JOURNAL DE LA SOCIÉTÉ STATISTIQUE DE PARIS

ALBERT JACQUARD

Génétique des populations et raisonnement probabiliste

Journal de la société statistique de Paris, tome 111 (1970), p. 223-229

<http://www.numdam.org/item?id=JSFS_1970__111__223_0>

GÉNÉTIQUE DES POPULATIONS ET RAISONNEMENT PROBABILISTE

Le titre de ce court exposé aurait pu avantageusement être remplacé par « Pascal et Mendel »; notre propos est en effet de montrer comment les originalités de pensées de l'un et de l'autre se complètent et permettent le développement de la « génétique des populations ».

I. — Pascal et Mendel

Les apports de Pascal, fondant dans ses lettres à Fermat du 29 juillet et du 24 août 1654 le raisonnement probabiliste, et de Mendel, apportant dans sa communication du 8 février 1865 à la Société d'Histoire naturelle de Brno l'explication logique de la reproduction sexuée, ont ceci de commun :

— qu'ils permettent de résoudre un paradoxe, donc d'échapper à une impasse dans laquelle la pensée scientifique s'était fourvoyée;

— et qu'ils sont contraire au « bon sens ».

L'objectif de Pascal, avec sa « règle des partis » est de montrer qu'il est possible de discourir avec rigueur à propos d'objets qui non seulement sont mal connus, mais qui peut-être n'auront jamais d'existence. S'affranchissant de la règle cartésienne qui fait avancer le raisonnement pas à pas, à partir de résultats certains, vérifiés par l'expérience, il met en évidence les propriétés d'un objet inexistant : la fin d'une partie de « pile ou face » définitivement interrompue.

Le calcul des probabilités ainsi créé permet d'échapper au paradoxe du comportement dans l'incertain : toute décision est prise, en pratique, dans une situation d'incertitude; elle ne peut donc être motivée par un raisonnement respectant les règles cartésiennes, et pourtant il faut décider.

En affectant des probabilités aux divers possibles, un lien entre le raisonnement logique et le comportement peut être retrouvé grâce à la prise en compte du « hasard » dans la décision.

L'hypothèse de Mendel, permettant d'expliquer ses observations sur la transmission des caractères héréditaires chez le pois, est fondamentalement contraire au « bon sens »; bien plus, elle est contraire au langage usuel : elle suppose que l'« individu », cet être « indivisible », ne transmet à chaque descendant que la moitié de son propre patrimoine génétique, que les deux moitiés ainsi reçues par chaque être vivant coexistent sans se mélanger et qu'elles se séparent à nouveau, au hasard, lors de la procréation suivante. Une conception aussi quantique de la réalité biologique ne pouvait être comprise au milieu du XIX^{e} siècle, ce qui explique en grande partie le peu d'audience du modèle mendélien jusqu'au début du XX^{e} siècle.

Cette hypothèse permettait cependant d'échapper au paradoxe auquel mène la théorie « naturelle » de l'hérédité continue, mélangée : si chez un enfant la mesure de tel caractère a pour espérance la moyenne arithmétique de sa mesure chez son père et chez sa mère, la variance de ce caractère dans la population doit diminuer de moitié à chaque géné-

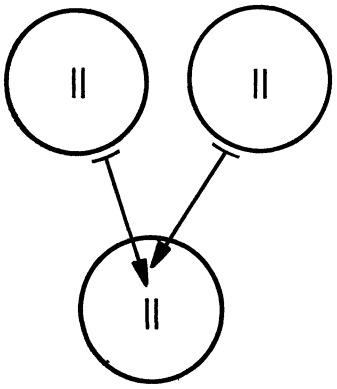
ration et donc tendre rapidement vers zéro. En effet si (avec des notations évidentes) : $x_f = \frac{1}{2}(x_p + x_m)$, on a :

$$V_F = \frac{1}{4}(V_P + V_m), \text{ soit } V_g = \frac{1}{2} V_{g-1}$$

Cette conclusion étant de toute évidence erronée, une autre explication est nécessaire; on a pu montrer que, sans recours à aucune expérience, le modèle mendélien aurait pu être imaginé en recherchant le modèle le plus simple permettant d'expliquer le maintien de la variabilité des caractères.

