Genome-wide coexpression dynamics: Theory and application

Ker-Chau Li†

Department of Statistics, University of California, Los Angeles, CA 90095-1554

Edited by Peter J. Bickel, University of California, Berkeley, CA, and approved October 15, 2002 (received for review August 3, 2002)

High-throughput expression profiling enables the global study of gene activities. Genes with positively correlated expression profiles are likely to encode functionally related proteins. However, all biological processes are interlocked, and each protein may play multiple cellular roles. Thus the coexpression of any two functionally related genes may depend on the constantly varying, yet often-unknown cellular state. To initiate a systematic study on this issue, a theory of coexpression dynamics is presented. This theory is used to rationalize a strategy of conducting a genome-wide search for the most critical cellular players that may affect the coexpression pattern of any two genes. In one example, using a yeast data set, our method reveals how the hormones associated with the urea cycle are expressed to ensure proper mass flow of the involved metabolites. The correlation between ARG2 and CAR2 is found to change from positive to negative as the expression level of CPA2 increases. This delicate interplay in correlation signifies a remarkable control on the influx and efflux of ornithine and reflects well the intrinsic cellular demand for arginine. In addition to the urea cycle, our examples include SCH9 and CYR1 (both implicated in a recent longevity study), cytochrome c1 (mitochondrial electron transport), calmodulin (main calcium-binding protein), PFK1 and PFK2 (glycolysis), and two genes, ECM1 and YNL101W, the functions of which are newly revealed. The complexity in computation is eased by a new result from mathematical statistics.

M icroarrays have generated an enormous amount of gene-expression data from a variety of biological studies (1–5). After proper preprocessing, the data can be stored as a matrix of real numbers with \( N \) rows and \( m \) columns. Rows represent the gene-expression profiles, and columns represent the cell types, time points, or environmental or other experimental conditions under which the mRNA samples are taken. To elucidate microarray data, most methods (6–10) rely on the notion of profile similarity as described for Fig. 1. It is argued that coexpressed genes are likely to encode proteins that participate in a common structural complex, metabolic pathway, or biological process (6, 10).

Despite the many successful applications reported in the literature, there is an important issue that is hard to address by profile-similarity analysis. As is known, all biological processes are interlocked, and many proteins have multiple cellular roles. Two proteins engaged in a common process under certain conditions may disengage and embark on activities of their own under other conditions, which implies that both the strength and pattern of association between two gene profiles may vary as the intrinsic cellular-state changes. Weaver (11) discussed two transcription factors, Max and thyroid hormone receptor (TR). They can serve either as activators or repressors depending on other molecules bound to them. Max can bind to Myc and form a Myc–Max dimer that acts as a transcription activator. But when bound to Mad, the Mad–Max dimer serves as a repressor. For the case of TR, it associates with retinoid X receptor (RXR) to form a TR–RXR dimer that serves as a repressor in the absence of thyroid hormone. In the presence of thyroid hormone, the TR–RXR dimer is converted into an activator. Histone deacetylation is involved in both repressing events. Thus for example, if we take \( X \) to be the expression profile of the gene encoding TR and take \( Y \) to be the profile of one of its target genes, then \( X \) and \( Y \) may be either positively or negatively correlated depending on the hormone level. If the hormone level fluctuates in an unspecified manner, the opposing directions of correlation may cancel out each other, and no similarity-based analysis may succeed in detecting the functional association between \( X \) and \( Y \).

A general issue arising from the above discussion is how to systematically study the coexpression patterns between functionally related genes as the cellular-state changes. The issue is compounded by the numerous intracellular and intercellular conditions that can alter the cellular state. A direct approach would be to specify a number of them and conduct more profiling accordingly, but this depends on the resource availability and the knowledge about what conditions are relevant.

*This paper was submitted directly (Track II) to the PNAS office.

Abbreviations: MIIPS, Munich Information Center for Protein Sequences; TR, thyroid hormone receptor; LA, liquid association; LAP, LA pair; PFK, 6-phosphofructokinase.

