Détails techniques des développements mathématiques de ma recherche

Charles-Elie Rabier

charles-elie.rabier@umontpellier.fr
http://charles-elie.rabier.pagesperso-orange.fr

Table des matières

1 Recherches

L'afflux d'informations moléculaires résultant notamment des nouvelles technologies de séquençage du génome est considérable : l'analyse des données relatives au génome requiert de mettre en oeuvre des techniques mathématiques issues du champ des statistiques et des probabilités. Mes recherches contribuent à proposer des outils mathématiques permettant d'extraire des informations pertinentes de ces données. Mon travail de thèse consistait à étudier et proposer des techniques statistiques propres à la détection et à la localisation de loci (i.e. emplacements physiques précis sur un chromosome) responsables de la variation d'un caractère quantitatif. On nomme ces loci, Quantitative Trait Loci (QTL). D'un point de vue théorique, la détection de QTL nécessite de travailler sur les processus Gaussiens et de Chi-Deux, les modèles de mélanges, et les plans d'expérience. Par la suite, lors de mon postdoctorat à l'université de Madison, WI, USA, je me suis intéressé aux techniques probabilistes en matière d'évolution. J'ai étudié en particulier les arbres aléatoires (processus de naissances et de morts), la reconciliation entre les arbres d'espèces et les arbres de gènes ainsi que les duplications entières du génome (WGDs). Nous sommes amenés à effectuer une optimisation stochastique par MCMC, en raison de l'immensité de l'espace des paramètres : à l'aide d'algorithmes permettant de proposer de nouveaux arbres, on échantillonne la distribution à posteriori de l'arbre de gènes sachant les séquences d'ADN.

Lors de ma deuxième expérience postdoctorale, j'ai été impliqué dans un projet portant sur la sélection génomique et la statistique en grande dimension. On ne se focalise plus sur les associations, l'objectif étant désormais d'effectuer des prédictions à l'aide de milliers de marqueurs. J'ai étudié en particulier les propriétés mathématiques de la prédiction dans le cadre de la régression Ridge avec un design aléatoire. Les applications s'appuyaient sur de nouvelles espèces séquencées (blé, raygrass). A l'université de Montpellier, je travaille sur les réseaux phylogénétiques dans le cadre du modèle de coalescence. Les applications portent sur le génome du riz qui a connu des étapes d'hybridation dans son histoire évolutive.

Statistique des processus / Détection de QTL

Je me suis tout d'abord intéressé à une technique d'amélioration de protocole : le Selective Genotyping. Puis, j'ai étudié le Génome Scan où l'on recherche des QTLs en balayant le génome. Enfin, plus récemment, j'ai combiné ces 2 thématiques en considérant le problème du Génome Scan en présence d'un Selective Genotyping.

Selective Genotyping

Dans cette étude, on se place uniquement en un point précis du génome : sur un marqueur génétique. De plus, on supposera que le QTL se trouve sur ce marqueur.

X désigne la variable aléatoire correspondant au génotype au QTL (i.e. au marqueur). On considère 2 génotypes au QTL : $+1$ avec probabilité p, et -1 avec probabilité $1-p$. Le caractère quantitatif (i.e. phénotype) Y vérifie : $Y = \mu + qX + \varepsilon$ où ε suit une loi normale de moyenne nulle et de variance σ^2. Un QTL est présent si et seulement si l'effet QTL q est différent de zéro. On considère un échantillon de n observations i.i.d.

La problème est le suivant : le génotypage, i.e. recueillir l'information marqueur X pour l'ensemble des individus, s'avérait très onéreux dans le passé. Dans un tel contexte, Lebowitz et al. (1987) proposèrent de génotyper uniquement les individus présentant un phénotype extrême (i.e. les plus petits et les plus grands Y), car ils remarquèrent que la plupart de l'information au sujet du QTL était présente dans les phénotypes extrêmes. Ainsi, pour une puissance donnée, une grande augmentation du nombre d'individus conduit à une diminution du nombre d'individus génotypés. En d'autre termes, le Selective Genotyping permet de réduire les coûts dus au génotypage tout en gardant une bonne puissance pour le test statistique, à condition que le nombre d'individus au total ait été augmenté. Par la

suite, Lander et Botstein (1989) formalisèrent cette approche et la nommèrent “Selective Genotyping". Aujourd'hui, bien que les coûts dus au génotypage aient largement diminué, le Selective Genotyping est encore un concept très intéressant car nous pouvons améliorer le processus de détection en se focalisant uniquement sur les individus extrêmes au lieu d'individus “aléatoires". La particularité du Selective Genotyping réside dans le fait que la théorie usuelle n'est pas applicable : le modèle n'est plus un “modèle d'analyse de variance", en raison des génotypes manquants. De plus, les phénotypes extrêmes correspondant à un génotype donné ne suivent pas une loi normale classique mais une loi normale tronquée.

Afin de modéliser le Selective Genotyping, on considère deux seuils fixes S_- et S_+. Pour un individu donné, on a connaissance de son phénotype Y mais on dispose de son génotype X uniquement si $Y \notin [S_-,\ S_+]$. Dans mon étude, je m'intéresse principalement à l'inférence statistique en Selective Genotyping. Je propose différents tests statistiques appropriés pour le Selective Genotyping et donne leurs distributions asymptotiques sous l'hypothèse nulle ($q = 0$) et sous des alternatives locales ($q = a/\sqrt{n}$), grâce notamment aux travaux de Le Cam. Leurs performances sont comparées en terme d'efficacité asymptotique relative (efficacité de Pitman). Je prouve que les phénotypes non extrêmes (i.e. les phénotypes pour lesquels les génotypes sont manquants) n'apportent aucune information pour l'inférence statistique. Je démontre également que nous devons génotyper symétriquement, c'est à dire le même pourcentage de grands et de petits phénotypes quelles que soient les proportions des deux génotypes dans la population. Des résultats similaires sont également obtenus dans le cadre d'un Selective Genotyping avec deux phénotypes corrélés. Pour finir, je prouve que la proportion optimale d'individus à génotyper est très sensible au ratio des coûts (i.e. coût dû au génotypage divisé par le coût dû au phénotypage). Cette étude du Selective Genotyping s'est traduit par une publication dans Journal of Statistical Planning and Inference (2014).

Génome Scan sous le modèle de Haldane

L'étude porte sur une population backcross : $A \times (A \times B)$, où A et B sont de pures lignées homozygotes. La problématique se veut la détection d'un QTL sur un chromosome. On ne se place donc plus comme précédemment en un point précis du génome. Le chromosome est représenté par un segment $[0, T]$. La distance sur $[0, T]$ est appelée distance génétique et se mesure en Morgans. Les mécanismes de la génétique, en particulier de la méïose, impliquent que parmi les deux chromosomes de chaque individu, l'un est purement hérité de A, alors que l'autre (le “recombinant") consiste en parties provenant de A et de B, en raison du phénomène de recombinaisons. Le génome $X(t)$ d'un individu prend la valeur $+1$ si, par exemple, le “chromosome recombinant" provient de A à l'emplacement t et prend la valeur -1 s'il provient de B. On utilise la modélisation de Haldane (1919) qui se représente de la manière suivante : $X(0)$ est un signe aléatoire et $X(t) = X(0)(-1)^{N(t)}$ où $N(.)$ est un processus de Poisson standard sur $[0, T]$. La modélisation de Haldane suppose l'absence d'interférence entre les recombinaisons, et le processus de Poisson $N(.)$ représente le nombre de recombinaisons qui se produisent durant la méïose.

