Markov Chain Monte Carlo pour la reconstruction de réseaux phylogénétiques / Domestication du riz

Intérêt Scientifique

Les réseaux phylogénétiques peuvent décrire les transferts horizontaux de gènes, les phénomènes d'hybridations ainsi que les introgressions (cf. Solis Lemus et Ané, 2016). Les introgressions représentent l'intégration d'allèles d'une population dans une autre existante. L'hybridation intervient lorsque des individus appartenant à 2 populations génétiquement différentes se reproduisent, créant une nouvelle population.

Au sein du projet GenomeHarvest (collaboration avec le CIRAD), nous nous intéressons à l'histoire évolutive du riz. Ce projet a pour objectif une meilleure exploitation de la diversité des génomes dans le cadre des programmes de sélection. Il existe principalement quatre types de riz, à savoir japonica (grains courts), indica (grains longs), aus, et aromatique (e.g. riz basmati). Le processus de domestication du riz s'avère controversé et débattu. Plusieurs thèses s'opposent (Huang et al., 2012, Civan et al. 2015, Choi et al., 2017, ...). Huang et al. (2012) infèrent que japonica aurait été domestiqué à partir d'un riz sauvage dans le sud de la Chine, puis croisé à un sauvage dans le sud est de la Chine, générant indica. A contrario, Civan et al (2015) suggèrent que indica, japonica et aus auraient été domestiqués séparément dans différentes parties d'Asie. Les méthodes utilisées dans ces publications sont d'ordre assez générique (Neighbour-Joining, Analyse en Composantes Principales). Dans notre étude, on s'intéressera à un modèle qui, outre le tri de lignées, considère explicitement les mutations et hybridations. Ainsi, par un apprentissage assez fin du réseau phylogénétique, il sera possible d'inférer plus précisément le processus de domestication. D'autre part, une fois le réseau phylogénétique connu, il est possible d'exploiter cette information, dans l'optique du chromosome painting, qui retrace l'histoire évolutive de chaque portion (i.e. quelques SNPs) du génome. On dispose dès lors d'un arbre de locus, pour chaque locus considéré.

Programme de travail / Méthodes employées/ Résultats obtenus

Une approche sur la base de modèles évolutifs en réseau phylogénétique a été adoptée. Dans un premier temps elle a consisté en filtrage et sélection de données SNPs de riz cultivés, puis en l'inférence indirecte de réseaux phylogénétiques : des phylogénies collectées lors d'un échantillonnage bayésien sur des données SNPs ont servi d'entrée à un logiciel existant d'inférence de réseaux phylogénétiques. En d'autres termes, on apprend par MCMC la distribution à posteriori d'arbre d'espèces sachant les données à l'aide du logiciel SNAPP (Bryant et al., 2012), puis résume cette distribution d'arbres d'espèces en réseau au moyen de SplitsTree (Huson et Bryant, 2006).

Ces premières expériences ont montré que l'on peut retrouver des scénarios proches de plusieurs scénarios proposés dans la littérature pour expliquer l'histoire évolutive des riz cultivés (cf. figures 1 et 2). Néanmoins cette première phase des travaux a mis en lumière l'existence de cas où cette approche indirecte peut passer à côté d'événements importants d'introgression.

En conséquence, nous avons proposé une méthode d'inférence directe de réseau phylogénétique à partir de données SNPs. Cette méthode, une généralisation de SNAPP, s'inscrit dans un cadre bayésien en étendant le modèle MultiSpecies Coalescent connu dans le cadre des phylogénies. Dans sa version originale, SNAPP intègre sur tous les arbres de locus et modélise la distribution a priori de l'arbre d'espèces par un processus de Yule (processus de naissances). La généralisation de SNAPP aux réseaux demeure délicate car les branches du réseau ne peuvent plus être traitées de manière indépendante, et des algorithmes dédiés aux réseaux doivent ainsi être proposés. Nous avons dérivé des équations permettant un calcul analytique de la vraisemblance de données SNPs (pour des variétés actuelles) étant donné un réseau phylogénétique. La distribution a priori du réseau est désormais un processus de Naissance-Hybridation (Zhang et al., 2017). A l'aide d'un échantillonnage par MCMC, on apprend la distribution a posteriori du réseau phylogénétique sachant les données SNPs. L'algorithme est implémenté dans le logiciel BEAST, très employé dans la communauté bioinformatique. Quelques étapes de l'algorithme sont décrites en figure 4. Notre approche s'inscrit dans un cadre généraliste. Rappelons tout d'abord que le niveau d'un réseau phylogénétique désigne le nombre maximal de noeuds avec plus d'un parent parmi toutes les composantes biconnexes du réseau. Contrairement à Solis-Lemus et Ané (2016) qui considèrent uniquement des réseaux de niveau un, notre méthode permet de traiter des réseaux de n'importe quel niveau. De plus, notre méthode prend en entrée des données SNPs et non pas des arbres de gènes estimés lors d'une étape préalable (e.g. Yu et al., 2012, Wen et al., 2016, Solis-Lemus et Ané, 2016).

