Identifier les segments génomiques expliquant les variations de fonctions de réponse: intérêt des équations différentielles stochastiques dans un contexte bayésien

Bénédicte Fontez [1,2] & Timothée Flutre [3] & Fabien Campillo [4] & Pierre Roumet [3]

[1] *Supagro Montpellier, UMR Mistea, F-34060 Montpellier, France, benedicte.fontez@supagro.fr*
[2] *INRA, UMR729 Mistea, F34060 Montpellier, France*
[3] *INRA, UMR AGAP, F34060 Montpellier, France, pierre.roumet@supagro.inra.fr ; flutre@supagro.inra.fr*
[4] *INRIA, équipe LEMON, F34060 Montpellier, France*

Résumé. Tout organisme vivant, quel qu'il soit, se développant au cours du temps, il est nécessaire, voire primordial, de prendre cette dimension en considération. Grâce à la montée en puissance des capteurs haut-débit, de plus en plus de caractères d'intérêt sont mesurés sous forme de fonctions de réponse et de courbes de croissance. Par exemple en agronomie, dans un but de sélection artificielle, il devient alors pertinent de chercher à identifier les segments génomiques expliquant les variations de fonctions de réponse au sein d'une population (*quantitative trait locus*, QTL). Habituellement, les individus sont d'abord utilisés séparément les uns des autres pour estimer, chez chacun, les coefficients d'une fonction de réponse, ceux-ci étant ensuite testés pour association avec les segments génomiques. Cette méthode *ad hoc* entraîne une perte importante d'information, d'autant plus que l'incertitude associée à l'estimation des paramètres est généralement négligée lors du test. Des auteurs comme Wu *et al.* (2007) ont proposé des approches fonctionnelles pour tenir compte de la dynamique sous-jacente à la croissance dans la détection de QTL. Nous proposons d'étendre cette approche à un modèle plus réaliste où la fonction de réponse est définie comme un processus aléatoire caractérisé par une équation différentielle stochastique (EDS). L'inférence est réalisée dans un cadre bayésien qui permet d'estimer l'effet du QTL et, simultanément, de sélectionner la période de temps durant laquelle le QTL est influent.

Mots-clés. données fonctionnelles, équation différentielle stochastique, QTL, bayésien

Abstract. As every living organism develops through time, it is necessary, even crucial, to take this dimension into account. Thanks to the improvement of high-throughput sensors, traits of interest are increasingly measured as response functions and growth curves. For instance in agronomy, for breeding purposes, it becomes relevant to search for

genomic segments explaining variations of response functions within a population (quantitative trait locus, QTL). Usually, individuals are analyzed separately to estimate, for each of them, the coefficients of a response function, these coefficients being then tested for association with genomic segements. Such *ad hoc* method leads to a substantial loss of information, especially as uncertainty associated with the parameters' estimation is commonly ignored for the test. Authors such as Wu *et al.* (2007) proposed functional approaches to take into account for QTL detection the dynamics underlying the growth. We propose to expand this approach to a more realistic model for which the response function is defined as a random process caracterised by a stochastic differential equation (SDE). Inference is performed in a Bayesian framework which allows to estimate the QTL effect and, jointly, to select the time period during which the QTL is active.

Keywords. functional data, stochastic differential equation, QTL, Bayesian

1 Introduction

En agronomie, la caractérisation des plantes se fait maintenant en temps continu générant des données de type fonctions de réponse, de croissance à partir des plateformes de phénotypage. Une question importante pour l'amélioration des plantes est de pouvoir relier ces fonctions au génome. Habituellement, les données fonctionnelles obtenues par phénotypage sont résumées par les coefficients d'un modèle ou par quelques points remarquables, ce qui entraîne une perte importante de l'information. Les fonctions de réponse (ou de croissance) utilisées sont souvent définies à partir d'équations différentielles ordinaires classiques comme la fonction logistique, Gompertz etc. Or, ces phénomènes biologiques sont souvent stochastiques par nature et non déterministes. Aussi, Donnet *et al.* (2010) ont proposé l'utilisation d'équations différentielles stochastiques pour améliorer la modélisation des courbes de croissance individuelles. Nous avons étudié le gain potentiel de cette approche pour la détection des QTL liés à une fonction de réponse. Notamment, nous proposons une modélisation qui permet de répondre aux questions biologiques suivantes : quels gènes sont influents sur quelles parties de la courbe de réponse ? Peut-on détecter de nouveaux gènes qui gouvernent la forme ou la variabilité de la fonction de réponse ?

Nous avons étudié et comparé les résultats d'une approche classique de détection de QTL (interval mapping) à l'approche fonctionnelle proposée par Ma *et al.* (2002). Nous nous sommes inspirés de cette dernière approche et des travaux de Donnet *et al.* (2010) pour généraliser la détection de QTL dans un cadre bayésien de modélisation à partir d'équations différentielles stochastiques. Le modèle bayésien proposé nous permet de détecter la présence de QTL, d'estimer l'effet du gène et la période de temps où le gène est influent. Les trois approches ont été comparées sur des données réelles de cinétique de résorption d'azote dans une feuille de blé dur et sur des données simulées.

