Uses of the Singular Value Decompositions in Biology

The Singular Value Decomposition is a very useful tool deeply rooted in linear algebra. Despite this, SVDs have found their way into many different areas of biology. Biological datasets, often large and noisy, lend themselves to the analysis SVDs offer. While the biological interpretation can be difficult, and the linear algebra can be overwhelming, the utility of the decomposition is making it very popular among biologists. This paper will review the SVD, and many of the applications biologists are using it for.

Explanation of SVD

In linear algebra terms, an SVD is the decomposition of a matrix ***A*** into three matrices, each having special properties. Written out explicitly:

$$\boldsymbol{A} = \boldsymbol{USV}^{\mathrm{T}} \qquad (1)$$

where $\boldsymbol{V}^{\mathrm{T}}$ is the transpose of the matrix ***V***. The transpose is defined as the operation on a matrix that switches the order of the subscripts. That is, element V_{ij} becomes element V_{ji}.

Matrix ***A*** is an *m x n* matrix with no specific properties. *m* is allowed to smaller or larger than *n*. There are two types of SVDs, a short form and a long form. In the long form, ***U*** is an *m x m* matrix and ***S*** is an *m x n* matrix while in the short form, ***U*** is an *m x n* matrix, and ***S*** is an *n x n* matrix. For both forms, ***V*** is an *n x n* matrix. The extra information contained in the long form often has little biological relevance, and is thus provides no extra utility in the perspective of the biologist. For the remainder of this paper I will concentrate on the short form version of the SVD.

The elements of ***S*** are only non-zero on the diagonal. That is ***S*** is a diagonal matrix with at most *n* distinct non-zero elements. The non-zero elements on the diagonal are called singular values, where by convention they are sorted so that the largest singular value is in the first row, the second largest value is in the second row, and so on. That is $S_{ii} \geq S_{(i+1)(i+1)}$ for all i. Furthermore, the singular values are always greater than zero, and the number of non-zero singular values is equal to the rank of the matrix A.

The condition on ***U*** is that the columns of ***U*** form an orthonormal basis.

$$\mathbf{u_i} \cdot \mathbf{u_i} = 1 \qquad (2)$$

for all i where $\mathbf{u_i}$ is the ith column of matrix ***U***.

$$\mathbf{u_i} \cdot \mathbf{u_j} = 0 \qquad (3)$$

for all i ≠ j.

The condition on ***V*** is that the rows of $\boldsymbol{V}^{\mathrm{T}}$ form an orthonormal basis. These conditions define the SVD.

The calculation of an SVD is not a trivial matter. An entire advanced computer science class can be taken on the subject. A good reference for the calculation of the SVD is (Golub *et al,* 1996). Many software packages contain blackbox SVD functions that can be used for those who don't want to understand the exciting intricacies of calculating the SVD. Matlab, for example, contains a function svd(***X***) that returns the singular decomposition of ***X***, with an option for the long form or short form of the SVD. The short form is called the economy size decomposition in matlab. An example of an SVD is given below:

$$\begin{vmatrix} 1 & 2 \\ 3 & 4 \\ 5 & 6 \\ 7 & 8 \end{vmatrix} = \begin{vmatrix} -.15 & -.82 \\ -.35 & -.42 \\ -.55 & -.02 \\ -.74 & .38 \end{vmatrix} \begin{vmatrix} 14.3 & 0 \\ 0 & .63 \end{vmatrix} \begin{vmatrix} -.64 & -.77 \\ .77 & -.64 \end{vmatrix} \qquad (4)$$

Here I show the matrix on the left decomposed into the multiplication of three matrices. A quick check shows the columns of the first matrix are indeed orthonormal and the rows of the third matrix are also orthonormal. There are two singular values, 14.3 and .63.