Notre étude se veut une concurrente directe du récent article Zhu et al. (2018) de l'équipe de Luay Nakhleh. L'approche suivie par Zhu et al. (2018) est basée sur une énumération des partitions possibles des lignées au sein du réseau et sur des calculs de probabilités. La complexité du calcul de vraisemblance (sachant le réseau) est de l'ordre de $O(mn^{4k+4})$, m désignant le nombre d'espèces, n le nombre maximal de lignées dans une espèce, k le niveau du réseau phylogénétique. A contrario, dans notre étude, chaque fois qu'un noeud ou une arête du réseau est traité, cela nécessite un temps en $O(n^{2k'+2})$, où k' représente le nombre maximal d'arêtes impliquées dans le calcul des probabilités conditionnelles (i.e. nombre maximal d'arêtes traversées par les différentes lignes violettes en figure 4). Ce qui conduit à une complexité de l'ordre de $O(sn^{2k'+2})$, où s représente la somme du nombre de noeuds et du nombre d'arêtes du réseau. A titre d'exemple, sur le réseau A en figure 3, notre algorithme présente une complexité en $O(n^6)$ alors que la méthode de Zhu et al. (2018) se veut en $O(n^8)$. De même, sur les réseaux B et C, notre complexité est de l'ordre $O(n^8)$, contrairement aux concurents en $O(n^{12})$. Afin de valider notre méthode, nous avons développé un logiciel SNAPPSimNet construit sur la base du simulateur SNAPPSim de Bryant et al. (2012) et qui génère des données SNPs à l'aide d'arbres de gènes évoluant à l'intérieur d'un réseau selon un processus de coalescence. Nous avons dès lors observé sur données simulées, que plus le vrai réseau s'avèrait compliqué, plus le nombre d'observations échantillonnées par MCMC devait être important afin d'atteindre un nombre d'observations indépendantes (ESS) d'au moins 200 (valeur diagnostique préconisée par la communauté MCMC, pour atteindre la convergence). Ainsi, on se doit d'évaluer la vraisemblance rapidement, afin de pouvoir proposer de nouveaux réseaux par MCMC, d'où l'intérêt de notre approche.

Collaborations Université de Montpellier (C. Scornavacca, F. Pardi, V. Berry), le CIRAD (J-C. Glazmann, J. Santos, A. D'hont), D. Bryant (université d'Otago), et C. Ané (université de Madison).

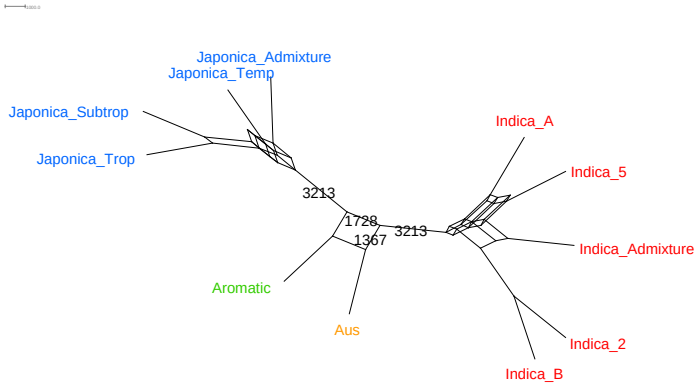


FIGURE 1 – Inférence indirecte de réseau phylogénétique (SNAPP + SplitsTree) à partir du Chromosome 6 du riz (44 variétés, conservation de 1550 SNPs, i.e. un SNP tous les 500 SNPs). Ce réseau serait plutôt en accord avec la thèse de Civan et al. (2015), à savoir Aromatic serait issu d'une hybridation de Japonica avec Aus.

