A Higher-Order Generalized Singular Value Decomposition for Comparison of Global mRNA Expression from Multiple Organisms

Sri Priya Ponnapalli, Michael A. Saunders, Charles F. Van Loan, Orly Alter

Abstract

The number of high-dimensional datasets recording multiple aspects of a single phenomenon is increasing in many areas of science, accompanied by a need for mathematical frameworks that can compare multiple large-scale matrices with different row dimensions. The only such framework to date, the generalized singular value decomposition (GSVD), is limited to two matrices. We mathematically define a higher-order GSVD (HO GSVD) for \( N \geq 2 \) matrices \( D_i \in \mathbb{R}^{m_i \times s} \), each with full column rank. Each matrix is exactly factored as \( D_i = U_i \Sigma_i V_i^T \), where \( V_i \), identical in all factorizations, is obtained from the eigensystem \( SV = VA \) of the arithmetic mean \( S \) of all pairwise quotients \( A_i A_j^{-1} \) of the matrices \( A_i = D_i^T D_i \), \( i \neq j \). We prove that this decomposition extends to higher orders almost all of the mathematical properties of the GSVD. The matrix \( S \) is nondefective with \( V \) and \( A \) real. Its eigenvalues satisfy \( \lambda_k \geq 1 \). Equality holds if and only if the corresponding eigenvector \( v_k \) is a right basis vector of equal significance in all matrices \( D_i \) and \( D_j \), that is \( \sigma_i v_k / \sigma_j v_k = 1 \) for all \( i \) and \( j \) and the corresponding left basis vector \( u_k \) is orthogonal to all other vectors in \( U_i \), for all \( i \). The eigenvalues \( \lambda_n = 1 \), therefore, define the “common HO GSVD subspace.” We illustrate the HO GSVD with a comparison of genome-scale cell-cycle mRNA expression from \( S. pombe \), \( S. cerevisiae \) and human. Unlike existing algorithms, a mapping among the genes of these disparate organisms is not required. We find that the approximately common HO GSVD subspace represents the cell-cycle mRNA expression oscillations, which are similar among the datasets. Simultaneous reconstruction in the common subspace, therefore, removes the experimental artifacts, which are dissimilar, from the datasets. In the simultaneous sequence-independent classification of the genes of the three organisms in this common subspace, genes of highly conserved sequences but significantly different cell-cycle peak times are correctly classified.

Introduction

In many areas of science, especially in biotechnology, the number of high-dimensional datasets recording multiple aspects of a single phenomenon is increasing. This is accompanied by a fundamental need for mathematical frameworks that can compare multiple large-scale matrices with different row dimensions. For example, comparative analyses of global mRNA expression from multiple model organisms promise to enhance fundamental understanding of the universality and specialization of molecular biological mechanisms, and may prove useful in medical diagnosis, treatment and drug design [1]. Existing algorithms limit analyses to subsets of homologous genes among the different organisms, effectively introducing into the analysis the assumption that sequence and functional similarities are equivalent [e.g., [2]]. However, it is well known that this assumption does not always hold, for example, in cases of nonorthologous gene displacement, when nonorthologous proteins in different organisms fulfill the same function [3]. For sequence-independent comparisons, mathematical frameworks are required that can distinguish and separate the similar from the dissimilar among multiple large-scale datasets tabulated as matrices with different row dimensions, corresponding to the different sets of genes of the different organisms. The only such framework to date, the generalized singular value decomposition (GSVD) [4–7], is limited to two matrices.

It was shown that the GSVD provides a mathematical framework for sequence-independent comparative modeling of DNA microarray data from two organisms, where the mathematical variables and operations represent biological reality [7,8]. The variables, significant subspaces that are common to both or exclusive to either one of the datasets, correlate with cellular programs that are conserved in both or unique to either one of the organisms, respectively. The operation of reconstruction in the
subspaces common to both datasets outlines the biological similarity in the regulation of the cellular programs that are conserved across the species. Reconstruction in the common and exclusive subspaces of either dataset outlines the differential regulation of the conserved relative to the unique programs in the corresponding organism. Recent experimental results [9] verify a computationally predicted genome-wide mode of regulation that correlates DNA replication origin activity with mRNA expression [10,11], demonstrating that GSVD modeling of DNA microarray data can be used to correctly predict previously unknown cellular mechanisms.

