iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 6 **Issue:** VII **Month of publication:** July 2018

DOI: http://doi.org/10.22214/ijraset.2018.7076

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VII, July 2018- Available at www.ijraset.com

Inferences of Transcriptional and Translational Regulatory Modules for Systems Biology: A Review

Suhas A Bhyratae[1], Dr. Neha Mangla[2]
[1]*Research Scholar, Dept. of Information Science & Engineering, Atria Institute of Technology, Bengaluru, India*
[2]*Research Supervisor, Dept. of Information Science & Engineering, Atria Institute of Technology, Bengaluru, India*

Abstract: ***Computational molecular biology is the superset of Bioinformatics. Areas such as genomics, mapping, sequencing, determination of sequence and structure are applied by computer and data science classically. Computational biology is the application of computational tools and techniques to molecular biology. It is a multidisciplinary area of study that combines biology, computer science and statistics. There are different types of problems in computational biology such as how to recognize genes in DNA sequences that contain regulatory information, to determine interaction and regulation of all genes, to predict the structure of proteins, to forecast the functions of newly discovered proteins, to cluster and classify proteins into families and to align similar proteins etc. computational biology driven by two factors, 1. The recent explosion in the amount of available biological data, 2. Commercial incentive to exploit the available biological data for drug discovery and other developments. The algorithms are computationally expensive and computational patterns range from regular to very irregular structured patterns. This outlines computational issues related to parallelism i.e., parallel programming approaches which involves increasing efficiency and creates accurate outcomes.***
Keywords: ***Gene regulatory networks, parallelization,***

I. INTRODUCTION

Inconsistency in gene expression is the necessary mechanism underlying animal development and cell differentiation. Genes need to be transcribed at the right time, in the right amount and in the right cells in order for development to proceed correctly. A significant portion of gene regulation results from the interaction of sequence-specific DNA-binding factor (or transcription factors (TFs)). A very close examination of genome helps to know how it varies with time. The most well-known and common applications of bioinformatics are sequence analysis [1] and phylogenetic analysis. Nucleotide or protein sequences used for bioinformatics may be collected in different ways. One technique is to go through a whole genome and search individual sequences to record and store it. Another method is to simply collect large amounts of fragments and perform comparisons on them, and then determining whole sequences by overlapping the unused fragments. The latter method discussed is known as shotgun sequencing [2], and is presently the most popular method because of its speed and the ease of use. By comparison of known sequences of a genome to specific changes, much information can be obtained about undesirable changes and diseases such as cancers. With the whole sequencing of the human genome, bioinformatics has played very important role in the research of cancers.

The composite cellular network formed by the cooperating macromolecules underlies an organism's phenotypes [3]. Biomolecules are often thought to shape into interacting modules (functional building blocks) for completing a specific biological process [4]. This opinion is supported by the fact that many observable phenotypic variances are frequently not determined by a single gene but by a set of interacting genes. Efficient reconstruction of a complete map of these interacting molecular modules is vital for understanding an organism's genetic architecture underlying phenotypes. Several methods have been developed to find functional gene modules by utilizing transcriptome data. Differential Expression (DE) analysis uses traditional statistical hypothesis testing-based approach, such as t-test, F-test, ANOVA or negative binomial test for assessing statistical significance of an observed expression change of each individual gene by comparing the between-conditions variation and within-condition variation, which can reveal the genes related to specific experimental conditions or sample types [5, 6]. However, differentially expressed genes are only a proxy for finding the key molecular modules related to our concerned biological questions because of highly dynamic transcriptome in different types of cells, tissues and experimental conditions [7]. Complementary with the DE analysis, differential gene co-expression analysis aims to identify a group of differently co-expressed genes under two or more conditions, which has been applied to discern condition-specific gene co-regulation patterns [8-10]. Bi-clustering analysis is an approach that performs simultaneous clustering on genes and conditions across a wide range of transcriptome experiments. This method can discern the groups of genes that demonstrate similar expression patterns underlying the specific conditions but behave independently under

