termes d'organes : dans certaines circonstances, l'hypophyse produit telle hormone qui, véhiculée par le sang, provoque telle réaction des glandes surrénales qui libèrent telle enzyme… L'approche au niveau cellulaire concernait par exemple le fonctionnement des cellules nerveuses : une impulsion électrique parcourt l'axone d'un neurone provoquant au niveau de la synapse l'émission d'acétylcholine perçue par les dendrites d'un second neurone…

Tout au long de ce siècle, on est entré davantage dans les mécanismes internes à la cellule : fonctionnement des organelles comme les mitochondries par exemple, description de voies métaboliques mettant en jeu toute une panoplie de protéines. Et, bien sûr, le devant de la scène s'est vu occuper par la biologie moléculaire.

Cette « descente » décompose un fait biologique en processus élémentaires plus faciles à décrire en termes de lois universelles « mathématisables ». On peut avoir une conception extrêmement matérialiste du vivant et caresser le rêve (ou le cauchemar) qu'un jour on modélisera « le vivant » comme aujourd'hui on modélise la trajectoire d'un satellite ou le fonctionnement d'une diode. Mais, même si l'on vise cet objectif, il est aujourd'hui « infiniment lointain » tant il est clair que le fonctionnement d'un organisme, même unicellulaire, intègre tant de composantes qu'il est vain de rêver à une appréhension globale,

Une modélisation déterministe semble peu réaliste – sauf approximation grossière (« la taille d'un colonie bactérienne est multipliée par un coefficient $r$ toutes les heures ») : une modélisation sans aléa présuppose que l'ensemble des facteurs intervenant dans un phénomène peuvent être pris en compte alors que la totalité des phénomènes biologiques étudiés font intervenir d'innombrables cofacteurs. Et il se trouve – une longue expérience l'a montré – que, pour comprendre comme pour prédire avec efficacité, il est très judicieux de regrouper l'ensemble des effets que l'on n'a pas été capable de décrire un à un dans une unique quantité et de traiter celle-ci comme l'une de ces variables aléatoires que décrivent les ouvrages de mathématique pure que sont les Traités de Probabilités – par exemple une gaussienne.

Ceci explique pourquoi les Probabilités et les Statistiques se sont montrées si efficaces dans les applications en sciences de la vie. Ce n'est d'ailleurs en quelques sorte que le remboursement d'une dette due par les probabilistes aux biologistes : l'essor moderne des probabilités s'est fondé sur des interactions avec des biologistes. Ce sont les questions relatives au vivant (remplaçant celles posées par les jeux de hasard) qui ont motivé les travaux des « pères » de la statistique moderne, Sir Ronald Fisher (1890-1962), Student (de son vrai nom W.S. Gosset, 1876-1937), J. Neyman

(1894-1981), K. Pearson (1857-1936) etc. et des probabilités modernes, comme A.N. Kolmogorov (1903-1987).

La question du rôle de l'aléa dans la modélisation d'un processus biologique reste posée. Si, par exemple, on cherche à calculer le risque R d'accident opératoire sur un individu en fonction de covariables, telles que l'âge, le sexe, la pression artérielle (PA), etc., une procédure statistique de choix de modèle pourra nous conduire à conclure que le risque opératoire est multiplié par 2 chaque fois que l'âge augmente de 10 ans, multiplié par 1,1 chaque fois que la PA augmente de 1 et est trois fois plus important chez un homme que chez une femme. On pourra avoir un résultat s'écrivant :

$$R = R_0\, 1{,}072^{\text{âge}}\, 1{,}1^{\text{PA}}\, (3 \times \mathbb{1}_{\{\text{homme}\}} + \mathbb{1}_{\{\text{femme}\}})$$

où $R_0$ est une constante à déterminer. Ce résultat, tout à la fois, décrit une réalité biologique qui « parle au médecin » en lui donnant explicitement les diverses sources du risque et permet dans un cas concret de prendre une décision thérapeutique.

Ce type de modèle aléatoire est souvent utilisé pour rendre compte d'une expérience unique portant sur un individu unique. On touche ici l'une des plus grandes difficultés de la modélisation aléatoire, difficulté d'ailleurs presque toujours occultée : si pour un individu donné on estime ce risque R à, par exemple, 30 %, cela ne veut pas dire *a priori* que si l'on pratique cette opération sur un grand nombre d'individus ayant l'âge, le sexe, la pression artérielle de l'individu considéré, il y aura accident dans 30 % des cas. D'abord parce que cette population de grande taille partageant ces covariables n'existe pas[3] : pour peu qu'une douzaine de covariables soient prises en compte, l'individu à opérer est sans doute à jamais l'unique individu de cette « population ». Cela ne veut bien sûr pas dire non plus que si l'on répète N fois l'opération sur cet individu, il y aura accident dans 30 % des cas. Poussée à l'extrême, l'approche fréquentiste des probabilités révèle son caractère absurde : le tirage de l'aléa dont on parle se fera une unique fois (si l'on opère).

