1 Rappel sur la méthode des lod-score

La lectrice est renvoyée au cours de Valérie Chaudru pour le contexte scientifique, la définition du lod-score, son interprétation... Nous nous contenterons ici d'égrener quelques exemples (poly en construction !).

La première section concerne l'estimation de taux de recombinaison (ou de la distance génétique !) entre deux marqueurs observés. Sa lecture doit aider à aborder la seconde section, qui traite du cas du taux de recombinaison entre un marqueur observé et un « locus maladie putatif », et est l'objet de l'analyse de liaison.

1.1 Estimer le taux de recombinaison entre deux marqueurs génétiques

1.1.1 Si la phase est connue

On suppose ici qu'on peut, lors du génotypage, observer « la phase », c'est-à-dire de déterminer quels allèles d'un marqueur autosomal sont sur un même chromosome, ou mieux : sont hérités d'un même parent, ce qui permet de parler de phase même quand les deux locus ne sont pas liés.

Commençons par une famille nucléaire.

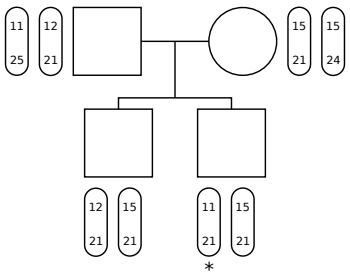


FIGURE 1 – Une famille nucléaire

Les chromosomes 15-21 hérités de la mère peuvent être recombinants ou ne pas l'être : ces transmissions ne sont pas informatives (nous allons y revenir). Du coté du père, on voit chez les enfants un chromosome recombinant (signalé par $*$ sur la figure) et un chromosome non recombinant.

On appelle θ le taux de recombinaison entre les deux locus considérés. La vraisemblance d'une valeur de θ, étant donnée cette observation, est proportionnelle à la probabilité de cette observation. On peut écrire

$$L(\theta) \propto \theta(1-\theta),$$

ce qui correspond à la probabilité d'avoir une recombinaison $\times$ la probabilité de n'avoir pas de recombinaison. Il y a plusieurs constantes (du point de vue du taux de recombinaison) qui sont négligées ici, en particulier les probabilités de transmission de chaque allèle qui ne dépendent pas de θ.

Le maximum de vraisemblance est pour $\theta = \frac{1}{2}$ (correspondant à une absence de liaison entre les deux marqueurs). Le lod-score $Z(\theta) = \log_{10} L(\theta) - \log_{10} L(0{,}5)$ a l'allure suivante.

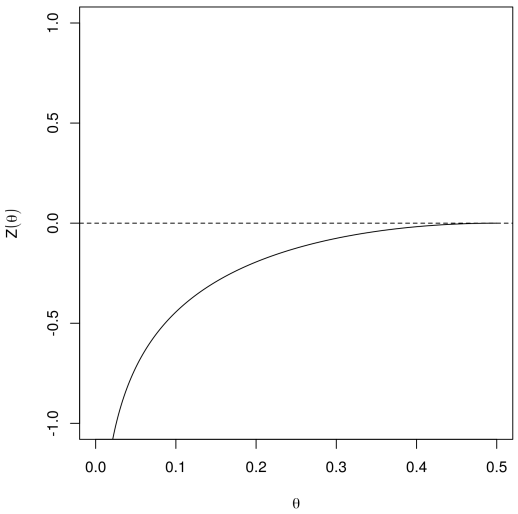


FIGURE 2 – Le lod-score pour la famille de la figure 1

Si la raison pour laquelle on n'a pas pris en compte les transmissions maternelles dans cet exemple n'est pas claire, nous pouvons refaire le calcul comme suit. Il est possible que les deux chromosomes par la mère transmis ne soient pas recombinants, ou qu'ils le soient tous les deux, ou encore qu'un seul des deux le soit (et ce, de deux façons différentes), cf figure 3. En prenant ces possibilités en compte, la vraisemblance est alors

$$L(\theta) \propto (1-\theta)^2 \times \theta(1-\theta) + \theta^2 \times \theta(1-\theta) + 2\theta(1-\theta) \times \theta(1-\theta)$$

et après avoir mis $\theta(1-\theta)$ en facteur, on retrouve la même chose que précédemment.

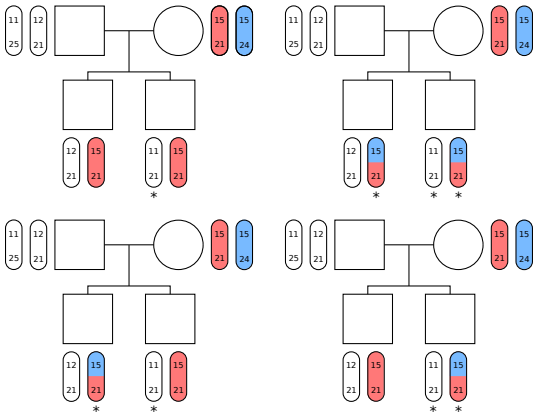


FIGURE 3 – Les quatre possibilités du côté maternel
(les recombinaisons sont signalées par $*$)

1.1.2 Si la phase n'est pas connue

De façon plus réaliste, supposons que la phase ne soit pas connue. Considérons à nouveau une famille nucléaire très similaire à la précédente, avec cette fois-ci trois enfants.

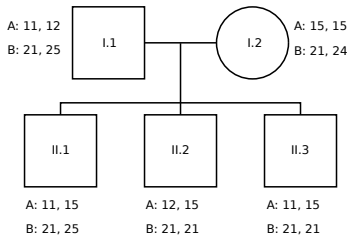


FIGURE 4 – Une famille nucléaire avec trois enfants

Il est facile ici de déterminer la phase des enfants : c'est forcément le père qui transmet A11 ou A12, la mère ne portant pas ces deux allèles ; la mère transmet toujours B21, l'allèle qui reste est l'allèle paternel. Comme dans l'exemple précédent, les transmissions maternelles ne sont pas informatives. Il reste deux possibilités pour la phase paternelle, cf figure 5.

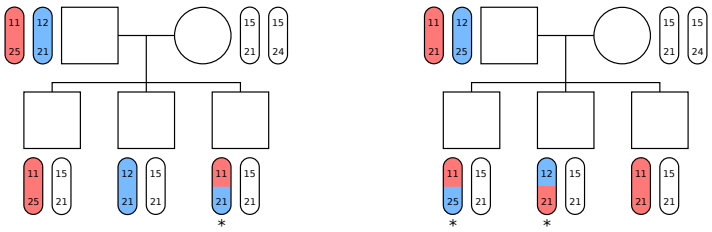


FIGURE 5 – Deux possibilités pour la phase

Une de ces possibilités implique une recombinaison pour trois méïoses (probabilité $\theta(1-\theta)^2$), l'autre deux recombinaisons pour trois méïoses (probabilité $\theta^2(1-\theta)$). La vraisemblance est donc

$$L(\theta) \propto \theta(1-\theta)^2 + \theta^2(1-\theta) = \theta(1-\theta).$$

Tout se passe comme si on n'observait que deux méïoses et non trois. Le maximum de vraisemblance est en $\hat{\theta} = \frac{1}{2}$.

Exercice Si on reprend l'exemple de la famille de la figure 1, en supposant la phase inconnue (ce qui revient à supprimer l'enfant III.1 de la famille de la figure 4), quelle vraisemblance obtient-on ? □

La connaissance des génotypes d'une génération supplémentaire peut tout changer. Si on dispose des génotypes des grands-parents paternels :

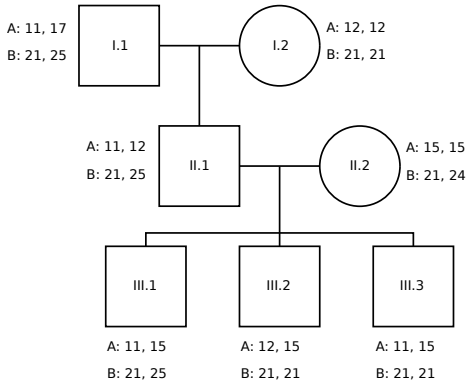


FIGURE 6 – Une génération supplémentaire

On a à nouveau deux possibilités, visibles sur la figure suivante :

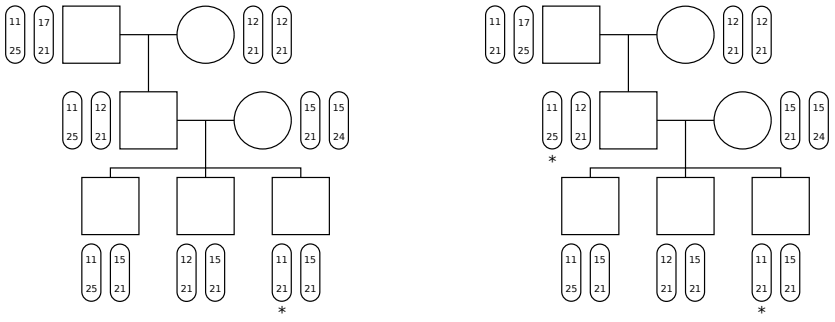


FIGURE 7 – Deux possibilités pour la phase

Cette fois-ci, la phase paternelle est connue, on a déterminé sans ambiguïté n'avoir qu'un enfant recombinant ; reste la possibilité que le père soit lui-même recombinant ou non, mais c'est sans importance :

$$L(\theta) \propto \theta(1-\theta)^3 + \theta^2(1-\theta)^2 = \theta(1-\theta)^2.$$

Le maximum de vraisemblance est en $\hat{\theta} = \frac{1}{3}$. Il est intéressant de constater que c'est l'introduction des grands-parents qui, permettant de « phaser le père », rend les trois enfants pleinement informatifs.

Voici l'allure du lod-score cette fois-ci.

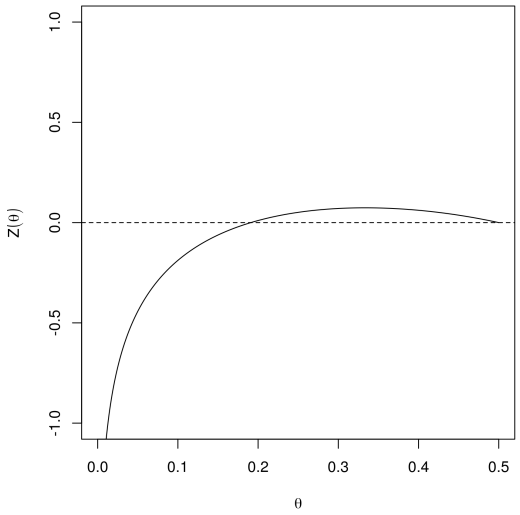


FIGURE 8 – Le lod-score pour la famille de la figure 6

1.1.3 Si certains génotypes ne sont pas connus

Si certains génotypes ne sont pas connus (pas d'ADN / échec du génotypage) il faut énumérer tous les génotypes possibles, et faire l'analyse pour chacun d'eux. Pour limiter le volume de l'exemple, nous allons reprendre la même famille en nous restreignant au cas où le génotype de l'individu II.2 est manquant pour le seul marqueur B :

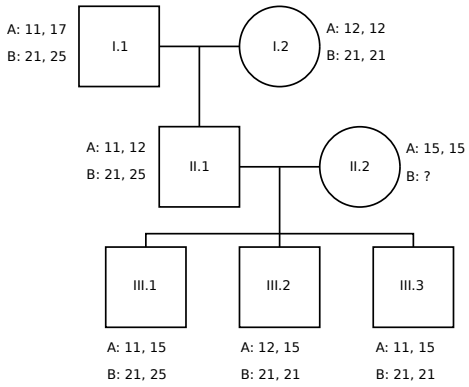


FIGURE 9 – Un génotype manquant

On peut déterminer la phase de deux des enfants, et un des allèles manquant chez II.2. Notons l'autre allèle b. Puisque II.2 est homozyote 15/15 au marqueur A, son génotype phasé est 15-21/15-b.

