Prediction of regulatory networks: identification of transcription factor-target relationship from gene ontology information and gene expression data

Cai, Yu-Dong and Jen, Chi-Hung and Qian, Jiang and Qian, ZiLiang and Muldoon, Mark

2007

ISSN 1749-9097

Prediction of regulatory networks: identification of transcription factor-target relationship from gene ontology information and gene expression data.

Yu-Dong Cai[1], Chih-Hung Jen[2], Jiang Qian[3], Ziliang Qian[4]
Mark Muldon[1*]

[1]Department of Mathematics, The University of Manchester, P.O. Box 88, Sackville Street, Manchester M60 1QD, UK

[2]School of Biochemistry and Microbiology, University of Leeds, Leeds, West Yorkshire, LS2 9JT, U.K

[3]The Wilmer Institute, Johns Hopkins University School of Medicine, Baltimore, MD 21287, USA

[4]Key Laboratory of System Biology, Shanghai Institutes for Biological Sciences, Chinese Academy of Sciences, 320 Yueyang Road, Shanghai 200031, China

*Corresponding to: M.Muldoon@Manchester.ac.uk (M. Muldon)

Short Title: regulatory networks prediction

Keywords: Gene ontology; regulatory networks; nearest neighbour algorithm; gene expression

ABSTRACT

Defining regulatory networks, linking transcription factors (TFs) to their targets, is a central problem in post-genomic biology. Here we apply an approach based on the Nearest Neighbour (NN) Algorithm to predict the targets of a transcription factor by combining gene ontology (GO) and gene expression data. In particular, we used NN algorithm to predict the regulatory targets for 36 transcription factors in the *Saccharomyces cerevisiae* (Qian J. et al., 2003, Bioinformatics. 19(15):1917-26) based on the gene ontology and microarray expression data from various physiological conditions. We trained and tested our NN algorithm on a data set which contains a number of both positive and negative examples. The overall success rate by the jackknife test for the dataset was 97%, and that for the regulatory targets(positive) was 58%, suggesting that such a hybrid approach particularly by incorporating the knowledge of gene ontology) may become a useful high-throughput tool in the area of regulatory networks modelling.

I. INTRODUCTION

The transcription of a gene in a cell can be regulated in different levels, such as alteration of DNA structure, DNA methylation, and DNA-RNA interaction etc. Among them, the transcription factor (TF) binding to the target (*cis*-regulatory element of a gene) is the most important key mechanism because it can alter the amount of the gene expression and further to affect the cell behaviours. Thus, identification of the TF-target relationship can help biologists to decipher the transcriptional regulatory networks and understand the processes of cell differentiation or cellular responses to the environmental conditions.

Currently, the TF-target relationship in the whole genome can be extensively determined by identification of the TF binding site (TFBS) on the *cis*-regulatory element of a gene, either using experimental or computational approaches. For experimental approaches, the most popular approach is ChIP-on-chip. It combines the chromatin immunoprecipitation and microarray technologies, and can directly identify *in vivo* target promoters for a specific TF (Orlando, 2000; Ren, *et al.*, 2000; Wang, *et al.*, 2002). For computational approaches, many algorithms have been developed in the past few years, and they can be generally classified into two main strategies: (1) For the TF with well known TFBS, the TF targets can be revealed by scanning for consensus binding sites or position weight metrics (PWM) in their promoter regions (Banerjee and Zhang, 2002; Kel, *et al.*, 2003; Matys, *et al.*, 2003); (2) For the TF without known TFBS, the motif-finding algorithms were used to find the novel binding sites among promoter regions of a group of genes with related functions or the TF correlated genes derived from gene expression data (Banerjee and Zhang, 2002; Haverty, *et al.*, 2004; Qiu, 2003; Zhu, *et al.*, 2002). That is based on the assumption that genes with the similar gene expression profile may under the similar transcriptional regulation.