Cette difficulté se retrouve dans tous les domaines (en économie, en météorologie) et bien sûr aussi en génomique.

## 2 – Les trois aléas de la génétique

Notons d'abord que le mot « génétique » est en train de changer complètement de sens. Il a longtemps désigné la science de la transmission des caractères (en grec genos signifie famille ou race). Aujourd'hui la génétique est surtout devenue la science de la détermination des caractères, celle qui relie le génotype de l'individu, c'est à dire l'ensemble des gènes qu'il porte, à son phénotype, que ce soit la couleur des yeux ou la susceptibilité de développer tel cancer.

La génétique est la seule science[4] où la notion d'aléa est présente dans le modèle lui-même ; en général l'aléa est introduit pour modéliser des erreurs de mesure, pour

(3) et si elle existait, le résultat fréquentiste annoncé serait conséquence d'un théorème qui aura mérité une démonstration (la Loi des Grands Nombres). La conclusion d'un théorème ne peut être confondue avec une définition, voire un axiome !

(4) Avec la Physique Quantique. Est-ce un hasard ? Les aléas de la biologie sont-ils, à l'origine, des aléas quantiques… ?

rendre compte d'un échantillonnage (lors d'un sondage électoral, par exemple), « au mieux » pour ajuster un modèle sur des données (c'est ce que l'on faisait dans l'exemple des risques opératoires). Au contraire, les lois dégagées vers 1880 par Mendel – qui ne connaissait ni gènes ni chromosomes – s'expriment ainsi (en termes modernes) :

> Les caractères phénotypiques (= apparents) dépendent de gènes. Chaque gène peut varier, l'ensemble des possibilités pour ce gène s'appelle ses allèles. Dans les espèces sexuées, chaque individu reçoit un allèle de son père et un allèle de sa mère.
>
> Lorsqu'un individu se reproduit, il transmet à chaque descendant l'allèle qu'il a reçu de son père avec probabilité 1/2 et l'allèle qu'il a reçu de sa mère avec probabilité 1/2.
>
> Ces aléas sont indépendants lors de la conception des différents descendants de cet individu.

Notons l'emploi du terme « probabilité » – aveu implicite de l'incapacité où l'on est de prédire quel allèle sera transmis dans le jeu de pile-ou-face de la méiose, et même de la notion « d'indépendance » qui est la caractéristique de la théorie des probabilités au sein de la théorie mathématique de la mesure.

À côté de ce premier aléa, que nous qualifierons de mendélien, deux autres sont pris en compte dans la transmission des caractères. Tout d'abord celui des recombinaisons : chacun de nous porte chacun de ses chromosomes en deux exemplaires, un reçu de son père et un reçu de sa mère. Or lors de la fabrication des gamètes (spermatozoïde ou ovule) deux chromosomes de la même paire peuvent se couper en un ou plusieurs emplacements pour fabriquer une mosaïque (crossing-over, cf. Fig. 1). Chaque bébé reçoit à sa naissance une telle mosaïque des génomes de ses grands-parents coupés et recollés en moyenne en une trentaine de positions.

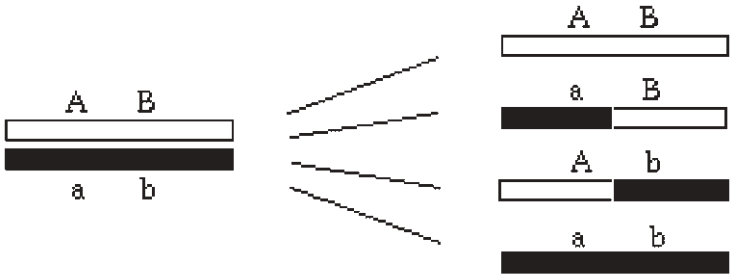


Figure 1 : Un individu porte un chromosome hérité de son père (en noir) et un chromosome hérité de sa mère (en blanc) ; il transmet à chaque descendant une mosaïque de ceux-ci et la recombinaison entre deux positions est d'autant plus probable que celles-ci sont éloignées.