On considère le modèle suivant pour le caractère quantitatif :

$$Y = \mu \ + \ X(t^\star)\, q \ + \ \sigma\varepsilon \tag{1}$$

où ε est un bruit blanc Gaussien et $t^\star$ est la vraie position du QTL.

En fait, l'information génome s'avère disponible uniquement à certaines positions données, $t_1 = 0 < t_2 < ... < t_K = T$ appelées marqueurs génétiques, et on observe n observations $(Y_j,\ X_j(t_1),\ ...,\ X_j(t_K))$ i.i.d.

Le point clé est dans le fait que conditionnellement à $X(t_1), \ldots, X(t_K)$, Y obéit au modèle de mélange aux poids connus :

$$p(t^\star) f_{(\mu+q,\sigma)}(Y) + \{1 - p(t^\star)\}\, f_{(\mu-q,\sigma)}(Y), \tag{2}$$

où $f_{(m,\sigma)}$ est la densité Gaussienne de paramètres (m,σ) et où la fonction $p(t)$ est la probabilité $\mathbb{P}(X(t)=1)$ conditionnellement aux marqueurs. Elle s'exprime en fonction de Haldane (1919).

Le challenge est que la vraie position $t^\star$ est inconnue. Si $t^\star = t$ était connue, le modèle serait un modèle régulier. Si on définit respectivement $\Lambda_n(t)$ et $S_n(t)$ comme le Test du Rapport de Vraisemblance (LRT) et la statistique du score correspondant à l'hypothèse nulle "q=0", il est bien connu que

$$\Lambda_n(t) = \ S_n^2(t) + o_P(1)$$

et que $S_n(t)$ est asymptotiquement Gaussienne. Lorsque $t^\star$ est inconnu, le maximum de $\Lambda_n(t)$ donne encore le LRT de "q=0". $\arg\sup_t \Lambda_n(t)$ s'avère dès lors un estimateur naturel de la position du QTL. Le choix comme statistique de test du maximum de ce processus, revient à effectuer un LRT dans un modèle où la localisation du QTL est un paramètre supplémentaire.

Dans notre étude, nous donnons la distribution asymptotique exacte de la statistique du LRT sous l'hypothèse nulle d'absence de QTL sur $[0,T]$ $(q=0)$ et sous l'alternative qu'il existe un QTL en $t^\star \in [0,T]$. Ces distributions ont été obtenues approximativement par Cierco (1998), Azaïs et Cierco-Ayrolles (2002), Azaïs et Wschebor (2009). Dans Rebaï et al (1995, 1994) et Chang et al. (2009), les auteurs étudient uniquement l'hypothèse nulle en utilisant des approximations. Des résultats théoriques sont présents également dans Chen et Chen (2005) sous des alternatives non contigües. Nous montrons que la distribution de la statistique de LRT est asymptotiquement celle du maximum du carré d'un "processus d'interpolation non linéaire renormalisé". La preuve est basée sur des résultats récents de Gassiat (2002) et Azaïs et al (2009), qui s'appuient sur la théorie des processus empiriques. Le second résultat important est que nous disposons d'une formule simple pour la distribution du maximum du carré du "processus d'interpolation non linéaire". Pour finir, on propose une nouvelle méthode appropriée quelle que soit la carte génétique, et utilisant des méthodes de Monte-Carlo Quasi Monte-Carlo (Genz, 1992), afin de calculer les seuils pour la détection de QTL. Nous montrons que notre méthode donne de meilleures performances que la méthode de Rebaï (1994) basée sur Davies (1977, 1987), et que la méthode de Feingold (1993) basée sur Siegmund (1985). A titre d'exemple, après avoir effectué 657 tests statistiques sur le génome, le pourcentage de Faux Positifs correspondant à notre méthode, s'avére proche des 5% attendus.

Notre méthode est distribuée au travers d'un package Matlab avec interface graphique (disponible sur http ://charles-elie.rabier.pagesperso-orange.fr/doc/articles.html). Ce travail a donné lieu à une publication avec J-M. Azaïs et C. Delmas dans Statistics (2012).

A propos du maximum du processus

Soit $Z(.)$ le processus Gaussien d'interpolation non linéaire renormalisé. Décrivons ici, notre nouvelle méthode pour le calcul du α% quantile du maximum du processus $Z^2(.)$, sous H_0. Si on note

$$\begin{aligned}
&\mathrm{Cov}\left(\mathrm{Z(t_k)}, \mathrm{Z(t_{k+1})}\right) = \rho(\mathrm{t_k}, \mathrm{t_{k+1}}) \\
&h(t_k, t_{k+1}) = \frac{Z^2(t_k) + Z^2(t_{k+1}) - 2\rho(t_k,t_{k+1})Z(t_k)Z(t_{k+1})}{1-\rho^2(t_k,t_{k+1})} 1_{\frac{Z(t_{k+1})}{Z(t_k)} \in]\rho(t_k,t_{k+1}), \frac{1}{\rho(t_k,t_{k+1})}[}
\end{aligned}$$

nous démontrons que nous avons la relation suivante

$$\sup_{t\in[0,T]} Z^2(t) = \max\left\{Z^2(t_1), Z^2(t_2), h(t_1,t_2), ..., Z^2(t_{K-1}), Z^2(t_K), h(t_{K-1},t_K)\right\}.$$

Ainsi, nous avons juste à nous focaliser sur la distribution de M où

$$M = \max\left\{Z^2(t_1), Z^2(t_2), h(t_1,t_2), ..., Z^2(t_{K-1}), Z^2(t_K), h(t_{K-1},t_K)\right\}$$

. On a $\forall c \in \mathbb{R}$

$$\mathbb{P}(0 \leq M \leq c^2) = \mathbb{P}\left\{-c \leq Z(t_1) \leq c, ..., -c \leq Z(t_K) \leq c\right\} \times$$
$$\mathbb{P}\left\{0 \leq h(t_1, t_2) \leq c^2, ..., 0 \leq h(t_{K-1}, t_K) \leq c^2 \mid -c \leq Z(t_1) \leq c, ..., -c \leq Z(t_K) \leq c\right\}.$$

Le premier terme est une intégrale sur la densité d'un vecteur Gaussien de dimension K. Il peut être calculé pour de grandes valeurs de K, en utilisant la fonction QSIMVNEF d'Alan Genz, qui est un programme Monte-Carlo Quasi Monte-Carlo (MCQMC). QSIMVNEF permet aussi le calcul du second terme. Les méthodes MCQMC de Genz (1992) sont très rapides. Comme le calcul numérique d'une loi normale multivariée est souvent un problème difficile, Genz décrit dans son papier, une transformation qui simplifie le problème et le place dans $[0,1]^K$. Cette forme permet l'utilisation d'algorithmes efficaces pour le calcul d'intégrales multiples. Il suggère en particulier l'utilisation d'algorithmes MCQMC. En effet, une simple méthode de Monte-Carlo (MC) utilisant N points présente une erreur en $O(1/\sqrt{N})$ alors que les méthodes Quasi Monte Carlo (QMC) ont des erreurs approximativement en $O(1/N)$ (Fox 1999, Sloan et al. 1994). Afin d'obtenir un intervalle de confiance, une étape de Monte-Carlo supplémentaire est nécessaire, d'où le MCQMC.
A noter qu'ici la fonction QSIMVNEF a été adaptée, et une méthode de Newton a été utilisée afin de trouver le seuil c_α^2 tel que $\mathbb{P}(0 \leq M \leq c_\alpha^2) = \alpha$.

