

Extracting Knowledge from Dynamics in Gene Expression

Ben Y. Reis,* Atul S. Butte,† and Isaac S. Kohane‡¹

*Harvard–MIT Division of Health Science Technology, Cambridge, Massachusetts 02139;

†Children's Hospital Informatics Program & MIT Division of Health Science Technology, Boston,

Massachusetts 02115; and ‡Children's Hospital Informatics Program, Boston, Massachusetts 02115

Received December 22, 2000; published online March 15, 2001

Most investigations of coordinated gene expression have focused on identifying correlated expression patterns between genes by examining their normalized static expression levels. In this study, we focus on the *dynamics* of gene expression by seeking to identify correlated patterns of changes in genetic expression level. In doing so, we build upon methods developed in clinical informatics to detect temporal trends of laboratory and other clinical data. We construct relevance networks from *Saccharomyces cerevisiae* gene-expression dynamics data and find genes with related functional annotations grouped together. While some of these associations are also found using a standard expression level analysis, many are identified exclusively through the dynamic analysis. These results strongly suggest that the analysis of gene expression dynamics is a necessary and important tool for studying regulatory and other functional relationships among genes. The source code developed for this investigation is freely available to all non-commercial investigators by contacting the authors. © 2001 Academic Press

Key Words: gene expression; clustering; dynamics; bioinformatics; clinical informatics; trends.

INTRODUCTION

To understand a system fully, one must study its dynamics. With the sequencing of the human genome completed last

¹To whom correspondence should be addressed at Children's Hospital Informatics Program, Harvard Medical School, 300 Longwood Avenue, Boston, MA 02115. Fax: (617) 355-3456. E-mail: isaac.kohane@harvard.edu.

year, the focus of the research community is shifting toward a functional understanding of the roles of and relationships between different genes. With advances in genetic expression profiling techniques [1, 2] enabling detailed genomic scale measurements of genetic activity, it is important for the purposes of knowledge discovery to extract all the meaningful information present in the data. To date, most analyses [3, 4] have focused on clustering genes based simply on correlated patterns of genetic expression, ignoring other relationships present in the data. In this report we propose that further identifying correlated patterns of gene expression *dynamics* reveals additional meaningful information in the data.

In pursuing the investigation of gene dynamics, we are recapitulating and building on a large body of work in clinical informatics dealing with the identification of temporal abstractions and trend analysis. The literature is replete with reports of the limitations of performing diagnosis or planning with atemporal data [5] and the leverage obtained by capturing the dynamics of biomedical processes [6–11]. Until recently, the application of these techniques in bioinformatics has been relatively limited, particularly in the analysis of gene expression, in part because of the paucity of data sets with sufficient time points. Those analyses that have been published have focused primarily on the use of signal processing techniques using the Fourier transform [12, 13].

Many techniques have been used in functional genomics

for clustering, including phylogenetic trees [4], self-organizing maps [14, 15], and relevance networks [16, 17]. These clustering techniques have relied on a variety of association metrics such as Euclidean distance, correlation coefficients, and mutual information. These different techniques and association measures have, to varying degrees, all proved successful in clustering genes known to be related in function. While many differences among these various approaches exist, all of them cluster according to the absolute level of genetic expression. In this study, we propose an alternate approach involving the dynamics of genetic expression, and formulate a methodology for clustering genes according to changes in genetic expression level.

Clustering Genes According to Expression Dynamics Has Important Advantages

We use the term *dynamics* to refer to the rate of change of genetic expression over time, calculated as the first-order difference of the genetic expression levels ($E_{t_2}-E_{t_1}$, $E_{t_3}-E_{t_2}$). This is different from the simple temporal pattern of genetic expression (E_{t_1} , E_{t_2} , E_{t_3}) that we refer to as *statics*.

The primary motivation for studying gene expression dynamics is that existing static techniques may not identify all the important relationships. Some genes may have associated dynamic behaviors but may not have associated static expression behaviors. A hypothetical example is shown in Fig. 1: Gene A codes for an enhancer protein that regulates the expression of gene B—a high level of gene A causes an up-regulation of expression in gene B. Since gene B can be at many possible expression levels before being affected by gene A, the enhancer-type relationship between the two genes cannot be noticed by simply examining the correlation of static expression patterns. Instead, one needs to examine the dynamics of gene expression—the way in which the expression level of gene A leads to a change in gene B—in order to detect the underlying dynamic relationship. We therefore hypothesize that this dynamic approach has the potential to discover relationships between genes that are not detectable using existing static techniques. It is the goal of this study to formulate, validate, and evaluate this dynamic approach for knowledge discovery in functional genomics.

METHODS

Experimental Data

We studied the *Saccharomyces cerevisiae* (Table 1) mRNA-expression data aggregated from several experiments reported by Eisen *et al.* [3] in which the response of

the yeast cells to several different stimuli is recorded. The data contain 79 data points in 10 time-series measured under different experimental conditions, shown in Table 1. Of over 6000 genes in the yeast genome, Eisen included only 2467 genes that had functional annotations. We analyze the same subset of genes.

Representing Gene Expression Dynamics

Slopes are calculated between each adjacent pair of expression data points, E_{t_n} and $E_{t_{n+1}}$:

$$\text{Slope}(n, n + 1) = \frac{\text{expression_level}_{n+1} - \text{expression_level}_n}{(\text{time}_{n+1} - \text{time}_n)}. \quad (1)$$

Since slopes are only calculated between data points within the same time series, the 79 data points in 10 time series are reduced to only 69 slope measurements. The units of the slope measurements are in normalized expression level units per minute.

Data Visualization and Analysis

Relevance networks are constructed for the purposes of analysis and visualization of the data. Relevance networks are reviewed briefly here and have been described in full previously [16–18].

Relevance networks help identify groups of interrelated genes. A metric of association is chosen for comparing patterns of genetic expression between genes. After all pairwise gene–gene association strengths are calculated, a statistically significant threshold level of association is determined. All connections weaker than this threshold are removed, leaving small interconnected islands of significantly related genes called relevance networks. The method for determining this threshold is outlined below, but as described in prior work [16–18] it involves permuting the entire original data set, to preserve the distribution of gene expression values, but breaking the link between expression value and a particular condition or tissue. The pairwise association strengths are recalculated for each permutation and the largest value of association obtained in the pairwise associations is recorded. After a large number of permutations, this maximum value becomes the minimum threshold value for any association in the unpermuted data sets.

