A Maximal Eigenvalue Method for Detecting Process Representative Genes by Integrating Data from Multiple Sources

Haixuan Yang[1] **Prajwal Bhat**[2,1] **Hugh Shanahan**[1] **Alberto Paccanaro**[1]
[1]Department of Computer Science [2]School of Biological Sciences
Royal Holloway University of London
{haixuan,p.bhat,hugh,alberto}@cs.rhul.ac.uk

Abstract

An important problem in computational biology is the identification of candidate genes which can be considered as representative of the different cellular processes taking place in the cell as it evolves through time. Multiple and very noisy data sources contain information about such processes and should therefore be integrated in order to obtain a reliable identification of such candidate genes. In this paper, we present a novel ranking algorithm which determines process representative genes by integrating a set of noisy binary relations between genes. We present some preliminary results on two artificial toy datasets and one real biological problem. In the biological problem, we use this method to identify representative genes of some of the fundamental biological mechanisms taking place during cellular growth in *A. thaliana* by integrating gene expression data and information from the gene GO annotation.

1 Introduction

Gene expression experiments measure the activity of thousands of genes in response to various conditions. In these experiments, genes involved in a particular biological mechanism tend to exhibit similar expression patterns and form groups. Selecting marker genes which can represent specific mechanisms is an important problem. These markers serve as readouts and help in making sense of the mechanisms, monitoring interactions between the mechanisms and also track any physiological effects they may exert. For example, as plants grow, the genetic data contained in them is converted into phenotype according to its genetic content and various environmental signals that are mediated by different types of hormones. Genes involved in various hormone pathways exhibit distinct similarity in expression patterns and form groups. Sensitive and specific markers which can track and report the dynamics of each group are essential for investigating the mechanisms of response to each hormone, cross-talk between hormone pathways and the relationship between hormones and phenotypic effects [3].

Multiple data sources exist for any set of genes that describe a particular biological mechanism. One such data type is expression similarity between two genes measured by a linear correlation coefficient (LCC). Besides gene expression data, one could also consider taxonomical data such as the one developed by the Gene Ontology (GO) Consortium [1]. The GO project is specifically designed for annotating gene products by a shared, structured and controlled vocabulary that can be applied to any organism. Within this ontology, terms are inter-related forming a DAG. The nodes of the GO tree represent terms with a specific biological meaning. Genes are better represented by lower nodes of the tree as these nodes describe biological concepts with higher precision. GO is divided into three independent categories - Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). Genes are annotated to GO terms in each of the three independent

categories. We can compute semantic similarity between pairs of gene products using methods developed for lexical taxonomies such as Resnik measure [5].

In this paper, we present a novel ranking algorithm which determines process representative genes by integrating a set of noisy binary relations between genes.

2 Problem Definition

2.1 Ranking on A Single Graph

Given a weighted graph $G = (V, E, W)$, where W_{ij} is the correlation between Gene i and Gene j, we want to rank on the nodes of G so that the node with a higher representative ability has a higher ranking score. Let us denote the ranking vector by $f = [f_1, f_2, \ldots, f_n]^T$. The representation ability is formally defined as:

- If W_{ij} is large, then the difference between ranking scores f_i and f_j is large.
- If $D_{ii} = \sum_k W_{ik}$ is large, then f_i is large.

This can be formulated as the optimization of the following objective function:

$$\max \sum_{ij} \frac{1}{2} W_{ij}(f_i - f_j)^2 + \mu(\sum_i D_{ii} f_i)^2 \;\; s.t. \; ||f||_2 = 1 \tag{1}$$

We notice that this function can be written in matrix form as:

$$\sum_{ij} \frac{1}{2} W_{ij}(f_i - f_j)^2 + \mu(\sum_i D_{ii} f_i)^2 = f^T L f + \mu f^T V * V^T f \tag{2}$$

where $L = D - W$, D is diagonal weight matrix, its entries are column (or row, since W is symmetric) sums of W, $D_{ii} = \sum_j W_{ji}$, and V is column vector consisting of D_{ii}. Here $L = D - W$ is the Graph Laplacian, which is a symmetric, positive semidefinite matrix, and so has only non-negative eigenvalues. The lagrangian of the above problem is

$$J = f^T L f + \mu f^T V * V^T f + \lambda(1 - f^T f). \tag{3}$$

By taking the derivative of J with respect to f, and setting it to be zero, we have $(L + \mu V * V^T) f = \lambda f$. Replacing $(L + \mu V * V^T) f = \lambda f$ in the objective function and considering the constraint $f^T f = 1$, we obtain $f^T L f + \mu f^T V * V^T f = f^T (L + \mu V * V^T) f = \lambda f^T f = \lambda$. Therefore, we only need to find the maximum eigenvalue and the corresponding eigenvector of the matrix $L + \mu V * V^T$. When $\mu \geq 0$, $L + \mu V * V^T$ is a symmetric, positive semidefinite matrix since both the Graph Laplacian $L = D - W$ and $V * V^T$ are positive semidefinite matrices. This property guarantees the existence of the maximum positive eigenvalue.

