The First International Symposium on Optimization and Systems Biology (OSB'07)
Beijing, China, August 8–10, 2007
, pp. 68–74

Gene Function Prediction with the Shortest Path in Functional Linkage Graph

Xing-Ming Zhao[1,2,3,*] Luonan Chen[1,3,4]
Kazuyuki Aihara[1,3]

[1] ERATO Aihara Complexity Modelling Project, JST, Tokyo 151-0064, Japan.
[2] Intelligent Computing Lab, Hefei Institute of Intelligent Machines, Chinese Academy of Sciences, Hefei, Anhui, 230031, China
[3] Institute of Industrial Science, The University of Tokyo, Tokyo 153-8505, Japan
[4] Department of Electrical Engineering and Electronics, Osaka Sangyo University, Osaka 574-8530, Japan

Abstract This paper presents a new technique for gene function prediction with the shortest path in functional linkage graph. With existing protein-protein interaction data, complex data and gene expression data, a weighted functional linkage graph is inferred. By finding the shortest path in the functional linkage graph, the functions of unknown genes can be predicted with the functions of those that have functional links to the unknown genes. The experiments show promising results and prove the efficiency of the proposed methods.

1 Introduction

One of the main goals in post-genomic era is to predict the biological functions of genes. In the literature, many methods have been developed for this problem, among which the straightforward way is applying PSI-blast [1] and FASTA [2] to find the homology relationships between known genes and the query gene, and transfer functions of the most similar known genes to the query gene. Furthermore, structure-based approaches are proposed because structure is more conserved than sequence, and proteins with similar structures are clustered into a structure space that is then mapped to the functional space [3]. The methods described above use the entire protein structure or sequence for gene function prediction. However, gene functions may be determined by only a few key residues, which are denoted as functional motifs here. In literature, it has been shown that the functional motifs, which may directly mediate catalysis or binding, are important for gene function [4]. With the local similarity among a small number of key residues, new genes can be annotated with the function of known genes.

Recently, with the advance in high-throughput biotechnologies, a large amount of biological data have been generated, such as yeast two-hybrid systems, protein

*Email: xmzhao@aihara.jst.go.jp

complex and gene expression profiles, etc. These data are rich sources for deducing and understanding gene functions. In literature, protein-protein interaction data has been widely used for gene function prediction with the assumption that interacting proteins have the same or similar functions, i.e. 'guilty by association' rule [5]. For example, Schwikowski et al. [6] have proposed neighbor counting method for gene function determination; Hinshigaki et al. [7] proposed the χ^2 statistics to infer protein function. Some other research works have utilized the Markov random fields [8] and simulated annealing [9] techniques for gene function prediction. In addition, gene expression data has also been widely used for gene function prediction with clustering where genes with similar expression values are assumed to have similar function [10].

Although the successful application of the high-throughput data for gene function prediction, the errors in the high-throughput data have not been handled well. In this paper, we propose a new technique for constructing a weighted functional linkage graph with protein interaction data, protein complex and gene expression profiles. By finding the shortest path in the functional linkage graph, we can infer the functional links between genes. With the inferred functional links, Support Vector Machines (SVMs) is utilized to predict the functions of unknown genes.

The rest of this paper is organized as following. Section 2 presents the datasets used in this work; Section 3 describes the proposed methods for gene function prediction; Section 4 presents the experimental results; Finally, conclusions are drawn in Section 5.

2 Data sets

In this study, to predict the function of genes, we use three kinds of data including protein-protein interaction data, microarray data and protein complex data. All of these data are integrated for function prediction of *S. cerevisiae* genes.

In this work, the functional annotation of *S. cerevisiae* genes was obtained from the FunCat 2.0 functional classification scheme, which can be downloaded from the Comprehensive Yeast Genome Database (CYGD) of MIPS [11]. The annotation data in FunCat are organized as a hierarchical, tree like structure with up to six levels of increasing specificity. In total, the FunCat includes 1307 functional categories. In this work, 13 general functional classes are selected, where each class has no less than 30 gene annotations. Table. 1 shows the selected functional classes and corresponding number of genes.

