Génétique statistique

Springer

Paris
Berlin
Heidelberg
New York
Hong Kong
London
Milan
Tokyo

Stephan Morgenthaler

Génétique statistique

Springer

Stephan Morgenthaler
EPFL FSB IMA
Station 8 - Bât. MA
CH-1015 Lausanne
Suisse
stephan.morgenthaler@epfl.ch

ISBN : 978-2-287-33910-3 Springer Paris Berlin Heidelberg New York

Maquette de couverture : Jean-François Montmarché

Collection
Statistique et probabilités appliquées
dirigée par Yadolah Dodge

Professeur Honoraire
Université de Neuchâtel
Suisse
yadolah.dodge@unine.ch

Comité éditorial :

Dans la même collection :

– *Statistique. La théorie et ses applications*
Michel Lejeune, avril 2004
– *Le choix bayésien. Principes et pratique*
Christian P. Robert, novembre 2005
– *Maîtriser l'aléatoire. Exercices résolus de probabilités et statistique*
Eva Cantoni, Philippe Huber, Elvezio Ronchetti, novembre 2006
– *Régression. Théorie et applications*
Pierre-André Cornillon, Éric Matzner-Løber, janvier 2007
– *Le raisonnement bayésien. Modélisation et inférence*
Éric Parent, Jacques Bernier, juillet 2007
– *Premiers pas en simulation*
Yadolah Dodge, Giuseppe Melfi, juin 2008

Avant-propos

Ce court recueil procède à une revue de diverses méthodes statistiques applicables à la génétique. Cette seconde science nous permet, mieux que nulle autre, de faire connaissance de la pensée probabiliste. Dans l'histoire de la statistique, la génétique a souvent été à l'origine d'idées nouvelles importantes. Nous livrons ici aux lecteurs dotés d'une formation mathématique quelques exemples tirés de cette discipline biologique dont les concepts sont définis au fur et à mesure de leur introduction. Aucune connaissance biologique préalable n'est donc nécessaire à la lecture de cet ouvrage.

Les lecteurs biologistes pourront eux aussi découvrir des modèles statistiques dans un contexte familier, mais il leur faudra posséder un certain niveau de connaissances mathématiques, ou faire preuve d'une réelle assiduité.

Les questions traitées dans les pages qui suivent constituent une sélection personnelle et ne prétendent pas à l'exhaustivité. Nous avons notamment laissé de côté l'analyse des données d'expressions géniques (« *microarray* »). De nombreux livres récents expliquent ce sujet de manière détaillée.

Cet ouvrage se fonde sur un cours de master (troisième ou quatrième année universitaire) que j'ai donné plusieurs fois à l'École polytechnique fédérale de Lausanne et à des étudiants en mathématiques, en informatique et en bioinformatique.

Les exercices à la fin des chapitres ont été élaborés par Andrei Zenide, Sandro Gsteiger, Sahar Hosseinian et Jean-Marc Nicoletti.

Lausanne, le 15 avril 2008
Stephan Morgenthaler

Sommaire

Chapitre 1

Introduction

La génétique est la science de la transmission des caractères héréditaires dans des populations d'êtres vivants. Elle occupe une place centrale au sein des sciences biologiques.

Les faits suivants représentent des points marquants dans le développement de la génétique : la publication de l'ouvrage de Ch. Darwin (*On the origin of species by means of natural selection*, London, John Murray, 1859), celle de l'article de G. Mendel intitulé *Versuche über Pflanzen-Hybriden* (1865), l'extraction d'ADN (acide désoxyribonucléique) de globules blancs (J.F. Miescher, 1869), l'observation du comportement des chromosomes lors de la division cellulaire par Th. Boveri (1888), la découverte portant sur le fait que les facteurs de Mendel sont liés physiquement aux chromosomes (Th. Boveri et W. Sutton, 1902), la découverte démontrant que la structure chimique de l'ADN pourrait en faire une substance porteuse de l'information génétique (F.H.C. Crick et J.D. Watson, 1953), le séquençage de la totalité du génome humain par une association internationale de chercheurs (*Nature* et *Science*, février 2001, voir aussi www.ornl.gov/sci/techresources/Human_Genome/home.shtml).

L'intérêt pour la génétique humaine est aujourd'hui extrêmement vif et les sciences du vivant sont perçues comme le moteur du développement futur de nos sociétés. Le fonctionnement de tout organisme vivant est fondé sur les gènes. Grâce à la collaboration entre gènes, il existe une richesse incroyable de propriétés et de fonctions. Une compréhension approfondie des propriétés des gènes est indispensable si nous souhaitons guérir les organismes des maladies, les protéger de dangers environnementaux, diagnostiquer des dysfonctionnements, etc.