II. — Le hasard dans la transmission héréditaire

En quoi consiste, fondamentalement l'explication mendélienne de l'hérédité? : chez tout individu chaque caractère élémentaire est gouverné par deux facteurs, les « gènes » reçus l'un du père, l'autre de la mère. Ces gènes coexistent chez chaque individu, l'action de l'un peut masquer celle de l'autre, mais ils demeurent inaltérés, insécables, sans échanges possibles entre eux.



Lorsqu'à son tour cet individu procrée, il transmet pour chaque caractère, la copie de l'un des deux gènes qu'il possède, le choix du gène copié étant réalisé « au hasard ».

Cette intervention du hasard est sans doute le caractère essentiel de la reproduction sexuée; elle permet, à chaque génération, l'insertion d'un facteur aléatoire, d'une indétermination.

Si nous connaissons parfaitement les gènes de P et de M, disons (AB) et (CD), nous ne connaissons qu'en probabilité les divers génotypes possibles de leur fils F : (AC), (AD), (BC) ou (BD). A la connaissance totale a succédé l'ignorance partielle.

Dans ces conditions il paraît normal que la génétique fasse appel aux formes de raisonnement permettant de traiter des objets mal connus ou des processus non déterministes, c'est-à-dire au raisonnement probabiliste. Étrangement les premiers travaux de génétique des populations ont utilisé presque exclusivement les méthodes statistiques en mettant au premier plan le recours aux concepts de variance ou de corrélation. Ce n'est que depuis 1948 que, dans la direction marquée par le professeur Gustave Malecot, les recherches utilisant les méthodes probabilistes se sont développées.

Si le hasard intervient ainsi au niveau de l'individu, son rôle est plus large encore lorsque l'on considère une population dans son ensemble.

L'héritage génétique collectif transmis de génération en génération est plus ou moins modifié à la longue selon les conditions dans lesquelles s'opère cette transmission. Si tel gène est favorisé, ou est transmis par hasard à de nombreux enfants, si les individus qui les portent sont plus féconds ou moins soumis aux risques de mortalité, il se répand progressivement dans la population aux dépends de tel autre gène qui, peu à peu, disparaît. D'une génération à l'autre le patrimoine essentiel qu'est pour un groupe humain la collection de gènes dont il est porteur, peut ainsi se modifier profondément.

Le but de la « génétique des populations » est d'étudier la liaison entre ces modifications et les divers facteurs qui peuvent intervenir : système de mariages, migrations, taux de natalité ou de mortalité, etc.

Son objet est la « structure génétique » d'une population définie comme l'ensemble des fréquences ou des probabilités de présence soit des divers gènes gouvernant tel caractère, soit des divers génotypes, associations deux à deux de ces gènes chez les individus qui composent la population. Selon les problèmes, on devra donc étudier l'évolution :

— soit de la « structure génique réelle », ensemble des fréquences des gènes;
— soit de la « structure génique en probabilité », ensemble des probabilités des divers gènes ;
— soit de la « structure génotypique réelle », ensemble des fréquences des divers génotypes;
— soit enfin de la « structure génotypique en probabilité », ensemble des propabilités des divers génotypes.

Remarquons que lorsqu'un caractère est gouverné par des gènes de n types, les structures géniques comportent n éléments, et les structures génotypiques $\frac{n(n+1)}{2}$ éléments.

Nous nous bornerons ici à évoquer deux problèmes mettant en cause ces structures : la mesure de l'apparentement entre deux individus, la dérive des populations de faible effectif.

III. — La liaison entre les apparentés

La ressemblance entre individus apparentés a été sans doute le point de départ de la réflexion génétique : l'observation la plus superficielle montre que les enfants « ressemblent » à leurs parents ou parfois plus encore à un ascendant lointain, que les membres d'une même fratrie se « ressemblent », mais que le degré de ressemblance est variable selon les caractères considérés.

La transmission simultanée des gènes par les deux parents est évidemment la cause fondamentale de cette ressemblance; cette dualité des rôles du père et de la mère n'est cependant admise que depuis un siècle et demi : la querelle entre « ovistes » prétendant que seul l'ovule féminin est à l'origine de l'enfant, et « spermatistes » partisans de la thèse opposée n'a été clause qu'aux début du XIX^e siècle (1).

La mesure de la liaison entre individus apparentés n'a pas seulement pour but la satisfaction d'une curiosité désintéressée concernant la part de l'hérédité dans la manifestation d'un caractère, elle a aussi un objectif médical en constituant un élément de diagnostic.

