Research

Vector algebra in the analysis of genome-wide expression data

Finny G Kuruvilla*, Peter J Park† and Stuart L Schreiber*

Addresses: *Howard Hughes Medical Institute, Bauer Center for Genomics Research, Department of Chemistry and Chemical Biology, Harvard University, Cambridge, MA 02138, USA. †Department of Biostatistics, Harvard School of Public Health, Informatics Program, Children's Hospital, Harvard Medical School, Boston, MA 02115, USA.

Correspondence: Stuart L Schreiber. E-mail: sls@slsiris.harvard.edu

Published: 13 February 2002

Genome ***Biology*** 2002, **3(3)**:research0011.1–0011.11

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/3/research/0011

Received: 20 August 2001
Revised: 14 December 2001
Accepted: 11 January 2002

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Abstract

Background: Data from thousands of transcription-profiling experiments in organisms ranging from yeast to humans are now publicly available. How best to analyze these data remains an important challenge. A variety of tools have been used for this purpose, including hierarchical clustering, self-organizing maps and principal components analysis. In particular, concepts from vector algebra have proven useful in the study of genome-wide expression data.

Results: Here we present a framework based on vector algebra for the analysis of transcription profiles that is geometrically intuitive and computationally efficient. Concepts in vector algebra such as angles, magnitudes, subspaces, singular value decomposition, bases and projections have natural and powerful interpretations in the analysis of microarray data. Angles in particular offer a rigorous method of defining 'similarity' and are useful in evaluating the claims of a microarray-based study. We present a sample analysis of cells treated with rapamycin, an immunosuppressant whose effects have been extensively studied with microarrays. In addition, the algebraic concept of a basis for a space affords the opportunity to simplify data analysis and uncover a limited number of expression vectors to span the transcriptional range of cell behavior.

Conclusions: This framework represents a compact, powerful and scalable construction for analysis and computation. As the amount of microarray data in the public domain grows, these vector-based methods are relevant in determining statistical significance. These approaches are also well suited to extract biologically meaningful information in the analysis of signaling networks.

Background

The three goals of most microarray experiments are description, classification, or characterization. For example, microarrays have been used to describe comprehensively the diauxic shift, progression through the cell cycle, sporulation and the effects of treatment with a small molecule [1-4]. They have been used to classify the cancer type of a given sample or classify groups of co-regulated genes [5-8]. Finally, microarrays have been used to characterize biological systems using comparisons of wild-type and mutant cells, with the goal of obtaining mechanistic insights [9-11].

Searching for concerted, dramatic changes in gene expression or searching for differential expression of a given gene has been a successful method in analyzing transcription-profiling data, especially in description or characterization [4,12]. Often, investigators take a manual approach to accomplishing these tasks. However, manual approaches to data analysis are sometimes impractical or cumbersome, inspiring the development of tools to accomplish the three goals described above. A variety of techniques such as hierarchical clustering, *k*-means clustering and self-organizing maps have been implemented with success, especially in classification [13].

As the number of publicly available profiles in *Saccharomyces cerevisiae* alone now exceeds 500, a great need exists to exploit this information properly to understand cell function. At least three independent international projects have been set up to serve as database-driven repositories of genome-wide expression data [14]. A major effort is being made to systematize data storage, especially involving XML (extensible markup language), to ensure interoperability of these databases and associated analysis tools.

A related need that has been less addressed is the systematization of expression data analysis. This requirement extends not only to analysis but also to pedagogy and to practical aspects of algorithm implementation. Various studies in the literature have successfully implemented tools from vector algebra in analyzing genome-wide expression data [11,15,16]. However, a framework for the analysis of transcription profiles using vector algebra has not yet been codified. Here we present such a framework. Common statistical measures have natural counterparts in vector algebra that have visual interpretations and are easily implemented on a computer. Within this framework, the analysis of genome-wide expression data is converted to the study of high-dimensional vector spaces. The many powerful theorems that have been developed in vector algebra can be applied to these spaces, and these theorems offer biologically relevant insights. Elements of the vector space can also be analyzed statistically. This construction has analytic and pedagogic appeal.

Results and discussion

Constructing expression vectors

Transcription-profiling experiments offer different kinds of measurements depending on the technology used. One type of technology (using noncompetitive hybridization) measures absolute gene expression, while a second type (using competitive hybridization) measures relative gene expression. A microarray that uses competitive hybridization yields a list of fold changes for each gene between the conditions or cell types measured. The data from a microarray that uses noncompetitive hybridization can be 'divided' (after proper normalization) into another microarray of the same type to produce the fold change of each gene from one condition to another. In practice, even those investigators who use noncompetitive technology platforms generally interpret and publish the fold changes between conditions or strains rather than absolute gene-expression levels. Therefore this study will focus on analysis of fold-change values, hereafter called a transcription or expression profile.

There are three common types of values that can be associated with the fold change of a gene. The first, a signed fold change (such as a +1.6 or -2.3 fold change, corresponding to induction or repression, respectively) has the most intuitive appeal but has a discontinuity spanning from -1.0 to +1.0 that can be problematic. The second type, an unsigned fold change (such as 1.6 or 0.43 fold change, again corresponding to induction or repression, respectively) has no such discontinuity but is bounded on the left by zero and is unbounded on the right. This asymmetry about unity hinders analysis. The third type, a logarithm (base 2, base 10, or natural) of the unsigned fold change, is undoubtedly the most tractable. No fold change in expression is represented by zero, induction is positive and repression is negative. Most importantly, there are no discontinuities or asymmetries.

Once a set of transcription profiles is obtained (where each profile represents a set of fold changes for a set of genes), a data matrix can be generated with genes as rows and experiments as columns (Figure 1a). With this picture in mind, generating vectors from the data matrix (the logarithms of the unsigned fold changes) is a natural operation for which there is an obvious construction.

Definition. For p transcription profiles measuring the expression of n genes, let s_{ij} represent the logarithm of the expression of the ith gene in the jth experiment. The vector

$$[s_{i1}\, s_{i2} \ldots s_{ip}]^{T}$$

is defined as a 'gene vector'. Similarly,

$$\begin{bmatrix} s_{1j} \\ s_{2j} \\ \vdots \\ \vdots \\ s_{nj} \end{bmatrix}$$

is defined as an 'experiment vector'. By construction, a gene vector is p-dimensional and an experiment vector is n-dimensional. An 'expression vector' represents either a gene vector or an experiment vector. The superscript ^{T} designates transposition.

It can easily be seen that with respect to the data matrix, a gene vector consists of the elements of a row and an experiment vector consists of the elements of a column (Figure 1a). By convention, vectors are treated as columns and not as rows. Experiment vectors are already columns, so gene vectors are transposed from rows to also become columns.

