

PhD proposal

New methods for mining biological information in Next Generation Sequencing data

PhD advisors : Dominique Lavenier (DR CNRS), Pierre Peterlongo (CR INRIA), Claire Lemaitre (CR INRIA)

Location : Symbiose Group – Irisa/Inria Rennes

Contact: pierre.peterlongo@inria.fr , claire.lemaitre@inria.fr

web : http://www.irisa.fr/symbiose/pierre_peterlongo

Context

One fundamental question in biology is to understand sequence variations among individuals of a same species, that is polymorphism. These variations can involve single nucleotides differences (SNP) or longer stretches of DNA that can be duplicated, deleted, reversed or relocated in the genome (structural variants). They are often associated to specific traits of the individuals and can lead to adaptation and evolution of the species or conversely they can also be involved in diseases and cancers.

Polymorphism studies entered an unprecedented deep change a few years ago with the arrival of Next Generation Sequencers (NGS). This new technology enables to sequence biological material with a flow much higher than before, for a price now accessible to most biological lab. However this generates huge amounts of data of a new type that can only be treated by sophisticated computational methods. Indeed, the main drawback of these data is the small size of the generated sequences (called reads) which makes it difficult to reconstruct the full DNA molecule they come from. Therefore most of the current pipelines to detect polymorphism in these data go through a pre-processing step of read mapping on a reference genome (that is a fully sequenced and assembled genome for a closely related organism). However this step can lead to the loss of biologically relevant information and the requirement of a reference genome limits its application to a few model organisms. As more and more sequencing projects do not benefit from a "good" reference genome, new methods enabling to extract biological information solely from the raw data are urgently needed.

The Symbiose group is leader in this field and developed such a new approach. KisSNP is a new method that can find single nucleotide polymorphism (SNP) between two NGS raw data samples. It is based on the exploration of the commonly used data structure, the De Bruijn Graph, which can efficiently store large sets of short reads. The idea is that biological variants, such as SNP, generate specific motifs in this graph that can be detected without any pre-processing step.

PhD proposal

The aim of this proposal is to extend the existing method KisSNP to the detection of all types of polymorphisms in one or between several samples. In particular, structural variants are more complex variants which have been rarely studied in non model organisms and for which original methods free of reference genome are critically missing. The work will consist first in formalizing models generated in the De Bruijn Graph by each type of biological variants. Topological and combinatorial properties will be investigated to characterize the variants and other data structures could also emerge to better fit the biological data (dealing with pairs of reads instead of single reads). Then, new algorithms will be proposed to detect efficiently the variants. Finally, the methods will be validated on real NGS samples in collaboration with biologists closely linked to the Symbiose team and Genouest Bioinformatics platform.

References

- Next-generation gap. J D McPherson. Nat Methods, 2009, 6:S2-S5
- Assembly algorithms for next-generation sequencing data. J R Miller, S Koren & G Sutton. Genomics, 2010, 95:315-327
- Computational methods for discovering structural variation with next-generation sequencing. P Medvedev, M Stanciu & M Brudno. Nat Methods, 2009, 6:S13-S20
- Identifying SNPs without a reference genome by comparing raw reads. P Peterlongo, N Schnel, N Pisanti, MF Sagot & V Lacroix. In proceedings of String Processing and Information Retrieval, 2010