Génome Scan sous le modèle d'interférence

Dans la section précédente, nous avons modélisé les recombinaisons à l'aide de la modélisation de Haldane qui suppose que le nombre de recombinaisons suit un processus de Poisson. Cependant, il s'avère que parfois une recombinaison dans un intervalle donné puisse inhiber la formation d'une autre recombinaison à proximité. On appelle ce phénomène, "phénomène d'interférence" (Sturtevant 1915, Muller 1916). Nous nous intéressons ici à ce phénomène, à l'aide du modèle de Rebaï et al. (1995). Le modèle peut se décrire de la manière suivante : une seule recombinaison est autorisée au maximum sur l'intervalle délimité par deux marqueurs, et lorsqu'il y a recombinaison, celle-ci se produit uniformément sur l'intervalle. Par conséquent, les poids du modèle de mélange ne sont plus les mêmes que sous Haldane (1919). Conditionnellement aux marqueurs, Y obéit au modèle de mélange suivant dont les poids s'avèrent connus :

$$\tilde{p}(t^\star) f_{(\mu+q,\sigma)}(Y) + \{1 - \tilde{p}(t^\star)\} f_{(\mu-q,\sigma)}(Y), \tag{3}$$

où $f_{(m,\sigma)}$ désigne la densité Gaussienne de paramètres (m, σ) et où la fonction $\tilde{p}(t)$ peut s'exprimer à l'aide de Rebaï (1995).

Dans mon étude, je démontre que le maximum du processus de LRT converge en loi vers le maximum du carré d'un processus d'interpolation linéaire renormalisé. Ce processus d'interpolation linéaire est centré sous l'hypothèse nulle d'absence de QTL sur $[0,T]$, et décentré d'une fonction moyenne sous l'alternative qu'il existe un QTL en $t^\star \in [0,T]$. C'est une généralisation des résultats présents dans Rebai (95) et Rebai (94), où les auteurs ne s'intéressent qu'à l'hypothèse nulle et caractérisent le processus uniquement par sa covariance.

Un résultat également obtenu est le suivant : le maximum du carré du processus d'interpolation linéaire et le maximum du carré du processus d'interpolation non linéaire, sont identiques sous l'hypothèse nulle. En d'autres termes, le seuil de rejet est le même que l'on utilise le modéle de Haldane ou le modèle d'interférence. D'autre part, sous l'hypothèse alternative, je prouve que la fonction moyenne du processus est totalement différente que celle obtenue avec le modèle de Haldane. Un résultat original et intéressant pour les biologistes, est que l'on dispose de beaucoup plus de puissance statistique lorsque l'on considère le modéle d'interférence que lorsque l'on considère le modéle de Haldane (i.e. sans interférence). Ce travail s'est traduit par une publication dans TEST (2013).

Robustesse du LRT au modèle de recombinaison

On s'intéresse ici à la robustesse des tests statistiques en détection de QTL. On considère deux modèles possibles de recombinaison (Haldane et le modèle d'interférence), et on étudie, sous le vrai modèle de recombinaison, la distribution de la statistique de LRT construite à partir du faux modèle de recombinaison.

En supposant Haldane comme le vrai modèle, le résultat principal obtenu est le suivant : la distribution de la statistique de LRT, $\sup \tilde{\Lambda}_n(.)$, construite à partir du modèle d'interférence (i.e. le faux modèle) est asymptotiquement celle du maximum du carré d'un "processus d'interpolation linéaire renormalisé". On montre également que, sous l'hypothèse nulle et sous des alternatives contigües, le maximum du carré de ce "processus d'interpolation linéaire renormalisé" est le même que celui du "processus d'interpolation non linéaire renormalisé" obtenu par Azaïs et al (2014). En d'autres termes, sous Haldane, nous avons la relation suivante :

$$\sup \tilde{\Lambda}_n(.) = \sup \Lambda_n(.) + o_P(1) \ , \tag{4}$$

où $\sup \Lambda_n(.)$ désigne la statistique de LRT basée sur Haldane (cf. Section Génome Scan sous le modèle de Haldane).

Par conséquent, il existe une certaine "robustesse asymptotique du LRT" : même si l'on choisit par erreur de modéliser les recombinaisons par le modèle d'interférence au lieu d'opter pour le modèle de Haldane, le test de rapport de vraisemblance sera asymptotiquement optimal.

Par la suite, on s'intéresse à la configuration inverse : le vrai modèle est le modèle d'interférence, et le faux modèle correspond à Haldane. On démontre que la formule (4) est encore vraie sous le modèle d'interférence. Ainsi, on peut vraiment utiliser la terminologie "robustesse asymptotique du LRT" en détection de QTL. Ce travail a donné lieu à publication dans Electronic Journal of Statistics (2014).

Génome Scan combiné à un Selective Genotyping

Lors de mon étude du Selective Genotyping, je me suis focalisé uniquement sur une position du génome donnée. Ainsi, la suite logique est de combiner l'approche Génome Scan avec un Selective Genotyping. Le fait de parcourir le génome (ie. Génome Scan) en considérant un Selective Genotyping, est tout à fait nouveau, et avec un intérêt à la fois théorique et pratique. A noter que dans la littérature, on ne dispose que d'une étude par simulation (Rabbee et al., 2004) présentant les puissances associées à certaines stratégies d'analyse de données en Selective Genotyping.

En Selective Genotyping, on peut prouver que la vraisemblance (phénotype Y et marqueurs flanquants), est proportionnelle au modèle de mélange suivant

$$\begin{aligned} &\Big[p(t^\star)\ f_{(\mu+q,\sigma)}(Y)\ 1_{Y\notin[S_-,S_+]} \ + \ \{1-p(t^\star)\}\ f_{(\mu-q,\sigma)}(Y)\ 1_{Y\notin[S_-,S_+]} \\ &+ \ \frac{1}{2}\ f_{(\mu+q,\sigma)}(Y)\ 1_{Y\in[S_-,S_+]} \ + \ \frac{1}{2}\ f_{(\mu-q,\sigma)}(Y)\ \ 1_{Y\in[S_-,S_+]}\Big] \end{aligned} \tag{5}$$

où $f_{(m,\sigma)}$ désigne la densité Gaussienne de paramètres (m,σ).

Comme précédemment, le challenge réside dans le fait que $t^\star$ est inconnu. Par conséquent, à chaque position $t \in [0,T]$, on effectue un Test du Rapport de Vraisemblance (LRT), $\Lambda_n(t)$, de l'hypothèse"q=0" dans la formule (5), basé sur n observations. On est ainsi amené à considérer un processus de LRT $\Lambda_n(.)$ et le maximum de $\Lambda_n(.)$ donne encore le LRT de "q=0".

Le résultat principal que j'ai obtenu est que sous le modèle de Haldane, le maximum du processus de LRT converge en loi vers le maximum du carré d'un "processus d'interpolation non linéaire". Sous l'hypothèse nulle, ce processus interpolé est exactement le même que celui obtenu précédemment pour un Génome Scan sans Selective Genotyping. Cependant, sous l'alternative, les fonctions moyennes des

deux processus s'avèrent différentes. Je prouve également que l'on doit génotyper le même pourcentage d'individus aux deux extrêmes de la population, et que les phénotypes non extrêmes n'apportent aucun gain de puissance dans l'analyse statistique. Enfin, je propose une formule simple afin de calculer le maximum du processus de LRT : il n'est plus nécessaire d'avoir recours à l'algorithme EM et l'on doit considérer uniquement quelques positions bien précises sur le génome. A noter qu'à travers cette étude, on prouve également que le seuil (i.e. valeur critique) reste le même que celui obtenu pour un Génome Scan sans Selective Genotyping. Ce travail a donné lieu à publication dans Statistics (2015).

