Quality of Solution of Massively Parallel MSA Based on Metaheuristic Social Behavioral Model

Plamenka Borovska[1, a)] and Maria Marinova[2, 3, b)]

[1]*Technical University of Sofia, 8 boul. Kliment Ohridsky, 1000 Sofia, Bulgaria*
[2] *Technical University of Sofia, Plovdiv Branch, 25 Tsanko Diustabanov St, 4000 Plovdiv, Bulgaria*

[a)]pborovska@tu-sofia.bg
[b)] m_marinova@tu-sofia.bg

Abstract. Due to their importance for computational biology, the methods and algorithms for multiple alignment of biological sequences have been the subject of continuous intensive research and innovation in the last 30 years. Within a PRACE research project we have designed and implemented metaheuristic algorithm for massively parallel MSA and based on it parallel software tool MSA_BG that has been ported on the European supercomputer Juqueen. In this paper we have investigated the Quality of Solution (QoS) of the parallel algorithm MSA_BG for multiple alignment of biological sequences on GPU-accelerated computing infrastructure based on 3 synthesized benchmarks of genetic sequences comprising the 8 segments of the swine virus AH1N1, the 9 genes of the severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human as well as the human genes BRCA1 and BRCA2 associated with the breast cancer issue. The analysis of the experimental results show that the QoS of the individual runs fall in the range of ∓ 10%, except for the case study of BRCA1 and BRCA2 where the range is within ∓1%.

THE PROBLEM AREA

Comparative analysis of molecular sequence data is essential in computational biology to reconstruct the evolutionary histories of species and to determine the nature and extent of the selective forces that shape the evolution of genes and species. Due to their importance, the methods and algorithms for multiple alignment of biological sequences [1] have been the subject of continuous intensive research and innovation in the last 30 years. Considering that multiple sequence alignment (MSA) is a NP - hard problem there exist a wide spectrum of exact and approximate algorithms for solving it such as algorithms based on dynamic programming, progressive alignment, iterative methods, consensus methods, maximum parsimony, alignment by blocks, heuristic and optimization methods. Practically, MSA is a combinatorial optimization problem [2] characterized by the effect of "combinatorial explosion" i.e. the increase of data size results in drastically exponential increase of computational time. In [3] an innovative parallel algorithm MSA_BG for multiple alignment of biological sequences that is highly scalable and locality aware has been proposed. The designed MSA_BG algorithm is iterative and is based on the metaphor of Artificial Bee Colony metaheuristics and the concept of algorithmic and architectural spaces correlation. In [4] we have presented experimental results for parallel performance and scalability evaluation of the parallel software tool MSA_BG based on the designed algorithm and ported on the European supercomputer JUQUEEN (PRACE project) which have shown speedup up to 70 times for 10^7 iterations and machine size of 512 computing cores. In [5] we have revealed the experimental results of code optimization of multiple sequence alignment software tool MSA-BG on GPU-accelerated computing infrastructure that have shown the estimated speedup of MSA_BG_CUDA versus MSA_BG_ABC (12 threads) for the case study of the swine virus dataset (H1N1) for 1000 iterations is up to 2.5.

The goal of this paper is to investigate the convergence and to estimate the Quality of Solution (QoS) of the parallel algorithm MSA_BG for multiple alignment of biological sequences based on 3 synthesized benchmarks of genetic sequences comprising the 8 segments of the swine virus AH1N1, the 9 genes of the severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human as well as the human genes BRCA1 and BRCA2 associated with the breast cancer issue.

METHODOLOGY FOR ESTIMATING PERFORMANCE AND QUALITY OF SOLUTION OF PARALLEL METAHEURISTICS ALGORITHMS

Due to the complexity of the problems and the limited computational resources for solving them (time, memory), the development of metaheuristics has become a major area in operations research. Metaheuristics ensures the finding of optimal or suboptimal solutions in a reasonable time. The main goal of metaheuristics is to solve combinatorial optimization problems, which are generally multidimensional and complex, using models represented by discrete variables. Metaheuristics provides strategies to guide the randomized search process, and the goal is to effectively explore the search space to find optimal or sub-optimal solutions within reasonable computational time. Metaheuristic algorithms are universal, approximate and usually nondeterministic. The basic concepts of metaheuristics can be described on an abstract level and are not related to solving a specific problem. It requires a dynamic balance between the use of two fundamental concepts: diversification and intensification. Diversification provides exploration of as many areas of the search space as possible. Intensification implies the use of the accumulated experience in the search, which in turn saves time not to check already researched solutions and allows the search to be concentrated in areas with good quality solutions. The balance between the two concepts is crucial - on the one hand for the rapid identification of areas of the search area with high quality solutions, and on the other hand - to limit the search time in areas with low quality solutions or to avoid repeated search in already searched areas. In this context, of particular importance in meta-heuristics are the exploration of new areas that have not been explored and the use of the experience gained from previous exploitation, which is stored in a certain type of memory. The metaphor used in parallel metaheuristics is as follows: the search space is determined by the combination of the specific problem to be solved, the neighborhood relations, and the guiding function.