For this study, we use a linear correlation coefficient as a measure of association. Slopes are calculated as described

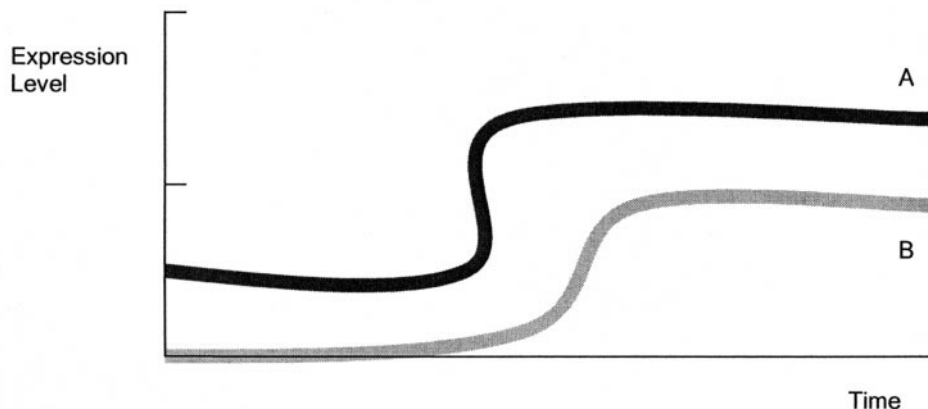
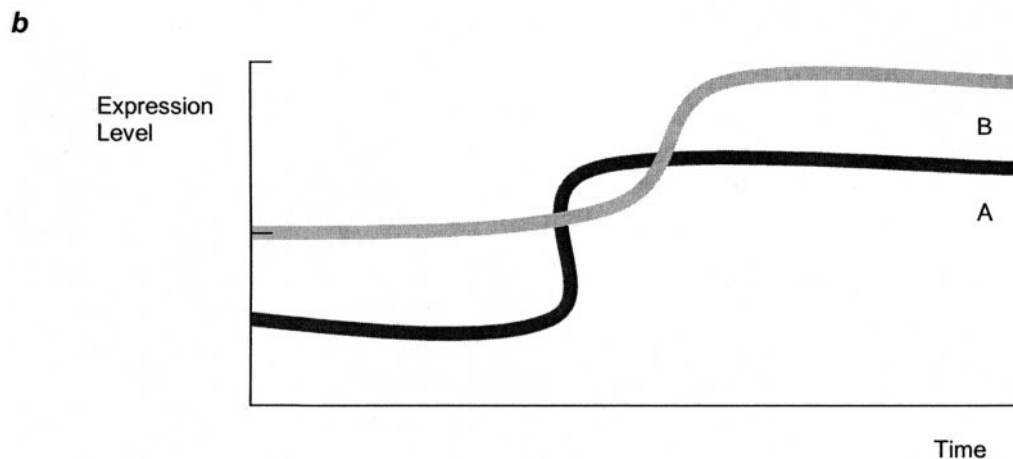
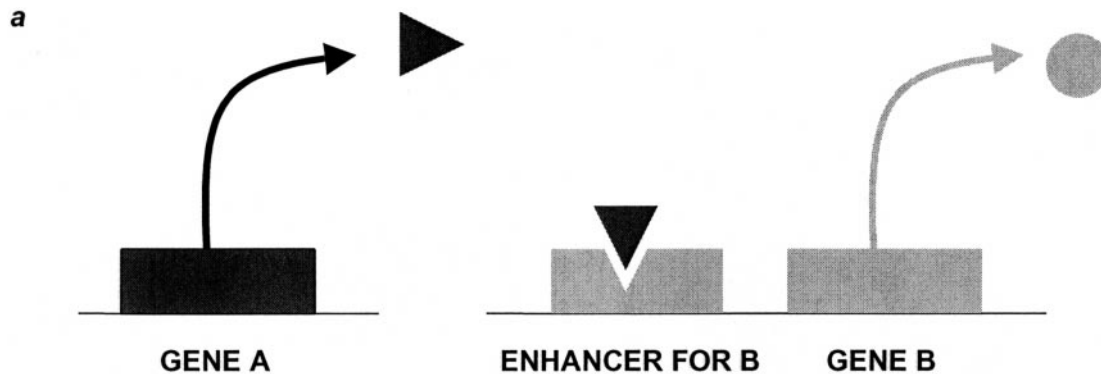


FIG. 1. Dynamic relationships between genes. (a) The expressed product of gene A binds an enhancer region that increases transcription of gene B. (b) B's initial expression level before being affected by A can vary throughout the experiment. As a result, measuring the correlation between the absolute levels of genes A and B will not reveal the underlying enhancement relationship between the two. Instead, this can only be done by analyzing the expression dynamics—the change in expression level of gene B in relation to the expression level of gene A.

TABLE 1
Experimental Conditions

Number of time points	Conditions
18	Cell cycle after Alpha factor arrest and release.
14	Cell cycle after elutriation.
15	Cell cycle for cdc 15 mutants after temperature-sensitive arrest and release.
6	Sporulation, Experiment 1.
3	Sporulation, Experiment 2.
2	Sporulation, Experiment 3.
6	Response to high-temperature shock.
4	Response to reducing shock.
4	Response to low-temperature shock.
7	Response to diauxic shift.

Note. The experimental conditions under which the gene expression measurements reported by Eisen *et al.* [4] were taken.

above, yielding a dataset where each row is the time series of a particular gene’s expression dynamics. Pairwise Pearson correlation coefficients are then calculated between all possible combinations of two rows. These are squared to yield the R^2 , after which the original sign is reappended to conserve the information of whether the genes are positively or negatively correlated [7]. We call this final signed value R^2 .

Correlation coefficients are sensitive to outlying values, which can bias downstream data analysis. Two symmetric outlying values may artificially raise the correlation coefficient of an otherwise nonlinear distribution. That is, in all except one or two microarrays, a gene will have a scatter within a small range and then due to an artifact of the hybridization process, the one or two microarrays will have a very high value for that gene. This is all the more striking in the data set on which this analysis has been performed where each data point belongs to a time series of a given stimulus and where the rest of the time series shows much smoother changes. We have had to apply the filter for these outlier values also in prior studies for the same reason [17]. It should be pointed out, nonetheless, that we will necessarily miss those few occasions where outlier values do represent quantum and dramatic change in expression. Consequently, an entropy-based filter is used to remove genes with outlying values in their distributions from the analysis. First, the individual entropies of the dynamics time series are calculated for each gene, with the entropy $H(A)$ defined as:

$$H(A) = \sum_n -p(A_n)\log_2 p(A_n). \quad (2)$$

The genes are ranked according to their entropies, and the bottom 5% (entropy threshold <2.1464) are excluded from the analysis.

Issues Specific to Dynamics

The inclusion of dynamics in the methodology introduces a number of important issues that will be addressed in turn. First, the issue of stasis: we observe that most genes do not change their expression levels most of the time. Figure 2 shows the distribution of slopes taken from all the points in the data set. The widespread stasis in the data can lead to seriously misleading analyses, as genes that remain stationary together can lead to an artificially high measure of association.