The protein interaction data used in our experiments were obtained from the BioGRID database [12]. The 2.0.20 version of the BioGRID is used in this work. The dataset contains 82,633 pairs of interactions between 5,299 yeast genes. The interaction network of genes can be denoted as a graph $G(V,E)$, where the vertexes V are genes and the edges E are interactions between genes.

The gene expression profiles used in this work were from the Rosetta Compendium [13], which includes 300 diverse mutations and chemical treatment experiments. The dataset contains 6298 genes, of which, 4,376 genes are among the 5,299

Table 1: The functional categories and genes used in this work.

Functional category	Number of proteins
01 metabolism	1292
02 energy	354
10 cell cycle and dna processing	812
11 transcription	534
12 protein synthesis	325
14 protein fate (folding, modification, destination)	905
20 cellular transport, transport facilities and transport routes	684
30 cellular communication/signal transduction mechanism	205
32 cell rescue, defense and virulence	663
34 interaction with the environment	394
40 cell fate	37
42 biogenesis of cellular components	545
43 cell type differentiation	357

genes in the protein-protein interaction dataset. Finally, the dataset used in this work contains 4,376 genes and 300 real value features for gene expression profiles.

The protein complex data were obtained from the MIPS database, including the data from [14] and [15]. The protein complex data is used here because genes occurring in the same complex are assumed to have the same or similar functions. Although it is not reasonable to infer interaction relationship from protein complex directly, the genes in the same complex have functional correlations. Hence, we assign functional relationship to the genes occurring in the same complex, where an edge is constructed for any pair of genes occurring in the same complex. Finally, 62,042 functional edges were assigned to our dataset.

3 Gene function prediction with shortest path in functional linkage graph

After getting the protein-protein interaction data, gene expression profiles and protein complex data, we first construct a graph $G(V,E)$ with the protein interaction and protein complex data. In the graph, a pair of genes will be assigned an edge if they interact or appear in the same complex. Furthermore, the absolute expression correlation $C_{i,j}$ of gene expression is used as the weight of edge E in graph G. Therefore, we get a functional linkage graph. The edge length between vertices i and j is defined as $d_{i,j} = (1 - C_{i,j})^{\alpha}$, where α is a parameter used to enlarge the difference between edge lengths. We use Dijkstra's algorithm [16] to find the shortest path between a vertex to all the other vertices in the functional linkage graph. With the shortest paths available, each gene i can be expressed as a vector as following:

$$f_i = [\text{st\_path}_{i1}, \ldots, \text{st\_path}_{in}] \quad (1)$$

where st_path$_{ij}$ is the shortest path length between gene i and gene j, and n is the total number of genes in the functional linkage graph.

In addition, the Radial Basis Function (RBF) kernel is used to measure the similarity between a pair of genes with:

$$K(x,y) = \exp(-\gamma||x-y||^2) \tag{2}$$

where x and y are vectors for two genes, respectively. Furthermore, the kernel matrix is normalized as following:

$$K(x,y) = \frac{K(x,y)}{\sqrt{K(x,x)K(y,y)}}. \tag{3}$$

After getting the kernel matrix for genes, the SVMs classifier is utilized for gene function prediction, where one classifier is constructed for each functional class, respectively. Given an unknown gene, it will be assigned to the class that has the biggest decision value.

4 Experimental results

In this experiment, we applied our proposed methods to gene function prediction. To evaluate the performance of our proposed methods and other methods, we adopt the following indexes of *Sensitivity*, *Specificity*, and AUC score that are defined as:

$$Specificity = \frac{TN}{TN+FP}, \tag{4}$$

$$Sensitivity = \frac{TP}{FN+TP}, \tag{5}$$

where TP is the number of true positives, FP is the number of false positives, TN is the number of true negatives, FN is the number false negatives, and AUC score is the area under the ROC curve (AUC). Furthermore, all the methods are evaluated with 3-fold cross-validation.