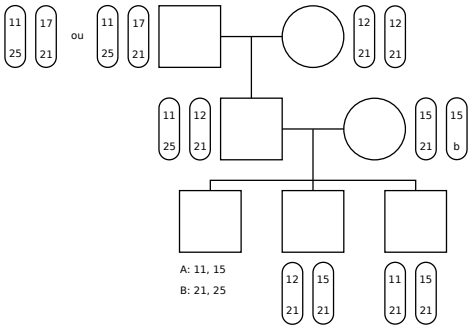


FIGURE 10 – On peut déjà déterminer pas mal de choses...

Il y a deux cas à distinguer. L'allèle indéterminé est l'allèle 25 avec probabilité $\mathbb{P}(b = 25)$ (qui correspond à sa fréquence en population), auquel cas la phase de III.1 ne peut être déterminée :

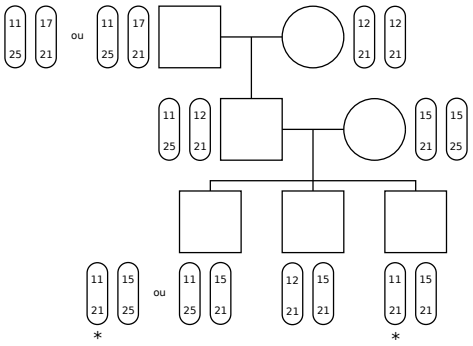


FIGURE 11 – Si $b = 25$ on ne peut pas déterminer la phase de III.1

Les seules transmissions informatives sont les transmissions paternelles à III.2 et III.3 (un non-recombinant et un recombinant). On a

$$L(\theta | b = 25) = \theta(1 - \theta).$$

Si $b \neq 25$, tout se passe comme à la section précédente où tous les génotypes étaient observés.

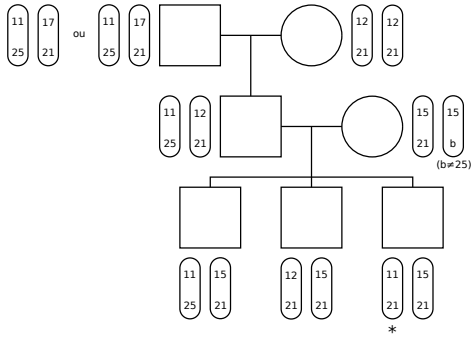


FIGURE 12 – Si $b \neq 25$ on retrouve trois transmissions informatives

$$L(\theta | b \neq 25) = \theta(1-\theta)^2.$$

Au final on a

$$\begin{aligned} L(\theta) &= L(\theta|b=25)\mathbb{P}(b=25) + L(\theta|b\neq 25)(1-\mathbb{P}(b=25)) \\ &= f_{25}\theta(1-\theta) + (1-f_{25})\theta(1-\theta)^2 \\ &= \theta(1-\theta)\big(1-(1-f_{25})\theta\big) \end{aligned}$$

où f_{25} est la fréquence de l'allèle 25 au locus B. Si cet allèle est très rare (f_{25} est très petit), on retrouve $L(\theta) \simeq \theta(1-\theta)^2$: cela revient à négliger la possibilité que $b = 25$. Si au contraire cet allèle est très fréquent (f_{25} est proche de 1), on a $L(\theta) \simeq \theta(1-\theta)$: le cas de la figure 11 est le plus probable et III.1 n'est pas informatif.

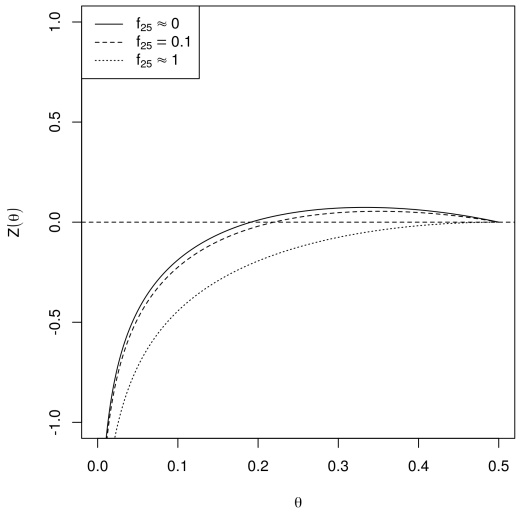


FIGURE 13 – Le lod-score pour diverses valeurs de f_{25}

1.2 Estimer le taux de recombinaison entre un marqueur génétique et un locus maladie

On n'observe pas directement le génotype au locus maladie ; on a cependant une information à travers le phénotype.

Dans le cas d'une maladie dominante causée par un locus di-allélique A/a, si on sait que l'« allèle morbide » a est suffisament rare pour négliger la probabilité qu'un individu soit de génotype aa, tous les atteints sont Aa et les non-atteints sont AA. Ainsi, pour la famille représentée sur la figure suivante :

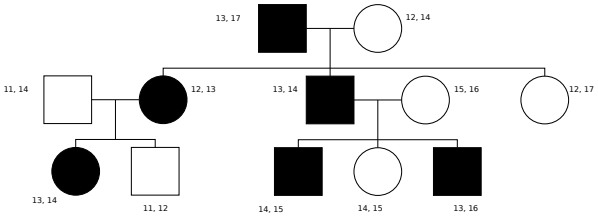


FIGURE 14 – Un pedigree avec une maladie dominante

On n'a que deux cas à envisager, selon la phase de l'individu fondateur dont tous les allèles a observés dans ce pedigree proviennent :

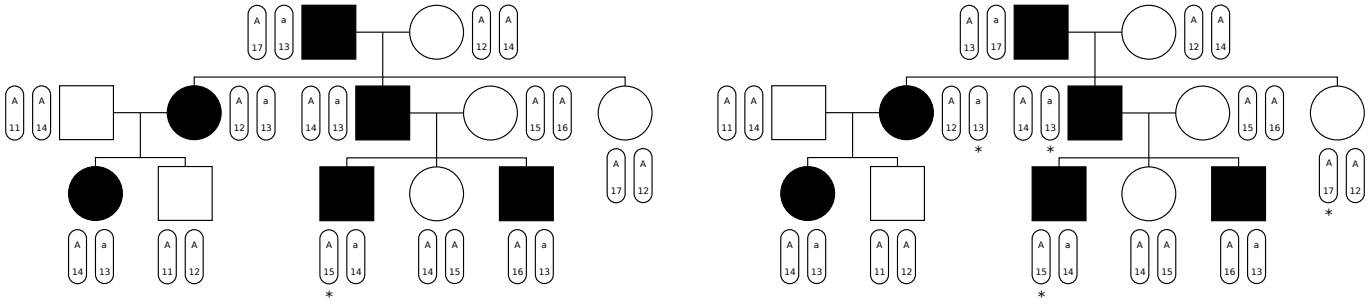


FIGURE 15 – Les deux cas possibles

Dans le premier cas, on a une recombinaison pour 8 méïoses chez un individu porteur de l'allèle a, dans le deuxième cas, on a 4 recombinaisons, d'où une vraisemblance

$$L(\theta) = \theta(1-\theta)^7 + \theta^4(1-\theta)^4 = \theta(1-\theta)^4\left((1-\theta)^3 + \theta^3\right).$$

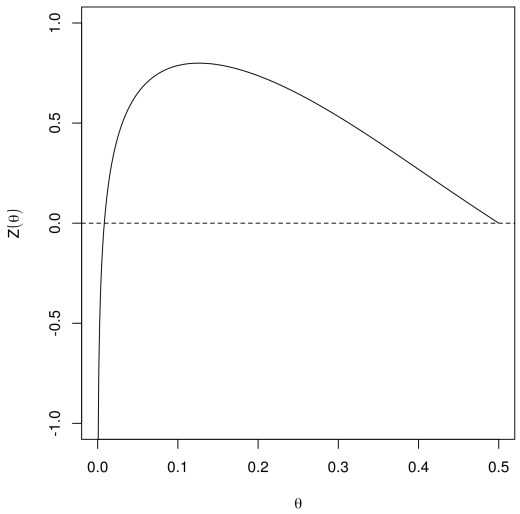


FIGURE 16 – Le lod-score pour la famille de la figure 6

Pour les maladies récessives, la combinatoire est plus complexe, car il n'est pas toujours possible de déterminer, parmi les individus sains, lesquels sont de génotypes AA et lesquels sont de génotypes Aa. Il faut donc envisager beaucoup plus de combinaisons génotypiques. La présence d'individus dont on n'a pas les génotypes non plus au marqueur, la taille des pedigrees, rendent les calculs très difficiles à faire à la main.

On peut élargir le cadre à la recherche d'un « gène majeur », modélisé par les paramètres $q = \mathbb{P}(a)$, $f_{\mathrm{AA}} = \mathbb{P}(\text{atteint}|\mathrm{AA})$, $f_{\mathrm{A}a} = \mathbb{P}(\text{atteint}|\mathrm{A}a)$ et $f_{aa} = \mathbb{P}(\text{atteint}|aa)$. Si la pénétrance est incomplète, il faudra envisager encore plus de combinaisons génotypiques. Il est nécessaire d'estimer les paramètres de pénétrance en amont, par une étude de ségrégation, on ne peut pas mettre des paramètres « au doigt mouillé » : l'expérience montre que le résultat obtenu est très sensible aux valeurs de ces paramètres.

2 Rappel sur le MLS

2.1 Motivation

La méthode des lod-scores demande une estimation préalable du modèle génétique (fréquence de l'allèle à risque, pénétrance des génotypes), afin de pouvoir calculer la probabilité de chaque génotype possible, conditionnellement au phénotype. Ceci est possible dans le cas des maladies mendéliennes ; cela devient plus difficile dans le cas des maladies monogéniques avec des pénétrances incomplètes (cas où il existe un « gène majeur ») ; pour les maladies complexes, c'est à peu près impossible.

Les résultats de la méthode dépendent beaucoup de cette spécification du modèle, ce qui la rend à peu près inutilisable pour les maladies complexes.

On a donc recherché des méthodes d'analyse de liaison qui se passent de modèle (*disease-model free*, également dites « non paramétriques »). La méthode du MLS est (historiquement) la première d'entre elles. Elle ne permet de traiter « que » les paires de germains atteints (*affected sib-pairs*).

2.2 Le MLS

2.2.1 Quand il n'y a pas d'ambiguïté sur les états IBD

Si on observe N_0, N_1 et N_2 paires de germains atteints avec IBD 0, 1 et 2, on réalise un test de χ^2 « de conformité » à deux degrés de liberté, les effectifs attendus étant, sous H_0, $\frac{1}{4}N$, $\frac{1}{2}N$ et $\frac{1}{4}N$, avec $N = \sum_i N_i$.

