Universität Bielefeld

Technische Fakultät
Computational Metagenomics

Universität Bielefeld | Postfach 10 01 31 | 33501 Bielefeld

Dr. Alexander Sczyrba

Raum: M3-111
Tel.: 0521.106-2910
Fax: 0521.106-152910
asczyrba@techfak.uni-bielefeld.de
www.cebitec.uni-bielefeld.de/cmg

Bielefeld, 26. April 2019

Review of the Doctoral Dissertation of Nurk Sergey Yurievich "Assembling genomes of non-cultivable microorganisms from high-throughput sequencing data" submitted for defense of the degree of candidate of physico-mathematical sciences, speciality 03.01.09 "mathematical biology, bioinformatics"

Scope of the thesis

High throughput sequencing has become a standard technology in life sciences today. Although tens of thousands of microbial genomes are sequenced every year now, the majority of the microbial world remains unexplored due to the inability to grow these organisms in pure culture: about 99% of microbial organisms cannot be cultured under standard laboratory conditions. However, two complementary approaches try to shed light on the "microbial dark matter", metagenomics and single cell genomics. Sergey Nurk addresses these highly relevant approaches in his dissertation thesis and presents novel assembly approaches to obtain the genomic sequences of prokaryotic genomes from single cell and metagenomic sequencing. Both approaches combined with high-throughput sequencing have revolutionized microbiology, providing access to genomic information formerly inaccessible.

While assembly algorithms have been available for a while, only very few have addressed the specific problems arising from the data obtained by metagenomics and single cell genomics approaches. While sequencing isolate genomes generates relatively even coverage of the genome, the multiple displacement amplification (MDA) used for amplification of the single cell DNA introduces a highly non-uniform read coverage of the genome and also formation of chimeric reads. In metagenomics, the different abundances of the genomes in the sample lead to very different read coverage of the corresponding genomes. Sergey Nurk approaches these problems by implementing novel approaches for processing sequencing data from metagenomics and single cell sequencing.

09/2 - 128 am 26.04.2019

Background work

The thesis gives a short introduction (chapter 1) into the topic of high-throughput whole genome sequencing and assembly of prokaryotic genomes. Here, Mr. Nurk introduces the main concepts of genome sequencing, assembly, assessment of assembly quality, as well as using de Bruijn graphs for genome assembly. Also, the concepts of single cell sequencing and metagenomics sequencing are introduced.

SPAdes – a genomic assembler and new approaches to processing de Bruijn graphs

In chapter 2 Mr. Nurk presents a novel method for compact representation of de Bruijn graphs in memory. A special focus of this chapter are methods for the construction and simplification of assembly graphs for single cell sequencing. This includes methods for removing tips and bulges, as well as processing complex bulges in the assembly graph. Also, methods for aligning sequences to the assembly graph and incorporating information about paired read linkage are presented. These methods are the underlying solutions for the SPAdes assembler.

metaSPAdes – a metagenomic assembler

In chapter 3 the metagenomic assembler metaSPAdes is presented. Proven solutions from single cell sequencing assemblies are used to address challenges of metagenomic assembly, including assembly of different abundant genomes and closely related microbial strains. metaSPAdes incorporates SPAdes functionality for the iterative construction of de Bruijn graphs. Using heuristic modification procedures consensus assemblies for closely related strains are produced.

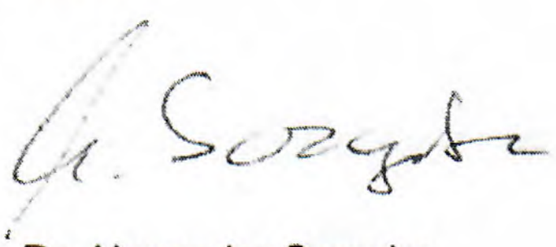
Assembler Benchmarks

For both SPAdes and metaSPAdes results of computational experiments demonstrating advantages over competing solutions are presented at the end of chapter 2 and 3, respectively. These benchmarking results against other popular assemblers demonstrate the advancements for single cell and metagenomic assembly presented in the thesis by Sergey Nurk.

Summary

In his dissertation project, Sergey Nurk has demonstrated proficiency in determining worthwhile challenges in sequence analysis and proved his ability to implement algorithms to tackle challenges in assembling single cell genomes and metagenomes. The presented work is highly relevant, novel and accurately presented. The thesis itself is well structured and written. The main results of the work has been published in three papers, cited as [7,8,9] in the thesis. Mr. Nurk proved that the novel approaches he implemented in the two assemblers SPAdes and metaSPAdes have been significant advancements in the field of metagenomics and single cell genomics. The assemblers are highly accepted and widely used in the bioinformatics community.

Thus, the dissertation of Sergey Nurk meets the requirements necessary for granting of the degree of candidate of physico-mathematical sciences, speciality 03.01.09. I support this action without hesitation.



Dr. Alexander Sczyrba