To address this issue, we filter out the stationary data points, including only the more dynamic ones in the analysis. This involves setting an exclusion range, or hole, around the zero slope range. We choose thresholds of ± 0.02 normalized slope units per minute, for a total hole size of 0.04. This approach removes approximately 70.0% of the original data points, and allows us to study the genes that *change* in a coordinated fashion, while avoiding the misleading identification of genes that simply remain stationary together. We automatically evaluated a range of hole sizes and picked the value of 0.04 based on maximizing the number of retained data points and minimizing the threshold association level in the permuted data (described below).

This solution leads to the second complication: Since many data points are removed, the remaining data can be very sparse. To ensure that all correlation coefficient calculations are based on enough points to avoid too many spurious associations (as defined by our permutation analysis, below), we set a threshold requiring a minimum of five data points for a calculation to be valid: any pairwise distribution having fewer data points than this threshold is excluded from the analysis. This thresholding approach is similar to the one used in work on clinical data relevance networks [18].

The last step in the relevance networks methodology involves determining a threshold association level that represents a likely nonspurious association between genes. We determine this level by permuting the time points within each gene and obtaining the distribution of pairwise correlation coefficient values. We perform this permutation 100 times and then compare the average permuted distribution with the distribution obtained from the original data set. It is clear from Fig. 3 that much stronger correlations are present in the original dataset. We comfortably place the threshold

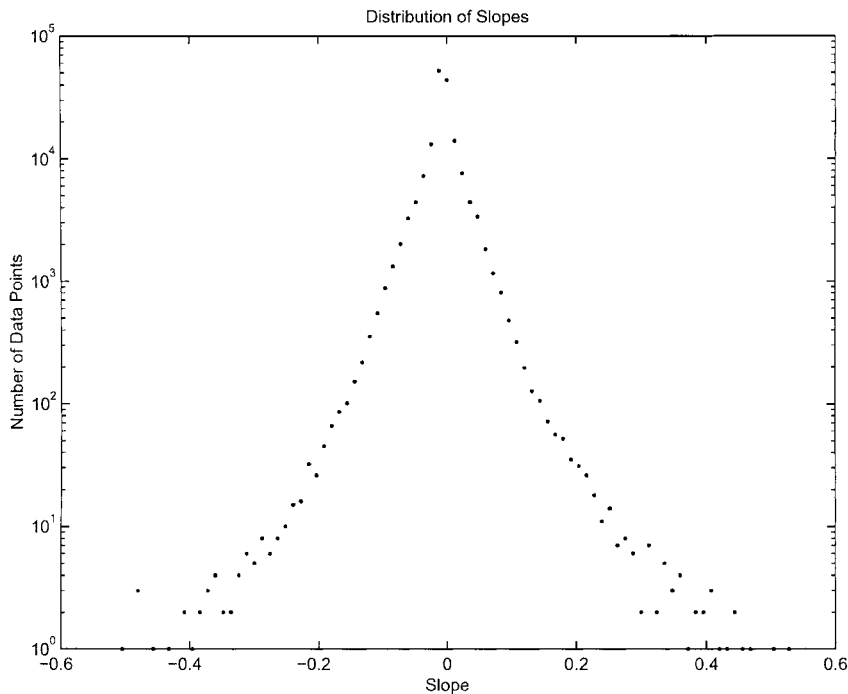


FIG. 2. Slope measurements of gene expression. A semilog plot of the distribution of slopes derived from the yeast genetic expression dataset reported by Eisen *et al.* Almost all of the slopes are near zero, illustrating that most genes remain stationary most of time. The complications caused by this widespread stasis must be addressed with in analyzing gene expression dynamics data, as described in the text.

level of significance at ± 0.78 , as no permuted data points are able to achieve an R^2 value greater than that.

RESULTS

The relevance networks generated from the dynamics analysis are presented first. These are then evaluated in the context of the networks generated from a static analysis below.

Dynamic Relevance Networks

Using a threshold of $R^2 = 0.78$, the dynamic analysis yields 71 relevance networks consisting of 348 nodes (Fig. 4). Of the 3,041,811 possible gene–gene connections, only 371 (0.012%) are above this threshold. A box labeled with the gene name represents each gene. The width of each box represents its indegree—the number of other genes connected to it.

There are far too many associations to discuss each one individually. We therefore present the strongest associations found, as well as some of the more interesting negative associations. The full dataset and analysis are available at <http://www.chip.org/chip/people/kohane/papers/dynamics/readme.html>.

Of the 71 networks, the largest one contains 154 nodes with 238 links and consists mostly of ribosomal proteins and related genes, such as RNA helicases, RNA polymerases, translation initiation proteins, and other translational regulators. All of these genes are directly related in function to protein synthesis. A smaller network, with 14 nodes and 19 links, also consists of mostly ribosomal proteins.

The gene with the highest indegree is *RRP4*, a 3' \rightarrow 5' exoribonuclease involved in a diverse array of RNA processing [19]. It is linked with 12 other genes, including RNA helicases, RNA polymerases, and other translational regulators. Its high connectivity suggests that it is coregulated with many of the genes involved in protein synthesis and appears to interact with these genes in a dynamic manner.

Of all the dynamic associations found, Table 2 shows the 10 with the highest R^2 values. The gene pairs found are