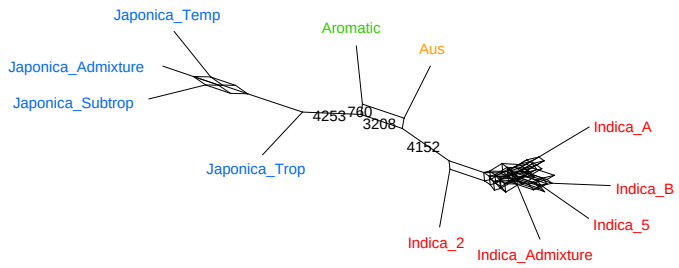
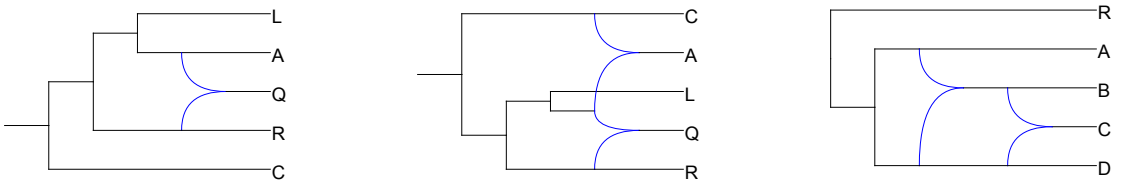
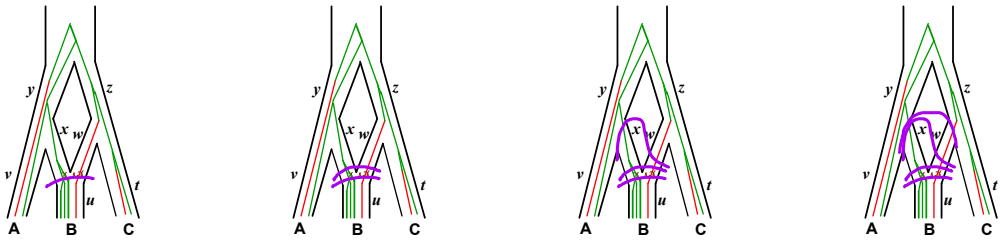


FIGURE 2 – Inférence indirecte de réseau phylogénétique (SNAPP + SplitsTree) à partir du Chromosome 10 du riz (44 variétés, conservation de 1089 SNPs, i.e. un SNP tous les 500 SNPs). Ce réseau serait plutôt en accord avec la thèse de Huang et al. (2012), à savoir Aromatic serait issu de Japonica.



Réseau A Réseau B Réseau C

FIGURE 3 – Les différents réseaux étudiés. Les réseaux A et B sont tirés de la figure 3 de Zhu et al. (2018).



Etape 1 Etape 2 Etape 3 Etape 4

FIGURE 4 – Les quatre premières étapes de notre algorithme sur un réseau très simple (réseau phylogénétique en noir, marqueurs bialléliques rouge/vert, arbre de locus indiquant les mutations au cours du temps). Calcul de la vraisemblance (conditionnelle) des données aux feuilles situées sous la ligne violette la plus ancienne. Le nombre de lignées dans chaque espèce est connu. Détail des calculs par étape : 1) calcul de $\mathbb{P}(r_B = 1 \mid n_{u^{\text{Top}}}, r_{u^{\text{Top}}})$; 2) calcul de $\mathbb{P}(r_B = 1 \mid n_{x^{\text{Bot}}}, r_{x^{\text{Bot}}}, n_{w^{\text{Bot}}}, r_{w^{\text{Bot}}})$; 3) calcul de $\mathbb{P}(r_B = 1 \mid n_{x^{\text{Top}}}, r_{x^{\text{Top}}}, n_{w^{\text{Bot}}}, r_{w^{\text{Bot}}})$; 4) calcul de $\mathbb{P}(r_B = 1 \mid n_{x^{\text{Top}}}, r_{x^{\text{Top}}}, n_{w^{\text{Top}}}, r_{w^{\text{Top}}})$; où r_B désigne le nombre d'allèles rouges dans l'espèce B, $n_{x^{\text{Top}}}$ (resp. $n_{x^{\text{Bot}}}$) le nombre de lignées en haut (resp. bas) d'une branche notée x, et $r_{x^{\text{Top}}}$ (resp. $r_{x^{\text{Bot}}}$) le nombre d'allèles rouge en haut (resp. bas) d'une branche notée x.

Références

Arvestad, L., Lagergren, J., Sennblad, B. (2009). The gene evolution model and computing its associated probabilities. *Journal of the ACM*, **56**, 1-44.

