Prof. dr hab. inż. Franciszek Seredynski
Cardinal Stefan Wyszynski University in Warsaw
Department of Mathematics and Natural Sciences
Wóycickiego 1/3, 01-938 Warsaw, Poland
Email: f.seredynski@uksw.edu.pl

Warsaw, 21.04.2015

REVIEW REPORT

on Ph.D. dissertation authored by

Michal Kierzynka

entitled
„GPU-ACCELERATED GRAPH CONSTRUCTION FOR THE WHOLE GENOME ASSEMBLY"

1. Problem and its impact

The main scientific problem that is considered in the thesis concerns the DNA de novo assembly. The purpose of the DNA de novo assembly is to reconstruct computationally the sequence of a genome that has not been known before. This is an indispensable element of DNA sequencing workflow, as current sequencing machines cannot read the whole genome at once, only its very short fragments, called reads.

The problem of DNA de novo assembly has been a vital combinatorial optimization problem ever since the sequencing technologies became available. The first graph-based approaches to this problem were published in 1980s. However, at that time the sequencing data were obtained through the hybridization method. This technique was very popular until around 2005 when so-called next-generation sequencing (NGS) methods started to attract more and more attention. The advantages of NGS include low costs of sequencing and the possibility of sequencing longer genomes. On the other hand, one of the main problems associated with this type of sequencing is the very large size of data sets that need to be processed. Therefore, the efficiency aspect of this problem plays a crucial role, especially when the problem was shown to be computationally hard, more exactly NP-hard. Therefore, researchers around the world are constantly improving methods for DNA assembly on both computational complexity and accuracy fronts.

In this thesis, the author proposed a novel algorithm for an efficient construction of the DNA overlap graphs, which are widely used in the DNA de novo assembly process. The work focuses on both an accuracy and efficiency of the algorithm. The quality of the results was deeply analyzed in multiple experiments, whereas the high performance was achieved due to highly parallelization of the proposed algorithm obtained by using graphics processing units (GPUs). Moreover, an open source implementation of the proposed method makes it a very practical tool that may be used to solve the outlined problem.

2. Contribution

The main original contribution of the author is the design and implementation of a novel heuristic algorithm for DNA overlap graph construction. In order to accomplish this goal several smaller problems have been addressed. In particular, the author proposed a new efficient algorithm for selection of similar sequences which is based on alignment-free sequence comparison. He also introduced to the literature the notion of *k-mer characteristics* which are used to estimate the similarity between sequences. Moreover, the author proposed several post-processing routines that increase the overall accuracy of the algorithm for graph construction (e.g. method based on paired-end reads). As a consequence, the quality of constructed graphs is very high. Finally, the proposed algorithm was parallelized in an interesting way, using the graphics processing units. As a result the execution time of the algorithm is reasonably short, which makes it a very practical tool for scientists dealing with DNA assembly.

The high quality of the research may be also confirmed by numerous scientific publications of the author, e.g. in *Journal of Parallel and Distributed Computing, BMC Bioinformatics or Bulletin of the Polish Academy of Sciences, Technical Sciences*. Some of the publications have been cited multiple times, and according to Web of Science the H-index of the author is equal to 4.

3. Correctness

The results presented in the thesis are new and interesting. The arguments are correct, reasoning is clear and the methodology is scientifically correct. A strong point of the thesis is the experimental part. The author presents a number of practical tests and experiments that support the correctness of the proposed algorithm. Both real and "in silico" data sets were used to investigate different aspects of proposed algorithm. Moreover, the author proposed an interesting way of measuring the quality of the constructed graph, which is based on confusion matrix. Therefore, the presented results are very accurate and meaningful. As a consequence the reader may trust what is claimed in dissertation.

4. Knowledge of the candidate

The thesis consists of 127 pages. It is divided into nine chapters, also contains appendix and bibliography. It was written in English and is preceded by a short two page abstract in Polish.

The first chapter is a general introduction to the problem considered in the thesis. It also formulates goals and scope of the thesis.

Chapter two introduces the reader to the biological aspects of the work and provides relevant definitions.

The third chapter constitutes an introduction to the theory of graphs, computational complexity of combinatorial problems and dynamic programming. Moreover, algorithms for pairwise sequence alignment are outlined.

Chapter four explains the details regarding both current and historical genome sequencing technologies.

In chapter five the author provides detailed description of different graph models. State-of-the-art algorithms for graph construction in the context of DNA assembly are compared and discussed in this chapter. The author justifies his choice of the graph model.

Chapters six to eight constitute the main part of the thesis, i.e. author's original findings.

Chapter six presents a novel designed and implemented software library oriented on GPU and CUDA programming model and tailored to needs related to the problem solved in the thesis.

Chapter seven contains a description and analysis of the proposed original heuristic algorithm for the DNA overlap graph construction, which is a solution of the problem posed in the thesis.

The results of experimental study of the proposed algorithm are presented in Chapter eight. In silico data sets and two real data sets coming from the Illumina Genome Analyzer II sequencer were used in experiments.

The last chapter concludes the thesis. It describes all the original achievements and work that was accomplished. It provides suggestions for future research.

The Appendix A explains the practical aspects related to the software implementation and usage.

The chapters are of good quality and seem to be complete in terms of the scope of the thesis. The standard of the English language is very high. The general knowledge of the candidate may be also confirmed by a long and complete list of references, which include also those articles that were published quite recently.

5. Other remarks

Minor remarks, ambiguities and mistakes:

- Page 12: In the introductory section the author claims that the graphics cards may increase not only the speed but also the accuracy of the heuristic algorithms. It is not explained in what way, and the reader must go through the rest of the thesis to find it out.
- Page 24: "A" at the beginning of a sentence should be started from a new line.
- Page 74: The "SLI" algorithm could be illustrated with more examples, as it is quite difficult to follow.
- Page 76: The formatting of the brackets looks quite odd, e.g. "(cf. 1) in Figure 7.8)".
- Page 105: "Testing environment" could be described before the author presents the results of the tests.

All the above mentioned remarks are minor and do not reduce the overall quality of presented thesis, which is very high. They do not affect the core idea that is described in this thesis either.

6. Conclusion

Taking into account what I have presented above and the requirements imposed by Article 13 of *the Act of 14 March 2003 of the Polish Parliament on the Academic Degrees and the Academic Title* (with amendments)[1], my evaluation of the dissertation according to the three basic criteria is the following:

A. Does the dissertation present an original solution to a scientific problem? (the selected option is marked with **X**)

[X]	[ ]	[ ]	[ ]	[ ]
Definitely YES	*Rather yes*	*Hard to say*	*Rather no*	*Definitely NO*

B. After reading the dissertation, would you agree that the candidate has general theoretical knowledge and understanding of the discipline of **Computing**, and particularly the area of **Bioinformatics**?

[X]	[ ]	[ ]	[ ]	[ ]
Definitely YES	*Rather yes*	*Hard to say*	*Rather no*	*Definitely NO*

C. Does the dissertation support the claim that the candidate is able to conduct scientific work?

[X]	[ ]	[ ]	[ ]	[ ]
Definitely YES	*Rather yes*	*Hard to say*	*Rather no*	*Definitely NO*

Moreover, taking into account the high quality of the results obtained in the dissertation confirmed by top international journals and citation data I **recommend to distinguish** the dissertation for its quality.

Signature

[1] http://www.nauka.gov.pl/g2/oryginal/2013_05/b26ba540a5785d48bee41aec63403b2c.pdf