Acknowledgements

I thank M. Nei, I. Rogozin, A. Rooney, and H. Rosenberg for discussions.
This work was partly supported by the NIH and NSF research grants to M. Nei.

References

1 Das, S. *et al.* (1997) Biology's new Rosetta stone. *Nature* 385, 29–30
2 Tatusov, R.L. *et al.* (1997) A genomic perspective on protein families. *Science* 278, 631–637
3 Brown, J.R. and Doolittle, W.F. (1997) *Archaea* and the prokaryote-to-eukaryote transition. *Microbiol. Mol. Biol. Rev.* 61, 456–502
4 Koonin, E.V. *et al.* (1997) Comparison of archaeal and bacterial genomes: computer analysis of protein sequences predicts novel functions and suggests a chimeric origin for the Archaea. *Mol. Microbiol.* 25, 619–637
5 Brown, J.R. *et al.* (1998) A bacterial antibiotic resistance gene with eukaryotic origins. *Curr. Biol.* 8, 365–367
6 Nelson, K.E. *et al.* (1999) Evidence for lateral gene transfer between Archaea and Bacteria from genome sequence of *Thermotoga maritima*. *Nature* 399, 323–329
7 Chervitz, S.A. *et al.* (1998) Comparison of the complete protein sets of worm and yeast: orthology and divergence. *Science* 282, 2022–2028
8 Johnson, N.L. and Kotz, S. (1970) *Distributions in Statistics: Continuous Univariate Distributions*, Houghton Mifflin
9 Duret, L. and Mouchiroud, D. (1999) Expression pattern and, surprisingly, gene length shape codon usage in *Caenorhabditis*, *Drosophila*, and *Arabidopsis*. *Proc. Natl. Acad. Sci. U. S. A.* 96, 4482–4487
10 Nei, M. (1987) *Molecular Evolutionary Genetics*, Columbia University Press
11 Wolfe, K.H. and Sharp, P.M. (1993) Mammalian gene evolution: nucleotide sequence divergence between mouse and rat. *J. Mol. Evol.* 37, 441–456

Regulation of adjacent yeast genes

Semyon Kruglyak
kruglyak@hto.usc.edu

Haixu Tang
tanghx@hto.usc.edu

Department of Mathematics, University of Southern California, Los Angeles, CA, USA.

In the genomes of prokaryotes many cases are known where a single regulatory system controls two or more functionally related genes[1]. Although important differences exist between the regulatory systems of prokaryotes and eukaryotes, it has been suggested that multi-gene regulation also exists in eukaryotic genomes. We explore this notion through the analysis of gene-expression data because observed changes in gene-expression levels quantify the effect of the regulatory mechanisms. The yeast *Saccharomyces cerevisiae* is our model organism because the whole genome has been sequenced, all open reading frames have been identified, and much expression data is available.

In our discussion of the regulatory system, we are referring to the transcription-factor-binding sites that interact with various proteins to regulate gene expression at the transcription level. It is widely accepted that yeast regulatory systems are typically located within several hundred base pairs (bp) upstream of the genes they control[2–4]. Typical searches encompass 800 bp upstream of the start codon[5]. Zhang and Smith have found many functionally related genes that are located near one another in the yeast genome[6] and Cho *et al.*[7] show the existence of adjacent genes the expression of which is initiated in the same phase of the cell cycle. A disproportionate fraction of these genes are transcribed on opposite strands, away from each other[7]. This would suggest that a regulatory system that was located between the two genes could control the expression of both.

This is interesting in itself, and it might lead to increased understanding of the regulatory system. Finding transcription-factor-binding sites[8] and deciphering their interactions[9] are important and difficult problems. The case of a single regulatory system controlling a gene pair is easier to analyse than the general case because the search for transcription-factor-binding sites can be narrowed to the region between the two genes. There are many examples where this region is less than 400 bp in length (see Table 1). In addition, the fact that two genes must be controlled could impose some symmetry on the regulatory system. For example, symmetry might be reflected in the interaction of transcription factors or by the presence of palindromic binding sites. The known binding sites in the region of interest can be identified by resources such as the TRANSFAC database[10].

In this work, we present additional evidence for the existence of multi-gene regulation in yeast, and we give a list of candidate gene pairs that are likely to be controlled in this manner. Our conclusions are based on expression data from cell cycle[7], diauxic shift[11] and sporulation experiments[12]. In these experiments, microarray technology has been used to measure the expression level of every yeast gene at a series of time points, and in each experiment, the time horizon spans one of the biological processes mentioned above. We compare the expression patterns of adjacent genes in these data sets.

