HAL open science

Coopération entre Optimisation Combinatoire et Statistiques pour la Sélection animale

Julie Hamon, Clarisse Dhaenens, Julien Jacques, Gaël Even

▶ **To cite this version:**

Julie Hamon, Clarisse Dhaenens, Julien Jacques, Gaël Even. Coopération entre Optimisation Combinatoire et Statistiques pour la Sélection animale. 13e congrès annuel de la Société française de Recherche Opérationnelle et d'Aide à la Décision, Apr 2012, Angers, France. hal-00749727v1

HAL Id: hal-00749727

https://inria.hal.science/hal-00749727v1

Submitted on 8 Nov 2012 (v1), last revised 5 Mar 2013 (v2)

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Coopération entre Optimisation Combinatoire et Statistiques pour la Sélection animale

Julie Hamon[1,3,4], Clarisse Dhaenens[1,3], Julien Jacques[2,3], Gaël Even[4]

[1] Université Lille 1 (LIFL, UMR CNRS 8022)
clarisse.dhaenens@lifl.fr, julie.hamon@inria.fr
[2] Université Lille 1 (Laboratoire Painlevé, UMR CNRS 8524)
julien.jacques@inria.fr
[3] INRIA Lille-Nord Europe
[4] GÈNES DIFFUSION 3595 Route de Tournai, BP 70023, 59501 DOUAI Cedex
g.even@genesdiffusion.fr

Mots-clés : *sélection génomique, optimisation, régression*

1 Introduction : problème de sélection animale

La génomique a grandement évolué avec le développement récent des technologies haut-débit en séquençage puis en génotypage. Dans le domaine animal, nous sommes aujourd'hui capables de lire les informations génomiques sur près de 800 000 marqueurs sur des ensembles d'individus de plus en plus larges (de 3 000 à 10 000). Ces données peuvent donner lieu à des études d'association entre les marqueurs (GWAS : Genome-Wide Association Studies) ou, plus récemment, à des modéles prédictifs utilisant l'information génomique. Outre les contraintes biologiques (stockage des échantillons, manipulations longues et coûteuses...), l'analyse de données (étude et extraction de connaissances) doit aussi être adaptée en terme de méthodologie et d'architecture matérielle et logicielle.

La Sélection Génomique est une méthode d'évaluation génétique des animaux à partir de leur ADN (suite à un prélèvement biologique de type sang, poils ou biopsie), qui utilise un nombre très élevé de marqueurs couvrant l'intégralité du génome. Le principe de base a été établi par Meuwissen, Hayes et Goddard en 2001 [5]. Dans ce contexte, une problématique importante de la sélection génomique consiste à rechercher des marqueurs explicatifs (ou combinaisons de marqueurs) pour un phénotype (trait caractérisant un animal) sous étude.

Le challenge consiste donc à élaborer des modèles prédictifs permettant, à partir de données génomiques, de déterminer les individus les plus performants selon certains critères quantitatifs.

2 Un problème d'extraction de connaissances

L'augmentation actuelle du nombre de marqueurs rend l'application de méthodologies séquentielles (analyse des marqueurs un par un par régression linéaire) non adaptée et extrêmement coûteuse en temps de calcul, et ne prend en compte aucune interaction éventuelle entre marqueurs (par exemple la redondance d'information entre deux marqueurs expliquant une même zone du génome). Parmi les méthodes classiquement utilisées pour aborder ce type de problème, nous pouvons citer LASSO [6], PLS [3] ou BLUP (modèle spécifique à la génétique animale) [2]. Cependant, afin d'obtenir un modèle prédictif compréhensible, et interprétable, une autre approche consiste à réaliser conjointement une sélection des attributs (marqueurs) pertinents et une recherche de modèle explicatif, ce qui se modélise de la façon suivante :

$$Y_i = \beta_0 + \sum_{j=1}^{p} (\beta_j z_j X_{ij}) + \epsilon_i \ ,$$

Avec Y un trait d'intérêt ($Y \in \mathbb{R}$), X_j les SNPs étudiés ($X_j \in \{-1, 0, 1\}$), ϵ_i i.i.d $\sim N(0, \sigma^2)$, et $z_j = 1$ si la variable j est sélectionnée, 0 sinon.