Population based methods use concepts and models inspired by nature. Consequently, iterative metaheuristic algorithms are created, and at each iteration many solutions (populations) are analyzed. Artificial Bee Colony metaheuristics is actually a SWARM intelligence algorithm based on the model of the distributed intelligent collective behavior of bees. SWARM metaheuristics use agents that can be simple or complex, with training, adaptive, recompiled, reactive or planned, with or without internal states, stochastic or deterministic, fully decentralized, or can use a form of centralized management. Common to all algorithmic frameworks is that agents interact (communicate).

The main objectives of the experimental study of parallel metaheuristic algorithms are verification (testing the performance of the parallel algorithm), evaluation of performance, scaling and quality of solutions of the parallel algorithm (based on statistical analysis). A major problem is the development of a methodological framework for the experimental evaluation of parallel metaheuristics. The aspects of evaluation include the design of the experiment, the discovery of good sources of data sets, the measurement of algorithmic performance in a meaningful way, meaningful analysis and a clear presentation of the results. The metrics to assess the performance of parallel metaheuristics are execution time, speedup, and efficiency. Execution time is measured on a wall clock basis (including the time for additional costs, as they reflect the cost of parallelisms). The most important measure for the parallel algorithm is the speedup - the ratio of the sequential execution time to the parallel execution time $S_n = T_1/T_n$, where n is the number of processors. However, this measure is not applicable to metaheuristics due to the nondeterministic nature. In parallel metaheuristics, acceleration is defined as the ratio of the *mean* time for serial execution to the *mean* time for parallel execution.

$$S_n = T_{1,mean}/T_{n,mean} \qquad (1)$$

The quality of the solution obtained as a result of the execution of the parallel metaheuristic algorithm is defined as:

- Number of hits or speed of success (provided that the optimal solution is known), i.e. the percentage of successful runs (% hits);
- Mean values of the quality of the solutions (provided that the optimal solution is not known or localized);

It is accepted that the evaluation of the quality of the solutions should be based on a specified number of runs of the parallel algorithm (it is recommendable not to repeat the initial configurations).

Because of the nondeterministic nature of the parallel metaheuristic algorithm designed we estimate the quality of the solution as the average value of the sum of the QoS's obtained at each run to the number of the runs.

$$QoS = {\sum_{n}^{1} QoS_i}/{n} \qquad (2)$$

Where QoS is the mean value of the quality of solution for a group of benchmark sequences of similar length and functionality, *i* is the consecutive number of the run and n is the total number of the experimental runs of the multiple sequences alignment tool.

Deviation from the mean value of the QoS is defined as the difference of the value of QoS obtained at a specific run and the mean value of the QoS for all runs.

$$QoS\_DEV_i = {QoS_i - QoS}/{QoS} \qquad (3|)$$

Where *i* is the number of the run, QoS_i is the quality of solution, obtained at run *i*, and QoS is the the mean value of the quality of solution for the group of benchmark sequences for all the conducted runs.

TECHNOLOGICAL EXPERIMENTAL FRAMEWORK

We have conducted a number of experiments on a GPU accelerated computing infrastructure in order to estimate experimentally the performance and Quality of Solution (QoS) parameters of our software tool for multiple sequence alignment MSA_BG_CUDA. The architectural specifics of the computing infrastructure with GPU accelerators [7] are shown in Fig.1.