The key idea is that two genes that are controlled by a single regulatory system should have similar expression patterns in any data set. We used the correlation coefficient as a measure of similarity to show that the expression patterns of adjacent genes are more often highly correlated than the expression patterns of randomly selected gene pairs. Because the correlation coefficient is highly sensitive to experimental variation in the data, we filtered the data sets to include only genes whose expression values undergo substantial changes during the time course of the experiment. Such genes are informative because their expression patterns have a high signal to noise ratio.

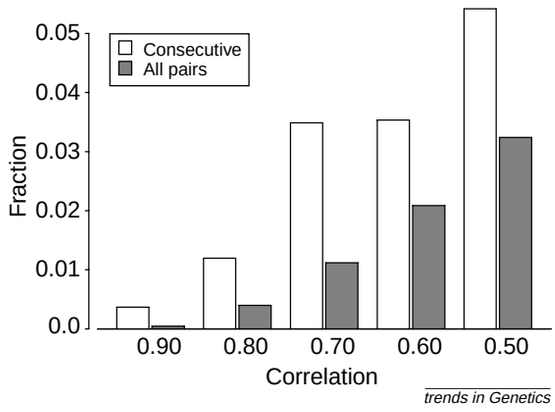
TABLE 1. Candidate pairs for control by a single regulatory system[a]

ORF	Sporulation	Diauxic shift	Cell cycle	Direction	Distance
YAR007C YAR008W	0.82	0.83	0.96	← →	345
YBR052C YBR053C	0.73	0.83	0.90	← ←	322
YDR229W YDR230W	0.76	0.83	0.77	→ →	86
YIL020C YIL019W	0.87	0.85	0.87	← →	282
YJL190C YJL189W	0.73	0.97	0.91	← →	630
YKR024C YKR025W	0.94	0.63	0.86	← →	397
YLL062C YLL061W	0.67	0.81	0.93	← →	342
YNL294C YNL293W	0.80	0.72	0.73	← →	379
YNL263C YNL262W	0.78	0.78	0.67	← →	371
YNL037C YNL036W	0.84	0.81	0.73	← →	811

[a]Candidate gene pairs that had correlated expression patterns in all three data sets. Direction refers to the direction of transcription of each gene in the pair, and distance is the number of base pairs between the genes.