In the most common scheme of comparing two profiles, a 'cloud' of genes is plotted on a scatterplot. This view is simply the display of n two-dimensional gene vectors (Figure 1b). It is well suited to detecting particular genes whose expression varies between experiments. As more experiments are added to the matrix, the dimension of the space grows but the number of points remains fixed. Considering a large number of experiments, genes that have similar biological regulation remain in nearby regions of high-dimensional

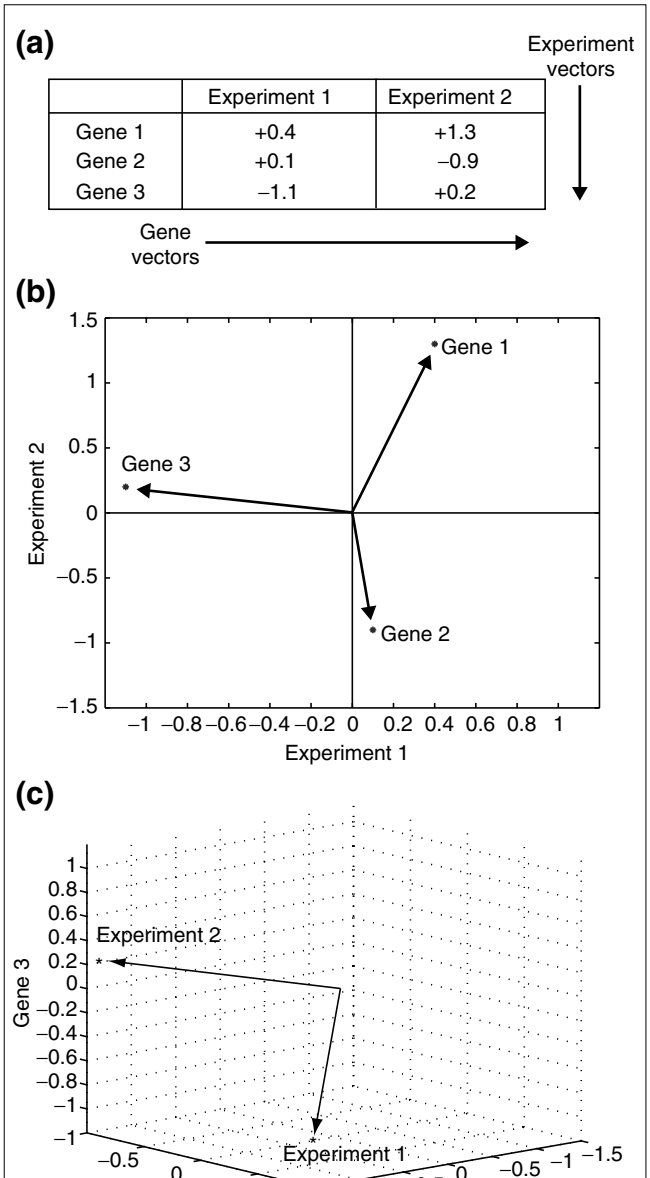

Figure 1
The two complementary methods of understanding a transcription profile. **(a)** Two transcription profiles of three genes are shown. Rows form gene vectors while columns form experiment vectors. **(b)** In a typical profile comparison, gene vectors are plotted in two dimensions, where the axes represent the experiments and the points are the genes. Data from (a) are plotted. Additional genes add points to the graph, but it remains in two dimensions. **(c)** In the vector-based approach, the axes are genes and the points are experiments. Data from (a) are also plotted here. Additional genes would not add any points to the graph (that is, there would always only be two vectors), but the space that the vectors reside in would increase in dimension.

spaces. Integrated with promoter-sequence analysis and consideration of chromatin structure, the study of gene vectors is useful in studies of transcription regulation.

There is another, complementary, way of thinking about expression profiles, or more generally, about multivariate statistical data [17]. Although analysis of gene vectors is a powerful tool, it is less efficient at demonstrating relationships between profiles. This is because overlaying clouds of thousands of gene vectors may not offer an insightful picture because of the many data points in view. In the alternative method of analysis, an expression profile is regarded as just one point in a high-dimensional space. Because this point captures the same information contained in the scatterplot of gene vectors, the point must reside in a much higher-dimensional space. In this space, the relationships between profiles become more apparent. Instead of plotting genes, one now plots experiments (Figure 1c). As new profiles are added to the data matrix, the dimension of the space remains fixed but the number of points increases, the reverse case of gene vectors. Although the high-dimensional space of experiment vectors cannot be visualized, the relationships of vectors in the space can often be understood using intuition from the two- or three-dimensional case. The power of vector algebra comes from its ability to scale - the concepts, equations and theorems move seamlessly (usually identically) to higher dimensions.

Assuming that there are n genes represented on the microarray, it is theoretically possible that each profile would yield an n-dimensional experiment vector. However, because of poor detection of scarce transcripts, flaws on the microarray or experimental error, an expression profile in practice gives fewer than n values. This leaves three basic options for how to carry out the analysis. The first is that a lack of data can be declared as no change in expression, and a zero inserted into the position of that gene. This approach is generally too simplistic for use in further analysis.

The second option is to estimate the missing value on the basis of other profiles and insert it into the vector. For example, we can look for the gene most similar to the missing gene as judged by all other experiments, and then replace the missing value with the value of the similar gene for that experiment. For robustness, we can also find the most similar k genes and use the average of those genes for that experiment as the replacement. A comparative study has found that this k-nearest neighbors method performs better than filling missing values with zeros or the averages of that gene in other experiments [18]. There are more complicated imputation methods in the statistics literature, but their usefulness, given the complex calculations involved, is not yet clear.

Finally, the vector space can be resized to a smaller dimension than n, where the expression profiles of interest have data for all genes in that space. As the missing data are not estimated, this is the most conservative method of analysis, although at the price of discarding some data.

Vector angle is the counterpart to the Pearson correlation coefficient

One of the most fundamental measures in statistics is the Pearson correlation coefficient (referred to as r). This represents a widely used measure of similarity between n-tuples. With expression vectors, the geometric picture of similarity is that the vectors are pointing in the same direction of expression space. This picture is equivalent to two vectors with a small angle between them. When expression vectors are uncorrelated, they are orthogonal. This leads to the following remark.

Remark 1. Correlation of expression vectors corresponds to the angle θ between those vectors. Low angle (r near 1) implies correlation, Angle near 90° (r near zero) implies uncorrelation, and angle near 180° (r near –1) implies anticorrelation.

This remark is motivated by consideration of the relationship between the Pearson correlation coefficient (r) and vector angle (θ). (We use the notation where < , > represents an inner product, also called a dot product, and where cov(,) is the covariance between ordered n-tuples.) For two vectors $\mathbf{x}$ and $\mathbf{y}$, the formulas for r and θ are:

$$r = \frac{\mathrm{cov}(\mathbf{x},\mathbf{y})}{\sqrt{\mathrm{cov}(\mathbf{x},\mathbf{x}) \cdot \mathrm{cov}(\mathbf{y},\mathbf{y})}} \qquad \theta = \cos^{-1}\left(\frac{\langle \mathbf{x},\mathbf{y} \rangle}{\sqrt{\langle \mathbf{x},\mathbf{x} \rangle \cdot \langle \mathbf{y},\mathbf{y} \rangle}}\right)$$

It is evident that r and the argument of the inverse cosine function are identical except that covariance and inner product are interchanged. In practice, the argument to the inverse cosine and r are nearly identical because the covariance is essentially a mean adjusted inner product and the means of expression vectors are near zero. (In fact, covariance and inner product are identical for vectors with zero mean.) Thus, there is no advantage of either measure because of nearly identical performance on actual expression data (Figure 2a).