A valid interpretation of the SVD is to view the it as a decomposition of the four vectors (1,2), (3,4), (5,6), and (7,8) onto an orthonormal basis (-.64, -.77), and (.77,-.64). The first matrix would be a weight matrix that describes how much of each orthonormal vector the original data vector contains. This orthonormal basis is the "natural basis" of the matrix and are often called the eigenrows. There are an infinite number of orthonormal basis that this matrix can be decomposed into. Some common basis would be the sine and cosine waves of a Fourier transform, or the exponential functions of a Laplace transform. However the SVD is the only decomposition that causes the columns of the weight matrix to be orthonormal.

The singular values are useful to determine which eigenrows are important to describing the matrix. The square of the singular values is proportional to the percent of the matrix variation that is explained by the eigenrow corresponding to it. This feature is very useful for image compression. An SVD can be performed on an image. Singular values smaller than a certain threshold can be thrown out, as most of the information is stored in the first few eigenrows that correspond to the largest singular values. This is also very important for any data that contains noise. If the signal to noise ratio is decently sized, then singular values smaller than a certain threshold can be thrown out. A common practice is to graph the singular values and look for an 'elbow' in the plot. The eigenrows corresponding to the noisy values have small singular values.

Applications of SVD:

Titration curves of chromophores

A useful example on the utility of the SVD is found in (Hendler and Shrager, 1994). They show a step by step example of how an SVD is applied to deconvolute a titration involving a mixture of three pH indicators. We can measure the absorbance spectra at many wavelengths at many different pHs. We can then arrange the data into an *m x n* matrix ***A*** where the columns are absorbance measurements at a given wavelength and the columns correspond to different pHs. There are then m pHs that the absorbance spectra was measured, and n different wavelengths that the spectra was measured across. There are three chromophores with two states whose relevant concentration is given by the Henderson-Hassalbach equation:

$$\mathrm{pH} = \mathrm{pK} + \log([\mathrm{salt}]/[\mathrm{acid}]) \qquad (5)$$

The spectrum of each chromophore is dependent upon the state, protonated or deprotonated. This problem can be reduced to a matrix problem:

$$\boldsymbol{D}\boldsymbol{F}^{\mathrm{T}} = \boldsymbol{A} \qquad (6)$$

where ***A*** is a matrix formed from our measurements as described above. ***D*** is a matrix formed from the difference spectrum from each chromophore in the mixture and from a spectrum of the mixture of the three deprotonoated chromophores. The matrix ***F*** contains information from the Henderson-Hasselbach equations for each of the transitions as well as ones for a constant reference spectrum. We would like to use ***A*** to determine both ***D*** and ***F***.

To accomplish this we first perform the SVD on A to give:

$$\boldsymbol{A} = \boldsymbol{U}\boldsymbol{S}\boldsymbol{V}^{\mathrm{T}} \qquad (7)$$

At this point it is important to analyze the results of the SVD to determine the important eigenrows and discard the superfluous noisy vectors. The procedure recommended by Hendler and Shrager is as follows:

A: The singular values start relatively high and then decrease to a low noisy plateau which represents the noise level. A visual inspection will indicate the rank of the matrix.

B: Perform an autocorrelation analysis, or keep the vectors of ***U*** and ***V*** that look like absorbance spectra or titration curves.

Keep the singular values and their respective eigenrows and eigencolumns that are significant. At this point, curve fitting can be performed on the columns of ***V*** to the Henderson-Hasselbach equation. This will yield 3 values of pK's and 4 baseline amounts, 1 for each chromophore and 1 for a total baseline. At this point, having an estimate for the ***F*** matrix allows us to calculate ***D*** by taking the pseudoinverse of $\boldsymbol{F}^{\mathrm{T}}$ and mupltiplying it by ***A.***

Small Angular Scattering

SVD has been used to detect structural intermediates. (Chen *et al*, 1996) used an SVD to detect a third structural state of lysozyme in biomolectular small-angle scattering in addition to the correctly folded and completely denatured states. The scattering curve of hen egg lysozyme was measured at multiple urea concentrations, a denaturing agent. The scattering data is intensity values measured at many different angles. The significant basis vectors were found by a combination of significance of the singular values, the autocorrelations of $\boldsymbol{U}$, the shape of the basis functions and the minimum fitting reduced chi-square. The SVD analysis combined with a denaturant binding model that was used to estimate thermodynamic parameters allowed the reconstruction of a scattering profile for the pure intermediate state. This intermediate was not apparent from the data without application of the SVD.