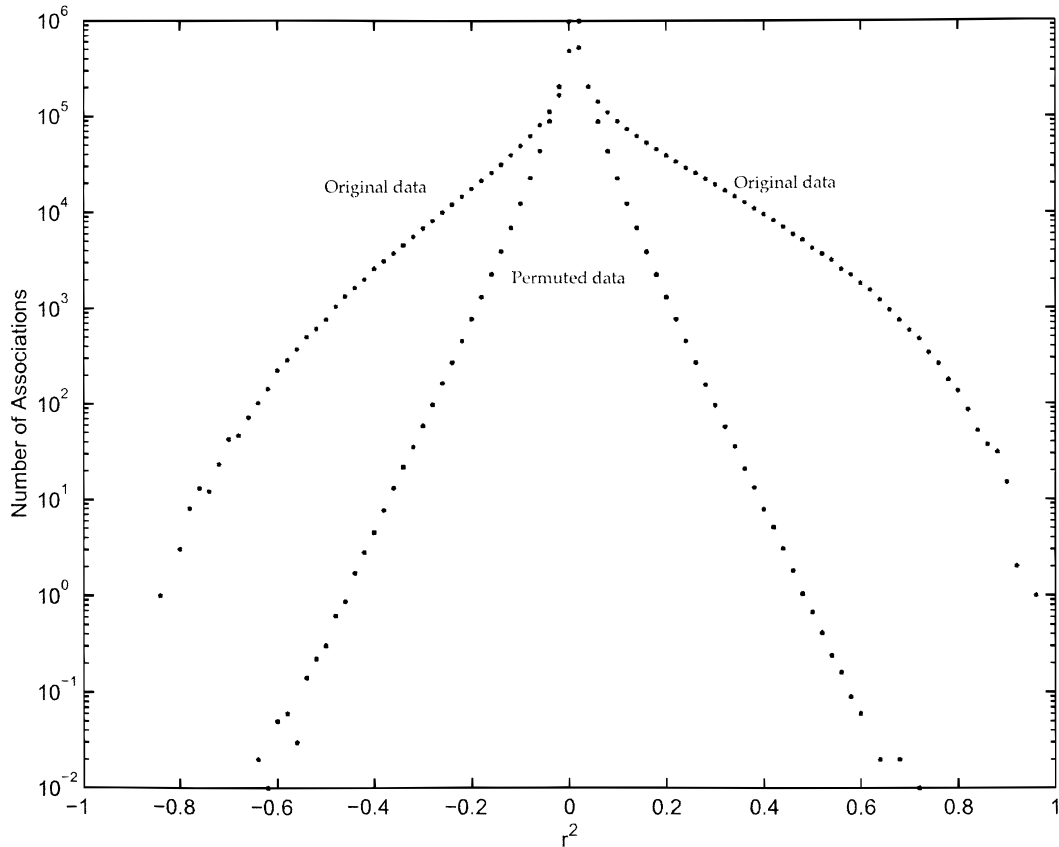


FIG. 3. Dynamic correlations of gene expression. A semilog plot of the distribution of R^2 calculated from the pairwise comparisons of the dynamic expression patterns of all the yeast genes. Plotted are the permuted and the unpermuted data. The permuted data points represent the average R^2 distribution derived from 100 permutations of the dataset. It is clear from the graph that the original data are able to achieve high R^2 values that are not achieved in any of the permuted runs.

closely related in function, including the four occurrences of *ASP3* (L-asparaginase II), which are redundantly present on the microarray used for making the measurements. These associations are shown graphically in Fig. 5A.

Negative Associations

We also examine two negative associations of interest. First, we look at *EXM2* and *MAD3* (Fig. 6A). *EXM2* is a protein involved in allowing cells to exit mitosis, while *MAD3* is a spindle-assembly checkpoint protein that prevents certain cells from leaving mitosis [20]. These two counteracting genes appear as negatively correlated in their dynamics with an R^2 of -0.797 . Meanwhile, they are/are not found to be strongly associated in the static analysis.

Figure 6B shows the distribution of slopes between *RAD6*

and *MET18*. *RAD6* is a ubiquitin-conjugating enzyme concentrated in the nucleus that is essential for mediating the degradation of amino-end rule-dependent protein substrates [21]. *MET18*, also known as *MMS19*, is a protein concentrated in the nucleus that affects RNA polymerase II transcription [22]. These are inversely related in their dynamics, with an R^2 of -0.791 . It is not surprising that a gene responsible for protein degradation has an inverse relationship to a gene responsible for RNA transcription leading to protein synthesis.

The circles in Fig. 6B represent the static data points that we filter out to avoid direct the analysis to finding correlated changes in gene expression. If these static points are included in the analysis, the R^2 shifts from -0.79 to -0.58 , far below the determined level of significance. This illustrative example highlights the methodological utility of the filtering

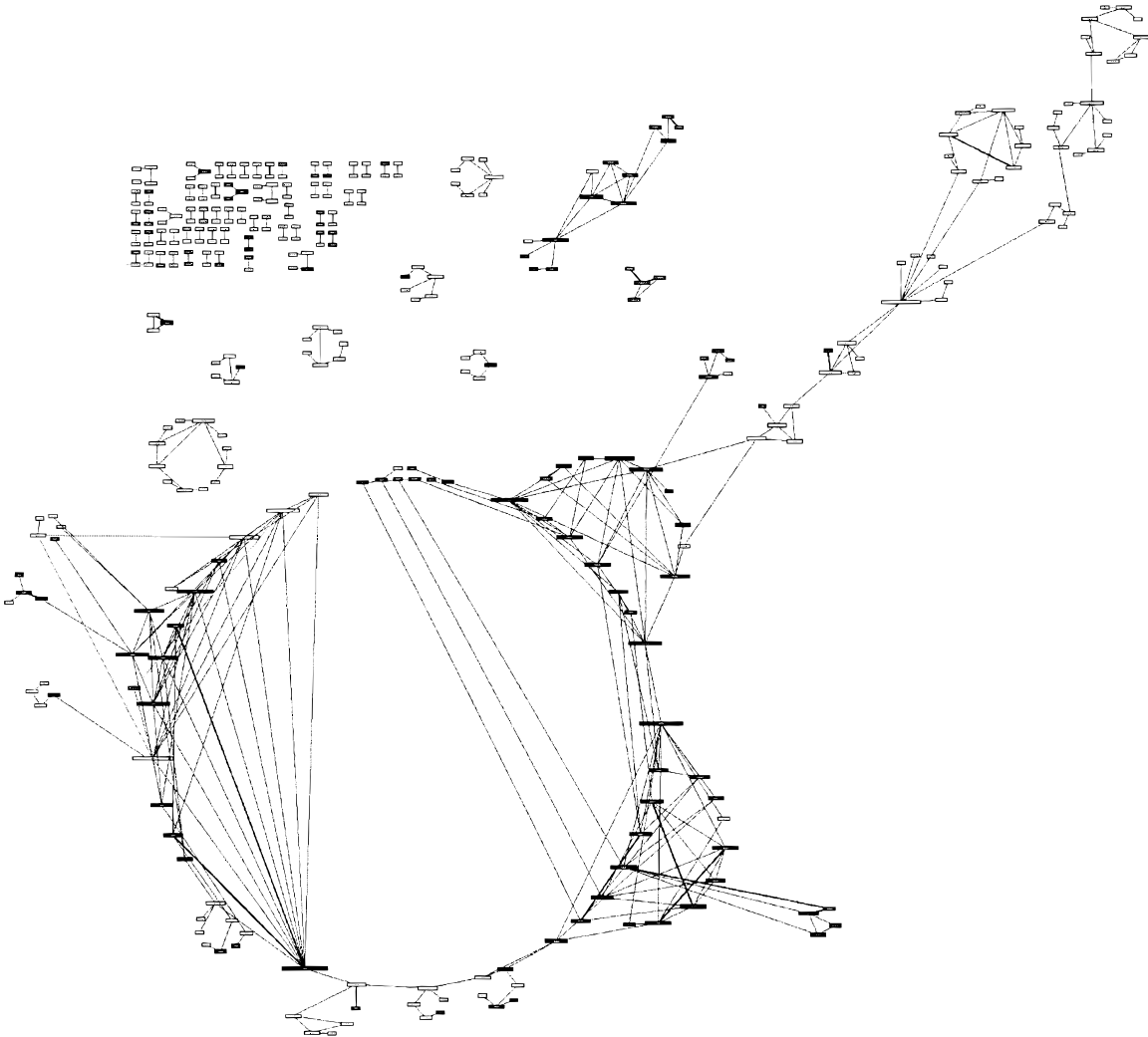


FIG. 4. Dynamic relevance networks. The relevance networks generated using the dynamics methodology proposed in this study. Each box represents a gene, labeled with its alphanumeric identification tag. The width of each box is determined by how many other genes it is connected to. The various groups of interconnected genes are called relevance networks. The shaded genes are those that are also found in the static analysis.

out the stationary data points when clustering according to gene expression dynamics.