In this experiment, we first compared the proposed methods to other three methods: SVMs trained on protein-protein interaction and protein complex data (denoted as PPI+complex), SVMs trained on gene expression profiles, and SVMs trained on the kernel integration of all the above three datasets. The parameter α in identifying the shortest path in the functional linkage graph is set to 1. For the protein-protein interaction and protein complex, the diffusion kernel is utilized and the parameter τ of the kernel is set to 5, and the kernel matrix is denoted as $K_{ppi+comp}$. For the gene expression profiles, the gaussian kernel is utilized and the parameter σ is set to 0.5, and this kernel matrix is denoted as K_{gene}. Both $K_{ppi+comp}$ and K_{gene} are normalized as described in Eq.3. Furthermore, the kernel-based integration of all the three data sources is utilized to predict gene function as described in [17], where the integrated

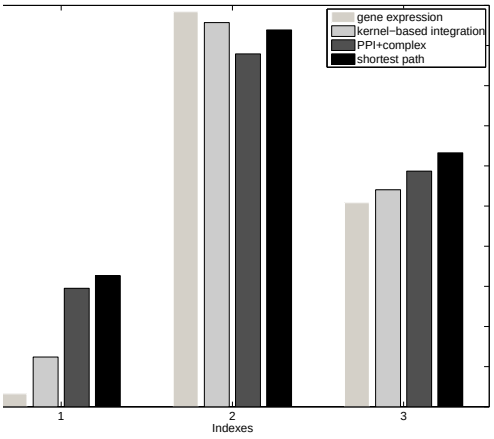

Figure 1: The performance of the four methods for gene function prediction, where X-axis denotes the performance indexes: 1 Sensitivity, 2 Specificity and 3 AUC score

kernel is denoted as K_{all} and defined as $K_{all} = K_{ppi+comp} + K_{gene}$ which has performed well in gene function prediction [17].

Fig.1 shows the results obtained via the four methods, where the results are averaged over 13 functional classes. It can be seen from Fig.1 that our proposed method outperforms all the other methods with overall performance. The results prove that the functional linkage graph constructed in this work is indeed useful for gene function prediction. It can also be seen from Fig.1 that the shortest path in the functional linkage graph can really capture the functional links between genes. Furthermore, our proposed method can better integrate different data sources than the kernel-based method for gene function prediction. The poor performance of the kernel-based integration technique may be due to the noise in the gene expression data.

Furthermore, we compared the proposed methods against the SVMs trained on protein-protein interaction and protein complex class by class. Fig. 2 shows the comparison of the two methods with respect to AUC scores. It can be seen from Fig. 2 that our proposed method outperforms the one using protein interaction and protein complex in nearly every class. The results prove again that the functional linkage graph is really useful for gene function prediction, and our proposed method is effective for integrating different data sources for gene function prediction.

5 Conclusions

In this paper, a new technique is proposed for gene function prediction. A functional linkage graph is constructed with the integration of the protein-protein interaction data, protein complex data and gene expression profiles. By finding the

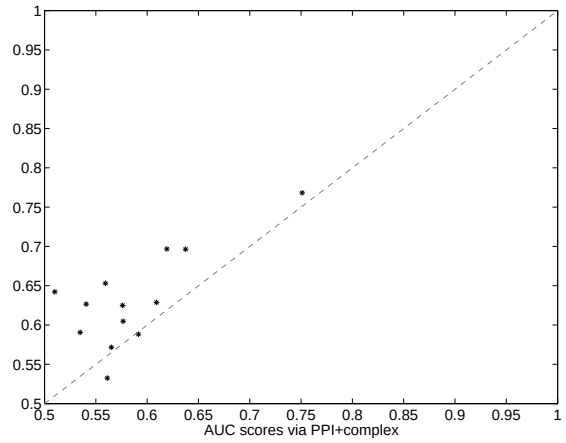

Figure 2: The comparison between PPI+complex method and shortest path method class by class

shortest paths in the functional linkage graph, the functional links between genes can be found. Furthermore, the SVMs are utilized to predict the function of unknown genes. Numerical experiments show that our proposed methods outperform those using single information source, e.g. gene expression profiles or protein interaction. In addition, our proposed methods also outperforms the kernel-base integration method, which proves the efficiency and effectiveness of the proposed methods.

