# *SVDMAN -- Singular value decomposition analysis of microarray data*

*Michael E. Wall, Patricia A. Dyck, Thomas S. Brettin**

*Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM 87545*

Running head: SVD microarray analysis using SVDMAN

Key words: global gene expression analysis; non-exclusive gene groups; self-consistency measure of confidence; threshold group membership, singular value decomposition

Accepted in *Bioinformatics* as an Application Note

* To whom correspondence should be addressed

## Abstract

***Summary**: We have developed two novel methods for singular value decomposition analysis (SVD) of microarray data. The first is a threshold-based method for obtaining gene groups, and the second is a method for obtaining a measure of confidence in SVD analysis. Gene groups are obtained by identifying elements of the left singular vectors, or gene coefficient vectors, that are greater in magnitude than the threshold $WN^{-1/2}$, where N is the number of genes, and W is a weight factor whose default value is three. The groups are non-exclusive and may contain genes of opposite (i.e. inversely correlated) regulatory response. The confidence measure is obtained by systematically deleting assays from the data set, interpolating the SVD of the reduced data set to reconstruct the missing assay, and calculating the Pearson correlation between the reconstructed assay and the original data. This confidence measure is applicable when each experimental assay corresponds to a value of parameter that can be interpolated, such as time, dose or concentration. Algorithms for the grouping method and the confidence measure are available in a software application called SVDMAN (SVD Microarray ANalysis). In addition to calculating the SVD for generic analysis, SVDMAN provides a new means for using microarray data to develop hypotheses for gene associations and provides a measure of confidence in the hypotheses, thus extending current SVD research in the area of global gene expression analysis.*



***Availability**: ftp://bpublic.lanl.gov/compbio/software*



***Contact**: brettin@lanl.gov*


***Supplemental Information**: http://public.lanl.gov/mewall/svdman*

February 14, 2001

**Introduction**

Principal component analysis (PCA), often performed by singular value decomposition (SVD), is a popular analysis method that has recently been explored as a method for analyzing large-scale expression data (Raychaudhuri *et al.*, 2000; Holter *et al.*, 2000; Alter *et al.*, 2000). Additionally SVD/PCA has been used to identify high-amplitude modes of fluctuations in macromolecular dynamics simulations (Garcia, 1992; Romo *et al.*, 1995), and identify structural intermediates in lysozyme folding using small-angle scattering experiments (Chen *et al*., 1996). The first published microarray results are those of Raychaudhuri *et al.* (2000), who used PCA to analyze time series yeast sporulation expression data (Chu *et al.*, 1998). Their study found that much of the sporulation data was explained by two principal components, and that previously defined gene clusters could be visualized using the PCA coefficients. Subsequent reports supported these results (Alter *et al.*, 2000; Holter *et al.*, 2000). Alter *et al.* (2000) analyzed yeast cell-cycle expression data (Spellman *et al.*, 1998), identified sinusoidal modes in the SVD which correspond to cell-cycle modes, and found that 641 out of 784 previously identified cell-cycle genes had at least 25% of their normalized expression signal due to cell-cycle modes. In similar work, Holter *et al.* (2000) analyzed cell-cycle data (Spellman *et al.*, 1998), sporulation data (Chu *et al.*, 1998), and data from serum-treated human fibroblasts (Iyer *et al.*, 1999), demonstrating that groups obtained by cluster analysis tend to cluster in the space of appropriately chosen SVD matrix elements.

Here we describe two novel methods for SVD analysis of microarray data. One is a threshold method for obtaining gene groups. Another is a method for measuring confidence in SVD analysis. We first give a brief overview of the anatomy of the SVD. Next we describe a computer program called SVDMAN (SVD Microarray ANalysis) that implements the methods. We have validated the performance of SVDMAN using publicly available microarray data and biology databases (Dyck *et al.*, 2000). Finally we contrast SVDMAN with two other analysis methods: clustering (see, e.g., Eisen *et al.*, 1998) and gene shaving, (Hastie *et al.*, 2000).

**Methods and Implementation**

*Singular Value Decomposition (SVD)*

We define the matrix of gene expression data as A, where A is a *N* x *M* matrix in which the *N* rows index the genes, and the *M* columns index the assays. The SVD theorem states (see Press *et al.*, 1992):

$$A = U\Sigma V^T$$

where U is a *N* x *M* matrix whose columns are the left singular vectors (*gene coefficient vectors*); $\Sigma$ is a *M* x *M* diagonal matrix of singular values (*mode amplitudes*); and $V^T$ is a *M* x *M* matrix whose rows are the right singular vectors (*expression level vectors*). The gene coefficient vectors form an orthonormal set, the expression level vectors form an orthonormal set, and $\Sigma$ is diagonal. The SVD represents an expansion of the original data in a coordinate system where the covariance matrix is diagonal. PCA identifies principal components by diagonalization of a matrix of variation such as the covariance matrix. SVD is thus a method for performing PCA by diagonalization of the covariance matrix.

