Journal of Computer and System Sciences 80 (2014) 118–136

Contents lists available at SciVerse ScienceDirect

Journal of Computer and System Sciences

www.elsevier.com/locate/jcss

Discovering gene association networks by multi-objective evolutionary quantitative association rules

M. Martínez-Ballesteros *, I.A. Nepomuceno-Chamorro, J.C. Riquelme

Department of Computer Science, University of Seville, Spain

ARTICLE INFO

Article history:
Received 16 July 2012
Received in revised form 26 November 2012
Accepted 14 March 2013
Available online 21 March 2013

Keywords:
Data mining
Multi-objective evolutionary algorithms
Quantitative association rules
Gene networks
Microarray analysis

ABSTRACT

In the last decade, the interest in microarray technology has exponentially increased due to its ability to monitor the expression of thousands of genes simultaneously. The reconstruction of gene association networks from gene expression profiles is a relevant task and several statistical techniques have been proposed to build them. The problem lies in the process to discover which genes are more relevant and to identify the direct regulatory relationships among them. We developed a multi-objective evolutionary algorithm for mining quantitative association rules to deal with this problem. We applied our methodology named GarNet to a well-known microarray data of yeast cell cycle. The performance analysis of GarNet was organized in three steps similarly to the study performed by Gallo et al. GarNet outperformed the benchmark methods in most cases in terms of quality metrics of the networks, such as accuracy and precision, which were measured using YeastNet database as true network. Furthermore, the results were consistent with previous biological knowledge.

1. Introduction

Since late 1990s, the interest in microarray technology has exponentially increased due to its ability to monitor the expression of thousands of genes simultaneously. Microarray technology has revolutionized the biological research because it allows to study thousand of genes or even whole genomes [1].

As molecular biology is rapidly evolving into a quantitative science, it increasingly relies on computational algorithms to make sense of high-throughput data. One of the main goals in Microarray analysis is the reconstruction of gene regulatory processes and a key task is the inference of regulatory interactions among genes from gene expression data [2]. Our aim is to infer the relationships between genes from an organism in a particular biological process. This relationships can be modeled in several levels of abstraction, these levels range from the detailed gene regulatory processes (where a chain of intracellular reaction activates a regulatory molecule, transcription factors until a protein is synthesized) to the high models of abstraction named gene association networks. In the reconstruction of gene regulatory processes, building gene association networks has been proven to provide useful insights for such task, the reconstruction of gene regulatory processes. A gene association network can be defined as a graph in which nodes represent genes and edges represent the influence between them. Our goal in this work is the inference of gene association networks from Microarray datasets.

There are several statistical methods to infer gene association networks from Microarray data. A microarray dataset is a bidimensional data structure where conditions are experiments or sources and the columns are gene expression values. In our problem the conditions will be the instances and the gene expression values will be the attributes or features. These methods range from relatively straightforward correlation-based methods to more sophisticated methods based on

* Corresponding author.
E-mail addresses: mariamartinez@us.es (M. Martínez-Ballesteros), inepomuceno@us.es (I.A. Nepomuceno-Chamorro), riquelme@us.es (J.C. Riquelme).

0022-0000/$ – see front matter
http://dx.doi.org/10.1016/j.jcss.2013.03.010

the concept of conditional independence. In general, these methods based on pairwise similarity measures are very useful to determine whether two genes have a strong global similarity under all conditions in the dataset. However, there could be strong local similarities over a subset of conditions, which could not be detected by them [3]. In this context, the discovery of Association Rules (AR), and particularly of Quantitative Association Rules (QAR), is a popular methodology that allows the discovery of significant and apparently hidden relations among attributes in a subspace of the instances from the dataset. Therefore, we developed a multi-objective evolutionary algorithm for mining QAR to favor the detection of localized similarities over a more global similarity. Furthermore, as it can be observed in the review of the state-of-the-art [4], methods based on the discovery of QAR have not been used to infer gene associations from microarray data. However, qualitative AR have been used to infer gene association networks but this approach needs a discretization step that our proposal avoids.