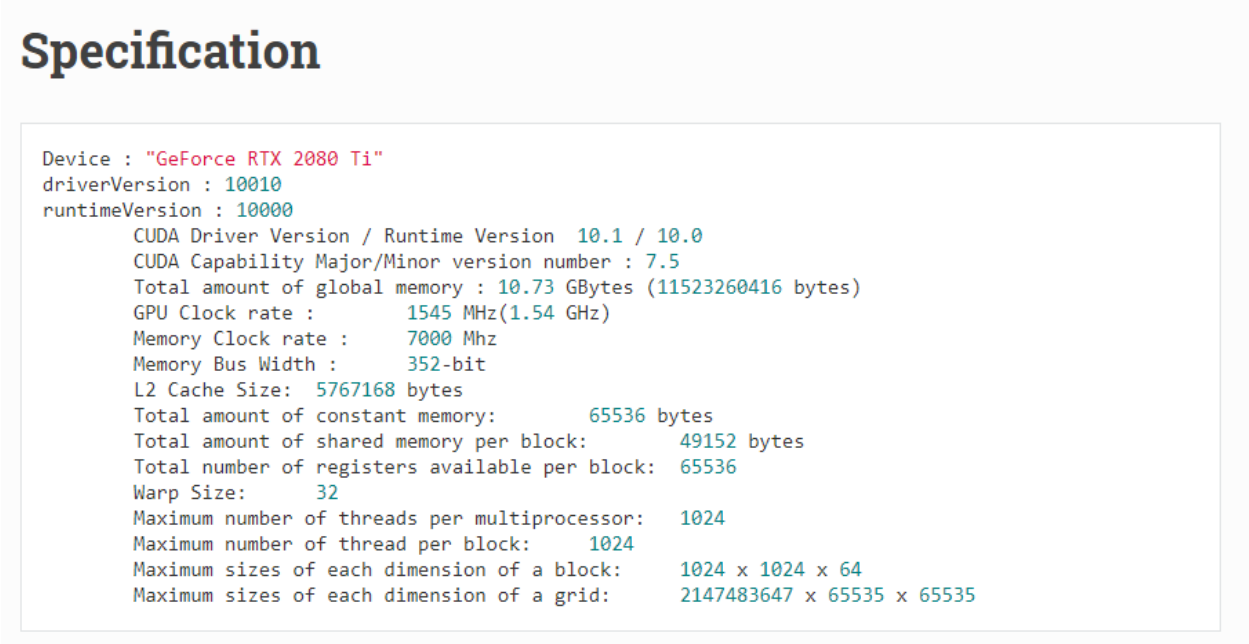

FIGURE 1. Architectural attributes of the experimental GPU accelerated computing infrastructure.

The experimental data sets comprise 3 categories of genetic sequences accessed from GenBank, NCBI: (1) the 9 genes of the severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human, (2) pandemic swine influenza virus AH1N1, and (3) the human genes BRCA1 and BRCA2 associated with the breast cancer issue.

SARS-CoV-2 is a new strain of coronavirus that has not been previously identified in humans [8]. Using different methods and tools it has been shown that the length of the genetic sequences is about 30 000 bp linear RNA comprising

9 genes. In order to overcome the problem of overloading the computational resources we have applied preprocessing of the genetic data of SARS-CoV-2 by performing gene finding within the genome via the GeneMark tool. GeneMark is a family of gene prediction programs developed at Georgia Institute of Technology, Atlanta, Georgia, USA [9]. We have used the module for Gene Prediction in Viruses, Phages and Plasmids - the self-training program GeneMarkS. In fig.2 we have shown the result of the application of GeneMark software for identifying the genes within MT276326.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/GA_2741/2020, complete genome [9]. Obviously, the lengths of the 9 genes within the SARS-CoV-2 genome isolate varies in a wide range for the different genes – from 13218 bp for gene 1 down to 186 bp for gene 6 (Fig.2).

The experimental data sets of the genetic sequence of the pandemic swine influenza virus AH1N1 are accessed from NCBI GenBank [11] and comprise 8 segments: PB2 PB1 PA HA NP NA MP NS.

The experimental data sets of the genetic sequence of the human genes BRCA1 and BRCA2 associated with the breast cancer issue are accessed from NCBI GenBank [12].

