Des mathématiques dans nos cellules ? Partie 1

Bernard Prum

▶ **To cite this version:**

Bernard Prum. Des mathématiques dans nos cellules ? Partie 1. Images des Mathématiques, 2009, http://images.math.cnrs.fr/Des-mathematiques-dans-nos.html. hal-00586146

HAL Id: hal-00586146
https://hal.science/hal-00586146

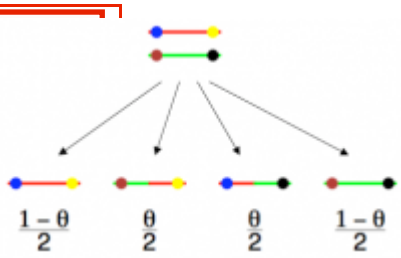
Submitted on 15 Apr 2011

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Des mathématiques dans nos cellules ?

To print higher-resolution math symbols, click the
Hi-Res Fonts for Printing button on the jsMath control panel



Partie 1

Le 14 septembre 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.

Ce premier article se place avant la lecture des séquences d'ADN (le séquençage) et traite de l'héritabilité des caractères au niveau des individus ou des espèces. Le second (ici), parcourera les génomes à la recherche des gènes. [1]

Introduction

LA baleine est-elle plus proche du requin ou de la vache ? du chat ou de la vache ?

Quel risque court chacun d'entre nous d'avoir un ADN suffisamment proche de celui d'un autre pour être accusé à tort d'un crime ?

Peut-on évaluer le risque pour qu'une jeune fille ait un jour un cancer du sein ?

Peut-on connaître la taille qu'aura un bébé quand il aura atteint l'âge adulte ? Peut-on calculer le rendement à l'hectare qu'aura une variété de blé ?

Peut-on calculer un médicament ?

Ces questions fascinantes n'ont pas, aujourd'hui, toutes la même réponse (voir **Section 5**) .

Mais les chercheurs comprennent chaque jour plus finement les règles qui conduisent de la biochimie, au cœur de nos cellules, jusqu'aux caractères observables et tout porte à croire que, d'ici quelques années, on pourra répondre positivement à toutes ces questions.

On a pu noter que ces questions relevaient du domaine de la biologie, mais que, incidemment, elles empruntaient leur vocabulaire aux mathématiques : « proche, risque, évaluer, calculer... ». De même que depuis longtemps la frontière est ténue entre physique et mathématiques, on constate de nos jours que celle entre biologie et mathématiques s'estompe sans cesse.

Si nous savons tous que la physique utilise sans cesse des mathématiques [2], nous avons tous l'idée que la biologie établit des catalogues d'espèces (*l'herbier* de notre jeunesse), décrit des organes (*dissections* de grenouilles et dessin de leur système nerveux), et ne fait pas ou peu

appel aux mathématiques [3].

Or depuis une trentaine d'année se produit une véritable révolution conceptuelle, due essentiellement aux progrès de la technologie. À côté des sciences toujours vivaces et fécondes (l'écologie, qui utilise des modèles spatiaux, la mécanique des os ou des fluides organiques, le sang dans les veines ou l'air dans les poumons, etc.), deux sciences jumelles, la *génétique* et la *génomique*, connaissent un développement qui envahit tout le champ. Et toutes deux font grand usage des mathématiques.

La génétique est d'abord la science de la transmission des caractères — nous appellerons désormais *phénotype* un caractère observable [4]. Le plus surprenant — de nos jours — est que l'on a longtemps fait de la génétique sans gènes — nous préciserons comment le sens de ce mot a évolué depuis 150 ans.

Avant de nous plonger dans l'univers des chromosomes, des réseaux de régulation des gènes et de la recherche des mécanismes du vivant, parlons un peu de génétique — ou, pourquoi pas, tant que l'on n'a pas de gène, de *phénétique*, science de l'héritabilité des phénotypes.

2. De la « phénétique » qualitative...

On n'a certes pas attendu **Mendel (1822-1884)** pour savoir que certains caractères étaient transmis des parents aux enfants (« il a le menton de son père et les yeux de sa grand-mère ! »).

À la fin du dix-neuvième siècle, **Francis Galton**, un cousin de **Darwin**, tenta par un modèle mathématique de prédire la taille des hommes (nobles anglais, il va sans dire) en fonction de celles de leurs pères (un eugéniste comme Galton comptait pour néant l'influence maternelle — l'idée dominante était que tout venait du père, à travers un animalcule lové depuis Adam dans le spermatozoïde !). Il s'est trouvé que les grands étaient moins grands que leurs pères, les petits moins petits que leurs pères, qu'il y avait une « régression » vers les tailles moyennes. De là le nom « droite de régression » que les statisticiens utilisent pour prédire une variable Y à partir d'une variable plus facile à mesurer, liée à X [5]. En eut-il été autrement que l'on parlerait peut-être aujourd'hui de « droite de progression » [6].

Si l'aspect humain de ces travaux choque aujourd'hui, il faut bien admettre que les sélectionneurs de plantes ou d'animaux perpétuent cette approche. Des millénaires de sélection empirique ont permis aux mésoaméricains de produire à partir du **téosinte** (épis de 2,5 cm, rendement de 1 quintal/hectare) le maïs cultivé (épis atteignant 30 cm, plus de 84 quintaux/hectare) — on parlerait de même de production de viande ou de lait.

Cette succession d'améliorations s'est faite par une modélisation de plus en plus précise. Les éleveurs « sentent » quel mâle croiser avec quelle femelle pour avoir de beaux produits (sic). Petit à petit cette démarche s'est quantifiée, comme nous allons le voir.

On a, dans l'élan du texte, écrit le mot *modélisation*. Il est au cœur de notre discours. Modéliser un phénomène, c'est en considérer une représentation à la fois abstraite et imparfaite. Donnons un exemple : supposons que le poids adulte Y d'une vache soit une fonction simple de celui X_P de son père et de celui X_M de sa mère, par exemple $Y = 0,10\,X_P + 0,90\,X_M$. On voit tout de suite qu'un tel modèle « ne tient pas la route », ne serait-ce que parce que toutes les vaches issues d'un même croisement auraient exactement le même poids. Il convient de le nuancer en

introduisant une part *non expliquée* :

$$Y = 0,10\, X_P + 0,90\, X_M + \varepsilon \qquad \textbf{(1)}$$

et le concept de *non expliqué* conduit à celui de *variable aléatoire* [7].