2.2.2 Quand il y a de l'ambiguïté sur les états IBD

C'est un test de maximum de vraisemblance. Le paramètre du modèle est $z = (z_0, z_1, z_2)$, avec $z_j = \mathbb{P}(\text{IBD} = j|\text{ASP})$ (où ASP = affected sib-pair). On a naturellement la contrainte $z_0 + z_1 + z_2 = 1$. Sous l'hypothèse nulle H_0 : pas de liaison entre le locus étudié et le locus maladie, on a

$$z = z_{H_0} = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right)$$

Dans la famille numéro i, la vraisemblance du modèle est

$$\begin{aligned} L_i(z) &= \mathbb{P}_z(\text{génotypes}|\text{ASP}) \\ &= \sum_{j=0}^{2} \mathbb{P}(\text{génotypes}|\text{IBD} = j)\mathbb{P}_z(\text{IBD} = j|\text{ASP}) \\ &= \sum_{j=0}^{2} z_j w_{ij}. \end{aligned}$$

Calculer la probabilité $w_{ij} = \mathbb{P}(\text{génotypes}|\text{IBD} = j)$ des génotypes observés pour une paire de germains donnés, conditionnellement à l'IBD, se fait par des raisonnements probabilistes simples : par exemple, si on ne connaît pas les génotypes des parents de deux germains atteints, on a, pour des allèles A et B distincts,

$$\mathbb{P}(G_1 = AB, G_2 = AB|IBD = j) = \mathbb{P}(G_1 = AB)\mathbb{P}(G_2 = AB|G_1 = AB, IBD = j)$$
$$= \begin{cases} (2f_A f_B) \times (2f_A f_B) & \text{si } j = 0 \\ (2f_A f_B) \times \left(\frac{1}{2}f_A + \frac{1}{2}f_B\right) & \text{si } j = 1 \\ (2f_A f_B) \times 1 & \text{si } j = 2 \end{cases}$$

Détaillons le calcul :

— pour $j = 0$, les deux germains ne partagent pas d'allèles IBD, et leurs génotypes sont indépendants :

$$\mathbb{P}(G_2 = AB|G_1 = AB, IBD = 0) = \mathbb{P}(G_2 = AB) = (2f_A f_B);$$

— pour $j = 1$, sachant que le premier germain est AB et qu'il y a un allèle partagé IBD : soit il s'agit de l'allèle A et le second germains a reçu (de façon indépendante) un allèle B ; soit c'est le contraire. Donc :

$$\begin{aligned}\mathbb{P}(G_2 = AB|G_1 = AB, IBD = 1) &= \mathbb{P}(\text{allèle partagé} = A, \text{autre allèle} = B \text{ ou allèle partagé} = B, \text{autre allèle} = A) \\ &= \mathbb{P}(\text{allèle partagé} = A)\mathbb{P}(\text{autre allèle} = B) + \mathbb{P}(\text{allèle partagé} = B)\mathbb{P}(\text{autre allèle} = A) \\ &= \frac{1}{2}f_B + \frac{1}{2}f_A;\end{aligned}$$

— pour $j = 2$, on a

$$\mathbb{P}(G_2 = AB|G_1 = AB, IBD = 2) = 1.$$

La table suivante récapitule tous les cas possibles pour une paire de germains avec génotypes parentaux inconnus (utiliser $\mathbb{P}(G_1 = g, G_2 = h) = \mathbb{P}(G_1 = h, G_2 = g)$ pour les cas non listés).

g	h	$j = 0$	$j = 1$	$j = 2$
AB	CD	$4f_A f_B f_C f_D$	0	0
AA	BC	$2f_A^2 f_B f_C$	0	0
AA	BB	$f_A^2 f_B^2$	0	0
AB	AC	$4f_A^2 f_B f_C$	$f_A f_B f_C$	0
AA	AB	$2f_A^3 f_B$	$f_A^2 f_B$	0
AB	AB	$4f_A^2 f_B^2$	$f_A f_B (f_A + f_B)$	$2f_A f_B$
AA	AA	f_A^4	f_A^3	f_A^2

TABLE 1 – Valeurs de $\mathbb{P}(G_1 = g, G_2 = h|IBD = j)$

On note $\hat{z}$ l'estimateur du maximum de vraisemblance : $\hat{z} = \arg\max_z \ell(z)$. La statistique de test du rapport de vraisemblance est

$$Z_{MLS} = 2\left(\ell(\hat{z}) - \ell\left(z_{H_0}\right)\right).$$

Elle suit (asymptotiquement) un $\chi^2(2)$ (deux degrés de libertés pour deux paramètres « libres »). Dans le cas non-ambigu, on retrouve un test asymptotiquement équivalent au χ^2 de conformité.

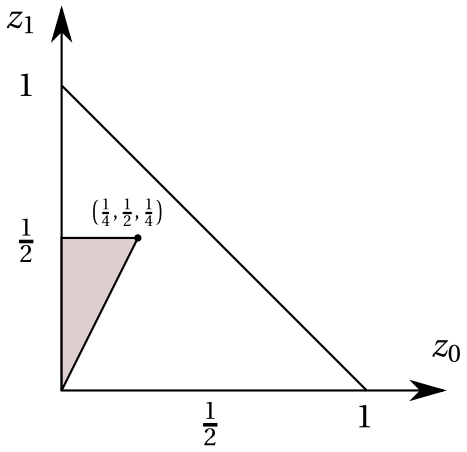


FIGURE 1 – On peut restreindre les paramètres au triangle grisé

2.3 Les contraintes du triangle (contraintes de Risch)

On peut contraindre les paramètres z_0, z_1, z_2 à vivre dans le petit triangle défini par $2z_0 \leqslant z_1 \leqslant \frac{1}{2}$ (cf figure 2.3). Ceci va permettre d'augmenter la puissance du test. Nous allons d'abord donner des arguments pour justifier cette contrainte.

2.3.1 Récurrence familiale

Notons λ_R la récurrence familiale, définie pour les relations d'apparentement $R = S, 1, MZ$ pour germain *(sibling)*, enfant ou parent, et jumeau monozygote, par

$$\lambda_R = \frac{K_R}{K},$$

où $K = \mathbb{P}(\text{Atteint})$ est la prévalence de la maladie et $K_R = \mathbb{P}(\text{Atteint}|\text{Apparenté R atteint})$ est la prévalence de la maladie chez les apparentés d'un atteint.

Dans un modèle monogénique, la probabilité d'être atteint quand on est IBD j au locus maladie avec un atteint est

$$\mathbb{P}(\text{Atteint}|\text{IBD} = j) = \begin{cases} K & \text{si } j = 0 \\ K\lambda_1 & \text{si } j = 1 \\ K\lambda_{MZ} & \text{si } j = 2 \end{cases}$$

En effet, le cas R = 1 correspond à IBD = 1 sur l'ensemble du génome, et le cas R = MZ à IBD = 2 sur l'ensemble du génome.

Ceci permet de calculer par exemple λ_S en fonction de λ_1 et λ_{MZ} :

$$\begin{aligned} K_S &= \frac{1}{4}\mathbb{P}(\text{Atteint}|\text{IBD} = 0) + \frac{1}{2}\mathbb{P}(\text{Atteint}|\text{IBD} = 1) + \frac{1}{4}\mathbb{P}(\text{Atteint}|\text{IBD} = 2) \\ \lambda_S K &= \frac{1}{4}K + \frac{1}{2}\lambda_1 K + \frac{1}{4}\lambda_{MZ} K \\ \lambda_S &= \frac{1}{4} + \frac{1}{2}\lambda_1 + \frac{1}{4}\lambda_{MZ} \end{aligned}$$

On peut aussi calculer la probabilité d'être IBD = j pour une paire de germains atteints :

$$z_j = \mathbb{P}(\text{IBD} = j|\text{ASP}) = \frac{\mathbb{P}(\text{ASP}|\text{IBD} = j)\mathbb{P}(\text{IBD} = j)}{\mathbb{P}(\text{ASP})} = \begin{cases} \frac{1}{4\lambda_S} & \text{si } j = 0 \\ \frac{\lambda_1}{2\lambda_S} & \text{si } j = 1 \\ \frac{\lambda_{MZ}}{4\lambda_S} & \text{si } j = 2 \end{cases} \tag{2.1}$$

2.3.2 Les contraintes du triangle

On suppose en outre qu'on a

$$1 \leq \lambda_1 \leq \lambda_S \leq \lambda_{MZ}, \tag{2.2}$$

ce qui correspond au fait que le risque relatif diminue avec la proximité génétique d'avec l'atteint. Ceci est vérifié dans le cas monogénique, et devrait l'être par tout modèle « raisonnable ».

Notons que dans le cas dominant, $\lambda_1 = \lambda_S = \frac{1}{2}$, alors que dans le cas récessif on a $K = q^2$ (où q est la fréquence de l'allèle morbide), $\lambda_1 \simeq \frac{q}{K} = \frac{1}{q}$ et $\lambda_S \simeq \frac{1/4}{K} = \frac{1}{4q^2}$, donc si q est petit on a bien $\lambda_1 < \lambda_S$. En d'autres termes, on peut être IBD = 2 avec sa sœur mais pas avec son père, avec lequel on est toujours IBD = 1 : dans le cas récessif, avoir une sœur atteinte confère donc un risque beaucoup plus important que d'avoir un père atteint. Cette possibilité d'une plus grande similarité entre germains qu'entre parent/enfant justifie (en général) l'inégalité $\lambda_1 \leq \lambda_S$.

En combinant 2.1 et 2.2, on obtient

$$2z_0 \leq z_1 \leq \frac{1}{2}. \tag{2.3}$$

Il ressort de (Risch 1990) que ces inégalités restent vraies dans une grande variété de modèles polygéniques. Holmans a donc proposé (Holmans 1993) de restreindre la recherche du maximum de vraisemblance au « triangle de Risch », ce qui augmente la puissance du test.

2.3.3 Le test sous contrainte

On utilise toujours une statistique de test

$$Z_{MLS} = 2\left(\ell(\hat{z}) - \ell(z_{H_0})\right).$$

Le point z_{H_0} correspondant à l'hypothèse nulle est au bord de l'espace des paramètres : on ne peut plus utiliser les résultats asymptotiques. La loi de la statistique de test (sous H_0) n'est plus un $\chi^2(2)$ mais un mélange entre la constante 0, un $\chi^2(1)$ et un $\chi^2(2)$, dans des proportions qui dépendent du nombre d'allèles au locus considéré et de leurs fréquences.

Ce test est plus puissant que le test non-contraint. La raison intuitive de ce gain de puissance est que la contrainte impose à la statistique de test, sous H_0, d'être généralement plus petite que la statistique non-contrainte ; sous H_1, la contrainte a moins d'influence. La valeur seuil est donc abaissée, et on la dépasse « plus facilement » sous H_1.

Éléments de bibliographie

Holmans P. (1993) Asymptotic properties of affected-sib-pair linkage analysis. *Am. J. Hum. Genet.* 1993 :52, 362–74.

Risch N. (1990) Linkage Strategies for Genetically Complex Traits. Trois articles publiés ensemble et numérotés I, II, III, *Am. J. Hum. Genet.* 1990 :46, 222–253.

3 NPL : Non Parametric Linkage

Le NPL est une extension du MLS : une statistique permettant de tester la liaison, sans modèle (dominant, récessif, etc) pour la maladie, d ans des pedigrees plus complexes que des paires de germains affectés.