Even though the results obtained from these computational approaches thus far are very promising to discover TF-target relationships, they still suffer few limitations, such as that not all genes shared common TFBS are the targets of a TF, and not all co-regulated gene promoters share common TFBSs. The later limitation is far more significant when using the second strategy described above based on microarray data sets, since gene expression relationship between a TF and its target is complex, dynamic, and non-linear. For example, TF and its targets do not have a correlated expression profile over a time course (Qian, *et al.*, 2001). To solve this issue, Qian (Qian, *et al.*, 2003) employed the support vector machines (SVMs) approach to identify the TF-target relationship, without identify TFBSs in the promoter region of the target gene, using budding yeast gene expression dataset. This tool can achieve 63% precision for the positive prediction and 93% accuracy for overall prediction.

Recently, rather than using gene group with correlated expression profiles, the over-represented individual and pairs of TFBSs in the proximal promoters of gene group with the similar Gene Ontology terms have been addressed (Cora, *et al.*, 2004; Long, *et al.*, 2004). This suggests that using Gene Ontology information could better increase the accuracy of determining the TF-target relationships in the genome. In the view of this, here a strategy is developed to represent a gene by the combination of the gene ontology composition and gene expression data. The combination makes allowance for bringing out the best in one another. With the approach, the Nearest Neighbour (NN) algorithm was employed to predict the 175 regulatory targets for 36 transcription factors in the *Saccharomyces cerevisiae* (Qian, *et al.*, 2003), and high success rates are obtained.

II. HYBRIDIZATION OF GENE ONTOLOGY, AND GENE EXPERSSION DATA

To improve the quality of predicting transcription factors targets, a logic step is to catch the core features of a gene. According to the Gene Ontology (GO) Consortium (Ashburner et al., 2000), the GO database was established based on the following criteria: **(a)** biological process referring to a biological objective to which the gene or gene product contributes; **(b)** molecular function defined as the biochemical activity of a gene product; and **(c)** cellular component referring to the place in the cell where a gene product is active. Since the above three criteria are not only the attributes of genes, gene products or gene-product groups, but also the core features reflecting the subcellular localization, it is anticipated that the prediction quality will be enhanced if using the GO database to define genes according to the following procedures.

By mapping of InterPro (Apweiler et al., 2001) entries to GO, one can get a list of data called “InterProt2GO” (ftp://ftp.ebi.ac.uk/pub/databases/interpro/interpro2go/), where each InterPro entrance corresponds to a GO number. The relationships between InterPro and GO may be one-to-many, “reflecting the biological reality that a particular protein may function in several processes, contain domains that carry out diverse molecular functions, and participate in multiple alternative interactions with other proteins, organelles or locations in the cell.” (Ashburner et al., 2000). For example, “IPR000003” corresponds to “GO:0003677”, “GO:0004879”, “GO:0005496”, “GO:0006355” and “GO:0005634”. Also, since the current GO database is far from complete yet, some InterPro entrances (such as IPR000001, IPR000002, and IPR000004) do not have the corresponding GO numbers in the InterProt2GO list.
Furthermore, the GO numbers in InerProt2GO are not increasing successively and orderly, and hence a reorganization and compression procedure was taken to renumber them. For example, after such a procedure, the original GO numbers GO:0000012, GO:0000015, GO:0000030, …, GO:0046413 would become GO_compress: 0000001,

GO_compress: 0000002, GO_compress: 0000003, ……, GO_compress: 0001930, respectively. The GO database thus obtained is called GO_compress database, whose dimensions were reduced to 1,930 from 46,413 in the original GO database. Each of the 1,930 entities in the GO_compress database will serve as a base to define a gene.

However, the current GO numbers do not give a complete coverage in the sense that some genes might not belong to any of the GO numbers. Although the problem will eventually no longer exist as GO increases in size, to cope with such a situation right now, a hybrid approach was introduced by combining GO with the gene expression data (http://rana.lbl.gov/EisenData.htm), as described below.