Our proposal, henceforth named GarNet (Gene–gene associations from Association Rules for inferring gene NETworks), is based on the well-known multi-objective evolutionary approach NSGA-II [5] to discover QAR with adaptive intervals without performing a previous discretization. NSGA-II algorithm has been selected instead of SPEA-II [6] algorithm because it performs better than SPEA-II due to the powerful crowding operator that keeps diversity in the population and generates a more uniform Pareto front. Furthermore, NSGA-II is considered as the paradigm within the MOEA research community [7]. GarNet carries out an inference process based on an iterative rule learning to extract gene–gene associations and builds gene networks by the intersection of the gene–gene associations retrieved from the QAR found in several input microarray datasets. To summarize, our proposal presents mainly two improvements: it favors the detection of localized similarities and avoids the discretization step of other approaches based on the discovery of AR [8].

In this work, we focus on the analysis of a set of genes that encode proteins important for cell-cycle regulation. We applied GarNet to a well-known microarray data of yeast cell cycle and we compared our approach against several benchmark methods focused on the same biological problem. For performance analysis we applied as benchmark methods a decision-tree-based method [9], a regression-tree based method [3], a probabilistic graphical model [10] and combinatorial optimization algorithm [11,12]. The performance analysis was organized in three steps similarly to the study performed by Gallo et al. [12]. GarNet outperformed the benchmark methods in most cases in terms of quality metrics of the networks, such as accuracy and precision, which were measured using YeastNet database as a true network. Furthermore, the results were consistent with previous biological knowledge.

The remainder of the paper is organized as follows. In Section 2, a summary of the benchmark methods to infer gene networks and to extract AR is presented. In Section 3, a detailed explanation of the methodology and the algorithm are presented. Section 4 reports the performance analysis, parameter settings and comparison analysis together with the biological relevance of the experiments. Finally, Section 5 summarizes the most relevant conclusions and future works.

2. Related work

The related work is divided into two parts: the first one describes the methods to infer gene networks from microarray data in the literature and the second one describes data mining techniques to build AR.

2.1. Inferring gene networks: a review

There are several methods to infer gene–gene association networks from gene expression data. These methods range from rather straightforward correlation-based methods to more sophisticated models, such as Bayesian network models.

One of the first approaches to the problem was clustering algorithms [13,14]. These approaches are based on a simple assumption, which is still used in functional genomic, called the guilt-by association heuristic. This assumption suppose that co-expression means co-regulation, i.e. if two genes show similar expression profiles, they are supposed to follow the same regulatory programme.

In order to formalize the idea of similar expression behavior, several statistical measures have been proposed in the literature. In correlation-based methods, gene–gene associations are built using correlation as a pairwise similarity measure between gene expression profiles over all the conditions in the dataset. In standard correlation-based methods, the Pearson or Spellman's coefficient has been used to identify gene–gene associations [15]. In this kind of methods, if the correlation between gene pairs is higher than a threshold value (usually 0.95), then it is assumed that these gene pairs interact directly in a relevant biological process or in a signaling pathway [16,17]. As shown in [18], the results provided by these methods are a framework for assigning biological functions to group of genes. In the literature, gene co-expression networks are also known as gene relevance, gene association or gene interaction networks. Different versions of the standard correlation-based method exist, such as one by Obayashi and Kinoshita in [19] that instead of correlation values uses correlation ranks.

Correlation-based methods are very useful to determine strong global similarity between two genes over all conditions in the dataset. This is a relevant constraint due to there might exist a strong local similarity over a subset of conditions which could not be detected with correlation measure. This constraint is taken into account in the model tree-based method proposed by Nepomuceno-Chamorro et al. [3], where they used regression trees as a way to detect linear dependencies localized over a subset of conditions. Similar to this approach, another rule-based method is presented in [9] in which the authors used decision trees as a way to extract dependencies. Inspired in these two techniques, the model tree-based method and the rule-based method, this work proposed a method to favor the detection of localized similarities over a more