Identification of Co-regulated Gene Network by Using Path Consistency Algorithm Based on Gene Ontology

Hanshi Wang[1], Chenxiao Wang[1], Lizhen Liu[1], Chao Du[1], and Jingli Lu[2]

[1] Information and Engineering College, Capital Normal University, Beijing, China
necrostone@sina.com, {wangchenxiao0329,liz_liu,dc0213}@126.com
[2] Agresearch Ltd, Hamilton, New Zealand
Janny.jingll@gmail.com

Abstract. Recently, the reconstruction of co-regulated gene network has become increasingly popular in the area of bioinformatics. It tries to find the associated information and network topology among genes through large numbers of biological data. In this paper, we proposed a novel method PC-GO by using path consistency (PC) algorithm based on gene ontology (GO). GO provides a basis for measuring the conditional semantic similarity between genes, and then PC algorithm is applied to remove links between genes with less correlation in the network. We successfully applied our algorithm to yeast data. Experimental results show that the accuracy and integrity of the co-regulated network acquired by our method outperforms previous methods.

Keywords: Gene ontology, conditional semantic similarity, path consistency algorithm, co-regulated gene network.

1 Introduction

High-throughput techniques have produced vast amounts of sequence, expression and structure data [1]. As the data source of bioinformatics, gene expression data is now in an increasingly important position. Increasing evidences suggest that interactions between genes have an impact on the regulation of gene expression [2]. Currently, many scholars are committed to find the association of genes [3-5] since during getting the associated information between genes, one of the major challenges is the identification of co-regulated gene network. Gene co-regulated network, which aims to find the associated information among genes through large numbers of biological data expressed by genes and visualize the network topology representing gene interactions, as well as reveals the complex reaction mechanism among genes, is regarded as one of the most important objectives in the field of bioinformatics.

Numerous analytical methods have been developed to identify gene co-regulated network from gene expression profiles [6]. Genes that are part of the same operon in prokaryotes, or have the same expression pattern in eukaryotes, are co-regulated transcriptionally [7]. Generally speaking, we considered a network of genes co-regulated if the percentage that has one or more common TFs is above 80%. Researches on the regulation relationship and regulatory mechanism provide the opportunity for

Z. Shi et al. (Eds.): IIP 2014, IFIP AICT 432, pp. 278–283, 2014.

understanding the underlying and predicting genes function, which can help to systematically characterize the process of life activities.

Gene Ontology (GO) [8] is a standard vocabulary of functional terms and allows for coherent annotation of gene products. These annotations provide a basis for new methods to compare the similarity of genes and gene products regarding their molecular function and biological role. The semantically similarity of annotation information on genes can be an evidence for functionally similarity of genes. Meanwhile, gene products that participate in the same biochemical reaction, have similar biological functions [9]. Constructing a gene co-regulated network automatically based on GO is still a big challenge at present.

In this paper, the Yeast dataset was used as the test data. We present a new PC algorithm using the CSS based on GO. It consists of two parts: the first part is to calculate the CSS between genes based on GO, and the second is to remove links between genes based on pair-wise similarity by using PC algorithm. Experimental results show that our method has improved the accuracy and integrity of the co-regulated network.

2 Method

In this section, we will introduce the CSS calculation method, as well as the PC-GO method for identifying co-regulated gene network.

2.1 Conditional Semantic Similarity (CSS) of Genes

The function similarity between genes can be determined by comparing the semantic similarity. To improve the accuracy of the semantic similarity of genes, we must consider the semantic similarity of GOs annotating genes, and the key is to measure the semantic similarity of GO terms. This is achieved by considering the Wang's method [10] of computing semantic similarity between gene pairs. Based on Wang's method, we proposed a new concept – conditional semantic similarity (CSS), which can be applied to the PC algorithm.