 PII: S0168-9525(99)01941-1

FIGURE 1. Correlation of expression patterns for consecutive gene pairs

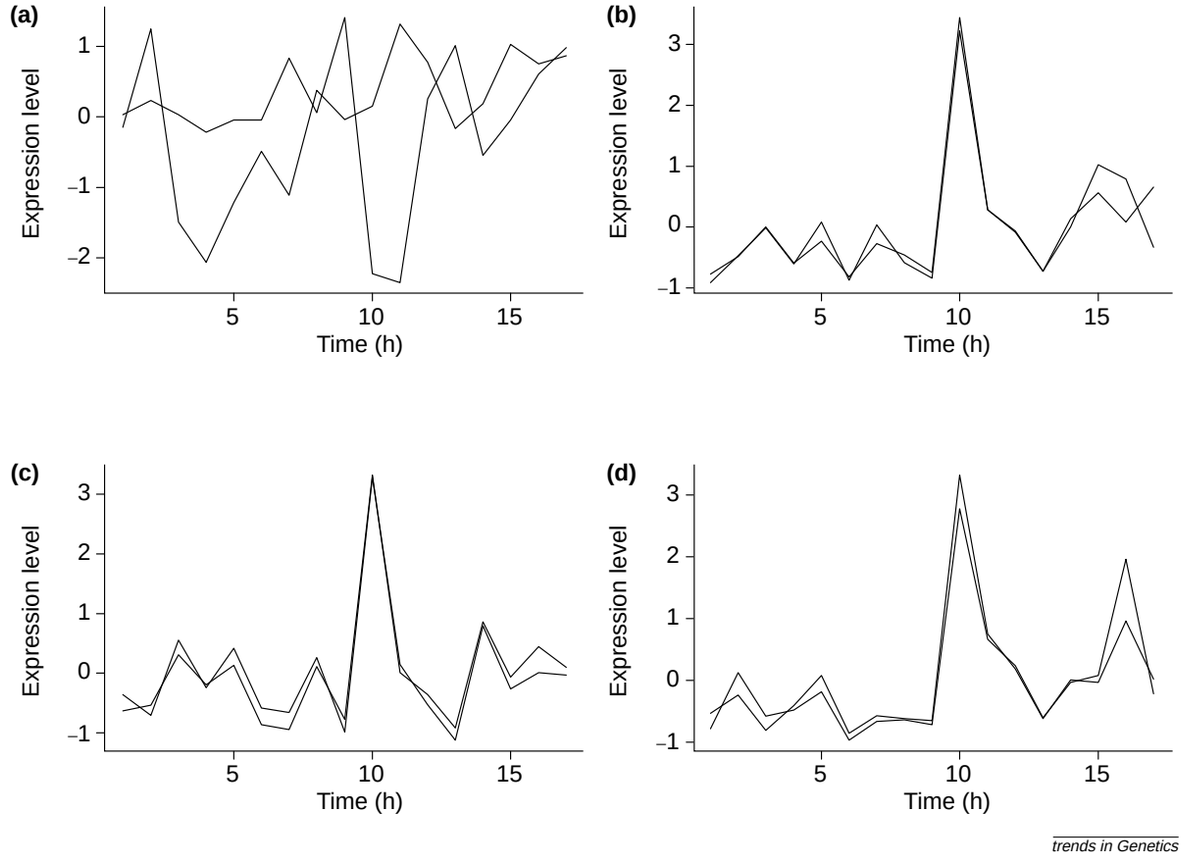


Fraction of highly correlated expression patterns for consecutive gene pairs versus all gene pairs. For example, the bars at 0.80 show that 1.2% of consecutive genes have expression patterns with a correlation between 0.80 and 0.90, compared 0.40% of all gene pairs.

Figure 1 shows a histogram of the fraction of adjacent genes that have a high correlation of expression patterns in the cell cycle data set, versus the fraction of all gene pairs that have high correlations in the same data set. The fraction of adjacent genes that have a correlation of 0.90 or higher is more than eight times greater than the corresponding fraction of all gene pairs. The high correlation between adjacent genes could be due to physical overlap of the ORFs on the chromosome. However, this is unlikely because experimental protocol for microarrays calls for careful probe design that avoids overlapping regions. As a precautionary measure, we have excluded overlapping ORFs from our search results reported in Table 1 and the web page (http://www-hto.usc.edu/~tanghx/yeast-gene-pair.html).

Despite the significant increase in correlation, <10% of adjacent genes have an expression pattern correlation of 0.60 or higher in the cell cycle data set. Visual inspection shows that expression pattern pairs below this correlation level have little in common. The low fraction of adjacent genes with highly correlated expression patterns indicates that proximity is not sufficient for finding gene pairs that are controlled by a single regulatory system. For example, Zhang and Smith[6] note that the genes YKR085C and YKR086W are 176 bp apart. They state that the genes share a common promoter region and conjecture that some functional relation exists[6]. However, expression data do not support this conjecture. Figure 2a shows the standardized expression patterns, obtained using oligonucleotide microarrays, for the two genes during two cell cycles. The patterns

FIGURE 2. Standardized expression data for gene pairs



(a) Standardized expression data for YKR085C (MRPL20) and YKR086W (PRP16). Despite their proximity, expression patterns are uncorrelated. (b)–(d) Standardized expression data for three pairs of genes coding for ribosomal proteins.

have a correlation of −0.15 (there are many similar examples). In certain cases, the patterns can be different owing to differences in mRNA stability – for example, the same amount of mRNA might be transcribed, but differential degradation rates could affect the expression measurements of the microarrays. However, in the many examples of completely different expression pattern pairs, it is more likely that separate regulatory systems control the genes.

Considering the gene pairs YBR189W–YBR191W, YJL190C–YJL189W and YOL040C–YOL039W in more detail, the first pair is transcribed on the same strand, whereas the other two pairs are transcribed away from each other on opposite strands. In the first case, the regulatory system of both genes could be located upstream of YBR189W. In the other cases, it might be located between the genes. The expression patterns from the cell cycle data are shown in Fig. 2b–d. These genes code ribosomal protein, and as the genes are functionally related, it is not surprising that all six patterns are similar. However, the expression patterns for the genes in each pair are almost identical. The average correlation between adjacent genes is 0.95, whereas the average correlation between the groups is 0.85. In the diauxic shift data, the corresponding correlations are 0.96 and 0.93. The values are 0.83 and 0.67 in the sporulation data. The correlation values support the notion that the consecutive genes are controlled by a single regulatory system.