As shown in Fig. 3, there are far more strong positive associations than strong negative associations. Figure 7 shows the distribution for the strongest positive association (two ribosomal proteins *RPL42* and *RPS24*, $R^2 = 0.957$), and for the strongest negative association (two RNA polymerase genes *RPO31* and *SRB8*, $R^2 = -0.854$). In general, the distributions with an extremely tight linear fit are all positive. It could be argued that these tight correlations represent more direct relationships between genes, such as two genes occurring in the same step of a biological pathway—two

ribosomal proteins that are always up-regulated or down-regulated together. It could further be argued that there are no extremely strong negative correlations because negative feedback in biological systems occurs mostly through multistep signal cascades. These are by definition more indirect and thus result in less tight linear relationships between negatively correlated genes.

Comparison of Dynamics and Static Analyses

In this work we have formulated a methodology for clustering genes according to gene expression dynamics. To

TABLE 2
Dynamic Associations

Gene name	Category	Gene description	R^2
RPL42B	Protein synthesis	Ribosomal protein L42b	0.957
RPS24B	Protein synthesis	Ribosomal protein L24B	
ASP3	Asparagine utilization	L-Asparaginase II	0.910
ASP3	Asparagine utilization	L-Asparaginase II	
RRP4	rRNA processing	Exoribonuclease / rRNA processing	0.905
SUA5	Protein synthesis	Translation initiation protein	
LOS1	tRNA splicing	Involved in tRNA splicing	0.900
NIP1	Nuclear protein targeting	Subunit of translation initiation	
RPL5	Protein synthesis	Ribosomal protein	0.897
RPS0A	Protein synthesis	Ribosomal protein	
NMD3	mRNA decay	Required for stable ribosomal subunit formation	0.897
NSR1	Nuclear targeting protein	NLS binding protein/rRNA processing	
RPL9A	Protein synthesis	Ribosomal protein	0.894
RPS8B	Protein synthesis	Ribosomal protein	
HHF2	Chromatin structure	Histone H4	0.894
HTB1	Chromatin structure	Histone H2B	
ASP3	Asparagine utilization	L-Asparaginase II	0.893
ASP3	Asparagine utilization	L-Asparaginase II	
IMG1	Protein synthesis	Mitochondrial ribosomal protein	0.893
RSC6	Chromatin structure	Chromatin remodeling complex subunit	

Note. The strongest associations between genes found using the dynamics analysis.

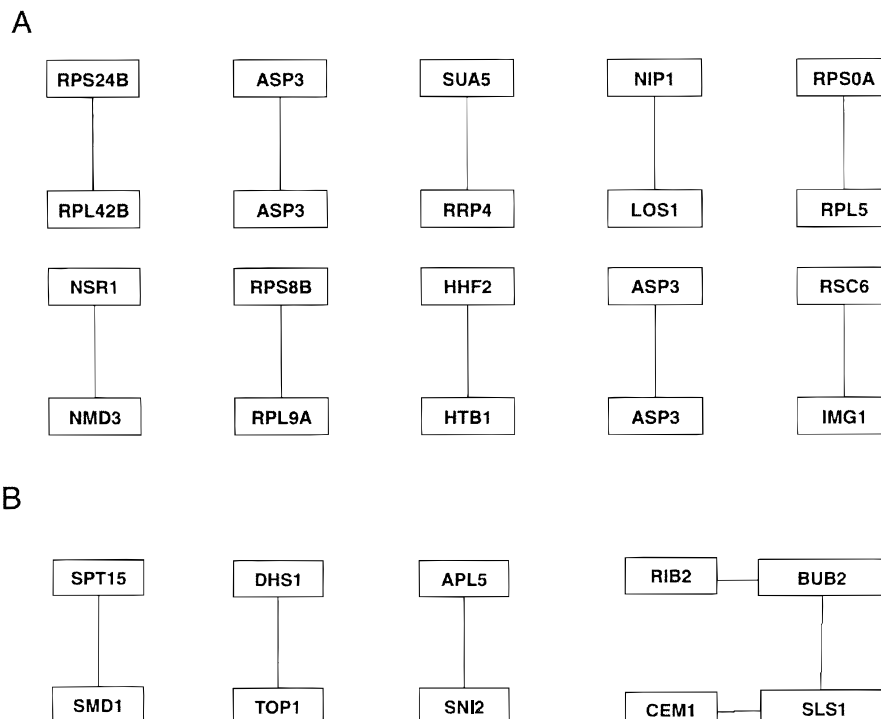


FIG. 5. (A) The strongest associations found using the dynamics analysis. (B) Selected networks found using the dynamics analysis, but not the static analysis.

