Artificial Intelligence in Medicine 70 (2016) 41–61

Contents lists available at ScienceDirect

Artificial Intelligence in Medicine

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

Effective gene expression data generation framework based on multi-model approach

Utku Sirin [a], Utku Erdogdu [b], Faruk Polat [b,*], Mehmet Tan [c], Reda Alhajj [d]

[a] School of Computer and Communication Sciences, École Polytechnique Fédérale de Lausanne, Route Cantonale, 1015 Lausanne, Switzerland
[b] Department of Computer Engineering, Middle East Technical University, 06800 Ankara, Turkey
[c] Department of Computer Engineering, TOBB University of Economics and Technology, Sogutozu Cd., 06560 Ankara, Turkey
[d] Department of Computer Science, University of Calgary, 2500 University Dr. NW, Calgary, Alberta, Canada T2N 1N4

ARTICLE INFO

Article history:
Received 20 October 2015
Accepted 27 May 2016

Keywords:
Multi-model approach
Probabilistic Boolean networks
Ordinary differential equations
Genetic algorithm
Hierarchical Markov models
Gene expression data generation
Gene regulation network modeling

ABSTRACT

Objective: Overcome the lack of enough samples in gene expression data sets having thousands of genes but a small number of samples challenging the computational methods using them.
Methods and material: This paper introduces a multi-model artificial gene expression data generation framework where different gene regulatory network (GRN) models contribute to the final set of samples based on the characteristics of their underlying paradigms. In the first stage, we build different GRN models, and sample data from each of them separately. Then, we pool the generated samples into a rich set of gene expression samples, and finally try to select the best of the generated samples based on a multi-objective selection method measuring the quality of the generated samples from three different aspects such as compatibility, diversity and coverage. We use four alternative GRN models, namely, ordinary differential equations, probabilistic Boolean networks, multi-objective genetic algorithm and hierarchical Markov model.
Results: We conducted a comprehensive set of experiments based on both real-life biological and synthetic gene expression data sets. We show that our multi-objective sample selection mechanism effectively combines samples from different models having up to 95% compatibility, 10% diversity and 50% coverage. We show that the samples generated by our framework has up to 1.5*x* higher compatibility, 2*x* higher diversity and 2*x* higher coverage than the samples generated by the individual models that the multi-model framework uses. Moreover, the results show that the GRNs inferred from the samples generated by our framework can have 2.4*x* higher precision, 12*x* higher recall, and 5.4*x* higher *f*-measure values than the GRNs inferred from the original gene expression samples.
Conclusions: Therefore, we show that, we can significantly improve the quality of generated gene expression samples by integrating different computational models into one unified framework without dealing with complex internal details of each individual model. Moreover, the rich set of artificial gene expression samples is able to capture some biological relations that can even not be captured by the original gene expression data set.

1. Introduction

Gene expression data has vital importance for genome research. In fact, most of the research in this field is driven by the gene expression data. Gene expression data sets are hard to construct, and it is generally difficult to come up with data sets that have large number of samples. Furthermore, for small data sets, the number of genes is very high compared to the number of samples. This decreases the confidence levels of the computational methods using the gene expression data for various purposes, e.g., modeling a gene regulatory network (GRN), or classifying diseased samples [1–5]. Although a data set reflects an accurate picture of the underlying genome mechanics, some application domains such as health informatics and molecular biology suffer from the small number of samples making the knowledge discovery task difficult and challenging. For example, in the cancer biomarkers prediction problem, the predictive gene sets constructed by different research groups contain quite small number of samples compared to what is needed to construct reliable predictive gene lists [6].

* Corresponding author.
E-mail addresses: utku.sirin@epfl.ch (U. Sirin), polat@ceng.metu.edu.tr (F. Polat), alhajj@ucalgary.ca (R. Alhajj).

http://dx.doi.org/10.1016/j.artmed.2016.05.003
0933-3657/

There are both direct and indirect approaches proposed for mitigating the small number of samples problem of gene expression data sets. The studies in [7–10] are examples of indirect approaches. They obtain a formula for the number of required samples based on various experimental parameters, and apply the most appropriate experiment to enlarge the gene expression data set. The study described in [7] proposes repeating the microarray experiments to increase the number of samples. The studies described in [11–13], on the other hand, are examples of direct approaches. They combine different generative models to enrich the gene expression data sets. However, these studies have important deficiencies. Firstly, their sample selection mechanism is not multi-objective but single-objective. Secondly, they determine the quality of the generated samples based on a linear combination of the metrics that cannot be transformed to each other. Third, their compatibility and diversity metric formulations are based on the same idea of Euclidean distance. Forth, in their experimental evaluations, they do not refer to dividing the data set into separate training and test sets, but directly use the same data set both for training and test purposes. Lastly, their experimental evaluation is restricted to the defined metric results.