Bien que certains caractères tels que le groupe sanguin soient déterminés par des facteurs purement génétiques, d'autres ne le sont que partiellement ou même pas du tout. Même si deux individus sont génétiquement identiques, ils ne le sont pas dans leurs comportements sociaux, leurs intérêts culturels, et même au niveau de leurs physiologies.

La diversité génétique entre humains n'est, dans un certain sens, pas très

importante. Nos génomes sont identiques à 99,9 %. Et pourtant, la statistique s'intéresse avant tout aux différences. Elle essaie de comprendre l'origine de la différence entre les individus ainsi que son impact.

1.1 Données génétiques

Les données issues d'une étude génétique sont très variées. Les caractères que l'on observe sur un individu tels que sa taille, la couleur des yeux ou la présence d'une maladie sont des variables *phénotypiques*, tandis que l'information interne et héritable d'une cellule est *génotypique*. Les variables biochimiques telles que la concentration d'une protéine dans le sang, la présence d'une mutation sur l'ADN ou la concentration de microorganismes dans un échantillon d'eau sont des *biomarqueurs*. La liste suivante donne quelques exemples de variables ou biomarqueurs qui peuvent se présenter dans une étude :

- un caractère complexe, tel que la production laitière d'une vache ;
- un biomarqueur simple, tel que le groupe sanguin ;
- le génotype par rapport à un groupe de gènes, c'est-à-dire les allèles dont un individu est porteur ;
- le taux d'activité d'un ou de plusieurs gènes, mesuré dans un échantillon de tissus provenant d'un organe ;
- une séquence d'ADN ;
- les relations familiales d'un ensemble d'organismes.

Les mesures sont effectuées parfois au moyen de cultures de cellules (*in vitro*) et parfois avec des cellules prises sur des individus (*in vivo*). Dans le second cas, les individus peuvent former un échantillon sélectionné au hasard parmi une population. Dans d'autres situations, il s'agit d'individus ayant des relations familiales et une généalogie connue.

Parmi les objectifs de l'analyse statistique des données génétiques, on trouve les suivants :

- trouver des associations entre phénotypes et génotypes, par exemple, des facteurs de risque génétiques ;
- déterminer l'arrangement d'un ensemble de gènes sur un chromosome (« *physical mapping* » en anglais) ;
- élucider la liaison évolutive entre espèces ;
- identifier les dispositions génétiques sources de maladies ;
- déterminer la fonction d'un gène dans les processus cellulaires ;
- modéliser le processus à l'origine des mutations ;
- décrire l'interaction entre gènes.

Les données et les questions étant très variées, les méthodes statistiques utilisées dans l'analyse de telles données le sont aussi. La génétique a souvent été à l'origine de nouvelles méthodes statistiques. Ce petit livre en détaillera quelques-unes.

1.1.1 Expérience de Mendel

Pour analyser de manière scientifique la transmission de phénotypes d'une génération à l'autre, G. Mendel a effectué des expériences avec des plantes *pisum sativum* (petit pois). Les phénotypes qu'il choisissait étaient, entre autres, l'apparence (lisse ou ridée) et la couleur (jaune ou verte) des graines. En croisant de multiples fois des plantes qui produisaient des graines lisses ou ridées, il a, par selection, produit des plantes pure-souche du type « *lisse* » et « *ridée* ». Ces plantes formaient la génération parentale P_1 de l'expérience génétique de Mendel. Il a ensuite créé des plantes hybrides en croisant une plante lisse avec une plante ridée. Ces hybrides sont les descendants F_1, la première génération filiale. Mendel a observé que leurs graines étaient toutes lisses. En 1865, la théorie génétique contemporaine affirmait que, dans la fécondation, les caractères se mélangeaient. Interprétée de manière naïve, cette théorie était en contradiction avec les résultats de Mendel, car les plantes F_1 étaient d'un seul et unique type.

Mendel souhaitait voir plus clair et a poursuivi ses expériences en croisant les plantes de la population F_1. En faisant ainsi, on obtient la génération F_2 et à ce stade, les deux types parentaux, lisse et ridée, réapparaissent. En chiffres, la génération F_2 a produit 5 474 graines lisses et 1 850 graines ridées, ce qui correspond au rapport de cotes de 74,74 % : 25,26 % ou bien $\frac{3}{4} : \frac{1}{4}$.