†E-mail: kcli@stat.ucla.edu.
The correlation of \((X, Y)\), and the LA score is positive. Thus when comparing the low with the high expression levels of a positive LA-scouting gene, the scouted LAP is likely to change from being contraexpressed to being coexpressed. For a negative LA-scouting gene, the change goes in the opposite direction: from coexpression to contraexpression. In general, an LA-scouting gene serves only as a whistle blower, a surrogate for the intrinsic-state variable that facilitates the LA activity. The protein encoded by an LA-scouting gene may not have the direct physical contact with its LAP or the proteins encoded. The word “scouting” may be replaced by “monitoring,” “mediating,” or their synonyms with the above note in mind.

For the genome-wide study, there is still a computational hurdle to overcome, because there are too many combinations for choosing three genes from \(N\) genes. With \(N = 5,878\), the number is already hovering above 33 billion. We surely cannot afford to inspect every scatter plot to find all triplets with LA patterns, which is why an LA score is needed and must be easy to compute. Aided by a statistical theory as outlined next, this turns out to be possible. In fact, only two simple steps are required.

\(i\) Standardize each gene-expression profile with a normal score transformation. Specifically, the \(m\) values in the profile are compared with each other and their ranks \(R_1, \ldots, R_m\) are recorded. The ranks are then used to obtain the transformed profile, \(\Phi^{-1}(R_1/(m + 1)), \Phi^{-1}(R_2/(m + 1)), \ldots, \Phi^{-1}(R_m/(m + 1))\), where \(\Phi(.)\) is the cumulative normal distribution.

\(ii\) Compute the average product of the three transformed profiles,

\[
(X_1 Y_1 Z_1 + \cdots + X_m Y_m Z_m)/m,
\]

which is the LA score that we need.

As illustrated in the flow chart (Fig. 3), there are several ways of using the LA approach. Two of them are given in this article:
(i) screen the genome to find the LA-scouting genes for a given pair of genes (an order of $N$ comparisons is required), or
(ii) screen the genome to find the LAPs of a given gene (an order of $N^2$ comparisons is required).

Statistical Development of LA Score. Change of the cellular state is most likely a continuous process. It is convenient to present the theory in terms of continuous random variables. Suppose $X$, $Y$, and $Z$ are already standardized to have mean 0 and variance 1. Then the correlation coefficient between $X$ and $Y$ is equal to $E(XY)$. By conditioning, $E(XY) = E(E(XY|Z)) = Eg(Z)$, where $g(z) = E(XY|Z = z)$ denotes the conditional expectation of $XY$ given $Z = z$. The notion of LA now arises from a straightforward dynamic perspective. We regard $g(z)$ as the coexpression measure between gene $X$ and gene $Y$ when $Z$ is at level $z$ and ask how $g(z)$ varies as $z$ increases. Denote the derivative of $g(z)$ with respect to $z$ by $g'(z)$, which leads to the following definition for measuring the average amount of coexpression change.

Definition: Suppose $X$, $Y$, and $Z$ are random variables with mean 0 and variance 1. The LA of $X$ and $Y$ with respect to $Z$, denoted by $LA(X,Y|Z)$, is

$$LA(X,Y|Z) = Eg'(Z),$$  

where

$$g(z) = E(XY|Z = z).$$

This definition is fairly general. It can be estimated by applying one of the many existing nonparametric regression techniques to estimate the curve $g(z)$ first, but here is a shortcut. We assume that $Z$ follows a normal distribution and obtain the following.

Theorem. If $Z$ is standard normal, $LA(X,Y|Z) = E(XYZ)$.

Using the celebrated Stein Lemma (12), the theorem follows from

$$Eg'(Z) = Eg(Z)E = E(XYZ).$$

This theorem justifies why we need the normal score transformation as the first step in computing the LA score. Another advantage is that the interference of abnormally large values is tempered, because only the ranks of the expression levels are used in the transformation. More statistical discussion is given in Supporting Text, which is published on the PNAS web site, www.pnas.org.

We use a permutation test to tell whether the LA score is statistically significant. The procedure generates as many as $10^3$ or $10^4$ artificial profiles $Z^*$ of $Z$ by randomly permuting $(Z_1, \ldots, Z_n)$ and computes their LA scores. The $P$ value is obtained by counting how often the LA score of $(X,Y|Z^*)$ exceeds the LA score of $(X,Y|Z)$.