Notons pour être complet un dernier phénomène biologique que l'on modélise par des modèles probabilistes, celui des mutations : un gène peut être « mal copié » lors de sa transmission parent-enfant et cette « erreur de copie » est traitée de façon semblable à toute autre erreur, comme un aléa. C'est d'ailleurs le seul aléa intervenant pour les espèces asexuées, depuis les bactéries jusqu'à certains vers.

Nous n'aurons pas la place ici de développer le traitement mathématique de ces mutations. C'est ce traitement qui permet de reconstruire des « arbres phylogénétiques », c'est à dire décrivant l'évolution des espèces depuis l'apparition de la vie sur Terre. On cherchera par exemple à construire l'arbre qui permet l'observation, aujourd'hui, des séquences chromosomiques d'un certain nombre d'espèces (par exemple des Primates) au prix d'un nombre minimum d'erreurs de copie. Les aspects statistiques, combinatoires et informatiques des algorithmes inventés pour ce faire sont loin d'être totalement explorés aujourd'hui.

## 3 – Recombinaison et localisation de gènes

Pour comprendre réellement pourquoi tel gène intervient dans la détermination de tel caractère ou dans la susceptibilité à telle maladie (par exemple le cancer du sein), il est nécessaire d'en connaître la nature chimique – ce qui est souvent encore difficile aujourd'hui. Mais cette connaissance n'est absolument pas indispensable pour conclure au caractère héréditaire d'un trait : observer que le trait est concentré dans des familles peut suffire (il faut être précautionneux : les individus d'une même famille partagent un même environnement, un même mode alimentaire, qui peut aussi être la cause d'une fréquence élevée du trait dans cette famille).

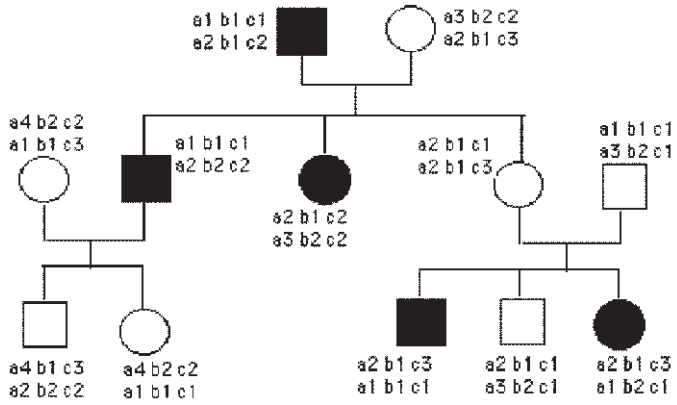


Figure 2 : Un exemple de données. Les cercles représentent des femmes, les carrés des hommes. Sont indiqués en noir les individus malades. On a « typé » les deux grands parents, leurs trois enfants (et leurs conjoints) et les cinq petits enfants en trois locus marqueurs a, b et c. À côté de chaque individu on a indiqué les numéros des allèles qu'il porte ; par exemple au locus a, on a observé quatre allèles différents, numérotés (arbitrairement) a1, a2, a3 et a4, le grand père porte a1 et a2, la grand mère a2 et a3.

Questions : peut-on estimer les probabilités de recombinaison (donc, en un certain sens, les distances) entre les locus marqueurs a, b et c ? Peut-on conclure qu'il y a un gène de la maladie et le situer par rapport aux marqueurs ?

(Réponse : non, bien sûr, on dispose ici de trop peu de données ; il faudrait quelques dizaines de familles comme celle dessinée ici pour pouvoir travailler).

On n'a pas non plus besoin de savoir ce qu'est (chimiquement) un gène pour le localiser, au moins approximativement ! Il suffit de s'appuyer sur la notion, rappelée ci-dessus, de recombinaison. Depuis une vingtaine d'années on connaît des allèles

« facilement » identifiables soit au travers d'un phénotype (le groupe sanguin est un exemple), soit par un séquençage partiel. On appellera ces allèles des « marqueurs ». L'important est que l'on peut « typer » les marqueurs d'un individu à peu de frais.