Azaïs, J. M. and Cierco-Ayrolles, C. (2002). An asymptotic test for quantitative gene detection. *Ann. I. H. Poincaré*, **38**, **6**, 1087-1092.

Azaïs, J. M. and Wschebor, M. (2009). *Level sets and extrema of random processes and fields.* Wiley, New-York.

Blanc, G., Wolfe, K.H. (2004). Widespread paleopolyploidy in model plant species inferred from age distributions of duplicate genes. *The Plant Cell Online*, **16**, 1667-1678.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. A., RoyChoudhury, A. (2012). *Inferring species trees directly from biallelic genetic markers : bypassing gene trees in a full coalescent analysis.* Molecular biology and evolution, **29(8)**, 1917-1932.

Choi, J. Y., Platts, A. E., Fuller, D. Q., Wing, R. A., Purugganan, M. D. (2017). The rice paradox : Multiple origins but single domestication in Asian rice. *Molecular biology and evolution.* **34**, (4), 969-979.

Churchill, G. A., Doerge, R. W. (1994). Empirical threshold values for quantitative trait mapping. *Genetics.* **138**, (3), 963-971.

Civan, P., Craig, H., Cox, C. J., Brown, T. A. (2015). Three geographically separate domestications of Asian rice. *Nature plants.* **1**, 15164.

Cui, L., Wall, P.K., Leebens-Mack, J.H., Lindsay, B.G., Soltis, D.E., Doyle, J.J., Soltis, P.S., Carlson, J.E., Arumuganathan, K., Barakat, A., et al.. (2006). Widespread genome duplications throughout the history of flowering plants. *Genome Research.* **16**, 738-749.

Csűrös, M., Miklós, I. (2006). A Probabilistic Model for Gene Content Evolution with Duplication, Loss, and Horizontal Transfer. *Lecture Notes in Computer Science.*

Davies, R.B. (1987). Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*, **74**, 33-43.

De Bie, T., Cristianini, N., Demuth, J. P., Hahn, M. W. (2006). CAFE : a computational tool for the study of gene family evolution. *Bioinformatics*, **22**, (10), 1269-1271.

D'Hont, A. and Denoeud, F. and Aury, J.M. and Baurens, F.C. and Carreel, F. and Garsmeur, O. and Noel, B. and Bocs, S. and Droc, G. and Rouard, M. and others (2012). The banana (Musa acuminata) genome and the evolution of monocotyledonous plants. *Nature.*

Ferrao, L. F. V., Ferrao, R. G., Ferrao, M. A. G., Fonseca, A., Carbonetto, P., Stephens, M., Garcia, A. A. F. (2018). Accurate genomic prediction of Coffea canephora in multiple environments using whole-genome statistical models. *Heredity.*

Freeling, M. (2009). Bias in plant gene content following different sorts of duplication : tandem, whole-genome, segmental, or by transposition. *Annual review of plant biology*, **60**, 433-453.

Friedman, J., Hastie, T., Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of statistical software*, **33**, (1), 1.

Genz, A. (1992). Numerical computation of multivariate normal probabilities. *J. Comp. Graph. Stat.*, 141-149.

Hahn, M., De-Bie, T., Stajich, J., Nguyen, C., Cristianini N. (2005). Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Research.* **15**, 1153-1160.

Huang, X., Kurata, N., Wang, Z. X., Wang, A., Zhao, Q., Zhao, Y., ... Lu, Y. (2012). A map of rice genome variation reveals the origin of cultivated rice. *Nature.* **490**, (7421), 497.

Jannink, J.L., Lorenz, A.J. & Iwata, H. (2010). Genomic selection in plant breeding : from theory to practice. *Briefings in functional genomics.* **9**, (2), 166-177.

Jiao, J., ..., dePamPhilis, C.W (2011). Ancestral polyploidy in seed plants and angiosperms. *Nature*, **473**, 99-102.

Kellis, M., Birren, B., and Lander E. (2004). Proof and evolutionary analysis of ancient genome duplication in the yeast saccharomyces cerevisiae. *Nature*, **428**, 617-624.

Kubatko, L.S. (2009). Identifying hybridization events in the presence of coalescence via model selection. *Syst. Biol.*, **58**, 478488.

Kumar, S., Chagné, D., Bink, M.C., Volz, R.K., Whitworth, C. & Carlisle, C. (2012). Genomic selection for fruit quality traits in apple (Malus× domestica Borkh.). *PloS One.* **7**, (5), e36674.