2.2 Ranking on Multi Graphs

Often multiple graphs are available which contains complementary information. For example, in Fig. 1, both Fig. 1(b) and Fig. 1(c) are subgraphs of the graph in Fig. 1(a), which therefore contains more information than the other two graphs. It is not surprising to observe that the ranking algorithm cannot find the best result on both of the subgraphs. However, it can be expected that a combination of some subgraphs (multiple data sources) can supplement each other so that a resulting graph can better model the complete information hidden in each individual graph. According to the above consideration, we propose to combine each subgraph by the following method.

Given a set of weighted graphs $G^k = (V^k, E^k, W^k)$ $(k = 1, 2, \ldots, I)$, we construct a graph $G = (V, E, W)$, where $V = \bigcup V^k$, $E = \bigcup E^k$, $W = \sum_{i=1}^{I} \alpha_k \underline{W}^k$, and $\underline{W}^i$ is an extended matrix constructed from W^i such that $\underline{W}^k_{ij} = W^k_{ij}$ if $(i, j) \in E^k$, and zero otherwise. We can apply to G the algorithm described in the previous section.

Note that α_k $(k = 1, 2, \ldots, I)$ are not necessarily nonnegative because some data sources may be contained in other data sources, and so they are redundant. Therefore these redundant data sources should be subtracted from the summation, which may cause the negative coefficients.

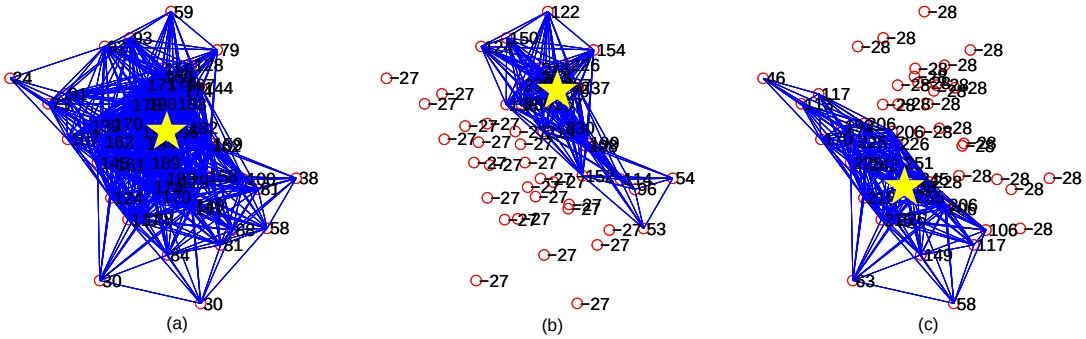

Figure 1: Dataset containing 50 points are generated by a normal distribution with mean $(0, 0)$ and variance I_2 (the 2×2 identity matrix). The symbol $\star$ in each graph highlights the highest rank node, and numbers show the ranking values by the solution to Eq. (1) where $\mu = 1$. (a) There is an edge between i and j if the distance d_ij between them is less than 2. The weight is calculated by $W_{ij} = e^{-d_{ij}}$ if there is an edge between i and j, and zero otherwise. (b) and (c) are two subgraphs of the graph in (a), in which some edges are removed.

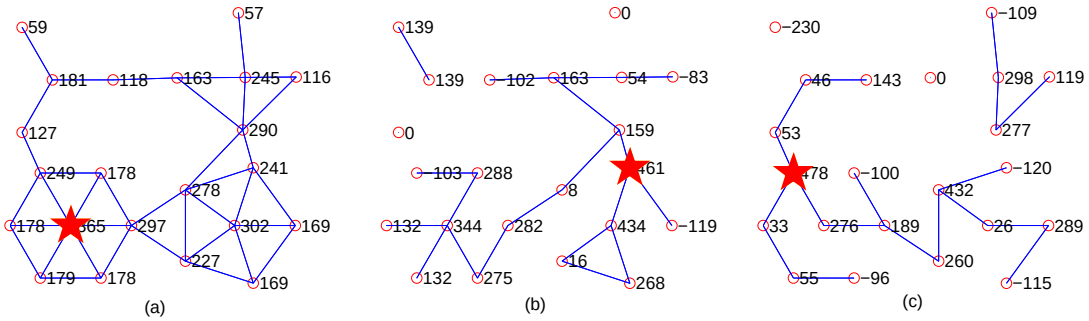

Figure 2: The symbol $\star$ in each graph shows the highest rank node, and numbers show the ranking values by the solution to Eq. (1) where $\mu = 1$. (a) The graph has 22 nodes. (b), (c) are subgraphs of the graph in (a).