Il est alors facile de suivre les marqueurs dans des familles – de voir avec quelle fréquence il y a recombinaison entre deux marqueurs, ce qui permet de les localiser les uns par rapport aux autres sur les différents chromosomes : on a alors dressé une « carte génétique ». Il est ensuite possible de chercher quels sont les marqueurs dont la transmission n'est pas indépendante de celle de la maladie et ainsi de localiser un gène impliqué dans la maladie. Une étape ultérieure consistera à séquencer précisément ce gène chez les malades et les individus sains pour savoir quels allèles sont liés au fait d'être atteint et peut-être aboutir à une compréhension « chimique ».

Il est savoureux de constater que tous les modèles décrits ci-dessus se fondent sur un nombre limité de tirages en oui/non (des variables de Bernoulli) : un parent transmet soit l'un soit l'autre de ses allèles ; entre deux marqueurs il y a ou non recombinaison. Néanmoins, même pour des structures familiales assez simples, on arrive à une complexité telle que ni le mathématicien théoricien ni l'ordinateur n'en viennent à bout.

## 4 – Chaînes de Markov et annotation

On sait maintenant que les génomes sont des textes écrits (avec une orientation début-fin) dans un alphabet constitué de quatre molécules que nous désignerons par leurs initiales, {a, c, g, t}. La première chose consiste à « lire » ces textes. Cette opération, appelée le séquençage, produit plus de 20 millions de nouvelles lettres par jour ! Un virus compte quelques milliers de lettres (9817 pour HIV) ; un génome de bactérie quelques millions (4,6 millions pour le colibacille, de son nom savant *Escherichia coli*) ; le génome humain compte environ 3,3 milliards de lettres. On dispose d'une cinquantaine de génomes complets et il en arrive un ou deux nouveaux par mois. Ce rythme ne cesse d'ailleurs de progresser. Leur étude par les généticiens est donc impensable sans un traitement automatique par l'ordinateur – donc sans procédure statistique fondant ces traitements.

L'une des tâches premières à exécuter sur ces séquences est leur « annotation » : chez les organismes dont les cellules ont un noyau (les eucaryotes – en particulier chez tous les organismes pluricellulaires), seul un faible pourcentage du génome est constitué de gènes. Chez l'Homme ce pourcentage n'excède pas 5 %. Le reste, qualifié d'intergénique contient certes quelques signaux (« précurseurs » annonçant un gène, complexes où se fixent les mécanismes de lecture des gènes, etc.) mais le rôle de la quasi totalité de cet intergénique reste obscur – et l'idée dominante est qu'il ne sert à rien. Qui plus est la plupart des gènes ne sont pas « écrits en un seul bloc » mais sont « écrits en pointillés » : une partie du texte, constituée de segments baptisés introns, est détruite avant la traduction en protéines, de sorte que l'information nécessaire pour fabriquer les protéines n'occupe que le reste du texte, les exons (voir figure 3).

Trouver les gènes dans la masse gigantesque de séquences arrivant dans les banques de données ne peut se faire qu'à l'aide d'un outil informatique adapté. La première idée (qui n'est pas si mauvaise) est que les proportions des quatre lettres, a,

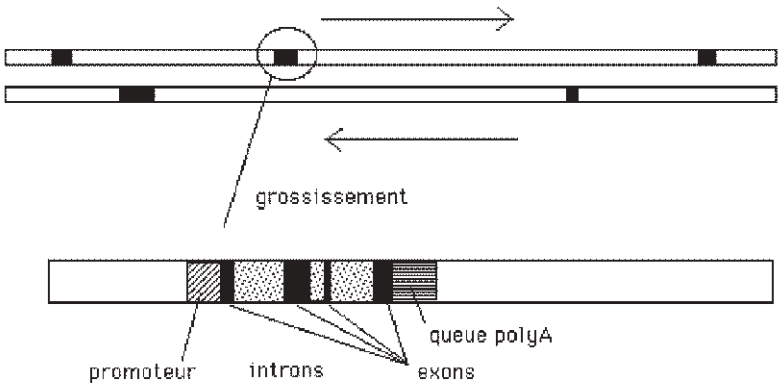


Figure 3 : La grande majorité du texte est de type intergénique (représenté en blanc sur la figure) ; les deux brins de l'ADN sont orientés en sens inverses ; chacun porte des gènes (représentés en noir) ; vu de plus près, un gène est composé d'un promoteur et d'une queue polyA entre lesquels se trouve le texte qui sera copié en ARN. Avant d'être traduit en protéine, l'ARN subit un processus, dit de maturation, au cours duquel certains segments intermédiaires (les introns) disparaissent de sorte que seule l'information portée par les autres (les exons) sert à coder les protéines.

c, g et t n'ont aucune raison d'être les mêmes dans les gènes et dans l'intergénique, ce qui donne un point d'attaque. On va améliorer cette idée en se disant que les fréquences des seize mots de deux lettres, {aa, ac, ag, …, tg, tt} elles aussi doivent différer. C'est ce qui va nous conduire à utiliser des chaînes de Markov.