other conditions. Gene co-expression meta-analysis is another powerful method, which adopted the all experimental conditions to build co-expression network, when compared with bi-clustering analysis [11-13]. This document also surveys the computational strategies followed the execution of processes (Transcription and Translation) are carried out simultaneously. The term parallel computing refers to the simultaneous execution of processes. Structured parallel programming is recognized as a viable and effective means of tackling parallel programming problems [14]. Typically, the most common simulation technique for parallel computing is used as stochastic simulation [15]. This technique comes as a parallel approach to ordinary differential equations (ODEs) and is able to describe transient and multi-stable behaviors of biological systems that do not appear with ODEs modeling. The stochastic simulation uses Monte Carlo methods applied to a number of independent instances, in order to obtain a result that is statistically significant. Because of the independence, these instances make the simulation an disconcertingly the parallel problem. Groups of instances can be executed independently. This does not irrelevantly extend to the whole simulation-analysis pipeline, which exhibits data dependencies between the two stages of the processes.

II. RELATED WORK

The following table stretches the related work towards the Transcriptional modules, Translational modules and parallelization processes.

Table 1:An overview of Transcription, Translation and Parallelization processes.

Authors	Purpose	Description
Haitao Guo et. al., (2017)	Modelling the regulatory structure of CRMs	Proposed a CRM discovery algorithm called ComSPS
Julie Dubois et. al., (2017)	Regulatory activities	Used mouse liver CRMs involved in regulatory activities of the hepatic TR, NR1H4 by multiple TRs at CRMs.
Luo Jiawei et. al., (2017)	Detect co-regulatory modules	RWRRGM first identifies regulator and gene modules by greedily expanding seed nodes and then walks on the identified modules randomly. Finally, functional homogeneous regulator and gene modules are integrated to form co-regulatory modules.
Hua Yu et. al., (2017)	Establishing and analysing the RNA-seq-based GCNs	Explored the regulatory mechanism of the modules by enrichment of the known 19 cis-elements, transcription factors and mRNA targets.
Yosvany López et. at., (2017)	study of the regulatory regions of genes expressed	Proposes the use of a combination of structural features, such as positioning of individual motifs relative to the transcription start site, orientation, pairwise distance between motifs, and presence of motifs anywhere in the promoter for predicting gene expression from structural features of promoter sequences.
Wenbin Guo et. at., (2017)	Proposed a novel network inference methods based on Relevance Low order Partial Correlation (RLowPC).	The proposed RLowPC method effectively reduces the indirect edges predicted as regulatory relationships and increases the precision of top ranked predictions. P
Andrea Bracciali et. al., (2014)	addressed an optimisation problem	The parallel implementation on multi-core architectures allows for a relevant reduction of the execution time for haplotyping.
Fabio Tordini et. al., (2016)	propose a model for multi-omic data integration	Describes how driver and passenger mutations accumulate during the development of diseases providing which exploits a multi-layer network approach to analyse, visualize and obtain insights for which is designed according to the Fast Flow pattern-based approach on such biological information.

Marco Aldinucci et. al., (2015)	introduce a new pool evolution parallel pattern models	The parallel evolution of a population subject to mutations and evolving in such a way that a given fitness functions is optimized.
Fabio Tordini et. al., (2015)	introduce NuChart-II, a tool for Hi-C data analysis	This provides a gene-centric view of the chromosomal neighbourhood in a graph-based manner
Marco Aldinucci et. al., (2014)	Presents a multi-core parallelization of WhatsHap	WhatsHap moves complexity from fragment length to fragment overlap and is hence of particular interest when considering sequencing technology's current trends
Ivan Merelli et. al., (2015)	Proposed a novel sequencing technique called Chromosome Conformation Capture	The analysis of the chromosome organization in the cell's natural state
Marco Aldinucci et. al., (2013)	advocate the high-level software design as a vehicle for building efficient and portable parallel simulators for the cloud	which is designed according to the Fast Flow pattern-based approach
Marco Aldinucci et. al., (2013)	parallel programming	Demonstrated how the same parallel building block set in parallel programming abstractions
Marco Aldinucci et. al., (2013)	advocated the high-level design of simulators for stochastic systems	Fast Flow has been extended to support also clusters of multi-cores with minimal coding effort, assessing the portability of the approach.
Sander Pronk et. al., (2013)	Model complex bio molecular interaction and function in a manner directly testable by experiment.	GROMACS supports several implicit solvent models, as well as new free-energy algorithms, and the software now uses multithreading for efficient parallelization even on low-end systems, including windows-based workstations.

