

Integrating mRNA Decay Information into Co-Regulation Study

Liang Chen¹ and Hong-Yu Zhao²

¹Department of Molecular, Cellular, and Developmental Biology, Yale University, New Haven, CT 06520, U.S.A.

²Department of Epidemiology and Public Health, Yale University, New Haven, CT 06520, U.S.A.

E-mail: liang.chen@yale.edu; hongyu.zhao@yale.edu

Revised November 4, 2004.

Abstract Absolute or relative transcript amounts measured through high-throughput technologies (e.g., microarrays) are now commonly used in bioinformatics analysis, such as gene clustering and DNA binding motif finding. However, transcription rates that represent mRNA synthesis may be more relevant in these analyses. Because transcription rates are not equivalent to transcript amounts unless the mRNA degradation rates as well as other factors that affect transcript amount are identical across different genes, the use of transcription rates in bioinformatics analysis may lead to a better description of the relationships among genes and better identification of genomic signals. In this article, we propose to use experimentally measured mRNA decay rates and mRNA transcript amounts to jointly infer transcription rates, and then use the inferred transcription rates in downstream analyses. For gene expression similarity analysis, we find that there tends to be higher correlations among co-regulated genes when transcription-rate-based correlations are used compared to those based on transcript amounts. In the context of identifying DNA binding motifs, using inferred transcription rates leads to more significant findings than those based on transcript amounts. These analyses suggest that the incorporation of mRNA decay rates and the use of the inferred transcription rates can facilitate the study of gene regulations and the reconstruction of transcriptional regulatory networks.

Keywords mRNA decay rate, transcription rate, transcript amount, DNA binding motif, microarray

1 Introduction

Transcription initiation and mRNA degradation are important steps in gene expression regulation. Together with other factors (e.g., those involved in mRNA splicing), they determine the transcript amount for each gene in the genome. To date, there is generally more focus on the transcription initiation process than the mRNA decay process. However, in the study of transcription initiation, transcript amounts are commonly used to infer the initiation regulation process, although transcription rates may be more relevant. For example, in gene expression data analyses based on microarrays, the observed absolute/relative gene expression levels are routinely used although transcription rates directly related to RNA synthesis may be of the ultimate interest and relevance in most analyses. For example, gene clusters are mostly constructed based on gene-expression-based similarity measures^[1]. Gene expression levels have also been used in combination with other information, e.g., DNA-protein binding, to infer transcriptional regulatory networks^[2,3]. In addition, gene expression profiles have also been used to identify transcription factor binding motifs under the assumption that genes with similar expression levels are under similar regulatory controls. One approach is to cluster genes according to their expression profiles and to search for over-represented short words (motifs) in the upstream regions of those genes in the same cluster^[4]. The other approach directly models gene expression levels as a function of the occurrence of

candidate DNA binding motifs in the upstream regions. For example, REDUCE (Regulatory Element Detection Using Correlation with Expression) assumes a linear relationship between gene expression and the number of functional motifs to identify transcription factor binding motifs^[5]. In all these analyses, especially the identification of DNA binding motifs, transcription rates, not transcript amounts, are more relevant to the objectives of the analyses. Therefore, it seems more appropriate and informative to use transcription rates in all the above-mentioned analyses.

Although high-throughput technology is not available yet to directly measure transcription rates, they may be inferred indirectly from two data sources that are currently available: time-course gene expression levels and mRNA decay rates^[6,7]. In this article, we propose to a simple method to infer transcription rates based on these two data types, and illustrate the potential benefit of using these inferred rates in various bioinformatics analyses, including gene regulation analysis and DNA binding motif identification.

2 Methods and Results

2.1 Data Sources: Data on Expression Levels and Decay Rates

In the present study, we use Cho *et al.*'s yeast cell cycle data set on gene expression levels^[8]. This dataset includes measured gene expression levels at 17 time points for synchronized yeast cells. The RNA expression is

*

This work was supported by NSF under Grant No.DMS 0241160.

measured by Affymetrix chips. Cell cycle arrested yeast cells are synchronized by being held at a restrictive temperature and then released at the same time.