For gene vectors, angles are a measure of how similar two genes are expressed across experiments. For experiment vectors, angles are a measure of how similar two experiments are across genes. For experiment vectors, a quantitative sense of the distribution of θ can be obtained by examining a collection of diverse transcription profiles. A set of 300 transcription profiles in *S. cerevisiae* from a wide variety of gene deletions and treatments was obtained and made publicly available [19]. We converted these profiles into experiment vectors and computed all possible angles between the vectors (a total of 44,850 angles). (Computer source written in Matlab™ is publicly available at our website [20]) These angles gave a distribution with a mean of 88.4° (close to orthogonal, as expected) and a standard deviation of 7.4° (Figure 2b). In practice, we have found 60° and 120° as useful cutoff values for significance of correlation or anticorrelation, respectively.

The vector algebra approach is illustrated using a publicly available data set. The small molecule rapamycin has been transcriptionally profiled by three research groups in four separate studies [4,11,21,22]. It has a very dramatic expression profile in which hundreds of genes are rapidly (within minutes) up- or down-regulated. The protein targets of rapamycin (the Tor proteins) are known to sense nutrients. When treated as an experiment vector, other experiment vectors were searched for that had a low angle with the rapamycin vector. Against the same set of 300 diverse expression profiles described earlier [19], not a single vector could be found that had an angle less than 60° with the rapamycin experiment vector (data not shown). However another study identified two vectors with smaller angles to the rapamycin expression vector (angles of 44° and 47°) [11]. These similar vectors were those corresponding to the removal of high-quality carbon or nitrogen from the media. Thus, it can be inferred that the Tor proteins regulate the responses to carbon- and nitrogen-quality of the medium. This is an example of using angles to identify the functions of uncharacterized proteins.

It is apparent that angles provide a quantitative measure for asserting genome-wide similarities. This measure is important when making claims about the target of a small molecule, effects of gene deletion or specificity of a drug. Orthogonal angles (implying uncorrelation) when similarity is expected or claimed raise questions about the validity of the hypothesis under examination.

A ratio of magnitudes is a natural second metric for comparing two expression vectors

An advantage of vector algebra is the extension from angles into ratios of magnitudes for which other formulations have awkward or no counterparts. The two four-dimensional vectors $\mathbf{x} = [1\ 0\ 2\ -1]^T$ and $\mathbf{y} = [2\ 0\ 4\ -2]^T$ have $\theta = 0$ (implying $r = 1$). However, there is an important difference between the vectors not identified by these measures, as we can observe that $\mathbf{y} = 2\mathbf{x}$. This information is captured by computing the ratio of vector magnitudes.

Remark 2. Besides angle, the ratio of vector magnitudes (α) is a second measure of similarity of two expression vectors. For two vectors $\mathbf{x}$ and $\mathbf{y}$, this is calculated as:

$$\alpha = \sqrt{\frac{\langle \mathbf{y},\mathbf{y} \rangle}{\langle \mathbf{x},\mathbf{x} \rangle}}$$

With the two four-dimensional vectors listed above, α is easily seen to be 2. By interchanging < , > and cov(,) as was done with r and θ, it is tempting to think that the quantity $\beta = (\mathrm{cov}(\mathbf{y},\mathbf{y})/\mathrm{cov}(\mathbf{x},\mathbf{x}))^{1/2}$ might behave similarly

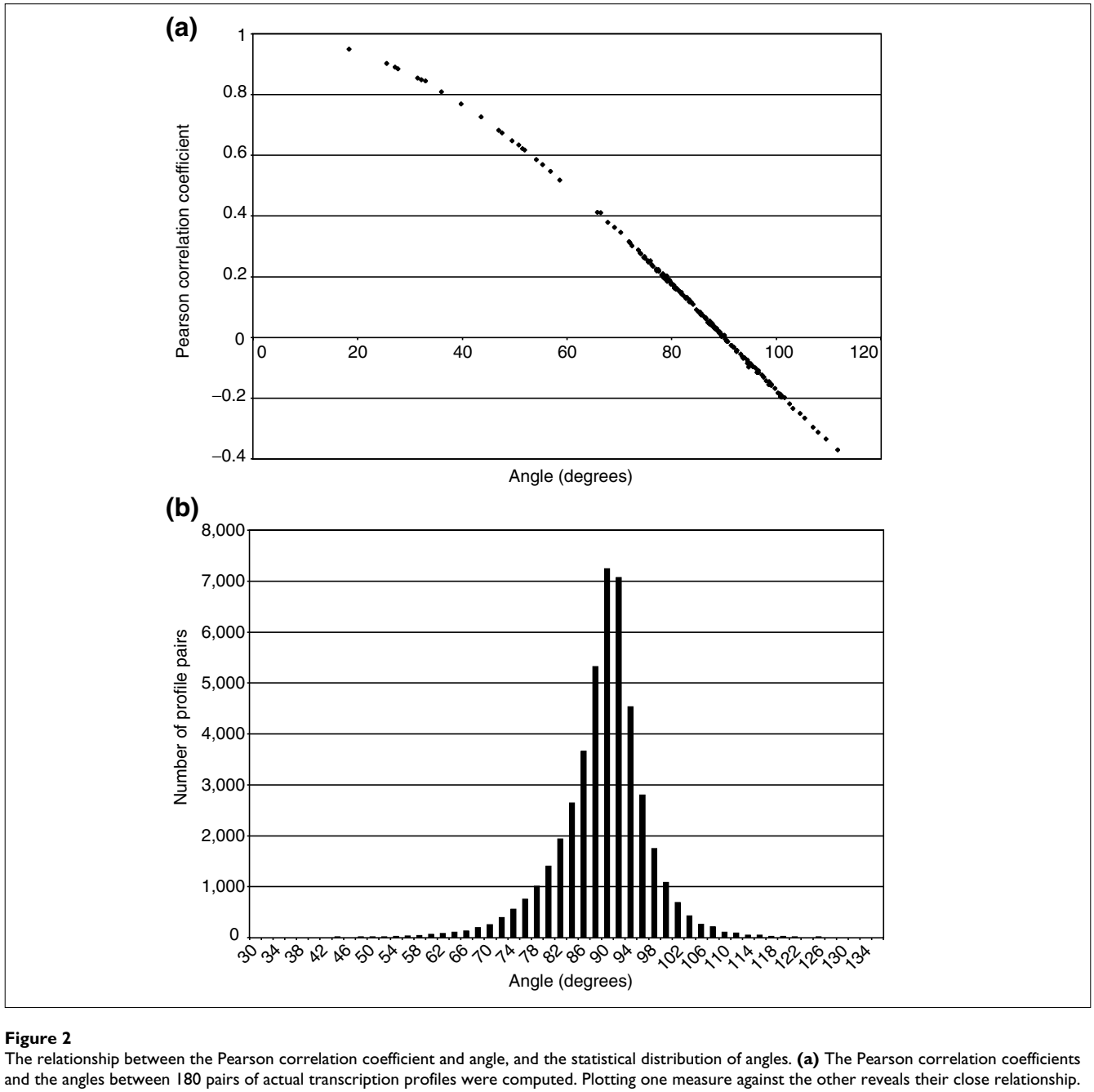

Figure 2
The relationship between the Pearson correlation coefficient and angle, and the statistical distribution of angles. **(a)** The Pearson correlation coefficients and the angles between 180 pairs of actual transcription profiles were computed. Plotting one measure against the other reveals their close relationship. **(b)** The statistical distribution of angles in a large, diverse data set (300 profiles) [19] was computed by calculating angles between all possible pairs (a total of 44,850 angles) of expression vectors. Data were taken from the Rosetta Inpharmatics website [34] and the software used to compute the angles is publicly available for download from our website [20].