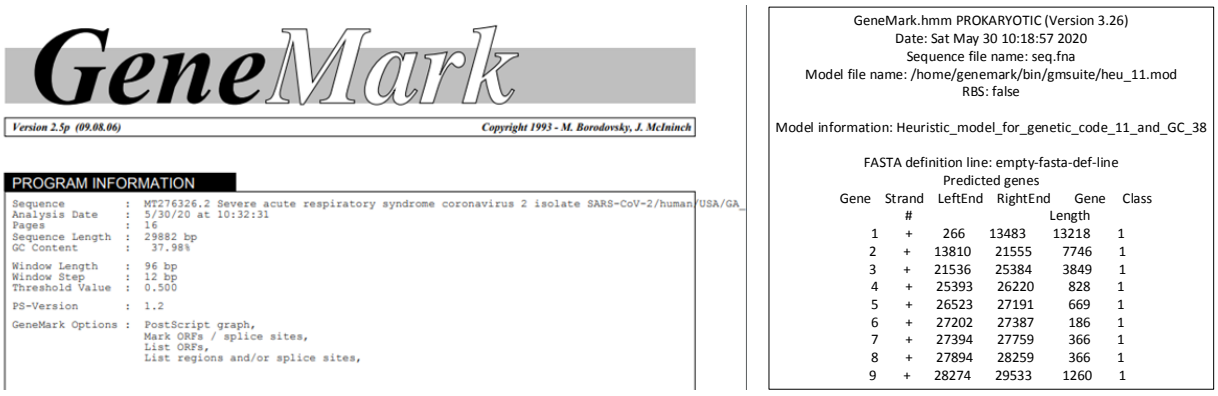

FIGURE 2. The result of the application of GeneMark software for identifying the genes within MT276326.1 Severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human/USA/GA_2741/2020, complete genome

EXPERIMENTAL RESULTS AND ANALYSIS

We have conducted a series of experiments of multiple sequence alignment on 9 benchmark groups each comprising 100 nucleotide sequences of isolates of the respective genes (Table 1). In Fig.3 and Fig. 4 the evaluations of QoS for MSA of SARS-CoV-2 gene 1 and gene 2 are shown, respectively. Fig.5 presents the results for the evaluations of QoS for MSA of SARS-CoV-2 gene 3 through gene 9 for 5 consecutive runs. Fig.6 depicts the deviation of the QoS of each run from the mean value measured in bp, while Fig.7 presents the deviation in percent. Fig.8 presents the results of QoS evaluation of pandemic swine influenza virus AH1N1, and Fig. 9 shows the results for breast cancer associated human genes BRCA1 and BRCA2.

The analysis of experimental results show that for the case study of SARS-CoV-2 genes the deviation of the QoS obtained is much less than 10% except for gene 8 meaning difference in about 20 in terms of the number of nucleotides. For the case study of MSA of pandemic swine influenza virus AH1N1 nucleotide sequences the deviation is within the range +11% and -9.88%. For the investigation of human gene BRCA1 the observed deviation of the quality of solution is within the range of +1.13% to -1.13%, while for the case study of BRCA2 deviation is within the range of +0.45% to -0.25%, which are quite satisfactory results and confirm the efficiency of our MSA_BG_CUDA software tool.

Table 1: Experimental results for SARS-CoV-2 genetic data set

SARS-CoV-2	Run #1	Run #2	Run #3	Run #4	Run #5	Mean Value
	Grade_Out	Grade_Out	Grade_Out	Grade_Out	Grade_Out	Grade_Out
Gene 1	5509	5540	5504	5500	5500	5511
Gene 2	1181	1181	1098	1200	1150	1162
Gene 3	259	300	262	259	259	267.8
Gene 4	368	366	360	345	344	356.6
Gene 5	290	280	307	298	291	293.2
Gene 6	76	80	76	76	76	76.8
Gene 7	88	79	86	88	84	85
Gene 8	38	31	31	23	33	31.2
Gene 9	138	134	120	120	145	131.4