Acknowledgment

This work was partly supported by the 863 Project of National High Technology of China (2006AA02Z309).

References

[1] Altschul, S., Madden, T., Schafer, A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. (1997) Gapped blast and psi-blast: A new generation of protein data. *Nucleic Acids Research*, **25**, 3389–3402.

[2] Pearsom, W. and Lipman, D. (1998) Improved tools for biological sequence comparison. *Proc. Natl Acad.Sci.USA*, **85**, 2444-2448.

[3] Hou, J.T., Jun, S.R., Zhang, C. and Kim, S.H. (2005) Global mapping of the protein structure space and application in structure-based inference of protein function. *Proc. Natl Acad.Sci.USA*, **102**, 3651–3656.

[4] George, R.A., Spriggs, R.V., Bartlett, G.J., Gutteridge, A., MacArthur, M.W., Porter, C.T., Al-Lazikani, B., Thornton, J.M., and Swindells, M.B. (2005) Effec-

tive function annotation through catalytic residue conservation. *Proc Natl Acad Sci USA*, **102**, 12299–12304.

[5] Nabieva, E., et al. (2005) Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, **21**, Suppl. 1, i302–i310.

[6] Schwikowski, B., Uetz, P. and Fields, S. (2000) A network of protein-protein interactions in yeast. *Nat. Biotechnol.*, **18**, 1257–1261.

[7] Hishigaki, H., Nakai, K., Ono, T., Tanigami, A. and Takagi, T. (2001) Assessment of prediction accuracy of protein function from protein-protein interaction data. *Yeast*, **18**, 523–531.

[8] Letovsky, S. and Kasif, S. (2003) Predicting protein function from protein/protein interaction data: a probabilistic approach. *Bioinformatics*, **19**, Suppl. 1, i197–204.

[9] Vazquez, A., et al. (2003) Global protein function prediction from protein ‐ protein interaction networks. *Nat. Biotechnol*, **21**, 697–670.

[10] Zhou, X., Kao, M.J. and Wong, W.H. (2002) Transitive functional annotation by shortest-path analysis of gene expression data, *PNAS*, **99**, 12783–12788.

[11] Ruepp, A., Zollner, A., Maier, D., Albermann, K., Hani, J., Mokrejs, M., Tetko, I., Guldener, U., Mannhaupt, G., Munsterkotter, M. and Mewes, H.W. (2004) The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes, *Nucl. Acids Res.*, **32**(18), 5539–5545.

[12] Stark, C., Breitkreutz, B.J., Reguly, T., Boucher L., Breitkreutz A., and Tyers, M. (2006) BioGRID: A General Repository for Interaction Datasets. *Nucleic Acids Res.*, **34**(Database issue), D535–D539.

[13] Hughes, T. R., Marton, M. J., Jones, A. R., Roberts, C. J., Stoughton, R., Armour, C. D., Bennett, H. A., Coffey, E., Dai, H., He, Y. D., et al. (2000) Functional Discovery via a Compendium of Expression Profiles *Cell*, **102**, 109–126.

[14] Gavin, A., Bosche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J., Michon, A. and Cruciat, C. (2002) Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, **415**, 141–147.

[15] Ho, Y., Gruhler, A., Heilbut, A., Bader, G.D., Moore, L., Adams, S., Millar, A., Taylor, P., Bennett, K., Boutilier, K. et al. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. *Nature*, **415**, 180–183.

[16] Dijkstra E. W. (1959) A note on two problems in connexion with graphs. In: Numerische Mathematik. 1: 269–271

[17] Lanckriet, G.R., et al. (2004) Kernel-based data fusion and its application to protein function prediction in yeast. *Pac. Symp. Biocomput*, 300–311.