Reconstruction of Gene Regulatory Networks Based on Two-Stage Bayesian Network Structure Learning Algorithm

Gui-xia Liu, Wei Feng, Han Wang, Lei Liu, Chun-guang Zhou

College of Computer Science and Technology, Jilin University, Changchun 130012, P. R. China

Abstract

In the post-genomic biology era, the reconstruction of gene regulatory networks from microarray gene expression data is very important to understand the underlying biological system, and it has been a challenging task in bioinformatics. The Bayesian network model has been used in reconstructing the gene regulatory network for its advantages, but how to determine the network structure and parameters is still important to be explored. This paper proposes a two-stage structure learning algorithm which integrates immune evolution algorithm to build a Bayesian network .The new algorithm is evaluated with the use of both simulated and yeast cell cycle data. The experimental results indicate that the proposed algorithm can find many of the known real regulatory relationships from literature and predict the others unknown with high validity and accuracy.

Keywords: gene regulatory networks, two-stage learning algorithm, Bayesian network, immune evolutionary algorithm

1 Introduction

The reconstruction of genetic regulatory networks based on microarray gene expression data is one of the most challenging tasks in bioinformatics. By constructing the gene regulatory networks, we can identify the complicated regulatory relationships, uncover the regulatory patterns in the cell, and gain the systematic view for biological process. Nowadays, many computational approaches have been proposed to solve this problem, such as Boolean networks[1], differential equations[2], linear combination and weighted[3], Bayesian networks[4] and neural networks[5]. Each method has its own strengths and weakness[6]. Boolean network model is a discrete model which is brief and clear, but relatively rough fixed and the inferring results have low accuracy. Differential equation model is the continuous network model, which uses the form of differential equations to describe the continuous impact and varying relationships among genes. Although this model can reflect the continuous dynamic relations among genes, the right side of the differential equation is difficult to define. The genetic relationship established by the linear combination model and the weighted matrix model is linear, but in fact, the relationship among genes is a very complex and

nonlinear. The Bayesian network model used in this paper is one of the constructing gene regulatory network models, which has the following advantages: first, it is suitable for processing the data with such random factors as disturbances and delay; second, it is able to process noise and lost data and at the same time add implicit variable information; third, it not only can describe regulatory relationship among genes, but also can get the extent and probability of regulation. Thus in this paper, a two-stage learning algorithm is proposed to determine the Bayesian network structure and optimize the network parameters by using the immune evolutionary algorithm. The method simplifies the traditional three-stage algorithm and the experimental result is more accord with the real biological regulatory network.

2 Two-stage learning algorithm

Cheng's three-phase algorithm[7] is based on dependence analysis and also it uses conditional mutual information as independent testing. When $I(X_i, X_j | C)$ is less than a certain threshold value ε, X_i and X_j are considered as condition independence. This algorithm can achieve fairly good results, but it needs ordered nodes which cannot be obtained in the microarray data in advance. Therefore, this algorithm cannot be used to

Corresponding author: Chun-guang Zhou
E-mail: cgzhou@jlu.edu.cn

Liu *et al.*: Reconstruction of Gene Regulatory Networks Based on Two-Stage
Bayesian Network Structure Learning Algorithm

87

construct the genetic regulatory network directly.

The immune algorithm's initial population is constructed by some simple networks, and the sketch which is built by the first phase of three-phase algorithm can be used as the initial population of immune evolutionary algorithm. Thus, this paper will combine the three-phase algorithm and immune algorithm to propose a new algorithm: two-stage learning algorithm. First, this algorithm uses the first phase of the three-phase algorithm to construct the simple network, and then the immune algorithm is used to optimize it until we find the best network[8].

2.1 Two-stage learning algorithm flow

Two-stage learning algorithm flow is as follows:
(1) Stage one: construct the network using the first phase of the three-phase algorithm
{Begin
  V={all the letters in the table};
  E={};// including the edge relations among nodes
  QUEUE L;
  IF $(v_i,v_j)\bullet$ V && $i \neq j$ && mutual information $I(v_i,v_j) > \varepsilon$
    THEN
    L.ADD$((v_i,v_j))$;
  END IF
  Sort in descending order;
  NODE node = L.REMOVE(the first node);
  E.ADD(node's edge);
  WHILE L is not empty DO
    Begin
     Take out a pair of nodes (v_i,v_j);
     IF there's no open pathway between two nodes
      THEN
      E.ADD(two nodes' edges);
      L.REMOVE(node pair$((v_i,v_j))$);
     END IF
    End
  END WHILE
End}
(2) Stage two: use the immune evolutionary algorithm to optimize network and find the best network structure
{Begin
  A_0 = {};//the initial network population
  Compute fitness(A_{0i});
  IF MAX(fitness(A_{0i})) THEN
    Set A_{0i} as antibody, vaccine extraction
  END IF

WHILE not over DO
  Begin
    IF the current population contains the best individual THEN
     EXIT(0);
    END IF
    ELSE
     Make crossover operation with probability Pc and get a new population B_k.
     Make mutation operation with probability Pm and get a new population C_k.
     Make vaccination operation to the population C_k and get a new population D_k.
     Make immune selection operation to population D_i and get a new generation father population A_{k+1}
    END ELSE
  End
  END WHILE
End}

2.2 The theory of static multiplication

The theory of static multiplication is introduced to the immune evolutionary algorithm in the optimization process in this paper. According to this theory we compare the fitness of offspring chromosome with the parent's after crossing and mutating operations. If the offspring chromosome fitness value is higher than the parent's, the parent's chromosome will be replaced by the offspring's, otherwise the offspring chromosome would be reserved[9,10]. The chromosomes' immunity is boosted and the average fitness of chromosomes is improved by using adaptive vaccine. As a result, the loss of optimum solution is avoided, searching space is reduced and evolution speed is increased, then the optimal solution can be achieved earlier.

2.3 The construction of fitness function

Fitness function reflects the match degree of network and data, it would bring punishment to the complex networks so that we can avoid searching outside the complex networks. The fitness function used in this paper is the Minimum Description Length (MDL)[11]. MDL theory which is based on the coding theory is one of the optional measure functions in the study of the Bayesian network structure. It can take into account both the complexity of network structures and the descriptive