On comprendra peut-être mieux si l'on revient à un langage qui nous est plus familier, la langue française. Dans cette langue, la lettre « q » est presque toujours suivie d'un « u », un « m » souvent suivi d'un « p » ou d'un « b », etc. C'est ce que l'on pourrait appeler un style. Un simple coup d'œil à une page où ces règles ne sont pas suivies nous convaincra qu'il ne s'agit pas d'un texte en français(5).

**Chaînes de Markov et Chaînes de Markov Cachées**

Dans bien des situations, on est amené à considérer des variables aléatoires mesurées à des « instants » successifs, 1, 2, 3, …, $n$, … Prenons comme exemples le temps qu'il fait au jour $n$, $X_n \in$ {beau, couvert, précipitations} ou le cours $X_n$ du dollar à la bourse de Paris. Comme on ne sait pas prédire $X_n$ avec certitude, on a recours à un modèle probabiliste. Chacun s'attend à ce que, pour prédire $X_n$, ce modèle tienne compte des valeurs $X_m$ pour les jours $m$ précédents ($m < n$) : les réalisations de X aux jours successifs ne sont pas indépendantes.

Le modèle le plus simple pour formaliser cette dépendance consiste à supposer que, pour prédire $X_n$, il n'est pas nécessaire de connaître tout le passé mais seulement

(5) Bien sûr cette remarque est encore plus vraie si l'on regarde les fréquences des mots de 3, 4 lettres ou davantage. Les styles de chaque langue se différencient de mieux en mieux quand on considère des mots plus longs. La réalité biologique demande de considérer des mots de 5 ou 6 lettres, par exemple, ce qui conduit à des modèles markoviens d'ordre supérieur à 1. Nous n'en parlerons pas ici.

$X_{n-1}$. C'est le modèle markovien[6] :

$$P(X_n = v \mid X_m,\ m < n) = P(X_n = v \mid X_{n-1}).$$

Ce modèle est donc caractérisé par les probabilités qu'une valeur $v$ succède à une valeur $u$, quantités notées $\pi(u, v)$ et qualifiées de transitions.

La quantité $\pi(u, v)$ (qui est la probabilité, sachant qu'à une position on a la lettre « $u$ », d'avoir un « $v$ » à la position suivante) sera estimée de façon très naturelle par

$$\hat{\pi}(u,v) = \text{proportion des « } u \text{ » de la séquence suivis d'un « } v \text{ »}$$

(et le statisticien montre que c'est « le meilleur estimateur » que l'on peut inventer). Donc si l'on note N($uv$) le nombre de fois où le mot « $uv$ » apparaît dans la séquence, et N($u$+) le nombre de fois où « $u$ » apparaît, suivi de n'importe quoi [donc $N(u+) = \sum_w N(uw)$ ], on a

$$\hat{\pi}(u,v) = \frac{N(uv)}{N(u+)}.$$

Il y a donc équivalence entre connaître tous les N($uv$) et connaître toutes les transitions $\pi(u, v)$. Prendre en compte les nombres d'apparition des mots de deux lettres, c'est, de fait, se placer dans un modèle de Markov.

Par exemple dans le texte depuis le début de cet article, il y a 571 fois la lettre « c ». 153 fois elle est suivie d'un « e » ; donc la transition de « c » vers « e » peut être estimée par 153/571 = 26,80 % ; 115 « c » sont suivis d'un « o » donc $\hat{\pi}(c,o)$ = 115/571 = 20,14 % etc… et 2,63 % des « c » sont suivis d'un blanc : $\hat{\pi}(c,\ )$ = 2,63 %.

La question est : si l'on copiait au milieu de cet article une ou deux pages écrites par mon collègue T. Tournesol – dont le style est certainement différent du mien –, l'ordinateur serait-il capable de retrouver le passage « collé » uniquement par le décompte des mots de deux lettres, ce que nous avons appelé le style ? Autrement dit, peut-on découvrir ce qui correspond au modèle de Markov de transition $\pi$ (celle de cet article, dont on vient de donner l'estimation de quelques termes) et ce qui correspond au modèle de Markov de transition, disons $\pi_T$, propre au Professeur Tournesol ?

