Combining Combinatorial Optimization and Statistics to Mine High-throughput Genotyping Data

Julie HAMON[1,3], Clarisse DHAENENS[1], Julien JACQUES[2] and Gaël EVEN[3]

[1] LIFL, Université Lille 1 / INRIA Lille-Nord Europe
clarisse.dhaenens@lifl.fr, julie.hamon@inria.fr
[2] Laboratoire Painlevé, UMR CNRS 8524 & Université Lille 1 / INRIA Lille-Nord Europe
Julien.Jacques@inria.fr
[3] GENES DIFFUSION, 3595 Route de Tournai, BP 70023, 59501 DOUAI Cedex
g.even@genesdiffusion.com

Keywords genomic selection, optimization, regression.

Coopération entre Optimisation Combinatoire et Statistiques pour l'Analyse de données de Génotypage haut-débit

Depuis quelques années, la génomique a grandement évolué avec le développement de nouvelles technologies telles que le séquençage et le génotypage haut-débit. En ce qui concerne 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 (Genome-Wide Association Studies). Outre les contraintes biologiques (stockage des échantillons, manipulations longues et coûteuses...), la partie analyse de données (extraction de connaissances) doit aussi être adaptée en terme de méthodologie et d'architecture matérielle et logicielle. L'objectif est d'élaborer des modèles prédictifs permettant, à partir des données génomiques, de déterminer les individus les plus performants selon certains critères quantitatifs de sélection animale. Pour cela, l'objectif théorique est à terme de définir de nouvelles méthodes permettant la coopération entre statistique et optimisation combinatoire spécifiquement dédiées aux données issues de génotypage haut débit en vue d'une implémentation.

1 La Sélection Génomique dans le Domaine Animal

En génétique, on admet que plusieurs zones chromosomiques, portant un ou plusieurs gènes, sont impliquées dans le contrôle de caractères quantitatifs (production de lait, fertilité...), et que de nombreux allèles (identifiés sous forme de marqueurs) sont responsables de la variabilité. On appelle ces zones QTL : Quantitative Trait Loci. La Sélection Génomique est une méthode d'évaluation génétique des animaux grâce à 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 établie par Meuwissen, Hayes et Goddard en 2001 [1]. Elle ne prend pas en compte les régions chromosomiques (QTL) mais exploite une densité de marqueurs suffisante si bien que chaque QTL se trouve en déséquilibre de liaison avec au moins un marqueur. Les effets des SNP (Single Nucleotide Polymorphism) sont estimés en associant les génotypes aux valeurs phénotypiques d'animaux déjà indexés. Grâce à leur détection, on peut calculer l'index propre d'un animal. Dans ce contexte, une problématique importante de la sélection génomique consiste à rechercher des marqueurs explicatifs (ou combinaisons de marqueurs) pour un phénotype sous étude. Il est à noter que l'augmentation actuelle du nombre de marqueurs (777 000 marqueurs en bovins) 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. Nous proposons d'aborder ce problème en combinant deux approches de la littérature.

2 Approches Statistiques Existantes

Deux types de modèles statistiques sont généralement utilisés pour prédire un trait quantitatif à partir d'un grand nombre de marqueurs génétiques [2] :

- les méthodes de régression pénalisées :
  - la régression Ridge qui consiste à imposer une pénalité L^2 sur les coefficients de la régression.
  - la régression LASSO (Least Absolute Shrinkage and Selection Operator), qui en imposant une pénalité L^1, réduit des coefficients à 0, et donc sélectionne des variables (marqueurs génétiques) [3].
  - la régression Oscar dont la pénalité conduit à annuler certains coefficients de régression et à en regrouper d'autres en mettant leurs coefficients égaux [4].
- Les méthodes de régression sur combinaison des variables d'entrées :
  - PCA (Principal Components Analysis) - MCA (Multiple Correspondance Analysis)
  - SPCA (Sparse Components Analysis) : intégration d'une pénalité de type LASSO dans la détermination des composantes principales.
  - PLS (Partial Least Square)

On notera que les méthodes de régression sur combinaison des variables d'entrées ne permettent pas de sélectionner un nombre réduit de SNP et sont difficilement interprétables.

3 Optimisation Combinatoire

Les problématiques d'analyse liées aux données génomiques peuvent également être vues, dans la plupart des cas, comme des problèmes d'optimisation combinatoire. L'utilisation de méthodes d'optimisation combinatoire pour l'extraction de connaissances permet d'accélérer l'analyse en présélectionnant des sous-ensembles d'attributs intéressants, et de proposer de nouvelles méthodes permettant d'identifier des zones d'intérêts. Etant donnée la taille de l'espace de recherche (constitué de toutes les combinaisons de marqueurs possibles), seules des méthodes de types méta-heuristiques peuvent être mises en œuvre de façon efficace. Parmi ces méthodes, nous pouvons citer les algorithmes de recherche locale (descente, recherche tabou...) et les algorithmes à base de population (algorithme génétique) qui ont déjà fait leur preuve dans ce domaine [5].

4 Approche proposée : Optimisation Combinatoire et Statistique

L'objectif de ce travail est de définir un modèle de prédiction des traits des animaux à partir des marqueurs génétiques, utilisant à la fois la puissance exploratoire des algorithmes d'optimisation combinatoire et la spécificité des modèles statistiques de régression [1]. Nous choisissons comme première approche d'effectuer une sélection d'attributs en combinant une méthode d'optimisation de type recherche locale avec une régression RIDGE. A chaque étape de la recherche locale, nous évaluons la sélection d'attributs à l'aide d'un critère de type CVE (Cross Validation Error) calculé sur les prédictions par un modèle de régression, pour au final converger vers une sélection d'un nombre réduit de SNP, et vers un modèle de régression sur ces SNP.

Afin d'évaluer la qualité de la méthode nous utiliserons les données de XII QTLMAS 2008 et comparerons nos résultats et performances avec ceux des méthodes de sélection génomique présentées lors de ce Workshop.

Pour ce faire, nous utiliserons la plateforme ParadisEo développée par des membres de l'équipe DOLPHIN-INRIA, en C++. La solution proposée pourra également être parallélisable et déployable sur des architectures de calcul haute-performance (Cluster ou Grille de Calcul).

Références

[1] 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.

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

[3] T.T. Wu, Y.F. Chen, T. Hastie, E. Sobel and K. Lange, Genome-wide association analysis by lasso penalized logistic regression, *bioinformatics*, vol. 25, no. 6, 2009.

[4] H.D. Bondell and B.J. Reich, Simultaneous regression shrinkage, variable selection and clustering of predictors with OSCAR, *Biometrics*, vol. 64, p. 115-123, 2008.

[5] C. Dhaenens, *Optimisation Combinatoire Multi-Objectif : Apport des Méthodes Coopératives et Contribution à l'Extraction de Connaissances*, HDR, Université Lille 1, 2005.