We now define a higher-order GSVD (HO GSVD) for the comparison of N ≥ 2 datasets. The datasets are tabulated as N real matrices $D_e \in \mathbb{R}^{m \times n}$, each with full column rank, with different row dimensions and the same column dimension, where there exists a one-to-one mapping among the columns of the matrices. Like the GSVD, the HO GSVD is an exact decomposition, i.e., each matrix is exactly factored as $D_i = U_i \Sigma_i V_i^T$, where the columns of $U_i$ and $V_i$ have unit length and are the left and right basis vectors respectively, and each $\Sigma_i$ is diagonal and positive definite. Like the GSVD, the matrix $V$ is identical in all factorizations. In our HO GSVD, the matrix $V$ is obtained from the eigensystem $SV = VA$ of the arithmetic mean $S$ of all pairwise quotients $A_i A_j^{-1}$ of the matrices $A_i = D_i^T D_i$, or equivalently of all $S_i = \frac{1}{2} (A_i A_j^{-1} + A_j A_i^{-1})$, $i \neq j$.

To clarify our choice of $S$, we note that in the GSVD, defined by Van Loan [5], the matrix $V$ can be formed from the eigenvectors of the unbalanced quotient $A_i A_j^{-1}$ (Section 1 in Appendix S1). We observe that this $V$ can also be formed from the eigenvectors of the balanced arithmetic mean $S_{12} = \frac{1}{2} (A_i A_j^{-1} + A_j A_i^{-1})$. We prove that in the case of $N = 2$, our definition of $V$ by using the eigensystem of $S = S_{12} = \frac{1}{2} (A_i A_j^{-1} + A_j A_i^{-1})$ leads algebraically to the GSVD (Theorems S1–S5 in Appendix S1), and therefore, as Paige and Saunders showed [6], can be computed in a stable way. We also note that in the GSVD, the matrix $V$ does not depend upon the ordering of the matrices $D_1$ and $D_2$. Therefore, we define our HO GSVD for $N \geq 2$ matrices by using the balanced arithmetic mean $S$ of all pairwise arithmetic means $S_i$, each of which defines the GSVD of the corresponding pair of matrices $D_i$ and $D_j$, noting that $S$ does not depend upon the ordering of the matrices $D_i$ and $D_j$.

We prove that $S$ is nondefective (it has n independent eigenvectors), and that its eigensystem is real (Theorem 1). We prove that the eigenvalues of $S$ satisfy $\lambda_k \geq 1$ (Theorem 2). As in our GSVD comparison of two matrices [7], we interpret the kth diagonal of $\Sigma_i = \text{diag}(\sigma_{i,k})$ in the factorization of the $i$th matrix $D_i$ as indicating the significance of the $k$th right basis vector $v_k$ in $D_i$ in terms of the overall information that $v_k$ captures in $D_i$. The ratio $\sigma_{i,k}/\sigma_{j,k}$ indicates the significance of $v_k$ in $D_i$ relative to its significance in $D_j$. We prove that an eigenvalue of $S$ satisfies $\lambda_k = 1$ if and only if the corresponding eigenvector $v_k$ is a right basis vector of equal significance in all $D_i$ and $D_j$, that is, $\sigma_{i,k}/\sigma_{j,k} = 1$ for all $i$ and $j$, and the corresponding left basis vector $u_k$ is orthonormal to all other vectors in $U_i$ for all $i$. We therefore mathematically define, in analogy with the GSVD, the “common HO GSVD subspace” of the $N \geq 2$ matrices to be the subspace spanned by the right basis vectors $v_k$ that correspond to the $\lambda_k = 1$ eigenvalues of $S$ (Theorem 3). We also show that each of the right basis vectors $\{v_k\}$ that span the common HO GSVD subspace is a generalized singular vector of all pairwise GSVD factorizations of the matrices $D_i$ and $D_j$ with equal corresponding generalized singular values for all $i$ and $j$ (Corollary 1).

Recent research showed that several higher-order generalizations are possible for a given matrix decomposition, each preserving some but not all of the properties of the matrix decomposition [12–14] (see also Theorem S6 and Conjecture S1 in Appendix S1). Our new HO GSVD extends to higher orders all of the mathematical properties of the GSVD except for complete column-wise orthogonality of the left basis vectors that form the matrix $U_i$ for all $i$, i.e., in each factorization.

