MODELISATION DE L'INCERTITUDE DES CARTES COMPAREES ET APPLICATION A L'ETUDE DES ASSEMBLAGES GENOMIQUES

SERVIN Bertrand[1], FARAUT Thomas[1] et de GIVRY Simon[2]
[1] *UMR de Génétique Cellulaire, 31426 Castanet-Tolosan*
[2] *Laboratoire de Mathématiques et Informatique Appliquées, 31426 Castanet-Tolosan*

INTRODUCTION

L'arrivée de puces de génotypage SNP à haut débit chez plusieurs espèces animales offre d'importantes perspectives pour la caractérisation de gènes d'intérêt agronomique et l'étude de l'évolution des populations d'animaux d'élevage. Cependant, avant toute exploitation génétique de ces données, il est nécessaire de cartographier les SNP, c'est à dire de déterminer leur ordre sur le génome de l'espèce. Le fait que ces marqueurs soient très nombreux pose un problème particulier pour leur cartographie. En attendant que les séquences complètes des génomes nous fournissent les cartes ultimes des espèces, il faut recourir à des méthodes expérimentales telles que la cartographie par hybrides d'irradiation (RH).

Pour améliorer les cartes RH, Faraut et al. (2006) ont proposé une méthode de cartographie comparée permettant de prendre en compte l'information d'ordre des marqueurs sur un génome de référence. Dans ce travail nous proposons une extension de cette approche permettant d'évaluer l'incertitude existante sur la carte ainsi obtenue..

MATÉRIEL et MÉTHODES.

Le principe de la méthode de cartographie comparée proposée par Faraut et al. (2006) est de rechercher l'ordre O* qui maximise la probabilité a posteriori P(O|Y,Oref), proportionnelle a P(Y| O)P(O|Oref), où O est un ordre, Y sont les données RH et Oref un ordre chez un génome de référence. P(O|Oref) est un a priori sur l'ordre des marqueurs basé sur l'ordre de référence et P(Y|O) est la vraisemblance RH. Dans la méthode de Faraut et al. le maximisation de cette probabilité est obtenue en transformant le problème en problème du voyageur de commerce.

Pour étendre cette approche et mesurer l'incertitude sur la carte obtenue, nous cherchons ici à caractériser la distribution des cartes P(O| Y,Oref) autour de la carte optimale O*. A cette fin nous avons développé un algorithme MCMC pour échantillonner des cartes dans cette distribution.

Pour évaluer la performance de notre approche, nous avons effectué des simulations. Nous avons simulé une zone chromosomique de 25 Mb correspondant à une distance de cassure de 5 Rays, sur lequel nous avons placé aléatoirement 100 marqueurs. L'ordre de référence a été simulé en effectuant 4 réarrangements évolutifs. Le panel d'hybride simulé comportait 100 hybrides.

Nous présentons des données obtenues par le groupe d'André Eggen du LGBBC (Jouy-en-Josas) portant sur le typage d'une puce 60K du génome bovin sur deux panels RH distincts.

RÉSULTATS

Les résultats issus de nos simulations montrent que l'incertitude estimée par notre algorithme est globalement bien estimée. Par exemple, parmi les couples de marqueurs dont la probabilité d'être adjacents est estimée à p, les couples de marqueurs réelement adjacents sont effectivement en fréquence p. Nous montrons également comment identifier les zones dont l'orientation par rapport à la carte optimale est incertaine.

Finalement, les résultats obtenus sur les données bovines suggèrent que la qualité de l'assemblage bovin actuel varie beaucoup entre les différents chromosomes. Nous présentons des exemples pour certains chromosomes.

DISCUSSION

La méthode présentée permet d'apporter une information supplémentaire a la carte construite en utilisant la méthodologie de Faraut et al. (2006) en permettant de distinguer les zones de la carte bien supportées par les données de zones plus sujettes a caution. Une importante application de notre approche, que nous sommes en train d'approfondir est l'évaluation des assemblages basée sur l'évaluation de l'incertitude ainsi caractérisée.

REFERENCES

Faraut T., S. de Givry, P. Chabrier, T. Derrien, F. Galibert, C. Hitte et T. Schiex, 2006, Bioinformatics 23 p. e50-e56.