For mRNA decay rates, we use the data from Wang *et al.*^[6], who measured the decay rate of almost all yeast mRNAs by using a temperature-sensitive RNA polymerase II strain. The polymerase II activity is blocked by temperature shifting at time 0, and mRNA amounts at different time points are measured. The half-life of each mRNA is estimated by the mRNA decay profiling analysis. The half-lives of mRNAs in yeast range from about 3 minutes to more than 90 minutes.

In this article, instead of analyzing all the genes in the yeast genome, we focus on a subset of genes that have substantial presence and appreciable variations across the cell cycle. These genes are selected based on the following criteria: 1) the maximum intensity across all time points is larger than 1,000; 2) the ratio of the maximum intensity to the minimum intensity is larger than 2; and 3) the gene's decay rate is detected by the experiment. A total of 745 genes satisfy all three criteria.

2.2 Inference of the Transcription Rates

To estimate the transcription rates, we use the following model to combine the observed expression data (transcript amount), the mRNA decay data, and the (unobservable) transcription rates,

$$\frac{dm}{dt} = K_p - K_d m, \quad (1)$$

where K_p is the transcription rate that needs to be inferred during the time interval Δt , K_d is the mRNA decay rate, m_0 is the transcript amount at time 0, and m is the transcript amount at time t . Here we assume a well-mixed system and negligible stochastic effects. K_p is assumed to be constant during the time interval Δt . For different time intervals, K_p may be different. We assume that K_d is constant across all the time intervals, and exponential decay. From (1), we can estimate K_p for the time interval Δt ,

$$K_p = \frac{K_d(m - m_0 e^{-K_d \Delta t})}{1 - e^{-K_d \Delta t}}. \quad (2)$$

We note that there may be a time delay for the estimated K_p because of the mRNA splicing process and the mRNA nucleo-cytoplasmic transport process. However, we still use the above equation because of the difficulty of estimating time delays.

2.3 Gene Co-Regulation Study

We first compare similarity analyses for genes under common regulation based on the observed transcript amounts and the inferred transcription rates. To identify genes under common regulation, we use a

database assembled by Young and colleagues for regulation targets of yeast transcription factors maintained at http://web.wi.mit.edu/young/regulatory_network. Among the 103 regulators in the database, 51 regulators have more than two target genes in our filtered dataset.

The transcription rates are estimated by the measured gene expression levels and mRNA decay rates through equation^[2]. The pair-wise transcript amount correlations and the pair-wise transcription rate correlations among genes in each regulator group are then calculated. Genes having both transcript amount correlations and transcription rate correlations less than 0.5 with more than 80% other genes in the same group are removed from the analysis. This is because such genes are likely to be mostly regulated through other transcription factors or through the interactions between this regulator and other regulators. This filtering results in 16 regulator groups having more than 5 genes within each group.

In Fig.1, we plot the transcript amount correlations and the transcription rate correlations for these 16 groups. For the majority of the groups, the correlations based on the transcription rates are higher than those based on the transcript amounts. This suggests that correlations between transcription rates better represent the co-regulation relationships.

In order to test the significance of the difference between the transcription rate correlations and the transcript amount correlations, a hypergeometric test is performed and the results are summarized in Table 1. Among the 745 filtered genes, there are 152,583 pairs with transcription rate correlations larger than that based on the transcript amount and 124,557 pairs with transcription rate correlations smaller than that based on the transcript amount. The statistical significance is assessed through the hypergeometric test.

Table 1. Hypergeometric Tests for the 16 Regulator Groups

Regulator	Number of pairs with larger correlations	Total number of pairs	P-value for hypergeometric test	Average of the transcription rate correlations	Average of the transcript amount correlations
TYE7	9	10	0.003	0.864	0.801
GCR1	61	91	0.008	0.866	0.843
HSF1	15	21	0.039	0.637	0.551
HAP3	11	15	0.043	0.570	0.471
GCR2	19	28	0.058	0.831	0.811
HAP4	7	10	0.100	0.553	0.531
MOT3	7	10	0.100	0.332	0.238
YAP1	18	28	0.120	0.639	0.616
ABF1	22	36	0.185	0.759	0.738
MSN4	6	10	0.267	0.675	0.505
RLM1	34	66	0.676	0.521	0.527
HAP2	4	10	0.740	0.606	0.624
RAP1	24	55	0.941	0.859	0.868
LEU3	7	21	0.963	0.583	0.693
GLN3	3	15	0.994	0.350	0.532
CBF1	1	10	0.996	0.365	0.563