On cherche à découvrir deux chaînes de Markov cachées au sein du texte : on parle donc de modèle de chaînes de Markov cachées.

**Et en génomique ?**

Développons la réponse en revenant à notre problème de génomique. Imaginons un instant que la suite des {a, c, g, t} le long d'un chromosome suive un modèle de chaîne de Markov – nous reviendrons sur cette hypothèse au § 6. Il n'y a aucune raison pour que la matrice de transition au sein des gènes (notons la $\pi_G$) soit la même que la transition dans l'intergénique (notons celle-ci $\pi_I$)

(6) La notation P(A | B) se lit « probabilité de A sachant que l'on a B » ou bien « probabilité de A conditionnellement à B ».

Toute la démarche est fondée sur deux idées simples :
• prendre en considération les N($uv$), c'est à dire combien de fois chaque mot de deux lettres apparaît, c'est se placer dans un modèle de Markov ;
• les transitions ne sont pas les mêmes dans les gènes et dans l'intergénique.

Le défi semble à première vue impossible à relever : on demande de trier les positions en deux catégories (génique/intergénique) sans nous dire comment reconnaître l'un de l'autre, sans nous dire ce que valent les matrices $\pi_G$ et $\pi_I$ ! Néanmoins, les mathématiciens ont trouvé une solution : ils ont étendu au cas des chaînes de Markov une idée surprenante et efficace qu'un collègue anglais, M. Dempster, avait eue en 1977, l'algorithme EM.

**L'algorithme EM**

Décrivons-le sur un exemple simple. Supposons que l'on veuille estimer la taille moyenne des hommes et la taille moyenne des femmes dans une population et que l'on ait mesuré la taille des individus – mais sans avoir noté quelles sont les mesures relatives à un homme et celles relatives à une femme(7).

1°) Si l'on savait quelles mesures $X_i$ correspondent à un homme, il serait élémentaire d'estimer $\mu_H$ par la moyenne de l'échantillon des hommes $\overline{X_H} = \sum_{\text{hommes}} X_i$, comme il serait élémentaire d'estimer la variance de cet échantillon. Les quantités analogues s'obtiendraient également chez les femmes.

2°) Si l'on connaissait les paramètres, on connaîtrait les densités de la gaussienne-homme et de la gaussienne-femme en chaque valeur $X_i$ observée, et en les comparant on pourrait estimer la probabilité $p_i(H)$ pour que l'individu $i$ soit un homme (et donc $p_i(F) = 1 - p_i(H)$ pour que ce soit une femme).

Mais si chacun de ces calculs est accessible séparément, on ne peut pas mener les deux simultanément : chacun nécessite le résultat de l'autre pour être fait !

Dempster a proposé de commencer par répartir au hasard les données en deux paquets.

**Pas E** : On « fait comme si » on avait correctement trié hommes et femmes ; on sait donc calculer les paramètres des deux gaussiennes.

**Pas M** : On décide que l'individu $i$ est un homme avec probabilité $p_i(H)$, sinon c'est une femme. On procède donc à un tirage au sort auxiliaire, qui « réaffecte » chaque individu à « Homme » ou « Femme »(8).

Et l'on peut retourner au Pas E. Que se passe-t-il si l'on alterne le pas E (pour « estimation ») et le pas M (pour « maximisation ») ? Seule une démonstration

(7) On modélisera la taille des hommes par une loi gaussienne $\mathcal{N}\left(\mu_H;\sigma_H^2\right)$ et celle des femmes par une loi $\mathcal{N}\left(\mu_F;\sigma_F^2\right)$.

(8) Cet algorithme est qualifié de « stochastique » à cause de ce tirage aléatoire (algorithme SEM) ; il existe d'autres variantes : par exemple l'individu $i$ sera compté chez les hommes en proportion $p_i(H)$ et chez les femmes en proportion $1 - p_i(H)$…

mathématique peut répondre à cette question. Dans le cas simple présenté ici, les démonstrations montrent que les estimations successives des espérances $\mu_H$ et $\mu_F$ et des variances $\sigma_H^2$ et $\sigma_F^2$ tendent vers « ce que l'on peut faire de mieux », les valeurs du maximum de vraisemblance et que l'on finit par « bien évaluer » les probabilités $p_i(H)$ et $p_i(F)$ (voir figure 4).