*SVDMAN (SVD Microarray ANalysis)*

SVDMAN calculates the SVD of the matrix A, generates gene groups, and calculates a confidence measure. The input matrix A contains tab-delimited ASCII data, with an extra column of gene labels at the left, and an extra row of assay labels at the top. The SVD is calculated using the CLAPACK linear algebra library (http://www.netlib.org/clapack/). The gene grouping algorithm makes use of a novel threshold that is similar to a $3\sigma$ statistical significance cutoff. Each element of each gene coefficient vector is compared to the value $WN^{-1/2}$, where *N* is the number of genes and *W* is a weight factor whose recommended value is 3. If the magnitude of the element is greater than $WN^{-1/2}$, the corresponding gene is placed in the group corresponding to the gene coefficient vector. Each column of U defines a unique group. The significance of the threshold is that it provides a scale-independent way to determine group membership. The sign of the gene coefficient indicates whether a gene has a positive or negative response to the expression level vector. This means that if a transcriptional regulator has a promoter activity for gene A and a repressor activity for gene B, SVDMAN can place genes A and B in the same group while preserving the promoter/repressor distinction.

SVDMAN uses a novel confidence measure to evaluate results. The program iterates through all columns $\{a_j\}$ (i.e. assays) but the first and last in A. A new matrix A' is created by deleting $a_j$. The SVD of A' is calculated, and the elements of an extra column $v^i_j$ of $V'^T$ are calculated by cubic spline interpolation (see Press *et al.*, 1992). A reconstructed column $a^i_j$ of A is calculated by multiplying U'Σ' times $v^i_j$.. The confidence measure is the Pearson correlation between $a_j$ and $a^i_j$. A low value can suggest high information content due to excess noise or undersampling (see Discussion and Conclusion), which would produce a relatively flat spectrum of mode amplitudes in Σ. A high value is consistent with good signal-to-noise and sampling, but is not sufficient to validate these conditions.

SVDMAN is executed in the Unix shell (a Windows version is under development) by typing "`svdman –i input.dat –o output_base`" where `input.dat` is a tab-delimited matrix of microarray data in a format like that in the Stanford Microarray Database (http://genome-www4.stanford.edu/MicroArray/SMD/restech.html), and `output_base` is the base name for all output files. Output files are tab-delimited ASCII flat files that enable easy importing into spreadsheets for visualization and parsing by scripts. Seven assays on 6000 genes can be analyzed in a one second on a 500 MHz processor.

**Discussion and Conclusion**

Here we contrast SVDMAN with two other methods of gene expression analysis: clustering (see, e.g., Eisen *et al.*, 1998) and gene shaving (Hastie *et al.*, 2000). In clustering, which does not make use of SVD, groups of genes are defined by having similar expression patterns according to a rigorous distance metric (e.g., correlation). Common clustering methods generate mutually exclusive groups, a limitation considering that a gene product (e.g. a kinase) may participate in several cellular reaction networks. Gene shaving does make use of PCA/SVD, but analysis is restricted to the first principal component, and higher-order groups are obtained by orthogonalization of the expression data with respect to an averaged total expression profile for the group. Groups are generated by iterative exclusion of a fixed fraction of genes and evaluation of the optimal group size using a "gap statistic." As in cluster analysis, genes in a group

have similar overall expression patterns, but they are not mutually exclusive. In SVDMAN, all principal components are considered, with higher-order groups coming from higher-order components. A group is formed by identifying all genes whose expression patterns are significantly influenced by a given mode, generating a hypothesis that the genes participate in a common genetic network. Genes in a group do not necessarily have similar overall expression patterns, and groups are not mutually exclusive. There is also a self-consistency measure to evaluate the analysis. Clustering and gene shaving are well-suited for categorization of comparative clinical data and detecting correlations between gene expression patterns (e.g., classifying disease states). SVDMAN may be useful for similar applications, but it is particularly well-suited to analysis of time-series data, where insight into modulation of genetic networks is desired.

