How much expression divergence after yeast gene duplication could be explained by regulatory motif evolution?

Zhongqi Zhang, Jianying Gu and Xun Gu

Department of Genetics, Development & Cell Biology, Center for Bioinformatics and Biological Statistics, Iowa State University, IA 50011, USA

We used the yeast genome sequences of gene families, microarray profiles and regulatory motif data to test the current wisdom that there is a strong correlation between regulatory motif structure and gene expression profile. Our results suggest that duplicate genes tend to be co-expressed but the correlation between motif content and expression similarity is generally poor, only ~2–3% of expression variation can be explained by the motif divergence. Our observations suggest that, in addition to the cis-regulatory motif structure in the upstream region of the gene, multiple trans-acting factors in the gene network can influence the pattern of gene expression significantly.

Predicting the transcriptional regulation network from genomic data is a major challenge [1–5]. Many computational methods are based on the expression-motif-conservation hypothesis that co-expressed genes are likely to share regulatory motifs that are conserved during evolution [6–14]. Several studies have tested this idea [15,16]. For instance, Yu et al. [15] found that (evolutionarily unrelated) yeast gene pairs targeted by shared transcription factors (TFs, also known as target gene pairs) are co-expressed more frequently than expected by chance. One would expect a similar co-expression between duplicated genes because they usually retain similar functions or regulatory patterns.

We used Saccharomyces cerevisiae gene families to test this prediction and examined 202 yeast two-member gene families (Box 1). Using methods modified from Yu et al. [15], we have shown that 7% of duplicate gene pairs are co-expressed (Box 1), which is approximately seven times greater than random expectation (P-value = 1.9 × 10^{-5}). This can be compared with the results of Yu et al. [15] that ‘overall, 3.3% TF-target gene pairs are co-expressed, which is four times greater than random expectation.’ Our analysis indicates that duplicate gene pairs might have an even higher co-expression level compared with the TF-target gene pairs. Thus, we thought it was worthwhile exploring a new motif-search strategy by combining co-expressed cluster analysis and gene family phylogeny to increase the signal-to-noise ratio.

References
6. Dupuy, D. et al. (2002) SCAN domain-containing 2 gene (SCAND2) is a novel nuclear protein derived from the zinc finger family by exon shuffling. Gene 289, 1–6

0168-9525/$ - see front matter © 2004 Elsevier Ltd. All rights reserved.
doi:10.1016/j.tig.2004.06.013
Box 1. Computational analysis for yeast genome datasets

Datasets
(i) The complete sequences for 43 genomes of bacteria, archaea and Saccharomyces cerevisiae are available at NCBI (http://www.ncbi.nlm.nih.gov/COG/), where gene families are classified as clusters of orthologous groups (COG). Amino-acid alignments of 202 COGs containing two yeast duplicate genes were downloaded.
(ii) The whole genome sequence of S. cerevisiae was downloaded from NCBI (http://www.ncbi.nlm.nih.gov/Ftp/index.html). The upstream regions (500 bp) of each yeast gene in the COG family were extracted. According to the list of 50 ‘known’ regulatory motifs that were compiled by Kellis et al. [14] and verified by experimentation (among those motifs, 50 are unique motif sequences), we used Perl to scan along the upstream region of each yeast gene to identify all of the motifs on both strands (Figure I). Kellis et al. had also predicted 72 ‘discovered’ motifs, including 30 known motifs. The motif-expression (predicted) correlation is lower because of the apparent prediction errors (not shown).
(iii) A total of 276 microarray data points were collected from the online database [18–22]. As commonly suggested, we used the fold-change after normalization to representing the gene expression level. Because we are focused on the expression divergence between duplication pairs, potential cross hybridization [35] might give us a more conserved result.
(iv) Chromatin immunoprecipitation (ChIP) data [26] were downloaded from the Young laboratory homepage http://web.wi.mit.edu/young/regulator_network/. Each yeast gene in the 202 families was used as target gene to retrieve its corresponding transcription factor(s) from the database.

Association study for co-expression pattern of duplicates
The statistical significance of co-expression (top 1% of largest correlation coefficients of all possible gene pairs in the genome) between duplicate genes is calculated using a cumulative binomial distribution:

\[ P_{c\leq c_0} = \sum_{c=c_0}^N \binom{N}{c} p^c (1-p)^{N-c} \]

(Eqn I)

where \( N \) is the total number of gene pairs of interest, \( c_0 \) is the number of co-expressed gene pairs and \( p \) is the probability of finding a co-expressed gene pair within the whole genome. (For more information, see Ref. [15].)

