Prof. Dr Pascal Bouvry
University of Luxembourg
6 rue Coudenhove-Kalergi
L-1359 Luxembourg
e-mail: Pascal.Bouvry@uni.lu

Luxembourg, the 27th of April, 2015

Reviewer's opininion on Ph.D. dissertation authored by Michal Kierzynka entitled
"GPU-accelerated Graph Construction for the whole Genome Assembly"

1. Problem and its impact

In recent decades the advances in molecular biology allowed to conduct new types of biochemical experiments. Their main purpose is to better understand the nature in general and in particular the mechanisms governing the animate matter including the human being. However, the analysis of information gathered from biochemical experiments would not be possible without computer science and dedicated algorithms. Bioinformatics is a branch of science that has emerged as a natural consequence of interdisciplinary research conducted on the verge of computer science, molecular biology and medicine. Importantly, over the last several years the amount of genetic information obtained through biochemical experiments has increased significantly. Therefore, most of the software tools for data processing and analysis that were sufficiently good a few years ago no longer meet the requirements of modern science.

One of the problems in this area that is affected by the "big data" syndrome is the DNA *de novo* assembly. The ultimate goal of this problem is to reconstruct the original sequence of the sequenced genome without any reference information regarding its structure. The DNA assembly is one of the most important problems in molecular biology as it allows to answer the question about the "shape" of a genome which is the basic carrier of genetic code. This is also vital in modern, personalized medicine which more and more often looks at the individual's genetic code and its variations to answer different questions regarding humans' health. From the computational perspective, the DNA *de novo* assembly problem belongs to the NP-hard class, and therefore is extremely hard to solve, especially in the context of massive amount of sequencing data.

The most important problem discussed in this thesis concerns the construction of the DNA overlap graphs that are used in a basic step of the DNA *de novo* assembly. The author presented an original, parallel algorithm that efficiently deals with this problem. His approach is undoubtedly scientific and presented results confirm the practical relevance of proposed method.

2. Contribution

The main original contributions of the author in the presented thesis are:

- development of a new algorithm for the DNA overlap graph construction,
- a new method of alignment-free sequence comparison and sorting based on full and partial kmer characteristics,
- detailed analysis of the algorithm with respect to parallelization and very efficient implementation of the most time consuming part of the algorithm, i.e. sequence alignment, on graphics processing units,
- an interesting method of detailed analysis of accuracy of the graph based on confusion matrix.

The presented results are new and interesting. The author has not only presented but also proved the advantage of his method over the other algorithms known from literature. The high quality of research may be confirmed by a number of scientific publications, e.g. in *BMC Bioinformatics* or in *Journal of Parallel and Distributed Computing*. Additionally, according to Web of Science, the author has been cited 38 times in total. Finally, the work has also a practical aspect, as the implementation of the proposed algorithm has been released on GPL license, and may be used by scientists working on DNA assembly or other related topics.

3. Correctness

The arguments presented in the thesis in question are scientifically correct. The results obtained on real as well as specially prepared artificial data sets confirm the correctness of the proposed method for DNA overlap graph construction and other considered subproblems. The extensive tests presented by the author are far from being a “black box”, since the testing methodology is also thoroughly described. The design of the experiments is adequate to the outlined problem and the good results were feasible to obtain. Moreover, the quality of the software delivered as part of this thesis seems to be very high, as its performance (speed, ability to handle real and large data sets, accuracy) has been validated in a transparent way.

4. Knowledge of the candidate

The thesis consists of nine chapters and one appendix, all of them written in a clear and understandable way. The work is well structured and its layout is transparent. The author presents also a good command of English. The first five chapters have a form of introduction or tutorial and present the general knowledge in the field of: computer science, molecular biology, sequencing and DNA assembly. More specifically, the first chapter is a general introduction to the thesis. The second presents the basic notions in molecular biology, including the genetic code and sequence alignment. The third chapter outlines the problems of algorithms theory and computational complexity, with a special emphasis on graph theory and dynamic programming that are widely used in this thesis. The chapter four is a short compendium of historical and current sequencing technologies. The fifth chapter in turn presents state-of-the-art graph approaches that are used for DNA assembly. The author compares them and comments on their pros and cons to choose finally the best graph representation of the problem. All these chapters are important to understand the rest of the thesis which presents the original work and findings of the author. A novel implementation of pairwise sequence alignment on a GPU is presented in chapter six. The next chapter covers the topic of a new algorithm for DNA overlap graph construction, which is thoroughly tested in chapter eight. The last regular chapter concludes the thesis. Finally, appendix A provides details regarding software usage.

All these chapters confirm a strong background of the author not only in the area of computer science (computing discipline) but also in bioinformatics, and especially in next-generation sequencing. Additionally, the author cites over a hundred top-class papers from these fields and the list of references seems to be complete.

5. Other remarks[1]

Even though the thesis is written in a proper way, the author made a few minor mistakes:

- Page 21, Figure 2.4: the caption of the figure is not self-explanatory and the reader must look for the explanation in the chapter. The same applies for the colours in the figure.
- Page 73: Table 7.2: it is not obvious why there are only twelve nucleotide alphabets. This is correct, but the author could add more exhaustive explanation of this fact.
- Page 78, Figure 7.10: the presented fragment of the source code may be not entirely clear to the reader who is not familiar with programming.
- Page 85: "in case" should be replaced by "in the case".

Nevertheless, these drawbacks do not affect the overall quality of the presented research and certainly do not disqualify it in any way.

6. Conclusion

Summing up, the PhD thesis presented by Michal Kierzynka is written technically correct and with due care. It presents an original solution to a scientific problem. The author demonstrated an extensive knowledge and excellent skills in formulating and solving difficult problems in computer science, and in particular in bioinformatics. The main goal of the thesis was achieved successfully, which supports the claim that the candidate is able to conduct scientific work. As a result of opinion expressed above, I consider that the thesis entitled "GPU-accelerated graph construction for the whole genome assembly" meets the high standards defined by the current doctoral dissertation law.

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)[2], 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*

[1] Optional

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

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*

Prof. Dr Pascal Bouvry
Computer Science and Communication Research Unit