Protein dynamics

SVD analysis has been used to characterize protein dynamics. (Romo et al, 1995) used an SVD to analyze the movement of myoglobin. Using molecular dynamics methods, Romo measured the atomic positions of all atoms sampled during the simulation. According to Romo, the SVD provided a method for decomposing a molecular dynamics trajectory into fundamental modes of atomic motion. The eigenrows are projections of the protein conformations onto these modes showing the protein motion in a generalized low-dimensional basis. Analysis of the singular values gives information on the extent of configuration space sampled by myoglobin. The results showed that much of configuration space was not actually explored by the protein. The visualization of the trajectory of the protein also showed a very interesting trajectory that was described as resembling beads on a string.

(Ozkan *et al,* 2002) also used SVDs to study if two-state proteins fold by pathways or funnels. In their study, they solved the exact dynamics of a simple model (16mer amino acid chain) using a master equation formalism. Hidden intermediates had been discovered, and their existence was argued as not being consistent with the funnel model of protein folding. Ozkan, however, showed that a funnel model could produce hidden intermediates. A huge multiplicity of trajectories were observed at the microscopic level, but these collapsed to hidden intermediates at the macroscopic level running in parallel. The energy landscapes produced by these simulations of protein folding are complex and have a high dimensionality. An SVD was used to process this data down to a level that could be analyzed visually. A 32 x M matrix was created out of M conformations of the 16mer amino acid chain. The eigenrows of the two largest singular values were used as axis for a 3-dimensional plot with the z-axis being the energy. Thus, a two dimensional representation of the energy landscape was developed using the SVD. This representation (Fig 1) shows that the energy landscape is actually a funnel. As more and more contacts between amino acids are formed, the energy landscape becomes more and more like a funnel.