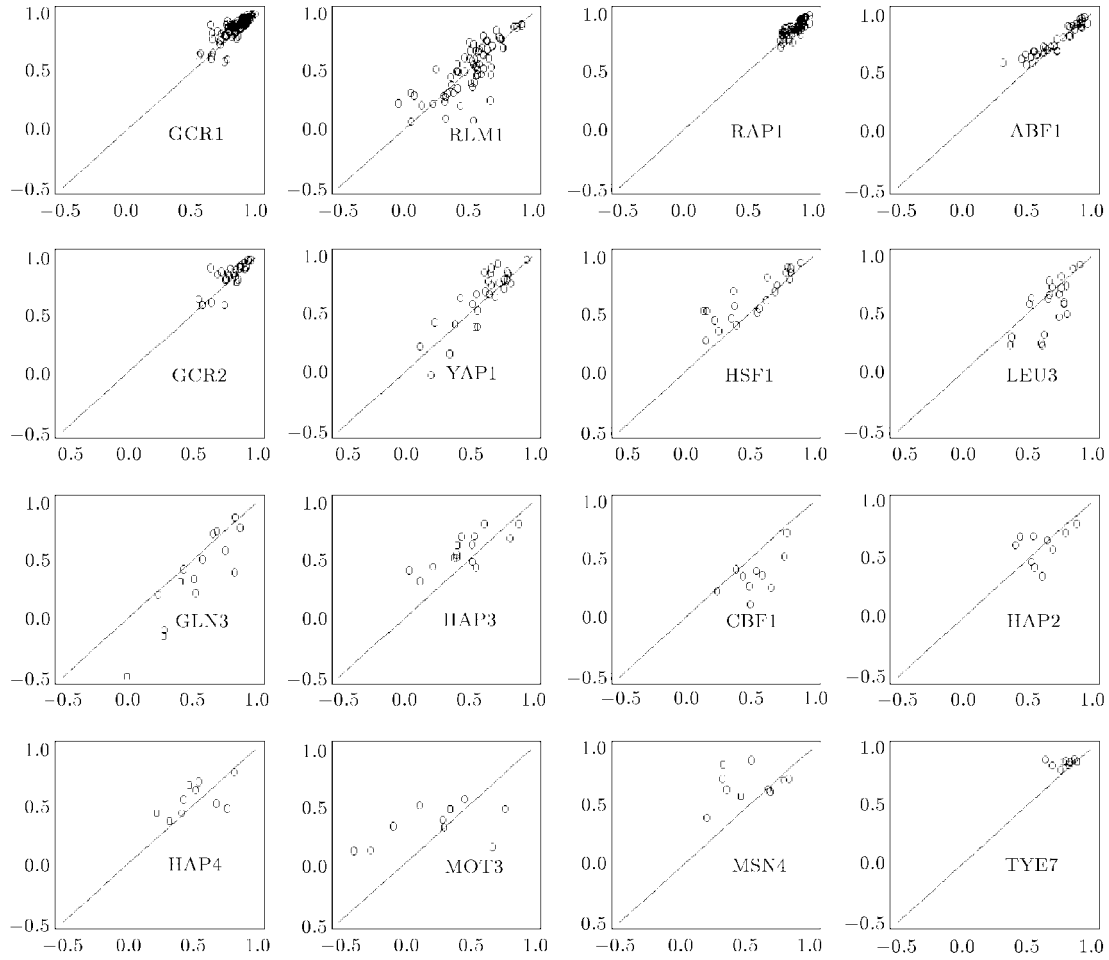


Fig.1. Transcription rate correlations versus transcript amount correlations in the same regulator group. In this figure, Y-axis shows the correlations between transcription rates and X-axis shows the correlations between transcript amounts. The solid line is $y = x$, and the regulators' names are shown on each plot.