Pour modéliser cette expérience, nommons A le facteur qui cause le caractère « *graines lisses* » et a le facteur qui cause le caractère « *graines ridées* ». Pour évaluer dans quelle proportion les facteurs a et A étaient représentés dans les plantes F_2, Mendel a pratiqué des autofécondations. Les plantes F_2 étant munies du caractère « *graines ridées* », les descendants possédaient dans tous les cas ce même caractère, ce qui démontrait que ces plantes ne contenaient pas le facteur A. L'autofécondation de plantes F_2 de caractère a montré un autre résultat. Parfois, tous les descendants possédaient le caractère « *graines lisses* » et, parfois, ils étaient des deux types. Parmi ses plantes à caractère « *graines lisses* » de la génération F_2, Mendel a observé 193 hybrides pure-souche A et 372 hybrides mixtes A et a. Cela correspond au rapport 34,16 % : 65,84 % ou bien $\frac{1}{3} : \frac{2}{3}$. Parce que $\frac{3}{4}$ des plantes F_2 avaient le caractère « *graines lisses* », ce résultat montre que $\frac{1}{4}$ des plantes F_2 étaient pure-souche a et $\frac{2}{4}$ étaient des hybrides mixtes.

Les conclusions de G. Mendel étaient les suivantes. Premièrement, trois types de plantes existent dans la génération F_2, pure-souche A, pure-souche a et Aa mixte. Parce que les descendants des plantes mixtes peuvent être aussi bien A que a, elles doivent être porteuses des deux facteurs a et A. Dans un souci de cohérence, il faut postuler que les plantes pure-souche contiennent également deux copies des facteurs, mais deux fois le même, AA ou aa. Deuxièmement, les trois types de plantes étaient présents en proportions presque exactement égales à $\frac{1}{4} : \frac{1}{2} : \frac{1}{4}$. Si l'on suppose que les deux facteurs d'une plante peuvent se séparer durant la formation d'ovules et de pollens, on obtient le schéma de la figure 1.1. On constate que les plantes de la génération F_1 sont toutes du type mixte Aa. Leurs descendants sont avec probabilité $\frac{1}{4}$ du type AA, avec

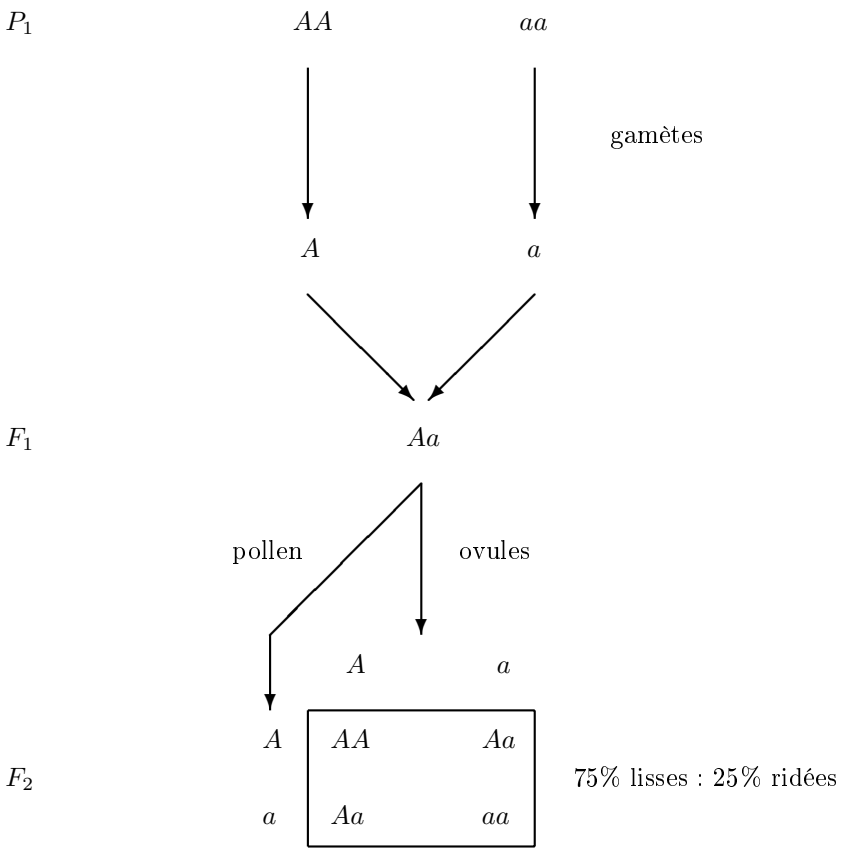

Figure 1.1 – Ce schéma décrit les expériences génétique de G. Mendel et fournit en même temps une explication des résultats.

probabilité également $\frac{1}{4}$ du type aa et avec probabilité $\frac{1}{2}$ du type Aa. Ces chiffres expliquent à merveille les observations de G. Mendel.