We illustrate the HO GSVD with a comparison of cell-cycle mRNA expression from S. pombe [15,16], S. cerevisiae [17] and human [18]. Unlike existing algorithms, a mapping among the genes of these disparate organisms is not required (Section 2 in Appendix S1). We find that the common HO GSVD subspace represents the cell-cycle mRNA expression oscillations, which are similar among the datasets. Simultaneous reconstruction in this common subspace, therefore, removes the experimental artifacts, which are dissimilar, from the datasets. Simultaneous sequence-independent classification of the genes of the three organisms in the common subspace is in agreement with previous classifications into cell-cycle phases [19]. Notably, genes of highly conserved sequences across the three organisms [20,21] but significantly different cell-cycle peak times, such as genes from the ABC transporter superfamily [22–28], phospholipid B-encoding genes [29,30] and even the B cyclin-encoding genes [31,32], are correctly classified.

### Methods

#### HO GSVD Construction

Suppose we have a set of $N$ real matrices $D_e \in \mathbb{R}^{m \times n}$ each with full column rank. We define a HO GSVD of these $N$ matrices as

$$
D_1 = U_1 \Sigma_1 V_1^T,
D_2 = U_2 \Sigma_2 V_2^T,
\vdots
D_N = U_N \Sigma_N V_N^T,
$$

where each $U_i \in \mathbb{R}^{m \times n}$ is composed of normalized left basis vectors, each $\Sigma_i = \text{diag}(\sigma_{i,k}) \in \mathbb{R}^{n \times n}$ is diagonal with $\sigma_{i,k} > 0$, and $V$ is identical in all matrix factorizations, is composed of normalized right basis vectors. As in the GSVD comparison of global mRNA expression from two organisms [7], in the HO GSVD comparison of global mRNA expression from $N \geq 2$ organisms, the shared right basis vectors $v_k$ of Equation (1) are the “genelists” and the $N$ sets of left basis vectors $u_{i,k}$ are the $N$ sets of “arraylists” (Figure 1 and Section 2 in Appendix S1). We obtain $V$ from the eigensystem of $S$, the arithmetic mean of all pairwise quotients $A_i A_j^{-1}$ of the matrices $A_i = D_i^T D_i$, or equivalently of all $S_{ij} = \frac{1}{2} (A_i A_j^{-1} + A_j A_i^{-1})$, $i \neq j$:

$$
S = \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1, j \neq i}^{N} (A_i A_j^{-1} + A_j A_i^{-1})
= \frac{2}{N(N-1)} \sum_{i=1}^{N} \sum_{j=1}^{N} S_{ij},
$$

$$
SV = VA,
V = (v_1, \ldots, v_N), \quad A = \text{diag}(\lambda_k),
$$

with $\|v_k\| = 1$. We prove that $S$ is nondefective, i.e., $S$ has $n$ independent eigenvectors, and that its eigenvectors $V$ and eigenvalues $A$ are real (Theorem 1). We prove that the eigenvalues of $S$ satisfy $\lambda_k \geq 1$ (Theorem 2).

Given $V$, we compute matrices $B_i$ by solving $N$ linear systems:

$$
VB_i^T = D_i^T,
B_i = (b_{i1}, \ldots, b_{in}), \quad i = 1, \ldots, N,
$$
genelets, i.e., right basis vectors mathematically define, in analogy with the GSVD, the “common HO GSVD subspace” of the matrix $V^T$ transformed to the reduced diagonalized matrices $U_i$, $U_j$, and $U_3$ and the shared 17-genelets 17-arrays transformation matrix $V^T$. We prove that with our particular $V$ of Equations (2)–(4), this decomposition extends to higher orders all of the mathematical properties of the GSVD except for complete column-wise orthogonality of the arraylets, i.e., left basis vectors that form the matrices $U_i$, $U_j$, and $U_3$. We therefore mathematically define, in analogy with the GSVD, the “common HO GSVD subspace” of the $N=3$ matrices to be the subspace spanned by the genelets, i.e., right basis vectors $v_k$ that correspond to higher-order generalized singular values that are equal, $s_{i,k} = s_{j,k} = s_{g,k}$, where, as we prove, the corresponding arraylets, i.e., the left basis vectors $u_{1,k}$, $u_{2,k}$, and $u_{3,k}$, are orthonormal to all other arraylets in $U_i$, $U_j$, and $U_3$. We show that like the GSVD for two organisms [7], the HO GSVD provides a sequence-independent comparative mathematical framework for datasets from more than two organisms, where the mathematical variables and operations represent biological reality: Genelets of common significance in the multiple datasets, and the corresponding arraylets, represent cell-cycle checkpoints or transitions from one phase to the next, common to human. Simultaneous reconstruction and classification of the three datasets in the common subspace that these patterns span outline the biological similarity in the regulation of their cell-cycle programs. Notably, genes of significantly different cell-cycle peak times [19] but highly conserved sequences [20,21] are correctly classified.