to α. With the same vectors as above, $\beta = \alpha = 2$. When considering the two vectors $\mathbf{x} = [2\ 2\ 2\ 2]^T$ and $\mathbf{y} = [1\ 2\ 1\ 2]^T$, it can be calculated that $\alpha = 0.79$, the value expected as $\mathbf{y}$ is a 'smaller' vector. However $\beta = 0.5/0 = \infty$. Indeed, the mean subtraction component of the covariance fundamentally alters the behavior of β compared to α, making it an unsuitable measure of expression vector magnitude. An elegant feature of regarding microarray data as expression vectors is that both angles and magnitudes are natural measurements with important biological meaning, as will be discussed in later sections.

Searching subspaces is an important step in analyzing expression vectors

Another advantage to treating expression data as vectors in a space rather than as n-tuples is the associated notion of

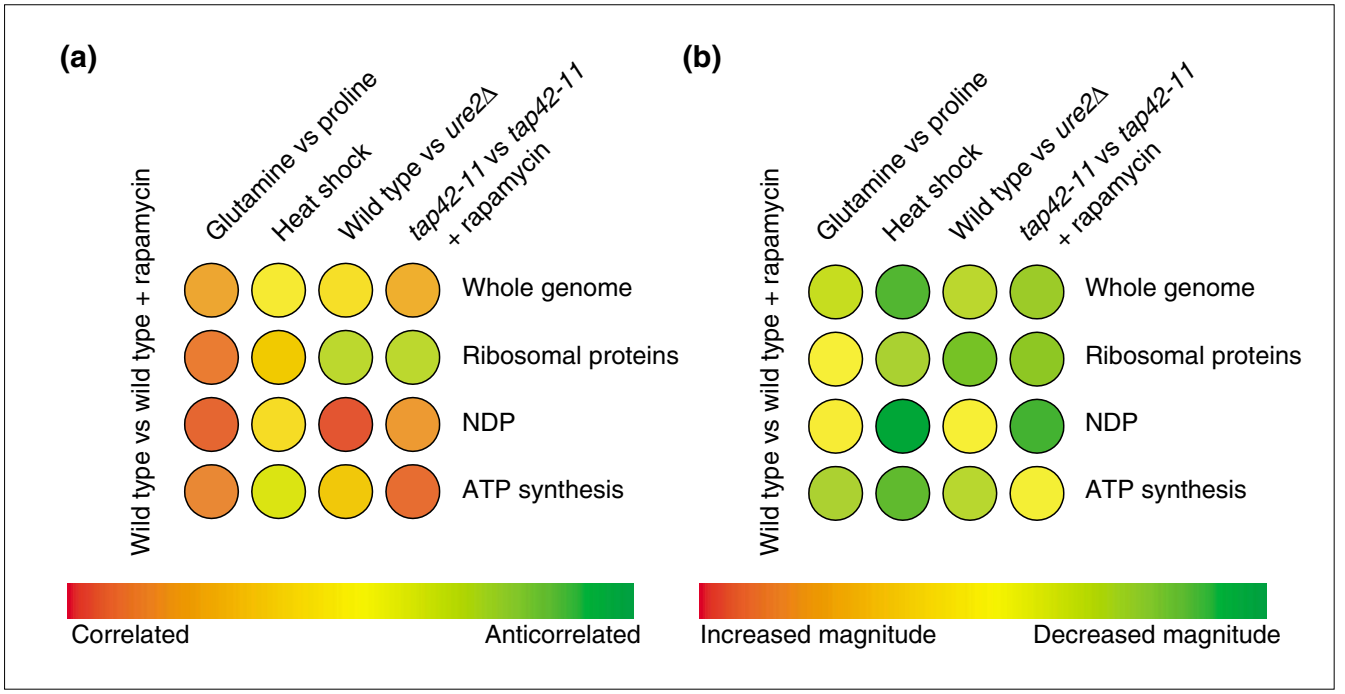

Figure 3
An illustration of the use of the colorimetric comparison array (CCA) to infer information about a cellular signaling network. **(a)** Using competitive hybridization data from a study of the target of rapamycin (Tor) proteins) [11], the angles between a reference expression vector of a wild-type strain treated with rapamycin and four other expression vectors are represented on a CCA. The profile of cells shifted from glutamine to proline shows that at a whole-genome level, and within the subspaces shown, there is strong similarity (44° at a whole-genome level) to a profile of cells treated with rapamycin. Cells that have undergone heat shock, however, have an orthogonal (uncorrelated) expression vector. Deleting the gene *URE2* (*ure2*Δ) does not generate whole-genome correlation, but does generate correlation within the subspace of nitrogen-discrimination pathway (NDP) genes. Finally, cells containing the *tap42-11* allele of *TAP42* that are treated with rapamycin illustrate that Tap42p is downstream of the Tor proteins in the control of ribosomal protein gene expression. **(b)** A similar analysis to that in (a) can be carried out using ratios of vector magnitudes (α). The overall greenness of the CCA reflects the fact that these four expression vectors have a smaller magnitude than the reference rapamycin-treatment profile. The magnitude of the whole-genome expression vector of heat shock is considerably smaller. Deleting *URE2*, however, increases the magnitude of expression of the NDP subspace of genes to levels comparable to those of cells treated with rapamycin. When *TAP42* is mutated to the *tap42-11* allele, NDP gene induction is severely impaired, implicating Tap42p as an important regulator of NDP gene expression downstream of the Tor proteins. Iterating this analysis over many effectors and many subspaces of genes reveals a wealth of information about the transcriptional network downstream of the Tor proteins [11].

subspaces. With experiment vectors, this concept is particularly valuable. Experiment vector space can be divided into subspaces, most naturally where the subspaces correspond to co-regulated or functionally related genes. This can most easily be done using annotated lists where genes are classified by function. The Munich Information Center for Protein Sequences (MIPS) provides a commonly used functional grouping for genes [23].

Considering subspaces is nearly always an essential part of analysis. From the vantage point of certain subspaces, many different experiment vectors may 'collapse' into identical vectors. This enables the decomposition of vectors into sums of common and distinct components. For example, yeast stress responses (such as DNA damage or heat shock) are composed of a general response that stresses share and a specific response for the particular stress faced. Identification of these subspaces remains a goal in biology. Another example is that of the Tor proteins, which have many effectors, one of which is called Ure2p. Comparing the expression vectors of rapamycin treatment with a *URE2* deletion reveals that the genome-wide angle between these two vectors is 76° and thus the vectors show only weak correlation. However, calculating the angle of these same vectors within the subspace of genes controlled by Ure2p gives an angle of 13° [11]. With respect to this subspace of genes, treatment with rapamycin and deletion of *URE2* are nearly identical. Experiments reveal that treatment with rapamycin leads to dephosphorylation of Ure2p, corroborating this low angle [4,21]. Thus it is important to consider not only whole-genome angles, but also those for important subspaces.