3.1 Quand il n'y a pas d'ambiguïté sur les méioses

Nous traitons ici le cas où « chaque méiose est informative », c'est-à-dire qu'on sait déterminer, pour chaque individu non-fondateur, quel allèle lui ont transmis chacun de ses parents.

3.1.1 Paires de germains

On suppose que toutes les familles considérées sont constituées de deux parents et deux enfants atteints. Dans la famille i, on pose $S_i = 0,1,2$ selon l'IBD des deux germains au point considéré. Sous l'hypothèse nulle H_0 : « pas de liaison », on a

$$E(S) = \mu_0 = 1$$

$$\text{var}(S) = \sigma_0^2 = \frac{1}{2}$$

On pose $Z_i = \frac{1}{\sigma_0}(S_i - \mu_0)$, la variable centrée réduite associée. Sous H_0, on a $E(Z_i) = 0$; les $Z_i > 0$ plaident en faveur de H_1 : « liaison entre le locus et la maladie ».

Pour l'ensemble des familles, on pose

$$Z = \frac{1}{\sqrt{\sum_i \gamma_i^2}} \sum_i \gamma_i Z_i,$$

où γ_i est le poids attribué à la famille i – dans le cas présent, on prendra $\gamma_i = 1$ pour toutes les familles. Nous verrons plus loin quels autres choix sont possibles pour des structures familiales plus complexes.

Sous H_0, la variable Z est centrée et réduite ; quand les familles sont assez nombreuses, on a (par une version forte du Théorème central de la limite) $Z \sim \mathcal{N}(0,1)$.

Sous H_1, on attend $Z > 0$, ce qui correspond à un excès d'allèles IBD. On réalise donc un test unilatéral, dont le degré de signification pour une valeur observée z est

$$p = \mathbb{P}(Z > z).$$

3.1.2 Familles complexes

Tournons-nous vers le cas général.

La statistique de test est construite de la même façon : on calcule une statistique S_i dans chaque famille, on y associe une statistique centrée réduite Z_i, et la statistique pour l'ensemble des familles sera $Z = \frac{1}{\sqrt{\sum_i \gamma_i^2}} \sum_i \gamma_i Z_i$,

qui sera à nouveau asymptotiquement distribuée selon une loi $\mathcal{N}(0,1)$.

Nous allons tout d'abord définir le vecteur d'hérédité d'une famille.

Vecteur d'hérédité

Dans chaque famille, on définit f et n comme suit :
- f est le nombre d'individus fondateurs ; les $2f$ allèles fondateurs sont numérotés par $j = 1, \ldots, 2f$;
- n est le nombre d'individus non-fondateurs.

Pour chaque individu d'une famille, on ordonne les allèles au locus considéré
- de façon arbitraire pour les fondateurs (par exemple, le premier allèle est celui transmis au premier descendant) ;
- pour les non-fondateurs, on met d'abord l'allèle paternel puis l'allèle maternel.

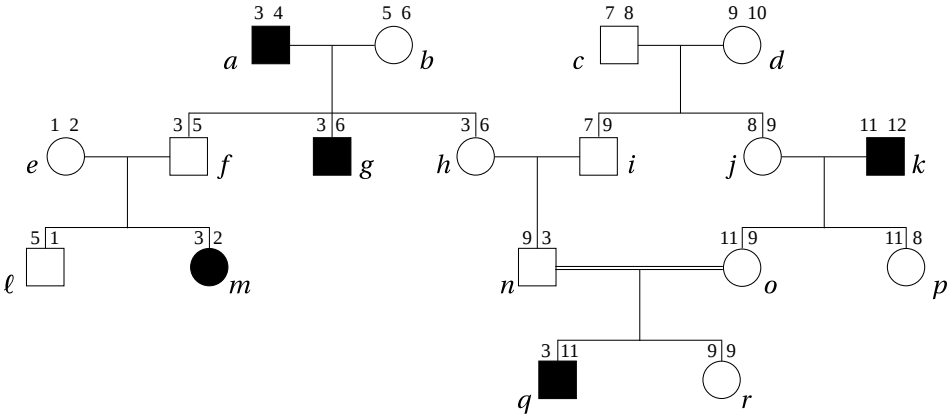


FIGURE 1 – Un exemple de généalogie

Exemple : Dans le pedigree représenté figure 1, on a 6 individus fondateurs : a, b, c, d, e et k, dont les allèles ont été numérotés de 1 à 12. On a indiqué les allèles des 12 individus non-fondateurs, en respectant la convention d'ordre (paternel, maternel) ci-dessus. □

On construit, pour chaque individu non-fondateur dont les allèles sont j_1, j_2 (dans cet ordre : j_1 est l'allèle paternel, j_2 est l'allèle maternel), un vecteur à deux composantes (a,b) où
- $a = 0,1$ selon que j_1 est le premier ou le second allèle du père,
- $b = 0,1$ selon que j_2 est le premier ou le second allèle de la mère.

Le vecteur d'hérédité est construit en concaténant ces vecteurs (on a ordonné les individus de façon arbitraire). Sa longueur est $2n$ où n est le nombre d'individus non-fondateurs de la famille ; il résume le résultat de toutes les meïoses observées. Il y a 2^{2n} vecteurs d'hérédité possibles, tous a priori équiprobables (sous H_0), de probabilité $c = 2^{-2n}$.

Exemple : Pour le pedigree de la figure 1, on a $v_f = (0,0)$, $v_g = (0,1)$, $v_h = (0,1)$, $v_i = (0,0)$, $v_j = (1,0)$, $v_\ell = (1,0)$, $v_m = (0,1)$, $v_n = (1,0)$, $v_o = (0,1)$, $v_p = (0,0)$, $v_q = (1,0)$ et $v_r = (0,1)$. Le vecteur d'hérédité (de longueur $2n = 24$) est

$$\begin{array}{rcl} v & = & (v_f, \quad v_g, \quad v_h, \quad v_i, \quad v_j, \quad v_\ell, \quad v_m, \quad v_n, \quad v_o, \quad v_p, \quad v_q, \quad v_r) \\ & = & (0,0, \quad 0,1, \quad 0,1, \quad 0,0, \quad 1,0, \quad 1,0, \quad 0,1, \quad 1,0, \quad 0,1, \quad 0,0, \quad 1,0, \quad 0,1) \end{array}$$

La connaissance du vecteur d'hérédité permet de trouver quels allèles fondateurs ont été transmis à tous les individus de la famille.

Vecteur d'hérédité condensé

En utilisant la convention qui ordonne les allèles des fondateurs en plaçant en premier l'allèle transmis au premier descendant, on peut se contenter de vecteurs d'hérédité de longueur $2n-f$. Cette nuance n'a que peu d'importance pour le présent exposé, mais elle permet en pratique d'alléger les calculs.

Si on reprend l'exemple précédent, cette convention impose que v_f et v_i sont toujours égaux (0,0), et que le premier élément de v_o et le deuxième élément de v_ℓ sont toujours 0. On obtient un vecteur d'hérédité condensé de longueur $2n-f=18$

$$\begin{array}{rcccccccccccccc} v & = & (v_f, & v_g, & v_h, & v_i, & v_j, & v_\ell, & v_m, & v_n, & v_o, & v_p, & v_q, & v_r) \\ & = & (& 0,1, & 0,1, & & 1,0, & 1, & 0,1, & 1,0, & 1, & 0,0, & 1,0, & 0,1) \end{array}$$

Définir S dans chaque famille

Dans chaque famille i, on définit une statistique S_i. Oublions momentanément le numéro de la famille, et parlons simplement de la statistique S. Il y a plusieurs définitions concurrentes pour S. Les deux plus couramment utilisées sont S^{pairs} et S^{all}.

La définition de S^{pairs} est très simple : on somme les états IBD de toutes les paires d'atteints.

$$S^{pairs} = \sum_{u,v \text{ atteints}} IBD(u,v).$$

Dans le cas de l'exemple de la figure 1, on a

$$\begin{array}{rcccccccccc} S^{pairs} & = & IBD(a,g) & + & IBD(a,k) & + & IBD(a,m) & + & IBD(a,q) \\ & = & & + & IBD(g,k) & + & IBD(g,m) & + & IBD(g,q) \\ & = & & & & + & IBD(k,m) & + & IBD(k,q) \\ & = & & & & & & + & IBD(m,q) \end{array}$$

et donc $S^{pairs} = 1+0+1+1+0+1+1+0+1+1 = 7$.

La statistique S^{all} prend en compte tous les apparentés à la fois. Pour la définir, nous avons besoin de quelques notations supplémentaires :

- m est le nombre d'atteints dans la famille ;
- $\mathscr{H}$ est l'ensemble de tous les vecteurs obtenus en prenant un allèle de chacun des atteints ; les éléments de $\mathscr{H}$ sont donc des vecteurs de longueur m, il y en a 2^m en tout ;
- pour tout $h \in \mathscr{H}$ et $j = 1,\dots,2f$, on note $\psi(j,h)$ le nombre de fois où l'allèle fondateur j apparaît dans h.

Enfin, on pose

$$S^{all} = \frac{1}{2^m} \sum_{h \in \mathscr{H}} \prod_{j=1}^{2f} \psi(j,h)!$$

Dans le cas de l'exemple de la figure 1, on a 32 vecteurs h possibles. Les valeurs de $\psi(j,h)$ sont non nulles pour $j = 2,3,4,6,11$ et 12. Le calcul de S^{all} est détaillé dans la table 1.

Remarque : La valeur de S dépend uniquement (quand la structure familiale est fixée) du vecteur d'hérédité v : en effet, la valeur de v permet de retrouver tous les allèles portés par les non-fondateurs. On peut considérer $S = S(v)$ comme une fonction de v.

Tous les vecteurs d'hérédité étant équiprobables, on peut calculer $\mu_i = E(S_i)$ comme la moyenne des $S_i(v_1)$, $S_i(v_2)$,… (tous les vecteurs possibles). On peut de même calculer $\sigma_i^2 = \text{var}(S_i)$. Ceci va permettre de définir une statistique centrée réduite Z_i associée à la famille i, puis la statistique Z.

h	$\psi(2,h)$	$\psi(3,h)$	$\psi(4,h)$	$\psi(6,h)$	$\psi(11,h)$	$\psi(12,h)$	$\prod_j \psi(j,h)!$
(3, 3, 11, 3, 3)	0	4	0	0	1	0	24
(3, 3, 11, 3, 11)	0	3	0	0	2	0	12
(3, 3, 11, 2, 3)	1	3	0	0	1	0	6
(3, 3, 11, 2, 11)	1	2	0	0	2	0	4
(3, 3, 12, 3, 3)	0	4	0	0	0	1	24
(3, 3, 12, 3, 11)	0	3	0	0	1	1	6
(3, 3, 12, 2, 3)	1	3	0	0	0	1	6
(3, 3, 12, 2, 11)	1	2	0	0	1	1	2
(3, 6, 11, 3, 3)	0	3	0	1	1	0	6
(3, 6, 11, 3, 11)	0	2	0	1	2	0	4
(3, 6, 11, 2, 3)	1	2	0	1	1	0	2
(3, 6, 11, 2, 11)	1	1	0	1	2	0	2
(3, 6, 12, 3, 3)	0	3	0	1	0	1	6
(3, 6, 12, 3, 11)	0	2	0	1	1	1	2
(3, 6, 12, 2, 3)	1	2	0	1	0	1	2
(3, 6, 12, 2, 11)	1	1	0	1	1	1	1
(4, 3, 11, 3, 3)	0	3	1	0	1	0	6
(4, 3, 11, 3, 11)	0	2	1	0	2	0	4
(4, 3, 11, 2, 3)	1	2	1	0	1	0	2
(4, 3, 11, 2, 11)	1	1	1	0	2	0	2
(4, 3, 12, 3, 3)	0	3	1	0	0	1	6
(4, 3, 12, 3, 11)	0	2	1	0	1	1	2
(4, 3, 12, 2, 3)	1	2	1	0	0	1	2
(4, 3, 12, 2, 11)	1	1	1	0	1	1	1
(4, 6, 11, 3, 3)	0	2	1	1	1	0	2
(4, 6, 11, 3, 11)	0	1	1	1	2	0	2
(4, 6, 11, 2, 3)	1	1	1	1	1	0	1
(4, 6, 11, 2, 11)	1	0	1	1	2	0	2
(4, 6, 12, 3, 3)	0	2	1	1	0	1	2
(4, 6, 12, 3, 11)	0	1	1	1	1	1	1
(4, 6, 12, 2, 3)	1	1	1	1	0	1	1
(4, 6, 12, 2, 11)	1	0	1	1	1	1	1
						Total	146