(1) Use the program IPRSCAN (Apweiler et al., 2001) to search InterPro (release 6.1) database (Apweiler et al., 2001) for a given gene, if there is a hit corresponding to the *i*th number of the GO_compress database, then the *i*th component of the gene in the 1930D GO_compress space is assigned 1; otherwise, 0. Thus, the gene can be formulated as

$$\mathbf{P} = \begin{bmatrix} a_1 \\ a_2 \\ \vdots \\ a_i \\ \vdots \\ a_{1930} \end{bmatrix}, \qquad (1)$$

where

$$a_i = \begin{cases} 1, & \text{hit found in GO\_compress} \\ 0, & \text{otherwise} \end{cases} \qquad (2)$$

(2) If no hit (i.e., no corresponding GO number) was found at all in the entire 1930D GO_compress space, the gene should be defined in the 80D gene expression space (http://rana.lbl.gov/EisenData.htm), as given below

$$\mathbf{P} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_j \\ \vdots \\ b_{7785} \end{bmatrix}, \qquad (3)$$

where b_j is the jth gene expression time points.

To encode the regulatory network prediction problem into a form suitable for training of a machine learning method, we construct TF-target pairs. These pair a known transcription factor S and a putative target gene T that maybe regulated by this factor. For instance, the pairing (S => T) means transcription factor S regulates gene T. To connect this paring with gene ontology and gene expression information, we note that each gene in the pair is characterized by its GO and expression information, which comprise data from samples collected at various time points during several biological procedures. In total, we used a 1930D GO vector and 80D gene expression vector to characterize each gene. Then putative TF-target pairing corresponds to a 3860D=2*1930D GO vector in which the first 1930D vector for the TF while the second 1930 are for the regulated gene, or a 160D=2*80D gene expression vector in which the first 80 data points for the TF while the second 80 are for the regulated gene.

III. THE NEAREST NEIGHBOUR ALGORITHM

The Nearest Neighbour (NN) Algorithm (Cover & Hart, 1967; Friedman et al., 1975) tries to classify the new patterns into their class membership by comparing the features of the unknown new patterns with the features of the patterns which have already been classified. It is particularly useful in the situations when the distributions of the patterns and the categories of the patterns are unknown. The approach will weight heavily the evidence derived from the nearby patterns. It is attractive because it is simple to implement and has a low probability of error.

Suppose there are N pairs ($\mathbf{P}_1$, $\mathbf{P}_{2,}$ …, $\mathbf{P}_N$) which have been classified into categories 1, 2, …, μ. Now, for a query pair **P**, how can we predict which category it belongs to? According to the nearest neighbour principle, the prediction can be formulated as follows. First, let us define a *generalized distance* between **P** and ***P**_k* (*k* = 1, 2, …, N) given by

$$D(\mathbf{P},\mathbf{P}_k) = 1 - \frac{\mathbf{P}\cdot\mathbf{P}_k}{\|\mathbf{P}\|\|\mathbf{P}_k\|}, \qquad (k=1,\ 2,\ \cdots,\ N) \tag{4}$$

where P, P_k is the dot product of vectors P and P_k, and $\|P\|$ and $\|P_k\|$ their modulus, respectively. Obviously, when $P \equiv P_k$, we have $D\left(P,P_k\right)=0$ Generally speaking, $D\left(P,P_k\right)$ is within the range of 0 and 1; *i.e.*, $0 \le D\left(P,P_k\right) \le 1$. Accordingly, the NN algorithm can be expressed as follows. If the generalized distance between P and $P_k\left(k=1,2,\cdots,\textbf{or } N\right)$ is the smallest; i.e.

$$D(\mathbf{P},\mathbf{P}_k) = \textbf{Min}\left\{D(\mathbf{P},\mathbf{P}_1), D(\mathbf{P},\mathbf{P}_2),\cdots,D(\mathbf{P},\mathbf{P}_N)\right\}, \tag{5}$$

then the query pair **P** is predicted belonging to the same category as of $\mathbf{P}_k$. If there is a tie, the query pair is not uniquely determined, but cases like that rarely occur.