The effects of mutations on vector angles and magnitudes

When studying a particular process by transcription profiling, it is common also to transcriptionally profile the same process after deleting one or more genes known to be involved. For example, this has been carried out with MAP

kinase deletion mutants after pheromone treatment [10], deletions of effectors of the Tor proteins after rapamycin treatment [11], and *lexA* deletions after UV irradiation [24].

When an effector of some process is deleted, the resulting vector is often not uncorrelated in the subspace involving that effector, but is reduced in magnitude. For example, when there are two transcription factors with overlapping specificity for some genes, deletion of one can be compensated by the presence of the second with a similar (though weaker) induction of the same genes. If an effector truly is an exclusive controller of some set of genes, when it is deleted, the resulting experiment vector in that subspace would be expected to be orthogonal (uncorrelated) with the original vector. However, biological networks generally exhibit redundancy in their control of genes - a reduction in the vector magnitude is to be more commonly expected than rotation toward perpendicularity (uncorrelation). Thus it is important to rely on the coordinate use of angles and magnitudes in expression-profiling analysis, something that is not always performed in the literature.

Combined use of angles and ratios of vector magnitudes is a tool for biological discovery

When one is characterizing a transcription profile of an uncharacterized gene deletion or treatment with a small molecule, the first operation that should ordinarily be completed is a comparison of that profile as an experiment vector with all other available experiment vectors from the same organism. By performing such a comparison, the gene can be partially characterized or the target of a small molecule can be identified [19]. By coloring the results of a comparison, mimicking the false-coloring schemes usually found on microarrays, one can quickly scan a large set of profiles for those correlated, uncorrelated, or anticorrelated with a reference profile that is being tested [11]. This color scheme is red to represent correlation, yellow to represent uncorrelation, and green to represent anticorrelation (Figure 3). This overall depiction is called a colorimetric comparison array (CCA). It can also be carried out with ratios of vector magnitudes where red represents greater magnitude, yellow unchanged magnitude, and green reduced magnitude.

To understand the most general relationships, it is useful in the first row of a CCA to compare whole-experiment vectors. This gives an important genome-wide characterization of the profiles being studied - assertions of global behavior should be substantiated in these values. As previously discussed, the angles and magnitudes within subspaces are important to analyze - in subsequent rows, therefore, the same experiment vectors are compared in various subspaces of the genome. By examining a range of subspaces, this analysis can be used to partition a transcription profile and define which effectors are responsible for particular patterns of expression. By looking at whole-genome angles as well as angles within subspaces of interest, the CCA reveals a wealth of information about the expression vectors being studied (Figure 3). A CCA can also be used to identify the structure of signaling networks.

Singular value decomposition in expression-profile analysis

We next explore ideas relating to dimensionality reduction of expression vectors. Dimensionality reduction is a common problem in many disciplines of science and engineering. Questions include: what are the essential pieces of information within a transcription profile and what is noise? If the full set of genes (around 6,000 in *S. cerevisiae*) carries redundant information, how small a subset can we choose? Are most transcription profiles merely combinations of a smaller number of transcription profiles? The first tool we discuss is singular value decomposition (SVD).

SVD is a matrix factorization that reveals many important properties of a matrix. It is a standard tool in many areas of the physical sciences, and many algorithms in matrix algebra make use of SVD. Given an $n \times p$ rectangular data matrix $\mathbf{A}_{n\times p}$ of expression profiles (again where n is the number of genes, p is the number of experiments), we can obtain the following factorization:

$$\mathbf{A}_{n\times p} = \mathbf{U}_{n\times n}\,\mathbf{D}_{n\times p}\,\mathbf{V}^{\mathrm{T}}_{p\times p}$$

where $\mathbf{U}$ and $\mathbf{V}$ are orthonormal and $\mathbf{D}$ is diagonal. (Mathematical details of the SVD can be found in Materials and methods.) The intuition behind the SVD is straightforward. Let $\mathbf{u}_i$ be the ith column of $\mathbf{U}$. Note that this column vector is n-dimensional. The best vector (in a sense that can be made mathematically precise, see Materials and methods) that captures (spans) the experiment vectors of $\mathbf{A}$ is $\mathbf{u}_1$. Similarly, the best two column vectors that span the experiment vectors of $\mathbf{A}$ are $\mathbf{u}_1$ and $\mathbf{u}_2$. In the limit, all n vectors $\mathbf{u}_i$ exactly span all the experiment vectors of $\mathbf{A}$. In the same way, the columns of $\mathbf{V}$, in descending order, are p-dimensional vectors that best span the gene vectors of $\mathbf{A}$. These ideas are best illustrated with an example.

$$\mathbf{A} = \begin{bmatrix} 2 & 4 \\ 1 & 3 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}$$

$$\mathbf{U} = \begin{bmatrix} 0.82 & -0.58 & 0 & 0 \\ 0.58 & 0.82 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix},\ \mathbf{D} = \begin{bmatrix} 5.47 & 0 \\ 0 & 0.37 \\ 0 & 0 \\ 0 & 0 \end{bmatrix},\ \mathbf{V} = \begin{bmatrix} 0.40 & -0.91 \\ 0.91 & 0.40 \end{bmatrix}$$

The first column of $\mathbf{U}$, $\mathbf{u}_1$, best spans (that is, when multiplied by a constant) the columns of $\mathbf{A}$. Together with the second column, $\mathbf{A}$ is exactly captured by taking linear combinations of $\mathbf{u}_1$ and $\mathbf{u}_2$. In this case, $\mathbf{u}_3$ and $\mathbf{u}_4$ are not

needed to capture **A**. Similarly, the first column of **V**, $\mathbf{v}_1$, best spans the gene vectors (rows) of **A**. The first value along the diagonal of **D** measures how much contribution $\mathbf{u}_1$ and $\mathbf{v}_1$ make to capturing **A**, and the next value measures how much contribution $\mathbf{u}_2$ and $\mathbf{v}_2$ make. In this case, the large size of the first singular value indicates that $\mathbf{u}_1$ and $\mathbf{v}_1$ capture most of the information in **A**. These values along the diagonal of **D** are called the singular values of **A**. Matrices that contain mostly redundant columns (low-rank matrices) have singular values that rapidly decay, because the matrix is efficiently spanned by a small number of vectors. However, matrices containing very independent columns (high-rank matrices) have singular values that slowly decay. In the limit, the identity matrix, which has the highest possible rank of *n*, has singular values that do not decay at all - they all are one.

For microarray data, SVD can be used to bring out dominant underlying behaviors. For example, examining the cell-cycle data [25] using SVD revealed that among the first **U** vectors ($\mathbf{u}_1$, $\mathbf{u}_2$, ...), the first corresponds to a steady-state and the subsequent ones correspond to the oscillatory behavior that one would expect from such data [15,16]. SVD can also be used to 'de-noise' profiles by recomputing the data matrix using only significant $\mathbf{u}_i$ vectors (significance determined by reading the *i*th singular value along the diagonal matrix **D**.) The SVD can also be used to estimate missing values.