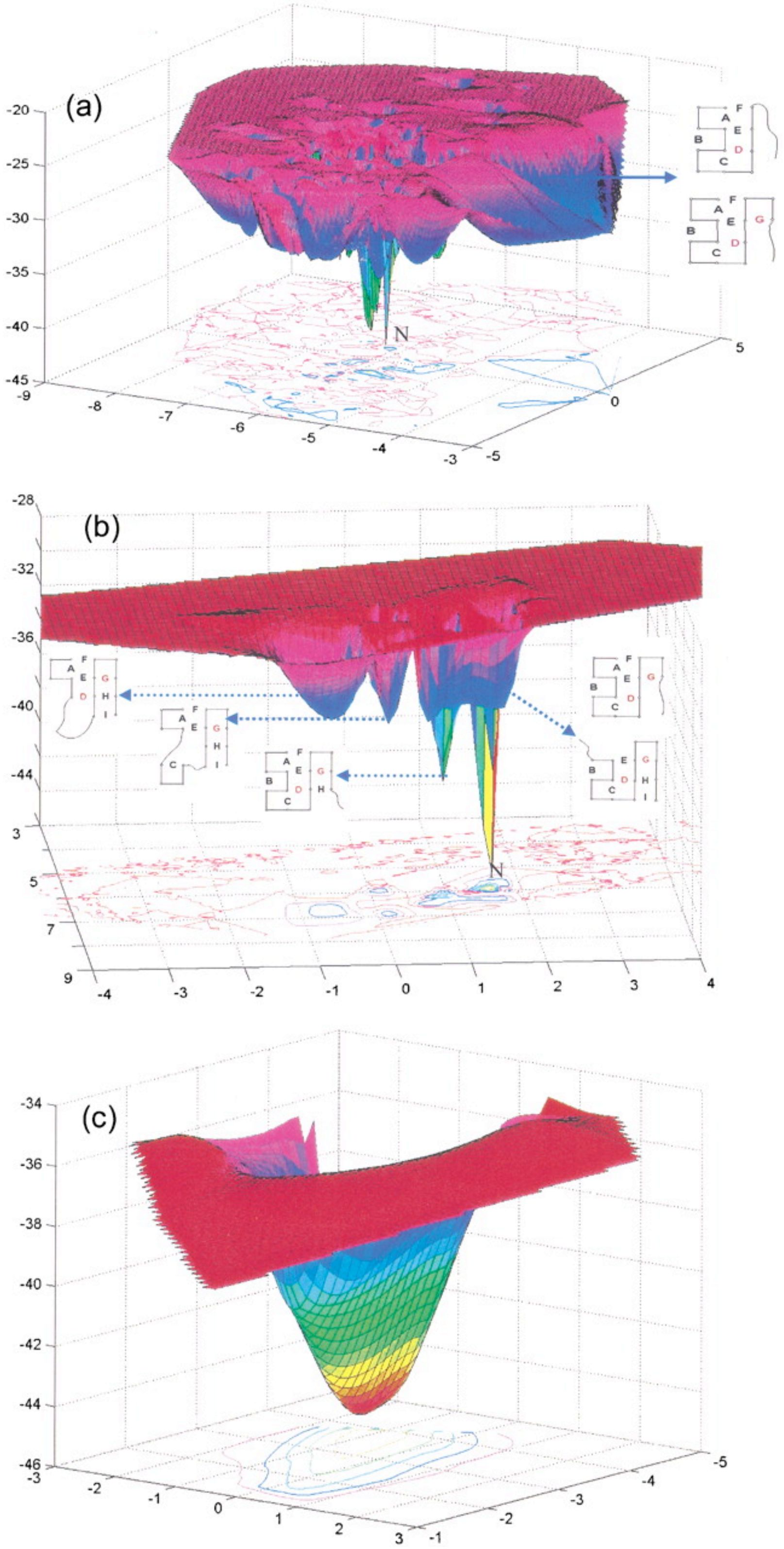

Fig 1: Shows the landscape for configurations with (a) 4 or more contacts (b) 5 or more contacts (c) 6 or more contacts

Analysis of microarray data

Gene expression data are particularly well suited to SVD analysis. Microarrays are often performed on thousands of genes and tens to hundreds experiments are often performed. (Wall *et al*, 2003) breaks down studies into two broad classes, systems biology applications, and diagnostic applications. In both cases, microarray data is organized into a data matrix ***X*** with the *n* columns corresponding to assays and the *m* rows corresponding to genes. The data from the microarrays often will be preprocessed before being placed in the matrix. For example, logarithms are often taken on the ratios so that a 2-fold change in both directions produce the same numeric difference as no change. The data can also be scaled and normalized to a zero mean and unit standard deviation. An SVD is then often performed on ***X.*** For gene expression data, the orthonormal columns of ***U*** are often called eigenassays, while the orthonormal rows of ***V*** are called eigengenes. For the remainder of this section I will refer to these orthonormal basis by these names.

An SVD was applied to budding yeast cell-cycle data set generated by (Cho *et al*, 1998). The SVD was performed by (Wall *et al*, 2003*)*. In the Cho experiment, 6200 yeast genes were monitored for 17 time points taken at ten-minute intervals. The data was pre-processed by taking the logarithm of both sides and normalizing each genes response to have a zero mean and unit standard deviation. An auto-correlation test was performed (Kanji, 1993) to filter out ~ 3200 of the genes that showed random fluctuations. Plots of the three most significant eigengenes revealed interesting features. The first eigengenes was an exponential decay. The second and third eigengenes showed very obvious cell cycle fluctuation and appeared very close to a sine and cosine wave. These three interesting patterns account for over 60% of the variation in the data. Assigning biological relevance to these three eigengenes could possibly provide important clues to cell cycle progression.

