Computers in Biology and Medicine 48 (2014) 55–65

Contents lists available at ScienceDirect

Computers in Biology and Medicine

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

A review on the computational approaches for gene regulatory network construction

Lian En Chai, Swee Kuan Loh, Swee Thing Low, Mohd Saberi Mohamad *, Safaai Deris, Zalmiyah Zakaria

Artificial Intelligence and Bioinformatics Research Group, Faculty of Computing, Universiti Teknologi Malaysia, Skudai, 81310 Johor, Malaysia

ARTICLE INFO

Article history:
Received 16 September 2013
Accepted 17 February 2014

Keywords:
Gene regulatory network
Computational approaches
Gene expression data
Bayesian network
Dynamic Bayesian network
Boolean network
Probabilistic Boolean network
Ordinary differential equation
Neural network

ABSTRACT

Many biological research areas such as drug design require gene regulatory networks to provide clear insight and understanding of the cellular process in living cells. This is because interactions among the genes and their products play an important role in many molecular processes. A gene regulatory network can act as a blueprint for the researchers to observe the relationships among genes. Due to its importance, several computational approaches have been proposed to infer gene regulatory networks from gene expression data. In this review, six inference approaches are discussed: Boolean network, probabilistic Boolean network, ordinary differential equation, neural network, Bayesian network, and dynamic Bayesian network. These approaches are discussed in terms of introduction, methodology and recent applications of these approaches in gene regulatory network construction. These approaches are also compared in the discussion section. Furthermore, the strengths and weaknesses of these computational approaches are described.

Contents

* Corresponding author.
E-mail address: saberi@utm.my (M.S. Mohamad).

http://dx.doi.org/10.1016/j.compbiomed.2014.02.011
0010-4825

1. Introduction

Human DNA contains approximately twenty per cent of protein-encoding genes, with the other eighty per cent encodes retrotransposons, transposons and pseudogenes. Retrotransposons copy themselves to RNA and then back to DNA by reverse transcription [1]. Transposons are called "jumping genes" as they can change position within the genome [2]. Both transposons and retrotranposons create mutations and alter or amplify the genome size. DNA also encodes many types of small RNA. In order to understand the nature of cellular function, it is necessary to study the behaviour of genes from a holistic perspective instead of individual manner. The expression and activity of genes are independent from each other.

Many crucial molecular processes and cellular pathways are based on the interactions among genes. The genes in living cells regulate each other to control the production of gene products. The regulation network of gene expression in an organism is very complex [3]. Computational approaches have been developed for the construction of genetic interaction models. Gene X is said to be regulated by gene Y if a change in the expression of gene Y also induces change in the expression of gene X. This regulation can be either up-regulation or down-regulation. Up-regulation results in activation and down-regulation results in inhibition. One of the most famous paradigms in gene regulation is the *lac* operon in prokaryotes.

Gene regulatory networks describe control at the gene expression level and could be inferred from microRNAs (miRNAs), regulatory motifs, expression profiles and interactions between regulatory targets [4]. Gene regulation has become important as information of molecular regulatory interactions has become increasingly available. The importance of gene regulatory networks is evident for all biological species and system [5] as they play important role in maintaining the biological functions of living organism [6]. The inference of regulators is the core factor in interpreting the actual regulatory conditions in gene regulatory networks [7]. This provides a clearer blueprint on the relationship between target genes and regulator genes. Each gene expression level in the genome is modified by a central dogma control process. Central dogma is the process that comprises transcription and translation. The expression of the gene products affects the expression rate of other genes, thus giving rise to a complicated network structure. Genes and proteins do not act independently. Some proteins function independently; however, others only react actively in a complex form [8].

Gene expression data is crucial for gene regulatory network construction. It serves as the raw input data to provide information on gene expression. According to the information provided by gene expression data, gene–gene interactions can be inferred to construct a gene regulatory network. Gene expression data allows biologists to observe the expression level of genes on a large scale. This includes DNA microarray, RNAseq [9] and SAGE (Serial analysis of gene expression) [10]. There are two types of gene expression data used for gene regulatory network construction: time series and perturbation experiments. Time series expression data enables biologists to investigate the temporal pattern in biological networks. Perturbed expression data provides the information on interactions directions. The perturbation treatments are divided into two classes: gene manipulation (i.e. gene deletion, over-expression, temperature sensitive or kinetic mutation) [11] or external treatments (i.e. environmental stresses) [12].

Gene regulatory network construction is a major focus in biological research since it can provide important information that is useful for drug design or medical-related fields. The resultant network or module would serve as a working model for researchers to form novel hypotheses and assist in experimental design [13]. For example, Madhamshettiwar et al. [14] applied gene regulatory network construction to ovarian cancer, generated a series of testable hypotheses, and finally discovered a list of potential drug targets. Nowadays, researchers use not only gene expression data but also integrate prior biological knowledge such as DNA sequence data or annotation data [15–17]. The confluence of expression data and prior biological information can improve the accuracy and the quality of the gene regulatory network [13]. In the past, gene regulatory networks were constructed by using the clustering approach [18]. However, this approach failed to identify significant transcriptional network interactions. Hence, many computational approaches have been developed for constructing gene regulatory networks more effectively. Some review papers on gene regulatory network construction exist [19,20], but this review focuses on reviewing some of the commonly used computational approaches for gene regulatory network construction, which include Bayesian network, dynamic Bayesian network, Boolean network, probabilistic Boolean network, ordinary differential equation and neural network. Also, this review discusses recent applications, development, and comparison of the aforementioned approaches, as well as future challenges. Readers are encouraged to consult review on recent attempts by biologists to apply predicted gene regulatory networks to bench experiments for new findings [13], as well as a review on the performance of recent gene network inference software [21].

2. Bayesian network

Bayesian network (BN) models the qualitative properties of gene regulatory networks by using a combination of two mathematical areas: probability and graph theory [22]. Bayesian network allows visualisation of the independent relationships between genes through a directed acyclic graph (DAG).

2.1. Bayesian network methodology

According to Kaderali and Radde [8], Bayesian network is a directed and acyclic graph where $G=(X, A)$ with a set of local probability distributions P. X represents the nodes, $\{x_1, \ldots, x_n\}$, which represent the gene variables. However, A indicates the directed edge that corresponds to probabilistic dependence interactions among genes. A DAG is a network graph with directed edges and no cycles. In a DAG, the edge pointing from one node to another represents the regulatory relationship between parent node and child node. For instance, if X_1 is the parent of X_2 and X_2 is the parent of X_3, the assumptions are that X_1 is the ancestor of X_3 and X_3 is the descendant of X_1. The interaction between two nodes linked by an undirected edge is described by the neighbour term or adjacent term [23]. The joint distribution of a set of variables is showed below:

$$P\{x_1, x_2, \ldots x_n\} = \prod_{i=1}^{n} P\{x_i | \text{ parents } (x_i)\} \quad (1)$$

where x_i is a given gene node. n is the total number of genes involved. The parents (x_i) is the parent gene that regulates gene x_i. $P(x_i|$ parents $(x_i))$ for node i is denoted as Conditional Probability