Incorporating genomic variation information into DNA sequencing data analysis

Norbert Dojer
Institute of Informatics, University of Warsaw

dojer@mimuw.edu.pl

In the majority of DNA sequencing experiments the first step of analysis consists of mapping sequencing reads onto a so-called *reference* genome, which represents the consensus of genomic sequence of the species of interest. Currently reference genomes are available for thousands of species and much effort is devoted to the analysis of genomic diversity among them. This is especially visible in the case of human genomics, where the development is driven by the perspective of application to personalized medicine. However, current pipelines of sequencing data analysis are unable to utilize this knowledge to reduce the bias and the noise caused by differences between reference and actual genomes [1].

I am currently starting the project that addresses this problem. The proposed approach is based on the concept of *reference multi-genome* (RMG), i.e. the reference modeling multiple variants of particular genomic loci. The project aims at: (1) developing software mapping reads onto RMG, (2) incorporating RMG-based mapping in DNA sequencing analysis pipelines, (3) performing a case study of data analysis using RMG-based mapping approach.

Possible internship projects include:

1. RMG construction for small genomes.
2. RMG-based read mapping.
3. Annotation of RMG-extracted genome.

References

[1] The Computational Pan-Genomics Consortium. Computational pan-genomics: status, promises and challenges. *Briefings in Bioinformatics*, 2016.