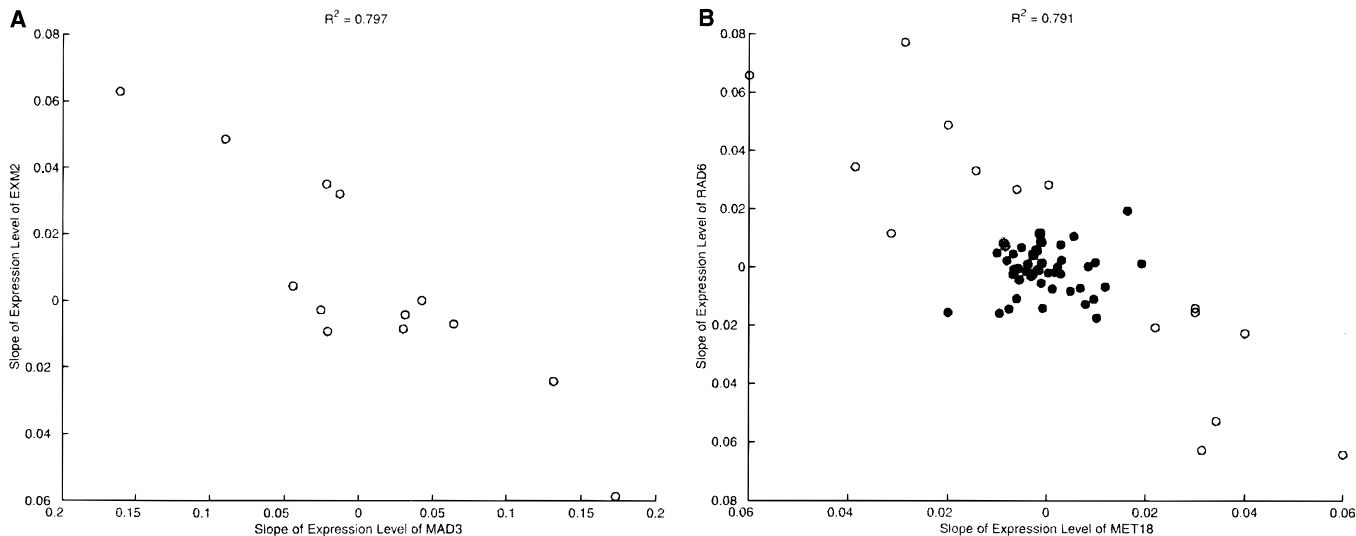


FIG. 6. Negative dynamic correlations. (A) The distribution of slopes of MAD3 and EXM2 plotted one against another. (B) The distribution of slopes of MET18 and RAD6 plotted one against another. The filled points in the middle are those static points excluded in the analysis to ensure identification of only truly dynamic relationships between genes.

evaluate this methodology in the context of existing techniques, we construct a second set of relevance networks based on a static analysis of the same dataset.

The static analysis is performed as above, with a few key differences. First, we use the original gene expression data, and not the first difference of gene expression. Second, while the genes are still ranked by entropy value and the bottom 5% (entropy threshold <2.2187) are removed, there is no

need to filter out any “stationary” data points since this is a static analysis.

For purposes of comparison, we set the threshold R^2 to 0.70, creating a set of relevance networks with a similar number of genes as that seen in the dynamic analysis (356 static vs 348 dynamic). Of the 3,041,811 possible gene–gene connections, 4872 (0.16%) are found to be above this threshold.

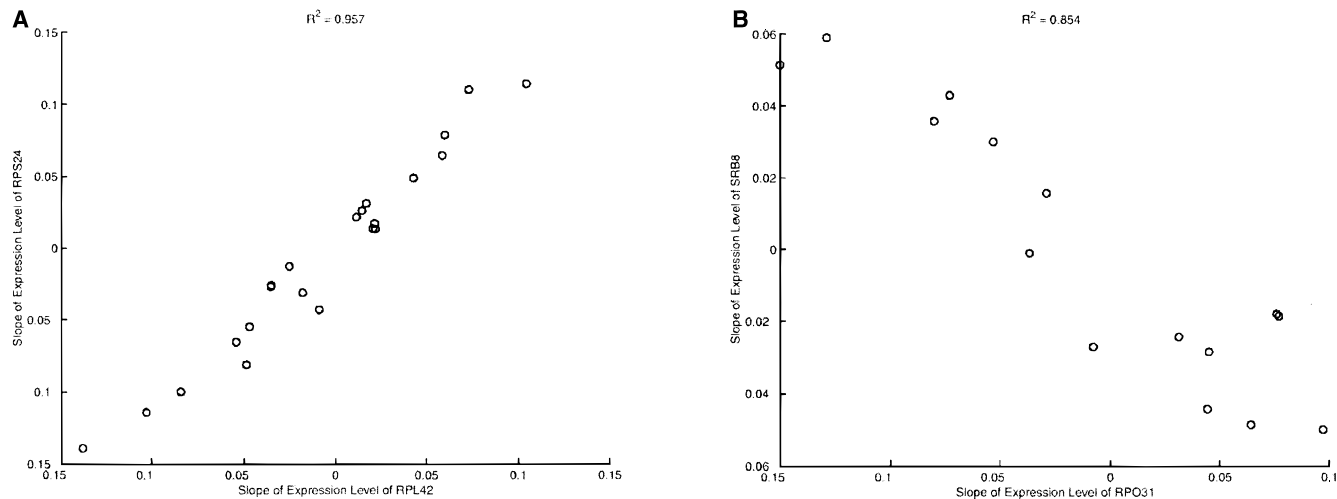


FIG. 7 Positive and negative correlations. Slope–slope distributions of the strongest positive correlation (A) and the strongest negative correlation (B). On the whole, the strongest positive correlations were more tightly linear than the strongest negative ones, as explained in the text.

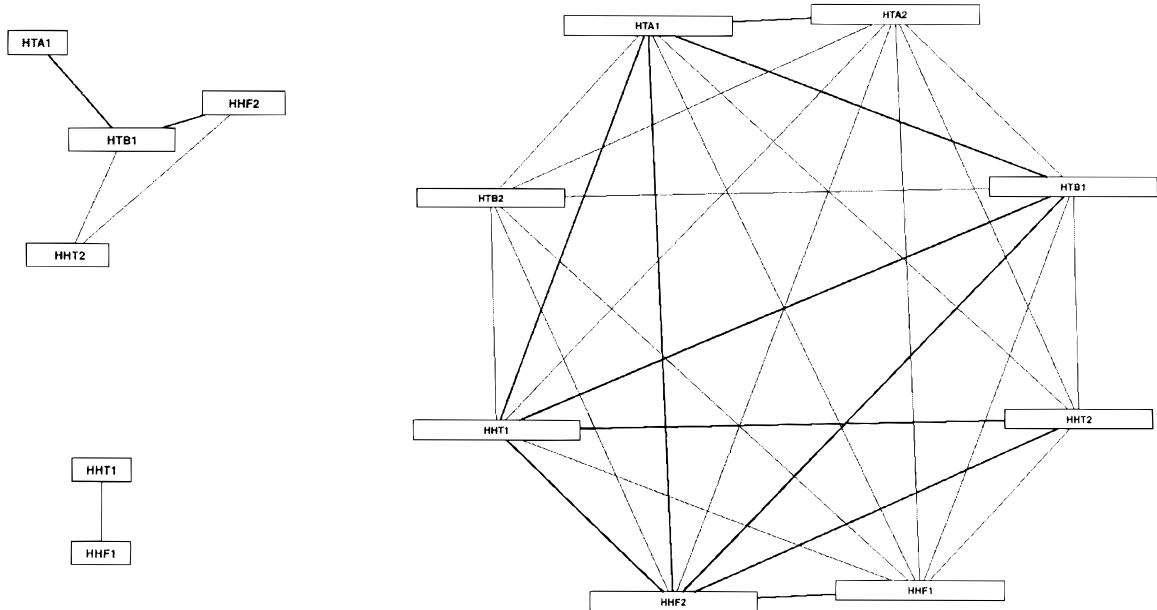


FIG. 8 Comparison of networks. A group of histone genes grouped together by both the static and dynamic analyses. The dynamics analysis (left) found few dynamic connections compared to the almost fully connected clique formed by the static analysis.