Since the semantic of a GO term are determined by its location in the GO graph and semantic relations with all of its ancestor terms, the directed acyclic graph (DAG) starting from the specific GO term and ending at the root GO term (a sub-DAG of an ontology) is used to show all the relationship of this specific GO term in the ontology. Formally, a GO term A can be represented as $DAG_A = (A, T_A, E_A)$, where T_A is the GO terms set in DAG_A, including term A and all of its ancestor terms in sub-DAG, and E_A is the set of edges (semantic relations) connecting the GO terms in sub-DAG. The semantic value of GO term A is defined as the aggregate contribution of all terms in sub-DAG. The contribution of a GO term (including A) to the semantic meaning of A is defined as S-value. S-value is calculated by:

$$\begin{cases} S_A(A) = 1 \\ S_A(t) = \max\left\{\omega_e * S_A(t') \mid t' \in childrenof(t)\right\} & t \neq A \end{cases} \tag{1}$$

where ω_e is the semantic contribution factor of two edge types of semantic relationship – "is a" and "part of". Through large numbers of repeated experiments, Wang obtained that the semantic contribution factors for "is - a" and "part - of" are 0.8 and 0.6, respectively.

Given $DAG_A = (A, T_A, E_A)$ and $DAG_B = (B, T_B, E_B)$ for GO terms *A* and *B* respectively, the semantic similarity, $S_{GO}(A,B)$, is defined as

$$S_{GO}(A,B) = \frac{\sum_{t \in T_A \cap T_B} S_A(t) + S_B(t)}{SV(A) + SV(B)} \tag{2}$$

where *t* is the intersection GO term of sub-DAG of *A* and *B*, $SV(A)$ and $SV(B)$ are the semantic value of GO term *A* and *B*. The semantic value of GO term *M* is calculated by $SV(M) = \sum_{t \in T_M} S_M(t)$.

As each gene is annotated by GO terms, the semantic similarity between two genes *A* and *B* can be represented by the semantic similarity of two GO terms sets, which annotated the corresponding gene. Assuming $GO_1 = \{go_{11}, go_{12}, \cdots, go_{1m}\}$ and $GO_2 = \{go_{21}, go_{22}, \cdots, go_{2n}\}$ are two GO terms sets that annotate genes G_1 and G_2 respectively, their semantic similarity is as follows:

$$\mathrm{Sim}(\mathrm{G}_1, \mathrm{G}_2) = \frac{\sum_{1 \le i \le m} \mathrm{Sim}(\mathrm{go}_{1i}, \mathrm{GO}_2) + \sum_{1 \le j \le n} \mathrm{Sim}(\mathrm{go}_{2j}, \mathrm{GO}_1)}{\mathrm{m} + \mathrm{n}} \tag{3}$$

In this formula, $Sim(go, GO) = \max_{1 \le i \le k} S_{GO}(go, go_i)$, that is, $Sim(go, GO)$ is defined as the maximum semantic similarity between term go and any of the *k* terms in set GO.

In this work, we proposed a new concept - conditional semantic similarity (CSS), which indicates the similarity of two genes given the state of another gene. CSS between three genes can be written as:

$$Sim(G_1, G_2 \mid G_3) = \frac{|\, Sim(G_1, G_2) - Sim(G_1, G_3) * Sim(G_2, G_3) \,|}{\sqrt{1 - Sim^2(G_1, G_3)} * \sqrt{1 - Sim^2(G_2, G_3)}} \tag{4}$$

2.2 The PC-GO Method

After we obtain the semantic similarity and CSS through formulation (3) and (4), the PC algorithm is used to remove the edges, which the semantic similarity or CSS value is smaller than the threshold θ given in advance. The θ value is experimentally set 0.8 for the reason that genes in a co-regulated network are highly functionally related.

The process of PC-GO begins with a complete graph and attempts to remove as many links as possible. First, for adjacent gene pair *i* and *j*, compute the semantic similarity $Sim(i, j)$ (0-order). If the gene pair *i* and *j* has a lower semantic similarity, it presents functionally irrelevant, then we delete the edge between genes *i* and *j*. Second, for adjacent gene pair *i* and *j*, select the adjacent gene *k* of them and compute

CSS value $Sim(i, j \mid k)$ (1-order). If the gene pair i and j has a low CSS, delete the edge between them. The next step is to compute higher order CSS until there are no more adjacent edges. At each round, the number of neighbors in the conditional set increases one by one. A conceptual representation of this approach is presented in Figure 1. In this figure, $Sim(\cdot,\cdot)$ is the semantic similarity and $Sim(\cdot,\cdot \mid \cdot)$ is the CSS. The semantic similarity and CSS lower than the given threshold represent functionally irrelevant between genes.

