Prof. dr. Marnix H. Medema
Professor of Bioinformatics

(+31) 317484706
marnix.medema@wur.nl

Bioinformatics Group
Wageningen University
Droevendaalsesteeg 1
6708PB Wageningen
The Netherlands

July 24th, 2023

To whom it may concern,

The PhD thesis of Tatiana Dvorkina addresses three interesting and highly relevant challenges in the field of computational metagenomics:

1. Alignment of biological sequences to assembly graphs
2. Identification of complex multi-domain proteins from assembly graphs
3. Automatic analysis of centromere sequence assemblies

Dvorkina's work is highly relevant, as massive amounts of short-read metagenome data are available that are often hard to functionally annotate and interpret. This thesis makes an important contribution to this field. In my review, I will focus on the first two challenges outlined above, as these are where my expertise lies.

The SPAligner algorithm allows mapping of long reads and functional genes to short-read-derived metagenomic assembly graphs, which can facilitate the identificiation of known genes in complex metagenomic datasets. The latter is nicely proven by reconstructing beta-lactamase-encoding genes from a metagenomic assembly that would be missed by a conventional search due to them being split across multiple contigs. I do wonder how often this problem is really relevant (how often are such proteins split across multiple contigs due to unresolved assembly graphs rather than insufficient sequencing depth) in actual metagenomic data, but regardless the approach is elegant and useful, also as part of larger metagenomic data analysis algorithms.

With her ORFograph algorithm, Dvorkina further expands this work by combining SPAligner with the HMM-based PathRacer algorithm to accurately identify complex multi-domain proteins from metagenomic assemblies. While it is not always entirely clear which of the high-scoring paths through a metagenome graph will represent actual proteins (and this might be complemented in the future by reference-based approaches), the technique represents a very useful discovery tool, and this strategy can indeed be further extended to other families of multi-domain proteins.

Compared to other PhD theses I have seen, the quality and impact of Tatiana's work is very good. The methods are innovative, the results are of high impact, and the chapters are clear and well-written. Also, all chapters have been published in highly esteemed journals and have been well received by the scientific community.

Beyond a few minor spelling errors, I have no major concerns about the quality of the work, nor on the appropriateness/correctness of the methodology. I look forward to an exciting scientific discussion with the soon-to-be Dr. Dvorkina on July 25th.

Yours sincerely,

Prof dr. Marnix H. Medema