En utilisant la terminologie actuelle, on peut définir un diagnostic comme un « traitement d'information ». Cette information provient certes pour l'essentiel de mesures faites sur l'intéressé (tension, rythme cardiaque, température...), mais elle peut provenir également d'individus qui lui sont apparentés. Notre objectif est de répondre à la question : « Quelle information apporte au sujet d'un individu A la connaissance (certaine ou en probabilité) de telle caractéristique de l'individu B, ayant avec A des liens de parenté connus? »

L'existence de liens de parenté entre deux individus correspond au fait que parmi les ascendants de l'un, figurent un ou plusieurs ascendants de l'autre, ou l'autre lui-même.

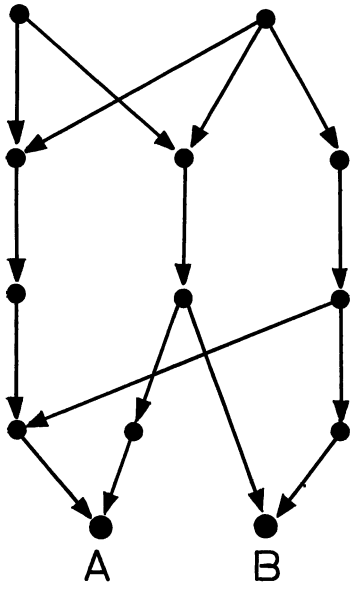
1. Ainsi, en latin, l'adjectif « consanguineus » ne concerne que les liens avec les apparentés de la lignée paternelle.

Désignons par « réseau d'ascendance » d'un individu, l'ensemble ordonné de ses ascendants, l'ordre étant celui des rapports de filiations; deux individus sont « apparentés » lorsque leurs deux réseaux d'ascendance ont une partie commune.

La complexité et la diversité des liens de parenté possibles sont telles qu'une classification et, plus encore, une mesure paraissent à première vue hors de portée; mais le raisonnement peut progresser de façon parfaitement naturelle et rigoureuse, si l'on a recours au concept de l'« identité » des gènes.

Mesure de l'apparentement

Quelles que soient la nature et la complexité des liens de parenté entre deux individus A et B, l'effet de ces liens sur leurs constitutions génétiques résulte de la possibilité pour les gènes de A comme pour les gènes de B d'être la réplique d'un même gène d'un de leurs ancêtres communs.



Deux gènes sont dits « identiques » s'ils proviennent, par duplications successives de génération en génération, d'un même gène ancêtre. Il en résulte que deux gènes identiques sont obligatoirement homologues, c'est-à-dire représentent le même allèle, puisque nous n'envisageons pas ici la possibilité de mutations.

Supposons connu le génotype de A, les informations que nous possédons concernant son parent B sont :

— le fait que B appartient à une certaine population caractérisée par sa structure génotypique;
— les liens parentaux que B a avec A, liens qui rendent plus ou moins probable l'identité de chacun des gènes de B avec chacun des gènes de A.

La liaison entre A et B peut ainsi être décrite, du point de vue génétique, par les probabilités d'identité de leurs gènes.

En un locus donné, (c'est-à-dire pour un caractère élémentaire donné) deux individus A et B possèdent au total quatre gènes : G_A transmis à A par son père, G_A^* transmis à A par sa mère, G_B et G_B^* (dans cette étude nous n'envisageons que les gènes autosomaux non portés par les chromosomes sexuels). Ces gènes peuvent, en fonction de la configuration des réseaux d'ascendance, être identiques ou être non identiques deux à deux.

Mais, dans le cas où, par exemple, G_A est identique à G_A^* et où G_A^* est identique à G_B^*, ce dernier est obligatoirement identique à G_A; autrement dit la relation d'identité est transitive : deux gènes identiques à un troisième sont identiques entre eux.

Compte tenu de cette transitivité on constate que 15 cas sont possibles pour le regroupement des 4 gènes de A et B en ensembles de gènes identiques. Ces 15 cas sont des « situations d'identité » auxquelles il est possible, en analysant le réseau d'ascendance commun de A et de B d'attribuer des probabilités; ces 15 probabilités constituent l'ensemble des « coefficients d'identité » de A et B; ils représentent une description résumée des liens parentaux de A et B permettant de répondre à toutes questions concernant leur liaison génétique.