Si l'on considère désormais un modèle d'interférence associé à un Selective Genotyping, le processus devient alors un processus d'interpolation linéaire renormalisé (publication dans Journal of Statistical Planning and Inference 2014).

Pour finir, si on génotype uniquement les individus présentant un phénotype non extrême (i.e. $Y \in [S_-, S_+]$), alors le processus de LRT s'avère toujours être le carré d'un processus interpolé. Cette étude a été publiée dans les Annales de la Faculté des Sciences de Toulouse (2014). D'après cette étude, la puissance du test est maximum lorsque les génotypes manquants sont situés dans une queue de la distribution Gaussienne. Par conséquent, cela prouve que la plupart de l'information est présente dans les extrêmes, comme suggéré par Lebowitz et al. (1987).

Processus de Chi-deux d'Ornstein-Uhlenbeck

On considère une population avec une structure de famille. Chaque famille et constituée de descendants d'un père (i.e. design fille). En effet, en pratique, un QTL ne peut être détecté dans une famille que si le père est hétérozygote au QTL. Par conséquent, les généticiens ciblent plusieurs familles. Dans un article publié dans Statistical Papers (2016, coauteurs JM Azaïs, JM Elsen et C Delmas), nous donnons la distribution asymptotique du processus de LRT sous l'hypothèse nulle d'absence de QTL dans aucune des familles, et sous l'alternative locale où un QTL est présent à $t^\star \in [0, T]$ dans au moins une famille.

Nous montrons que le processus de LRT converge dès lors vers la somme de processus Gaussiens interpolés au carré. Le nombre de processus correspond dès lors au nombre de familles. A noter que lorsque le nombre de marqueurs génétiques tend vers l'infini, le processus de LRT converge vers un processus Chi-deux d'Ornstein-Uhlenbeck (nombre de degrés de libertés = nombre de familles) sur $[0, T]$. Par conséquent, afin de prendre une décision quant à la présence d'un QTL sur $[0, T]$, la valeur critique pour le processus de Chi-deux d'Ornstein-Uhlenbeck doit être calculée. Pour rappel, un processus de Chi-deux $Z(.)$ avec d degrés de libertés, vérifie :

$$Z(t) = V_1(t)^2 + \cdots + V_d(t)^2 \tag{6}$$

où les $V_i(t)$ sont indépendants pour chaque t et distribués selon une loi Normale standardisée sous l'hypothèse nulle. Le processus de Chi-deux d'Ornstein-Uhlenbeck présente la particularité que les processus $V_i(.)$ sont indépendants et $\mathrm{Cov}\left\{\mathrm{V_i(t)}, \mathrm{V_i(t')}\right\} = \mathrm{e}^{-2|\mathrm{t}-\mathrm{t'}|}$.

Afin de poursuivre dans cette thématique, j'ai travaillé en collaboration avec Alan Genz, professeur à l'université de Washington State (USA). Dans un article publié dans Methodology and Computing in Applied Probability (2013), on propose une formule théorique basée sur les travaux de Delong, permettant d'obtenir la valeur critique pour le maximum d'un processus de Chi-deux d'Ornstein-Uhlenbeck. On s'attarde d'une manière plus générale sur les valeurs critiques pour le maximum de processus de Chi-deux. Nous proposons une borne inférieure (due à la discrétisation du processus) obtenue par des techniques Monte-Carlo Quasi Monte-Carlo (MCQMC) et utilisant une transformation décrite dans Genz (1992). A noter qu'une simple méthode de Monte-Carlo (MC) utilisant N points présente une erreur en $O(1/\sqrt{N})$ alors que les méthodes Quasi Monte Carlo (QMC) ont des erreurs approximativement en $O(1/N)$ (Fox 1999, Sloan et al. 1994). Afin d'obtenir un intervalle de confiance, une étape de Monte-Carlo supplémentaire est nécessaire, d'où le MCQMC. Pour finir, nous

comparons notre borne inférieure obtenue par MCQMC avec la borne supérieure de Davies (1987) pour des processus de Chi-deux suffisament réguliers.

Nouvelle méthode de sélection de variables basée sur le processus empirique

Résumé : Nous introduisons une nouvelle méthode de selection de variables, nommée SgenoLasso, qui prend en compte les données extrêmes. Notre méthode est basée sur la construction d'un test statistique spécifique, une transformation des données et par la connaissance de la corrélation entre régresseurs. Cela s'avère approprié en génomique car une fois la carte génétique construite, cette corrélation est parfaitement connue. Cette nouvelle technique est inspirée des processus stochastiques en provenance de la statistique génétique. Nous prouvons que le rapport signal bruit est largement augmenté en considérant les extrêmes. Notre approche ainsi que les méthodes existantes sont comparées sur données simulées et réelles. Ceci valide notre nouvelle approche.

Contexte et description de la méthode : Généralement, plusieurs loci sont responsables de la variation d'un caractère quantitatif. Ainsi, dans ce qui suit, on suppose que le caractère quantitatif Y est affecté par m QTLs. q_s et $t_s^\star$ désignent respectivement l'effet et la position du s-ème QTL sur le chromosome. De plus, on impose $0 < t_1^\star < ... < t_m^\star < T$. On suppose dès lors le modèle d'analyse de variance

$$Y = \mu \ + \ \sum_{s=1}^{m} X(t_s^\star)\, q_s \ + \ \sigma\varepsilon \tag{7}$$

où ε désigne un bruit Blanc Gaussien.

La problèmatique est la suivante : considérer comme statistique de test le supremum du processus est approprié lorsqu'il existe uniquement un QTL sur le chromosome, mais cela devient inapproprié lorsque plusieurs QTLs sont en ségrégation sur le chromosome. Par conséquent, une approche différente doit être adoptée.

Dans un papier publié dans Statistics (Rabier et Delmas, 2021), nous calculons la distribution asymptotique du processus de LRT sous l'hypothèse générale qu'il existe m QTLs sur $[0, T]$. Conditionnellement aux observations aux marqueurs, la réponse est désormais un mélange à 2^m composantes. Plus précisément, ce mélange est de la forme :

$$\sum_{(u_1,...,u_m)\in\{-1,1\}^m} w_{\vec{t}^\star}(u_1,\ ...,\ u_m)\ f_{(\mu+u_1q_1+...+u_mq_m,\sigma)}(Y)$$

où $w_{\vec{t}^\star}(u_1,\ ...,\ u_m)$ est la probabilité $\mathbb{P}\left\{X(t_1^\star) = u_1, ..., X(t_m^\star) = u_m\right\}$ conditionnellement aux observations aux marqueurs.

On montre par contiguïté, que sous cette alternative générale, le processus de LRT converge vers le carré d'un processus Gaussien dont la fonction moyenne dépend du nombre de QTLs, leurs positions et leurs effets. A noter que nous calculons également la distribution asymptotique du processus de LRT sous l'hypothèse générale qu'il existe m QTLs additifs et $\tilde{m}$ QTLs en intéractions sur $[0, T]$. Nous prouvons que la fonction moyenne du processus ne dépend pas des QTLs en intéractions.