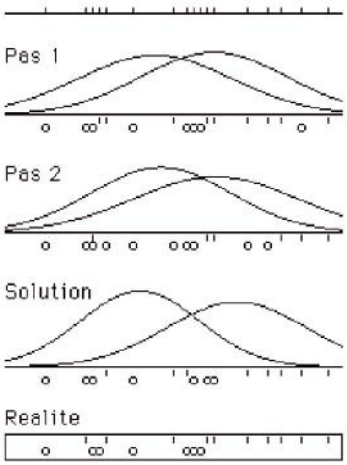


Figure 4 : on mesure la taille de 17 personnes sans savoir s'il s'agit d'hommes (H) ou de femmes (F) (ligne du haut).
Pas 1 : on répartit au hasard ces points entre H (indiqués |) et F (indiquées o). On ajuste une loi gaussienne sur les H et une autre loi gaussienne sur les F.
Pas 2 : on compare pour chaque mesure la probabilité qu'elle soit H ou F et l'on réaffecte cette mesure à H ou F selon ces probabilités.

On réitère cette procédure un grand nombre de fois et l'on fait la moyenne des résultats trouvés. Cette moyenne se stabilise autour d'une solution. Ici la réalité était connue, ce qui permet de la comparer à notre solution.

**Où l'on parvient enfin à annoter**

Reste à transposer cet exemple à celui des chaînes de Markov cachées – et à reprendre les démonstrations dans un cadre techniquement beaucoup plus complexe. Chaque année plusieurs thèses de Mathématiques s'intéressent à ce problème et à ses variantes. C'est dire que tout n'est pas fini et que la recherche en Mathématiques doit encore apporter des résultats théoriques dans ce domaine.

Ensuite une telle recherche n'a de sens que si elle débouche sur des logiciels mis à la disposition des biologistes pour analyser leurs données. En pratique on travaille avec beaucoup plus que deux types (génique/intergénique) :
– d'une part les chromosomes comportent deux brins qui sont orientés en sens inverses (voir la figure 3). Les gènes peuvent être écrits sur l'un ou l'autre brin, alors que seul le texte d'un brin est conservé dans les banques de données (on retrouve l'autre brin si l'on sait que face à un « a » figure toujours un « t » et que face à un « c » figure toujours un « g »).
– d'autre part, on l'a vu, chez les organismes autres que les bactéries les gènes sont composés d'exons (qui codent vraiment pour les protéines) et d'introns (qui sont éliminés avant traduction en protéines). Bien sûr ce qui intéresse le biologiste c'est le texte pertinent, celui des exons : il indiquera quelle protéine est codée, ce qui est un pas sur la compréhension du rôle du gène dans l'organisme.

L'un des programmes réalisant ce travail, *Genscan*, utilise 27 types de chaînes de Markov et des modèles markovien d'ordre 5 (fondés sur les décomptes des mots de 6 lettres). Le problème est rendu « plus facile » en pratique – mais « plus difficile »

quand il s'agit de démontrer la pertinence de ce que l'on fait – si l'on prend en compte un certain nombre de signaux : par exemple un gène commence toujours par un triplet START « atg » et s'achève par un triplet STOP (« tag », « tga » ou « taa »).

Une étape ultérieure pour le biologiste consistera à rechercher dans les banques de données si ce qu'il vient d'identifier comme un gène – ou comme un ensemble d'exons – « ressemble » à des séquences déjà connues, annotées, identifiées, voire dont la fonction biologique est connue (procédures d'alignement). Ici encore chaînes de Markov et chaînes de Markov cachées interviennent, et ici aussi beaucoup de travail théorique et pratique reste à faire.

## 5 – Transferts horizontaux

Signalons une autre application des chaînes de Markov cachées en génomique, application dont on peut penser qu'elle va être de plus en plus employée dans les années à venir.

Un phénomène a été omis dans la description des différents aléas de la génomique que nous avons présentée au § 2, celui, considéré aujourd'hui comme très répandu, du transfert horizontal(9) d'un grand morceau de chromosomes (disons de quelques centaines de lettres à un million de lettres !) d'un emplacement vers un autre, soit au sein d'un même organisme, soit d'un organisme à un autre. Les bactéries, par exemple, peuvent se coller les unes aux autres et réaliser des transferts de matériel génétique. Les virus peuvent aussi « copier » un texte d'ADN chez un hôte pour ensuite aller le « coller » chez un autre. Et si une bactérie « invente » par le hasard des mutations un gène utile (pour résister à un poison, pour digérer une nouvelle matière plastique, …) le mécanisme que nous décrivons fera que ce gène sera peut-être rapidement copié par d'autres bactéries. C'est de l'espionnage industriel au niveau des unicellulaires !

