# Application des techniques du voyageur de commerce à la production de cartes génétiques

P. Chabrier, C. Gaspin, S. de Givry et T. Schiex

*INRA / BIA, Chemin de Borde Rouge, BP 27, 31326 Castanet-Tolosan cedex*
*(chabrier, gaspin, degivry, tschiex)@toulouse.inra.fr*

**Mots-clefs** : optimisation combinatoire, problème du voyageur de commerce (TSP), ordonnancement de marqueurs, ordre de vraisemblance maximum.

## 1 Introduction

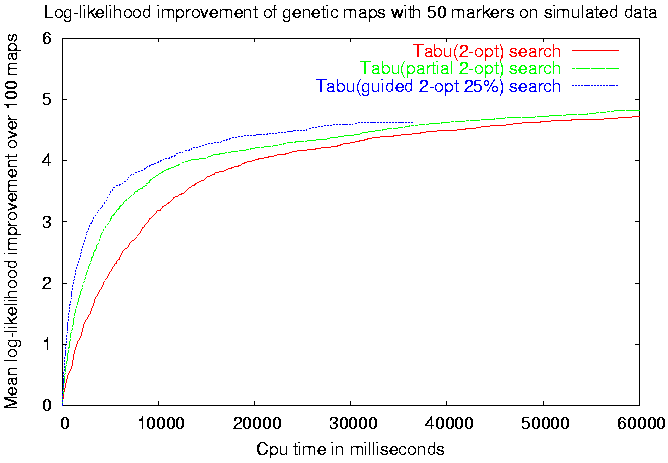
La **génétique** [3] est la branche de la biologie qui étudie les caractères héréditaires, leur variabilité et transmission. En simplifiant, les caractères, par exemple la couleur des cheveux, sont codés et transmis à la descendance par les **gènes**. La valeur d'un caractère, par exemple blond ou châtain, est appelée **allèle**. Les gènes sont regroupés en séquence pour former des **chromosomes**. Tout être vivant est formé de cellules contenant chacune un lot de chromosomes. La division cellulaire produisant les cellules sexuelles (méiose) effectue un brassage des chromosomes assurant la diversité des individus d'une même espèce. En particulier des morceaux de paire de chromosomes issus du père et de la mère, contenant les mêmes gènes mais pas les mêmes allèles, sont échangés (recombinaison d'allèles).

La construction d'une **carte génétique** est une étape importante de l'analyse des gènes d'une espèce. Il s'agit de repérer la position des gènes sur les chromosomes. Cette information peut être obtenue par **analyse de pedigree**. Les gènes sont identifiés par l'existence, dans une descendance donnée, d'une forme normale et d'une forme mutée donnant lieu à deux allèles observables dont la distribution peut être suivie de génération en génération. Du fait du brassage lors de la méiose, deux gènes proches sur un chromosome apparaîtront transmis « en bloc » à la descendance (les allèles correspondants resteront associés). Au contraire, deux gènes très éloignés ou sur des chromosomes différents seront transmis indépendamment. La fréquence de recombinaison des allèles augmente avec leur éloignement. A partir d'observations généralement incomplètes des allèles de plusieurs individus, le problème de l'**ordonnancement de marqueurs** (associés aux gènes identifiables) consiste à déterminer un ordre des marqueurs qui maximise la vraisemblance (*likelihood* en anglais) des observations. Dans le cas d'observations complètes, ce problème admet une reformulation comme une variante du problème de voyageur de commerce symétrique (*TSP* en anglais) [6] : les villes sont les marqueurs, les distances entre deux villes sont calculées en fonction de la fréquence de recombinaison entre les deux marqueurs et l'objectif est de trouver un chemin *hamiltonien* (et non un cycle), visitant chaque ville exactement une fois, qui minimise la somme des distances du chemin. Dans le cas d'observations incomplètes, la reformulation n'est plus possible. En particulier, la distance entre deux villes dépend du chemin[1]. Cependant [6] fait l'hypothèse que les deux problèmes gardent une structure proche et que les techniques de recherche locale appliquées au TSP devraient également donner de bons résultats pour l'ordonnancement de marqueurs. La *recherche tabou* [1] implémentée dans [6] semble donner de bons résultats en temps limité. Le but de nos travaux a été d'améliorer l'algorithme existant en se servant des résultats de [1] et [4] et en utilisant des connaissances sur le problème pour guider la recherche.

## 2 Recherche tabou améliorée

L'algorithme existant utilise un voisinage ***2-opt*** [1]. La taille de ce voisinage est $S=N.(N-1)/2 -1$, avec $N$, le nombre de marqueurs. Chaque mouvement est évalué par l'algorithme EM (coûteux en temps). Le meilleur mouvement *2-opt* n'appartenant pas à la *liste tabou* est retenu comme nouvel ordre à chaque itération de l'algorithme. L'ordre initial étant produit par diverses heuristiques dont l'heuristique du plus proche voisin (*NN*) qui utilise des distances calculées en ignorant les observations manquantes. La liste tabou mémorise les $T.S$ derniers mouvements retenus, avec $0 \leq T \leq 1$ variant à chaque itération de façon stochastique dans un intervalle donné. Une exception à la règle du tabou a lieu si le mouvement conduit à un ordre meilleur que le dernier meilleur ordre rencontré (critère d'aspiration). Le nombre d'itérations, i.e. d'examens complets d'un voisinage, est défini à $2.N + L$, avec $L$, une valeur augmentée dynamiquement à chaque nouveau meilleur ordre.