A technique called principal components analysis (PCA) [26] is closely related to the SVD. In PCA, the factorization is applied to the covariance matrix of the data rather than the original data matrix. If the data are mean-adjusted, both SVD and PCA give the same information. If the data are from a time series, then the principal components may correspond to derivatives of the data [27].

Towards a basis of expression vectors

One of the most powerful ideas in linear algebra is the notion of basis vectors. Besides constructing the space, basis vectors can be chosen to highlight important features of the data or simply to store the data efficiently. For example, the function sin *x* can be expressed either in a polynomial or exponential basis,

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \frac{x^7}{7!} + \ldots$$

$$\sin x = \frac{1}{2}\left(e^{ix} - e^{-ix}\right)$$

The first basis is clearly awkward - it is infinite-dimensional and offers little insight. However, the second basis captures the sine wave more compactly and provides more insight. Many of the successes of vector algebra in image compression, smoothing and signal detection come from the identification of appropriate basis functions.

This concept has important implications in the analysis of transcription-profiling data. Considering the data as expression vectors, what is the right basis to express the data in? The data is originally in the basis of genes. But is that the best basis? We have already examined one change of basis - the SVD. In this case, the $\mathbf{u}_i$ and $\mathbf{v}_i$ vectors are basis vectors for the spaces of experiment vectors and gene vectors, respectively. While very efficient basis vectors, the vectors themselves are completely artificial and do not correspond to actual profiles.

An important question in biology is how many different fundamental cellular states or transitions exist. It may be that there are a relatively small number, and that cellular states or transitions are essentially superpositions of a finite number of basic states or transitions. This question is easily posed in expression space because, if the hypothesis holds, the complete set of expression profiles can be examined for the smallest number of basis vectors that can construct the nearly identical entire set. If a profile is added to the data matrix and cannot adequately be constructed by other existing profiles, then it becomes a basis vector. The collection of basis vectors formed by this procedure can be thought of as containing the building blocks of cellular state or transition space. Thus, it would be interesting to try to find basis vectors for all experiment vectors, using actual experiment vectors and not artificial bases that offer little insight.

First, how can one estimate how many basis vectors are required? This problem corresponds to estimating the rank of the data matrix. This problem of finding an approximate rank of a matrix, *k*, is a common one. SVD, as described above, is often employed. Using SVD, one typically looks for a sharp drop in the singular values to estimate the rank. Let us suppose that the singular values fall below a defined threshold at some value. This value serves as a logical choice for *k*. In fact, the first *k* vectors $\mathbf{u}_i$ of **A** serve as the best possible basis to span **A**. In our case, however, we desire the basis vectors to be a subset of the original vectors of **A**. This is a more difficult problem.

We investigated these concepts using two large data sets of transcription profiles generated in *S. cerevisiae*: the set of 300 expression profiles previously analyzed and a second set of over 170 expression profiles of various yeast environmental responses [19,28]. We converted these sets to data matrices of experiment vectors and computed the singular values of each matrix. For each matrix, the singular values decayed rapidly (Figure 4). This decay shows that the columns of the matrix are far from orthogonal and a basis containing a small number of $\mathbf{u}_i$ vectors can efficiently capture the original matrix. Interestingly, the singular value decay of the environmental responses was much faster than the more diverse 300 profile set (compare Figure 4b with a). This is due to the limited number of responses that yeast cells possess when facing stress. This

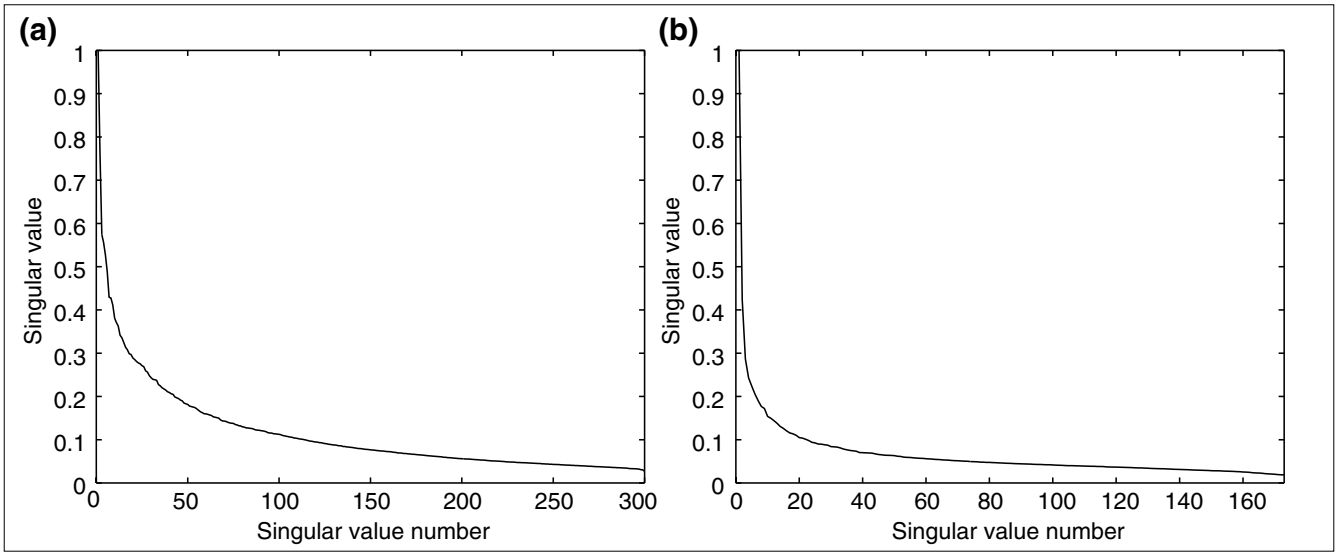

Figure 4
Using the SVD to estimate the number of fundamental expression vectors required to efficiently span a profile set. **(a)** A set of 300 diverse expression profiles [19] was converted to a matrix of expression vectors and the singular values were computed, normalizing such that the first singular value was equal to 1. **(b)** A set of 173 profiles of various yeast environmental responses [28] were converted to a matrix of expression vectors and processed as in (a).

further suggests that there are a small number of building-block transitions that yeast can undergo and that other behaviors are, at least partially, superpositions of these transitions.

But how can one find these building-block, fundamental expression vectors in the profile set? Instead of using the $\mathbf{u}_i$ vectors from SVD (which are linear combinations of the experiment vectors) it would be useful to identify which actual experiment vectors are most useful to form a basis for the matrix. That is, it would be useful to express profiles in terms of other profiles (basis vectors), rather than simply as lists of genes (Figure 5). There are a number of possible algorithms for identifying these basis experiment vectors. We tested a variety of algorithms, especially those

Basis 1			Basis 2	
YAL001C	–0.18	→	Glutamine→proline	+0.60
YAL002W	–0.05		Glucose→ethanol	+0.36
YAL005C	+0.14		Heat shock	+0.13
YAL017W	+0.99		*ure2*Δ steady state	+0.17
YAL029C	–0.33		*tap42-11* steady state	+0.06
...			...	

