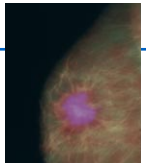
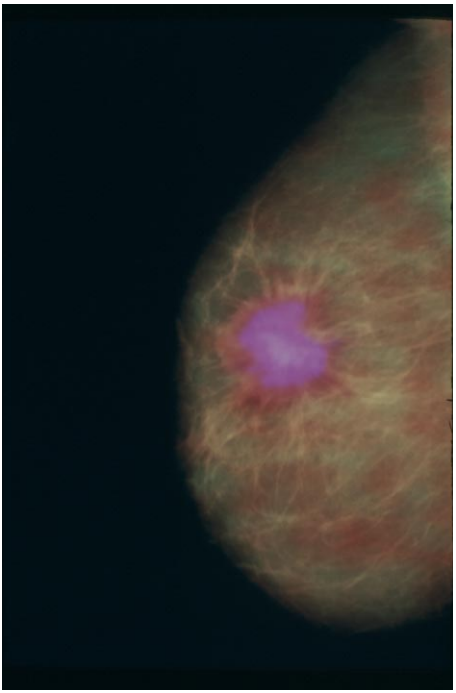
Trouver un gène responsable de cancer



Bernard Prum

Les développements de la biologie moderne, et notamment ceux de la génétique moléculaire, exigent de nouveaux outils mathématiques. Exemple avec la statistique et son rôle dans la recherche d'un gène lié au cancer du sein.

D'innombrables maladies ont une composante héréditaire : le risque d'être atteint est plus ou moins élevé chez un individu selon qu'il est porteur ou non d'un gène dit de *susceptibilité* à la maladie en question. C'est pourquoi la génétique d'aujourd'hui cherche à comprendre le rôle des différents gènes, et en particulier leur rôle dans l'étiologie des maladies — dans l'espoir de mettre au point un jour une thérapie. Prenons comme exemple le cancer du sein qui, en France, touche ou touchera environ une femme sur huit. À côté de divers facteurs de risque (alimentation, tabac, exposition aux radiations, etc.), on a identifié il y a quelques années un gène dont les mutations sont impliquées dans un pourcentage élevé de femmes atteintes d'un tel cancer. Ce gène a été baptisé BRCA1 (pour *breast cancer 1*). Un tel résultat, de nature biomédicale, n'a pu être obtenu que par une succession d'analyses statistiques qui, nous allons le voir, ont permis de localiser le gène de façon de plus en plus précise.



Dans cette mammographie en fausses couleurs, une tumeur cancéreuse est visible en rose. Une partie des recherches sur les cancers du sein sont consacrées à leur aspect génétique. La théorie des statistiques y joue un rôle capital. (Cliché Kings College School/SPL/Cosmos)

La génétique a longtemps ignoré la nature matérielle des gènes. Ce n'est que depuis une vingtaine d'années que l'on a accès massivement aux séquences d'ADN, la chaîne moléculaire qui matérialise l'information génétique transmise des parents aux enfants. Pour autant, l'ignorance de la composition chimique des gènes n'a nullement empêché d'obtenir des résultats fins sur l'hérédité de tel ou tel trait.

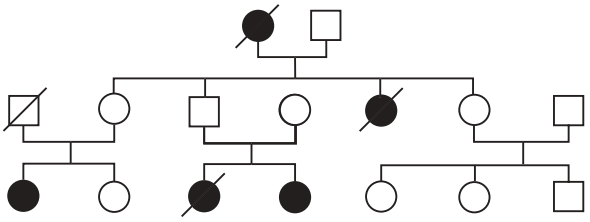


Figure 1. Une famille où l'on observe une concentration de cancers du sein. Les carrés indiquent les hommes, les cercles les femmes. Un individu est indiqué en noir s'il est atteint, barré s'il est décédé. On constate que la grand-mère, une de ses filles et trois de ses petites filles ont eu un cancer. Bien sûr, chez d'autres membres de la famille, la maladie peut encore se déclarer. C'est à partir de tels pedigrees que les généticiens sont conduits à supposer l'existence de gènes de susceptibilité à la maladie.

La première question que l'on se pose face à une maladie comme le cancer du sein est : « est-ce une maladie génétique, existe-t-il des gènes qui prédisposent à cette maladie ? ». Pour les cancers, la réponse a longtemps été incertaine. On s'attend à une réponse positive si l'on constate des concentrations familiales de la maladie, si l'on peut attribuer à la fille ou la sœur d'une femme atteinte un risque plus grand que celui encouru par l'ensemble de la population. Et pendant longtemps, le statisticien généticien a eu pour données de base des *pedigrees* comme celui de la figure 1.