Results

Urea Cycle. Fig. 4 shows the key enzymes and intermediates in the urea cycle. First, we consider two functionally associated genes, $GLN3$ and $CAR1$. CAR1 encodes arginase, the enzyme catalyzing the hydrolysis of arginine into urea, whereas GLN3 is a transcription factor known to activate nitrogen-catabolic genes including $CAR1$ (13). Intuitively we may expect a positive correlation between the two profiles, but this is not the case; very little similarity exists between the two expression profiles. Perhaps this merely reflects the complexity due to many cis elements and trans-acting factors for $CAR1$ (14), the localization of Gin3p, and its interactions with other factors (15). To apply the LA methodology, we take $(GLN3,CAR1)$ as the pair $(X,Y)$ in Fig. 2, take any gene as $Z$, and compute the LA score as defined earlier. We then rank all 5,878 genes according to their abilities to serve as a positive or negative LA-scouting gene by ordering their LA scores. It turns out that at the eighth place leading on the negative end is $ARG4$, the gene immediately ahead of $CAR1$ in the urea cycle. $ARG4$ and its two preceding genes in the flow chart, $ARG1$ and $ARG3$, share some mild degree of profile similarity. However, these three genes are not coexpressed with $ARG2$, the gene encoding acetylglutamate synthase, which catalyzes the first step in synthesizing ornithine from glutamate. Ornithine and carbamoyl phosphate are the substrates of the enzyme ornithine transcarbamoylase, encoded by ARG3. Carbamoyl phosphate synthetase is encoded by CPA1 and CPA2. ARG1 encodes argininosuccinate synthetase, ARG4 encodes argininosuccinase, CAR1 encodes arginase, and CAR2 encodes ornithine aminotransferase.

Fig. 4. Urea cycle/arginine biosynthesis pathway. ARG2 encodes acetylglutamate synthase, which catalyzes the first step in synthesizing ornithine from glutamate. Ornithine and carbamoyl phosphate are the substrates of the enzyme ornithine transcarbamoylase, encoded by ARG3. Carbamoyl phosphate synthetase is encoded by CPA1 and CPA2. ARG1 encodes argininosuccinate synthetase, ARG4 encodes argininosuccinase, CAR1 encodes arginase, and CAR2 encodes ornithine aminotransferase.
encodes a serine pathway (18). Because the nitrogen activation of the fermentable growth median-induced and treating them with heat stress for 1 h and the superoxide-long-lived mutants after transposon-mutagenizing yeast cells and longevity physiology. likely association between the efficient usage of the urea cycle nondividing life span than the wild type (17), this brings out a SCH9 case when the expression of also be negatively correlated with CAR1 ARG2 a way that not only positive correlation exists among entire urea cycle genes should be expressed coherently in such a thesis. To continue the discussion, we take note that ideally, the above study about yeast web site). Indeed, this rather unexpected finding has led us to the 8, which is published as supporting information on the PNAS web site. For further study, we compile a database that contains the information about the most significant LAPs detected from each of the 5,878 genes under study. After ranking all possible pairs (>17 million pairs in total) by their LA scores, we kept only the top 20 positive LAPs and top 20 negative LAPs. This LAP database can be accessed at www.stat.ucla.edu/~kcl.

**LAP Database.** For further study, we compile a database that contains the information about the most significant LAPs detected from each of the 5,878 genes under study. After ranking all possible pairs (>17 million pairs in total) by their LA scores, we kept only the top 20 positive LAPs and top 20 negative LAPs. This LAP database can be accessed at www.stat.ucla.edu/~kcl.

Longevity. A recent article (17) reveals the role of SCH9 in longevity of yeast. Checking the LA database that we compiled, the pair (ARG1, ARG2) appears as a negative LAP of SCH9 (Fig. 8, which is published as supporting information on the PNAS web site). Indeed, this rather unexpected finding has led us to the above study about yeast’s rational control of arginine biosynthesis. To continue the discussion, we take note that ideally, the entire urea cycle genes should be expressed coherently in such a way that not only positive correlation exists among ARG1, ARG2, ARG3, and ARG4, but concomitantly these genes should also be negatively correlated with CAR1. This turns out to be the case when the expression of SCH9 is low (see Fig. 6). SCH9 encodes a serine/threonine protein kinase. It is required for the nitrogen activation of the fermentable growth median-induced pathway (18). Because the sch9-delete mutant has a longer nondividing life span than the wild type (17), this brings out a likely association between the efficient usage of the urea cycle and longevity physiology.