III.BODY

A. Flaws in Existing Systems

Non-recurrent gene network architectures have been proposed in the past as mechanisms of information integration and storage (Van Der Putten, 2017), associative learning (McGregor et al., 2012; Sorek et al., 2013), and cellular decision making (Bates et al., 2015; Filicheva et al., 2016). However, processing of time-dependent information requires recurrent topologies. The proposed nonconventional computation framework also implies that the integration of information is distributed across the network in large and diffuse structures with well-defined functional roles.

A hierarchical recurrent neural network (HRNN) that identifies time-delayed gene interactions using time-course data (Mina et al, 2017). A customized genetic algorithm (GA) was used to optimize hierarchical connectivity of regulatory genes and a target gene. The proposed design provides a non-fully connected network with the flexibility of using recurrent connections inside the network. These features and the non-linearity of the HRNN facilitate the process of identifying temporal patterns of a GRN. Then further demonstrated the capability of the method in reconstructing GRNs synthetic network for reverse-engineering and modeling approaches (IRMA). HRNN method to be superior in terms of accuracy for nonlinear data sets with higher amounts of noise.

The power of sequencing RNA lies in the fact that the twin aspects of discovery and quantification can be combined in a single high-throughput sequencing assay called RNA-sequencing (RNA-seq) (Ana Conesa, 2017). RNA-seq can be used solo for transcriptome profiling or in combination with other functional genomics methods to enhance the analysis of gene expression. Finally, RNA seq can be coupled with different types of biochemical assay to analyze many other aspects of RNA biology, such as RNA–protein binding, RNA structure, or RNA–RNA interactions.

The execution time of data intensive bioinformatics applications can be significantly improve by parallelization of highly time consuming portion of whole applications and data structure used in bioinformatics application (Binay, 2011). Weighted suffix tree as data structure and parallelize the weighted suffix tree construction on all possible parallel architecture will be taken up in the future.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VII, July 2018- Available at www.ijraset.com

The parallel computing is a type of working out through which many data or the execution connected with processes are finished concurrently as well as scheduling along with source of information permitting so that to optimize efficiency standards within multi-cluster heterogeneous situations is acknowledged for NP-hard problems (Davinderjit, 2011). The flaw has shown the various meta heuristic techniques which has proved their usefulness to find the optimum schedule around large-scale allocated circumstances. It also shows the comparison of Meta heuristic techniques which evaluates the real workload trace as well as shows the advantages and disadvantages when it comes to other well-known approaches outlined inside literature.

B. Description of the Proposed Algorithm:

Aim of this review is to maximize the performance of the parallel computing with respect to transcription and the translation regulatory modules. Parallel computing comprises the high performance computation in these processes. The hybrid method which is combining the high performance computing with neural networks. This novel network inference methods based on a two-step approach to select the top ranked genes from an initial regulatory modules and then the next partial module gives information about protein which is associated with specific gene for the analysis of any cancerous data set.

The Architecture is shown below in the figure 1.