Que faire d'un tel pedigree ? On sait, presque depuis Mendel, qu'un caractère héréditaire est souvent déterminé par un « gène » pouvant prendre plusieurs formes, appelées ses *allèles*. Chaque individu hérite un allèle de son père et un allèle de sa mère ; il transmet à chacun de ses enfants l'un de ces deux allèles au hasard. Le généticien propose alors, pour la transmission de la maladie étudiée, un modèle, qui suppose l'intervention de certains gènes et allèles. Ce modèle, le statisticien doit le valider à l'aide de tests statistiques appropriés, qui permettront par exemple d'éliminer les hypothèses les plus simples, comme : « la maladie étudiée n'a aucune composante génétique ».

Dans le cas de plus en plus étudié des maladies à étiologie complexe (cas du cancer du sein), où interviennent des facteurs d'environnement ou bien dont l'incidence dépend de l'âge, il convient de traiter des données qui dépendent du temps ; on doit alors faire appel à la *statistique des processus*. C'est une branche mathématique élaborée, qui s'appuie en grande partie sur les résultats obtenus par l'école française de probabilités des années 1980 (P. A. Meyer, J. Jacod) et ceux de statistique dus à l'école scandinave.

Des statistiques pour déterminer le chromosome porteur du gène

Une fois établie par l'analyse des pedigrees l'existence d'un gène de susceptibilité au cancer du sein, la seconde étape consiste à le localiser, au moins grossièrement, sur l'un des 23 chromosomes humains. Pour cela, on dispose depuis les années 1980 de *marqueurs* ; ce sont de petites chaînes d'ADN bien déterminées que l'on peut « lire » à moindre coût, disons par une analyse chimique rapide. Balises

relativement faciles à localiser, les marqueurs permettent par exemple d'évaluer la ressemblance entre des régions de chromosomes examinées chez des personnes malades et apparentées. Plus grande est la similitude d'une même région de chromosome chez des personnes apparentées atteintes, plus élevée est la probabilité que cette région porte un gène impliqué dans la maladie.

Mais une telle analyse, statistique bien sûr, est compliquée par le fait que chaque parent ne transmet pas à ses enfants les chromosomes qu'il a lui-même hérités de ses parents, mais une *recombinaison* de ceux-ci (figure 2). Si l'on considère deux gènes situés au départ sur un même chromosome, ils pourront après recombinaison se retrouver sur deux chromosomes différents ; la probabilité que cela arrive est d'autant plus élevée que les deux gènes en question sont éloignés. Analyser le taux de similarité le long d'un chromosome, c'est donc étudier un processus aléatoire. Grâce à la statistique des *processus*, on peut donc délimiter un intervalle dans lequel se trouve un gène de susceptibilité. L'emploi des marqueurs a ainsi permis à l'équipe américaine de Jeff M. Hall, à Berkeley, de localiser en 1990 le gène BRCA1 sur le chromosome 17.

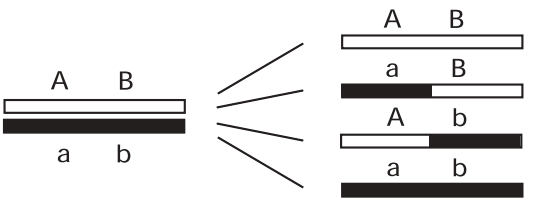


Figure 2. Pour chaque paire de chromosomes d'un individu, un chromosome est hérité de son père (en noir) et l'autre hérité de sa mère (en blanc). Un parent transmet à chaque descendant un seul chromosome de chaque paire. Mais avant la transmission, les chromosomes de chaque paire peuvent s'échanger des morceaux, au hasard. Ce processus dit de recombinaison fait que le parent transmet à son enfant un chromosome recombiné (l'une des quatre possibilités indiquées dans la figure, où l'on suppose que les chromosomes s'échangent deux régions).

Lire la molécule d'ADN pour décrire complètement le gène et ses formes anormales

Il s'agit ensuite de localiser précisément le gène et de déterminer sa structure. On sait que l'ADN, le matériau génétique, est une longue chaîne moléculaire « écrite » dans un alphabet de 4 « lettres » (*a*, *c*, *g* et *t*, initiales des quatre types de molécules dont est formée la chaîne d'ADN). Les banques de données génétiques répertorient plusieurs milliards de telles lettres (il en arrive quelque 25 millions par jour…).