In their longevity study, Fabrizio et al. (17) screened for long-lived mutants after transposon-mutagenizing yeast cells and treating them with heat stress for 1 h and the superoxide-generating agent paraquat for 9 days. In fact, two mutants were isolated and the affected genes were SCH9 and CYR1 (encoding adenylate cyclase). The longevity regulation by SCH9 and CYR1 is consistent with the popular oxidative-damage theory of aging (19). When checking for the LAPs of CYR1, we found an oxidative protection gene, TTR1/GRX2 (glutaredoxin), and two age-related genes, SIM1 and HST3. HST3 is a homolog of the gene SIR2, which encodes NAD-dependent histone deacetylase and regulates the replicative life span of yeast (20). Interestingly, ARG4 was found in one LAP of SIR2. In addition, from the LAPs of the last urea cycle gene, CAR1, we found GLN3, CPA1, and CPA2, which took us back to what was discussed earlier. On the other hand, we notice that glutamate, a product of most amino acid deamination, is at the head of the ornithine/arginine biosynthesis/urea cycle pathway (Fig. 4). From the LAPs of
CYR1, we find five genes involved in glutamate metabolism and nitrogen utilization: PUT2, LYS9, AR09, GAPI, and MEP2. Standard profile-similarity analysis also helps; SCH9 is positively correlated with CAR2. Thus, lowering the SCH9 expression is concomitant with the down-regulation of CAR2, thereby preventing the leakage of ornithine from the urea cycle. Before leaving this topic, we note that what is described here is not about the physiology for deletion mutants in the nondividing state. For our data, the yeast cells harvested for mRNA samples have just re-entered the cell cycle from the arrested states; they are growing and dividing.

Electron Transport. Mitochondria generate energy via the process of oxidative phosphorylation (21). During the process, electrons are passed along a series of respiratory enzyme complexes located in the inner mitochondrial membrane by using the released energy to pump protons across the membrane. The proton gradient then enables the making of ATP by ATP synthase (F1Fo ATPase, or complex V). Many of the encoding genes in the pathway indeed are coexpressed. One interesting example is cytochrome e1, which acts as the last leg for complex III (cytochrome bc1) in relaying electrons to cytochrome c. For S. cerevisiae, cytochrome c1 is encoded by a single gene, CYT1. From the profile-similarity analysis, we find one gene, QCR8 (ubiquinol cytochrome reductase subunit 8), from complex III and two genes, COX5a and COX8, from complex IV (cytochrome oxidase) to be among the top 20 genes with expression profiles most similar to CYT1. Complementing this finding, the LA method links CYT1 to genes further down the electron-transport pathway. From the 40 most significant LAPs of CYT1 (Table 2, which is published as supporting information on the PNAS web site), we find that two of them, ATP1 and ATP5, encode subunits from the mitochondrial ATP synthase. In fact, ATP1 (F1 α subunit) appears 11 times.

Calmodulin. Calmodulin is a ubiquitous Ca2+-binding protein that regulates a wide range of proteins and processes in all eukaryotes. It is encoded by CMD1 in S. cerevisiae (22). The best known binding target of calmodulin in mitosis is Nuf1p. Nuf1p is a component of spindle pole body, the yeast microtubule-organizing center wherein calmodulin is localized. Despite the functional association between their gene products, CMD1 and NUF1 do not have similar expression profiles. NUF1 does not appear in the 40 LAPs of CMD1, either (Table 3, which is published as supporting information on the PNAS web site). Instead, we find it paired with an unknown gene, YGL149W, as a positive LAP of CMK1, in the top 10 genes encoding the calmodulin-regulated protein kinase. More interestingly, YGL149W also appears in the 20 positive LAPs of CMD1. Indeed, although the profiles of CMD1 and CMK1 are not similar, they do share nine genes in their LAPs. Furthermore, YGL149W has the protein–protein interaction with CRM1p (β-karyopherin, involved in exporting certain proteins from nucleus) that in turn interacts with NUF1p. Consistent with this connection, we find KAP120 (a member of karyopherin family) in the LAPs of CMD1 and SXM1 (putative β-karyopherin) in the LAPs of CMK1. From the joint LAP lists for CMD1 and CMK1, we further notice the convergence of a number of genes involved in the signal transduction pathways for various types of growth morphogenesis such as sporulation, pheromone response, pseudohyphal growth, RAS protein signal transduction, mitogen-activated protein kinase, cAMP-protein kinase A, and high-osmolality glycerol-response pathways: IME1, RIM13, SPS1, SOK2, MUC1, TEC1, PTP3, PTC1, SLP4, CDA2, IME4, KAR4, KAR2, DOC1, AXL1, OPY1, and ERF2.