[1] Un algorithme itératif d'optimisation statistique, **EM** (*Expectation/Maximization*) [2], prend en compte des données incomplètes. Pour un jeu de données et un ordre fixé des marqueurs, il détermine les fréquences de recombinaison entre deux marqueurs consécutifs qui maximisent la vraisemblance des observations. L'objectif de l'ordonnancement de marqueurs est de trouver un ordre qui maximise la vraisemblance calculée par EM.




Une première amélioration porte sur le choix de l'heuristique fournissant l'ordre initial. L'heuristique *Greedy* [1], qui tente d'insérer d'abord les paires de marqueurs les plus proches, s'avère meilleure que l'heuristique NN en moyenne sur des jeux de données générés aléatoirement (ordres 30 fois plus vraisemblables pour *N=50*). En pratique, *Greedy* est également plus rapide en temps de calcul (7 fois plus rapide pour *N=50*). Une seconde amélioration concerne l'exploration du voisinage. L'exploration se termine prématurément dès qu'un mouvement améliore le meilleur ordre trouvé jusque là [4]. Cette modification mineure (notée *partial 2-opt* dans la figure) procure une nette amélioration du profil *anytime* de la recherche comme le montre la figure ci-contre. Enfin nous introduisons une connaissance du domaine : les ordres de forte vraisemblance ont tendance à minimiser la somme des taux de recombinaison entre marqueurs consécutifs. Nous choisissons d'ordonner avant de les évaluer tous les mouvements du voisinage en fonction des taux de recombinaison de l'ordre précédemment retenu. Un mouvement est évalué en priorité si la moyenne des taux de recombinaison des deux paires de marqueurs « cassées » par *2-opt* est élevée. Seuls les *X* meilleurs mouvements sont évalués. En pratique *X=25%* donne les meilleurs résultats (notés *guided 2-opt 25%* dans la figure). De plus la recherche termine environ 4 fois plus tôt pour *N=50*. Notons que les résultats obtenus n'améliorent pas la vraisemblance de l'ordre aux termes des recherches tabou. Par contre l'amélioration est effective dans un contexte de temps limité, lorsque la recherche tabou est appliquée à des problèmes réels de grande dimension (nécessitant plusieurs jours de traitement).

# 3 Discussion

Un logiciel en domaine publique, ***Carthagene*** (http://www.inra.fr/bia/T/carthagene) [6], permet de construire et fusionner des cartes génétiques à partir de jeux de données variés (*backcross*, *intercross*, *outbred* et *hybrides irradiés*). Il se compare favorablement avec ses principaux concurrents *MAP-MAKER* [5] et *JOINMAP* [7]. Nous avons expérimenté d'autres voisinages (*2-opt* & insertion de sommets ; *3-opt* bridé) sans obtenir de gain. La difficulté vient du fait que la fonction objectif est assimilée à une boîte noire (l'algorithme EM). L'évaluation d'un mouvement n'est pas incrémentale, car on ne dispose pas d'une méthode analytique pour calculer les taux de recombinaison pour toute paire de marqueurs. Ainsi certains critères utiles dans les TSP font défauts ici. Par exemple, le critère de croissance du gain dans la construction d'un mouvement *k-opt* pour la procédure *Lin-Kernighan* [4]. Une stratégie de recherche intéressante serait d'augmenter progressivement le paramètre *X* (contrôlant la taille du voisinage exploré) pour améliorer le profil *anytime* à la fin de la recherche tabou. Une autre approche serait de résoudre successivement une approximation du problème d'ordonnancement traitée de manière efficace et l'algorithme EM. Par exemple, en travaillant avec des distances entre marqueurs précalculées (hypothèse de données complètes). Puis en mettant à jour les distances en fonction du résultat de l'algorithme EM appliqué au meilleur ordre obtenu dans le problème approché. En pratique, les biologistes ne se contentent pas du critère de vraisemblance maximum. Ils ont également besoin de savoir si l'ordre est robuste, dans le sens qu'il n'existe pas d'autres ordres de vraisemblance proche. *Carthagene* propose de maintenir un ensemble de solutions. Il serait intéressant d'évaluer l'apport d'une phase de diversification dans la recherche tabou.

# Références


[1] E. Aarts et J.K. Lenstra (1998). *Local Search in Combinatorial Optimization*. John Wiley & Sons, Chapitre 8 : *The traveling salesman problem : a case study*, 215-311.

[2] A. Dempster, N. Laird et D. Rubin (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. R. Statist. Soc. Ser.* 39:1-38.

[3] C. Gaillardin (2000). *Gènes et génomes*. Encyclopédie Clares.

[4] K. Helsgaun (2000). An Effective Implementation of the Lin-Kernighan Traveling Salesman Heuristic. *European Journal of Operational Research*, 126(1), 106-130.

[5] E. Lander, P. Green, J. Abrahamson, A. Barlow, M.J. Daly, S.E. Lincoln et L. Newburg (1987). MAP-MAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations. *Genomics*, 1:174-181.

[6] T. Schiex et C. Gaspin (1997). Carthagene : Constructing and Joining Maximum Likelihood Genetic Maps. *Proc. of ISMB'97*, Grèce.

[7] P. Stam (1993). Constructing of integrated genetic linkage maps by means of a new computer package:JOINMAP. *The Plant Journal*, 3(5):739-744.