La précision de la méthode des marqueurs permet au mieux de localiser un gène sur une séquence d'ADN comptant quelque 4 millions de lettres. Pour savoir exactement quel allèle, ou quelle mutation est responsable, par exemple, du cancer du sein, il faut « lire » ces séquences chez les sujets sains et malades pour les comparer. Cela revient à trouver une « faute de frappe » dans un texte de 4 millions de caractères, disons un livre de 2000 pages – ou plutôt dans autant de livres de 2000 pages que l'on a d'individus à étudier. Cette tâche est lourde, même avec des moyens informatiques puissants. Or chez l'homme, les gènes ne constituent pas plus de 3 % des chromosomes. Le reste du matériel chromosomique est qualifié d'*intergénique*. Si l'on parvient à limiter la recherche des fautes de frappe aux seuls gènes, on réduit la séquence à explorer à une trentaine de pages, ce qui devient accessible à tout ordinateur.

Mais comment distinguer les gènes

du reste ? Il s'avère que le « style » dans lequel sont écrits les gènes diffère du style intergénique : les fréquences de successions de lettres ne sont pas les mêmes. On peut chercher à exploiter cette différence de style pour *annoter* la séquence et distinguer les gènes de la partie intergénique. Le défi est ardu. On doit faire appel à des modèles statistiques appelés *chaînes de Markov cachées* et développés dans les années 1980, en liaison notamment avec des problèmes de reconnaissance automatique de la parole ; ils ont dû être adaptés à la génomique, en même temps que l'on mettait au point des algorithmes capables à la fois de caractériser les différents styles et d'attribuer un style à chaque position sur le chromosome.

C'est ainsi que l'on a fini par localiser précisément BRCA1. On peut désormais le lire facilement chez chaque malade. Ce gène de susceptibilité au cancer du sein compte 5 592 lettres et l'on en connaît plus de 80 allèles. Reste un nouveau travail pour le statisticien : établir les relations entre les divers allèles et la prévalence de ce cancer.

La biologie offre aux mathématiques un nouveau terrain d'action

L'exemple du gène BRCA1 le suggère, la biologie jouera probablement vis-à-vis des mathématiques le rôle détenu par la physique au cours d'une bonne partie du XX^e^ siècle : offrir un champ d'application aux outils théoriques récents et susciter l'élaboration de nouveaux outils (nous avons évoqué ici les outils statistiques, mais on pourrait évoquer d'autres domaines des mathématiques comme les systèmes dynamiques, l'optimisation, jusqu'à la géométrie — la conformation spatiale des molécules joue, on le sait, un rôle essentiel dans leur fonction). Un nouveau défi est aujourd'hui lancé au statisticien : on est actuellement capable de placer quelques milliers de réactifs sur une surface de verre d'un centimètre carré (les « puces ») et de savoir ainsi quels gènes travaillent dans quels tissus, dans quelles conditions expérimentales ou… dans quelles cellules cancéreuses. Les mesures effectuées en laboratoire, selon des centaines de conditions diverses, fournissent aux chercheurs un nombre considérable de données numériques, qui caractérisent l'expression de milliers de gènes. À ce jour, seules des analyses statistiques peuvent prétendre les traiter et préciser ainsi les liens entre gènes et maladies.

Bernard Prum
Laboratoire Statistique et Génome
(UMR CNRS 8071),
La Génopole, Université d'Évry

Quelques références :

- B. Prum, « Statistique et génétique » dans *Development of Mathematics 1950-2000* (sous la dir. de J.-P. Pier, Birkhäuser, 2000).
- C. Bonaïti-Pellié, F. Doyon et M. G. Lé, « Où en est l'épidémiologie du cancer en l'an 2001 », *Médecine-Science, 17,* pp. 586-595 (2001).
- F. Muri-Majoube et B. Prum, « Une approche statistique de l'analyse des génomes », *Gazette des mathématiciens, n° 89,* pp. 63-98 (juillet 2001).
- B. Prum, « La recherche automatique des gènes », *La Recherche, n° 346,* pp. 84-87 (2001).
- M. S. Waterman, *Introduction to computational biology* (Chapman & Hall, 1995).