Exemple 1.1 *Mendel a également pratiqué des expériences avec deux caractères. D'un côté, l'apparence des graines et, de l'autre côté, leur couleur. En croisant une plante à graines lisses et jaunes avec une plante à graines ridées et vertes, il a constaté que les plantes de la génération F_1 sont des plantes à graines lisses et jaunes. En effectuant des autofécondations de telles plantes F_1, Mendel a obtenu 315 plantes à graines lisses et jaunes, 108 à graines lisses et vertes, 101 à graines ridées et jaunes et 32 à graines ridées et jaunes. Comment expliquer ces chiffres ?*

1.1.2 Test de Pearson

La méthodologie developpée par K. Pearson pour tester si une classification de n objets dans k types peut être expliquée par une répartition théorique est liée aux données de Mendel. Les expériences de Mendel ont résulté en une classification de $n = 565 = 193 + 372$ plantes dans deux classes qui ont des probabilités théorique de $\frac{1}{3}$ et $\frac{2}{3}$. Pour tester si les données sont en accord avec la théorie, K. Pearson a proposé la statistique du khi-deux

$$S = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}, \tag{1.1}$$

ou O_i est le nombre d'objets du i$^{\text{e}}$ type et $E_i = np_i$ le nombre espéré sous la théorie. La statistique S est une variable aléatoire. Si l'hypothèse nulle

$$H_0 : \text{probabilité de l'i}^{e} \text{ classe} = p_i \quad (i = 1, 2, \ldots, k)$$

est vraie, le résultat de la classification observée se situe dans un intervalle raisonnable autour de la classification théorique. Sous cette condition, la loi de S est approximativement une loi khi-deux avec $k - 1$ degrés de liberté. C'est ce qu'on appelle la *loi nulle* de ce test.

Dans l'exemple, on trouve

$$S = \frac{(193-565/3)^2}{565/3} + \frac{(372-2\times565/3)^2}{1130/3} = 0,173.$$

Cette valeur correspond au quantile $0,322$ de la loi khi-deux avec un seul degré de liberté, χ^2_1. Si on suppose que la répartition théorique soit la vraie répartition, l'événement $S = 0,173$ n'est donc pas du tout surprenant et montre que l'accord entre les données et la théorie de Mendel est tout à fait satisfaisant.

Si la théorie est fausse, la valeur de S devient grande car O_i et E_i peuvent être assez différents. On dit qu'une valeur de S est *significative*, si

$$\text{p-valeur} = P(X > S) < 0,05,$$

ou $X \sim \chi^2_{k-1}$ suit la loi nulle. Cela se produit lorsque S est loin dans la queue de la distribution χ^2_{k-1}.

1.1.3 Gènes, allèles, phénotypes et génotypes

Mendel appelait les causes génétiques des facteurs. Aujourd'hui, on les appelle *gènes*. Les caractères que Mendel choisissait sont appelés des *phénotypes*. Les copies des facteurs sont les *allèles*. Le mot allèle est utilisé pour indiquer deux choses. D'une part, un allèle est tout simplement une copie d'un gène. Ainsi, chaque individu est porteur de deux allèles chacun de nos deux parents nous a transmis un gène. D'autre part, le mot allèle signifie une variante d'un gène. Si j'ai le groupe sanguin O, par exemple, je sais que mes deux allèles

du gène ABO sont deux fois de la variante O. Deux allèles ne sont donc pas forcément égaux et si l'on a deux allèles différents d'un gène, on les note par exemple A et a ou A_1 et A_2, etc.

Les *gamètes* sont le véhicule de la transmission du génome de la génération parentale aux descendants. Les gamètes ont une seule copie du matériel génétique, ils sont dits *haploïdes*. Un individu est créé par la fusion de deux gamètes et chaque cellule (sauf les gamètes) contient donc deux copies de matériel génétique. Une cellule normale avec deux copies est appelée *diploïde*. La combinaison des deux variantes d'un gène que le descendant reçoit de ses parents est appelée son *génotype*. Le génotype d'un individu, pour un gène à deux variantes A et a, peut donc être soit AA, soit Aa, soit aa. Les deux types purs AA et aa sont dits *homozygotes*, l'autre étant dit *hétérozygote*. Par chance et par intuition, G. Mendel a choisi un gène dont le génotype a un effet immédiat et visible sur la plante adulte. L'apparence des graines est liée au génotype comme décrit au tableau suivant :

génotype	phénotype
aa	ridé
Aa, AA	lisse

Parce que Aa est lisse, même si une copie du gène a est présent, on dit que l'allèle a est *récessif*, tandis que l'allèle A est *dominant*.