TABLE 1 – Calcul de S^{all} pour la généalogie de la figure 1

La statistique de test Z

On pose $Z_i = \frac{1}{\sigma_i}(S_i - \mu_i)$, et la statistique globale est

$$Z = \frac{1}{\sqrt{\sum_i \gamma_i^2}} \sum_i \gamma_i Z_i.$$

Le degré de signification est là encore, pour une valeur observée z, $\mathbb{P}(Z > z)$ pour $Z \sim \mathcal{N}(0,1)$.

Contrairement au cas des paires de germains, il devient pertinent d'envisager des valeurs de γ_i différentes d'une famille à l'autre. Si le choix le plus simple reste $\gamma_i = 1$ pour tout i, on a proposé de prendre $\gamma_i^2 = \sigma_i$, de façon à donner plus de poids aux familles où la variation peut être la plus forte (ces familles étant a priori plus informatives).

3.2 Quand il y a de l'ambiguïté sur les méioses

La solution est d'écrire une vraisemblance. En cours, nous n'en dirons pas plus ! Les détails sont conservés ici pour les plus curieuses d'entre vous.

3.2.1 Première approche : le $\overline{Z}$

Nous mentionnons ici l'approche « naïve » qui a été tout d'abord proposée : utiliser l'espérance de la statistique conditionnellement aux génotypes observés.

Dans une famille donnée, en utilisant les génotypes observés $G_i = g_i$, on calcule

$$\begin{aligned}\overline{S}_i &= E(S_i | G_i = g_i) \\ &= \sum_v S_i(v) \mathbb{P}(v | G_i = g_i)\end{aligned}$$

où v parcourt tous les vecteurs d'hérédité possibles. Il s'agit donc de la « valeur moyenne » de S_i, compte tenu des information dont on dispose.

On pose ensuite

$$\overline{Z}_i = \frac{1}{\sigma_i}\left(\overline{S}_i - \mu_i\right)$$

et

$$\overline{Z} = \frac{1}{\sqrt{\sum_i \gamma_i^2}} \sum_i \gamma_i \overline{Z}_i.$$

On fait le test comme avant, en prenant comme degré de signification $\mathbb{P}(Z > \overline{z})$ où $Z \sim \mathcal{N}(0,1)$.

Mais le test ainsi construit est trop conservateur : en effet, sous H_0, la statistique construire est bien centrée, mais sa variance est < 1. C'est assez intuitif : la variance d'une moyenne est plus petite que la variance des quantités dont on fait la moyenne.

3.2.2 Détails sur la distribution du test

On va montrer qu'on a $E\left(\overline{S}_i\right) = E(S_i) = \mu_i$ et $\mathrm{var}\left(\overline{S}_i\right) < \mathrm{var}(S_i) = \sigma_i^2$. Dans la suite de cette section nous omettrons l'indice i.

Montrons d'abord que si on prend l'espérance (la moyenne sur toutes les génotypes possibles) de $\mathbb{P}(v|\mathrm{G}=g)$ (qui intervient dans le calcul de $\overline{\mathrm{S}}$), on obtient $\mathbb{P}(v|\mathrm{G}=g)=\mathbb{P}(v)$:

$$\begin{aligned}\mathrm{E}\big(\mathbb{P}(v|\mathrm{G}=g)\big) &= \sum_g \mathbb{P}(v|\mathrm{G}=g)\times\mathbb{P}(\mathrm{G}=g)\\ &= \mathbb{P}(v)\end{aligned}$$

par la formule des probabilités totales. La somme est sur tous les génotypes g envisagables. Notons que tous les vecteurs d'hérédité sont a priori équiprobables ($\mathbb{P}(v)=c$).

Montrons le résultat annoncé.

— La statistique est centrée : on a bien

$$\begin{aligned}\mathrm{E}\left(\overline{\mathrm{S}}\right) &= \mathrm{E}\left(\sum_v \mathbb{P}(v|\mathrm{G}=g)\mathrm{S}(v)\right)\\ &= \sum_v \mathrm{E}\big(\mathbb{P}(v|\mathrm{G}=g)\big)\mathrm{S}(v)\\ &= \sum_v \mathbb{P}(v)\mathrm{S}(v)\\ &= \mu.\end{aligned}$$

— Sa variance est <1 : si on a des probabilités $p_1, p_2\ldots$ (avec $\sum_k p_k = 1$), alors $(\sum p_k x_k)^2 \leqslant \sum p_k x_k^2$ (et l'inégalité est stricte sauf quand un des $p_k = 1$ et tous les autres sont nuls). On a donc

$$\begin{aligned}\mathrm{E}\left(\overline{\mathrm{S}}^2\right) &= \mathrm{E}\left(\left(\sum_v \mathbb{P}(v|\mathrm{G}=g)\mathrm{S}(v)\right)^2\right)\\ &\leq \mathrm{E}\left(\sum_v \mathbb{P}(v|\mathrm{G}=g)\mathrm{S}(v)^2\right)\\ &= \mathrm{E}\left(\mathrm{S}^2\right)\end{aligned}$$

et donc $\mathrm{var}\left(\overline{\mathrm{S}}\right) = \mathrm{E}\left(\overline{\mathrm{S}}^2\right) - \mu^2 \leq \mathrm{E}\left(\mathrm{S}^2\right) - \mu^2 = \mathrm{var}(\mathrm{S})$. La variance de $\overline{\mathrm{S}}$ est plus petite que σ_i^2, et ce d'autant plus qu'il y a de vecteurs v compatibles avec les génotypes observés ; au final, la loi de $\overline{\mathrm{Z}}$ est centrée de variance plus petite que 1.

Remarque sur le calcul pratique de $\overline{\mathrm{S}}_i$

On a

$$\mathbb{P}(v|\mathrm{G}=g) = \frac{\mathbb{P}(\mathrm{G}=g|v)\mathbb{P}(v)}{\sum_w \mathbb{P}(\mathrm{G}=g|w)\mathbb{P}(w)}$$

D'autre part tous les vecteurs d'hérédité sont a priori équiprobables : pour tout w, $\mathbb{P}(w)=\mathbb{P}(v)$. On a donc

$$\mathbb{P}(v|\mathrm{G}=g) = \frac{\mathbb{P}(\mathrm{G}=g|v)}{\sum_w \mathbb{P}(\mathrm{G}=g|w)}.$$

Quand tous les génotypes sont observés, on a $\mathbb{P}(\mathrm{G}=g|w)=1$ ou 0 selon que le vecteur d'hérédité est compatible ou non avec les génotypes. Quand certains génotypes manquent, il faut utiliser les fréquences alléliques pour écrire $\mathbb{P}(\mathrm{G}=g|w)$, dont l'expression peut devenir assez complexe.

3.2.3 Deuxième approche : le $Z_{\ell r}$

On fixe S une mesure de l'IBD, par exemple $\mathrm{S}^{\mathrm{pairs}}$ ou $\mathrm{S}^{\mathrm{all}}$.

Soit v un vecteur d'hérédité (pour la famille i). On écrit un modèle à un paramètre δ ; ce modèle s'écarte d'autant plus de H_0 que $\delta>0$ est grand. On pose

$$\mathbb{P}_\delta(v) = c_i \times \big(1+\delta\gamma_i \mathrm{Z}_i(v)\big),$$

où γ_i est le poids de la famille i et c_i est la probabilité de chacun des vecteurs d'hérédité sous H_0 (hypothèse sous laquelle on rappelle qu'ils sont équiprobables). Notons que $\sum_v Z_i(v) = 0$ (car $Z_i(v) = \frac{1}{\sigma_i}(S_i(v) - \mu_i)$ avec $\mu_i = \sum_v c_i S_i(v)$), et donc on a bien pour tout δ, $\sum_v \mathbb{P}_\delta(v) = \sum_v c_i = 1$.

On voit que si $\delta > 0$, les vecteurs d'hérédité pour lesquels $S_i(v)$ (et donc $Z_i(v)$) est grand ont une probabilité plus importante. Si $\delta = 0$, on a $\mathbb{P}(v) = c_i$, et tous les vecteurs d'hérédité sont équiprobables : $\delta = 0$ correspond à l'hypothèse nulle à tester.

La vraisemblance de δ pour la famille i est la probabilité des génotypes observés $G_i = g_i$:

$$\begin{aligned} L_i(\delta) &= \mathbb{P}_\delta(G_i = g_i) \\ &= \sum_v \mathbb{P}(G_i = g_i | v)\mathbb{P}_\delta(v) \end{aligned}$$

On peut ensuite calculer $L(\delta) = \prod_i L_i(\delta)$, et réaliser – par exemple – un test du maximum de vraisemblance :

$$2(\ell(\hat{\delta}) - \ell(0)),$$

où $\hat{\delta} = \arg\max_\delta \ell(\delta)$, suit (sous H_0) un $\chi^2(1)$.

Cependant on veut faire un test unilatéral car sous H_1, on a $\delta > 0$: on pose donc

$$Z_{\ell r} = \text{signe}(\hat{\delta})\sqrt{(\ell(\hat{\delta}) - \ell(0))}.$$

Sous H_0, on a $Z_{\ell r} \sim \mathcal{N}(0,1)$.

Lien avec Z et $\overline{Z}$

On remarque tout d'abord qu'on a $L_i(0) = \mathbb{P}_0(G_i = g_i) = \sum_v \mathbb{P}(G_i = g_i | v)c_i$: la probabilité dans le modèle où tous les vecteurs d'hérédité sont équiprobables. On a donc

$$\begin{aligned} \mathbb{P}(G_i = g_i | v) &= \frac{\mathbb{P}(v|G_i = g_i)\mathbb{P}(G_i = g_i)}{\mathbb{P}(v)} \\ &= \frac{\mathbb{P}(v|G_i = g_i)L_i(0)}{c_i} \end{aligned}$$

On a :

$$\begin{aligned} L_i(\delta) &= \sum_v \mathbb{P}(G_i = g_i | v) \times c_i \times (1 + \delta\gamma_i Z_i(v)) \\ &= \sum_v L_i(0)\mathbb{P}(v|G_i = g_i)(1 + \delta\gamma_i Z_i(v)) \\ &= L_i(0)\left(1 + \delta\gamma_i \sum_v \mathbb{P}(v|G_i = g_i)\gamma_i Z_i(v)\right) \\ &= L_i(0)\left(1 + \delta\gamma_i \overline{Z}_i\right) \end{aligned}$$

Pour finir, la log-vraisemblance pour la famille i est

$$\ell_i(\delta) = \ell_i(0) + \log(1 + \delta\gamma_i \overline{Z}_i),$$

et la log-vraisemblance pour l'ensemble des familles est

$$\ell(\delta) = \ell(0) + \sum_i \log(1 + \delta\gamma_i \overline{Z}_i).$$

Le score associé (en $\delta = 0$) est donc $U = \sum_i \gamma_i \overline{Z}_i$. Si il n'y a pas d'ambiguïté on retrouve la statistique Z, et dans le cas général on retrouve $\overline{Z}$.