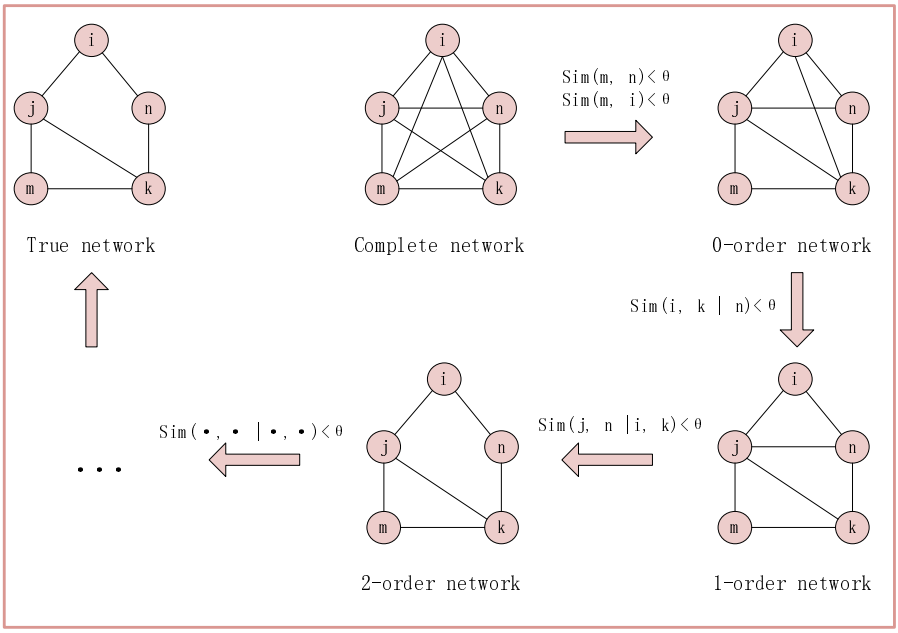

Fig. 1. Diagram of PC-GO method

3 Evaluation

Yeast Biochemical Pathway Database (YeastCyc) was used as the test data set. The YeastCyc Biochemical is a collection of manually curated metabolic pathway and enzymes of Saccharomyces cerevisiae. There are 217 metabolic pathways in the YeastCyc database. A total of 6381 yeast genes got from 217 pathways were selected for our investigation and GO terms search for all 6381 genes were performed based on the April 2014 releases of GO terms and gene annotations for Saccharomyces cerevisiae from the Gene Ontology Consortium[8]. We calculated the semantic similarities and CSS of gene pairs, and then partitioned genes into functionally related co-regulated networks.

Genes that share the same TF(s) are co-regulated, which can be a criterion to estimate whether genes are in a co-regulated network. The YEASTRACT database [11] (Yeast Search for Transcription Regulators And Consensus Tracking) is a curated

repository of more than 106000 regulatory associations between TFs and target genes in Saccharomyces cerevisiae, which can be used to search for common TFs of genes in each network. In the YEASTRACT database, we used the "Rank by TF" function and considered only expression evidence in the "documented" regulation to obtain the percentage of the genes in each network that are commonly regulated by one or more known TFs.

For each gene network, the summation of all the semantic similarity was calculated. Genes in a co-regulated network are considered to be co-regulated if the percentage of genes that have one or more common TFs is above 80%. Therefore, we eliminated networks containing less than five genes in order to achieve the 80% when one gene did not have a TF in common. Finally, 623 networks were obtained for analysis. All 415 networks were searched for documented TF-target relationships in the YEASTRACT database. Networks meet the rule that 80% genes have one or more common TFs account for 76.4% with the actual number of 476. Genes that in each network functionally participate in various biological processes, such as protein translation, RNA metabolism, carbohydrate metabolism, etc. Compared to other methods using gene expression profiles with the proportion of 52%, our method improved the accuracy significantly.