2 Présentation des données expérimentales

La population d' étude était constituée de la descendance issue d'hybridations entre 4 variétés de blé dur croisées 2 à 2. Chacune des 6 populations résultantes était représentée par 48 Lignées génétiquement fixées soit un total de 288 lignées qui ont génotypées avec 1383 marqueurs. L'expérimentation a été réalisée en serre ; tout au long de leur cycle les plantes ont bénéficié d'une alimentation hydrique et d'une nutrition minérale non limitantes. Le dispositif expérimental était constitué de 5 blocs complets au sein des desquels chacune des lignées était représentée. La teneur en azote de la dernière feuille, caractère lié à la composition chimique du grain à la récolte a fait l'objet d'un suivi longitudinal entre la floraison et la maturité de chaque plante (soit 20 à 30 mesures/ feuille). Le protocole d?acquisition des données, leur modélisation ont été détaillés dans une précédente publication par Vilmus *et al.* (2014).

3 Présentation des modèles

Le modèle logistique est usuellement utilisé pour les courbes de croissance sigmoïdes West (2001). Dans notre contexte, c'est le modèle de référence pour les données de résorption d'azote dans la feuille, Vilmus *et al.* (2014), mais aussi plus généralement pour faire le lien entre la croissance et le génome selon Wu *et al.* (2004), Wang *et al.* (2014). Notons Z_t la quantité d'azote au temps t. (Z_t) est un processus aléatoire défini à partir de l'équation différentielle du modèle logistique. Les données sont observées aux temps $T = t_1, \ldots, t_m$ pour chaque feuille i. On suppose que les observations $((y_{i\,t_l}, 0 < i \leq n, 0 < l \leq m)$ sont bruitées selon le modèle suivant:

$$y_{i\,t_l} = Z_{t_l}(\phi_i) + \varepsilon_{i\,t_l}$$

où les ε sont indépendants et de loi normale centrée $N(0, \sigma^2)$. Trois modèles ont été considérés pour le processus Z_t qui correspondent à trois façons de relier la fonction de réponse au gène :

1. Modèle simple et déterministe de la fonction logistique *Vilmus et al.* (2013)

$$\frac{dZ_t(\phi_i)}{dt} = C_i\, Z_t\, (1 - Z_{ti}/A_i)$$

   Les estimations individuelles des coefficients sont ensuite utilisées pour détecter les segments génomiques qui sont liés. Ces segments liés sont appelés Quantitative Trait Loci. La méthode de détection la plus simple consiste à tester segment par segment en utillisant l'interval mapping de Lander et Botstein (1989). Comme on n'observe pas la valeur du génotype au QTL, cette approche fait intervenir des variables latentes et un modèle de mélange. Le lien entre fonction de réponse et génome ne

se fait que pour les coefficients et en deux temps : estimation des coefficients de la fonction de réponse au niveau de l'individu puis détection par interval mapping au niveau du génome.

2. Modèle de mélange déterministe d'après la méthode de Ma *et al.* (2002) qui proposent de détecter directement les QTL en intégrant le modèle de mélange de l'interval mapping dans l'équation différentielle. Pour simplifier les notations, on suppose deux génotypes possibles à un QTL (notés $q = 1$ pour le génotype hétérozygote et $q = 0$ pour le génotype homozygote). Notons $F(Z_t, t, \psi_q) = C_q\, Z_t\, (1 - Z_t/A_q)$, alors :

$$\begin{aligned} \frac{dZ_t(\phi_i)}{dt} &= F(Z_t, t, \psi_{q=1})\, G_i + F(Z_t, t, \psi_{q=0})\, (1 - G_i) \\ G_i &\sim Bernoulli(p) \end{aligned}$$

où G_i désigne la valeur du génotype au QTL pour l'individu i. Cette variable est latente car non observée et on utilise un modèle de mélange. La probabilité p dépend d'un paramètre génétique (le taux de recombinaison) supposé connu. Dans Ma *et al.* (2002), plusieurs tests de type rapport de vraisemblance sont proposés pour tester l'influence d'un QTL sur la valeur des paramètres où son influence à des temps d'observation remarquables (comme au point d'inflexion). Notons, que Wu *et al.* (2004) propose une transformation both sides pour stabiliser la variance des données de croissance. Dans notre contexte, on effectuera un changement de variable $X_t = \ln Z_t$ dans l'équation différentielle.