Table 2. Hypergeometric Tests for the ChIP-Chip Binding Data (using a p -value threshold 0.001 to infer the binding)

Common regulators	Number of pairs with larger correlations	Total number of pairs	Pval for hypergeometric test	Average of the transcription rates correlations	Average of the transcripts amounts correlations
0	145,412	265,040	1.0	0.409	0.381
1	5,446	9,074	3.1×10^{-11}	0.610	0.588
2	1,607	2,763	4.4×10^{-4}	0.755	0.736
3	110	243	1.0	0.741	0.749
4+	8	15	0.45	0.582	0.645

Table 3. Hypergeometric Tests for the ChIP-Chip Binding Data (using a p -value threshold 0.005 to infer the binding)

Common regulators	Number of pairs with larger correlations	Total number of pairs	Pval for hypergeometric test	Average of the transcription rates correlations	Average of the transcripts amounts correlations
0	140,785	256,243	1.0	0.408	0.379
1	7,756	14,066	0.42	0.487	0.469
2	3,306	5,364	7.0×10^{-10}	0.718	0.693
3	514	969	0.89	0.741	0.734
4+	222	348	3.9×10^{-4}	0.757	0.758

Table 1 lists the p -value for the hypergeometric test for each regulator group. The table also lists the average transcription rate correlation and average transcript

amount correlation for each group. The p -values for the first 5 groups are significant or marginally significant.

In addition to using the online database, we also per-

form a similar study based solely on Young's ChIP-chip binding data^[2]. The gene pairs are grouped according to their common regulators. We use a threshold of p -value of 0.001 or 0.005 to infer DNA-protein binding. The hypergeometric tests are also performed to compare the correlations based on transcript amounts and those based on transcript rates.

The results are summarized in Tables 2 and 3. When a p -value threshold of 0.001 is used to infer DNA-protein binding, the correlations for groups of genes sharing 1 or 2 common regulators significantly increase after taking the decay information into account. When a p -value threshold of 0.005 is used to infer DNA-protein binding, the correlations for groups of genes sharing two common regulators significantly increase after taking the decay information into account. Different thresholds have different biological meanings. For the 0.001 threshold value, the criterion is very stringent, so the inference about the group with one or two common regulator is most accurate, and there is statistically significant evidence to show increased correlation for these groups. However, when a less stringent criterion is used, there are likely more false positive results in the inferred binding data, resulting in less significant effect for genes sharing one common regulator.

2.4 Transcription Factors' Binding Motif Study

As discussed in the Introduction section, there are, in general, two different computational strategies to identify DNA binding motifs based on gene expression data:

1) methods based on gene clusters and 2) those based on regression analysis. In this article, we consider the second approach, i.e., regression analysis. In addition to using transcript amounts, inferred transcription rates are also used in our analysis. All of the genes with measured decay rates (4,517 gene totally) are used. The non-cell-cycle related genes are also included to better infer the functional motifs. For each gene, the upstream 600bp sequence is used to evaluate the number of occurrence of candidate motifs. Table 4 summarizes the results. It can be seen that the motif results are similar between those based on the absolute transcript amounts and those based on the inferred transcription rates. However, more significant results are obtained when the inferred transcription rates are used in the regression analysis.

3 Discussion

In the study of transcription regulations, transcription rates, instead of transcript amounts, may be more biologically relevant. However, absolute or relative transcript amounts are commonly used in many studies to form gene clusters, identify DNA binding motifs, and reconstruct transcriptional regulatory networks. In this article, we have proposed a simple method to incorporate measured gene expression levels and mRNA decay rates to infer transcription rates. To illustrate the usefulness of using inferred transcription rates in genomic analyses, we have considered two types of analyses: coregulation studies and DNA binding motif identification.

Table 4. REDUCE Results for Time Points 1 to 8 of the Cell Cycle Data (We use the absolute transcript amounts and the inferred transcription rates in the regression analysis separately. The first 5 motifs are listed for each time point or time interval. $\Delta\chi^2$ represents a measure of motif significance to which a confidence or P value can be assigned. A larger $\Delta\chi^2$ corresponds to a smaller P value.)

