PhD Proposal

High Throughput Genomic Data Compression

Research team: IRISA/INRIA GenScale, Rennes
Supervisor: D. Lavenier, DR CNRS

Next generation sequencers (NGS) provide now in a single run very high volume of genomic data, which are stored into text files representing Tera bytes of information. Data represent the text of billion of short fragments, called reads, together with quality measure. The standard format is the FASTQ format derived from the famous FASTA format.

The size of these files is becoming a real problem. First, it saturates the space storage of bioinformatics centers. Second, transmitting Tera bytes of data through Internet takes a lot of time and is sometime impossible because of restricted bandwidth. In many cases, it is even faster to send physical hard drive. Standard compression algorithms allow the size to be reduced by a factor of 4 to 6.

However, NGS data are highly redundant. They result from shotgun sequencing techniques with significant coverage. In other words, there exist strong overlaps between a high majority of reads which can efficiently be exploited to reduce the overall description. In the ideal case of a single genome without sequencing errors and perfect coverage, the optimal reordering of the reads is the genome itself. Compression can be reduced to the text of the genome enhanced with read mapping information. The reality must however deal with sequencing error and imperfect coverage, leading to a much more complex compressing scheme.

Challenges are manifold: (1) find data structures allowing HTS genomic data to be efficiently compressed; (2) provide associated (parallel?) algorithms able to generate fast compression; (3) provide optimized API (Application Programming Interface) to extract data without decompressing files.

The PhD candidate will join the internal GenScale group working on this topic.

Contact: D. Lavenier. Please send a CV to lavenier@irisa.fr