Inférence démo-génétique : utilisation de l'homozygotie haplotypique pour estimer des changements de taille de population

Jean-Michel Marin[1,2], Coralie Merle [1,2,3] et François Rousset[2,4]

[1] *Institut Montpelliérain Alexander Grothendieck (IMAG), Université de Montpellier*
[2] *Institut de Biologie Computationnelle (IBC), Montpellier*
[3] *Centre de Biologie pour la Gestion des Populations (CBGP), INRA*
[4] *Institut des Sciences de l'Évolution - Montpellier (ISE-M), Université de Montpellier*

Résumé : Le séquençage des génomes de plus en plus complets des individus de divers organismes (bactéries, animaux, humains...) ouvre des perspectives nouvelles dans le domaine de la génétique des populations. Dans ce travail, nous montrons comment on peut tirer profit de longues séquences pour estimer des tailles de populations. Pour ce faire, nous utilisons la distribution empirique des longueurs de séquences conservées entre individus, appelée homozygotie haplotypique. Ces quantités portent beaucoup d'information quant à la taille de la population efficace correspondante.

Nous nous intéressons plus particulièrement à deux scénarios simples : pour le premier la taille de la population est constante ou cours du temps et pour l'autre il y a eu un seul changement de taille dans le passé. On se place dans le cas où l'on dispose du génome (quasi-complet) d'un individu diploïde échantillonné au temps présent dans une population isolée. Pour les deux modèles évoqués ci-dessus, nous proposons une procédure d'estimation de leurs paramètres démographiques : la taille de la population pour le premier, auquel on ajoute, pour le second, le temps du changement de taille et son intensité. Puis, nous introduisons une technique de choix entre les deux scénarios. Enfin, nous donnons des résultats numériques.

Mots-clés inférence démographique, homozygotie haplotypique, choix de modèles

Abstract. Sequencing more complete genomes for various organisms (bacteria, animals, humans...) opens new perspectives in the field of population genetics. We show how it is possible to use long sequences of few individuals to estimate population sizes. To do this, we use the empirical distribution of the lengths of conserved sequences between individuals called haplotype homozygosity. These quantities carry a lot of informations on the effective population sizes.

We are particularly interested in two scenarios : one for which the size of the population is constant over time and the other such that there is only one change in the population size. We assume that we have the genome of one diploid individual sampled at the present time in an isolated population. For the two models mentioned above, we propose a methodology to estimate the corresponding demographic parameters : the effective population size for the first one, and also the time and intensity of the change for the second. Then, we introduce a model choice technique between the two scenarios. Finally, we present some numerical results.

Keywords demographic inference, haplotype homozygosity, model choice

1 Introduction

L'utilisation de données génétiques pour inférer l'histoire démographique de populations est un champ de recherche extrêmement actif et depuis de nombreuses années. Traditionnellement, on échantillonne au temps présent un nombre plutôt faible d'individus pour lesquels l'information génétique disponible est portée par un nombre limité de locus (positions sur le génome). L'évolution des techniques de séquençage permet maintenant d'accéder à une information génétique de plus en plus importante. On dispose notamment de génomes presque complets pour les espèces modèles.

Dans le cas d'un organisme eucaryote diploïde, le génome d'un seul individu, comme mosaïque des génomes de tous les individus de la population dont il est issu, porte beaucoup d'information sur la taille efficace de cette population. Avec ce génome comme seule donnée pour inférer la taille efficace de la population, nous proposons une méthodologie basée sur la distribution des longueurs de séquences conservées, appelée Homozygotie Haplotypique (notée HH) .

Plusieurs approches ont été développées pour inférer l'histoire démographique d'une population à partir de données de type *HH*. Dans MacLeod et al. [2009] et MacLeod et al. [2013], les auteurs mettent en évidence des formules explicites, mais couteuses en temps de calcul, de la probabilité pour une paire d'haplotypes de partager un nombre donné de marqueurs adjacents identiques. Elles sont basées sur un modèle de coalescent approché. À partir de ces formules, ils proposent une estimation de l'histoire démographique par un algorithme itératif d'ajustement entre la version théorique et une estimation empirique calculée en utilisant une statistique de scan des données observées. La procédure statistique, qui est très heuristique, aboutit à la mise en évidence de modèles complexes. Harris and Nielsen [2013], quant à eux, proposent une estimation des paramètres démographiques par une technique de pseudo-vraisemblance basée également sur un coalescent approché mais, avec comme données, les spectres de segments identiques par états (IBS) attendus et observés. Par ailleurs, Palamara et al. [2012] estiment la taille de la population en utilisant le partage de segments identiques par descendance (IBD) observés dans un intervalle de longueur spécifique. Enfin, Browning and Browning [2013] présentent une méthode non paramétrique d'estimation de la taille efficace récente de la population basée sur les spectres IBS.