Ces résultats ont par la suite été généralisés au cas du Selective Genotyping. Lorsqu'il existe m QTLs, la loi de $\left(Y,\ \overline{X}(t_1),\ ...,\ \overline{X}(t_K)\right)$ est désormais proportionnelle au mélange à 2^m composantes

$$\begin{aligned}&\sum_{(u_1,...,u_m)\in\{-1,1\}^m} w_{\vec{t}^\star}(u_1,\ ...,\ u_m)\ f_{(\mu+u_1q_1+...+u_mq_m,\sigma)}(Y)\ 1_{Y\notin[S_-,S_+]}\\ &+ v_{\vec{t}^\star}(u_1,\ ...,\ u_m)\ f_{(\mu+u_1q_1+...+u_mq_m,\sigma)}(Y)\ 1_{Y\in[S_-,S_+]}\end{aligned}$$

où $v_{\vec{t}^\star}(u_1,\ ...,\ u_m)$ est la probabilité $P(X(t_1^\star) = u_1, X(t_2^\star) = u_2, \ldots, X(t_m^\star) = u_m)$ et où $w_{\vec{t}^\star}(u_1,\ ...,\ u_m)$ est la même quantité que précédemment. Nous prouvons que le processus de LRT converge vers le

carré d'un processus Gaussien, dont la fonction moyenne dépend non seulement de la position des QTLs et leurs effets, mais aussi d'un facteur lié au Selective Genotyping. Ce facteur joue le rôle de coefficient multiplicatif pour les effets QTLs : le signal peut ainsi être augmenté à condition que le nombre d'individus n ait été augmenté lors d'une experience avec Selective Genotyping.

Nous proposons d'estimer les paramètres inconnus (nombre de QTLs, leurs positions et leurs effets) par vraisemblance pénalisée (en particulier la pénalité L1), ce qui constitue une nouvelle méthode de recherche de QTLs. Cette nouvelle méthode, nommée SgenoLasso, diffère du Lasso classique (Tibshirani, 1996) car elle modélise explicitement les extrêmes. SgenoLasso présente toutes les propriétés statistique du Lasso classique car le problème a été replacé dans un cadre de pénalisation L1. Ce qui n'est pas le cas du Lasso en présence de données extrêmes. Comme son ancêtre Lasso, SgenoLasso a de nombreux cousins, chacun imposant sa propre pénalité sur les paramètres : on peut citer par exemple SgenoElasticNet (un mélange de pénalité L1 et L2) ainsi que SgenoGroupLasso (pénalité par groupe). Notre approche ainsi que les méthodes existantes sont comparées sur données simulées et réelles dans un contexte d'étude d'association (GWAS). SgenoLasso et ses cousins dominent largement le Lasso, Group Lasso, Elastic Net, RaLasso et BayesianLasso, en particulier quand le modèle est appris sur les individus élites (i.e. présentant les plus grands phénotypes). De la même manière, sur données simulées, nous montrons que les prédictions génomiques liées au SgenoLasso sont bien meilleures que celles basées sur les méthodes usuelles, lorsque le modèle de prédiction génomique est réactualisé. Ceci s'avère interessant pour la communauté (cf. Zhao et al. 2012). Comme les modèles usuels ne sont plus fiables dès lors qu'ils sont appris uniquement sur les individus élites, Brandariz et Bernardo (2018) préconisent de conserver quelques individus les moins performants dans le programme de reproduction afin de disposer de modèles fiables. Cependant, conserver de tels individus est non négligeable d'un point de vue économique. Notre approche ne présente pas de tels inconvénients.

Dans un article en cours de rédaction (Rabier et Delmas, 2022), nous nous intéressons à la version adaptative du SgenoLasso, le AdaptiveSgenoLasso. Ce nouveau concept repose sur un Selective Genotyping variant le long du génome. Rappelons que le Selective Genotyping, dans sa version originale, consiste à génotyper seulement les individus extrêmes, afin d'augmenter le signal. Cependant, comme le même taux de sélection est appliqué à tous les loci du génome, le signal est augmenté partout du même facteur proportionnel. En considérant un Selective Genotyping qui varie le long du génome, nous permettons ici aux généticiens de placer plus de poids sur des loci connus pour être responsables de la variation du caractère quantitatif. Le signal qui en résulte est désormais propre à chaque locus. Après avoir developpé la théorie pour AdaptiveSgenoLasso, nous montrons sur des données simulées la supériorité de cette nouvelle approche par rapport au Selective Genotyping classique.

Arbres aléatoires / Arbres phylogénétiques

Lors de mon postdoctorat à l'université de Wisconsin-Madison, je me suis tout d'abord intéressé au problème classique de reconciliation d'arbres d'espèces et de gènes. Puis, j'ai proposé deux nouvelles méthodes pour la détection de duplications entières du génome ("Whole Genome Duplications", e.g. Jiao et al., 2011, Sato et al., 2012) : l'une basée sur la taille des familles de gènes, l'autre s'appuyant sur les séquences d'ADN. Cette dernière méthode s'avère être une nouvelle méthode de reconstruction d'arbres de gènes en présence de duplications entières du génome.

Le problème classique de reconciliation d'arbres d'espèces et de gènes

Commençons par définir les notions d'arbre d'espèces, d'arbre de gènes et de famille de gènes. Un arbre d'espèces représente l'évolution d'un nombre d'organismes (cf. Arvestad et al. 2009) alors qu'un arbre de gènes représente l'évolution d'une famille de gènes. Les gènes appartenant à une famille présentent généralement des fonctions similaires et proviennent d'un même gène ancestral pas trop

distant. Les arbres de gènes sont souvent très différents suivant la famille de gène (Rokas et al. 2003), et par conséquent, de nombreux arbres de gènes ne sont pas en accord avec l'arbre d'espèces. La compréhension de ces incongruences entre arbres d'espèces et de gènes est aujourd'hui un défi majeur en phylogénie. D'une manière plus générale, même s'il y a congruence, on cherche à connaitre l'histoire qui a conduit à un arbre de gènes donné : on parle de reconciliation d'arbre d'espèces et de gènes.

Deux gènes d'une famille de gènes sont dits orthologues (Fitch 1970) si leur ancêtre le plus récent dans l'arbre de gènes est une spéciation. Les orthologues s'opposent aux paralogues dont l'ancêtre le plus récent est une duplication. Une fois l'arbre de gènes et l'arbre d'espèces reconciliés, on peut dès lors différencier les gènes paralogues et orthologues au sein de la famille de gènes. D'un point de vue biologique, cette distinction entre paralogues et orthologues est primordiale car, contrairement aux orthologues, les paralogues n'assurent pas les mêmes fonctions.

Dans notre étude, le modèle mathématique considéré est un modèle hiérarchique à plusieurs niveaux stochastiques. Tout d'abord, le modèle probabiliste de Arvestad et al. (2009) modélise l'arbre aléatoire d'une famille de gènes. En particulier, une famille de gènes évolue à l'intérieur d'un arbre d'espèces selon un processus de naissances et de morts (Kendall 1948 et 1949). Les naissances symbolisent les duplications, les morts représentent les pertes de gènes, tandis que les noeuds de bifurcation de l'arbre d'espèces représentent les spéciations. Le deuxième niveau stochastique réside dans le fait qu'il existe un modèle markovien modélisant les mutations dans les séquences d'ADN, le long des branches de l'arbre aléatoire. Les données étant uniquement les séquences d'ADN, il reste dès lors à reconstruire l'arbre de gènes à partir des séquences. Pour chaque topologie d'arbre de gènes inferé, on reconcilie l'arbre de gènes avec l'arbre d'espèces, i.e. on essaie de reconstruire le passé (pertes, spéciations, duplications) qui a conduit à un arbre de gènes donné. En d'autre termes, l'arbre de gènes doit à la fois "coller" aux séquences d'ADN mais doit également correspondre à une histoire vraisemblable du processus de naissances et de mort à l'intérieur de l'arbre d'espèces.