(Alter et al, 2003) performed an SVD on human and yeast cell-cycle data. This study actually used a generalized version of the SVD where data from both human and yeast was used to create eigengenes and eigenarrays. Use of each eigengene was compared and contrasted between the two species. This allowed the identification of processes specific and in common to both species. In that study, the first two eigengenes showed cyclic patterns. Two dimensional scatter plots, with the x and y axis being eigengenes one and two, showed that the cell-cycle genes tend to plot towards the perimeter of the disc. Figure 2 below is pulled from that paper. They also found that 641 of the 784 eigengenes identified in (Spellman et al., 1998) associate with the first 2 eigengenes. The distribution of these cell-cycle disc is approximately uniform, suggesting that gene regulation may be a very continuous process. An eigengene with exponential decay was also found, probably due to some sort of stress response due to the synchrony of the cells.

From Alter et al 2000

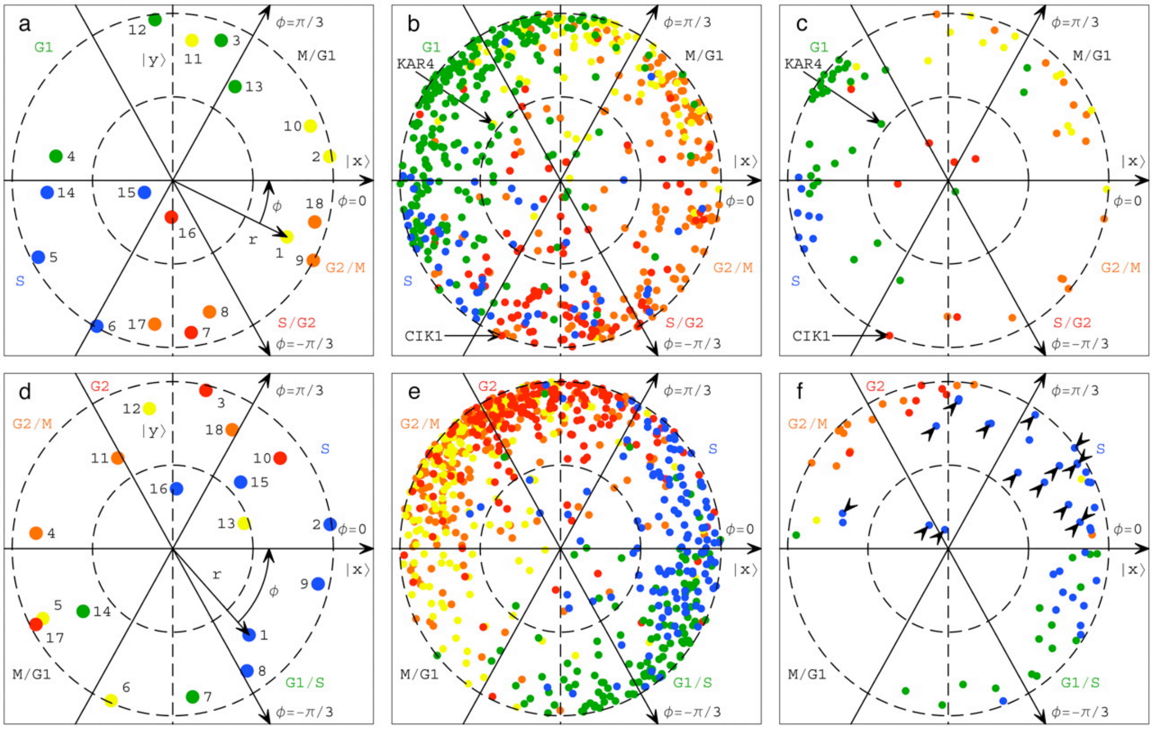

Fig. 2. Yeast (*a-c*) and human (*d-f*) expression reconstructed in the six-dimensional cell-cycle subspaces approximated by two-dimensional subspaces. (*a*) Yeast array expression, projected onto $\pi/2$-phase along the *y* axis vs. that onto 0-phase along the *x* axis and color-coded according to the classification of the arrays into the five cell-cycle stages: M/G_1 (yellow), G_1 (green), S (blue), S/G_2 (red), and G_2/M (orange). The dashed unit and half-unit circles outline 100% and 50% of added-up (rather than canceled-out) contributions of the six arraylets to the overall projected expression. The arrows describe the projections of the $-\pi/3$-, 0-, and $\pi/3$-phase arraylets. (*b*) Yeast expression of 603 cell cycle-regulated genes projected onto $\pi/2$-phase along the *y* axis vs. that onto 0-phase along the *x* axis and color-coded according to the classification by Spellman *et al.* (11) (*c*) Yeast expression of 76 cell cycle-regulated genes color-coded according to the traditional classification. (*d*) Human array expression color-coded according to the classification of the arrays into the five cell-cycle stages: S (blue), G_2 (red), G_2/M (orange), M/G_1 (yellow), and G_1/S (green). (*e*) Human expression of 750 cell cycle-regulated genes color-coded according to the classification by Whitfield *et al.* (12) (*f*) Human expression of 73 cell cycle-regulated genes color-coded according to the traditional classification; the arrows point to 16 human histones that were not classified by Whitfield *et al.* as cell cycle-regulated based on their overall expression.

Other experiments