Figure 5
The transcriptional profile of rapamycin expressed in two alternate bases. A part of the traditional basis for a transcription profile is shown on the left (only five genes out of around 6,000 are shown). The experiment vector of treatment with rapamycin was projected onto the five experiment vectors shown on the right. The coefficients for the rapamycin expression vector in this five-dimensional basis are shown on the extreme right.

utilizing projections of the $\mathbf{u}_i$ vectors on $\mathbf{A}$, to identify optimal candidate basis vectors. (The source code for one such algorithm is publicly available online at [20].) While small subsets of experiment vectors could span the entire set of experiment vectors, it was found that there were many such subsets of experimental vectors that had a comparable spanning ability toward $\mathbf{A}$ (data not shown). This suggests that the cell probably does have a small number of fundamental transitions, although many different groups of transitions (as represented by experiment vectors) appear nearly equivalently able to span the transcriptional range of cellular behavior.

Given this behavior, perhaps the ideal basis vectors for transition space would be the basis of experiment vectors in which only one signaling pathway is selectively modulated. This could be accomplished with small molecules like rapamycin that selectively modulate one signaling pathway with complete specificity. If one had a collection of small molecules that could selectively modulate every signaling pathway in the cell, then it might be possible to simply project an experiment vector on these small molecules' experiment vectors and determine which pathways are activated and to what degree. This would immediately characterize the expression profile in a very powerful basis that provides a great deal of biological insight.

Conclusions

We have described various techniques motivated by vector algebra for analyzing genome-wide expression data. Some of these techniques have been used previously, but we find that

by regarding expression profiles as vectors, either in the sense of genes or experiments, we gain a richer understanding of the existing techniques and can take these insights into new directions.

The angle between two vectors is essentially equivalent to the Pearson correlation coefficient and studying the changes in magnitudes as well as the angles provides important biological information. In particular, we find that decomposing the experiment vector space into a set of smaller subspaces, using functional categories for instance, results in a concise but useful description of the biological system (Figure 3). These types of genome-wide descriptions are essential for understanding large amounts of complex expression data. Thinking in terms of vector spaces naturally introduced the idea of optimal basis vectors, but with the constraint that these basis vectors belong to the original set of expression vectors. This may be a useful way of characterizing expression profiles. We have described the usefulness of the singular value decomposition as a tool for reducing the dimension of the problem and the insight it brings to the problem of finding basis vectors.

A complete analysis of genome-wide expression data involves numerous issues not discussed here. The importance of preprocessing and normalization to correct for various artifacts has been noted recently [29,30]. Artificially large fold changes caused by small values in ratios (red/green or vice versa) should be numerically bounded in competitive hybridization experiments. Assigning a level of significance to observed changes in expression levels is also an important issue and has been addressed by a number of different approaches, most often with a derivation of a statistical model for noise to obtain estimates for the fold changes [31,32]. For oligonucleotide arrays, there are similar and sometimes more complicated issues for the same steps. Finally, there are a variety of clustering and data-mining techniques for understanding the structure of the data in other ways. Appropriate methods should be chosen carefully, depending on the type of data and the questions to be answered. As the quantity of expression data increases rapidly, more tools such as the ones developed here will play an important role in understanding biological states and transitions.

Materials and methods

Software

Computer software was written in the Matlab™ programming language and is publicly available for download from our website [20].

Singular value decomposition (SVD)

SVD is the matrix factorization:

$$\mathbf{A}_{n\times p} = \mathbf{U}_{n\times n}\,\mathbf{D}_{n\times p}\,\mathbf{V}^{\mathrm{T}}_{p\times p}$$

where $\mathbf{D}$ is diagonal and $\mathbf{U}$ and $\mathbf{V}$ are orthonormal ($\mathbf{U}\mathbf{U}^{\mathrm{T}} = \mathbf{U}^{\mathrm{T}}\mathbf{U} = \mathbf{I}_{n\times n}$ and $\mathbf{V}\mathbf{V}^{\mathrm{T}} = \mathbf{V}^{\mathrm{T}}\mathbf{V} = \mathbf{I}_{p\times p}$). The diagonal entries of $\mathbf{D}$ are the singular values $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_p$ and the columns $\{\mathbf{u}_i\}_{i=1,\ldots,n}$ of $\mathbf{U}$ and $\{\mathbf{v}_i\}_{i=1,\ldots,p}$ of $\mathbf{V}$ are called left and right singular vectors, respectively. For microarray data, we assume that $n > p$ as we usually think of the data matrix as having n rows for genes and p rows for experiments, and we assume that the columns are linearly independent.

The geometric visualization of the SVD is the following. When $\mathbf{A}_{n\times p}$ is applied to a p-dimensional sphere, the result is an n-dimensional ellipse in which the principal axes have been stretched by factors $\sigma_1,\ldots,\sigma_p$ in directions $\mathbf{u}_1,\ldots,\mathbf{u}_p$ respectively. (The right singular vectors $\mathbf{v}_i$ are the axes of the sphere that get mapped to $\mathbf{u}_i$ through $\mathbf{A}\mathbf{v}_i = \sigma_i\mathbf{u}_i$.) The usefulness of the SVD is often based on the fact that it answers the following important question: what is the best approximation to $\mathbf{A}$ using a matrix of lower rank? (The term 'best' can be defined mathematically.) This is important because a comparable matrix with a smaller rank may make the underlying structure of the matrix more apparent. Another way of writing $\mathbf{A}=\mathbf{U}\mathbf{D}\mathbf{V}^{\mathrm{T}}$ is:

$$\mathbf{A} = \sum_{i=1}^{p} \sigma_i \mathbf{u}_i \mathbf{v}_i^{\mathrm{T}}$$

Then, it can be shown mathematically that the best approximation of dimension k is obtained by the partial sum of the first k terms [33]. Sometimes the terms $\sigma_i\mathbf{u}_i\mathbf{v}_i^{\mathrm{T}}$ are called the 'characteristic modes'. This representation is similar to that of a signal by a sum of its Fourier modes. The SVD can be used to de-noise the data by recomputing the sum with the small singular values set to zero.

Measuring basis quality

Let $\mathbf{A}$ be an $n \times p$ matrix where the columns of $\mathbf{A}$ represent a collection of p transcription profiles over n genes. k expression vectors are chosen from $\mathbf{A}$ as a basis for the complete set. Let $\mathbf{B}$ be the matrix whose columns consist of the k expression vectors ($\mathbf{B}$ is therefore $n \times k$). The ability of $\mathbf{B}$ to form a basis for the remaining $p - k$ vectors of $\mathbf{A}$ (designated $\mathbf{x}_i$) is tested by application of the projection theorem. Let $\mathbf{P}$ be the projection matrix for $\mathbf{B}$, that is $\mathbf{P} = \mathbf{B}(\mathbf{B}^{\mathrm{T}}\mathbf{B})^{-1}\mathbf{B}^{\mathrm{T}}$. A measure of the spanning ability of the basis $\mathbf{B}$ toward the $\mathbf{x}_i$ is:

$$\gamma = \sum_{i=1}^{p-k} \frac{|(\mathbf{I} - \mathbf{P})\mathbf{x}_i|}{|\mathbf{x}_i|}$$

Until γ is below a determined threshold for the original matrix $\mathbf{A}$, a different collection of k expression vectors can be tested or k can be increased to include more basis vectors.

