Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

journal homepage: www.elsevier.com/locate/aiim

A hybrid framework for reverse engineering of robust Gene Regulatory Networks

Mina Jafari, Behnam Ghavami*, Vahid Sattari

Department of Computer Engineering, Shahid Bahonar University of Kerman, Kerman, Iran

ARTICLE INFO

Article history:
Received 12 November 2016
Received in revised form 6 March 2017
Accepted 8 May 2017

Keywords:
Genomics
Gene Regulatory Network (GRN)
Feature selection
Information Gain (IG)
ReliefF
Pearson Correlation Coefficient (PCC)
Genetic Algorithm (GA)

ABSTRACT

The inference of Gene Regulatory Networks (GRNs) using gene expression data in order to detect the basic cellular processes is a key issue in biological systems. Inferring GRN correctly requires inferring predictor set accurately. In this paper, a fast and accurate predictor set inference framework which linearly combines some inference methods is proposed. The purpose of the combination of various methods is to increase the accuracy of inferred GRN. The proposed framework offers a linear weighted combination of Pearson Correlation Coefficient (PCC) and two different feature selection approaches, namely: Information Gain (IG) and ReliefF. In order to set the appropriate weights, Genetic Algorithm (GA) is used. Similarity measure is considered as fitness function to guide GA. At the end, based on the obtained weights, the best predictor set of GRN using three aforementioned inference methods is selected and the network topology is formed. Due to the huge volume of gene expression data, GRN inference algorithms should infer GRN at a reasonable runtime. Hence, a novel criterion is provided to evaluate GRNs based on runtime and accuracy. The simulation results using biological data indicate that the proposed framework is fast and more reliable compared to other recent methods [1–7].

1. Introduction

Each protein has its own unique amino acid sequence that is determined by the nucleotide sequence of the gene encoding this protein. Furthermore, proteins can act as transcription factors that regulate the expression of other genes; therefore, a living organism can be considered as a complex and interconnected network of molecules connected by biochemical reactions [8]. This regulatory mechanism forms a complex system of sending and receiving signals, which can be inquired to recognize the cell control mechanisms and the relationships among various biological entities. Understanding the relationships among genes and gene regulation through signal transmission is a crucial goal in biological systems [2,9,10].

The development of technologies to extract gene expression data like microarray DNA [11], SAGE [12] and also RNA Sequencing [13] have allowed to rapidly measure tens of thousands of gene expressions at once. By increasing the availability of these data, the researchers have focused on the interaction among genes and their functionality. Interactions among genes form a complex and interconnected network called Gene Regulatory Network (GRN). GRNs are essential to uncover details about key principles of biological systems and can be used to explain how cells control the expression of genes. Generally, GRN is a worthy approach to show the cell behavior through modeling relationships among genes and the effects of a set of genes on the another set of genes. The correct construction of GRN has various usages, some of the most remarkable of which are examining the behavior of a set of genes, identifying the occurrence of biological processes as well as faults in the processes (disease) and last but not least, prescribing the most effective drug treatment (removing faults). In a GRN, a 'predictor' regulates a target gene; moreover, a group of predictors that regulates a target gene is called 'predictor subset' and whole set of predictor subsets in a GRN is called 'predictor set' [10]. Fig. 1 shows the predictor subset and predictor set in a GRN.

The inference of GRN using gene expression data, which is also known as reverse engineering, is a crucial and difficult task [14]. Although many methods have been developed to infer GRNs from gene expression data [15], the major challenge in this research area is to infer GRNs based on purely observed gene expression data. Generally, the inference process without fault is impossible due to the lack of enough biological information. Some of the factors that make the inference process hard and challenging task are: lack of precision to measure the gene expressions that leads to create noisy

* Corresponding author.
E-mail addresses: m.jafari@eng.uk.ac.ir (M. Jafari), ghavami@uk.ac.ir (B. Ghavami), vsatari@uk.ac.ir (V. Sattari).

http://dx.doi.org/10.1016/j.artmed.2017.05.004
0933-3657/© 2017 Elsevier B.V. All rights reserved.

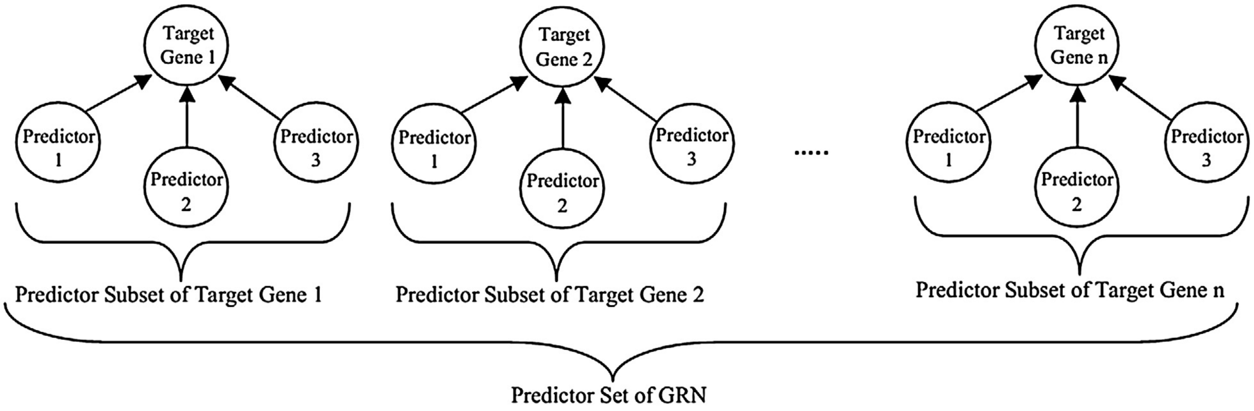

Fig. 1. The predictor subset of each target gene and predictor set of GRN.