Therefore, in this paper, we thoroughly extend the studies described in [12,13]. We build a comprehensive multi-model data generation framework that integrates different gene regulation models. Our goal is to generate high-quality gene expression samples that contain rich inner dynamics of different gene regulation models. Our rationale is that, whatever the model is, a single gene regulation model converges to its own system dynamics. By integrating different models, however, we can enrich the aggregate system dynamics, and use the most successful combination of different regulation models depending on the specific experimental setting and the data sets used. To do that, we combine four gene regulation models, namely ordinary differential equations (ODEs), probabilistic Boolean networks (PBNs), multi-objective genetic algorithm (GA) and hierarchical Markov model (HIMM). We firstly build the four gene regulation models. Then, we generate samples from each of the models, and pool the generated samples. This pool constitutes a rich source of gene expression samples exhibiting various system dynamics characteristics of different gene regulation models. Having pooled the generated gene expression samples, we employ a multi-objective sample selection mechanism to select the highest-quality, generated samples. The mechanism enables evaluating the generated samples from different aspects such as how much the generated samples look like to the original gene expression samples, i.e., compatibility, how much the generated samples are different than the original gene expression samples, i.e., diversity, and how much the generated samples cover the sample space, i.e., coverage. Each generated sample is evaluated from these three aspects, and sorted multi-objectively. At the final stage, the samples having the best scores are outputted as the final set of generated gene expression samples. The results show that our multi-model framework produces high quality gene expression samples in terms of their compatibility, diversity and coverage. Moreover, the produced samples contain valuable diverse inner dynamics of gene regulation such that the GRNs inferred from the generated artificial gene expression data are usually better than the GRNs inferred from the original gene expression data. Therefore, the artificial data sets produced by our multi-model framework are able to capture new biological relations that can even not be captured by real gene expression data sets. Furthermore, when we compare the generated data sets by a single computational model, and by our multi-model framework, we clearly see that integrating different computational models are much more effective than using a single computational model in terms of the quality of the generated samples. We also propose the number of required samples to train our multi-model framework in hoping to give a bound for the cost of the real life gene expression data generation experiments.

Table 1
List of variables commonly used throughout the paper, and their explanations.

Variable	Explanation
M	Number of generated samples
m	Number of original gene expression samples
n	Number of genes
i, j, k	Index variables

The rest of this paper is organized as follows. Section 2 describes the multi-model artificial gene expression data generation approach, defines the compatibility, diversity and coverage, and explains the multi-objective selection mechanism. Section 3 describes the formulation of the four generative models we use in our framework. Section 4 describes the data sets we used. Section 5 contains the experimental results justifying the effectiveness of our proposed sample generation method. Lastly, Section 6 concludes our work.

We provide the list of commonly used variables throughout the text in Table 1 for the convenience of the reader.

2. Multi-model approach

Multi-model gene regulation stands for integrating different gene regulation models into one unified framework. There are various gene regulation models such as ODE and PBN [14,15]. However, every model has its own intrinsic limitations on simulating the gene regulation dynamics. Therefore, under different experimental setting and data set, different models can behave very differently. For one type of data set, for example, ODE may simulate system dynamics more successfully than PBN, while, for another type of data set, PBN may simulate system dynamics better than ODE. In their study, Hurley et al. [16] and Marbach et al. [17] propose inferring GRNs (genetic regulatory networks) based on several GRN inference algorithms, such as ARACNE, BANJO, MIKANA and SiGN-BN [18–21]. They argue that comparing and combining different inferred networks would capture relationships more successfully. They refer this phenomenon as “wisdom of crowds” [17]. In our work, on the other hand, we propose combining different gene regulation models, not in terms of the inferred networks but the generated samples. We construct alternative gene regulation models, and use all of them concurrently to produce high quality gene expression data. Hence, our work is in accordance with the studies [16,17] in the sense of combining, not different GRN inference algorithms, but different gene expression data generation algorithms to improve available gene expression data sets from which hopefully more accurate networks are able to be inferred.

The block diagram of our proposed framework is shown in Fig. 1. For generating M samples, our multi-model framework generates M samples from each of the four generative models, ODEs, PBNs, multi-objective GA and HIMM. Then, it evaluates each generated sample based on their compatibility, diversity and coverage, and uses a multi-objective sample selection mechanism select the best M samples from the generated $4M$ samples combined from different generative models. Note that, as shown in Fig. 1, except the ODE model, all of the three models work on discrete domain. That is, they both use and produce discretized gene expression samples. Hence, we use binary discretized gene expression samples to generate samples, and convert back the generated binary discretized gene expression samples into their continuous values right before feeding them into the sample selection mechanism.

The discretization algorithm works as follows. It firstly finds the average, maximum and minimum gene expression value of