Some recent interesting uses of the SVD were performed by (Price *et al*, 2003) and (Kluger *et al*, 2003). Price used the SVD on an extreme pathways matrix in order to analyze the metabolic capabilities of two bacterial species. Analysis of the solution space and dominance of the first singular value revealed that Helicobacter pylori has a more rigid metabolic network than Haemophilus influenzae for the production of amino acids. The SVD was also able to identify key network branch points that may identify key control points for regulation.

Kluger developed a method to simultaneously cluster genes and conditions to find what they called "checkerboard" patterns in matrices of gene expression data. The checkerboard structures are found in eigenvectors corresponding to characteristic expression patterns across genes or conditions. These eigenvectors are identified by performing SVDs coupled with integrated normalization steps.

Reverse engineering

One of the most interesting papers to come out is (Yeung *et al*, 2002) involving the singular value decomposition to reverse engineer gene networks. The system was assumed to be operating near a steady state and the dynamics were approximated as:

$$\dot{x}(t) = \square\square x_i(t) + \sum_{j=1}^{N} W_{ij}x_j(t) + b_i(t) + \square(t) \quad \textit{for } i = 1,2,\ldots,N \qquad (8)$$

where x is the concentrations of the mRNAs that reflect the expression levels of the genes, □s represent the self-degradation rates, the Ws describe the type and size of the interactions between the genes, the bs are external stimuli, and the □s represent noise. An experiment applies a prescribed stimulus and uses a microarray to measure simultaneously the concentrations of all the N different mRNAs. Repeating this experiment M times allows us to estimate x and the time derivative of x. Equation 8 can be reorganized in matrix form as:

$$\dot{X} = AX + B \qquad (9)$$

where is a combination of the Ws as well as the □ self-interaction terms. The solution for ***A*** let us get to the Ws, exactly what we want.

As arrays are expensive, there are often less arrays performed than genes, which leaves us with an underdetermined problem. Often the SVD is taken of ***X***, allowing for a solution to equation number 9. Since the problem is underdetermined, many different solutions are possible of the form:

$$\boldsymbol{A} = \boldsymbol{A_0} + \boldsymbol{C}\boldsymbol{V}^{\mathrm{T}} \qquad (10)$$

Most often, the solution chosen is the one that minimizes the least squared errors. However, Yeung et al had an important insight. Namely gene networks are often sparse. That is most genes do not interact with each other, as only a few are able to regulate transcription. So instead of searching for the solution with the smallest squared error, they searched for the solution with the smallest number of connections, or the sparsest matrix. The also simulated results and tested the results from the two types. The errors are listed in Figure 3 below.