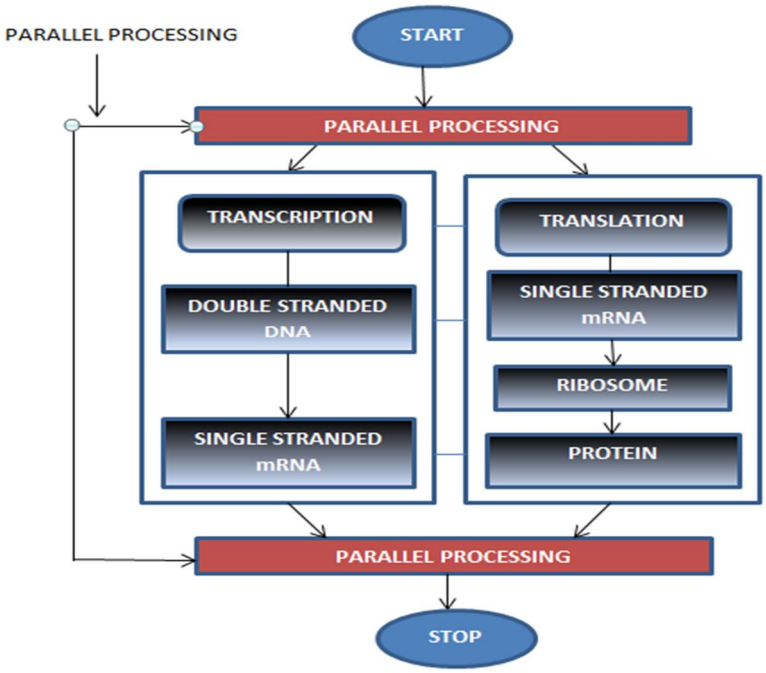


Figure 1: Architecture diagram

C. Pseudo Code

Step 1: Collect the sample Gene expression dataset from the population.
Step 2: Analyzing gene dataset and required gene-gene interaction.
Step 3: If interaction is more
Draw the network and analyze
Else
Select new sample from the population
End
Step 4: Then translate gene to protein
Step 5: Select the proteins with respect to the genes. And requires gene-protein interaction.
Step 6: Draw the network.
Step 7: go to step 2.
Step 8: End.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue VII, July 2018- Available at www.ijraset.com

IV. SIMULATION RESULTS

The simulation studies involve the deterministic small network topology. The proposed work can be implemented with MATLAB. We transcript same size of gene data through step 1 to step 6. Proposed algorithm can be compared between two different samples from the dataset. We considered the simulation time as a network lifetime and network lifetime is a time when no path is available to transcript and translate the data. Simulation time is calculated by the CPUTIME function of MATLAB. Our results shows that the more interaction with genes indicates the gene are sensitive to the cancer and need to regulate it through the network.

Figure 2: Network 1 for the sample 1.

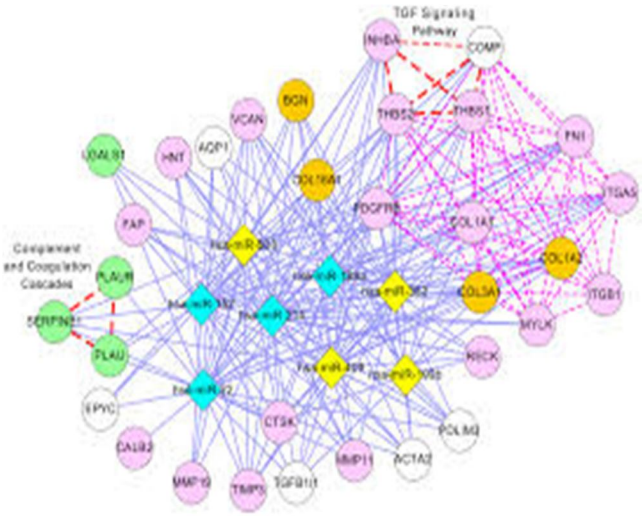


Figure 3: Network 2 for the sample 2.