Although both analyses contain a similar number of genes, there are more individual networks generated from the dynamics analysis (71 separate networks vs 45 in the static), while there are far more interconnections between genes in the static analysis (4872 links vs 371 in the dynamic). These results may indicate that slope–slope associations are less common biologically, or that they are more difficult to detect with this methodology than static associations.

We find that 133 genes appear in both the static and dynamic analyses, leaving 215 genes that are exclusive to the dynamics analysis. However, only about half of the 133 shared genes appear linked to the same genes in both analyses—most appear linked to other genes.

Relationships Appearing in Both Analyses

A number of links are found identically in both analyses, some of which are shown in Table 3. A large interconnected network of histone genes responsible for chromatin structure found by the static analysis appears broken up into two networks in the dynamics analysis (Fig. 8). This may indicate that certain associations are inherently more dynamic than others.

Some genes are found in both analyses, but appear associated with different genes. One particularly interesting example is discussed here. In the dynamic analysis, three genes involved in protein synthesis are grouped into a single network: *PRSI* is involved making PRPP, required for making amino acids [23]; *SIK1* is a nucleolar protein necessary for ribosomal subunit assembly [24]; *SUI2* codes for a subunit of a translation initiation factor [25]. All three of these genes appear in the static analysis as well, but none are linked to each other. In fact, while *PRSI* and *SIK1* do appear indirectly related in the same network in the static analysis, they are not found to be strongly directly linked to each other. These examples illustrate how using both static and dynamic approaches can attain a complementary view of gene–gene relations.

Associations Found Exclusively in the Dynamics Analysis

Most of the genes appearing in the dynamics analysis are not found using the static analysis. A selection is reviewed here (Fig. 5B).

One network grouped *SMD1*, involved in mRNA splicing [26], with *SPT15*, a gene involved in transcription [27].

TABLE 3
Shared Associations

Name	Category	Description
POL30	Replication	DNA polymerase processivity factor
RFA1	Replication	Replication factor A, 69 kDa subunit
RPN12	Protein degradation	26S proteasome regulatory subunit
RPN9	Protein degradation	26S proteasome regulatory subunit
CUP1	CU ²⁺ Ion homeostasis	Metallothionein
CUP1	CU ²⁺ ion homeostasis	Metallothionein
ASP3	Asparagine utilization	L-Asparaginase II
ASP3	Asparagine utilization	L-Asparaginase II
APT1	Purine biosynthesis	Adenine phosphoribosyltransferase
None	Protein synthesis	Tryptophan–TRNA ligase
HHF1	Chromatin structure	Histone H4
HHT1	Chromatin structure	Histone H3
HTB1	Chromatin structure	Histone H2B
HHF2	Chromatin structure	Histone H4
HHT2	Chromatin structure	Histone H3
HTA1	Chromatin structure	Histone H2A
RPS25B	Protein synthesis	Ribosomal protein S25B
RPS31	Protein synthesis	Ribosomal protein S31
RPL18A	Protein synthesis	Ribosomal protein L18A
RPL8B	Protein synthesis	Ribosomal protein L8B
RPL1B	Protein synthesis	Ribosomal protein L1B
RPS19B	Protein synthesis	Ribosomal protein S19B

Note. Selected associations found by both the dynamic and static analyses.

Another network grouped *TOPI*, involved in DNA replication [28], with *DHS1*, involved in DNA repair and recombination [29]. Yet another network grouped *APL5*, involved in vacuolar protein targeting [30], with *SNI2*, a gene involved in secretion [31].

Another interesting network consists of one cell cycle gene *BUB2* [22] and three genes localized in space in the mitochondria: *SLS1* is integral membrane protein involved in mitochondrial metabolism [32]; *CEM1* is a mitochondrial protein involved in fatty acid metabolism [33]; *RIB2* is involved in riboflavin synthesis, also localized to the mitochondria [34].

The intuitive nature of many of these relationships suggests that the dynamics analysis can identify meaningful associations that are not found using a static analysis.

DISCUSSION

Summary of Results

We have formulated and evaluated an analytic methodology for clustering genes according to gene expression dynamics. The relevance networks produced from the dynamics analysis reveal significant and meaningful relationships, indicating that the dynamics approach is useful for knowledge discovery in functional genomics. Furthermore, the fact that most of these relationships are not found using a comparable static analysis further suggests that the dynamic approach is actually necessary for a more complete picture of gene–gene interactions. It is argued that the inherent dynamic nature of certain gene–gene relationships requires this inherently dynamic approach for knowledge discovery.

A sizable number of associations are found using both the static and dynamic analyses. The similarity between the results of the dynamic analysis and those of the already established static analysis serves to further validate the proposed dynamic methodology.

There were clearly also relationships found with the static approach that were not found using the dynamic approach. From these results we conclude that to extract all the valuable information from gene expression measurements, one needs a full set of complementary analysis methodologies that capture the dynamics of these systems. With continuing work in this emerging and important area of research, and the continued decreased cost of massively parallel expression measurements, the dynamics approach is ready to take its place amidst the growing set of tools for knowledge discovery in functional genomics. We anticipate that many more of the techniques developed to handle “noisy” dynamic processes in clinical informatics will find ready and immediate application to functional genomics.

Future Work

The slope measurements reported here were measured between adjacent data points. Longer-term effects can be studied by measuring slopes between time points that are more distant from one another. The associations reported here were measured between simultaneous slopes. We are currently studying possible time-lagged associations between slopes, allowing for signal propagation times and other delays. This phase generalization expands the analysis methodology to extract even more information from the gene expression data.

