East Asian Journal on Applied Mathematics
doi: 10.4208/eajam.030511.060911a

Vol. **2**, No. 1, pp. 1-18
February 2012

On Construction of Sparse Probabilistic Boolean Networks

Xi Chen*, Hao Jiang and Wai-Ki Ching

Advanced Modeling and Applied Computing Laboratory, Department of Mathematics, The University of Hong Kong, Pokfulam Road, Hong Kong.

Received 3 May 2011; Accepted (in revised version) 6 September 2011

Available online 10 February 2012

Abstract. In this paper we envisage building Probabilistic Boolean Networks (PBNs) from a prescribed stationary distribution. This is an inverse problem of huge size that can be subdivided into two parts — viz. (i) construction of a transition probability matrix from a given stationary distribution (Problem ST), and (ii) construction of a PBN from a given transition probability matrix (Problem TP). A generalized entropy approach has been proposed for Problem ST and a maximum entropy rate approach for Problem TP respectively. Here we propose to improve both methods, by considering a new objective function based on the entropy rate with an additional term of L_α-norm that can help in getting a sparse solution. A sparse solution is useful in identifying the major component Boolean networks (BNs) from the constructed PBN. These major BNs can simplify the identification of the network structure and the design of control policy, and neglecting non-major BNs does not change the dynamics of the constructed PBN to a large extent. Numerical experiments indicate that our new objective function is effective in finding a better sparse solution.

AMS subject classifications: 65C20, 92B05

Key words: Probabilistic Boolean Networks, entropy, stationary distribution, sparsity, transition probability matrix.

1. Introduction

Coordinated interactions and regulations among genes and gene products form so-called gene regulatory networks, an important research topic in genomic research [3, 16] where inference from gene expression data plays an important role. In recent years, many formalisms have been proposed for modeling gene regulatory networks — including Bayesian networks [20], Boolean Networks (BNs) [18], multivariate Markov chain [7] and regression [31] models, and Probabilistic Boolean Networks (PBNs) [23, 24]. The various mathematical models are reviewed in Refs. [15, 25].

*Corresponding author. *Email addresses:* dlkcissy@hku.hk (Xi Chen), haohao@hkusuc.hku.hk (Hao Jiang), wching@hku.hk (Wai-Ki Ching)

The Boolean Network (BN) model and its Probabilistic Boolean Network (PBN) extension have received considerable attention, as they capture some fundamental characteristics of the gene regulations that occur in gene regulatory networks [28]. Consequently, one can understand a particular gene regulatory network and study the influence of different genes. In a BN model, first introduced by Kauffman [18, 19], each gene is represented as a node and each node can take two possible values (1 and 0). The value of a target node is determined by several input nodes (regulators) via a Boolean function. A BN model is deterministic, and randomness only arises from its initial state. Given this inherent deterministic directionality and also the finite number of possible states, the state transitions allow a BN network to enter a set of states and then cycle among them in a fixed order forever, so the set of states is an attractor. If the attractor contains only one state, it is called a singleton attractor; and if it contains more than one state, it is called an attractor cycle [1, 18, 19]. Since attractors represent stable states in a dynamic system, they can reflect the long term behavior of a BN. In particular, it has been demonstrated that attractors are associated with cellular phenotypes [28].

A BN is not only inherently deterministic but also a closed system and therefore has modeling limitations, but a PBN extension provides a stochastic aspect. A PBN consists of a cluster of BNs with selection probabilities assigned, and each BN can be considered a "context". At any given time instant, gene regulations are governed by one of the component BNs. At the next time instant, the system may switch to another BN with a certain switching probability, when the genes can interact under a different context. Thus a PBN model is more flexible than BN model, and it can be described via a Markov chain [8, 23, 24]. Since a PBN also has a finite number of states, its long term behavior can be characterized by the stationary distribution, providing a possible way to infer the PBN from gene expression data.

Time-independent gene expression data can be obtained from micro-array studies, usually by sampling steady states of the network. Using this data, one can estimate a stationary distribution of the network and hence consider building a PBN. This construction problem involves identifying all the component BNs and their corresponding selection probabilities, such that the long term behavior of the constituting PBN is consistent with the prescribed stationary distribution. There has been some preliminary work based on entropy theory [11, 12, 32], using the entropy rate as the objective function. We recall from information theory that the entropy can measure the amount of information missing before reception. Indeed, one can minimize the amount of missing information during the construction of PBNs from gene expression data, using entropy as the objective function. Motivated by the results in [12, 32], we tackle the inverse problem by splitting it into two different inverse problems — viz. (i) construction of a transition probability matrix from a given stationary distribution (Problem ST), and (ii) construction of a PBN from a given transition probability matrix (Problem TP). For the Problem ST, we propose to construct a transition probability matrix from the prescribed stationary distribution. The state transitions in a PBN can be regarded as a Markov chain, and our aim is to find a transition probability matrix that has the prescribed stationary distribution. For Problem TP the main aim is to construct a PBN from a given transition probability matrix by identifying all the