Lander, E.S., Botstein, D. (1989). Mapping mendelian factors underlying quantitative traits using RFLP linkage maps. *Genetics*, **138**, 235-240.

Lebowitz, R.J., Soller, M., Beckmann, J.S. (1987). Trait-based analyses for the detection of linkage between marker loci and quantitative trait loci in crosses between inbred lines. *Theoretical and Applied Genetics*, **73**, 556-562.

Lee, S. H., Weerasinghe, W. S. P., Wray, N. R., Goddard, M. E., Van Der Werf, J. H. (2017). Using information of relatives in genomic prediction to apply effective stratified medicine. *Scientific reports.* **7**, 42091.

Mangin, B., Siberchicot, A., Nicolas, S., Doligez, A., This, P., Cierco-Ayrolles, C. (2012). Novel measures of linkage disequilibrium that correct the bias due to population structure and relatedness. *Heredity*, **108(3)**, 285.

Manichaikul, A., Palmer, A. A., Sen, ?., Broman, K. W. (2007). Significance thresholds for quantitative trait locus mapping under selective genotyping. *Genetics*, **177(3)**, 1963-1966.

Meuwissen, T.H., Hayes, B., Goddard, M.E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics.* **157**, (4), 1819-1829.

Rasmussen, M., Kellis, M. (2010). A bayesian approach for fast an accurate gene tree reconstruction. *Molecular Biology and Evolution*, **28**, 273-290.

Rebaï, A., Goffinet, B., Mangin, B. (1994). Approximate thresholds of interval mapping tests for QTL detection. *Genetics*, **138**, 235-240.

Rincent R, Laloe D, Nicolas S, Altmann T, Brunel D, Revilla P, et al (2012). Maximizing the Reliability of Genomic Selection by Optimizing the Calibration Set of Reference Individuals : Comparison of Methods in Two Diverse Groups of Maize Inbreds (Zea mays L.). *Genetics*, **192(2)**, 715-728.

Sato, S., Tabata, S., Hirakawa, H., Asamizu, E., Shirasawa, K., Isobe, S., Kaneko, T., Nakamura, Y., Shibata, D., Egholm, M. and others (2012). The tomato genome sequence provides insights into fleshy fruit evolution. *Nature*, **485**, 635-641.

Scannell, D.R., Byrne, K.P., Gordon, J.L., Wong, S., Wolfe, K.H. (2006). Multiple rounds of speciation associated with reciprocal gene loss in polyploid yeasts. *Nature*, **440**, 341-345.

Shao, J, Deng, X. (2012). Estimation in high-dimensional linear models with deterministic design matrices. *The Annals of Statistics*, **40(2)**, 812-831.

Siegmund, D., Yakir, B. (2007). *The statistics of gene mapping.* Springer.

Solis-Lemus, C., Ané, C. (20016). Inferring phylogenetic networks with maximum pseudolikelihood under incomplete lineage sorting. *Plos Genetics*, **12(3)**, e1005896.

Spindel, J., Begum, H., Akdemir, D., Virk, P., Collard, B., Redoña, E., et al (2015). Genomic Selection and Association Mapping in rice (Oryza sativa) : Effect of trait genetic architecture, training population composition, marker number and statistical model on accuracy of rice genomic selection in elite, tropical rice breeding lines. *PLoS Genetics.* **11**, (2), e1004982.

Wen, D., Yu, Y., Nakhleh, L. (2016). Bayesian inference of reticulate phylogenies under the multispecies network coalescent. *PLoS genetics*. **12**, (5), e1006006.

Wu, Y. (2012). Coalescent-based species tree inference from gene tree topologies under incomplete lineage sorting by maximum likelihood. *Evolution*, **66(3)**, 763-775.

Yu, Y., Degnan, J. H., Nakhleh, L. (2012). The probability of a gene tree topology within a phylogenetic network with applications to hybridization detection. *PLoS Genetics*, **8(4)**, e1002660.

Zhang, C., Ogilvie, H. A., Drummond, A. J., Stadler, T. (2017). Bayesian inference of species networks from multilocus sequence data. *Molecular biology and evolution*, **35(2)**, 504-517.

Zhu, J., Wen, D., Yu, Y., Meudt, H. M., Nakhleh, L. (2018). Bayesian inference of phylogenetic networks from bi-allelic genetic markers. *PLoS computational biology*, **14(1)**, e1005932.