We choose two networks with the maximum and minimum summation if all the semantic similarity for verification. The verification result is shown in Table 1. As shown in Table 1, the network with maximum summation is a co-regulated network since Cst6 is the common TF of all the genes contained in this network while the network with minimum summation is not co-regulated as no more than 63.64% genes are regulated by one TF.

Table 1. Part of verification results

Network	Contained genes	Common TFs and percentages of commonly regulated	
Network with maximum summation	TFC3, EFB1, YAL004W, SPO7, CYS3, YAL018C, MAK16, POP5, YAL037W, ERV46, YAR023C, YAR066W, YAR075W, YBL077W	Cst6	100%
		Ace2	78.57%
		Spt20	71.43%
		Msn2	71.43%
		Sfp1	71.43%
		Gcr1	57.14%
		Snf2	50%
Network with minimum summation	ECM4, CSN9, ARI1, YAR1, ATG15, SAW1, GAL7, RPB9, PHM7, PCA1, COA4	Msn2	63.64%
		Msn4	63.64%
		Yap1	54.55%
		Bas1	54.55%
		Hsf1	54.55%
		Yhp1	45.45%
		Snf2	45.45%

4 Conclusions

We early proposed a novel PC-GO method by using path consistency algorithm based on gene ontology to identify gene co-regulated network, which provides a new way to the bioinformatics research. The effectiveness of the proposed approach has been experimentally demonstrated through its application to a well-researched yeast dataset. One of the strengths of our approach is that the prediction of co-regulated gene group does not require the availability of gene expression profiles. However, the result generated by the gene conditional semantic similarity calculation method may have an impact on the accuracy of PC algorithm, which require us to improve the accuracy of each procedure in the future research work.

Acknowledgements. This work was supported in part by National Science Foundation of China under Grants No. 61303105; the Humanity & Social Science general project of Ministry of Education under Grants No.14YJAZH046; the Beijing Educational Committee Science and Technology Development Planned under Grants No.KM201410028017; Academic Degree Graduate Courses group projects and the Beijing Key Disciplines of Computer Application Technology.

References

1. Obayashi, T., et al.: ATTED-II: a database of co-expressed genes and cis elements for identifying co-regulated gene groups in Arabidopsis. Nucleic Acids Research 35(suppl. 1), D863–D869 (2007)
2. Lanctôt, C., et al.: Dynamic genome architecture in the nuclear space: regulation of gene expression in three dimensions. Nature Reviews Genetics 8(2), 104–115 (2007)
3. Korbel, J.O., et al.: Systematic association of genes to phenotypes by genome and literature mining. PLoS Biology 3(5), e134 (2005)
4. Lee, I., et al.: Rational association of genes with traits using a genome-scale gene network for Arabidopsis thaliana. Nature Biotechnology 28(2), 149–156 (2010)
5. Perez-Iratxeta, C., Bork, P., Andrade, M.A.: Association of genes to genetically inherited diseases using data mining. Nature Genetics 31(3), 316–319 (2002)
6. Berri, S., et al.: Characterization of WRKY co-regulatory networks in rice and Arabidopsis. BMC Plant Biology 9(1), 120 (2009)
7. Teichmann, S.A., Babu, M.M.: Conservation of gene co-regulation in prokaryotes and eukaryotes. Trends in Biotechnology 20(10), 407–410 (2002)
8. Ashburner, M., et al.: Gene Ontology: tool for the unification of biology. Nature Genetics 25(1), 25–29 (2000)
9. Wei, H., et al.: Transcriptional coordination of the metabolic network in Arabidopsis. Plant Physiology 142(2), 762–774 (2006)
10. Wang, J.Z., et al.: A new method to measure the semantic similarity of GO terms. Bioinformatics 23(10), 1274–1281 (2007)
11. Teixeira, M.C., et al.: The YEASTRACT database: a tool for the analysis of transcription regulatory associations in Saccharomyces cerevisiae. Nucleic Acids Research 34(suppl. 1), D446–D451 (2006)