doi:10.1371/journal.pone.0028072.g001

and we construct $\Sigma_i$ and $U_i = (u_{i,1} \ldots u_{i,N})$ by normalizing the columns of $B_i$:

$$\sigma_{i,k} = \|h_{i,k}\|,$$

$$\Sigma_i = \text{diag}(\sigma_{i,k}),$$

$$B_i = U_i\Sigma_i.$$  

HO GSVD Interpretation

In this construction, the rows of each of the $N$ matrices $D_i$ are superpositions of the same right basis vectors, the columns of $V$ (Figures S1 and S2 and Section 1 in Appendix S1). As in our GSVD comparison of two matrices, we interpret the $k$th diagonals of $\Sigma_i$, the “higher-order generalized singular value set” $\{\sigma_{i,k}\}$, as indicating the significance of the $k$th right basis vector $v_k$ in the matrices $D_i$, and reflecting the overall information that $v_k$ captures in each $D_i$ respectively. The ratio $\sigma_{i,k}/\sigma_{j,k}$ indicates the significance of $v_k$ in $D_i$ relative to its significance in $D_j$. A ratio of $\sigma_{i,k}/\sigma_{j,k} = 1$ for all $i$ and $j$ corresponds to a right basis vector $v_k$ of equal significance in all $N$ matrices $D_i$. GSVD comparisons of two matrices showed that right basis vectors of approximately equal significance in the two matrices reflect themes that are common to both matrices under comparison [7]. A ratio of $\sigma_{i,k}/\sigma_{j,k} \ll 1$ indicates a basis vector $v_k$ of almost negligible
significance in $D_i$ relative to its significance in $D_j$. GSVD comparisons of two matrices showed that right basis vectors of negligible significance in one matrix reflect themes that are exclusive to the other matrix.

We prove that an eigenvalue of $S$ satisfies $\lambda_k = 1$ if and only if the corresponding eigenvector $v_k$ is a right basis vector of equal significance in all $D_i$ and $D_j$, that is, $\sigma_i, k / \sigma_j, k = 1$ for all $i$ and $j$, and the corresponding left basis vector $u_k$ is orthonormal to all other vectors in $U_i$ for all $i$. We therefore mathematically define, in analogy with the GSVD, the “common HO GSVD subspace” of the $N \geq 2$ matrices to be the subspace spanned by the right basis vectors $\{v_k\}$ corresponding to the eigenvalues of $S$ that satisfy $\lambda_k = 1$ (Theorem 3).

It follows that each of the right basis vectors $\{v_k\}$ that span the common HO GSVD subspace is a generalized singular vector of all pairwise GSVD factorizations of the matrices $D_i$ and $D_j$ with equal corresponding generalized singular values for all $i$ and $j$ (Corollary 1). Since the GSVD can be computed in a stable way [6], we note that the common HO GSVD subspace can also be computed in a stable way by computing all pairwise GSVD factorizations of the matrices $D_i$ and $D_j$. This also suggests that it may be possible to formulate the HO GSVD as a solution to an optimization problem, in analogy with existing variational formulations of the GSVD [33]. Such a formulation may lead to a stable numerical algorithm for computing the HO GSVD, and possibly also to a higher-order general Gauss-Markov linear statistical model [34–36].

We show, in a comparison of $N = 3$ matrices, that the approximately common HO GSVD subspace of these three matrices reflects a theme that is common to the three matrices under comparison (Section 2).

HO GSVD Mathematical Properties

**Theorem 1.** $S$ is nondefective (it has $n$ independent eigenvectors) and its eigensystem is real.

**Proof.** From Equation (2) it follows that

$$S = \frac{1}{N(N-1)}(H - NT),$$

and the eigenvectors of $S$ equal the eigenvectors of $H$.