Deux nouvelles méthodes pour la détection de duplications entières du génome

Il est bien connu que chez les animaux, fungi et autres organismes en particulier les plantes, a existé à un moment donné dans le temps, une duplication entière du génome (i.e. WGD). Ces WGDs peuvent être suivies d'une perte instantanée d'une copie du génome (e.g. Jiao et al., Nature 2011). A noter que de récentes études ont souligné la présence de WGDs chez la banane (D'Hont et al., Nature 2012) ainsi que chez la tomate (Sato et al., Nature 2012). Les WGDs inférées sont généralement basées sur la distribution de l'age des duplications (e.g. Blanc et al. 2004 , Cui et al. 2006), ou sur des méthodes de synthénie basées sur la conservation de l'ordre des gènes (e.g. Kellis et al. 2004). Cependant, les méthodes basées sur la distribution de l'age des duplications, ne permettent de détecter que les WGDs les plus récentes en raison des effets de saturation et de la difficulté à dater les anciennes duplications. De plus, les méthodes de synthénie sont sensibles aux réarrangements du génome. Par conséquent, l'objectif de notre étude est de proposer aux biologistes des méthodes pour la détection de WGD, ne présentant pas de tels inconvénients. Dans ce contexte, nous présentons deux méthodes rigoureuses basées sur des milliers de familles de gènes et sur un nouveau modèle probabiliste.

Pour chaque évènement de WGD, notre modèle suppose que tous les gènes entrant à la WGD sont instantanément copiés et que les copies supplémentaires peuvent être instantanément perdues avec une distribution de probabilité. On modélise ainsi les duplications à grande échelle et les duplications partielles (Jackson 2007, Freeling 2009). Ces pertes immédiates, incluses dans notre modèle, symbolisent la fragmentation, le mécanisme qui a tendance à ramener le nombre de gènes au nombre de gènes original avant la WGD (Langham 2004, Freeling 2009). Ces pertes immédiates permettent également de capter l'augmentation du nombre de pertes, sur une courte période, après un évènement de WGD (Scannell 2006, Konrad 2011).

A l'aide de notre modèle probabiliste pour les évènements de WGDs et le modèle probabiliste d'Arvestad et al (2009) pour les duplications à petite échelle et pertes, nous proposons deux méthodes

afin de tester la présence et la localisation des WGDs sur un arbre d'espèces connu. Ces deux méthodes nécessitent un ensemble de familles de gènes, échantillonnées aléatoirement à partir de l'ensemble des familles de gènes.

La première méthode, dite "méthode de reconciliation", s'appuie sur des alignements de séquences. Elle est inspirée de Rasmussen et Kellis (2011) mais incorpore les évènements de WGDs. La probabilité des alignements de séquences, D, d'une seule famille de gènes, est obtenue en intégrant la topologie T inobservée de l'arbre de gènes, sa reconciliation R avec l'arbre d'espèces, et les longueurs de branches l de l'arbre de gènes :

$$\mathbb{P}(D) = \sum_{T,R} \int_l \mathbb{P}(D, l, T, R) = \sum_{T,R} \int_l \mathbb{P}(D|l, T)\ \mathbb{P}(l|T, R)\ \mathbb{P}(T, R)\,. \tag{8}$$

Afin de calculer la probabilité à priori de la topologie réconciliée $\mathbb{P}(T, R)$, étant donné un arbre d'espèces avec des WGDs potentielles, nous utilisons les algorithmes puissants de Arvestad et al. (2003, 2009) et Rasmussen et Kellis (2011), modifiés afin d'autoriser la présence de WGDs. Cet algorithme traverse l'arbre par "post order" car $\mathbb{P}(T, R)$ peut être factorisé en probabilités de sous arbres de l'arbre de gènes, correspondant aux différentes branches de l'arbre d'espèces.

Notre deuxième approche ignore l'information apportée par les séquences d'ADN, et n'a pas pour but de reconstituer les arbres de gènes. Elle est basée sur la taille des familles de gènes (cf. Hahn et al. 2005, Csurös and Miklós 2006, mais incluant désormais des WGDs), i.e. le nombre de copies de gènes dans chaque espèce pour chaque famille de gènes. Afin de calculer la vraisemblance, des feuilles à la racine, nous calculons à chaque noeud v de l'arbre d'espèces les probabilités d'observer les données D_v chez les descendants de v, conditionnellement à ce que i gènes survivent en v. Ces calculs sont effectués au moyen de l'algorithme de Csurös et Miklós (2006) et adapté pour permettre la présence de WGDs.

La méthode de reconciliation et la méthode basée sur la taille des familles, ont par la suite été comparées à l'aide de données simulées et de données réelles. La méthode de reconciliation a été testée à l'aide de la technologie HTCondor, développée par les chercheurs en Informatique de l'université de Wisconsin-Madison, et permettant l'accès à un très grand nombre de processeurs sur le continent Américain. Nous avons au final eu recours à plus d'un million de processeurs. A noter qu'à l'heure actuelle, HTCondor est non seulement utilisé par Dreamworks pour la fabrication de ses films mais a aussi contribué à la mise en évidence du Boson de Higgs. Les deux méthodes présentent de bons résultats. La méthode basée uniquement sur les tailles de familles de gènes, et qui traite donc moins d'information, s'avère néanmoins plus performante que la méthode basée sur les séquences d'ADN. En effet, afin de tester de manière exacte la présence de WGD au moyen de la méthode de reconciliation, tous les arbres de gènes, reconciliations, longueurs de branches, auraient dû être considérés et intégrés. Cependant, cela s'avère impossible en pratique en raison de l'immensité de l'espace des paramètres. Nous avons échantillonné la distribution à posteriori de l'arbre sachant les séquences d'ADN, par MCMC, à l'aide d'algorithmes (NNI, SPR, Local Change ...) permettant de proposer de nouveaux arbres. Puis, nous nous sommes focalisés uniquement sur l'arbre de gènes correspondant au Maximum à Posteriori et à la reconciliation la plus parsimonieuse.

A contrario, la méthode basée sur la taille des familles et qui ne reconstruit pas par conséquent les arbres de gènes, considére, elle, l'ensemble des scénarios possibles. La méthode de reconciliation s'avère toutefois intéressante pour les biologistes, car elle renvoie l'arbre de gènes et permet de distinguer les gènes paralogues et orthologues. A l'aide de données simulées sous différentes configurations, nous avons estimé que la méthode de reconciliation, basée sur une optimisation stochastique (MCMC), reconstruit entre 82% et 88% des cas l'arbre de gènes simulé. Ces résultats ont été obtenus pour des séquences de longueurs 500 nucléotides.

Pour finir, nous avons illustré les performances de nos méthodes sur des données réelles de 16 espèces de fungi (9209 familles de gènes). Les deux méthodes furent capables de détecter la WGD

bien connue située dans la clade de Saccharomyces cerevisiae. Elles s'accordent sur un petit taux de rétention à la WGD, comme établi par les méthodes de synthénie (Kellis et al. 2004).

