

REVIEW

Of the Doctoral Dissertation of Andrey D. Przhibelskiy "Development of algorithms for de novo genome and transcriptome assembly" submitted for defense of the degree of candidate of physico-mathematical sciences, specialization 03.01.09 – Mathematical biology, bioinformatics.

Relevance

The dissertation describes the development of new algorithms for biological sequence assembly from genomes and transcriptomes. Recent advances in DNA sequencing technology have dramatically lowered the cost of whole-genome and transcriptome sequencing, making these technologies indispensable tools for research. Thus, the relevance of the dissertation topic is very high.

Sequence assembly is a foundational problem in genomics. Most DNA sequencing technologies can read only a short sequence of DNA at a time, and so computational methods are needed to reconstruct the original, longer fragments. The primary challenge of the sequence assembly problem is repeats, strings of DNA that occur in multiple regions of the genome. When attempting to reconstruct the original sequence from much shorter sequencing reads, these repeats confuse the process and can prohibit complete assembly. This ambiguity is typically represented as a graph of possible sequence reconstructions, and sequence assembly algorithms explore this graph to identify a correct reconstruction. Resolution of sequence assembly graphs is typically guided by assumptions on the data such as expected sequencing error, depth of sequence coverage, DNA fragment sizes, etc. However, some of these common assumptions are violated in the use cases explored by this dissertation. For example, in the context of single-cell sequencing, the genome must be heavily amplified before sequencing, resulting in a highly uneven depth of sequencing coverage. Similarly, in RNA sequencing experiments, differing transcription levels and alternative splicing result in uneven coverage depth across the multiple exons of a gene. Accurately assembling DNA and RNA sequences in the presence of such data characteristics is the primary contribution of the thesis.

The relevance of this work is immediately evident from the large community impact and high number of citations achieved by the associated publications. The candidate's contributions are organized into three main areas: (1) contributions to the widely used SPAdes assembly software, (2) development of the exSPAnDer algorithm within SPAdes, and (3) modifications of SPAdes for RNA transcript assembly, called rnaSPAdes. Google Scholar reports >8,000 citations for the 2012 SPAdes paper, on which the candidate was a co-author. The exSPAnDer paper, on which the candidate was first author, was awarded an Outstanding Student Paper award at bioinformatics' premier ISMB meeting in 2014, and forms a critical component of the SPAdes assembler. Lastly, rnaSPAdes was only recently published, but has already received multiple citations in the few months it has been available. Within the genomics community, 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. SPAdes has been 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. The

genomics community has clearly given this software its resounding approval. rnaSPAdes was more recently published, and so has not yet had the opportunity to gain such wide acceptance, but it has been peer-reviewed and the benchmarking results presented here and in the publication are reasonable and thorough.

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 contains minor grammatical errors, but these are not substantial enough to distract from understanding. 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

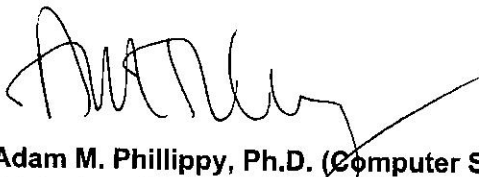
The primary contribution of the presented work is a set of algorithms for resolving repeats and artifacts within the context of sequence assembly graphs for both DNA and RNA sequencing. Although the candidate is not a first author on the SPAdes manuscript, resolving sequence repeats (the function of exSPAnder) is one of the most complex and important aspects of sequence assembly. Thus, his development of the algorithms behind exSPAnder were important to the success of SPAdes as a whole. In addition, the candidate developed novel adaptations of the software for the highly relevant topic of RNA transcript assembly.

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 Andrey Przhibelskiy includes work from five papers published in highly regarded and indexed journals, including one as first author and one as last author. Both the published papers and the dissertation itself demonstrate 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 Andrey Przhibelskiy 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.

21 February 2020, serving in my personal capacity,



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