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, specialization 03.01.09 – Mathematical biology, bioinformatics.

Relevance

The dissertation is devoted to the topic of microbial genome sequencing and assembly from single cells and metagenomes. Because of the difficultly of these problems, the proposed solutions also apply to the comparatively easier problem of assembling cultured microbial isolates. The relevance of these topics is extremely high.

Genome assemblies are the foundation of modern genomics. Because genomes cannot yet be read in a single, continuous pass, modern sequencing technologies must read many shorter fragments that are then computationally assembled into the original genome. Further complicating this problem, many microorganisms cannot be easily cultured in the laboratory, which makes isolating and extracting enough DNA for sequencing exceedingly difficult. As a solution, laboratory techniques have been developed that can amplify the genome of a single cell to many copies, making whole-genome sequencing feasible. However, this amplification process introduces its own artifacts, including chimeric sequencing reads and wildly uneven sequencing coverage, that must be dealt with during computational assembly. An alternative approach to single-cell amplification is to directly sequence the combined DNA of an environmental sample, known as a metagenome. In this context, the assembly problem is to reconstruct the genomes of an unknown number of microorganisms that constituted the original sample. Both the single-cell and metagenomic assembly problems are renowned for their difficulty. This dissertation addresses both problems with computationally sound and elegant solutions implemented in the well-engineered software SPAdes and metaSPAdes.

The relevance of this work is immediately evident from the large community impact and high number of citations achieved by the associated publications. For example, Google Scholar reports >5,000 citations for the 2012 SPAdes paper, on which the candidate was a co-first author; >300 citations for the 2013 on single-cell genome assembly, on which the candidate was first author; and >200 citations for the 2017 metaSPAdes paper, on which the candidate was the first and corresponding author. These citation counts are remarkable for the field of bioinformatics and are all well within the top 1% of all biomedical research, as reported by the US National Institutes of Health's citation statistics database. Furthermore, within the microbial genomics and bioinformatics communities, SPAdes is widely used, highly regarded for its expert engineering, and commonly viewed as a pinnacle of genome assembly software.

Validity

The validity of the work presented in the dissertation is unquestionable. Both SPAdes and metaSPAdes were thoroughly benchmarked against other state-of-the-art assembly programs, and the presented results demonstrate that the SPAdes tools are computationally efficient and reliably generate high-quality assemblies. Furthermore, there have been multiple, independently published evaluations of assembly software, and SPAdes is routinely a top performer in such studies. I have personally used both tools presented in the dissertation and found them to be of the highest quality.

09/2-127 om 26.04.2019

In the dissertation, both methods and results are presented in a clear and understandable manner, and appropriate validation has been performed to demonstrate the accuracy of the methods. The English portion of the defense is very well written and contains only a small number of minor typos that can be resolved by a thorough proofing. The introduction to the sequencing and assembly problems in chapter 1 is particularly well presented and shows a broad understand of the field. I am unable to assess the Russian portion of the dissertation, but based on the careful presentation of the English sections, I presume it is of equal quality.

Novelty

The primary contribution of the presented work is a comprehensive set of algorithms for building, simplifying, and traversing sequence assembly graphs for the purposes of single-cell, isolate, and metagenomic sequence assembly. In some cases, the chosen algorithms were adapted from prior works (e.g. minimal perfect hashing), and in other cases the algorithms were invented by the candidate for the problem at hand (e.g. simplifying complex assembly graph structures). This demonstrates a practical approach to problem solving that combines deep knowledge of the literature with a willingness to develop new solutions when no prior work exists. Such well-designed and practical tools are incredibly powerful for advancing research. They enable discovery across the field of genomics and amplify the impact of the dissertation. The contributions of this dissertation do not simply answer a single hypothesis; they enable others to answer countless hypotheses.

Conclusion

The dissertation of Sergey Nurk includes work from three papers published in highly regarded and indexed journals. Both the published papers and the dissertation itself demonstrate excellent scholarship and cite the relevant prior works. As detailed above, the presented work is highly relevant, valid, novel, and accurately presented. Thus, the dissertation of Sergey Nurk meets the requirements necessary for the granting of the degree of candidate of physico-mathematical sciences, specialty 03.01.09, and I support this action without hesitation.

The quality of this dissertation cannot be overstated. The presentation is thorough, precise, and demonstrates a profound grasp of the topic, and the developed methods have had a tremendous impact on the field. I commend the candidate on his achievements, and I thank him for the contributions he has made to the field.

26 April 2019, serving in my personal capacity,

Adam M. Phillippy, Ph.D. (Computer Science)
Investigator and Head, Genome Informatics Section
Computational and Statistical Genomics Branch
National Human Genome Research Institute
National Institutes of Health
Bethesda, MD USA