Many genes that are controlled by separate regulatory systems might have highly correlated expression patterns. This could be the case if the genes are functionally related, or it might occur by random chance. If genes are controlled by a single regulatory system, then their expression patterns should be highly correlated in any data set. Each data set that becomes available can be used to generate a list of adjacent genes with highly correlated expression patterns. By intersecting these lists we obtain the most likely candidates for control by a single regulatory system. Table 1 presents some of the best candidates based on the three expression data sets. A complete candidate list can be found at http://www-hto.usc.edu/~tanghx/yeast-gene-pair.html, and code for conducting searches in other data sets is available from the authors. Once strong candidates have been found, the issue can be resolved by experimental analysis of yeast strains having a mutation in the putative site of the single regulatory region. If both genes are affected, then they are controlled by a single regulatory system.

Acknowledgements

We thank L. Kruglyak and E. Hubbell for valuable comments and feedback. We were supported in part by NSF grant DBI 9504393 and NIH grant R01 GM36230. We thank S. Tavaré and M. Waterman for useful discussions and for providing us with the resources for this project. We also thank the referees for their valuable comments.

References

1 Reznikoff, W.S. (1972) The operon revisited. *Annu. Rev. Genet.* 6, 133–156
2 Ptashne, M. and Gann, A. (1997) Transcriptional activation by recruitment. *Nature* 386, 569–577
3 Struhl, K. (1989) Molecular mechanisms of transcriptional regulation in yeast. *Annu. Rev. Biochem.* 58, 1051–1077
4 Struhl, K. (1996) Chromatin structure and RNA polymerase II connection: implications for transcription. *Cell* 84, 179–182
5 Helden, J. *et al.* (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J. Mol. Biol.* 281, 827–842
6 Zhang, X. and Smith, T.F. (1998) Yeast 'operons'. *Microb. Comparative Genomics* 3, 133–140
7 Cho, R.J. *et al.* (1998) A genome-wide transcriptional analysis of the mitotic cell cycle. *Mol. Cell* 2, 65–73
8 Tavazoie, S. *et al.* (1999) Systematic determination of genetic network architecture. *Nat. Genet.* 22, 281–285
9 Arnone, M.I. and Davidson, E.H. (1997) The hardwiring of development: organization and function of genomic regulatory systems. *Development* 124, 1851–1864
10 Heinemeyer, T. *et al.* (1999) Expanding the TRANSFAC database towards an expert system of regulatory molecular mechanisms. *Nucleic Acids Res.* 27, 318–322
11 DeRisi, J.L. *et al.* (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* 278, 680–686
12 Chu, S. *et al.* (1998) The transcriptional program of sporulation in budding yeast. *Science* 282, 699–705

Complex evolution of the inositol-1-phosphate synthase gene among archaea and eubacteria

The presence of inositol and its metabolites is widely observed in eukaryotes. However, their occurrence in prokaryotes, with few exceptions, is uncommon. The origin of inositol metabolism in prokaryotes thus remains uncertain[1,2]. The first known enzymatic step in the *de novo* biosynthesis of inositol is mediated by the *INO1* gene, which encodes 1L-*myo*-inositol-1-phosphate synthase (I1-P synthase)[2]. In the presence of nicotinamide adenine dinucleotide (NAD^+), this enzyme catalyzes the conversion of glucose-6-phosphate (G-6-P) into inositol-1-phosphate, which is subsequently dephosphorylated to *myo*-inositol by a specific I1-P phosphatase. In yeast and fungi, genetic evidence indicates that I1-P synthase is essential for viability[3]. We recently identified I1-P synthase in *Mycobacterium tuberculosis H37Rv*, leading to the first evidence for the presence of this enzyme in a prokaryote[2]. To obtain insights into the origin of inositol metabolism among archaea and eubacteria, we have analyzed the evolution of this metabolically important enzyme.

Nandita Bachhawat
nandita@lion.imtech.ernet.in

Shekkar C. Mande
shekkar@bragg.imtech.ernet.in

Institute of Microbial Technology, Sector 39.A, Chandigarh 160036, India.