Transcript amount		Transcription rate		Transcript amount		Transcription rate	
Time point 1	$\Delta\chi^2$	Time interval 0-1	$\Delta\chi^2$	Time point 2	$\Delta\chi^2$	Time interval 1-2	$\Delta\chi^2$
AAAATTT	0.031	AAAATTT	0.064	AAAATTT	0.026	AAAATTT	0.033
CCGTACA	0.016	CGATGAG	0.009	CCGTACA	0.020	CGCG	0.027
AAAATT	0.020	AAAATT	0.032	AAAATT	0.018	AAAATT	0.024
ATCCGTA	0.015	CTCATC	0.012	ATCCGTA	0.016	ACGCG	0.024
AAAAATT	0.019	AAAAATT	0.031	AAAAATT	0.016	AAAAATT	0.020
Time point 3	$\Delta\chi^2$	Time interval 2-3	$\Delta\chi^2$	Time point 4	$\Delta\chi^2$	Time interval 3-4	$\Delta\chi^2$
AAAATTT	0.025	AAAATTT	0.032	AAAATTT	0.026	AAAATTT	0.040
ATCCGTA	0.016	CG	0.012	CCATACA	0.019	CCATACA	0.012
AAAATT	0.017	AAAATT	0.021	AAAATT	0.018	AAAATT	0.025
CCGTACA	0.016	CGC	0.011	CCGTACA	0.017	CG	0.012
AAAAATT	0.015	AAAAATT	0.019	AAAAATT	0.016	AAAAATT	0.027
Time point 5	$\Delta\chi^2$	Time interval 4-5	$\Delta\chi^2$	Time point 6	$\Delta\chi^2$	Time interval 5-6	$\Delta\chi^2$
AAAATTT	0.027	AAAATTT	0.020	AAAATTT	0.027	AAAATTT	0.044
CCATACA	0.014	GATGAGC	0.006	CCATACA	0.015	CCCGGGC	0.009
AAAATT	0.017	AAAAATT	0.008	AAAATT	0.019	AAAATT	0.029
CCGTACA	0.013	CGATGAG	0.007	CCGTACA	0.014	CGGGCGA	0.009
AAAAATT	0.016	AAATTT	0.006	AAAAATT	0.018	AAAAATT	0.026
Time point 7	$\Delta\chi^2$	Time interval 6-7	$\Delta\chi^2$	Time point 8	$\Delta\chi^2$	Time interval 7-8	$\Delta\chi^2$
AAAATTT	0.027	AAAATTT	0.041	AAAATTT	0.029	AAAATTT	0.037
CCGTACA	0.014	CAGGCCG	0.010	ATCCGTA	0.015	TGCGAA	0.006
AAAATT	0.018	AAAATT	0.023	AAAATT	0.019	AAAATT	0.020
ATCCGTA	0.013	CG	0.010	CCGTACA	0.015	CGATGAG	0.007
AAAAATT	0.017	AAAAATT	0.028	AAAAATT	0.018	AAAAATT	0.016

When the inferred transcription rates are used, we tended to observe higher similarity among genes under common regulations compared to those based on gene expression levels. In the context of identifying DNA binding motifs through regression analysis, we obtained more significant results when the inferred transcription rates were used. In principle, we can also use the inferred transcription rates (instead of transcript amounts) to cluster genes first and search for motifs in the upstream regions of the genes in the same cluster using established methods, such as AlignACE^[4]. Despite promising results we have obtained in this article, we must note that the mRNA decay rates are measured with much noise and the decay rates are likely to be strain and condition dependent. For example, the measured decay rates of temperature-sensitive RNA polymerase II strain may be different from a wild type strain, and the decay rates differ according to the environmental stimuli in diverse organism ranging from yeast to humans^[9–11]. In fact, the two data sources used in our analyses were collected by two different research groups. It is obvious that the more accurate and relevant of the measured decay rates are, the more accurate we can infer transcription rates.