Ce travail s'est traduit par une publication avec Cécile Ané et Tram Ta dans Molecular Biology and Evolution (2014). Les softwares correspondant aux méthodes, sont disponibles sur http ://charles-elie.rabier.pagesperso-orange.fr/doc/SPIMAPWGD.html. La méthode basée sur la taille des familles de gènes est distribuée au travers d'un package R, alors que la méthode de reconciliation est implémentée dans un programme C++, nommé SPIMAPWGD, et basé sur le software SPIMAP de Rasmussen et Kellis (2011).

Statistique en grande dimension / Sélection génomique

L'objectif de la sélection ("avec de la") génomique est de sélectionner des candidats sur la base de prédiction génomiques. Ces prédiction génomiques, utilisant une grande densité de marqueurs génétiques, se doivent d'être précises afin de selectionner les meilleurs candidats (i.e. ceux dont le futur phénotype à l'age adulte sera le plus grand par exemple). On se retrouve dans un cadre où le nombre de marqueurs K est bien plus grand que le nombre d'observations n (i.e. $K >> n$). J'étudie les propriétés mathématiques de la quantité suivante ρ_g appelée "accuracy" génotypique, définie de la manière suivante :

$$\rho_g := \frac{\mathrm{Cov}\left(\hat{\mathrm{Y}}_{\mathrm{new}},\ \mathrm{Y}_{\mathrm{new}}\right)}{\sqrt{\mathrm{Var}\left(\hat{Y}_{new}\right)\mathrm{Var}\left(Y_{new}\right)}} \tag{9}$$

où Y_{new} désigne la valeur phénotypique d'un nouvel individu (i.e. individu TEST) et $\hat{Y}_{new}$ est la valeur prédite (construite à partir d'un estimateur basé sur la régression Ridge et sur n individus d'entrainement). Rappelons que dans le cadre de la Ridge, la prédiction $\hat{Y}_{new}$ est la suivante :

$$\hat{Y}_{new} = m'_{new}M'V^{-1}\vec{Y} \text{ avec } V = MM' + \lambda I_n,\ \vec{Y} = (Y_1, \ldots, Y_n)' \text{ et}$$

— m_{new} : vecteur colonne de taille K contenant le génome aux marqueurs de l'individu new
— M : matrice de taille $n \times K$ contenant le génome aux marqueurs des n individus d'entrainement
— λ : paramètre de régularisation.

Dans mon étude, m_{new} est considéré comme aléatoire alors que M est fixe. Dans ce qui suit, $M^\star$ et $m^\star_{new}$ seront les analogues de M et m_{new} mais désormais pour le génome aux QTLs. De plus, Q désigne le nombre de QTLs alors que $\beta^\star$ désigne le vecteur colonne de taille Q contenant les effets QTLs. Ainsi, le "vrai" modèle est le suivant : $\vec{Y} = M^\star\beta^\star + \sigma\vec{\varepsilon}$, $Y_{new} = m^{\star\prime}_{new}\beta^\star + \sigma\varepsilon_{new}$. A noter que les QTLs ne sont pas nécessairement situés sur les marqueurs génétiques, ce qui constitue une originalité de cette étude.

Des résultats théoriques sur la régression Ridge, sont présents dans Goldenshluger et Tsybakov (2003), Dicker (2012), Shao et Deng (2012), Bühlmann (2013). Cependant, les auteurs s'intéressent principalement à l'erreur quadratique $\mathbb{E}\left\{\left\|\hat{Y}_{new} - Y_{new}\right\|^2\right\}$. Le critère d'accuracy, très peu étudié dans la littérature statistique, s'avère être un élément clé en génomique car il intervient dans le calcul du progrès génétique. Dans ce contexte, je prouve que, conditionnellement à M et $M^\star$,

$$\rho_g = \frac{\beta^{\star\prime}\,\mathbb{E}\left(m^\star_{new}m'_{new}\right)M'V^{-1}M^\star\beta^\star}{\left\{\sigma^2\mathbb{E}\left(\left\|m'_{new}M'V^{-1}\right\|^2\right) + \beta^{\star\prime}M^{\star\prime}V^{-1}M\mathrm{Var}\left(m_{new}\right)M'V^{-1}M^\star\beta^\star\right\}^{1/2}\Omega}\ . \tag{10}$$

Par la suite, on se replace dans un modèle sparse classique : on suppose que les QTLs se trouvent uniquement sur quelques marqueurs, mais toujours en considérant un très grand nombre de marqueurs.

Soit β le vecteur colonne de taille K constitué de zéros, et dont les composantes non nulles contiennent les différents effets QTLs lorsque le marqueur en question coïncide avec un QTL.

La décomposition en valeur singulières de la matrice M s'écrit :

$$M = UDW',$$

où U est une matrice $n \times r$ telle que $U'U = I_r$, W est une matrice $p \times r$ telle que $W'W = I_r$, et D est une matrice diagonale $r \times r$ de plein rang. On notera d_1 ,..., d_r les élements diagonaux de D (les valeurs singulières).

Nous montrons que l'accuracy génotypique peut être estimée par

$$\hat{\rho}_g = \frac{\hat{A}_1}{\left(\hat{A}_2 + \hat{A}_3\right)^{1/2} \left(\hat{A}_4\right)^{1/2}}$$

où

$$\hat{A}_1 = \frac{1}{n} \sum_{s=1}^{r} \frac{d_s^4}{d_s^2 + \lambda} \left\| W^{(s)} W^{(s)\prime} \beta \right\|^2 \ , \ \hat{A}_2 = \frac{\sigma^2}{n} \sum_{s=1}^{r} \frac{d_s^4}{(d_s^2 + \lambda)^2}$$
$$\hat{A}_3 = \frac{1}{n} \sum_{s=1}^{r} \frac{d_s^6}{(d_s^2 + \lambda)^2} \left\| W^{(s)} W^{(s)\prime} \beta \right\|^2 \ , \ \hat{A}_4 = \frac{1}{n} \sum_{s=1}^{r} d_s^2 \left\| W^{(s)} W^{(s)\prime} \beta \right\|^2 .$$

Ainsi, on peut remarquer l'influence des valeurs singulières, du paramètre de régularisation, et de la projection du signal sur l'espace engendré par les lignes de la matrice M.

Ce résultat est plus précis que ceux présentés par Shao et Deng (2012). En effet, notre estimation dépend de l'ensemble des valeurs singulières d_1, ..., d_r, et non pas uniquement de la plus grande et de la plus petite, comme dans Shao et Deng (2012). Il s'avère important de conserver l'ensemble des valeurs singulières car le signal β est un paramètre qui n'est pas lié à la décomposition SVD de la matrice M. Ainsi, on peut très bien imaginer que la projection de β soit maximale sur l'espace associé à la plus petite valeur singulière. En imposant des conditions de régularité sur les valeurs singulières ainsi que sur la projection du signal sur les différents sous espaces, nous étudions les convergences de $\hat{\rho}_g$ vers les cas extrêmes : accuracy oracle ou accuracy nulle.

Enfin, nous proposons d'améliorer la prédiction à l'aide d'un nouvel estimateur. Plus précisément, on propose de projeter le vecteur Y sur un sous espace bien choisi de l'espace engendré par les colonnes de X. Soient $1 \leq \tilde{r} \leq r$ et $\psi(.)$ une fonction $\psi : \{1, \ldots, \tilde{r}\} \to \{1, \ldots, r\}$ telle que $\psi(k) \neq \psi(k')$ pour $k \neq k'$.