REFERENCES

1. Cheung VG, Morley M, Aguilar F, Massimi A, Kucherlapati R, Childs G. Making and reading microarrays. *Nat Genet* 1999; 21(1 Suppl):15–19.
2. Botwell D. Options available—from start to finish—for obtaining expression data by microarray. *Nat Genet* 1999; 1 (Suppl):25–32.
3. Eisen MB, Spellman PT, Brown PO, Botstein D. Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci USA* 1998; 95(25):14863–8.
4. Michaels GS, Carr DB, Askenazi M, Fuhrman S, Wen X, Somogyi R. Cluster analysis and data visualization of large-scale gene expression data. *Pac Symp Biocomput* 1998; 42–53.
5. Schwartz WB, Patil RS, Szolovits P. Artificial intelligence in medicine: where do we stand? *N Engl J Med* 1987; 316(11):685–688.
6. Haimowitz I J, Le P P, Kohane I S. Clinical monitoring using regression-based trend templates. *Artificial Intelligence Med* 1995; 7:471–472.
7. Russ TA. Reasoning with time dependent data [PhD]: Massachusetts Institute of Technology; 1992.
8. Kohane IS. Temporal reasoning in medical expert systems. In: Salomon R, Blum B, Jørgensen M, editors. *MEDINFO 86/Fifth World Congress on Medical Informatics*; 1986; Washington, DC: Elsevier Science, 1986; 170–174.
9. Rutledge G, Thomsen G, Farr B, Tovar M, Sheiner L, Fagan L. VentPlan: a ventilator-management advisor. In: Clayton PD, editor. *Symposium on Computer Applications in Medical Care*, 1991. Washington, DC, 1991; 869–871.
10. Shahar Y, Tu S, Musen M. Knowledge acquisition for temporal abstraction mechanisms. *Knowledge Acquisition* 1992; 1(4):217–236.
11. Kahn MG, Fagan LMB, Sheiner L. Model-based interpretation of time-varying medical data. In: *Proceedings Symposium Computer Applications in Medical Care*, 1989; 1989. 28–32.
12. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, et al. Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell* 1998; 9(12):3273–97.
13. Chen T, He HL, Church GM. Modeling gene expression with differential equations. *Pac Symp Biocomput* 1999; 29–40.
14. Toronen P, Kolehmainen M, Wong G, Castren E. Analysis of gene expression data using self-organizing maps. *FEBS Lett* 1999; 451(2):142–6.
15. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, Dmitrovsky E, et al. Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. *Proc Natl Acad Sci USA* 1999; 96(6):2907–12.
16. Butte A, Kohane I. Mutual information relevance networks: functional genomic clustering using pairwise entropy measurements. In: Altman R, Dunker K, Hunter L, Lauderdale K, Klein T, editors. *Pacific Symposium on Biocomputing 2000*; Hawaii: World Scientific, 2000; 418–429.
17. Butte AJ, Tamayo P, Slonim D, Golub TR, Kohane IS. Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks [In Process Citation]. *Proc Natl Acad Sci USA* 2000; 97(22):12182–6.
18. Butte A, Kohane IS. Unsupervised Knowledge Discovery in Medical Databases Using Relevance Networks. In: Lorenzi N, editor. *Fall Symposium, American Medical Informatics Association*; 1999; Washington, DC: Hanley and Belfus, 1999; 711–715.
19. Mitchell P, Petfalski E, Shevchenko A, Mann M, Tollervey D. The Exosome: A conserved eukaryotic RNA processing complex containing multiple 3' → 5' exoribonucleases. *Cell* 1997; 91:457–466.
20. Hwang LH, Lau LF, Smith DL, Mistrot CA, Hardwick KG, Hwang ES, et al. Budding Yeast Cdc20: A Target of the Spindle Checkpoint. *Science* 1998; 279:1041–4.
21. Watkins JF, Sung P, Prakash S, Prakash L. The extremely conserved amino terminus of RAD6 ubiquitin-conjugating enzyme is essential for amino-end rule-dependent protein degradation. *Genes Dev* 1993; 7(2):50–61.
22. Lauder S, Bankmann M, Guzder SN, Sung P, Prakash L, Prakash S. Dual requirement for the yeast MMS19 gene in DNA repair and RNA polymerase II transcription. *Mol Cell Biol* 1996; 16:6783–93.
23. Carter AT, Beiche F, Hove-Jensen B, Narbad A, Barker P, Schweizer LM, Schweizer M. PRS1 is a key member of the gene family encoding phosphoribosylpyrophosphate synthetase in *Saccharomyces cerevisiae*. *Mol Gen Genet* 1997; 254:148–56.
24. Gautier T, Berges T, Tollervey D, Hurt E. Nucleolar KKE/D repeat proteins Nop56p and Nop58p interact with Nop1p and are required for ribosome biogenesis. *Mol Cell Biol* 1997; 17:7088–98.
25. Cigan AM, Pabich EK, Feng L, Donahue TF. Yeast translation initiation suppressor *sui2* encodes the alpha subunit of eukaryotic initiation factor 2 and shares sequence identity with the human alpha subunit. *Proc Natl Acad Sci USA* 1989; 86:2784–8.
26. Rymond BC. Convergent transcripts of the yeast PRP38-MSD1 locus encode two essential splicing factors, including the D1 core polypeptide of small nuclear ribonucleoprotein particles. *Proc Natl Acad Sci USA* 1993; 90:848–52.
27. Cormack BP, Struhl K. The TATA-binding protein is required for transcription by all three nuclear RNA polymerases in yeast cells. *Cell* 1992; 69:685–96.
28. Christman MF, Dietrich FS, Fink GR. Mitotic recombination in the rDNA of *S. cerevisiae* is suppressed by the combined action of DNA topoisomerases I and II. *Cell* 1988; 55:413–25.
29. Tishkoff DX, Boerger AL, Bertrand P, Filosi N, Gaida GM, Kane MF, Kolodner RD. Identification and characterization of *Saccharomyces cerevisiae* EXO1, a gene encoding an exonuclease that interacts with MSH2. *Proc Natl Acad Sci USA* 1997; 94:7487–92.
30. Cowles CR, Odorizzi G, Payne GS, Emr SD. The AP-3 adaptor complex is essential for cargo-selective transport to the yeast vacuole. *Cell* 1997; 91:109–118.
31. Lehman K, Rossi G, Adamo JE, Brennwald P. Yeast homologues of tomosyn and lethal giant larvae function in exocytosis and are associated with the plasma membrane SNARE, *sec9*. *J Cell Biol* 1999; 146:125–40.
32. Rouillard JM, Dufour ME, Theunissen B, Mandart E, Dujardin G, Lacroute F. SLS1, a new *Saccharomyces cerevisiae* gene

- involved in mitochondrial metabolism, isolated as a synthetic lethal in association with an *SSM4* deletion. *Mol Gen Genet* 1996; 252:700–8.
33. Harington A, Herbert CJ, Tung B, Getz GS, Slonimski PP. Identification of a new nuclear gene (*CEM1*) encoding a protein homologous to a β -keto-acyl synthase which is essential for mitochondrial respiration in *Saccharomyces cerevisiae*. *Mol Microbiol* 1993; 9:545–55.
34. Pallotta MLBC, Fratianni A, DeVirgilio C, Barile M, Passarella S. *Saccharomyces cerevisiae* mitochondria can synthesise FMN and FAD from externally added riboflavin and export them to the extramitochondrial phase. *FEBS Lett* 1998; 428(3):245–9.