In the expression for γ, the numerator is the 'residual vector', the part of the $\mathbf{x}_i$ that could not be captured by $\mathbf{B}$. The normalized magnitude of the residual vector measures how well the $\mathbf{x}_i$ can be constructed from the k columns of $\mathbf{B}$.

References

1. DeRisi JL, Iyer VR, Brown PO: **Exploring the metabolic and genetic control of gene expression on a genomic scale.** *Science* 1997, **278:**680-686.
2. Cho RJ, Campbell MJ, Winzeler EA, Steinmetz L, Conway A, Wodicka L, Wolfsberg TG, Gabrielian AE, Landsman D, Lockhart DJ, Davis RW: **A genome-wide transcriptional analysis of the mitotic cell cycle.** *Mol Cell* 1998, **2:**65-73.
3. Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I: **The transcriptional program of sporulation in budding yeast.** *Science* 1998, **282:**699-705.
4. Hardwick JS, Kuruvilla FG, Tong JK, Shamji AF, Schreiber SL: **Rapamycin-modulated transcription defines the subset of nutrient-sensitive signaling pathways directly controlled by the Tor proteins.** *Proc Natl Acad Sci USA* 1999, **96:**14866-14870.
5. Golub TR, Slonim DK, Tamayo P, Huard C, Gaasenbeek M, Mesirov JP, Coller H, Loh ML, Downing JR, Caligiuri MA, *et al.*: **Molecular classification of cancer: class discovery and class prediction by gene expression monitoring.** *Science* 1999, **286:**531-537.
6. Ross DT, Scherf U, Eisen MB, Perou CM, Rees C, Spellman P, Iyer V, Jeffrey SS, Van de Rijn M, Waltham M, *et al.*: **Systematic variation in gene expression patterns in human cancer cell lines.** *Nat Genet* 2000, **24:**227-235.
7. Perou CM, Sorlie T, Eisen MB, van de Rijn M, Jeffrey SS, Rees CA, Pollack JR, Ross DT, Johnsen H, Akslen LA, *et al.*: **Molecular portraits of human breast tumours.** *Nature* 2000, **406:**747-752.
8. Brazma A, Jonassen I, Vilo J, Ukkonen E: **Predicting gene regulatory elements *in silico* on a genomic scale.** *Genome Res* 1998, **8:**1202-1215.
9. Holstege FC, Jennings EG, Wyrick JJ, Lee TI, Hengartner CJ, Green MR, Golub TR, Lander ES, Young RA: **Dissecting the regulatory circuitry of a eukaryotic genome.** *Cell* 1998, **95:**717-728.
10. Roberts CJ, Nelson B, Marton MJ, Stoughton R, Meyer MR, Bennett HA, He YD, Dai H, Walker WL, Hughes TR, *et al.*: **Signaling and circuitry of multiple MAPK pathways revealed by a matrix of global gene expression profiles.** *Science* 2000, **287:**873-880.
11. Shamji AF, Kuruvilla FG, Schreiber SL: **Partitioning the transcriptional program induced by rapamycin among the effectors of the Tor proteins.** *Curr Biol* 2000, **10:**1574-1581.
12. Clark EA, Golub TR, Lander ES, Hynes RO: **Genomic analysis of metastasis reveals an essential role for RhoC.** *Nature* 2000, **406:**532-535.
13. Brazma A, Vilo J: **Gene expression data analysis.** *FEBS Lett* 2000, **480:**17-24.
14. Brazma A, Robinson A, Cameron G, Ashburner M: **One-stop shop for microarray data.** *Nature* 2000, **403:**699-700.
15. Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV: **Fundamental patterns underlying gene expression profiles: simplicity from complexity.** *Proc Natl Acad Sci USA* 2000, **97:**8409-8414.
16. Alter O, Brown PO, Botstein D: **Singular value decomposition for genome-wide expression data processing and modeling.** *Proc Natl Acad Sci USA* 2000, **97:**10101-10106.
17. Wickens TD: *The Geometry of Multivariate Statistics*. Hillsdale, NJ: Lawrence Erlbaum Associates; 1995.
18. Troyanskaya O, Cantor M, Sherlock G, Brown P, Hastie T, Tibshirani R, Botstein D, Altman RB: **Missing value estimation methods for DNA microarrays.** *Bioinformatics* 2001, **17:**520-525.
19. Hughes TR, Marton MJ, Jones AR, Roberts CJ, Stoughton R, Armour CD, Bennett HA, Coffey E, Dai H, He YD, *et al.*: **Functional discovery via a compendium of expression profiles.** *Cell* 2000, **102:**109-126.
20. **Stuart Schreiber laboratory** [http://www.schreiber.chem.harvard.edu]
21. Cardenas ME, Cutler NS, Lorenz MC, Di Como CJ, Heitman J: **The TOR signaling cascade regulates gene expression in response to nutrients.** *Genes Dev* 1999, **13:**3271-3279.
22. Bertram PG, Choi JH, Carvalho J, Ai W, Zeng C, Chan TF, Zheng XF: **Tripartite regulation of Gln3p by TOR, Ure2p and phosphatases.** *J Biol Chem* 2000, **275:**35727-35733.
23. Mewes HW, Heumann K, Kaps A, Mayer K, Pfeiffer F, Stocker S, Frishman D: **MIPS: a database for genomes and protein sequences.** *Nucleic Acids Res* 1999, **27:**44-48. [http://www.mips.biochem.mpg.de/]
24. Courcelle J, Khodursky A, Peter B, Brown PO, Hanawalt PC: **Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient *Escherichia coli*.** *Genetics* 2001, **158:**41-64.
25. Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B: **Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization.** *Mol Biol Cell* 1998, **9:**3273-3297.
26. Johnson RA, Wichern DW: *Applied Multivariate Statistical Analysis*. Upper Saddle River, NJ: Prentice Hall; 1998.
27. Raychaudhuri S, Stuart JM, Altman RB: **Principal components analysis to summarize microarray experiments: application to sporulation time series.** *Pac Symp Biocomput* 2000**,** 455-466.
28. Gasch AP, Spellman PT, Kao CM, Carmel-Harel O, Eisen MB, Storz G, Botstein D, Brown PO: **Genomic expression programs in the response of yeast cells to environmental changes.** *Mol Biol Cell* 2000, **11:**4241-4257.
29. Tseng GC, Oh MK, Rohlin L, Liao JC, Wong WH: **Issues in cDNA microarray analysis: quality filtering, channel normalization, models of variations and assessment of gene effects.** *Nucleic Acids Res* 2001, **29:**2549-2557.
30. Yang YH, Dudoit S, Luu P, Speed TP: **Normalization for cDNA microarray data.** In *Microarrays: Optical Technologies and Informatics*. Edited by Bittner ML, Chen Y, Dorsel AN, Dougherty ER. Bellingham: International Society for Optical Engineering, 2001, 141-152.
31. Chen Y, Dougherty ER, Bittner ML: **Ratio-based decisions and the quantitative analysis of cDNA microarray images.** *J Biomed Optics* 1997, **2:**364-374.
32. Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW: **On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data.** *J Comput Biol* 2001, **8:**37-52.
33. Golub GH, van Loan CF: *Matrix Computation*. Baltimore, MD: Johns Hopkins University Press; 1996.
34. **Rosetta Inpharmatics** [http://www.rii.com]