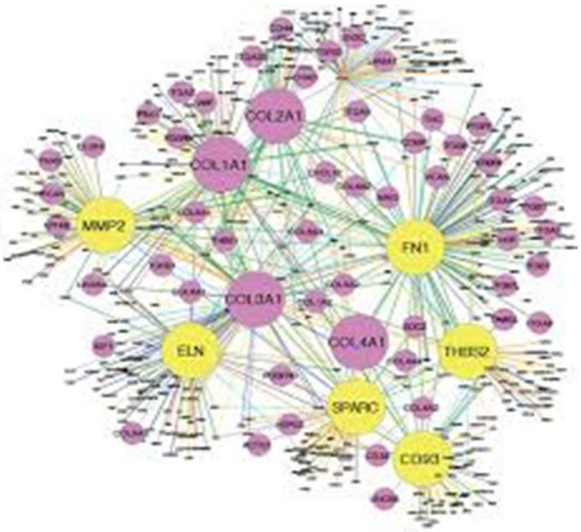


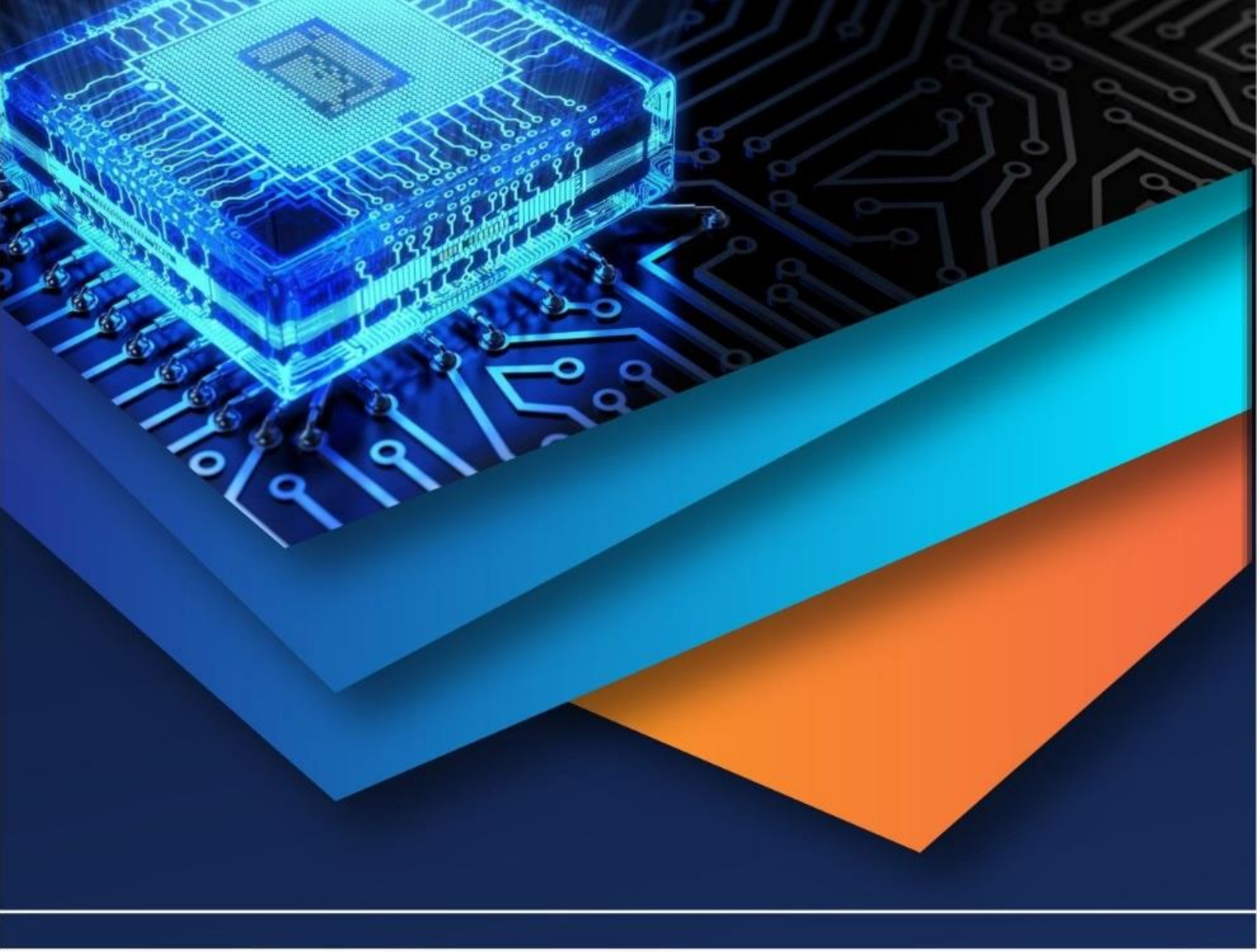
Figure 4: Network 3 for the sample 3.