Nowadays, the fluctuation of mRNA degradation has been recognized as an essential step to change gene expression pattern in response to developmental or environmental stimuli or environmental stresses. Furthermore, in carcinogenesis, some genes are up-regulated by the deregulated mRNA stability. For example, COX2 is over-expressed in colon carcinogenesis with the change of mRNA degradation^[12]. With a better understanding of the transcription initiation, we can distinguish the genes with changed mRNA degradation from the genes with changed transcription rates among the differentially expressed genes in tumor samples, which will benefit the therapeutic designs.

As a gene-silencing technology, RNA interference (RNAi) has been used successfully to investigate gene function. In RNAi, double-stranded RNA is introduced to cells and induces a sequence-specific RNA degradation mechanism that effectively silences a targeted gene. With a better understanding of the transcription initiation, we can better understand the RNAi and evaluate the effects of RNAi technology.

Although there are many caveats in the integrated analysis of gene expression data and mRNA decay data, before molecular technologies can be developed to directly measure transcription rates, our proposed method that incorporates both gene expression levels and mRNA decay rates may lead to better characterization of transcription regulations.

Acknowledge We thank two reviewers for their helpful comments.

References

- [1] Eisen M B, Spellman P T, Brown P O *et al.* Cluster analysis and display of genome-wide expression patterns. In *Proc. Natl. Acad. Sci., U.S.A.*, Dec. 8, 1998, 95(25): 14863–14868.
- [2] Bar-Joseph Z, Gerber G K, Lee T I *et al.* Computational discovery of gene modules and regulatory networks. *Nat. Biotechnol.*, Nov. 2003, 21(11): 1337–1342.
- [3] Lee T I, Rinaldi N J, Robert F *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science*, Oct. 25, 2002, 298(5594): 799–804.
- [4] Roth F P, Hughes J D, Estep P W, Church G M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.*, Oct. 1998, 16(10): 939–945.
- [5] Bussemaker H J, Li H, Siggia E D. Regulatory element detection using correlation with expression. *Nat. Genet.*, Feb. 2001, 27(2): 167–171.
- [6] Wang Y, Liu C L, Storey J D *et al.* Precision and functional specificity in mRNA decay. In *Proc. Natl. Acad. Sci., U.S.A.* Apr. 30, 2002, 99(9): 5860–5865.
- [7] Lam L T, Pickeral O K, Peng A C *et al.* Genomic-scale measurement of mRNA turnover and the mechanisms of action of the anti-cancer drug flavopiridol. *Genome Biol.*, 2001, 2(10): RESEARCH0041.
- [8] Cho R J, Campbell M J, Winzler E A *et al.* A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell.*, July 1998, 2(1): 65–73.
- [9] Malter J S. Posttranscriptional regulation of mRNAs important in T cell function. *Adv. Immunol.*, 1998, 68: 1–49.
- [10] Ross J. mRNA stability in mammalian cells. *Microbiol Rev.*, Sept. 1995, 59(3): 423–450.
- [11] Raghavan A, Ogilvie R L, Reilly C *et al.* Genome-wide analysis of mRNA decay in resting and activated primary human T lymphocytes. *Nucleic Acids Res.*, Dec. 15 2002, 30(24): 5529–5538.
- [12] Dixon D A, Kaplan C D, McIntyre T M *et al.* Post-transcriptional control of cyclooxygenase-2 gene expression. The role of the 3'-untranslated region. *J. Biol. Chem.*, Apr. 21 2000, 275(16): 11750–11757.



Liang Chen obtained her B.S. degree from Department of Biological Sciences and Biotechnology, Tsinghua University, Beijing, China in 2001. Now she is with the Department of Molecular, Cellular and Developmental Biology, Yale University, New Haven, Connecticut.



Hong-Yu Zhao obtained the B.S. degree from Department of Probability and Statistics, Peking University, China in 1990 and the Ph.D. degree from Department of Statistics, University of California at Berkeley, in 1995. From 1995–1996 he was adjunct assistant professor and assistant professor in residence in University of California at Los Angeles. From 1996–present he is assistant professor, associate professor, and Ira V. Hiscock associate professor in Yale University. He is associate editor of *Biometrics*; *Journal of Agricultural, Biological, and Environmental Statistics*; *Statistical Applications in Genetics and Molecular Biology*; *Pharmacogenomics*.