REVIEW
of the Doctoral Dissertation of Andrey Przhibelskiy
"Development of algorithms for de novo genome and transcriptome assembly"
submitted for defense of the degree of candidate of physico-mathematical sciences, specialty 03.01.09 "mathematical biology, bioinformatics"

Dissertation of Andrey Przhibelskiy consists of two main chapters describing solutions for a variety of genome and transcriptome assembly problems (Ch. 2 and Ch. 3, respectively). Needless to say, both problems are and in foreseeable future will be actual and relevant, as the cost and speed of sequencing constantly decrease, and larger and more complex dataset are generated by the researchers community.

The contribution of the author to these problems is of different kinds. Ch. 2 of the manuscript describes a variety of useful heuristics developed and implemented by the author to improve the performance of a popular and widely used assembler, SPAdes.; the most important of these is the exSPAnder module for the repeat resolution. At that, the exposition of this chapter is necessarily patchy, with few links between paragraphs.

Ch. 3, on the other hand, is dedicated to a single problem, that of transcriptome assembly, where the author has been the main contributor. Correspondingly, the description of the rnaSPAdes assembler is better integrated and more complete.

In any case, both projects are important and have led to new or radically improved tools. The author demonstrates good knowledge of the field, ability to identify problems and solve them, sharp mind and good algorithmic taste. He is the main author of two papers (the first author in Bioinformatics 2014, the last author in GigaScience 2019), and he has contributed to three more papers (J. Comp. Biol. 2012; J. Comp. Biol. 2013; Bioinformatics 2015); all are published in strong journals. By all standards, he deserves a PhD degree.

That said, turn to problems. The primary one is that the thesis writing seems to have followed the path of the least effort. The review sections of the dissertation are very short and largely superficial. This relates both to the general review (Ch. 1) and to sections describing the existing approaches to specific problems considered by the author (Sections 2.1 and 3.1). Further, it is not clear, where the description of the SPAdes general features end and the author's improvements start — what paragraphs in Section 2.3 describe the state of the package prior to the start of the project and what paragraphs describe the author's contribution? Same in Section 3.2). Both these aspects are unfortunate, as such layout provides almost no context and does not allow the reader to fully appreciate the novelty and originality of the author's solutions. I understand that a logical, complete description of a large project may yield a patchwork of individual contributions, but still there are ways to distinguish between the author's work and that of others, mentioned for completeness of the exposition (e.g. using different formatting of the text, fonts, margin sizes, etc.).

Secondly, the overall logic of the text is sometimes broken, notions and terms are used before they are defined etc. One such example may be found in Table 11 on page 44 that lists benchmark parameters defined only on page 55.

Sometimes the reader is referred for details to the author's publications (e.g. to [17] on page 48) — that defies the purpose of writing a thesis that by definition should as be comprehensive and self-contained as possible. At that, the entire section 3.3 reads like an excerpt from an autobiography rather than a self-contained scientific exposition.

Thirdly, the choice of particular parameters and procedures is not always well explained, for instance:

Page 19: *For read length of 100 bp SPAdes has three iterations with* k *equal to 21, 33, 55 <…> For single-cell datasets, however, the nearly optimal values remain 21, 33, 55.* — In what sense “optimal”? How do we know that?

Page 43: *One may notice that uneven coverage depth is the exactly the same challenge that was addressed in SPAdes genome assembler, since its primary purpose was assembly of single-cell data. In the view of such parallels between two unrelated sequencing data types, SPAdes was tested on the RNA-Seq data and compared with existing transcriptome assemblers* — the logic here is not transparent: the type of coverage non-uniformity is different, as coverage is more or less uniform within transcripts (notwithstanding alternative splicing isoforms)

Page 47: *rnaSPAdes collapses only simple bulges that consist of edges with less than 10% in length difference* — does that mean that many mutually exclusive exons (that often are of similar, if not identical length) will be collapsed?

Page 47: *Types of chimeric junctions detected by SPAdes are not typical for RNA-Seq data, and thus the majority of the algorithms for removing erroneous connections were excluded from rnaSPAdes pipeline* — what procedures were excluded?

Page 50: *If two isoforms of the same gene share exons longer than insert size of the sequenced library, it is not possible to reconstruct such isoform***<s>** *entirely. However, if these isoforms have significantly different expression levels, they can be resolved by their coverage depth* — this is based on an implicit assumption that selection of alternative variants should be correlated even for alternative splicing events at large distances. This is not necessarily true for the majority of cases of alternative splicing: in a normal situation, one should expect four (or at least three) different isoforms with no typical long-distance linkage.

Page 56: *Among all analyzed datasets, only two (***on** *one simulated and one real) were chosen for this manuscript* — why? How were they selected? Are they typical or best cases?

Finally, the manuscript has some language problems, contains misprints and minor inaccuracies. Here are some examples:

Page 4: *Although bacterial genomes are* **know <known>** *to be far less complex than large eukaryotic genomes in terms of* **<the>** *repeat structure, accurate and reliable algorithms for resolving repeats are still* **required <needed** or **necessary>** *to produce high-quality bacterial assemblies, especially for the case* **<of>** *single-cell data.*

Page 10: *Conventional whole genome sequencing (WGS) experiment requires millions of identical cells containing copy of a single genome. While this requirement usually sets no limitations for projects studying eukaryotic organisms, it may become a stumbling block in bacterial studies* — the author has in mind **multicellular** eukaryotes.

Page 16: *MismatchCorrector tool <…>* ***MismatchCorrection*** *is designed <…>*

Page 27: *2.5 Incorporating long-***rage <range>** *mate-pair libraries* — this is a really funny misprint in a heading.

Page 45: *A tip corresponding to a read with sequencing errors typically contains* ***a only few*** *mismatches* ***comparing*** **<compared>** *to* **<a>** *correct edge.*

Page 47: *One the most frequent difference between two alternative isoforms is exons inclusion/exclusion* — formally this is correct, but alternative splice sites are not less frequent.

Page 48. *During the analyses of misassemblies in various transcriptome assemblies generated by rnaSPAdes, it was revealed that* **<a>** *significant fraction of incorrect junction* **<junction>** *was* **<were>** ***cased*** **<caused>** *by using small* k *value of 21 during the first iteration* — it would be much better to rephrase: **The analysis of misassemblies <…> demonstrated that <…> had been caused <…>.**

The comments made above are purely editorial. Overall, the scholarship of Andrey Przhibelskiy, as proved by his dissertation, is excellent; his results are original and important; their relevance and validity are certain; the conclusions are convincing. The developed algorithms have been implemented in highly popular software, are published in first-class journals, and have been reported at major conferences. The dissertation meets all requirements of the St. Petersburg State University and the Russian Federation. Andrey Przhibelsky should be awarded the degree of candidate of physico-biological sciences in mathematical biology and bioinformatics (specialty 03.01.09).

Member of the Defense Council

Mikhail Gelfand, PhD (math), DSc (biol.), professor,
Deputy Director for Science,
A.A.Kharkevich Institute for Information Transmission Problems
(Russian Academy of Sciences),
B.Karetny per. 19, Moscow, 127051, Russia.
Tel: +7-916-6092971, e-mail: gelfand@iitp.ru

11.11.2019