V. CONCLUSION AND FUTURE WORK

The simulation results showed that the proposed algorithm performs better with the total transmission energy metric than the maximum number of hops metric. The proposed algorithm provides energy efficient path for data transmission and maximizes the lifetime of entire network. As the performance of the proposed algorithm is analyzed between two metrics in future with some modifications in design considerations the performance of the proposed algorithm can be compared with other energy efficient algorithm. We have used very small network of 5 nodes, as number of nodes increases the complexity will increase. We can increase the number of nodes and analyze the performance.

REFERENCES

[1] M. Aldinucci, S. Campa, M. Danelutto, P. Kilpatrick, and M. Torquati. Targeting distributed systems in fast flow. In Euro-Par 2012 Workshops, Proc. of the CoreGridWorkshop on Grids, Clouds and P2P Computing, volume 7640 of LNCS, pages 47–56. Springer, 2013.
[2] M. Aldinucci, S. Campa, M. Danelutto, P. Kilpatrick, and M. Torquati. Design patterns percolating to parallel programming framework implementation. International Journal of Parallel Programming, 42(6):1012–1031, 2014.
[3] I. Merelli, P. Li, and L. Milanesi, "NuChart: An R Package to Study Gene Spatial Neighborhoods with Multi-Omics Annotations," PLoS ONE, vol. 8, no. 9, Sep. 2013
[4] [C. Kwangmin, A. Saple and S. Kim, "Genome Data Type: a Vehicle to Deliver a Genome Comparison Sy stem on the Web", Sixth IEEE International Conference on Data Mining, Hong Kong (China), Page(s):142 – 146, 18-22 Dec- 2006.
[5] Y. Liu, T. Yan and R. Zhan g, "Protein Sequence Analysis Based on Smooth PST", Third International Conference on Bio informatics and Biomedical Engineering, Beijing (China), Page(s): 1-4, 11-13 June 2009
[6] Aldinucci M, Campa S, Danelutto M, et al. Targeting distributed systems in fast flow. In: Euro-Par 2012 Workshops, Proceeding. of the CoreGridWorkshop on Grids, Clouds and P2P Computing, 2013. Ser. LNCS, Springer, 7640:47–56
[7] V. C. Seitan, A. J. Faure, Y. Zhan, R. P. P. McCord, B. R. Lajoie, E. Ing-Simmons, B. Lenhard, L. Giorgetti, E. Heard, A. G. Fisher, P. Flicek, J. Dekker, and M. Merkenschlager, "Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments." Genome research, vol. 23, no. 12, pp. 2066–2077, Dec. 2013
[8] N. Servant, B. R. Lajoie, E. P. Nora, L. Giorgetti, C.-J. Chen, E. Heard, J. Dekker, and E. Barillot, "HiTC: exploration of high-throughput 'C' experiments," Bioinformatics, vol. 28, no. 21, pp. 2843–2844, Nov.2012
[9] V. C. Seitan, A. J. Faure, Y. Zhan, R. P. P. McCord, B. R. Lajoie, E. Ing-Simmons, B. Lenhard, L. Giorgetti, E. Heard, A. G. Fisher, P. Flicek,J. Dekker, and M. Merkenschlager, "Cohesin-based chromatin interactions enable regulated gene expression within preexisting architectural compartments." Genome research, vol. 23, no. 12, pp. 2066–2077, Dec. 2013
[10] N. Servant, B. R. Lajoie, E. P. Nora, L. Giorgetti, C.-J. Chen, E. Heard, J. Dekker, and E. Barillot, "HiTC: exploration of high-throughput 'C' experiments," Bioinformatics, vol. 28, no. 21, pp. 2843–2844, Nov. 2012.
