proceedings

MDPI

Proceedings

Parallelization of ARACNe, an Algorithm for the Reconstruction of Gene Regulatory Networks †

Uxía Casal *, Jorge González-Domínguez and María J. Martín

Grupo de Arquitectura de Computadores, CITIC, Universidade da Coruña, 15071 A Coruña, Spain; jgonzalezd@udc.es (J.G.-D.); mariam@udc.es (M.J.M.)

* Correspondence: uxia.casal.baldomir@udc.es

† Presented at the 2nd XoveTIC Conference, A Coruña, Spain, 5–6 September 2019.

Published: 31 July 2019

Abstract: Gene regulatory networks are graphical representations of molecular regulators that interact with each other and with other substances in the cell to govern the gene expression. There are different computational approaches for the reverse engineering of these networks. Most of them require all gene-gene evaluations using different mathematical methods such as Pearson/Spearman correlation, Mutual Information or topology patterns, among others. The Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNe) is one of the most effective and widely used tools to reconstruct gene regulatory networks. However, the high computational cost of ARACNe prevents its use over large biologic datasets. In this work, we present a hybrid MPI/OpenMP parallel implementation of ARACNe to accelerate its execution on multi-core clusters, obtaining a speedup of 430.46 using as input a dataset with 41,100 genes and 108 samples and 32 nodes (each of them with 24 cores).

Keywords: network reconstruction; ARACNe; High Performance Computing; MPI; OpenMP

1. Introduction

A Gene Regulatory Network (GRN) is a network that has been inferred from gene expression data, explaining how a collection of molecular regulators interact with each other and with other substances in the cell, to govern the gene expression levels and it is crucial for understanding normal cell physiology and complex pathological phenotypes. The importance of this kind of networks can be seen in [1], where the authors explain the concept of GRNs and its relevance in different fields.

The Algorithm for the Reconstruction of Accurate Cellular Networks (ARACNe) [2] represents one of the most used methods in the scientific community to reconstruct GRNs. It is based on information theory and a network pruning process called Data Processing Inequality (DPI) theorem, which is used to infer direct regulatory relationships among transcriptional factors and their genes. ARACNe has been extensively used for real-data biological works such as the ones shown in [3,4]. However, its main drawback is its quadratic complexity with the number of genes, resulting in a high computational cost that prevents its use for large datasets.

The aim of this work is to provide biologists a new version of ARACNe faster than the original one which will be able to infer GRNs of large datasets. The approach to decrease the runtime consists in parallelizing the code using MPI routines and OpenMP directives so that the new implementation increases performance in clusters of multi-core nodes, a type of systems that nowadays are widely used by biologists as each day it is easier to have access to them through a supercomputing center.

2. Parallel ARACNe

The ARACNe framework receives as input a matrix where rows represent genes or variables and columns represent samples. It returns as output an adjacency matrix that indicates gene-gene interactions. In this work, we have developed a hybrid OpenMP/MPI parallel implementation that works with the same input/output formats and guarantees the same results as the original tool but at significantly lower runtime. In our implementation, all MPI processes read the input matrix but the workload to calculate the adjacency matrix is distributed among them. We apply a cyclic distribution by rows. Moreover, our implementation uses a second level of parallelization so that the rows assigned to each process are distributed among the different OpenMP threads following a dynamic scheduling.

Once all processes have finished their work, they use MPI communications to store the whole output matrix in the memory of Process 0 (the only process that writes into the output file). Remark that we also needed to modify the datatype used to represent the output adjacency matrix because that datatype cannot be used for MPI communications.

3. Results and Conclusions

The performance analysis of the new version of ARACNe was carried out in 32 nodes of the Finis Terrae 2 (FT2) supercomputer, a computational system based on Intel Haswell processors (each node has 24 cores) that is installed in the Centro de Supercomputatión de Galicia (CESGA). We have used three datasets, GDS2767 (14,170 genes and 108 samples), GDS6248 (45,281 genes and 51 samples) and GDS5037 (41,000 genes and 108 samples). All of them have been downloaded from the Geo Expression Omnibus (GEO) Dataset Browser available at the National Center for Biotechnology Information (NCBI) website (https://www.ncbi.nlm.nih.gov/sites/GDSbrowser).

Table 1 shows the runtime and speedup (in brackets) for different number of nodes. In all the cases the parallel code was run using one MPI process per node and 24 threads per process, one per core. The tool was run using the default parameters except for the p-value, which was explicitly set to 0.01 to remove non-significant interactions. As can be observed, the performance is high, as the runtime decreases when the number of cores increases. For instance, for the dataset GDS5037 using the original framework the execution time is more than 38 h, while it is reduced to less than 6 minutes using 768 cores (32 processes, each one with 24 threads).

Table 1. Runtime (in seconds) and speedup of the parallel version of ARACNe using up to 32 nodes of the FT2.

Nodes	Cores	GDS2767	GDS6248	GDS5037
Original ARACNe		16552	41452	139077
1	**24**	801.56 (20.65)	1979.09 (20.95)	6563.16 (21.19)
2	**48**	415.84 (39.80)	1033.60 (40.10)	3343.16 (41.60)
4	**96**	223.25 (74.14)	561.39 (73.84)	1763.04 (80.11)
8	**192**	126.49 (130.86)	323.82 (128.01)	931.15 (149.36)
16	**384**	88.34 (187.36)	231.16 (179.33)	608.10 (228.71)
32	**768**	53 (312.31)	144.93 (286.02)	323.09 (430.46)

Author Contributions: conceptualization, J.G.-D. and M.J.M.; methodology, U.C., J.G.-D. and M.J.M.; software, U.C.; validation, U.C.; writing–original draft preparation, U.C.; writing–review and editing, J.G-D. and M.J.M.

Funding: This research was supported by the Ministry of Economy and Competitiveness of Spain and FEDER funds of the EU [Project TIN2016-75845-P (AEI/FEDER, UE)]; the Xunta de Galicia and FEDER funds of the EU (Centro Singular de Investigación de Galicia) [grant number ED431G/01]; and Consolidation Program of Competitive Research [grant number ED431C 2017/04].

Acknowledgments: We gratefully thank Galicia Supercomputing Center for providing access to the Finis Terrae 2 supercomputer.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Emmert-Streib, F.; Dehmer, M.; Haibe-Kains, B. Gene regulatory networks and their applications: understanding biological and medical problems in terms of networks. *Front. Cell Dev. Biol.* **2014**, *2*, 38.
2. Margolin, A.A.; Nemenman, I. ; Basso, K.; Wiggins, C.; Stoolovitzky, G.; Dalla Favera, R.; Califano, A. ARACNE: An Algorithm for the Reconstruction of Gene Regulatory Networks in a Mammalian Cellular Context. *BMC Bioinform.* **2006**, *7*, S1.
3. Mao, X.; Xue, X. ; Wng, L.; Wang , L.; Li, L.; Zhang, X. Hypoxia Regulated Gene Network in Glioblastoma Has Special Algebraic Topology Structures and Revealed Communications Involving Warburg Effect and Immune Regulation. *Cell. Mol. Neurobiol.* **2019**, 1–22, doi:10.1007/s10571-019-00704-5.
4. Clark, J.E.; Ng, W.-F.; Rushton, S.; Watson, S.; Newton, J.L. Network structure underpinning (dys)homeostasis in chronic fatigue syndrome; Preliminary findings. *Cancer Discov.* **2019**, *3*.

© 2019 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (http://creativecommons.org/licenses/by/4.0/).