Weighted correlation
The model for motif (M)-expression (E) regression can be written as:

\[ E = M_\beta + \alpha, M_\beta = \sum_{k=1}^2 \text{min}(c_{i_k}, c_{j_k}) \]

(Eqn II)

where \( c_{i_k}, c_{j_k} \) is the copy number of motif \( k \) in gene 1 (gene 2) of the duplicate pair. In particular, it becomes an unweighted correlation if \( c_{i_k} = 1 \) when motif \( k \) is found in the upstream region of a gene \( i \) or if \( c_{i_k} = 0 \) when a motif \( k \) is not found.

Identification of potential paralogous motifs
For each motif \( i \), its position in the upstream sequence of gene \( X \) is denoted by \( L_{i_X} \). Two motifs located upstream of each duplicate \( X \) or \( Y \) are paralogous if: (i) they are the same type of motifs; (ii) the positions of motifs appearing in the upstream region is 5.0 (±3.6) per gene, whereas the mean number of shared motifs between duplicate genes is 1.8 (±1.2) per pair. In addition, 183 gene families (91%) have at least one motif with more than one copy; on average, each regulatory motif is present in 1.3 copies in the upstream region. Meanwhile, the expression similarity between two duplicates for each gene family is computed based on 276 microarray data points under various experimental conditions [18–22].

According to the expression-motif-conservation hypothesis, a positive correlation is expected between the fraction of shared motifs and the gene expression similarity for duplicate gene pairs because both are negatively correlated with the age of the duplication event [16,17]. To minimize the error caused by falsely predicted TF-binding sites, we searched the upstream regions of duplicate genes for 50 ‘known’ regulatory motifs that were compiled by Kellis et al. [14] and verified by experimentation (Box 1). The mean number
Previous studies showed that multi-copied regulatory motifs are more likely to be active [4,25]. We have considered this effect by using the number of copies as a weighting factor (Box 1), resulting in a slightly higher statistical significance (Pearson correlation: $R = 0.17$, $P$-value $= 0.015$; Spearman rank: $\rho = 0.15$, $P$-value $= 0.032$).

Moreover, the motif-expression correlation becomes more meaningful for ‘paralogous’ motifs between duplicate pairs if the relative position of motifs in the regulatory region is taken into account. Tentatively, we defined two similar regulatory motifs located in each duplicate gene as ‘paralogous’ if: (i) their positions (relative to the transcription site) are close to each other; and (ii) they are in the evolutionary conserved yeast region (Box 1). In total, 93 shared paralogous motifs were detected within 72 gene families. The correlation (Figure 1b) between the fraction of shared paralogous motifs and expression similarity is improved slightly (Pearson correlation: $R = 0.18$, $P$-value $= 0.013$; Spearman rank correlation: $\rho = 0.16$, $P$-value $= 0.025$).

Because many potential TF-binding sites might not be functional, it would be beneficial to use those that are detected by large-scale chromatin immunoprecipitation (ChIP) experiments [26]. This complementary approach of determining functional interactions between TFs and upstream regions is not confounded by somewhat diverged motifs. We tested our result by using the dataset from Lee et al. (106 putative TFs from their ChIP experiment [26]) and found a higher motif-expression correlation (Pearson correlation $R = 0.26$, $P$-value $= 0.002$; Spearman rank correlation: $\rho = 0.26$, $P$-value $= 0.001$).

It would be desirable to exploit the assumption that paralogs possessed the same motif structures and expression profiles soon after gene duplication. To avoid the substantial rate variation of amino-acid distance among genes and the saturation of synonymous distance, we used the relative age of duplicates; the units used are defined by the Escherichia coli–yeast split (Box 1). Indeed, we have observed that young duplicates have stronger correlation between expression similarity and motif similarity, whereas a similar correlation for ancient duplicate pairs disappears. For instance, using 0.9 units as the cutoff, we divided the duplicate gene pairs into two same-size groups and found that the coefficient of motif-expression correlation for the ‘recent’ group is 0.19, whereas we found no correlation for the ‘ancient’ group ($R = 0$). Note that this pattern is insensitive to the cutoff value (not shown). Adjacent pairs of genes often show expression similarity, even without sharing the same set of regulatory elements [27], because local chromatin structure can also modulate gene expression profiles. However, after excluding a few adjacent duplicate pairs in our analysis, we obtained almost the same pattern (not shown).

