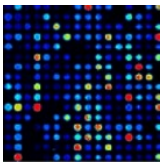
Séminaire – Math For Genomics

Séance du mercredi 10 janvier 2018. 10h30.
Evry. IBGBI. LaMME.

Problème d'assemblage de génomes



Véronique BRUNAUD (IPS2)

Title : Use case of an assembly with arabiodpsis transcriptome, limits and views.

I will show an application of the most popular algorithm (based on de Bruijn graphs) to evaluate the assembly of RNA-Seq samples. I will present the biological context, problems and why biologist need these tools.

Briefly, one has to cope with
- the fact that the genome or transcriptome is unknown
- the millions of short sequences (or reads) to merge in one transcript
- the fact that a locus does not correspond to a unique transcript.

The main objective is to detect new objects :
- a reference
- new gene annotations
- new transcipts...

but there are many limits with complex genomes.

Michel KOSKAS (AgroParisTech)

Titre : Exemples d'algorithmes utilisés pour le génome.

Cet exposé se compose de deux parties.

Les machines NGS (Next Generation Sequencing) ont révolutionné le séquençage en produisant de grands volumes de données pour un prix très modique. Ces données se présentent sous forme de *reads*, c'est-à-dire sous forme de petites séquences de paires de bases. Elles doivent être traitées entre autres pour deux grands problèmes : le mapping et le séquençage de novo. Le mapping consiste à tenter de placer les reads dans un génome de référence, en admettant des altérations des reads par rapport au génome. Le séquençage de novo consiste à tenter de reconstituer un génome entier à partir des seuls reads (en admettant là encore des altérations). La taille de ces données fait qu'il a fallu dééveloppper des algorithmes très sophistiqués pour tenter de venir à bout de ces problèmes. La première partie de cet exposé présentera des algorithmes et structures de données utilisés dans ce contexte.

Chez les organismes diploïdes, chaque individu de la population possède 2 copies de chaque gène. Si ce gène possède deux formes allèliques notées 0 et 1, le génotypage d'un individu par des techniques de séquençage permet de savoir si cet individu a 0, 1 ou 2 copies de l'allèle 1, et ce pour de multiples gènes (appelés marqueurs dans la suite) localisés le long de la séquence d'ADN. On dispose actuellement pour un grand nombre d'organismes de bases de données contenant des centaines d'individus ainsi séquencés. Cette connaissance est en réalité souvent très partielle : sur les millions de marqueurs disponibles dans les bases de données, seuls 30 ou 40% sont effectivement renseignés. L'objectif est alors d'inférer pour chaque individu l'information aux marqueurs non renseignés.

Les stratégies d'inférence reposent sur le fait que des marqueurs proches sur la séquence sont généralement dépendants (au sens statistique du mot). Ainsi chaque marqueur apporte un information partielle sur ses voisins. La dépendance entre marqueurs peut être modélisée par une chaîne de Markov, qui peut être estimée, et ensuite utilisée pour inférer - ou "imputer" - les marqueurs non renseignés.

Les méthodes d'imputation basées sur des chaînes de Markov ont déjà été appliquées avec succès chez l'humain et chez plusieurs espèces animales (Browning and Browning, 2013). Toutefois ces méthodes ont toutes été développées pour réaliser l'imputation chez les mammifères, et ne sont pas adaptées lorsque les individus considérés sont des lignées (de plantes ou de bactéries) pour lesquels on sait à l'avance que le nombre de copies de l'allèle 1 ne peut être que 0 ou 2 (les individus sont purement homozygotes). On a donc développé un logiciel prenant un compte le caractère homozygote des lignées considérées, obtenant ainsi un gain de performances substantiel en temps de calcul. C'est ce travail qui fera l'objet de la deuxième partie de cet exposé.