Glycolysis. The genome-wide LA-scouting ability of a gene can be assessed according to the number of its LAPs with scores exceeding a threshold. This ability varies substantially from gene to gene (Table 4, which is published as supporting information on the PNAS web site). Reasoning that each surviving species must evolve a delicate expression system to manage the intricate interplay in the gene products, we speculate that the LA-scouting ability of a gene should reflect the importance of its encoded protein. To support this statement, we compare the genes involved in the most well studied metabolic pathway, glycolysis. The principal rate-limiting enzyme in glycolysis is 6-phosphofructokinase (PFK), a hetero-octamer enzyme of four α and four β chains. It turns out that PFK1 and PFK2, encoding the α and β subunits of PFK, do detect far more LAPs than any other genes listed by the Munich Information Center for Protein Sequences (MIPS) in the LAP category of glycolysis and gluconeogenesis (see Table 5, which is published as supporting information on the PNAS web site).

LA-Scouting Leaders. Using some highly stringent criteria, we have selected a set of 66 genes with the highest LA-scouting ability (Table 6, which is published as supporting information on the PNAS web site). Some of these LA-scouting leaders are well studied genes [for instance, CYR1, PFK1, TPS1 (trehalose-6-phosphate synthase), TPS2 (trehalose-6-phosphate phosphatase), GSY1 (glycogen synthase isoform 1), GLC3 (α-1,4-glucan branching enzyme), ATP1, PPA1 (subunit of vacuolar ATPase), OCR9 (ubiquinol cytochrome c reductase subunit 9), CYC7 (cytochrome c, isoform 2), EGR13 (3-hydroxy-3methylglutaryl CoA synthase), AAT2 (aspartate aminotransferase), AP CI (largest subunit of anaphase-promoting complex), and MEI1 (positive regulator of sporulation genes)]. The selection of LA-scouting leaders is by no means definitive. GSY2 (dominant isoform of glycogen synthase) just misses the cut. PFK2 and YAP6 would have been included if not for much missing data in one of the four cell-cycle experiments. YAP6 encodes a transcription factor homologous to YAP1p (basic leucine zipper, yeast homolog of c-Jun). Overexpression of YAP6 increases resistance to cisplatin (23), one of the most widely used drugs for cancer chemotherapy. The LA-scouting leaders are often found to scout each other. For example, among the top 10 LA-negative and LA-positive scouting genes for (CYR1, PFK1), 7 are LA-scouting leaders (see also Fig. 9, which is published as supporting information on the PNAS web site).

Prediction. The LA system provides a portrait of the cellular context in which a gene is likely to be involved, which is useful for predicting the functions of little known genes. For example, consider one of the LA-scouting leaders, ECM1. The literature on ECM1 is quite thin, and its functional annotation is still vaguely put as “involved in cell wall biosynthesis” by MIPS and SGD (Saccharomyces Genome Database) (wording slightly different). A year ago when the LA idea was first conceived, we were wondering why, among the scouting targets of ECM1, several are involved in translation; for example, NIP7 (required for efficient 60S ribosome subunits biogenesis, delete-lethal), appears six times (Table 7, which is published as supporting information on the PNAS web site). Now we find this to be consistent with the recent finding that ECM1p tagged with GFP is localized in the nucleus with a mild enrichment in the nucleolus, and that ECM1p genetically interacts with MTR2p in the 60S ribosomal protein subunit export (24). Coincidently, we find that NIP7p appears as one of the 23 proteins copurified with NUG1p, the main gene characterized in ref. 24. The chance for this coincidence to occur by a fluke is <0.5%. This reassures the biological relevance of the LA approach in protein-function prediction.