The following self-consistency principle should be followed in practically using the hybridization approach. If a query pair was defined in the 3860D GO_compress space (see eq.1), then the prediction should be carried out based on those pairs in the training set that could be defined in the same 3860D space. If all of the components for the query pair in the 3860D Go_compress space were zero and hence it was defined by shifting to the 160D gene expression space (see eq.3), then the prediction should be conducted on the basis that all the rule parameters were derived from the same 160D space. Accordingly, the current NN predictor actually consists of two sub predictors: **(a)** the NN-3860D predictor that operates in the 3860D GO_compress space, **(b)** the NN-160D predictor that operates in the 160D gene expression space.

IV. POSITIVES AND NEGATAIVES

Positive examples were obtained from the reference(Qian, et al., 2003)

Negative examples were produced as the same method described in the reference(Qian, et al., 2003), in total, we constructed 3456 negative examples, which is about 20 times the number of positive examples.

V. RESULTS AND DISCUSSION

The computation was carried out in a Silicon Graphics IRIS Indigo workstation (Elan 4000). According to the search procedures as described in section II, we obtained the following results. For the 3631 pairs for both TF-target(positives) and Non-TF-target (negatives), 3543 got hits in the GO database and hence were defined in the 3860D GO_compress space, the left 88 pairs were defined in the 160D gene expression space. See Table 1 for the detail.

This means that, if only the GO database was used, 88 TF-target pairs would have no definition. That is why it is so important to hybridize with the gene expression data. Thus, the hybrid algorithm was operated according to the flowchart: if a query pair was defined in the GO_compress database, then the NN-3860D predictor was used to predict; if the query pair could not be defined in the GO_compress database but defined in the gene expression space, then the NN-160D predictor was used to predict.

The prediction quality was examined by the jackknife test. Compared with the independent dataset test and sub-sampling test often adopted in biology, the jackknife test is thought the most objective and effective method for cross-validation in statistics (Mardia et al., 1979). This is because in the independent dataset test, the selection of a testing dataset is quite arbitrary, and the accuracy thus obtained lacks an objective criterion unless the testing dataset is sufficiently large (Chou & Zhang, 1995; Zhou &

Assa-Munt, 2001). As for the sub-sampling test in which a given dataset is divided into several subsets, the problem is that the number of possible divisions might be too large to be handled. Hence in any practical sub-sampling tests as conducted by (Emanuellson et al., 2000), only a very small fraction of the possible divisions were investigated, and the results thus obtained could hardly avoid arbitrariness and might be overestimated. Accordingly, the jackknife test as adopted here is much more objective and rigorous. The overall success rates thus obtained are given in Table 2. For facilitating comparison, the rates obtained just by the gene expression vector are also listed in the same table. From Table 2 we can see the following. (1) The rates just based on the gene expression is so poor which indicate that the relationship of TF-Target in expression is in a too complex form to be captured. As a case study, pearson correlation of expression profiles (Spellman PT et al, 1998) for Fhl1, Rap1, Yap5 and RPS21B was calculated, where Fhl1, Rap1, Yap5 are the regulator of RPS21B (Lee et al, 2002). Figure 1 shows that no significant linear correlations were found among these TF-Targets., which indicate that too simple a mathematical model do not have the ability to capture TF-Target relationship in gene expression level In further we will do our effort to develop a new method which is better than NN when capturing these complex relationship. (2) The overall success rates obtained by the current approach, which has combined the gene product and gene expression information, are very high, indicating that the target of a transcription factor is closely related to its gene product and gene expression information. (3) The rates based on the hybrid information are much higher than those just based on the gene expression information, which means that GO information is helpful in deciphering TF-Target relationships.

V. CONCLUSION

From both the rationality of testing procedure and the success rates of test results, hybridization of the gene ontology approach and the gene expression approach can significantly improve the prediction quality of TF-target. This is fully consistent with the scientific logic because the current hybrid approach has combined the gene product and expression information. The gene product is closely correlated with the biological process, molecular function, and cellular component; while the gene expression data is closely correlated with the gene regulation. The introduction of the nearest neighboring algorithm, i.e. NN predictor, can make allowance for bringing out the best in one another and making one shining more brilliantly in the others' company. It has not escaped our notice that the hybridization approach can also be used to improve the prediction quality for other gene network attributes, such as protein-protein network and the metabolic network.

