The mapping task and its various applications in next-generation sequencing

Christian Otto

Abstract

The aim of this thesis is the development and benchmarking of computational methods for the analysis of high-throughput data from tiling arrays and next-generation sequencing. Tiling arrays have been a mainstay of genome-wide transcriptomics, e.g., in the identification of functional elements in the human genome. Due to limitations of existing methods for the data analysis of this data, a novel statistical approach is presented that identifies expressed segments as significant differences from the background distribution and thus avoids dataset-specific parameters. This method detects differentially expressed segments in biological data with significantly lower false discovery rates and equivalent sensitivities compared to commonly used methods. In addition, it is also clearly superior in the recovery of exon-intron structures. Moreover, the search for local accumulations of expressed segments in tiling array data has led to the identification of very large expressed regions that may constitute a new class of macroRNAs.

This thesis proceeds with next-generation sequencing for which various protocols have been devised to study genomic, transcriptomic, and epigenomic features. One of the first crucial steps in most NGS data analyses is the mapping of sequencing reads to a reference genome. This work introduces algorithmic methods to solve the mapping tasks for three major NGS protocols: DNA-seq, RNA-seq, and MethylC-seq. All methods have been thoroughly benchmarked and integrated into the `segemehl` mapping suite.

First, mapping of DNA-seq data is facilitated by the core mapping algorithm of `segemehl`. Since the initial publication, it has been continuously updated and expanded. Here, extensive and reproducible benchmarks are presented that compare `segemehl` to state-of-the-art read aligners on various data sets. The results indicate that it is not only more sensitive in finding the optimal alignment with respect to the unit edit distance but

also very specific compared to most commonly used alternative read mappers. These advantages are observable for both real and simulated reads, are largely independent of the read length and sequencing technology, but come at the cost of higher running time and memory consumption.

Second, the split-read extension of `segemehl`, presented by Hoffmann, enables the mapping of RNA-seq data, a computationally more difficult form of the mapping task due to the occurrence of splicing. Here, the novel tool `lack` is presented, which aims to recover missed RNA-seq read alignments using *de novo* splice junction information. It performs very well in benchmarks and may thus be a beneficial extension to RNA-seq analysis pipelines.

Third, a novel method is introduced that facilitates the mapping of bisulfite-treated sequencing data. This protocol is considered the gold standard in genome-wide studies of DNA methylation, one of the major epigenetic modifications in animals and plants. The treatment of DNA with sodium bisulfite selectively converts unmethylated cytosines to uracils, while methylated ones remain unchanged. The bisulfite extension developed here performs seed searches on a collapsed alphabet followed by bisulfite-sensitive dynamic programming alignments. Thus, it is insensitive to bisulfite-related mismatches and does not rely on post-processing, in contrast to other methods. In comparison to state-of-the-art tools, this method achieves significantly higher sensitivities and performs time-competitive in mapping millions of sequencing reads to vertebrate genomes. Remarkably, the increase in sensitivity does not come at the cost of decreased specificity and thus may finally result in a better performance in calling the methylation rate.

Lastly, the potential of mapping strategies for *de novo* genome assemblies is demonstrated with the introduction of a new guided assembly procedure. It incorporates mapping as major component and uses the additional information (e.g., annotation) as guide. With this method, the complete mitochondrial genome of *Eulimnogammarus verrucosus* has been successfully assembled even though the sequencing library has been heavily dominated by nuclear DNA.

In summary, this thesis introduces algorithmic methods that significantly improve the analysis of tiling array, DNA-seq, RNA-seq, and MethylC-seq data, and proposes standards for benchmarking NGS read aligners. Moreover, it presents a new guided assembly procedure that has been successfully applied in the *de novo* assembly of a crustacean mitogenome.