FIGURE 3. The evaluations of QoS for MSA of SARS-CoV-2 gene 1

FIGURE 4. The evaluations of QoS for MSA of SARS-CoV-2 gene 2

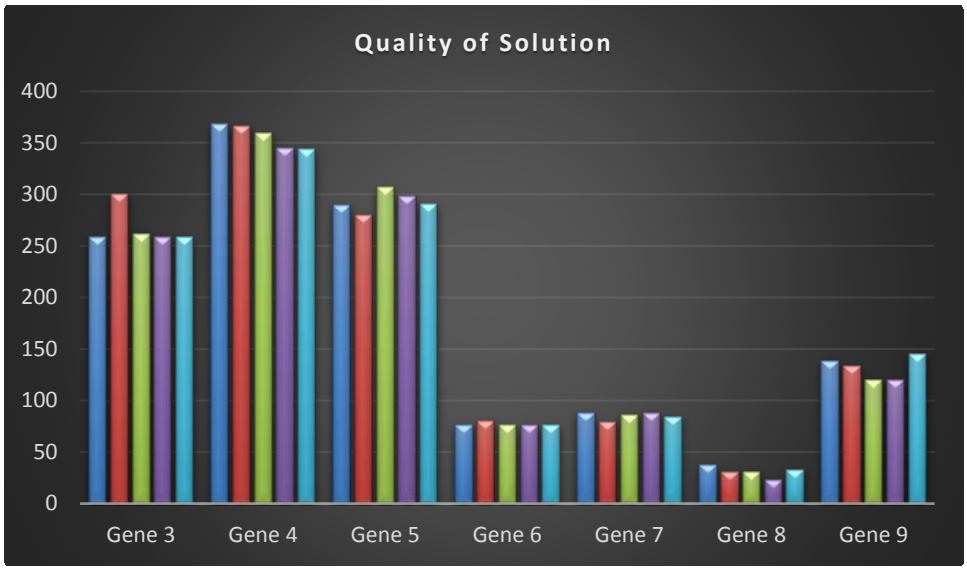

FIGURE 5. The evaluations of QoS for MSA of SARS-CoV-2 gene 3 through gene 9

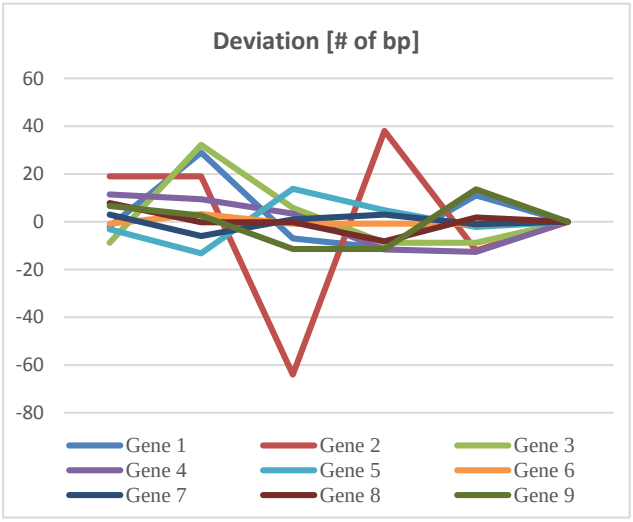

FIGURE 6. The deviation of QoS for MSA of SARS-CoV-2 genes in base pairs

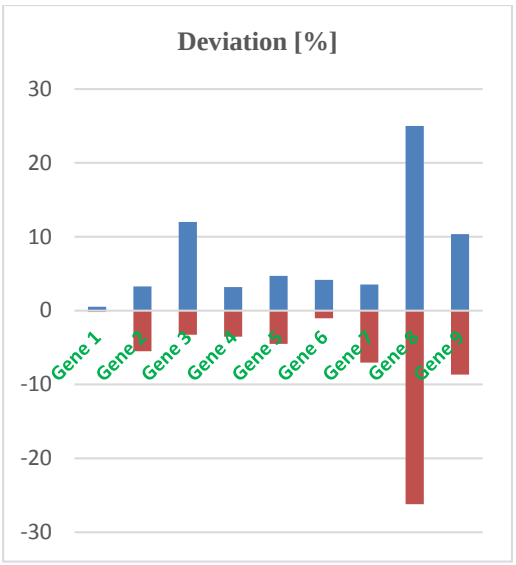

FIGURE 7. The deviation of QoS for MSA of SARS-CoV-2 genes in percentage

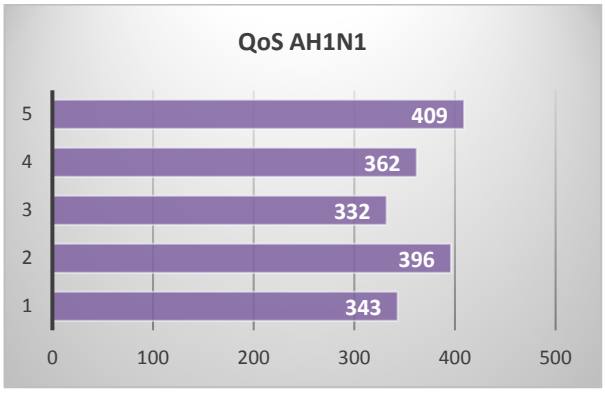

FIGURE 8. The evaluations of QoS for MSA of influenza swine virus AH1N1

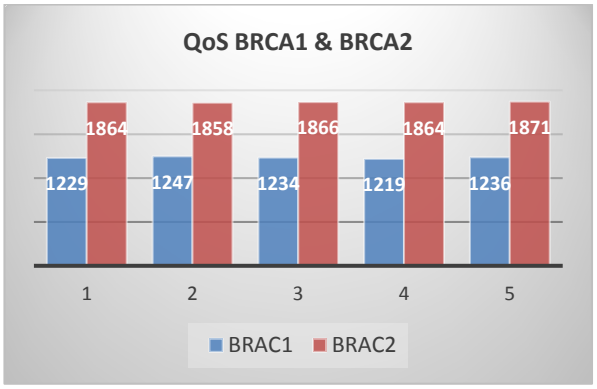

FIGURE 9. The evaluations of QoS for MSA of genes BRCA1 and BRCA2

CONCLUSION AND FUTURE WORK

In this paper we have investigated the quality of solution obtained by our software tool for multiple sequence alignment based on the intelligence of the collective behavior of artificial bee colonies. The experimentation has been conducted for 3 benchmark groups of nucleotide sequences: (1) the 9 genes of the severe acute respiratory syndrome coronavirus 2 isolate SARS-CoV-2/human, (2) pandemic swine influenza virus AH1N1, and (3) the human genes BRCA1 and BRCA2 associated with the breast cancer issue, involving 5 runs on a high performance GPU-accelerated computing infrastructure. The analysis of the experimental results show that the QoS of the individual runs fall in the range of $\mp$ 10%, except for the case study of BRCA1 and BRCA2 where the range is within $\mp$1%.

In future, our intention is to change the communication paradigm of migrating the elite ants from broadcast to queen bee to the island model, and thus slowing down the time of the parallel algorithm convergence and increasing the level of diversification with purpose of improving the quality of solution obtained.

ACKNOWLEDGMENTS

This paper presents the outcomes of research project “Intelligent Method for Adaptive In-silico Knowledge Discovery and Decision Making Based on Analysis of Big Data Streams for Scientific Research”, contract ДН07/24, financed by the National Science Fund, Competition for Financial Support for Fundamental Research – 2016, Ministry of Education and Science, Bulgaria.