Notre nouvel estimateur s'avère le suivant

$$\tilde{Y}_{new} = m'_{new} X' V^{-1} \tilde{U} \tilde{U}' Y \text{ où } \tilde{U} = \left(U^{\psi(1)}, \ldots, U^{\psi(\tilde{r})} \right).$$

Nous donnons une condition nécessaire et suffisante pour que l'accuracy construite sur $\tilde{Y}_{new}$ soit supérieure à celle basée sur $\hat{Y}_{new}$.

Ces travaux ont donné lieu à une publication appliquée dans PloS One (2016), ainsi qu'à une publication de nature plus théorique dans Scandinavian Journal of Statistics (2019, coauteurs S Grusea et B Mangin). Les applications portent sur des données réelles de Raygrass et de riz. Je renvoie à l'article pour les éléments théoriques de statistique en grande dimension.

Dans un article publié dans JRSS C (Rabier et Grusea, 2021), nous étudions l'"accuracy" génotypique (i.e. précision de la prédiction) en présence d'un déséquilibre de liaison incomplet. En effet, dans nos études précédentes, nous nous focalisions uniquement sur le cas du déséquilibre de liaison complet : les QTLs étaient localisés sur quelques marqueurs. Le déséquilibre de liaison incomplet est un sujet d'intérêt car pour certaines espèces, le nombre de marqueurs demeure trop faible pour

couvrir la très grande taille de génome. Dans un tel contexte, cette faible densité de marqueurs est incapable de cibler les QTLs.

Nous prouvons que la projection de la fonction de régression sur l'espace engendré par les colonnes de M est désormais un élément fondamental pour la précision de la prédiction. A partir de nos résultats généraux théoriques, nous sommes en mesure de retrouver les résultats obtenus dans le cas particulier du déséquilibre de liaison complet. Nous montrons également que l'accuracy oracle (i.e. optimale) est atteinte uniquement lorsque la limite d'un facteur de perte, lié au déséquilibre de liaison incomplet incomplet, est nulle. Les applications de cette étude portent sur la densité de marqueurs nécessaires afin d'obtenir une prédiction suffisament précise pour l'implémentation de la sélection génomique.

Arbres aléatoires / Réseaux phylogénétiques

Dans la continuité de mon travail en phylogénie aux USA, je m'intéresse à Montpellier, aux réseaux phylogénétiques, et non plus uniquement aux arbres d'espèces. Les réseaux phylogénétiques retranscrivent l'histoire évolutive de groupes d'individus (espèces ou populations), comportant des évènements de réticulation comme les phénomènes d'hybridation ou les transferts horizontaux de gènes.

Les méthodes basées sur la coalescence s'avèrent les plus classiques afin de modéliser un arbre de génes à l'intérieur d'un réseau. En particulier, le modèle "multispecies coalescent" modèlise le tri de lignées incomplet (ILS), une des sources principales de conflits entre arbres de gènes. Afin de calculer la probabilité d'un arbre de gènes à l'intérieur d'un réseau, le modèle de coalescence est considéré à l'intérieur de chaque branche du réseau, excepté celles conduisant à un noeud de réticulation. Un noeud de réticulation présente la particularité suivante : à ce type de noeud, une lignée génétique hérite du matériel génétique d'un de ces parents, avec probabilités γ et $1-\gamma$. Il existe deux modélisations différentes pour les réseaux (Kubatko 2009, Yu et al. 2012). Le modèle de Kubatko (2009) suppose que toutes les lignées génétiques à un locus proviennent du même parent au niveau du noeud de réticulation. A contrario, Yu et al. (2012) suppose que les lignées à un locus donné ne proviennent pas nécessairement du même parent.

D'autre part, le modèle de mutation est le suivant : nous considérons 2 allèles possibles rouge ou vert (loci bialléliques), et le processus de mutation est modélisé par un processus markovien le long de l'arbre aléatoire, avec taux u de muter de rouge à vert, et taux v de muter de vert à rouge. En d'autre termes, le processus de coalescence, qui remonte dans le temps, et le processus de mutation, qui avance dans le temps, agissent dans des directions opposées. A titre d'exemple, est représenté en figure 1, un arbre de gènes évoluant à l'intérieur d'un réseau phylogénétique à 4 espèces (B espèce hybride). Les mutations sont représentées le long des branches de cet arbre aléatoire.

Dans un article publié dans Plos Computational Biology (Rabier et al. 2021), nous proposons une nouvelle méthode Bayésienne de reconstruction de réseau phylogénétique, nommée SNAPPNET. Nous présentons une nouvelle façon de calculer la vraisemblance de marqueurs bialléliques étant donné un réseau phylogénétique. Ce calcul est au coeur de SNAPPNET, car il généralise la méthode SNAPP (Bryant el al., MBE 2012) dédiée à l'inférence d'arbre d'espèces. L'inférence de réseau demeure plus difficile que l'inférence d'arbres d'espèces, car certaines arêtes du réseau ne peuvent pas être traitées de manière indépendante. Ainsi, cette étude s'appuie sur une étude mathématique et sur de nouveaux algorithmes.

A noter que dans sa version originale, SNAPP modélise la distribution a priori de l'arbre d'espèces par un processus de Yule (processus de naissances). Au sein de SNAPPNET, la distribution a priori du réseau est désormais un processus de Naissance-Hybridation (Zhang et al., 2017). Les opérations sur les réseaux lors du MCMC, autorisent l'ajout ou la suppression de noeuds de réticulation (Zhang et al., 2017), synonymes d'évènements d'hybridation.

Notre algorithme de calcul de la vraisemblance s'avère beaucoup plus rapide que celui proposé au

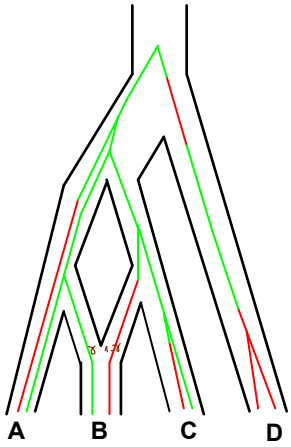

FIGURE 1 – Un arbre de gènes évoluant à l'intérieur d'un réseau phylogénétique (B espèce hybride)

sein de la récente méthode MCMCBiMarkers (Zhu et al, Plos Comput Biol 2018). Cet algorithme se base sur une nouvelle façon de parcourir les graphes dirigés acycliques (DAGs en anglais). De plus, le gain en terme de mémoire est substantiel. Ceci a été illustré à l'aide de données simulées, et par une analyse théorique de la complexité des 2 algorithmes. Pour rappel, l'espace des réseaux à explorer par MCMC s'avère très grand. Le calcul de vraisemblance étant de loin l'étape la plus coûteuse d'un point de vue computationnel, l'optimisation de ce calcul est donc primordiale afin de pouvoir traiter de grands jeux de données.

Notre méthode est distribuée au travers d'un package BEAST. La plateforme logicielle BEAST (Bouckaert et al., Plos Comput Biol 2014 et 2019) est très utilisée par la communauté bioinformatique dans le monde entier, d'où notre choix de cette plateforme. Notre package SNAPPNET bénéficie de l'interface graphique BEAUTI propre à BEAST, ce qui facilitera son utilisation par la communauté scientifique. Il est disponible sur https ://github.com/rabier/MySnappNet. Les applications portent sur le processus de domestication du riz. En particulier, SNAPPNET a suggéré un scénario évolutif jusqu'alors jamais proposé dans la communauté, mais qui s'avère compatible avec des scénarios plus simples présents dans la littérature.