Table 1. Breakdown of the TF-target pairs defined in the hybridization space of gene ontology, and gene expression

Dataset	3860D GO_compress space	160D gene expression space	Total
TF-target	166	9	175
Non TF-target	3377	79	3456

Table 2. Comparison of the predictor performances

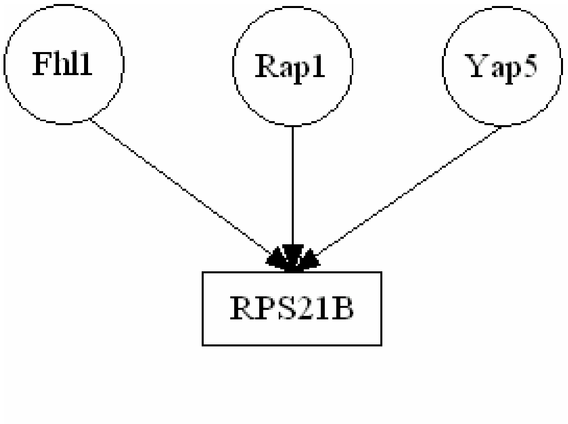
	3860D GO space	160D gene expression space	hybrid space	Just in 160D gene expression space
Overall	3447/3543=97.3%	79/88=89.8%	3526/3631=97.1%	3415/3631=94.1%
TF-target	99/166=59.6%	3/9=33.3%	102/175=58.3%	48/175=27.4%
Non-TF-target	3348/3377=99.1%	76/79=96.2%	3424/3456=99.1%	3367/3456=97.4%

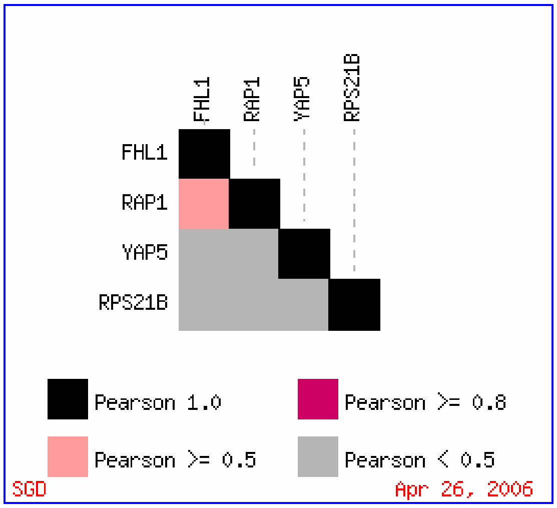
Figures

Figure-1. Pearson correlations between TF-Targets

Up, Transcription regulatory relationship adpted from (Lee et al, 2002).

Down, Pearson correlations of these genes adopted from SGD (http://db.yeastgenome.org/). No significant linear correlation was found between TF-Targets.





REFERENCES

Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D. R., Durbin, R., Falquet, L., Fleischmann, W., Gouzy, L., Hermjakob, H., Hulo, N., Jonassen, I., Kahn, D., Kanapin, A., Karavidopoulou, Y., Lopez, R., Marx, B., Mulder, N. J., Oinn, T. M., Pagni, M., Servant, F., Sigrist, C. J. A. & Zdobnov, E. M. (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Research* 29, 37-40.

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

Banerjee, N. and Zhang, M.Q. (2002) Functional genomics as applied to mapping transcription regulatory networks, *Curr Opin Microbiol*, **5**, 313-317.

Chou, K. C. & Zhang, C. T. (1995). Review: Prediction of protein structural classes. *Critical Reviews in Biochemistry and Molecular Biology* 30, 275-349.

Cora, D., Di Cunto, F., Provero, P., Silengo, L. and Caselle, M. (2004) Computational identification of transcription factor binding sites by functional analysis of sets of genes sharing overrepresented upstream motifs, *BMC Bioinformatics*, **5**, 57.