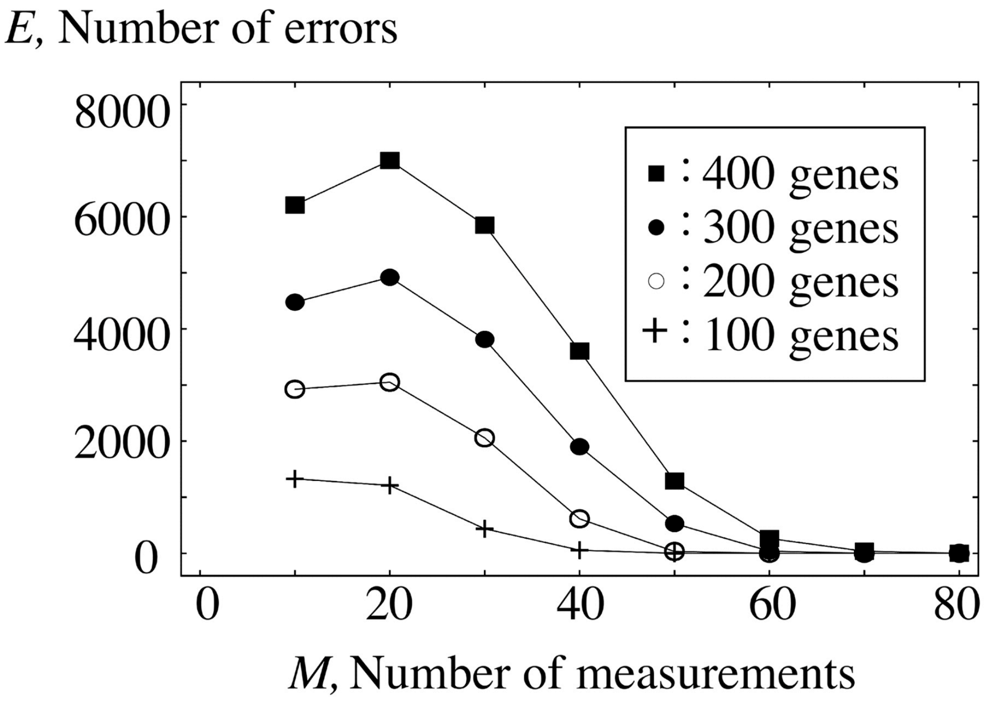

Fig. 3. Number of errors, *E*, made by the reverse engineering scheme as a function of *M*, the number of measurements, for four linear networks of the form **1** with different sizes *N*.