Sans doute est-il déjà merveilleux de pouvoir décrire, sans perdre d'information génétique, un réseau d'ascendance de quelque complexité que ce soit, au moyen de 15 nombres (ou plus précisément de 14 nombres puisque la somme des 15 coefficients d'identité est égale à 1.) Mais le plus souvent on souhaite une mesure plus simple, plus facilement manipulable. Si l'on consent à perdre un peu d'information, on peut ramener cette mesure à un seul nombre le « coefficient de parenté », avec la définition :

« Le coefficient de parenté φ_{AB} de deux individus A et B est la probabilité pour qu'un gène pris au hasard chez A soit identique à un gène pris au hasard au même locus chez B ».

Liaison entre les structures génétiques des apparentés

Nous avons vu que la structure génétique d'un individu pouvait être définie soit en fonction de la connaissance (certaine ou en probabilité) d'un gène de son génôme, soit en fonction de la connaissance des deux gènes de son génôme; dans le premier cas, il s'agit de sa « structure génique » s, vecteur des probabilités de présence, dans les gamètes émis, des divers allèles possibles au locus considéré; dans le second cas, il s'agit de sa « structure génotypique » S, ensemble des probabilités des diverses associations deux à deux de ces allèles.

Le problème posé est maintenant le suivant : « Si l'on connaît la structure génique s_A, ou génotypique S_A, d'un individu A, quelle est la structure s_B ou S_B d'un individu B, lié à A par les liaisons parentales connues et appartenant à une population panmictique de structure génique connue? »

Si A et B sont apparentés, la présence chez A d'un certain allèle accroît la probabilité de présence chez B de ce même allèle puisqu'ils peuvent l'avoir reçu l'un et l'autre d'un ancêtre commun. Les structures géniques en probabilité de deux individus apparentés ne sont donc pas indépendantes : une information au sujet de l'un constitue une information au sujet de l'autre.

On peut montrer que, dans le cas où le père et la mère de A ne sont pas apparentés, la structure génique en probabilité de B connaissant celle de A est donnée par :

$$s_{B/A} = 2\varphi\, s_A + (1 - 2\varphi)s$$

où, φ est le coefficient de parenté de A et B, et s la structure génique de la population.

Les relations concernant la structure génotypique de B sont naturellement plus complexes; on peut montrer que cette structure est donnée par une relation où les divers coefficients d'identité interviennent de façon linéaire. Malgré leur aspect rébarbatif ces relations sont d'une utilisation facile et peuvent permettre de résoudre des problèmes tel que celui-ci :

« Un enfant à naître a un double cousin atteint d'une tare héréditaire liée à un gène dont la fréquence dans la population est 1 %. Avec quelle probabilité cet enfant sera-t-il taré? Avec quelle probabilité sera-t-il porteur du gène de cette tare »?

Si l'on ignorait l'existence du double cousin ces probabilités seraient respectivement de 0,01 % et 1,98 %; l'information concernant ce double cousin les porte respectivement à 7,0 % et 38,9 % soit 174 et 19 fois plus.

On voit par cet exemple l'intérêt de ces raisonnements typiquement probabilistes dans le domaine médical.

Les relations établies entre les génotypes d'apparentés permettent d'utiliser une information concernant un individu pour en déduire une information concernant un autre individu; elles rendent par conséquent possible un élargissement du champ du diagnostic.

IV. — Dérive aléatoire des petites populations

Considérons une population comportant à chaque génération N individus. Chaque fois qu'un gène est transmis de la génération g à la génération $g + 1$, la probabilité pour qu'il représente un certain allèle a est égale à la fréquence p_g de cet allèle dans la génération g. Par conséquent le nombre de gènes a présents dans la génération $g + 1$ est une variable aléatoire dont la loi de répartition est une loi binôminale de paramètres p_g et $2\,N$. Ce nombre peut donc avoir toutes les valeurs depuis 0 jusqu'à $2\,N$, les valeurs extrêmes ayant une faible probabilité si l'effectif N est grand.

Ainsi à chaque passage d'une génération à la suivante se produit une variation purement aléatoire, imprévisible, « sans cause », de la fréquence de chaque gène; le risque que cette variation soit importante est d'autant plus grand que l'effectif de la population est plus limité.

Ce processus de variation aléatoire se poursuit sans qu'une force quelconque se manifeste pour ramener la fréquence vers sa valeur initiale. Si celle-ci est passée par exemple d'une valeur p_g à une valeur p_{g+1} inférieure à p_g, il n'y a aucune raison pour que la fréquence dans la génération suivante soit supérieure à p_{g+1} et se rapproche de p_g. A chaque instant l'évolution des fréquences est fonction de l'état instantané de la structure génique de la population, et en aucune manière de ses états antérieurs : autrement dit le processus est « strictement markovien ».