Tous ces travaux portent sur l'inférence complète de l'histoire démographique et conduisent à des estimations complexes. Typiquement, la taille est constante par morceaux et le nombre de changements est très important. Ce nombre est clairement associé au niveau de complexité du modèle : plus il est grand, plus le modèle est complexe. Les méthodologies existantes, évoquées ci-dessus, n'en tiennent pas compte et souffrent d'un problème de sur-ajustement : dans de nombreux cas, un modèle démographique beaucoup plus simple pourrait retenir les événements significatifs de l'histoire de la population considérée.

Nous proposons de palier cette difficulté en ciblant une question simple : existe-t-il un changement dans la taille de la population au cours du temps ou bien est-elle constante ? Nous proposons une procédure de choix de modèle entre ces deux configurations emboîtées. Pour ce faire, nous introduisons un critère pénalisé, critère basé sur la comparaison d'homozygoties haplotypiques empiriques et théoriques.

2 Estimation de l'histoire démographique

Ce travail se concentre sur un modèle démographique à une seule population isolée ayant subi des variations de taille dans le passé. Nous considérons deux scénarios démographiques : le modèle $\mathcal{M}_0$ pour lequel la taille efficace est constante et le modèle $\mathcal{M}_1$ pour lequel il y a un changement de taille. Pour $\mathcal{M}_0$, nous avons un seul paramètre à estimer : $\boldsymbol{\theta}_0 = N \in \mathbb{N}_*$ la taille efficace de la population. Pour $\mathcal{M}_1$, il y en a trois $\boldsymbol{\theta}_1 = (N_p, N_a, t) \in \mathbb{N}_*^3$, les tailles efficaces présentes et ancestrales et le temps du changement (en nombre de générations).

2.1 Estimation des paramètres

Soit $HH(i)$ la probabilité pour une paire d'haplotypes de longueur i d'être identiques. À partir d'un coalescent simplifié, pour un modèle démographique constant par morceaux, MacLeod et al. [2009] donne un attendu théorique noté $HH_{\text{th}}(\boldsymbol{\theta}, i)$ de $HH(i)$ où $\boldsymbol{\theta}$ contient les différentes tailles efficaces et les temps de changement.

Pour les modèles $\mathcal{M}_0$ et $\mathcal{M}_1$, nous comparons cet attendu théorique à une version empirique calculée à partir du génome d'un individu diploïde. La valeur empirique de $HH(i)$ notée $\widehat{HH}(i)$ est donnée par la proportion d'homozygotes entre les deux brins sur des fenêtres tirées aléatoirement dans chaque chromosome. C'est différent de la quantité utilisée par MacLeod et al. [2009] où toutes les fenêtres de longueurs i sont considérées, ceci afin d'éviter la dépendance trop forte entre les différentes quantités empiriques et l'effet de lissage associé.

Nous estimons $\boldsymbol{\theta}_0$ et $\boldsymbol{\theta}_1$ par

$$\widehat{\boldsymbol{\theta}}_0 \in \arg\min_{\boldsymbol{\theta}_0} \sum_{i \in \mathcal{I}} \left(\frac{HH_{\text{th}}(\boldsymbol{\theta}_0, i) - \widehat{HH}(i)}{\widehat{HH}(i)} \right)^2$$

$$\widehat{\boldsymbol{\theta}}_1 \in \arg\min_{\boldsymbol{\theta}_1} \sum_{i \in \mathcal{I}} \left(\frac{HH_{\text{th}}(\boldsymbol{\theta}_1, i) - \widehat{HH}(i)}{\widehat{HH}(i)} \right)^2$$

où $\mathcal{I}$ est un sous-ensemble de longueurs judicieusement sélectionné. L'évaluation de la fonction HH_{th} étant coûteuse en temps de calcul, nous utilisons un outil spécifique à l'optimisation de fonctions boîtes noires pour résoudre numériquement ces deux problèmes d'optimisation. La méthodologie mise en oeuvre est séquentielle : à chaque étape, on utilise une estimation par krigeage de la fonction boîte noire et on calcule de nouveaux points dans la zone du maximum du modèle réduit.

2.2 Choix de modèles

On souhaite maintenant sélectionner entre les modèles $\mathcal{M}_0$ et $\mathcal{M}_1$. Les modèles ayant des complexités différentes, $\mathcal{M}_0$ est emboîté dans $\mathcal{M}_1$, on ne peut pas utiliser le critère mis en oeuvre pour l'estimation des paramètres. Cela résulterait en le choix systématique du modèle $\mathcal{M}_1$.