Cover, T. M. & Hart, P. E. (1967). Nearest neighbour pattern classification. *IEEE Transaction on Information Theory* IT-13, 21-27.

Haverty, P.M., Hansen, U. and Weng, Z. (2004) Computational inference of transcriptional regulatory networks from expression profiling and transcription factor binding site identification, *Nucleic Acids Res*, **32**, 179-188.

Kel, A.E., Gossling, E., Reuter, I., Cheremushkin, E., Kel-Margoulis, O.V. and Wingender, E. (2003) MATCH: A tool for searching transcription factor binding sites in DNA sequences, *Nucleic Acids Res*, **31**, 3576-3579.

Long, F., Liu, H., Hahn, C., Sumazin, P., Zhang, M.Q. and Zilberstein, A. (2004) Genome-wide prediction and analysis of function-specific transcription factor binding sites, *In Silico Biol*, **4**, 395-410.

Matys, V., Fricke, E., Geffers, R., Gossling, E., Haubrock, M., Hehl, R., Hornischer, K., Karas, D., Kel, A.E., Kel-Margoulis, O.V., Kloos, D.U., Land, S., Lewicki-Potapov, B., Michael, H., Munch, R., Reuter, I., Rotert, S., Saxel, H., Scheer, M., Thiele, S. and Wingender, E. (2003) TRANSFAC: transcriptional regulation, from patterns to profiles, *Nucleic Acids Res*, **31**, 374-378.

Orlando, V. (2000) Mapping chromosomal proteins in vivo by formaldehyde-crosslinked-chromatin immunoprecipitation, *Trends Biochem Sci*, **25**, 99-104.

Qian, J., Dolled-Filhart, M., Lin, J., Yu, H. and Gerstein, M. (2001) Beyond synexpression relationships: local clustering of time-shifted and inverted gene expression profiles identifies new, biologically relevant interactions, *J Mol Biol*, **314**, 1053-1066.

Qian, J., Lin, J., Luscombe, N.M., Yu, H. and Gerstein, M. (2003) Prediction of regulatory networks: genome-wide identification of transcription factor targets from gene expression data, *Bioinformatics*, **19**, 1917-1926.

Qiu, P. (2003) Recent advances in computational promoter analysis in understanding the transcriptional regulatory network, *Biochem Biophys Res Commun*, **309**, 495-501.

Ren, B., Robert, F., Wyrick, J.J., Aparicio, O., Jennings, E.G., Simon, I., Zeitlinger, J., Schreiber, J., Hannett, N., Kanin, E., Volkert, T.L., Wilson, C.J., Bell, S.P. and Young, R.A. (2000) Genome-wide location and function of DNA binding proteins, *Science*, **290**, 2306-2309.

Wang, H., Tang, W., Zhu, C. and Perry, S.E. (2002) A chromatin immunoprecipitation (ChIP) approach to isolate genes regulated by AGL15, a MADS domain protein that preferentially accumulates in embryos, *Plant J*, **32**, 831-843.

Zhou, G. P. & Assa-Munt, N. (2001). Some insights into protein structural class prediction. *PROTEINS: Structure, Function, and Genetics* 44, 57-59.

Zhu, Z., Pilpel, Y. and Church, G.M. (2002) Computational identification of transcription factor binding sites via a transcription-factor-centric clustering (TFCC) algorithm, *J Mol Biol*, **318**, 71-81.

Lee, Tong Ihn, Rinaldi, Nicola J., Robert, Francois, Odom, Duncan T., Bar-Joseph, Ziv, Gerber, Georg K., Hannett, Nancy M., Harbison, Christopher T., Thompson, Craig M., Simon, Itamar, Zeitlinger, Julia, Jennings, Ezra G., Murray, Heather L., Gordon, D. Benjamin, Ren, Bing, Wyrick, John J., Tagne, Jean-Bosco, Volkert, Thomas L., Fraenkel, Ernest, Gifford, David K., Young, Richard A. Transcriptional Regulatory Networks in Saccharomyces cerevisiae *Science* 2002 298: 799-804

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B (1998) Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 9(12):3273-97