Sewall Wright a donné à cette évolution au hasard des fréquences le nom de « Dérive génétique ». Du fait de cette dérive la fréquence d'un gène se modifie légèrement à chaque génération, tantôt en accroissement, tantôt en diminution, sans que ces évolutions successives soient liées, puisqu'aucune « cause » ne les provoque. Il peut sembler que dans ces conditions aucune tendance à long terme ne puisse être définie, et que ces oscillations fassent passer la fréquence de chaque gène par toutes les valeurs possibles sans tendre vers une valeur asymptotique.

Mais un autre facteur intervient : si, au cours de cette dérive la fréquence d'un allèle a_i devient très faible, de l'ordre de $1/N$, le nombre de gènes de ce type présents dans la population est ramené à quelques unités ; la probabilité de ne copier aucun de ces a_i, lors de la procréation suivante, devient importante et cet allèle risque de ne plus être représenté dans la population. Un allèle qui, par exemple, n'existe plus qu'à un seul exemplaire chez les hommes et a disparu chez les femmes risque d'avoir totalement disparu à la génération suivante avec une probabilité : $P_e = \left(1 - \frac{1}{N}\right)$, soit, si N est suffisamment grand, $P_e \simeq e^{-1}$, la probabilité de disparition à la génération suivante d'un gène qui n'est représenté qu'à un exemplaire est ainsi de l'ordre de 4/10.

Par le simple effet de la dérive la fréquence d'un allèle peut donc s'annuler : mais dès que cette fréquence nulle est atteinte les oscillations au hasard s'arrêtent; l'allèle perdu ne peut plus être reconstitué (puisque par hypothèse il n'y a pas de mutation); la fréquence est définitivement fixée à 0. Symétriquement si la dérive amène, par hasard, la fréquence du gène a_i à une valeur proche de 1, le nombre des gènes autres que a_i est ramené à quelques unités et il peut se produire qu'à la génération suivante ils ne soient plus représentés, seul subsiste alors l'allèle a_i qui est, définitivement, le seul présent au locus considéré.

A mesure que le temps s'écoule, que les générations passent, la fixation de la fréquence aux deux valeurs extrêmes $p = 0$ et $p = 1$ devient de plus en plus probable. L'évolution

ne s'arrête que lorsque à chaque locus un seul allèle est représenté. La population, qui n'est plus constituée que d'homozygotes, atteint alors l'homogénéité totale.

Ce raisonnement purement qualitatif peut être précisé par une description du rythme de la dérive. Pour cela il suffit d'utiliser le concept de structure génotypique en probabilité d'une génération g à venir S_g; on peut montrer alors que l'on peut exprimer S_g en fonction de la structure génotypique de la génération initiale et de structures génotypiques de référence dites « Structures panmictiques » et « Structures hoozygotes » équivalentes.

Sans entrer dans les détails, contentons nous ici d'affirmer que le recours aux concepts probabilistes a permis de clarifier considérablement ce problème et de le traiter en toute rigueur.

***

La génétique des populations est une discipline trop ramifiée pour espérer en exposer même les rudiments en un court exposé. Notre désir aujourd'hui était surtout de montrer comment le recours au mode de pensée pascalien, énumérant les « possibles » et leur affectant des probabilités sans accorder un privilège exorbitant au « réel », permet de décrire avec rigueur les conséquences du modèle mendélien de l'hérédité.

Albert JACQUARD

N.B. : M. A. J. vient de publier aux Éditions Masson un livre sur la génétique des populations : *Structures génétiques des Populations.*

DISCUSSION

M. VESSEREAU. — Je voudrais demander à M. Jacquard si, et dans quelle mesure, les mutations peuvent expliquer que la tendance à l'uniformisation, annoncée par le calcul n'est en fait pas constatée?

RÉPONSE DE M. JACQUARD. — La description des effets de la « dérive génétique » que j'ai esquissée n'a de sens que dans l'hypothèse où seule cette dérive intervient. La réalité est évidemment beaucoup plus complexe; des mutations se produisent, des choix du conjoint bouleversent la répartition des gènes, et surtout des effets sélectifs favorisent certains gènes au détriment des autres. La prise en compte, simultanément, de tous ces facteurs complique étrangement la description de l'évolution du patrimoine génétique; il ne pouvait être question de présenter une telle complexité en un si court exposé.