[11] F. Ay, T. Bailey, and W. Noble, "Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts," 2014.
[12] "Intel Threading Building Blocks, project site," 2013, http://threadingbuildingblocks.org.F. Ay, T. Bailey, and W. Noble, "Statistical confidence estimation for Hi-C data reveals regulatory chromatin contacts," 2014
[13] "Intel Threading Building Blocks, project site," 2013, http://threadingbuildingblocks.org.
[14] M. Aldinucci, M. Danelutto, P. Kilpatrick, and M. Torquati, "Fastflow: high-level and efficient streaming on multi-core," in ProgrammingMulti-Core and Many-Core Computing Systems, S.Pllana and F. Xhafa, Eds., Parallel and Distributed Computing, chapter 13, JohnWiley & Sons, 2014
[15] M. J. Chaisson and G. Tesler, "Mapping single molecule sequencing reads using basic local alignment with successive refinement (BLASR): application and theory," BMC Bioinformatics, vol. 13, no. 1, article 238, 2012.
[16] D. Peters, K. Qiu, and P. Liang, "Faster short dna sequence alignment with parallel bwa," in American Institute of Physics Conference Series, I. Kotsireas, R. Melnik, and B.West, Eds., vol. 1368, pp. 131–134, 2011.
[17] R. AlSaad, Q. Malluhi, and M. Abouelhoda, "Efficient parallel implementation of the SHRiMP sequence alignment tool using MapReduce," in Qatar Foundation Annual Research Forum Proceedings, 2012.
[18] M. Aldinucci, C. Calcagno, M. Coppo et al., "On designing multicore aware simulators for systems biology endowed with on-line statistics," BioMed Research International, vol. 2014, Article ID 207041, 14 pages, 2014.
[19] C. Misale, "Accelerating Bowtie2 with a lock-less concurrency approach and memory affinity," in Proceedings of the 22nd Euromicro International Conference on Parallel, Distributed and Network-Based Processing (PDP '14), IEEE, Torino, Italy, February 2014.
[20] P. Klus, S. Lam, D. Lyberg et al., "BarraCUDA—a fast short read sequence aligner using graphics processing units," BMC Research Notes, vol. 5, article 27, 2012
[21] Y. Liu, B. Schmidt, and D. L. Maskell, "CUSHAW: a CUDA compatible short read aligner to large genomes based on the Burrows–Wheeler transform," Bioinformatics, vol. 28, no. 14, Article ID bts276, pp. 1830–1837, 2012.
[22] M. O. Carneiro, C. Russ, M. G. Ross, S. Gabriel, C. Nusbaum, and M. A.DePristo, "Pacific biosciences sequencing technology for genotyping and variation discovery in human data," BMC Genomics, vol. 13, no. 1, p. 375, 2012.
[23] J. Huddleston, S. Ranade, M. Malig et al., "Reconstructing complex regions of genomes using long-read sequencing
[24] V.Kuleshov, D. Xie, R. Chen et al., "Whole-genome haplotyping using long reads and statistical methods," Nature Biotechnology, vol. 32, no. 3, pp. 261–266, 2014.
[25] D. Sharon, H. Tilgner, F. Grubert, and M. Snyder, "A single molecule long-read survey of the human transcriptome," Nature Biotechnology, vol. 31, no. 11, pp. 1009–1014, 2013
[26] E. W. Loomis, J. S. Eid, P. Peluso et al., "Sequencing the unsequenceable: Expanded CGG-repeat alleles of the fragile x gene," Genome Research, vol. 23, no. 1, pp. 121–128, 2013.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY