FICHE TYPE POUR OFFRE DE THESE 2016
(les champs en rouge sont obligatoires)

Spécialité doctorale : Informatique

Inscription en thèse : Université de Montpellier

Date limite de validité de l'offre : fin 2016

ENCADREMENT DE LA THESE

Directeur de thèse (nécessairement HDR) : Berry Vincent

Co-encadrant éventuel : Celine Scornavacca

Correspondant/Contact :
Celine Scornavacca celine.scornavacca@umontpellier.fr 04 67 14 36 97

Titre en français : Nouvelles méthodes informatiques pour caractériser les génomes inter(sub)spécifiques de plantes cultivées

Titre en anglais : New computational methods to characterize inter(sub)specific crop genomes

Page web de l'offre (si elle existe) : http://www.lirmm.fr/~vberry/research.html

Financement prévu : sujet ouvert pour au concours pour attribution d'un contrat doctoral de l'ED

Profil(s) de candidats souhaité(s) : titulaire d'un diplôme de niveau Bac+5 en informatique (master, école d'ingénieur) obtenu en France ou à l'étranger. Le candidat aura une formation solide en informatique théorique (algorithmique, complexité) et/ou en mathématiques appliquées (maximum de vraisemblance, statistiques) et des compétences au moins minimales en programmation (C++ de préférence). Le candidat aura un goût pour la recherche pluridisciplinaire, même si aucune connaissance en biologie / agronomie n'est nécessaire pour candidater (ces connaissances seront apportées par les partenaires du travail de thèse).

Présentation détaillée en français : les génomes actuels de plantes cultivables (riz, tomate, citron, …) sont une mosaïque composée de plusieurs génomes ancestraux qui se sont hybridés plusieurs fois au cours de l'histoire de ces plantes. L'objectif du projet est de retracer ces événements d'hybridation ainsi que de façon générale l'histoire des plantes en question. Ceci nécessite la conception de nouveaux algorithmes pour la construction de réseaux phylogénétiques (graphes orientés acycliques ayant plusieurs sources).

Les données massives dont nous disposons maintenant permettent d'effectuer cette reconstruction, seules manquent les méthodes assez rapides (faible complexité algorithmique) pour analyser ces données et proposer des estimations de l'histoire de ces différentes plantes. La particularité des plantes étudiées ici réside dans le fait que leur génome est composé de restes du génome de plusieurs fondateurs, une spécificité qu'il faut prendre en compte dans la modélisation qui sera proposée. Cette modélisation rendra aussi explicite l'évolution des génomes par des événements tels que les duplications, pertes et transferts de gènes. Ces

phénomènes se modélisent traditionnellement pas comparaison d'un arbre de gène à un arbre d'espèces (ces arbres modélisant l'histoire évolutive, un peu à la manière d'un arbre généalogique). Il s'agit ici d'étendre ces méthodes comparatives aux cas où une des deux histoires est un graphe orienté et non plus simplement un arbre.

Pour plus de détails, voir la description en anglais du sujet de thèse

Présentation détaillée en anglais (non obligatoire mais recommandé) :
This PhD project issues from recent success stories of the Agropolis community in the production of crop reference genome sequences. These landmark achievements together with the current possibility for massive resequencing data production have opened new opportunities to understand the organization and dynamics of these genomes and the related keys to a more efficient exploitation of their diversity in breeding programs. However, this also brought up new challenges since the full exploitation of these data requires development of new biomathematic / bioinformatic concepts, methods and tools. We want to tackle these challenges thanks to the strength of Montpellier in computer science, mathematics and agricultural sciences.

The PhD will focus on aspects related to the frequent inter(sub)specific events involved in the history of crops and develop methodologies to decipher the inter(sub)specific structure of crop genomes.

The methods that have been used so far to characterize the mosaic structure in plant genomes consisted in analyzing the distribution of simple summary statistics along the genome (Wu et al. 2014, Curk et al. 2014). Although they have proved useful in some biological models such approaches do not make optimal use of the information contained in the data, and therefore lack statistical power in species where information from putative parental genomes is scarce or even missing. Moreover, these approaches will undoubtedly be affected by sequencing errors and/or uncertainties in variant calling.

The mosaic structure in plant genomes can be characterized by a phylogenomic approach. The PhD subject aims at extending phylogenetic methods and at applying them on genome-size data. More precisely, the student will couple the inference of hybridization networks, reconciliations and ancestral gene adjacencies.

Hybridization networks are often used to describe and explain how a few founders shaped current genomes through reticulate events such as rounds of hybridization or recombination. Such networks describe precisely the chain of major events that lead to current genomes. Their recovery is an important step to understand the broad evolutionary patterns that lead to the composition of current and ancestral genomes. The team at work shares some expertise in the field (Huson et al. 2009, Huson et al. 2010, Gambette et al. 2012, Huson et al. 2012, Kelk et al. 2012, El Baidouri et al. 2013, van Iersel et al. 2014). Such networks are currently inferred either from sequence data (that is, ordered character data, in which case the network is called an Ancestral Recombination Graph — ARG) (Gusfield 2014), or from a set of unordered gene trees (Huson et al. 2011).

Species tree / gene tree reconciliations is a technique often used to track the evolutionary history of genome fragments or gene sequences. Such methods allow to tag nodes and branches of gene phylogenies so as to infer gene duplications, losses and transfersthat can be used to model recombinations due to hybridization (Doyon et al. 2010, Doyon et al. 2011, El Baidouri et al. 2013, Nguyen et al. 2013a, Nguyen et al. 2013b, Scornavacca el al. 2013,

Scornavacca el al. 2015). Such events allow to explain the origin of current genome parts having identified homologues in other species or subspecies.

Last, the inference of ancestral gene adjacencies is a common technique to obtain a realistic picture of the mosaic of ancestral genomes from which current genomes evolved. Many methods exist that propose estimates of the gene order of ancient genomes (Braga et al. 2008, Chauve and Tannier 2008), yet most of them operate by translation and permutation operators on genome fragments, without accounting for the specific history of genome fragments or gene families. With several founders involved, accounting for such histories is a prerequisite to obtain the mosaic of ancestral genomes. Methods along this track just start to appear, being elaborated in part in Montpellier (Bérard et al. 2012) and Lyon (Patterson et al. 2013). These methods track current gene adjacencies backward in time being guided by species / gene tree reconciliations. It's also plausible that there will be a connection with works inferring synteny blocks in ancient genomes (see Lucas et al. 2014) for a preliminary work resorting on gene phlyogenies for two species only) or those relying on phylogenies to assemble contigs of fragments ancient DNA (Luhmann et al. 2014). Though the phylogenomic methods mentioned above pave the way to explain the mosaic of current and ancient genomes in general, they need further development in order to be applied to plants studied in this project. Methods for hybridization networks inference are still too slow to scale up to handle genome-size data as that considered here (e.g., the 3000 rice genome project). Moreover, reconciliation methods and gene adjacencies inference methods are only for species phylogenies that can be modeled as trees. They need to be extended to handle the case of species with multiple founders, such as polyploid plants considered in this project. Adapting these three approaches to our data and using them together will permit us to conceive a powerful framework to unravel mosaic structures of plant genomes in general.

INFORMATIONS SUPPLEMENTAIRES UTILES

Particularités de l'encadrement (par exemple : collaboration internationale, etc.) : ce sujet de thèse se situe dans le projet *GenomeHarvest*, financé par la fondation Agropolis. L'étudiant sera encadré par des informaticiens et interagira avec des chercheurs en agronomie dans le cadre de ce projet.

Partenariat industriel éventuel : la société de services *Bioversity International* est impliquée dans le projet. Des liens pourront être développés avec elle pour l'exploitation des résultats de ce travail.