Or il s'avère que les styles varient d'un organisme à un autre. Pour revenir aux langages humains, les styles (toujours en ce sens restrictif de fréquences des mots) varient entre la langue française et la langue allemande. Par exemple, les « tz », « sch », « ein » seront plutôt révélateurs d'un texte écrit en allemand.

Les chaînes de Markov cachées fournissent alors un outil pour rechercher les transferts horizontaux. Si l'on colle dans un ouvrage en français, *Les Misérables* disons, une page de Schiller, les algorithmes fondés sur les chaînes de Markov vont retrouver « ce qui n'est pas écrit dans le style de Victor Hugo » (peut être pas à la lettre près, mais en tout cas avec assez de précision pour qu'un œil humain puisse rapidement finir le travail).

Si l'on connaît le style d'une bactérie – appelons-la *Bacillus subtilis* – et si par nos algorithmes on détecte une région écrite dans un autre style, on pourra soupçonner qu'il s'agit du résultat d'un transfert horizontal. L'œil d'un biologiste pourra alors chercher d'autres caractéristiques des transferts (les « collages » laissent des traces visibles) et tâcher de comprendre pourquoi le gène ainsi transféré a été conservé par l'évolution. On a ainsi par exemple détecté des transferts chez *B. subtilis* d'un gène donnant une meilleure résistance à l'arsenic, gène apparu chez

(9) par opposition au « transfert vertical » parent-enfant.

un autre organisme et copié par ce bacille[10].

## 6 – Aléatoire et génomique

On peut s'interroger sur le rôle de l'aléatoire dans ces analyses. Est-ce raisonnable de traiter le génome d'une bactérie – ou le génome humain ! – comme une suite aléatoire ? Comme dans notre exemple de risque opératoire, on n'a pas de ces « répétitions d'expériences » chères au statisticien. On peut concevoir que l'évolution au cours de millénaires se soit faite de façon aléatoire – contrainte par la pression de la sélection – et qu'il en demeure une trace dans les génomes actuels.

Mais employer des modèles markoviens pour analyser les génomes, ce n'est pas prétendre que la Nature « tire » les lettres en respectant scrupuleusement des transitions. C'est prendre conscience qu'une analyse doit tenir compte de la fréquence des quatre lettres « a », « c », « g » et « t » – dans un génome « riche en a et pauvre en c », on attendra davantage de mots contenant de nombreux « a » que de mots contenant des « c » ; donc si l'on trouve un mot tel que « cctgctcac », c'est sans doute que l'organisme « en a besoin ». Le même raisonnement vaut pour les mots de deux lettres (ou plus) : « cg » est un mot biologiquement fragile, donc rare. Un mot tel que « cgacgcg », si on le rencontre, est sans doute indispensable à la survie de son porteur.

Et travailler conditionnellement aux décomptes des mots de deux lettres, c'est, qu'on le dise ou non, travailler dans un modèle markovien. Il est sans doute alors judicieux de ne pas faire, comme Monsieur Jourdain, du modèle de Markov sans le savoir et de s'appuyer sur une théorie abondamment étudiée dans la littérature mathématique.

L'emploi de chaînes de Markov cachées pour localiser les gènes ou les exons – de même que des transferts horizontaux – sera de plus en plus systématique au fur et à mesure que les données de séquençage vont se multiplier : trouver quelque 10 000 nouveaux gènes par mois dans un ensemble de nouvelles séquences approchant le milliard de lettres et provenant d'innombrables espèces ne pourra se faire sans une aide informatique fondée sur des algorithmes soigneusement optimisés – donc évalués mathématiquement.

## Courte bibliographie

Muri-Majoube F., Prum B. : Une approche statistique de l'analyse des génomes. *La Gazette*, *revue de la Société Mathématique de France*, Juillet 2001.

Pevzner P.P. *Computational molecular biology*. MIT Press, 2000.

Prum B. : Statistique et Génétique, deux sciences n'ayant cessé de s'enrichir mutuellement in *Development of Mathematics 1950-2000*, J.P. Pier ed. Birkhäuser, 2000.

Waterman M.S. *Introduction to computational biology*. Chapman Hall, 1995.

(10) Searching gene transfers on Bacillus Subtilis using hidden Markov models, Bize, L., Muri, F., Samson, F., Rodolphe, F., Ehrlich, S.D., Prum, B. and Bessieres, P. *Recomb'99 Proceedings*, 43-49.