1.2 Modèles stochastiques

La modélisation génétique fait appel de manière très naturelle à des processus aléatoires, car la sélection de deux gamètes avant leur union semble être une aventure pleine d'aléas. L'aléatoire joue un grand rôle tout d'abord dans la sélection des deux parents, ensuite – comme nous allons le voir plus tard – dans les détails de la construction des gamètes et, enfin, dans la vie quotidienne du nouvel être. Le modèle fondamental utilisé dans ce contexte est une simplification de la réalité, mais il est déjà assez riche.

Exemple 1.2 *Imaginons une population de taille constante, comprenant N individus dont les générations ne se chevauchent pas, donc ayant un rythme de vie parfaitement cyclique tel que les plantes annuelles. Les gamètes produits par les individus d'une génération s'unissent de manière complètement aléatoire pour créer les individus de la prochaine génération. On peut décrire ce processus par un schéma d'urne (fig. 1.2). L'urne contient tous les allèles d'une génération, donc un total de $2N$ boules. La prochaine génération est créée en tirant avec remise $2N$ fois dans cette urne. Le processus stochastique qui en résulte est dit le processus de Wright-Fisher.*

D'autres effets naturels, à part le mélange des génotypes, ont un caractère aléatoire, par exemple l'influence de l'environnement sur un individu et une

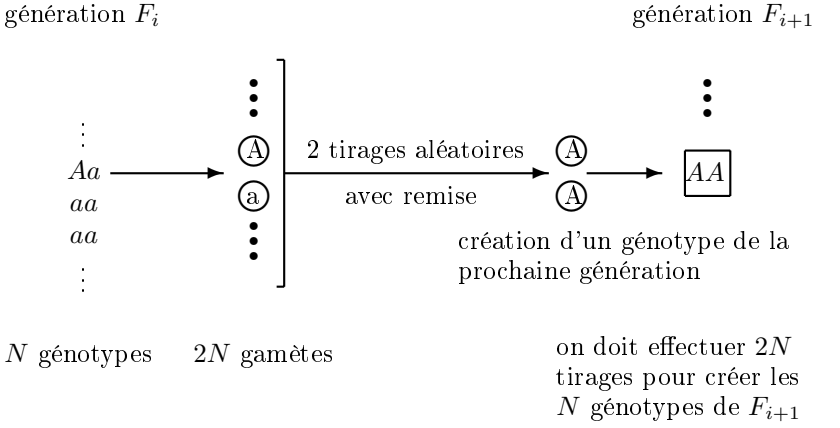

Figure 1.2 – Chaque individu de la présente génération est représenté par les deux gamètes qu'il peut produire. Les individus de la génération suivante ont un génotype créé par tirage aléatoire parmi tous ces gamètes. Le schéma ci-dessus montre la création d'un individu à génotype AA lors du passage entre générations i et $i+1$. Ce modèle simplifié de reproduction est nommé d'après S. Wright et R. A. Fisher.

population. Du fait de telles influences, des mutations se produisent dans le génome d'un individu. De telles mutations peuvent être bénéfiques en protégeant l'individu, dommageables en produisant des maladies, ou bien neutres. Modéliser l'émergence de mutations dans un individu et leur répartition et survie dans une population fait donc appel à des processus stochastiques.

Ces processus mutationnels peuvent également influencer la vie d'un individu. Dans le deuxième chapitre, nous étudierons le développement de tumeurs. Presque 90 % des patients qui souffrent d'une tumeur des poumons ont fumé. Mais seulement à peu près 10 % des fumeurs développent un tel cancer. Une explication de ces chiffres consiste à postuler un effet aléatoire assez important dans le développement de cette maladie.

1.3 Exercices

1. Dans son travail publié en 1865, Gregor Mendel a étudie la ségrégation de deux traits héréditaires de pois : la couleur (A jaune, a vert) et la forme (B lisse, b ridé). Ces génotypes différents donnent lieu à des phénotypes différents. Il a croisé le génotype AA, BB avec aa, bb, ce qui donnait une progéniture F_1 constituée uniquement de hétérozygotes dans les deux loci. En croisant la génération F_1 avec elle-même, il a obtenu pour la génération F_2 les fréquences suivantes :