Remarque finale

On a

$$J(\delta) = \sum_i \frac{\left(\gamma_i \overline{Z}_i\right)^2}{\left(1+\delta\gamma_i \overline{Z}_i\right)^2},$$

et donc $J(0) = \sum_i \left(\gamma_i \overline{Z}_i\right)^2$. Le calcul de l'information de Fisher se heurte comme précédemment au fait qu'on ne connaît pas la variance des $\overline{Z}_i$.

Une solution simple est d'utiliser l'information observée (en $\delta = 0$) à la place de l'information de Fisher, ce qui mène à la statistique de test

$$\frac{U(0)}{\sqrt{J(0}} = \frac{\sum_i \gamma_i \overline{Z}_i}{\sqrt{\sum_i \gamma_i^2 \overline{Z}_i^2}}.$$

On voit que sous H_0, le dénominateur $\sqrt{\sum_i \gamma_i^2 \overline{Z}_i^2}$ est bien un estimateur de l'écart-type de $\sum_i \gamma_i \overline{Z}_i$; et toujours sous H_0 on peut montrer que cette statistique est approximativement égale à $Z_{\ell r}$. Cependant, en présence de liaison, la statistique $Z_{\ell r}$ prend (en espérance) des valeurs plus grandes que celle-ci, et produit donc un test plus puissant.

Éléments de bibliographie

Kong A. et Cox N.J. (1997) Allele-Sharing Models : LOD Scores and Accurate Linkage Tests. *Am. J. Hum. Genet.* 1997 :61, 1179–1188.

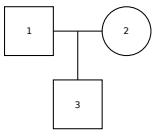
4 Algorithmes de calcul

4.1 Elston-Stewart

4.1.1 Problématique

L'algorithme d'Elston-Stewart répond au problème suivant.

— On considère un pedigree dont les individus indexés par $i = 1, \dots, n$ sont de génotype (possiblement non observé) G_i ;
— on dispose d'un ensemble $\mathscr{G}_i$ de génotypes possibles pour chaque individu i,
— on dispose d'un modèle pour les probabilités $\mathbb{P}(G = g)$ chez les individus fondateurs (on utilise généralement les proportions d'Hardy-Weinberg),
— les génotypes des autres individus ne dépendent que de ceux de leurs parents et on dispose de probabilités « de transition » entre les génotypes des enfants et ceux des parents



$\mathbb{P}(G_3 = g_3 | G_1 = g_1, G_2 = g_2)$ est connu

— on a des quantités observées Y_i pour chaque individu (des phénotypes), et un modèle faisant dépendre Y_i de G_i uniquement (conditionnellement à G_i, Y_i est indépendant de toutes les autres quantités)

$$\mathbb{P}(Y_i | Y_1, \dots, Y_{i-1}, Y_{i+1}, \dots, Y_n, G_1, \dots, G_n) = \mathbb{P}(Y_i | G_i) \quad \text{(valeur connue)}$$

Problème : calculer de façon efficace

$$\begin{aligned} &\mathbb{P}(G_1 \in \mathscr{G}_1, \dots, G_n \in \dots \mathscr{G}_n, Y_1 = y_1, \dots, Y_n = y_n) \\ &\qquad = \sum_{g_1 \in \mathscr{G}_1, \dots, g_n \in \mathscr{G}_n} \mathbb{P}(G_1 = g_1, \dots, G_n = g_n, Y_1 = y_1, \dots, Y_n = y_n). \end{aligned} \tag{4.1}$$

Notons qu'il ne s'agit pas forcément ici de géntoypes en un unique locus. On peut considérer le cas où G_i contient les génotypes en deux locus ; on peut même considérer des phénotypes phasés. Usuellement toutes les probabilités supposées connues dans la description ci-dessus peuvent dépendre de paramètres — notamment d'un taux de recombinaison entre deux marqueurs, pour les probabilités de transition de génotypes *phasés* : l'algorithme d'Elston-Stewart permet de calculer le lod-score.

4.1.2 Exemple : calcul du lod-score

Reprenons l'exemple du premier chapitre pour une analyse de liaison avec une maladie dominante, en supposant cette fois les génotypes des deux parents inconnus :

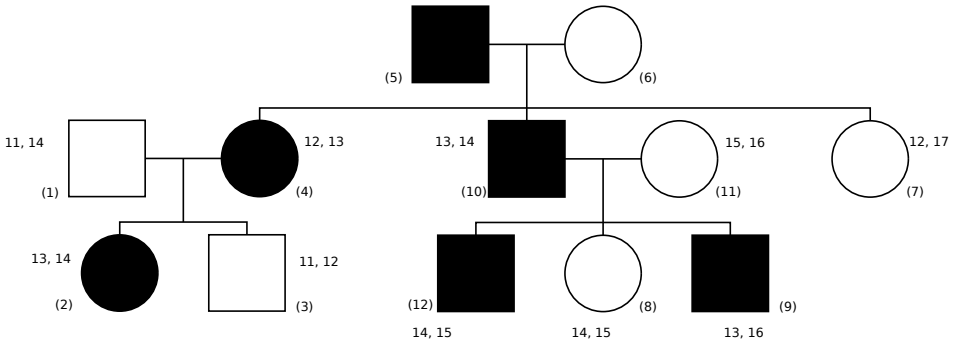


FIGURE 1 – Un calcul de lod-score

Les observations Y_i sont les phénotypes (on prendra souvent Y = 1 pour « atteint » et Y = 0 pour « sain »). On considère pour chaque individu des génotypes phasés, comprenant le marqueur considéré et le locus maladie. Ainsi G_4, le génotype de l'individu (4), peut-il être a priori un des quatre génotypes phasés

$$(12-\mathrm{A}, 13-\mathrm{A})$$
$$(12-\mathrm{A}, 13-a)$$
$$(12-a, 13-\mathrm{A})$$
$$(12-a, 13-a)$$

(si le modèle est dominant sans phénocopie on peut facilement restreindre l'éventail des possibilités à $(12-\mathrm{A}, 13-a)$ et $(12-a, 13-\mathrm{A})$). Une rapide analyse permet de restreindre l'ensemble des génotypes possibles pour les individus (5) et (6) à douze possibilités :

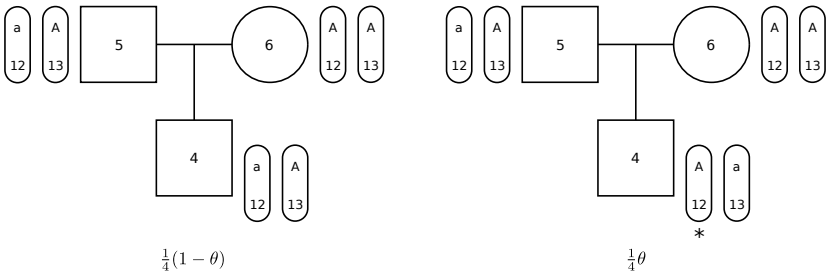
(12-A, 13-A)	(12-A, 13-a)	(12-a, 13-A)	(12-a, 13-a)
(12-A, 14-A)	(12-A, 14-a)	(12-a, 14-A)	(12-a, 14-a)
(13-A, 17-A)	(13-A, 17-a)	(13-a, 17-A)	(13-a, 17-a)

On peut encore diminuer le nombre de possibilités pour chacun des individus, en considérant le phénotype des deux individus (toujours dans un modèle dominant sans phénocopie).

Les probabilités de transition dépendent du taux de recombinaison θ entre le marqueur et le locus maladie. Ainsi par exemple

$$\mathbb{P}\big(G_4 = (12-a, 13-\mathrm{A}) \big| G_5 = (12-a, 13-\mathrm{A}), G_6 = (12-\mathrm{A}, 13-\mathrm{A})\big) = \frac{1}{4}(1-\theta)$$
$$\mathbb{P}\big(G_4 = (12-\mathrm{A}, 13-a) \big| G_5 = (12-a, 13-\mathrm{A}), G_6 = (12-\mathrm{A}, 13-\mathrm{A})\big) = \frac{1}{4}\theta$$

etc



Bien sûr certaines probabilités de transition seront nulles, ce qui traduit l'incompatibilité des combinaisons de génotypes envisagées.

Les probabilités d'« émission » sont faciles à écrire pour une maladie dominante (ci-dessous $*$ est un joker) :

$$\begin{aligned}
\mathbb{P}(\mathrm{Y}=0|\mathrm{G}=(*-\mathrm{A},*-\mathrm{A}))&=1\\
\mathbb{P}(\mathrm{Y}=1|\mathrm{G}=(*-\mathrm{A},*-\mathrm{A}))&=0\\
\mathbb{P}(\mathrm{Y}=0|\mathrm{G}=(*-*,*-a))&=0\\
\mathbb{P}(\mathrm{Y}=1|\mathrm{G}=(*-*,*-a))&=1
\end{aligned}$$

Il est facile de généraliser au cas de la recherche d'un gène majeur (possibilité de pénétrance incomplète, ou de phénocopies).