We have examined whether the expression similarity and the fraction of shared motifs are negatively correlated with the age of gene duplication [16,17,28]. Figure 1c
shows a highly significant negative correlation between the expression correlation and the age of gene duplication (Pearson correlation: $R = -0.42$, $P < 0.0001$; Spearman rank correlation: $\rho = -0.44$, $P < 0.0001$). Similarly, Figure 1d reveals a negative correlation between the fraction of shared motifs and the age of gene duplication (Pearson correlation: $R = -0.22$, $P = 0.001$; Spearman rank correlation: $\rho = -0.19$, $P = 0.008$). Using amino-acid distance as an estimate of the time of divergence gives almost the same results (not shown).

Gene expression is influenced by many trans-factors

One might be puzzled by the discrepancy between the strong co-expression pattern and the relatively weak motif-expression correlation between duplicates. Actually, it can be understood in terms of how much co-expression can be explained by the correlation between expression and motif divergences. Recall that Yu et al. [15] claimed that 3.3% of gene pairs that share the same TF(s) are co-expressed, whereas our analysis showed that 7% of duplicate pairs are co-expressed. Meanwhile, the coefficient of correlation between expression divergence and regulatory motif divergence ($R = 0.15$–0.18) suggests that $\sim 2.3$%–3.2% (i.e. $0.15^2$–$0.18^2$) of the expression variance can be explained by the regulatory motif divergence, even in the case of ChIP, it is only 6.7% ($0.26^2$). Therefore, the interpretation from these two studies is biologically consistent, although the significance level ($P$-value) might differ dramatically because the null hypotheses of two methods, and the sample size, are different. One explanation might involve the non-linear co-expression property or the shared motif combinations rather than individual motifs [16,29] – do the older duplicates have a lower expression similarity because they have lost particular motif combinations? We have studied a more general linear or non-linear (e.g. logit) regression model but we found no considerable improvement, partly because of the small sample size. One solution is to study the motif-expression correlation in the phylogeny of a large gene family, which can be tested by using more-sophisticated statistical methods [30].

Caveats notwithstanding, our analysis shows that experimental error is unlikely to be the only cause for the low explanatory power, supporting the notion that gene expression can be influenced by many trans-factors at different levels of gene networks. In addition to the cis-regulatory motif structure, other factors, such as motif–motif interactions, TF co-evolution and the chromatin structure [29,31,32], can affect the expression levels without changing the regulatory motifs.

The weak motif-expression correlation between duplicates might be the consequence of two phenomena. First, duplicates with divergent motif structures actually have similar expression profiles (e.g. among the ten most highly co-expressed duplicate gene pairs none of which has a fraction of shared motifs > 50%). This could be due to the neutral turnover of binding sites (i.e. randomly generated binding sites might cause the shift of TF from its original
binding site) or convergent evolution (i.e. the binding between different TFs and their corresponding motifs might have a similar impact on gene expression regulation). Second, duplicates with a similar set of regulatory motifs have low expression similarities because of the context-dependence of trans-transcriptional regulation, which requires the combinatorial interaction of other bound proteins to function [29]. Further study is certainly needed when more data are available. However, the evolution of regulatory motifs might not necessarily result in expression divergence because of genetic robustness or because of an alternative regulatory pathway in which the original binding sites were not used. Furthermore, our analysis implies that ancient duplicate pairs might have more chances to be affected by the reshuffling of gene networks. Consequently, the motif-expression correlation might disappear because of the evolutionary reorganization of gene networks, resulting in considerable expression changes even when the same motifs are conserved [29]. In addition, there might be other trans-acting factors (other than the TFs) that have an influence on gene regulation [33].

Concluding remarks
If our interpretation is correct, some evolutionary models need to be revisited because the whole gene network could be involved in shaping the expression divergence after gene duplication. For example, the duplication-degeneration-complementation (DDC) model of gene duplication [34] assumes that the status (presence or absence) of regulatory motifs (or motif modules) dominates the status of gene expression, which seems to be oversimplified according to our analysis. (For more information, see Wagner [28].) The important information conveyed by our results is that the identification of motif structures is only one step towards discovering the mechanism underlying gene expression regulation and evolution, and the effect of gene networks other than the cis-factors should not be neglected.

Acknowledgements
We thank Molley Vanhouten for assistance in collecting the microarray gene expression data. This work was supported by an NIH grant to X.G.

References
1 Ren, B. et al. (2000) Genome-wide location and function of DNA binding proteins. Science 290, 2306–2309