Cette problématique de sélection d'attributs très combinatoire est NP-difficile. Nous proposons une approche alliant méthodes statistiques et méthodes d'optimisation combinatoire.

3 Approche : optimisation combinatoire et statistique

L'objectif est d'estimer les β_j et z_j. z est un vecteur discret, dans un ensemble fini $\{0, 1\}^p$ et est donc impossible à estimer quand p (nombre d'attributs) est grand. L'utilisation alternée d'un algorithme d'optimisation et d'une méthode statistique permet d'une part, d'optimiser z connaissant β et σ à l'aide de la méthode d'optimisation et d'autre part, d'optimiser β et σ connaissant z à l'aide de la méthode statistique.

La première approche proposée consiste en une hybridation entre une recherche locale itérée (ILS - Iterated Local Search) et une régression. La recherche locale proposée travaille sur l'espace des sous-ensemble d'attributs représenté par un vecteur binaire indiquant si l'attribut est sélectionné ou non. L'exploration du voisinage des solutions consiste alors à ajouter ou supprimer un attribut (voisinage de type bit-flip) pour au final converger vers une sélection d'un nombre réduit d'attributs (marqueurs génétiques). Lorsqu'un optimum local est atteint une étape de perturbation permet de prolonger la recherche. Chaque solution (sélection d'attributs) est évaluée à l'aide d'un critère statistique calculé à partir de l'estimation d'un modèle de régression. Deux critères d'évaluation, calculés suivant le schéma ci-dessous, ont été comparés : BIC (Bayesian Inference Criterion) et CVE (Cross-Validation Error).

$$z_j \xrightarrow{Regression} \text{Vraisemblance} \xrightarrow{BIC} \text{fitness}$$

$$z_j \xrightarrow{CVE(Regression)} \text{fitness}$$

4 Validation

Afin d'évaluer la qualité de la méthode nous utilisons des données simulées et les données du workshop QTLMAS 2008 [4]. Nous comparons nos résultats et performances, avec trois approches statistiques généralement utilisées pour prédire un trait quantitatif à partir d'un grand nombre de marqueurs [1] : une méthode de régression avec sélection de variables pas à pas, une méthode de régression pénalisée (LASSO [6]) et une méthode de régression sur combinaison des variables (PLS [3]). Les résultats obtenus montrent que l'approche proposée permet d'identifier des modèles explicatif intéressants.

Références

[1] T. Hastie, R. Tibshirani, J. Friedman. The Elements of Statistical Learning, 2009.

[2] C. R. Henderson. Best Linear Unbiased Estimation and Prediction under a Selection Model. *Biometrics*, 31(2) :423-447, 1975.

[3] N. Long, D. Gianola, G.J.M Rosa, and K. A Weigel. Dimension reduction and variable selection for genomic selection : application to predicting milk yield in holsteins. *Journal of Animal Breeding and Genetics*, 128(4) :247-257, August 2011.

[4] M. S. Lund, G. Sahana, D-J. de Koning, G. Su and Ö. Carlborg. Comparison of analyses of the QTLMAS XII common dataset. I : Genomic selection. *BMC Proceedings*, 3(Suppl 1) :S1, 2009.

[5] T. H. E. Meuwissen, B. J. Hayes and M. E. Goddard. Prediction of Total Genetic Value Using Genome-Wide Dense Marker Maps. *Genetics Society of America*, 2001.

[6] T. T. Wu, Y. F. Chen, T. Hastie, E. Sobel and K. Lange. Genome-wide association analysis by Lasso penalized logistic regression. *Bioinformatics*, 25(6), 2009.