3 Experiments

In Fig. 1 and 2 we show the result obtained on two problems. These figures show that the ranking algorithm can find the most representative node, and that by combining parts of a larger graph, the problem of information incompleteness can be addressed (see figure captions for details).

In order to test our method on real biological data we chose 400 genes which were known to belong to 10 different biological mechanisms (clusters) which are activated by cellular growth in *A. thaliana* [2]. Four different fully connected graphs were available, where these genes represented the nodes, and the weight on the edges corresponded to the linear correlation (LCC) measured during gene expression experiments (from [2]); and the distances between the gene GO annotations according to the three GO categories of Molecular Function (MF), Biological Process (BP) and Cellular Component (CC). These distance were measured using [4].

We will evaluate our ranking algorithm by comparing the relevance score (RS) of the top N highest ranking genes using the following formula.

$$RS_N = \sum_{i=1}^{N} Rank(i) * C\_Cluster(i),$$

Table 1: The statistics for the number of genes appearing in each cluster among the top N genes

Top N	# C1	# C2	# C3	# C4	# C5	# C6	# C7	# C8	# C9	# C10
10	1	0	1	4	2	0	0	0	1	1
20	2	4	1	5	5	0	0	0	1	2
30	6	4	1	5	5	2	2	1	2	2
40	8	4	3	6	7	4	3	1	2	2
50	8	4	5	7	10	6	4	1	2	3

where $Rank(i)$ is the i-th highest ranking score, and $C\_Cluster(i)$ is the linear correlation coefficient between the expression of gene i and the mean of the expression of the genes in the same cluster [2]. Therefore, RS_N measures how good the top N results of the ranking list are relevant to the linear correlation coefficients.

The RS_{50} scores obtained using separately the LCC, CC, BP, and MF matrix were 2.5379, 2.4986, 3.1995, and 3.1102 respectively. For a linear combination of LCC, CC, BP and MF, we can achieve a RS_{50} score 3.8992, which is increased by 21.87% over the best one using single data source. The corresponding coefficients are $\alpha_1 = -0.2, \alpha_2 = -0.2, \alpha_3 = 0.9, \alpha_4 = 0.5$.

We found that optimal results are obtained when LCC is weighed the least and BP is assigned the highest weight. The fact that semantic similarity of BP terms contribute the most to the performance of the algorithm is in agreement with biological knowledge since BP is essentially an ontology that attempts to describe the various biological mechanisms in the organism. The -0.2 weight of LCC suggests that BP may render some of the LCC data redundant.

Next, we demonstrate the representative ability of the ranking algorithm. The best result should be that the top 10 genes with highest ranking score should be selected from 10 different clusters. By Table 1, we show that, although our algorithm cannot achieve the ideal case, it can cover all the 10 clusters by top 30 genes with highest ranking scores.

4 Conclusions

We have presented a novel ranking algorithm that has the ability to select the most representative node in a graph. Through implementing the ranking algorithm on some toy data, and some real biological data, we have also showed that with multiple data sources, the algorithm performance can be greatly increased. In the future work, we attempt to learn the mixing coefficients for the multiple data sources.

Acknowledgment

We thank L. Bögre, E. López-Juez, A. Devoto, and D. Zervas for their helpful discussions.

References

[1] M. Ashburner, C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, and G. Sherlock. Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat. Genet.*, 25(1):25–29, May 2000.

[2] E. López-Juez, E. Dillon, Z. Magyar, S. Khan, S. Hazeldine, de Jager SM, J. A. Murray, G. T. Beemster, L. Bögre, and H. Shanahan. Distinct light-initiated gene expression and cell cycle programs in the shoot apex and cotyledons of arabidopsis. *Plant Cell*, 20:947–968, 2008.

[3] J. L. Nemhauser, F. Hong, and J. Chory. Different plant hormones regulate similar processes through largely nonoverlapping transcriptional responses. *Cell*, 126(3):467–475, 2006.

[4] R. Gentleman Using go for statistical analyses. *Bioconductor Vignettes*, 2005.

[5] P. Resnik. Semantic similarity in a taxonomy: An information-based measure and its application to problems of ambiguity in natural language. *JAIR*, 11:95–130, 1999.