Nous proposons d'introduire une pénalisation basée sur le calcul d'indices de sensibilités de Sobol. L'un des enjeux majeurs de l'analyse de sensibilité est d'identifier les paramètres les moins influant pour réduire la dimension d'un modèle, et quantifier l'erreur ainsi commise (voir Sobol

et al. [2007], Saltelli et al. [2010]). Notre idée est de mesurer, pour chaque i, la part de variance de $HH_{\text{th}}(i)$ expliquée par chacun des paramètres du modèle le plus complexe, à savoir $\mathcal{M}_1$.

Pour cela, on considère les paramètres d'entrée comme des variables aléatoires et HH_{th} comme une fonction à évaluer sur ces paramètres. On décompose la variance de la sortie en la somme des variances attribuées à chaque paramètre d'entrée : indices de Sobol de premier ordre, et aux interactions entre les différents paramètres : indices de Sobol d'ordre supérieur. Ils sont compris entre 0 et 1 est leur somme vaut 1. Nous utiliserons uniquement les indices de Sobol de premier ordre correspondant à chaque paramètre. Ces indices sont estimés sous le modèle le plus complexe $\mathcal{M}_1$ grâce à la bibliothèque R sensitivity.

Notre critère de choix de modèle est basé sur une formule des moindres carrés dans laquelle, pour chaque i, l'écart de la prédiction $HH_{\text{th}}(\widehat{\boldsymbol{\theta}}_j, i)$ à la valeur observée $\widehat{HH}(i)$ va être affecté d'un poids dit de sensibilité $w_j(i)$. Nous choisissons de prendre pour $w_j(i)$ la somme des indices de Sobol d'ordre 1 des paramètres du modèle $\mathcal{M}_j$ sur lequel on évalue le critère. Pour le modèle $\mathcal{M}_0$, le poids $w_0(i)$ correspond à l'indice de Sobol d'ordre un relatif au paramètre N. Pour le modèle $\mathcal{M}_1$ en revanche le poids $w_1(i)$ correspond à la somme des indices de Sobol d'ordre un des trois paramètres $(N_p, N_a, t) \in \mathbb{N}_*^3$. L'idée est de ne comptabiliser que la part d'erreur que le modèle avait la capacité d'expliquer. C'est une forme de pénalisation du modèle en lien avec sa complexité. En effet, plus on réduit la complexité du modèle, en fixant des paramètres par rapport au modèle le plus complexe, et plus la somme des indices de Sobol des paramètres restants décroit. Une même erreur relative pénalisera ainsi plus fortement un modèle plus complexe qu'un modèle plus simple. Le critère de choix de modèle est finalement :

$$\arg \min_{j \in \{0,1\}} \sum_{i \in \mathcal{I}} w_j(i) \left(\frac{HH_{\text{th}}(\widehat{\boldsymbol{\theta}}_j, i) - \widehat{HH}(i)}{\widehat{HH}(i)} \right)^2$$

3 Résultats numériques

Nous présenterons des résultats numériques sur un modèle de contraction de la taille de la population dans le passé, sur des données simulées pour différentes dates et intensités de la contraction. Nous analyserons également un jeu de donnée issu d'un génome bovin de la race Holstein.

Remerciements

Ces travaux ont été en partie financés par le LabEx NUMEV (Solutions Numériques, Matérielles et Modélisation pour l'Environnement et le Vivant, ANR-10-LABX-20) et le LabEx CeMEB (Centre Méditerranéen de l'Environnement et de la Biodiversité).

Références

B. Browning and S. Browning. Improving the Accuracy and Efficiency of Identity by Descent Detection in Population Data. *Genetics*, 194(2) :459–471, 2013.

K. Harris and R. Nielsen. Inferring demographic history from a spectrum of shared haplotype lengths. *PLoS genetics*, 9(6) :e1003521, 2013.

I. MacLeod, T. Meuwissen, B. Hayes, and M. Goddard. A novel predictor of multilocus haplotype homozygosity : comparison with existing predictors. *Genetics research*, 91(6) :413–426, 2009.

I. MacLeod, D. Larkin, H. Lewin, B. Hayes, and M. Goddard. Inferring Demography from Runs of Homozygosity in Whole-Genome Sequence, with Correction for Sequence Errors. *Molecular Biology and Evolution*, 30(9) :2209–2223, 2013.

P. Palamara, T. Lencz, A. Darvasi, and I. Peer. Length distributions of identity by descent reveal fine-scale demographic history. *The American Journal of Human Genetics*, 91(5) : 809–822, 2012.

A. Saltelli, P. Annoni, I. Azzini, F. Campolongo, M. Ratto, and S. Tarantola. Variance based sensitivity analysis of model output. Design and estimator for the total sensitivity index. *Computer Physics Communications*, 181(2) :259–270, 2010.

M. Sobol, S. Tarantola, D. Gatelli, S. Kucherenkoc, and W. Mauntz. Estimating the approximation error when fixing unessential factors in global sensitivity analysis. *Reliability Engineering & System Safety*, 92(7) :957–960, 2007.