3. Modèle de croissance stochastique d'après Donnet *et al.* (2010) couplé au modèle de mélange. Le mélange entre les deux génotypes dépend maintenant d'une fonction du temps pour permettre la sélection des périodes de temps où le segment génomique est influent. On note T^* l'intervalle de temps sur lequel la fonction de réponse est définie, dans notre exemple $T^* = [0, 1200]$ en degré-jours. On partitionne l'intervalle T^* en K parties $h_k = [a_k, b_k], k = 1, \ldots, K$ disjointes dont l'union vaut T^*.

$$\begin{aligned} dZ_t(\phi_i) &= C_i Z_t\, (1 - Z_t/A_i)\, dt \; + Z_t\, \theta_t(\alpha_i, \beta) dt\, G_i + \gamma\, Z_t\, dW_t \\ \theta_t(\alpha_i, \beta) &= \sum_{k=1}^{K} \alpha_{ik}\, \beta_k\, 1_{h_{k(t)}} \\ G_i \mid p &\sim Bernoulli(p) \\ (\phi_i, \alpha_i) \mid \mu, \Sigma &\sim N_{K+2}(\mu, \Sigma) \\ \beta_k \mid n &\sim Bernoulli(n/K) \\ n &\sim Binomial(K, p_{nbactivations}) \end{aligned}$$

où W_t est un processus de Wiener standard (ou mouvement brownien), $W_t \sim N(0, t)$. Des lois a priori classiques sont choisies pour les paramètres μ (normal),

Σ, Ω (Wishart) et σ^2 (inverse gamma). Dans ce modèle, on utilise un modèle hiérarchique pour prendre en compte la variabilité individuelle. On utilise une base d'histogramme pour estimer les périodes de temps où le segment génomique serait influent. En comparant ce modèle complet au sous modèle où $G_i = 0$, on teste la présence d'un QTL. On peut ensuite regarder sur quelles périodes le segment génomique est influent et on a une représentation plus ou moins fine (selon la valeur de K) de son influence.

Pour stabiliser la variance, on a effectué le changement de variable $X_t = \ln Z_t$ qui se calcule à partir des formules d'Ito présentées dans Øksendal (1995). Chaque feuille i avec le traitement j a une fonction de réponse, aléatoire, notée $(Z_t(\phi_i))$. A partir de la formule d'Ito, on obtient une équation non pas pour Z_t mais pour la variable transformée X_t. On définit ensuite une version discrétisée aux points d'observation. L'estimation du modèle a été réalisée à partir d'un échantillonneur de Gibbs. Certains termes ont nécessité d'implémenter un algorithme de Metropolis Hastings dans l'échantillonneur de Gibbs.

4 Détection de QTL

Ces modèles ont été mis en oeuvre sur le jeu de données de résorption d'Azote dans la feuille. La qualité de l'ajustement a été vérifiée à partir de l'étude des résidus. Pour le modèle bayésien on a testé en plus la sensibilité du modèle à la loi a priori. Enfin, les résultats ont été mis en regard des connaissances actuelles (biologiques, physiologiques et génétiques) pour discuter de l'intérêt du modèle stochastique par rapport aux modèles déterministes existant. Pour compléter cette étude, on prévoit de comparer, sur des jeux de données simulées, la capacité selon les modèles à retrouver les vrais QTL. Enfin, dans cette étude on s'est surtout intéressé à l'apport d'une modélisation stochastique de la fonction de réponse dans un cas simple : effet additif, peu de marqueurs espacés sur le génome. Mais, ces travaux sont une première marche dont les résultats sont utiles pour réfléchir la génétique d'association en grande dimension (des dizaines à centaines de milliers de marqueurs génétiques).

Bibliographie

[1] Donnet, S. Foulley, J-L et Samson, A. (2010), Bayesian analysis of growth curves using mixed models defined by stochastic differential equations *Biometrics*, 66, 733–741.
[2] Lander, E. S. et Botstein, D. (1989), Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps *Genetics*, 21, 185–199.
[3] Ma, C. X. Casella, G. et Wu, R. (2002), Functional mapping of quantitative trait loci underlying the character process : a theoretical framework, *Genetics*, 161, 1751–1762.

[4] Øksendal, B. (1995), *Stochastic differential equations - an introduction with applications*, Springer.
[5] Vilmus, I. Ecarnot, M. Verzelen, N. et Roumet, P. (2014), Monitoring nitrogen leaf resorption kinetics, by near-infrared spectroscopy during grain filling in durum wheat in different nitrogen availability conditions *Crop Science*, 54, 284–296
[6] Wang, Z. Pang, W. Wu, W. Wang, J. Wang, Z. et Wu, R. (2014), Modeling phenotypic plasticity in growth trajectories: a statistical framework, *Evolution*, 68, 81–91.
[7] West, G. B. Brown, J. H. et Enquist, B. J. (2001), A general model for ontogenetic growth, *Nature*, 34, 67–89
[8] Wu, R. Ma, C. X. Lin, M. Wang, Z. et Casella, G. (2004), Functional mapping of quantitative trait loci underlying growth trajectories using a transform both sides logistic model, *Biometrics*, 60, 729-738.
[9] Wu, R. Ma, C. X. et Casella, G. (2007), *Statistical genetics of quantitative traits : linkage, maps and QTL*, Springer Verlag.