4.1.3 Algorithme

L'agorithme récursif « naturel » (mais inefficace) consiste à « effeuiller » le pedigree : on choisit une feuille, c'est-à-dire un individu sans descencance — par exemple, pour le pedigree de la figure 1, l'individu (12). On utilise le fait que G_{12} ne dépend que des génotypes parentaux G_{10} et G_{11}, et que Y_{12} ne dépend que de G_{12}, pour écrire

$$\begin{aligned}
&\mathbb{P}(\mathrm{G}_1=g_1,\dots,\mathrm{G}_{12}=g_{12},\mathrm{Y}_1=y_1,\dots,\mathrm{Y}_{12}=y_{12})\\
&=\mathbb{P}(\mathrm{G}_1=g_1,\dots,\mathrm{G}_{11}=g_{11},\mathrm{Y}_1=y_1,\dots,\mathrm{Y}_{11}=y_{11})\times\mathbb{P}(\mathrm{G}_{12}=g_{12}|\mathrm{G}_{10}=g_{10},\mathrm{G}_{11}=g_{11})\times\mathbb{P}(\mathrm{Y}_{12}=y_{12}|\mathrm{G}_{12}=g_{12})
\end{aligned}$$

On peut alors réorganiser la somme de l'équation 4.1 ainsi :

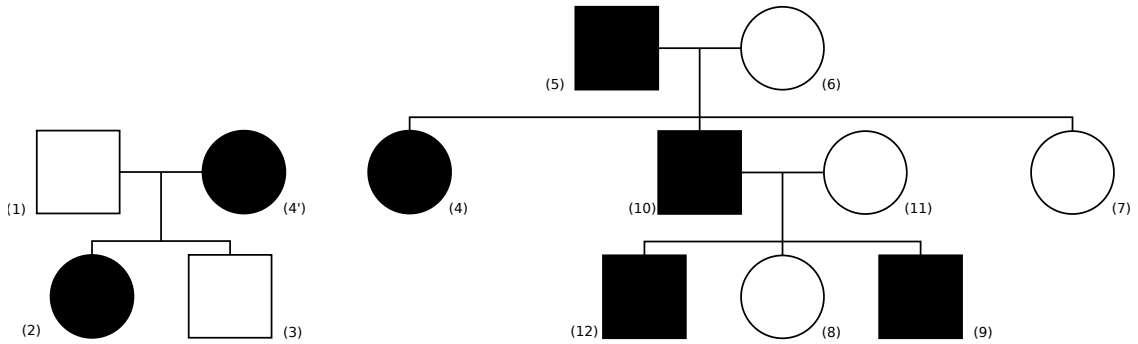
$$\begin{aligned}
&\sum_{g_1,\dots,g_{12}}\mathbb{P}(\mathrm{G}_1=g_1,\dots,\mathrm{G}_{12}=g_{12},\mathrm{Y}_1=y_1,\dots,\mathrm{Y}_{12}=y_{12})\\
&\qquad=\sum_{g_{10},g_{11}}\left[\left(\sum_{g_1,\dots,g_9}\mathbb{P}(\mathrm{G}_1=g_1,\dots,\mathrm{G}_{11}=g_{11},\mathrm{Y}_1=y_1,\dots,\mathrm{Y}_{11}=y_{11})\right)\right.\\
&\qquad\qquad\left.\times\sum_{g_{12}}\left(\mathbb{P}(\mathrm{G}_{12}=g_{12}|\mathrm{G}_{10}=g_{10},\mathrm{G}_{11}=g_{11})\times\mathbb{P}(\mathrm{Y}_{12}=y_{12}|\mathrm{G}_{12}=g_{12})\right)\right],
\end{aligned}$$

c'est-à-dire qu'on énumère tous les génotypes possibles pour G_{10} et G_{11}, et pour chacune de ces possibilités on s'est ramené d'une part à un calcul sur le trio (10) (11) (12) et d'autre part à un calcul sur un pedigree plus petit (un individu de moins) qu'on peut traiter de la même façon ; on recommence jusqu'à s'être ramené à un trio.

Ce que cet algorithme réalise en fait est l'énumération de toutes les possibilités. Bien qu'il soit relativement facile à programmer, il sera en pratique inefficace. L'astuce à la base de l'algorithme d'Elston-Stewart consiste à couper le pedigree en deux sur la base d'un individu pivot : un individu ayant des parents *et* une descendance. Dans notre exemple, on a deux pivots possibles, les individus (4) et (10). Choisissons (4). Si on raisonne conditionnellement à $\mathrm{G}_4=g_4$, les individus (1), (2) et 3) d'une part, et les individus (5) à (12) d'autre part, deviennent indépendants — tout en dépendant de G_4. On peut réorganiser l'équation 4.1 ainsi :

$$\begin{aligned}
&\sum_{g_1,\dots,g_{12}}\mathbb{P}(\mathrm{G}_1=g_1,\dots,\mathrm{G}_{12}=g_{12},\mathrm{Y}_1=y_1,\dots,\mathrm{Y}_{12}=y_{12})\\
&\qquad=\sum_{g_4}\left(\sum_{g_1,g_2,g_3}\mathbb{P}(\mathrm{G}_1=g_1,\mathrm{G}_2=g_2,\mathrm{G}_3=g_3,\mathrm{Y}_1=y_1,\mathrm{Y}_2=y_2,\mathrm{Y}_3=y_3|\mathrm{G}_4=g_4)\times\right.\\
&\qquad\qquad\left.\sum_{g_5,\dots,g_{12}}\mathbb{P}(\mathrm{G}_4=g_4,\dots,\mathrm{G}_{12}=g_{12},\mathrm{Y}_4=y_4,\dots,\mathrm{Y}_{12}=y_{12})\right)
\end{aligned}$$

Pour obtenir un algorithme récursif, on remarque que cela revient à dupliquer l'individu (4) ainsi :



et à écrire

$$\sum_{g_1,\dots,g_{12}} \mathbb{P}(\mathrm{G}_1 = g_1,\dots,\mathrm{G}_{12} = g_{12},\mathrm{Y}_1 = y_1,\dots,\mathrm{Y}_{12} = y_{12})$$
$$= \sum_{g_4} \frac{1}{\mathbb{P}(\mathrm{G}_{4'} = g_4)} \left(\sum_{g_1,g_2,g_3} \mathbb{P}(\mathrm{G}_1 = g_1,\dots,\mathrm{G}_{4'} = g_4,\mathrm{Y}_1 = y_1,\dots,\mathrm{Y}_4 = y_4) \times \right.$$
$$\left. \sum_{g_5,\dots,g_{12}} \mathbb{P}(\mathrm{G}_4 = g_4,\dots,\mathrm{G}_{12} = g_{12},\mathrm{Y}_4 = y_4,\dots,\mathrm{Y}_{12} = y_{12}) \right).$$

On s'est ramené au calcul des termes $\sum_{g_1,g_2,g_3} \mathbb{P}(\mathrm{G}_1 = g_1,\dots,\mathrm{G}_{4'} = g_4,\mathrm{Y}_1 = y_1,\dots,\mathrm{Y}_4 = y_4)$ et $\mathbb{P}(\mathrm{G}_4 = g_4,\dots,\mathrm{G}_{12} = g_{12},\mathrm{Y}_4 = y_4,\dots,\mathrm{Y}_{12} = y_{12})$ qui est un problème analogue, mais de complexité moindre. La méthode s'écrit de façon naturelle comme un algorithme récursif. Notons que dans une famille nucléaire, il n'y a pas de pivot ; on peut traiter leur cas de façon élémentaire, par exemple en énumérant les génotypes des fondateurs.

4.2 Lander-Green

La méthode de Lander-Green tire profit d'une méthodologie pré-existante très générale, qui s'applique aux modèles « à chaîne de Markov cachée ». Les algorithmes développés pour ces modèles permettent dans notre cas de calculer les probabilités des différents vecteurs d'hérédité en tout point du génome, conditionnellement aux génotypes observés. Il est ensuite facile de calculer une des statistiques de l'analyse de liaison, puisqu'elles ne dépendent que du vecteur d'hérédité.

4.2.1 Modèles à chaîne de Markov cachée

Chaîne de Markov

On dit que des variables aléatoires $(\mathrm{S}_1,\mathrm{S}_2,\dots)$ forment une chaîne de Markov si la pour tout i, la loi de S_i conditionnellement à $(\mathrm{S}_1,\dots,\mathrm{S}_{i-1})$ vérifie

$$\mathbb{P}(\mathrm{S}_i = s_i|\mathrm{S}_1 = s_1,\dots,\mathrm{S}_{i-1} = s_{i-1}) = \mathbb{P}(\mathrm{S}_i = s_i|\mathrm{S}_{i-1} = s_{i-1}).$$

Il suffit de connaître la valeur prise par S_{i-1} pour connaître toute l'information apportée sur S_i par l'ensemble des valeurs prises par les S_j $(j < i)$.

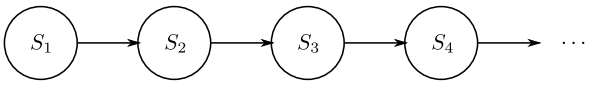


FIGURE 2 – Représentation schématique d'une chaîne de Markov

Exemple 1 (état IBD de deux germains) On considère une paire de germains, et des locus successifs sur le génome indicés par $i = 1, 2, \ldots$. Le taux de recombinaison entre le locus $i-1$ et le locus i est θ_i (pour $i \geqslant 2$). On note $S_i = 0, 1, 2$ leur état IBD ; les S_i forment une chaîne de Markov.

Un changement d'état IBD entre deux locus successifs se produit dès qu'une recombinaison a eu lieu chez (au moins) un des parents, ce qui arrive chez chacun d'eux avec probabilité θ_i. Ainsi, si on a l'état IBD = 0, la probabilité de ne pas changer d'état est $(1-\theta_i)^2$ (pas de recombinaison), celle de passer à IBD = 1 est $2\theta_i(1-\theta_i)$ (une recombinaison chez un parent ou l'autre), et celle de passer à IBD = 2 est θ_i^2 (deux recombinaisons, une chez chacun des parents).

$$\begin{array}{lll}
\mathbb{P}(X_i = 0|X_{i-1} = 0) = (1-\theta_i)^2 & \mathbb{P}(X_i = 1|X_{i-1} = 0) = 2\theta_i(1-\theta_i) & \mathbb{P}(X_i = 2|X_{i-1} = 0) = \theta_i^2 \\
\mathbb{P}(X_i = 0|X_{i-1} = 1) = \theta_i(1-\theta_i) & \mathbb{P}(X_i = 1|X_{i-1} = 1) = (1-\theta_i)^2 + \theta_i^2 & \mathbb{P}(X_i = 2|X_{i-1} = 1) = \theta_i(1-\theta_i) \\
\mathbb{P}(X_i = 0|X_{i-1} = 2) = \theta_i^2 & \mathbb{P}(X_i = 1|X_{i-1} = 2) = 2\theta_i(1-\theta_i) & \mathbb{P}(X_i = 2|X_{i-1} = 2) = (1-\theta_i)^2
\end{array}$$

Notons que si $\mathbb{P}(X_i = 0) = \frac{1}{4}$, $\mathbb{P}(X_i = 1) = \frac{1}{2}$ et $\mathbb{P}(X_i = 2) = \frac{1}{4}$, on retrouve à partir de ces probabilités de transition $\mathbb{P}(X_{i+1} = 0) = \frac{1}{4}$, $\mathbb{P}(X_{i+1} = 1) = \frac{1}{2}$ et $\mathbb{P}(X_{i+1} = 2) = \frac{1}{4}$. On pourra prendre $\mathbb{P}(X_0 = 0) = \frac{1}{4}$, $\mathbb{P}(X_0 = 1) = \frac{1}{2}$ et $\mathbb{P}(X_0 = 2) = \frac{1}{4}$, et la loi marginale de tous les X_i sera identique, la chaîne de Markov servant à spécifier la façon dont les états IBD sont corrélés en divers points du génome. On peut également montrer que, même si on prend une autre loi pour X_0 (par exemple, $\mathbb{P}(X_0 = 0) = 1$ qui correspond à la certitude d'un état IBD nul au début d'un chromosome), pour i suffisament grand (dès qu'on est assez loin du point initial), on a $\mathbb{P}(X_i = 0) \simeq \frac{1}{4}$, $\mathbb{P}(X_i = 1) \simeq \frac{1}{2}$ et $\mathbb{P}(X_i = 2) \simeq \frac{1}{4}$.

Exemple 2 (vecteur d'hérédité) On considère un pedigree et son vecteur d'hérédité (condensé) qui a été défini au chapitre précédent. Comme dans l'exemple précédent, on a des locus successifs indicés par $i = 1, 2, \ldots$, avec un taux de recombinaison entre les locus $i-1$ et i égal à θ_i. On prend pour S_i le vecteur d'hérédité au locus i ; comme à l'exemple précédent, les S_i forment une chaîne de Markov.