REFERENCES

1. Wang L, Jiang T., "On the complexity of multiple sequence alignment". J Comput Biol. **1** (4): 337–348, 1994 doi:10.1089/cmb.1994.1.337.
2. Just W. "Computational complexity of multiple sequence alignment with SP-score", J Comput Biol. **8** (6): 615–624, 2001
3. Borovska, P., Gancheva, V., Landzhev, N., “Massively parallel algorithm for multiple biological sequences alignment”, 2013 36th International Conference on Telecommunications and Signal Processing, TSP 2013, Rome; Italy; ISBN: 978-147990404-4, DOI: 10.1109/TSP.2013.6614014, IEEE Xplore.
4. Borovska, P., Gancheva, V., Georgiev, I., Ivanova, D., Hybrid parallel multiple sequence alignment based on artificial bee colony on the supercomputer JUQUEEN, Proceedings - 2017 European Conference on Electrical Engineering and Computer Science, EECS 2017, Bern; Switzerland, Pages 47-51, ISBN: 978-153862085-4, DOI: 10.1109/EECS.2017.18, IEEE Xplore
5. Borovska, P., Marinova, M., Tsanov, V., Code optimization of multiple sequence alignment software tool MSA-BG on GPU-accelerated computing infrastructure, AIP Conference Proceedings, Volume 2172, 13 November 2019, 45th International Conference on Application of Mathematics in Engineering and Economics, AMEE 2019; Sozopol; Bulgaria; 2019; ISBN: 978-073541919-3 ISSN:0094-243XE-ISSN:1551-7616 pp. 020006-1 - 020006-10
6. D. Karaboga, “An Idea Based On Honey Bee Swarm for Numerical Optimization,” Technical Report-TR06, Erciyes University, Engineering Faculty, Computer Engineering Department, 2005, https://www.researchgate.net/publication/255638348_An_Idea_Based_on_Honey_Bee_Swarm_for_Numerical_Optimization_Technical_Report_-_TR06.
7. https://www.gamersnexus.net/guides/3364-nvidia-turing-architecture-technical-deep-dive
8. GeneMark http://opal.biology.gatech.edu/GeneMark/
9. https://www.ncbi.nlm.nih.gov/nuccore/MT276326
10. http://exon.gatech.edu/GeneMark/
11. https://www.ncbi.nlm.nih.gov/genomes/FLU/SwineFlu.html
12. The BRCA1 and BRCA2 Genes https://www.cdc.gov/genomics/disease/breast_ovarian_cancer/genes_hboc.htm