data, the huge volume of genes and the small sample size [16]; thus, still there is a need of efficient methods to infer reliable GRNs. There are several recent initiatives to overcome data limitations by incorporating other biological information to discover the dependency of genes [9,17].

The purpose of this paper is to provide a fast and efficient framework to infer the predictor set in the GRN, taking into account the limitations of gene expression data. The process of predictor set inference consists of realizing the dependence of target genes and their potential predictors. In this work, a novel framework is proposed to infer predictor set in GRN which linearly combines some inference methods. As a matter of fact, the proposed framework offers a linear weighted combination of Pearson Correlation Coefficients (PCC), Information Gain (IG) and ReliefF scores. To fine-tune the weights, Genetic Algorithm (GA) is used and moreover similarity criterion as the fitness function is utilized to guide the GA. The proposed framework is investigated using biological data based on weights inferred from GA. Besides, a novel evaluation criterion is proposed based on runtime and accuracy of GRN inference methods. Experimental results on biological data reveal that despite the large number of genes and the small size of samples, the proposed framework can infer predictor set of the GRN in a robust manner.

The main contribution of this paper can be stated as follows:

- Utilizing an ensemble filter feature selection method (the combination of IG and ReliefF) combining with PCC instead of using only one feature selection method results in higher accuracy in terms of inferring the predictor set.
- In order to infer the weight each of the aforementioned methods executes the GA on a small subset of the data. This results in a higher accuracy avoiding over fitting when analyzing big datasets and in reduced runtime.
- Proposing a novel measure to consider time and accuracy when the algorithm is evaluated because time is a remarkable issue especially in big data.

To put it in a nutshell, the aforementioned points bring in a better accuracy and lower runtime regarding to other similar methods.

The rest of this paper is organized as follows. Section 2 introduces the related works for GRN inference. The proposed method is introduced in Section 3. Section 4 performs a preliminary comparison between the proposed framework and a recent similar method. Finally, Section 5 draws the conclusions and presents some future works.

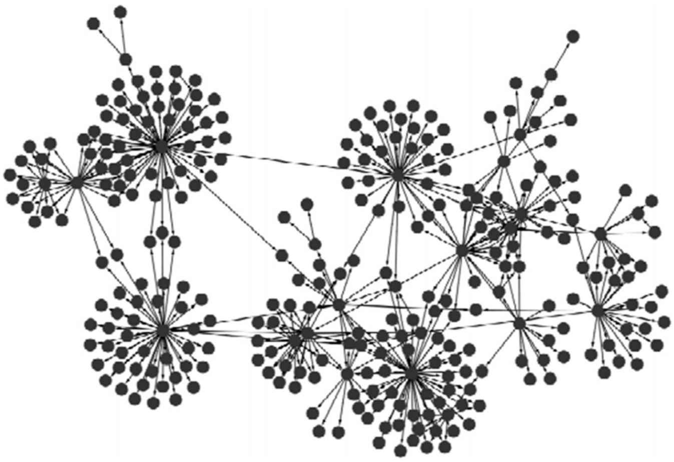

Fig. 2. A GRN from the *E. coli* with 300 genes and 439 regulatory relations [18].

2. Gene Regulatory Network inference

Genes interact indirectly with each other and with other substances in the cell. As a matter of fact, there is a dense set of associations among biological entities that need to be modeled and represented to increase the knowledge in molecular biology [18]. As previously mentioned, with the advent of high-throughput technologies, it has become relatively possible to measure gene expression very quickly [19]. Over the last decade, many researchers have turned to such methodologies with the aim of identifying the interactions and connections among genes and their regulatory functions [15]. The relationships among genes can be represented as a network; thus, it is needed to build the GRN. The best representation of such complex networks comes from graph theory. A GRN consists of a number of nodes and edges between the nodes. Every node represents a gene and the edges represent relations between genes [18]. Fig. 2 represents a sample of GRN. It is important to represent the GRNs in an appropriate way and provide an efficient algorithm to infer GRNs from gene expression data. Several approaches are introduced in literature for GRN inference using gene expression data where each of them has their own strength and weakness. In the rest of this section, we will review the related works to infer GRNs.

There are many heuristic methods for inferring GRNs from gene expression data, Mendoza et al. [20] have presented a method using GA to infer GRNs. The Tsallis entropy criterion is used to assess chromosomes. Jimenez et al. [7,21] have used a GA for each target gene, separately. Consequently, for each target gene an initial population is considered, the population is not generated randomly; in fact,