Nous avons donc introduit deux notions :

- celle de *modèle* : l'équation (1) est-elle raisonnable ?, est-elle confortée par les observations ? ne serait-il pas pertinent de la compliquer en :

$$Y = 0,10\, X_P + 0,90\, X_M + 0,001\, X_P X_M + \varepsilon \qquad \textbf{(2)}$$

voire en une équation plus compliquée ?

- celle de *variable aléatoire*, quantité destinée à modéliser... le non-modélisable ; de fait cet ε regroupe d'innombrables variables (non mesurées) intervenant dans la détermination de Y (alimentation, infections, etc.). Il est toujours surprenant de constater que ces modèles aléatoires, fondés sur la reconnaissance d'une incapacité à bien expliquer les observations, se montrent d'une redoutable efficacité pratique [8].

Nous n'entrerons pas ici dans les multiples tâches qui incombent alors au mathématicien — en l'occurrence au statisticien : ajuster au mieux les coefficients figurant devant X_P et X_M — nous avons mis $0,10$ et $0,90$, peut-on faire mieux ? — (estimation), choisir plutôt le modèle (1) ou plutôt le modèle (2) (test).

Gregor Mendel, un moine de l'Abbaye augustinienne de Saint Thomas à Brno (dans l'actuelle République Tchèque), a révolutionné la génétique en lui donnant un support concret (à vrai dire plus conceptuel qu'observé, à cette époque). En termes modernes, il a conclu de ses fameuses expériences sur les pois (portant tout de même sur près de $30\,000$ plantes !) que chaque caractère phénotypique [9] était dû à deux *allèles* — deux formes d'un *gène* — hérités l'un de son père, l'autre de sa mère. Cette audace explique sans doute pourquoi son œuvre est restée oubliée durant 40 ans.

Notons classiquement **a** et **A** ces allèles. Si un individu est *homozygote*, c'est à dire qu'il porte deux allèles identiques (il « est » **aa** ou bien il est **AA**), c'est bien sûr cet allèle qu'il transmettra à des descendants. Mais s'il est *hétérozygote* (à savoir **aA**), alors :

- *Loi numéro 1* : il transmet à chacun de ses descendants (directs) soit l'allèle **a** avec probabilité $1/2$, soit l'allèle **A** avec probabilité $1/2$.

Cette loi a ceci de fantastique qu'elle est la première a introduire la notion de probabilité au sein de la réalité elle même : la transmission des allèles dans les ovules et les spermatozoïdes. Jusqu'alors, les Probabilités semblaient être utiles aux joueurs (de dés, de cartes, du Loto,...), avec la génétique elles trouvaient un domaine d'application « sérieux » [10] et l'essentiel de la Statistique moderne s'est construite, autour de **Fisher**, à cette interface [11].

Mendel a complété cette loi par deux autres, l'une encore admise, l'autre complètement réfutée (ce qui sera la base de l'analyse de liaison) :

- *Loi numéro 2* : Ces transmissions aléatoires sont indépendantes pour un descendant et

pour un autre ;

- *Loi numéro 3* : Ces transmissions aléatoires sont indépendantes pour un phénotype et pour un autre.

Ce qu'il convient de prendre en compte ce n'est plus le poids du père X_P et celui de la mère X_M, mais les allèles qu'ils portent. Un modèle (simpliste) où l'on expliquerait le poids Y de la vache par un modèle à deux allèles serait :

$$Y = \begin{cases} P_0 + \varepsilon & \text{si l'individu est } \mathbf{aa} \\ P_1 + \varepsilon & \text{si l'individu est } \mathbf{aA} \\ P_2 + \varepsilon & \text{si l'individu est } \mathbf{AA} \end{cases} \quad \textbf{(3)}$$

Ceci, combiné avec les lois 1 et 2 de Mendel, permet de générer (p.ex. de *simuler* sur un ordinateur) des familles où le poids de chaque individu est connu. Reste au statisticien à estimer les paramètres « poids » moyens P_0, P_1 et P_2 au vu d'un échantillon de telles familles, la difficulté nouvelle étant que les « variables explicatives » ne sont plus observées (comme l'étaient X_P et X_M), mais qu'il convient de les deviner, ou plutôt d'attribuer des probabilités aux faits que chaque individu soit **aa**, **aA** ou **AA**.

En écho au terme *phénotype* (l'apparence extérieure de l'individu), appelons *génotype* les allèles portés par un individu en une position donnée du génome [12].

2.1 ... à la génétique qualitative

Nous avons surtout évoqué jusqu'ici des phénotypes quantitatifs, comme le poids d'un individu. Le généticien (et le citoyen en général) porte aussi un grand intérêt aux phénotypes qualitatifs, surtout quand il s'agit du caractère malade/non malade pour une maladie donnée. On sait depuis des siècles que certaines de ces maladies ont une dimension génétique (on disait plutôt « sont héréditaires ») : l'hémophilie [13], la chorée de Huntington [14], la mucoviscidose [15], les myopathies [16], comme la myopathie de Duchenne, sont parmi les plus connues [17].

Savoir si une maladie a une dimension génétique est un problème complexe : il est difficile de trier ce qui est dû aux gènes transmis de ce qui est dû à un environnement partagé par les membres d'une même famille (on pourrait conclure hâtivement que « posséder un château », « aimer le poisson » ou « parler anglais » sont des phénotypes transmis génétiquement). C'est aussi une question lourde de conséquences, depuis le « conseil génétique », préconisant à un couple de ne pas procréer, jusqu'à engager, on le verra aux sections suivantes, une recherche pour comprendre le déterminisme génétique et tâcher d'y porter remède.

Depuis Mendel on dispose de la notion, conceptuelle on l'a dit, d'allèle. On peut bâtir un modèle, par exemple à nouveau fondé sur un seul gène bi-allélique :

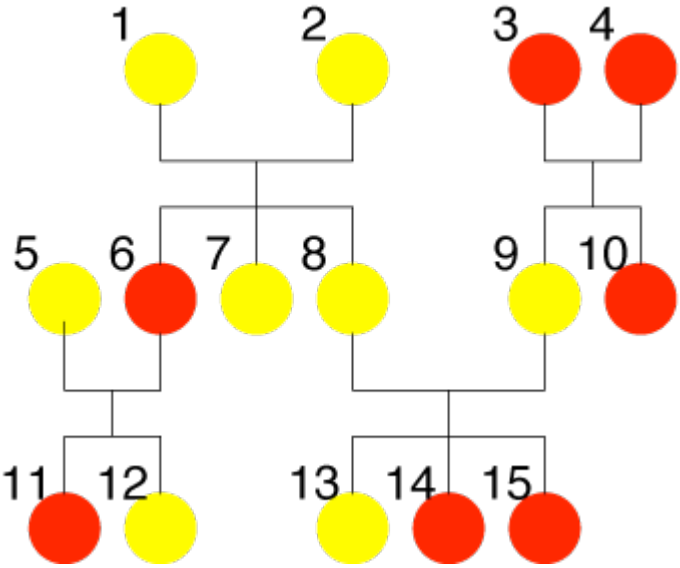
$$\begin{cases} \mathbb{P}(\text{malade}) = f_0 & \text{si l'individu est } \mathbf{aa} \\ \mathbb{P}(\text{malade}) = f_1 & \text{si l'individu est } \mathbf{aA} \\ \mathbb{P}(\text{malade}) = f_2 & \text{si l'individu est } \mathbf{AA} \end{cases} \quad \textbf{(4)}$$

Les probabilités d'être malades selon les allèles portés s'appellent les *pénétrances* et la question

posée au vu de familles comportant un certain nombre de malades est : « Les pénétrances de **aa**, **aA** et **AA** sont-elles les mêmes (pas de caractère génétique pour ces allèles) ou non ? ». Et, comme la version quantitative décrite il y a un instant, les génotypes des invidus font partie (pour le moment) des choses à deviner.

3. À la recherche d'un modèle génétique de maladie

Le graphe suivant représente une famille, dont les membres ont été numérotés ; le sexe des individus n'est pas indiqué. **1** et **2** ont eu trois enfants, **6**, **7** et **8** ; **6** a épousé **5** et ils ont eu pour enfants **11** et **12**, etc... Ont été marqués d'un rond rouge les individus malades, d'un rond jaune les individus sains.



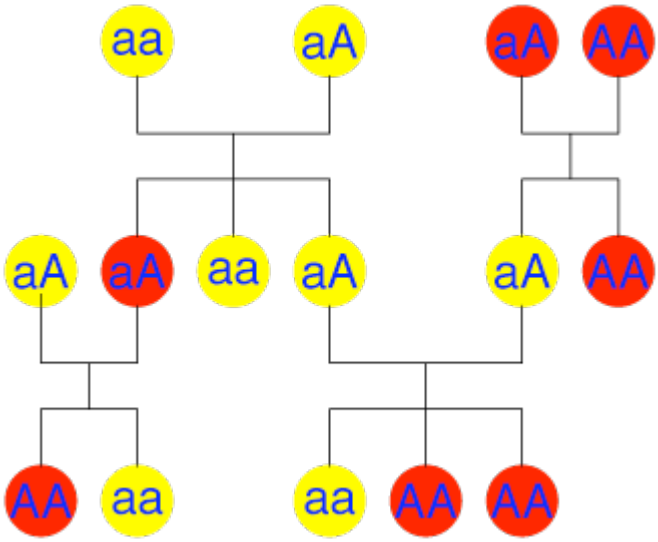
QUESTION : Admettant que cette maladie est monogénique (équation(4)), trouver (estimer !) :

- les trois pénétrances f_0, f_1 et f_2 ;
- le génotype **aa**, **aA** ou **AA** porté par chaque individu.

Attaquer ce problème par la « force brute », c'est considérer toutes les combinaisons de **aa**, **aA** et **AA** compatibles avec la transmission des allèles : si par exemple on suppose que **1** et **2** sont tous les deux **aa**, il en est nécessairement de même de **6**, **7** et **8** ; et même si **5** porte (un ou deux) **A**, on sait que **11** et **12** ne peuvent pas être **AA**.

L'énumération de toutes les possibilités est déjà un problème : il y a $190\,425$ combinaisons possibles (essayez de les dénombrer !)— et encore ici, on ne se fonde que sur *une* famille de 15 individus ; dans la réalité on dispose heureusement (?) de dizaines ou de centaines de familles, souvent de milliers d'individus [18].

Cette approche demande ensuite pour chaque combinaison de calculer la probabilité d'observer ce que montre la figure comme fonction de (f_0, f_1, f_2) et à maximiser cette fonction (appelée la *vraisemblance* [19]). Convenons que cette seconde étape est aisée, comme le montre l'exemple de configuration de la figure 2 :



dans laquelle on compte :

	aa	aA	AA
sains	4	4	0
malades	0	2	5
total	4	6	5

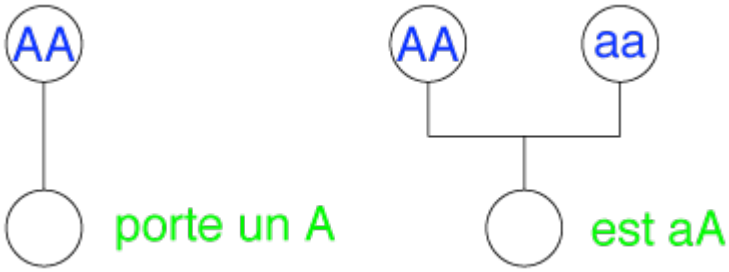
et un calcul élémentaire conduit à estimer f_0 par 0 ; f_1 par $1/3$; et f_2 par 1.

Le problème est donc de parcourir intelligemment l'ensemble des combinaisons, en écartant par exemple dès que possible de gros paquets de combinaisons.

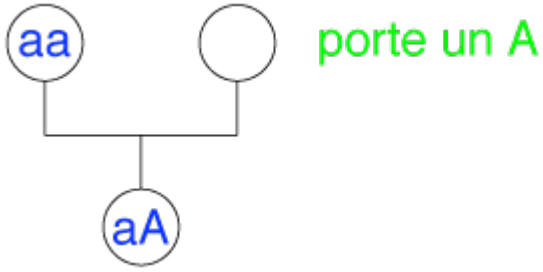
Quand on doit chercher le maximum d'une fonction $f(x_1, x_2, \ldots, x_n)$, sauf cas exceptionnellement simple [**20**], on procède par approximations successives : on choisit comme on peut un point M_1 et l'on cherche dans quelle direction se déplacer pour améliorer f. Toute une branche des mathématiques traite du problème du choix de la direction à prendre et détermine par exemple « de combien il faut se déplacer dans cette direction » avant de chercher à nouveau « vers où » aller, pour être assuré d'arriver au maximum.

Ici l'ensemble à explorer est l'ensemble fini (mais très gros) de configurations — on a convenu que, pour chaque configuration le choix optimal des pénétrances était « aisé ». On peut penser à partir d'une configuration « choisie comme on peut », changer le génotype d'un individu, regarder si la vraisemblance a augmenté et recommencer. Malheureusement, si l'on change le génotype d'un seul membre de la famille, on a toutes les chances de trouver une configuration impossible (deux parents **AA** pour un enfant **aa**, par exemple).

Il faut donc savoir propager l'information « quel est le génotype en un nœud du graphe » aux nœuds voisins. On constate sans surprise que ceci va se faire à l'aide de règles entre parents et enfants, telles celles-ci (on suppose connaître les génotypes indiqués en bleu — on en déduit la conclusion indiquée en vert) :



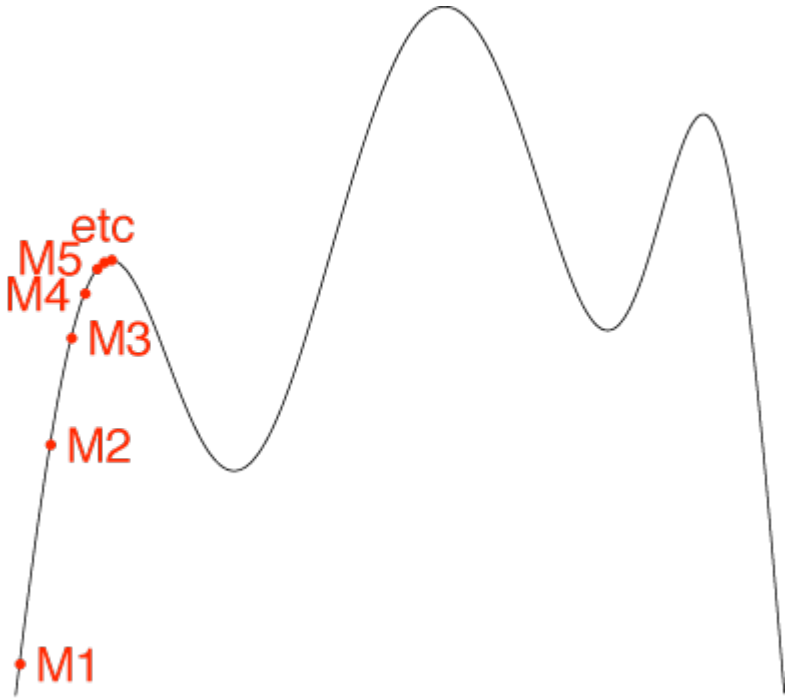
mais aussi entre parents d'un même enfant, par exemple [**21**] :



L'idée est alors de partir d'une configuration, de choisir au hasard un individu (ou seulement un ancêtre dans une famille), de modifier son génotype et de propager cette information pour obtenir une configuration cohérente. Si elle est meilleure (plus vraisemblable), on la garde et on recommence, sinon on essaye de modifier le génotype d'un autre individu.

Notons d'abord que cette propagation doit, elle aussi, être aléatoire : si les règles évoquées concluent « tel individu porte un **A** », il faut décider si on lui attribue le génotype **AA** ou **aA**, et l'algorithme tirera entre les diverses possibilités, par exemple, avec même probabilité $1/2$.

À chaque itération de ce processus, la vraisemblance augmente (ou reste la même). Comme elle est bornée (il n'y a qu'un nombre fini de configurations), elle finira par se stabiliser. Un critère bien choisi [**22**] permet de l'arrêter.



Mais il est bien connu des randonneurs — et des mathématiciens — que monter sans cesse selon la plus grande pente ne conduit pas nécessairement au sommet de la plus haute montagne. Comme sur la figure 3, la suite des points M_n visités risque de se faire piéger en un maximum local.

Et la même mésaventure peut survenir à notre algorithme de visite de configurations.

La solution est la même dans les deux cas : quand on trouve une configuration *moins bonne* que la précédente, au lieu de la rejeter sans appel, on l'accepte avec une certaine probabilité ε_n. Au début de l'algorithme (le numéro n du pas est petit), on autorise facilement cette attitude qui semble contre-productive (ε_n est grand, il vaut par exemple $50\,\%$) : si l'on part dans « une mauvaise bosse », on a de bonnes chances d'en sortir. Et, au fur et à mesure que n augmente, on fait diminuer ε_n pour stabiliser le processus.

On trouvera, en cliquant **ici**, un « dessin animé » montrant comment fonctionne cette procédure.

Et c'est là que commencent les mathématiques difficiles. Comment choisir la vitesse avec laquelle on fait tendre ε_n vers zéro pour être *assuré* que la suite des points M_n grimpe finalement tout en haut ? Et cette assurance acquise, comment y arriver le plus vite possible ? On conçoit que cette vitesse optimum dépend des pénétrances : si seuls les **AA** sont malades, cas dit récessif, ou si seuls les **aa** se portent bien, cas appelé dominant, on aura tôt fait de s'en apercevoir ; si les trois pénétrances sont voisines, il y aura plein de maximum locaux et il faudra être très minutieux dans la décroissance de ε_n. Et donc, comment gérer cette vitesse *en fonction des pénétrances* (ou plutôt de leurs estimations, itération après itération).

Pour conclure, notons, comme vous l'aviez tous deviné, que la **figure 2** et les pénétrances indiquées alors donnent la solution du problème posé par la **figure 1**.

4. Les marqueurs

Depuis les années 1990, on dispose de *marqueurs* le long des génomes. Ce sont des variants que l'on peut mesurer facilement et à bon marché [23]. On peut alors suivre à la trace la transmission de ces marqueurs dans des familles et chercher quel marqueur est *significativement* plus souvent transmis en même temps que la maladie [24].

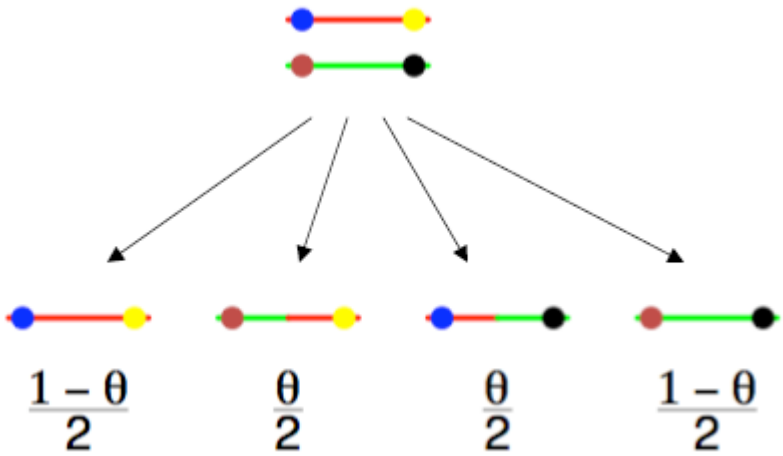
Trouver un tel marqueur, ce n'est pas répondre que le marqueur est responsable de la maladie, mais qu'il est « à côté », sur le chromosome, d'un locus responsable le la maladie, on parlera d'un locus *étiologique*. Pour comprendre ceci, il faut introduire les recombinaisons.

4.1 Les recombinaisons

On sait maintenant que le génome est constitué de *chromosomes*, longues chaînes linéaires de molécules accrochées les unes derrière les autres. Ces molécules sont de quatre types, tymine, cytosine, adénine et guanine, repérés par leurs initiales **t**, **c**, **a**, **g**. [25] Ces chaînes ont, chimiquement, une orientation, de sorte qu'un chromosome peut être considéré comme un texte, par exemple ... **atgccttgatccttctga**... [26].

Nous pouvons alors découvrir la vraie nature des allèles : un changement dans ce texte. Il suffit qu'**une** lettre soit changée pour que la signification du texte change [27]. Deux allèles sont donc deux gènes dont les textes diffèrent de une ou plusieurs lettres.

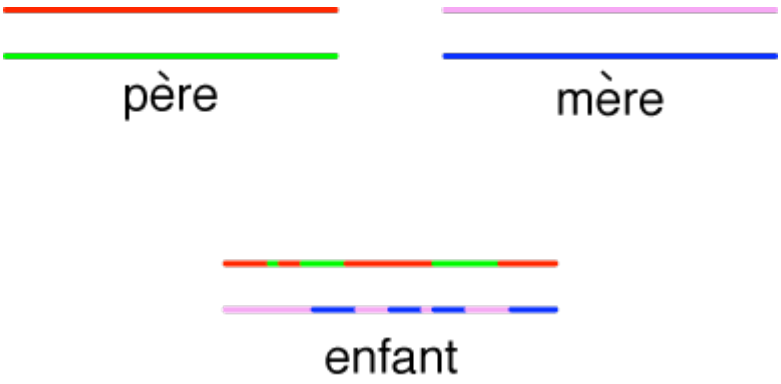
Lors de la reproduction, chez l'Homme par exemple, chaque parent transmet un chromosome de chaque paire. Mais ce n'est pas en général, tel quel, un chromosome hérité de son propre père ou de sa propre mère, mais un mélange ! Sur le schéma suivant figure (des segments d') une paire de chromosomes d'un (seul) parent : un chromosome rouge portant un allèle bleu et, en une autre position, un allèle jaune. L'autre chromosome, peint ici en vert, porte, aux mêmes positions, un allèle rouge et un allèle noir.



Avec une certaine *probabilité* θ, lors de la fabrication d'un gamète (ovule ou spermatozoïde), ces deux chromosomes se recombinent et l'un des deux produits de cette recombinaison (début rouge-fin verte ou début vert-fin rouge) est transmis dans le gamète. Avec la probabilité complémentaire, $1-\theta$, c'est le chromosome rouge *ou* le chromosome vert qui est transmis. Le

schéma indique alors les probabilités des quatre transmissions possibles — y compris les allèles concernés.

De fait, une paire de chromosomes « subit » de multiples combinaisons : si un père, qui porte un chromosome rouge et un chromosome vert, et une mère dont les chromosomes (de la même paire) sont violet et bleu, ont un enfant, voici ce que pourrait être le chromosome de leur enfant ;



On voit bien le rôle sur la diversité biologique que jouent ces recombinaisons. En moyenne, pour chaque enfant, il y a environ 30 recombinaisons « chez le père », et tout autant chez la mère.

La probabilité pour qu'une telle recombinaison ait lieu entre deux positions — ce que nous avons noté θ — est, comme on pouvait s'y attendre, d'autant plus grande que ces positions sont éloignées sur le chromosome [28]. On définit une distance sur le génome, dont l'unité, le centiMorgan (noté cM), correspond à une espérance de nombre de recombinaisons de $0,01$.

Cette notion de gènes répartis sur des segments (pouvant recombiner) met clairement en pièces la « 3ème loi de Mendel » : deux phénotypes gérés par des gènes proches s'hériteront de concert.

4.2 La localisation des gènes de maladies

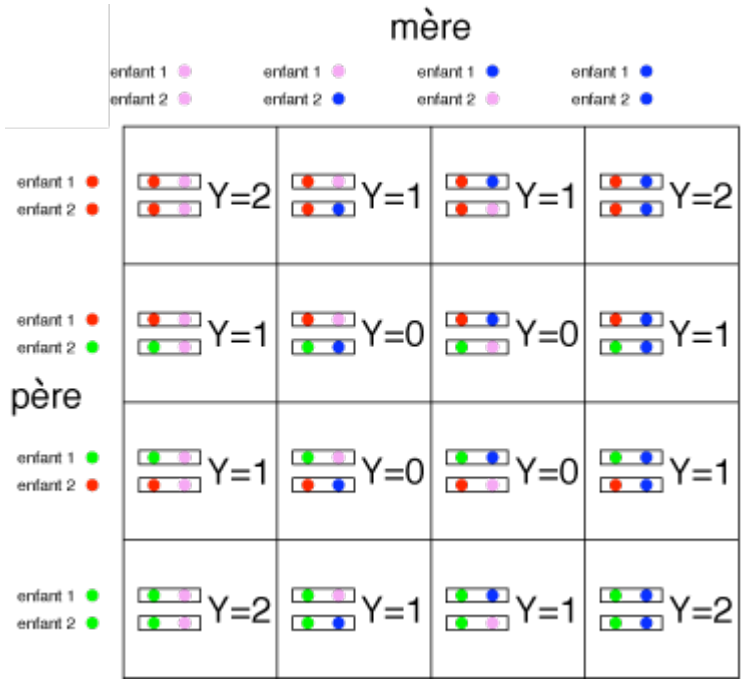
Si en deux positions voisines (θ est petit) il y a, d'un côté un marqueur, de l'autre un gène impliqué dans l'étiologie d'une maladie, les allèles du marqueur se transmettront plus souvent en même temps que les allèles de la maladie que s'ils se transmettaient indépendamment — et ça, le statisticien sait le tester.

Montrons comment sur un exemple simple, celui des *paires de germains atteints* [29]

Considérons le cas (le plus « informatif ») d'une position de marqueur très polymorphe, où l'on sait que les parents portaient en cette position quatre allèles (disons « rouge » et « vert » pour le père, « mauve » et « bleu » pour la mère, comme sur la figure précédente).

Appelons Y le nombre d'allèles partagés par deux enfants atteints. La figure décrit les 16 cas possibles, équiprobables si cette position est *indépendante* de la maladie ; on en déduit alors :

$$\mathbb{P}(Y=0)=\frac{1}{4} \qquad \mathbb{P}(Y=1)=\frac{1}{2} \qquad \mathbb{P}(Y=2)=\frac{1}{4}$$



Sur un échantillon de n couples de germains, on s'*attend* à ce que $Y=0$ pour $n/4$ couples, $Y=1$ pour $n/2$ couples, $Y=2$ pour $n/4$ couples ; on observe respectivement N_0, N_1, N_2 couples de chaque catégorie et, comme on l'apprend en première année de Statistique, un test du χ^2 permet de rejeter ou non l'indépendance.

Ce test très simple (mais, à vrai dire, pas très puissant) permet de conclure « on peut penser qu'il y a un gène lié à la maladie dans un voisinage de la position du marqueur ». Comme aujourd'hui on dispose de marqueurs disons tous les 2 ou 3 cM, on peut ainsi déceler des *régions candidates* à porter un gène étiologique.

Croisant ceci avec la connaissance des gènes situés dans cette région, voire avec une idée de leurs rôles biologiques, on peut concentrer la recherche sur de courts segments de chromosomes, que l'on peut séquencer... pour se ramener à la démarche de la **section 3**.

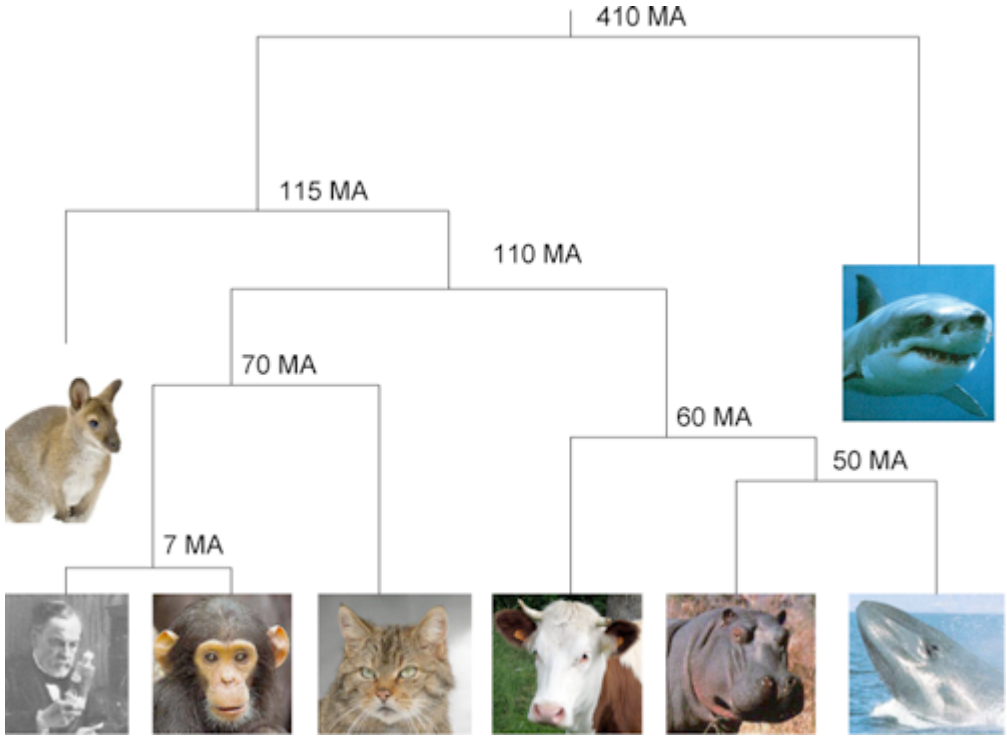
L'exemple historique de cette démarche est la localisation des gènes *brca*1 et *brca*2 (pour *breast cancer*), situés respectivement sur les bras longs des chromosomes 17 et 13, qui augmentent considérablement le risque de cancer du sein des femmes portant une mutation sur l'un de ces gènes [**30**] .

5. Quelques réponses ...

...aux questions de l'introduction.

La baleine, un cétacé bien sûr, avait, il y a environ 60 millions d'années (MA), un ancêtre commun avec les ruminants (dont la vache) ; elle a fait encore chemin commun pendant

quelque 10 MA avec les hippopotames. Pour trouver un ancêtre commun à toutes ces espèces et au chat, il faut remonter à l'apparition des préptothériens (110 MA), dont descendent tous les mammifères, sauf les ornithorynques, les kangourous et les tatous. La division avec les poissons (le requin) est bien sûr plus ancienne (410 MA).



La sous-discipline permettant ces réponses est la *phylogénie*. Elle se propose de dessiner des arbres décrivant comment, à partir d'un ancêtre commun, un phénomène de *spéciation* a séparé deux branches, qui se sont séparées à leur tour, jusqu'à donner les espèces contemporaines (*arbres phylogénétiques*). Le problème est en quelque sorte dual de celui de la génétique qualitative de la **section 3** : au lieu de connaître les filiations et d'en déduire l'héritabilité d'un caractère (une maladie, par exemple), on connaît les caractères et l'on recherche la filiation. Ces caractères ont longtemps été phénotypiques (Linné classait ensemble des plantes dont les feuilles étaient « du même vert » !), ce sont plutôt aujourd'hui des marqueurs [**31**].

La grosse différence est que, fondamentalement, on n'observe que les espèces contemporaines, au moins pour ce qui est de leurs marqueurs ADN [**32**]. Les méthodes *de distance* (regrouper les espèces dont les marqueurs sont les plus proches ; puis les paquets d'espèces dont les marqueurs sont les plus proches, jusqu'à n'avoir finalement qu'un seul paquet) laissent peu à peu la place à des maximisations de vraisemblance, comme celle décrite à la **section 3** :

Le modèle suppose que, au cours de l'évolution, un marqueur m_1 « mute » en un marqueur m_2 en une unité de temps (p.ex. 1 million d'années) avec une probabilité notée $\pi(m_1 ; m_2)$. On peut alors calculer la vraisemblance d'un arbre phylogénétique (caractérisé par sa forme et la longueur de ses branches).

Il s'agit alors d'explorer l'ensemble des arbres phylogénétiques possibles « au dessus » des espèces étudiées [**33**], et d'attribuer des dates (dans le passé) à chaque nœud, afin de maximiser cette vraisemblance. Comme pour la génétique qualitative, il n'est pas question de parcourir toutes les possibiltés, mais d'être extrêmement astucieux pour trouver l'arbre le plus

vraisemblable — et dater les spéciations baleine/vache ou baleine/requin.

L'identification des criminels (ou des pères putatifs) se fonde aujourd'hui sur un ensemble de, par exemple, 10 marqueurs fortement polymorphes : ce ne sont plus, comme à la **section 4** des marqueurs bi-alléliques, mais des marqueurs comptant, disons, une dizaine d'allèles différents, choisis parce que l'allèle le plus fréquent est, pour chaque marqueur, porté par moins de $20\,\%$ de la population. Ces marqueurs sont aussi supposés indépendants, de sorte que la probabilité pour que l'assassin et vous [34] portiez le même génotype est inférieure à $0,20^{10} \approx 10^{-7}$.

Sauf bien sûr si l'assassin est un de vos proches parents, ou, pire, votre frère jumeau !

Le risque d'être atteint d'une maladie donnée dépend (pour les maladies ayant une dimension génétique) du génotype de l'individu examiné. Distinguons soigneusement deux situations éthiquement fort différentes :

Pour une maladie génétique grave se pose le problème du « conseil génétique » : calculer pour des parents ayant un enfant atteint (ou appartenant à une famille où la maladie est fréquente) le risque pour un futur enfant qu'il soit atteint. Les méthodes esquissées à la **section 3** et la **section 4**, généralisées à plus d'un gène ou plus d'un marqueur, permettent d'estimer pour l'éventuel futur enfant les probabilités des divers génotypes et, au travers de leurs pénétrances, la probabilité pour qu'il soit atteint. Quand cette probabilité dépasse un seuil trop élevé et que la maladie est très handicapante, la question de la procréation se pose (aux parents).

À l'inverse, les maladies en population (donc dans les familles qui ne regroupent pas déjà de nombreux cas), sont le plus souvent très polygéniques (des dizaines de gènes participent à leur étiologie) ; la démarche consistant à typer un (au maximum deux) marqueurs pour annoncer — moyennant substancielle finance « votre risque de MMM est 10 fois plus élevé que pour l'ensemble de la population » a deux conséquences :

- enrichir les aigrefins à qui vous avez envoyé (à l'étranger) votre ADN — sans maîtriser d'ailleurs l'usage qui en sera fait — et vos sous — idem ;
- vous paniquer inutilement, en vous cachant que le risque moyen dans la population est de 1 sur 10 millions, le vôtre de 1 sur 1 million, bien moindre que le risque d'un accident mortel sur l'autoroute Paris-Lyon.

La génétique quantitative, qui prédirait la taille future d'un bébé ou le rendement d'une variété de blé est beaucoup moins avancée que la génétique qualitative (souvenez vous, Mendel avait déjà de l'avance sur Galton) : faute d'avoir compris les mécanismes de l'étiologie quantitative, on en est encore aux modèles phénotypiques de la **section 2**, tout au plus à la sélection assistée par marqueur. Mais le temps où l'on saura répondre à ces questions sur la base des génomes n'est pas loin... pour le meilleur ou pour le pire.

Quant au calcul d'un médicament, il se fera aussi, le jour où l'on maîtrisera les calculs de chimie-physique quantique capables de rendre compte des interactions moléculaires (pourquoi telle position d'un gène ou d'une protéine est un *site de fixation* qui permet à une enzyme d'entrer dans une cascade de réactions, et quels paramètres faut-il changer et de combien pour faciliter ou pour empêcher cette réaction). On peut rêver que, alors, au lieu de cribbler des milliers de molécules pour chercher celle qui est active pour telle myopathie, on demandera à un logiciel complexe de calculer la formule chimique et la conformation spatiale du médicament

recherché.

Mais là aussi, il faudra beaucoup de modélisation, beaucoup de statistique — et beaucoup de géométrie. Bref, beaucoup de maths.

Annexe : le recuit simulé

Si, au lieu de chercher un maximum on cherchait un minimum — vous admettrez que mathématiquement « c'est la même chose » —, on aurait la figure 3 « à l'envers » — disons une cuve avec plusieurs trous, comme sur la figure ci-dessus. Si on lâche une bille au hasard, elle va après quelques allers-retours se stabiliser dans un minimum local, pas forcément dans le plus profond.

Si on secoue fortement la cuve, la bille va sauter de trou en trou. Si on la secoue de moins en moins fort, elle va finir par rester piégée dans un trou. Si on s'y prend bien, à un moment du processus on secouera assez pour qu'elle quitte les trous pas trop profonds, mais pas le plus grand. Elle finira au fond du fond.

Si on laisse se dérouler l'animation ci dessus c'est ce qui se passe : après quelques angoisses « cette fois-ci, c'est perdu », un sursaut ramène toujours la bille vers le trou le plus profond.

Bien sûr, quand on a une fonction d'une seule variable, comme ici, il est toujours possible (et

plus rapide) de tracer la courbe et pas besoin d'être grand mathématicien pour trouver son minimum. Mais quand on a une fonction de centaines ou de milliers de variables, il devient impossible d'explorer systématiquement l'espace des possibilités et les solutions fondées sur ces méthodes d'agitation de moins en moins violentes (on parle de *recuit simulé* en allusion à une agitation thermique que l'on ferait tendre vers zéro) se révèlent les plus efficaces.

Notes

[▲1] Je remercie Christophe et Manon Ambroise, Catherine Matias et Pierre Latouche, qui ont relu ces articles d'un œil critique et constructif.

[▲2] Au lycée nous rencontrons des équations en cinétique ($\frac{1}{2}\, g\, t^2$), en thermodynamique ($pV = nRT$) ou en optique ($sin(i) = n\, sin(r)$), et l'on sait que la physique moderne se fonde sur des mathématiques très complexes, que ce soit de la relativité d'Einstein à la physique quantique de Planck : équations de Schrödinger, théorie de Yang-Mills fondée sur des opérations non commutatives sur des espaces plaçant un univers derrière chaque point de l'espace-temps, géométrie inimaginable au sein des protons, etc.

[▲3] par exemple pour décrire des modèles proie-prédateur (plus il y a de lapins, plus il y a de renards pour les manger, plus il y a de renards, moins il y a de lapins, a-t-on équilibre ou cycle ?).

[▲4] voire un ensemble de caractères observables.

[▲5] X peut aussi être mesurable aujourd'hui (le nombre de cigarettes fumées par jour) et servir à prédire un Y futur (un risque de cancer, une durée de survie).

[▲6] Galton, né comme Mendel en 1822, a même cherché le support matériel de l'hérédité. Croyant qu'il était dans le sang (noblesse oblige), il transfusait des lapins pour transmettre « horizontalement » leurs caractères physiques — sans grand succès !

[▲7] Nous ne définirons pas les bases de Probabilités utilisées dans ce texte, une compréhension intuitive suffisant toujours.

Pour les spécialistes, disons que l'on supposera par exemple que ε est gaussien $\varepsilon \sim \mathcal{N}(0\,;\,\sigma^2)$; les ε relatifs aux différentes vaches sont supposés indépendants.

[▲8] Voir Prum *Le hasard* in *Le Notionnaire* de l' *Encyclopedia Britanica*, 2005.

[▲9] La chance de Mendel est de s'être intéressé à des caractères qualitatifs : ridé/non ridé pour les pois, rouge/blanc pour les fleurs, alors que Galton cherchait à expliquer des phénotypes quantitatifs, tels que la taille des nobles ou des lapins

[▲10] Il faudra attendre la physique quantique pour que l'aléatoire trouve un nouveau domaine d'applications intensives.

[▲11] L'aléatoire est sinon utilisé, comme aux équations (1) ou (2), pour modéliser une observation incomplète de la réalité, une erreur de mesure ou un choix aléatoire d'échantillon — songez aux sondages. Au lieu d'être — comme ici — partie prenante du phénomène réel, il est alors introduit par le modélisateur

[▲12] ou parfois les allèles portés en un ensemble de position.

[▲13] Défaut de coagulation sanguine, « touchant les garçons, transmis par les femmes ».

[▲14] « Chorée » vient, comme chorégraphie, de la danse — on parlait autrefois de « danse de Saint-Guy » — les malades effectuant des mouvements incontrôlés.

[▲15] qui provoque l'accumulation du mucus dans les voies respiratoires.

[▲16] Dégénérescence des fibres musculaires — y compris cardiaques.

[▲17] On connaît aujourd'hui des milliers de maladies génétiques — et il est de plus en plus clair que les susceptibilités aux cancers, au paludisme ou au Sida, par exemple, ont une dimension génétique.

[▲18] Citons la famille de rêve pour le généticien : les 273 habitants de l'île anglaise de Tristan da Cunha, dans l'Atlantique sud, tous parents de multiples façons, tous atteints de diverses maladies, asthme, glaucomes, etc., et dont les génomes ont été relevés lors d'une évacuation en urgence de l'île à l'occasion d'une éruption volcanique en 1961.

[▲19] Le Statisticien tient ce raisonnement : « Plus le choix de paramètre donne une grande probabilité à ce que l'on observe, plus on croit en ce choix ». N'entrons pas dans le débat philosophique que ceci peut susciter : des théorèmes montrent que c'est la bonne façon de procéder.

[▲20] ceux de nos cours de lycée, où l'on annulait des dérivées en résolvant des équations du premier degré.

[▲21] Stefen Lauritzen, le père de cette théorie ne manque pas d'humour, puisqu'il a proposé de lier ces parents par une arête et d'appeler le graphe obtenu le *graphe moral* du précédent, puisqu'on y avait uni tous les couples parentaux.

[▲22] par exemple : la vraisemblance n'a plus crû depuis 1000 itérations

[▲23] Définition haïssable pour le mathématicien, puisque les technologies évoluant à une vitesse stupéfiante, ce qui est très difficile et hors de prix un jour se fait en routine et pour quelques euros quatre ans plus tard.

[▲24] Une autre question est : « ce marqueur est-il *significativement* plus présent chez les malades que chez les bien portants ? » — nous ne la traiterons pas ici.

[▲25] Voir l'article **« À la recherche de mots de fréquence exceptionnelle dans les génomes »** de Sophie Schbath.

[▲26] Une bactérie porte typiquement un seul chromosome, de quelques millions de lettres. L'Homme porte deux jeux de chromosomes, l'un hérité de son père, l'autre de sa mère — on parle de *paires* de chromosomes ; chaque jeu comporte 23 chromosomes dont la longueur totale dépasse les 3 milliards de lettres — il y a donc 23 paires.

[▲27] comme en français : les mots *maison*, *raison* ou *saison* ont des sens fort éloignés. On parle alors de SNP (acronyme de l'anglais *Single Nucleotide Polymorphism*.)

[▲28] on peut supposer que les recombinaisons sur des segments disjoints sont indépendantes — au sens que les Probabilistes donnent à ce mot ; on a alors un *processus de Poisson*.

[▲29] en français le mot « germains » désigne des frères-ou-sœurs ; en anglais, on parle de *affected sib pairs*.

[▲30] Ils interviennent tous deux dans la réparation de l'ADN, qui en a souvent besoin, suite à des irradiations, des dommages chimiques, ou tout simplement des ruptures consécutives à des tensions excessives lors de ses déformations — destinées par exemple à rendre accessible un gène enfoui au cœur d'une pelote compacte.

Hall & al. (1990) Linkage of early onset familial breast and ovarian concer to chromosome 17q21, *Science* 250, p. 1684 – 1689.

Wooster R.& al. (1994) Localization of the susceptibility gene *brca2* to chromosome 13q12-13, *Science* 265, p. 2088 – 2090.

[▲31] ou même des séquences d'ADN — nous y reviendrons dans l'article suivant.

[▲32] *Jurassic Park* n'est crédible qu'au cinéma — tout au plus garde-t-on des traces d'ADN quelques milliers d'années : les tout derniers néanderthal, les mammouths...

[▲33] Petit exercice : montrer qu'au dessus de n espèces, il y a $1 \times 3 \times 5 \times 7 \ldots \times (2n-3)$ arbres possibles : 8 espèces 135 135 arbres, 12 espèces, près de 14 milliards d'arbres.

[▲34] On suppose — n'est-ce pas ? — que l'on parle de deux personnes différentes.

Crédits images

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