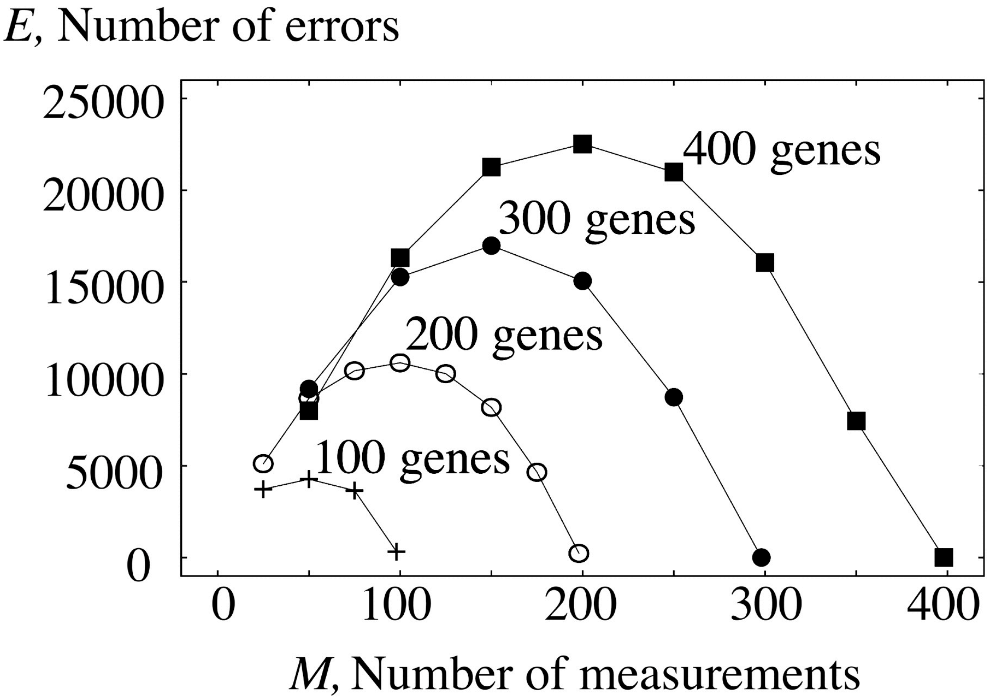

Fig. 4. Number of errors, *E*, made by the reverse engineering scheme as a function of *M*, the number of measurements, for four linear networks of the form **1** with different sizes

Conclusion

The SVD has become a very prevalent part of biology, mainly because of its utility. A quick search shows that 81 papers in pub-med since 2002 have the words singular value decomposition somewhere in the paper. As biologists become more comfortable with this technique, its use will continue to increase for the near future.

Bibliography

Alter O., Brown P.O., Botstein D. Singular value decomposition for genome-wide expression data processing and modeling. Proc Natl Acad Sci USA 2000; 97:10101-06.

Chen L., Hodgson K.O., Doniach S. A lysozyme folding intermediate revealed by solution X-ray scattering. J Mol Biol 1996; 261:658-71

Cho R.J., Campbell M.J., Winzeler E.A., Steinmetz L., Conway A., Wodicka L., Wolfsberg T.G., Gabrielian A.E., Landsman D., Lockhart D.J., Davis R.W. A genome-wide transcriptional analysis of the mitotic cell cycle. Mol Cell 1998; 2:65-73.

Golub G., Van Loan C., *Matrix Computations*. Baltimore: Johns Hopkins Univ Press, 1996.

Hendler R.W., Shrager R.I. Deconvolutions based on singular value decomposition and the pseudoinverse: a guide for beginners. J Biochem and Biophys Meth 1994; 28: 1-33.

Kanji G.K., *100 Statistical Tests*. New Delhi: Sage, 1993.

Kluger Y., Basri R., Chang J.T., Gerstein M. Spectral biclustering of microarray data: coclustering genes and conditions. Genome.org 2003; 13: 703-716.

Ozkan S.B., Dill, K.A., Bahar I. Fast-folding protein kinetics, hidden intermediates, and the sequential stabilization model. Protein Science 2002; 11:1958-1970

Price N.D., Reed J.L., Papin J.A., Famili I., Palsson B.O. Analysis of metabolic capabilities using singular value decomposition of extreme pathway matrices. Biophysical Journal 2003; 84: 794-804.

Romo T.D., Clarage J.B., Sorensen D.C., Phillips G.N., Jr. Automatic identification of discrete substates in proteins: singular value decomposition analysis of time-averaged crystallographic refinements. Proteins 1995; 22:311-21.

Spellman P.T., Sherlock G., Zhang M.Q., Iyer V.R., Anders K., Eisen M.B., Brown P.O., Botstein D., Futcher B. Comprehensive identification of cell cycle-regulated genes of the yeast Saccharomyces cerevisiae by microarray hybridization. Mol Biol Cell 1998; 9:3273-97.

Wall M.E., Rechtsteiner A., Rocha L.M. Singular value decomposition and principal component analysis. Kluwer: Norwell, MA, 2003: 91-109.

Yeung M.K., Tegner J., Collins J.J. Reverse engineering gene networks using singular value decomposition and robust regression. Proc Natl Acad Sci USA 2002; 99:6163-68.