We now describe a method for using our confidence measure to detect undersampling. If the data are highly oversampled, the SVD will have an uneven distribution of mode amplitudes, with most of the data accounted for by a small fraction of the modes. If the data are excessively noisy or have high information content, the SVD will have a flat distribution of mode amplitudes. In the case of high oversampling, any single deleted assay can be reconstructed with great accuracy. As the sampling becomes sparser, the correlations remain good until the sampling interval becomes comparable to the finest feature in the data. If the number assays with poor correlation values is small compared to the total number of assays, the distribution of mode amplitudes may remain flat. The confidence measure thus should enable localized sampling problems to be detected where the mode amplitudes fail to indicate a problem. If the problem assays correspond to parameter values (e.g. time points) of biological interest, the researcher can choose to acquire data that more finely sample the relevant range. To use this method consistently, it will be necessary to use well-characterized data (either experimental or simulated) to systematically study how the degree of undersampling gives rise to specific correlation values.

The main biological significance of SVDMAN is its novel means of generating gene groups, which provide testable hypotheses for gene associations. Previously published work has demonstrated that biologically significant results can be obtained from SVD analysis of microarray data (Raychaudhuri *et al.*, 2000, Holter *et al.*, 2000, Alter *et al.*, 2000). In addition we have validated the SVDMAN methods both computationally

and biologically using publically available microarray data combined with biology databases (Dyck *et al.*, 2000). Our purpose in distributing SVDMAN is twofold: to give biologists a useful application for gene expression analysis; and to make the software available to developers who wish to use SVDMAN in their own applications, e.g., as a component of a larger framework for analysis of expression data.

## Acknowledgments

We gratefully acknowledge Luis Rocha for assistance with implementation of the SVD algorithm, and Michael Altherr for critically reading the manuscript. This work was performed under the auspices of the Department of Energy under contract to the University of California and was supported by the Molecular Foundations of Pathogenesis project, funded by Laboratory Directed Research and Development at Los Alamos National Laboratory.

**References**

Alter O, Brown PO, Botstein D. (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc Natl Acad Sci U S A*, **97**, 10101-6

Chu S, DeRisi J, Eisen M, Mulholland J, Botstein D, Brown PO, Herskowitz I. (1998) The transcriptional program of sporulation in budding yeast [published erratum appears in science 1998 nov 20;282(5393):1421]. *Science*, **282**, 699-705

Chen L, Hodgson KO, Doniach S. (1996) A lysozyme folding intermediate revealed by solution X-ray scattering . *J Mol Biol*, **261**, 658-671

Dyck PA, Wall ME, Brettin T. (2000) Validation of singular value decomposition of global expression data. *J Comp Biol*, **7**, 636

Eisen MB, Spellman PT, Brown PO, Botstein D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc Natl Acad Sci U S A*, **95**, 14863-8

Garcia A. (1992) Large-amplitude nonlinear motions in proteins. *Phys Rev Lett*, **68**, 2696-2699

Hastie T, Tibshirani R, Eisen MB, Alizadeh A, Levy R, Staudt L, Chan WC, Botstein D, Brown P. (2000) 'Gene shaving' as a method for identifying distinct sets of genes with similar expression patterns. *Genome Biol*, **1**, research0003.1-03.21

Holter NS, Mitra M, Maritan A, Cieplak M, Banavar JR, Fedoroff NV. (2000) Fundamental patterns underlying gene expression profiles: Simplicity from complexity. *Proc Natl Acad Sci U S A*, **97**, 8409-14

Iyer VR, Eisen MB, Ross DT, Schuler G, Moore T, Lee JCF, Trent JM, Staudt LM, Hudson J, Jr., Boguski

MS, Lashkari D, Shalon D, Botstein D, Brown PO. (1999) The transcriptional program in the response of human fibroblasts to serum. *Science*, **283**, 83-7

Press WH, Teukolsky SA, Vetterling WT, Flannery BP. (1992) Numerical recipes in C, second edition. Cambridge University Press (Cambridge, UK)

Raychaudhuri S, Stuart JM, Altman RB. (2000) Principal components analysis to summarize microarray experiments: Application to sporulation time series. *Pac Symp Biocomput*, 455-66

Romo TD, Clarage JB, Sorensen DC, Phillips GN, Jr. (1995) Automatic identification of discrete substates in proteins: Singular value decomposition analysis of time-averaged crystallographic refinements. *Proteins*, **22**, 311-21

Spellman PT, Sherlock G, Zhang MQ, Iyer VR, Anders K, Eisen MB, Brown PO, Botstein D, Futcher B. (1998) Comprehensive identification of cell cycle-regulated genes of the yeast *Saccharomyces cerevisiae* by microarray hybridization. *Mol Biol Cell*, **9**, 3273-97