YNL101W is another LA-scouting leader that is still considered as an unknown gene in SGD and MIPS at this writing. From the list of LAPs for YNL101W (Table 8, which is published as supporting information on the PNAS web site), we find several genes involved in autophagy, protein degradation, and transport:
**AU77** (essential for autophagy, appearing six times), **PRE6** and **PUP1** (20S proteasome subunits), **RPT6** (26S proteasome-regulatory subunit), **CLC1** (clathrin light chain), **YPT52** (GTP-binding protein of the RAB family), and **UBP2** (ubiquitin-specific proteasine). This seems consistent with the newly identified role of **YNL101W** (newly named **AVT4**) as a membrane transporter responsible for the efflux of tyrosine and other large neutral amino acids from the vacuole (25).

**Discussion**

From the modeling point of view, how to infer the underlying cellular program from the mRNA data is extremely challenging because (i) microarray measurements are surely noisy, (ii) protein abundance may not be reflected well enough by the mRNA level, because other factors such as tRNA charging, protein stability, and so on are not considered, and (iii) intrinsic variables such as the protein phosphorylation or other modification status, localization of transcription factors, and quantities of important molecules such as ATP, NADPH, CAMP, etc. are difficult to predict. Because mathematical models seem intractable, statistical approaches become popular. Many similarity-based methods have proved very useful in elucidating the expression data.

Taking one step further, the LA method conducts a genome-wide search and identifies the most critical cellular players that may affect the coexpression pattern for any two genes. This method can be applied in any large microarray study. The aim of LA is to explore and exploit the dynamic, as opposed to the static, aspect of gene expression in cells. Our method eliminates the need to specify the cellular state before applying it. Instead, the method provides results for portraying the intrinsic state that facilitates the coexpression changes. This in turn can be used to guide the specification of experimental conditions for conducting more microarray profiling.

We are able to demonstrate some applications of LA using the yeast data. We show how the proper flow of ornithine in the urea cycle is mediated through a delicate switch between coexpression and contraexpression of **ARG2** and **CAR2**. The switch depends on the expression level of **CFA2**, and it reflects well the cellular need for arginine. We also find that the regulation of **GLN3** on **CAR1** is linked to **ARG4**. Moreover, the expression of **SCH9** is associated with the change in the coregulation pattern of **ARG1** and **ARG2**. A possible connection of efficient expression of the urea cycle to yeast longevity is inferred.

For mammals, the urea cycle is activated in the liver to excrete excessive nitrogen resulting from the metabolic breakdown of amino acids. Our body cannot use the arginine synthesized in the liver because it is immediately cleaved to form urea, which then is sequestered by the kidney for secretion in the urine. In contrast, arginine biosynthesis is very important for single-celled organisms such as yeast. Thus it would be interesting to find out how different the coexpression pattern is in the liver if such expression data are available in the future.

This work points to a new source of information hidden in the microarray data. Methodologically, it can be viewed as one that offers a strategy of data refinery. The original expression data are processed, and the information about the LA activities is distilled. For **N** genes, the algorithm returns a huge amount of message, in the order of **N³**, that can be stored and used in a variety of ways to meet different researchers’ needs. In our illustrations, we only use a small portion of high-scoring LAPs. In general, the better the LA score is, the more likely we can detect the LA pattern when visually examining the profile plots. It is in this sense that leading LA-scouring genes are better surrogates of the relevant intrinsic cellular-state variable. But how we use these surrogates to infer the cellular state depends on the available biological knowledge. The state variable can serve as a conceptual device for bringing out plausible biological hypotheses that can be cross-examined by using other bioinformatic resources such as the transcription factor database TRANSFAC (26) or the protein–protein interaction/complex database of MIPS.

The postrefinery statistical analysis can be extended in several directions. In addition to **P** values and visual inspection, one may bring in methods from multiple comparison/false discovery, an area with renewed interest fueled by microarray analysis. Similarity-based methods such as principal component analysis, also known as singular value decomposition, can be applied to high LA-scoring genes. In addition to the yeast data, our method can be applied to other large microarray studies on cancers, cell lines, and drug sensitivity (Supporting Text).

I am grateful to David Eisenberg, Samuel Karlin, and Michael Waterman for inspiring discussion; Chun-Houh Chen and Robert Yuan for computing assistance; Don Ylivisker and Jack and Jennie Li for editing; and two referees for useful suggestions on presentation. This work is supported in part by National Science Foundation Grants DMS-9901549, 0104038, and 0201005.