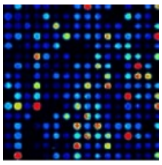
Séminaire – Math For Genomics

Séance du mercredi 1er avril 2020. 10h30.
Evry. IBGBI. LaMME.

Metagenomics: from biological data to statistical associations



Coline Billerey (Bioinformaticienne - Enterome)

Analysis of new methods of sequencing applied to metagenomics

The gut microbiota is made up of more than 500 different bacteria species. To identify them, we perform short-reads sequencing analysis on the DNA extracted from feces samples, the reads are then assembled in contigs from which genes are predicted. By homology with known sequences present in the public databases, we can assign the reconstructed genes to a taxonomy. But due to the lack of completeness in these databases, only half of the genes obtained can be assigned. In addition, the short-reads sequencing doesn't make it possible to produce complete genomes and limits the possibility of analysing structural genomic variation. The new methods of sequencing based on long-reads like Nanopore technology or cloud-reads like 10X may be a way to improve metagenomic assembly. The presentation will be focused on the analysis of data obtained with these two sequencing methods and the comparison of the resulting assemblies, as well as the limitations of the methods and improvements to be made.

Antoine Bichat (LaMME/Enterome)

Ornstein-Uhlenbeck process on a tree to detect differentially abundant species

In metagenomics, differential abundances studies are commonly used to identify species whose abundances are associated to a phenotype of interest, e.g. healthy versus diseased.

Most state of the art methods carry out independent tests to know whether each species is differentially abundant or not. However, such procedures do not take into account evolutionary relationships between species which may result in correlated abundances and then correlated p-values. These relationships are captured by the species phylogenetic tree.

This work proposes a method to detect differentially abundant species based on the phylogeny. p-values are seen as a function of an Ornstein-Uhlenbeck process with shifts on the phylogenetic tree. Our method estimates the positions and intensities of those shifts. Shifts are then propagated back to the leaves -i.e. species- in order to identify the differentially abundant ones.