Considérons par exemple ce pedigree avec $f = 3$ fondateurs et $n = 3$ non-fondateurs.

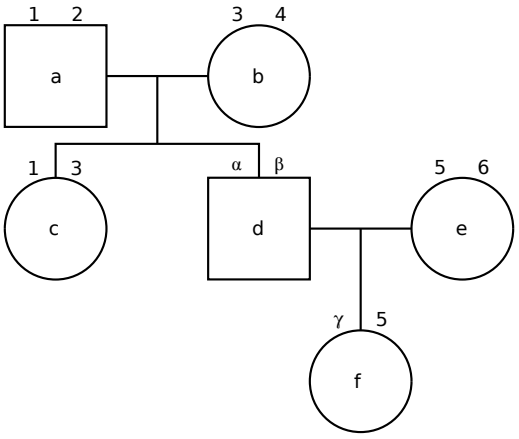


FIGURE 3 – Un pedigree avec $n = 3$ et $f = 3$.

On va considérer les vecteurs d'hérédité condensés, en ordonnant en tout point les allèles des fondateurs en plaçant en premier l'allèle transmis au premier descendant. On doit ainsi considérer des vecteurs à $2n - f = 3$, composantes :

- la première composante vaut 0 si $\alpha = 1$ (les individus c et d reçoivent le même allèle de leur père), 1 si $\alpha = 2$ (allèles paternels différents)
- la deuxième composante vaut 0 si $\beta = 3$ (les individus c et d reçoivent le même allèle de leur mère), 1 si $\beta = 4$ (allèles maternels différents)
- la troisième composante vaut 0 si $\gamma = \alpha$ (d transmet à f l'allèle qu'il a reçu de a) et 1 si $\gamma = \beta$ (d transmet à f l'allèle qu'il a reçu de b)

Quand on passe d'un locus $i-1$ au locus suivant $i-1$, ce qui détermine si on a un changement de la première composante c'est la présence ou non de recombinaisons dans les transmissions de a à c et d ; pour la

deuxième composantes, ce sont les transmissions de b et à c et d qu'il faut considérer ; et pour la troisième composante c'est la transmission de d à f qui importe seule.

Ainsi pour la première composante, on a une transition $0 \to 0$ avec probabilité $(1-\theta_i)^2+\theta_i^2$ (aucune recombinaison, ou bien recombinaison à la fois dans la transmission de a à c et dans la transmission de a à d), une transition $0 \to 1$ avec probabilité $2\theta_i(1-\theta_i)$ (recombinaison dans une ou l'autre transmission), et de même une transition $1 \to 0$ a une probabilité $2\theta_i(1-\theta_i)$ et $1 \to 1$ une probabilité $(1-\theta_i)^2+\theta_i^2$.

L'analyse est la même pour la seconde composante ; pour la troisième composante, on a plus simplement une transition $0 \to 0$ avec probabilité $(1-\theta_i)$, une transition $0 \to 1$ avec probabilité θ_i, une transition $1 \to 0$ avec probabilité θ_i, une transition $1 \to 1$ avec probabilité $(1-\theta_i)$.

Les recombinaisons considérées sont indépendantes les unes des autres, ce qui permet d'écrire toutes les probabilités de transition d'un vecteur à l'autre à partir des probabilités données ci-dessus. Par exemple

$$\begin{aligned}
\mathbb{P}(\mathrm{S}_i = 000 | \mathrm{S}_{i-1} = 000) &= \left((1-\theta_i)^2+\theta_i^2\right) \times \left((1-\theta_i)^2+\theta_i^2\right) \times (1-\theta_i) \\
\mathbb{P}(\mathrm{S}_i = 001 | \mathrm{S}_{i-1} = 000) &= \left((1-\theta_i)^2+\theta_i^2\right) \times \left((1-\theta_i)^2+\theta_i^2\right) \times \theta_i \\
\mathbb{P}(\mathrm{S}_i = 010 | \mathrm{S}_{i-1} = 000) &= \left((1-\theta_i)^2+\theta_i^2\right) \times (2\theta_i(1-\theta_i)) \times (1-\theta_i) \\
\mathbb{P}(\mathrm{S}_i = 011 | \mathrm{S}_{i-1} = 000) &= \left((1-\theta_i)^2+\theta_i^2\right) \times (2\theta_i(1-\theta_i)) \times \theta_i \\
&\vdots
\end{aligned}$$

Chaîne de Markov cachée

On dit que des variables aléatoires $(\mathrm{S}_1, \mathrm{S}_2, \ldots)$ et $(\mathrm{Y}_1, \mathrm{Y}_2, \ldots)$ forment une chaîne de Markov si $(\mathrm{S}_1, \mathrm{S}_2, \ldots)$ forme une chaîne de Markov, et si pour tout i la loi de Y_i conditionnellement à toutes les autres variables vérifie

$$\mathbb{P}(\mathrm{S}_i = s_i | \mathrm{S}_1 = s_1, \mathrm{Y}_1 = y_1, \ldots, \mathrm{S}_{i-1} = s_{i-1}, \mathrm{Y}_{i-1} = y_{i-1}, \mathrm{S}_i = s_i, \mathrm{S}_{i+1} = s_{i+1}, \mathrm{Y}_{i+1} = y_i, \ldots) = \mathbb{P}(\mathrm{Y}_i = y_i | \mathrm{S}_i = s_i).$$

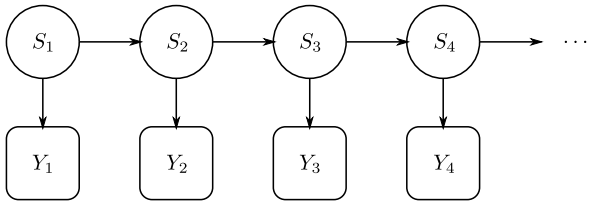


FIGURE 4 – Représentation schématique d'une chaîne de Markov cachée

On appelle les S_i les états cachés du modèle.

Exemple 1 (état IBD) On peut reprendre l'exemple d'une paire de germains, avec S_i l'état IBD en un point du génome : Y_i sera le vecteur formé par les génotypes des deux germains. En pratique, seul Y_i est observé, et on peut (plus ou moins facilement) inférer S_i à partir des Y_i. Pour que Y_i ne dépende que de S_i, il faut que les marqueurs considérés ne soient pas en déséquilibre de liaison.

Exemple 2 (vecteur d'hérédité) On peut de même reprendre l'exemple d'un pedigree étendu ; les états cachés S_i sont les vecteurs d'hérédités, et les Y_i sont les vecteurs de génotypes. Ici aussi il est crucial que les marqueurs considérés ne soient pas en déséquilibre de liaison.

Algorithme de Baum-Welch

Des méthodes spécifiques ont été créées pour l'étude des modèles à chaîne de Markov cachée. L'algorithme de Baum-Welsch permet notamment de calculer la vraisemblance d'observations $\mathrm{Y}_1 = y_1, \ldots, \mathrm{Y}_\mathrm{N} = y_\mathrm{N}$, de l'optimiser au besoin en certains paramètres inconnus, et de calculer les probabilités a posteriori des états cachés,

$$\mathbb{P}(\mathrm{S}_i = s_i | \mathrm{Y}_1 = y_1, \ldots, \mathrm{Y}_\mathrm{N} = y_\mathrm{N}).$$

4.2.2 L'analyse de liaison

Pour l'analyse de liaison, on reprendra l'exemple 2 donné ci-dessous : on peut calculer la probabilité des différents vecteur d'hérédité possibles en une série de marqueurs couvrant l'ensemble du génome, conditionnellement aux génotypes observés. Il est important que ces marqueurs ne soient pas en déséquilibre de liaison !

La statistique de liaison doit ensuite être calculée à partir des vecteurs d'hérédité.

On peut même calculer les probabilités des vecteurs d'hérédité en des points intermédiaires entre les marqueurs pour avoir une statistique de liaison en tout point du génome.

Analyse de liaison « paramétrique »

Alors qu'il est très complexe de prendre en compte plusieurs marqueurs simultanément dans l'algorithme d'Elston-Stewart pour calculer un lod-score « multipoint », cela devient beaucoup plus simple par cette méthode. Tous les marqueurs d'un chromosomes sont pris en compte pour le calcul des probabilités des vecteurs d'hérédité.

On calculera le lod-score en tout point d'un chromosome, en supposant $\theta = 0$ (le locus maladie est au point considéré). On calcule donc simplement la probabilité des phénotypes observés sous l'hypothèses où le locus maladie est au locus considéré

$$L(0) = \sum_v \mathbb{P}(\text{phénotypes}|v)\mathbb{P}(v|Y = y)$$

(la probabilité des vecteurs d'hérédité est conditionnée aux phénotypes observés) et leur probabilité sous l'hypothèse où le locus maladie est en un point quelconque du génome

$$L(0.5) = \sum_v \mathbb{P}(\text{phénotypes}|v)\mathbb{P}(v)$$

(on prend tous les vecteurs d'hérédité équiprobables), et le lod-score est

$$\log L(0) - \log L(0.5).$$

Analyse de liaison « non-paramétrique »

Le calcul du ZLR se fait de façon directe quand on connaît la probabilité des différents vecteurs d'hérédité, grâce à la formule (cf chapitre précédent)

$$\ell(\delta) = \text{constante} + \sum_i \log(1 + \delta\gamma_i\overline{Z}_i);$$

où $\overline{Z}_i$ est la statistique moyenne dans la famille i (on peut écrire $\overline{Z}_i = \sum \mathbb{P}(v|Y = y)Z_i(v)$).

4.3 Avantages respectifs des algorithmes et difficultés diverses

L'algorithme d'Elston-Stewart permet de traiter des familles de grande taille, mais faire de l'analyse de liaison multipoint est très complexe ; on ne peut pas utiliser plus d'une poignée de marqueurs à la fois. Le logiciel Fastlink semble être une bonne référence (dernière mise à jour en 1999 !) : https://www.ncbi.nlm.nih.gov/CBBresearch/Schaffer/fastlink.html

La méthode de Lander-Green devient très lourde à mettre en œuvre quand la taille $2n-f$ des vecteurs d'hérédité devient trop grande (le nombre de vecteurs à envisager étant 2^{2n-f}) ; il n'y a par contre pas de problème pour traiter beaucoup de marqueurs. Il existe des méthodes pour « découper » les grands pedigrees afin de les analyser morceau par morceau par la méthode de Lander-Green.

En TD, vous utiliserez Merlin http://csg.sph.umich.edu/abecasis/Merlin, qui implémente la méthode de Lander-Green.

Nous avons signalé le problème potentiel causé par la présence de déséquilibre de liaison entre les marqueurs. À l'époque héroïque de l'analyse de liaison on utilisait au grand maximum de l'ordre d'un millier de microsatellites répartis sur le génome, sans déséquilibre de liaison. Aujourd'hui on préférerait utiliser des cartes denses de SNPs ! Merlin propose une solution, en créant des « super-marqueurs » avec des haplotypes de SNPs en déséquilibre de liaison.

D'autres méthodes et logiciels plus récents ont été proposés. Nous citerons simplement, pour l'analyse de liaison paramétrique, Superlink et Superlink-online (Fisehlson et Geiger 2002) et Morgan (Thompson et Wijsman, 1993 à 2011) (méthode basée sur une approximation stochastique).