Let the SVD of the matrices $D_i$ appended along the $n$-columns axis be

$$[D_1 \ldots D_N] = [U_1 \ldots U_N] \Sigma \hat{V}^T,$$

and the eigenvalues of $H$ equal the eigenvalues of $\hat{H}$.

A sum of real, symmetric and positive definite matrices, $\hat{H}$ is also real, symmetric and positive definite; therefore, its eigensystem

$$\hat{Y}^T \hat{Y} = \text{diag}(\mu_k)$$

is real with $\hat{Y}$ orthogonal and $\mu_k > 0$. Without loss of generality let $\hat{Y}$ be orthonormal, such that $\|y_k\| = 1$. It follows from the similarity of $H$ with $\hat{H}$ that the common HO GSVD subspace can also be computed in a stable way by computing all pairwise GSVD factorizations of $H$.

From Equation (5), $S$ is nondefective with real eigenvalues. Thus, the eigenvalues of $S$ satisfy

$$\lambda_k = \frac{1}{N(N-1)}(\mu_k - N),$$

where $\mu_k > 0$ are the eigenvalues of $H$ and $\hat{H}$. Thus, the eigenvalues of $S$ are real. \(\Box\)

**Theorem 2.** The eigenvalues of $S$ satisfy $\lambda_k \geq 1$.

**Proof.** Following Equation (9), asserting that the eigenvalues of $S$ satisfy $\lambda_k \geq 1$ is equivalent to asserting that the eigenvalues of $H$ satisfy $\mu_k \geq N^2$.

From Equations (6) and (7), the eigenvalues of $H$ satisfy

$$\mu_k \geq \min_x \sum_{j=1}^N |x^T (U_j^T U_j)x|^{-1},$$

under the constraint that

$$\sum_{j=1}^N x^T (U_j^T U_j)x = 1,$$

where $x$ is a real unit vector, and where it follows from the Cauchy-Schwarz inequality [37] (see also [4,34,38]) for the real nonzero vectors $(U_j^T U_j)x$ and $(U_j^T U_j)^{-1}x$ that for all $j$

$$x^T (U_j^T U_j)^{-1}x \geq |x^T (U_j^T U_j)x|^{-1}.$$

With the constraint of Equation (11), which requires the sum of the $N$ positive numbers $x^T (U_j^T U_j)^{-1}x$ equal to one, the lower bound on the eigenvalues of $H$ in Equation (10) is at its minimum when the sum of the inverses of these numbers is at its minimum, that is, when the numbers equal

$$x^T (U_j^T U_j)x = x^T (U_j^T U_j)x = N^{-1}$$

for all $i$ and $j$. Thus, the eigenvalues of $\hat{H}$ satisfy $\mu_k \geq N^2$. \(\Box\)

**Theorem 3.** The common HO GSVD subspace. An eigenvalue of $S$ satisfies $\lambda_k = 1$ if and only if the corresponding eigenvector $v_k$ is a right basis vector of equal significance in all $D_i$ and $D_j$, that is, $\sigma_i, k / \sigma_j, k = 1$ for all $i$ and $j$, and the corresponding left basis vector $u_k$ is orthonormal to all other vectors in $U_i$ for all $i$. The “common HO GSVD subspace” of the $N \geq 2$ matrices is, therefore, the subspace spanned by the right basis vectors $\{v_k\}$ corresponding to the eigenvalues of $S$ that satisfy $\lambda_k = 1$.

**Proof.** Without loss of generality, let $k = n$. From Equation (12) and the Cauchy-Schwarz inequality, an eigenvalue of $H$ equals its minimum lower bound $\mu_n = N^2$ if and only if the corresponding
The common HO GSVD subspace we define (Theorem 3) can also correspond to the right basis vector quotients $\tilde{v}_k = \tilde{w}_k^{-1}V\Sigma y_k$ for all $k$ [37], where, from Equation (13), the corresponding eigenvalue equals $N^{-1}$,

$$\tilde{U}_j^T \tilde{U}_j \tilde{Y} = (\tilde{U}_j^T \tilde{U}_j) \tilde{y}_1 \tilde{y}_2 \cdots \tilde{y}_N = \tilde{y}_n.$$  \hspace{1cm} (14)

Given the eigenvectors $V = (\Sigma \tilde{Y}) W^{-1}$ of $S$, we solve Equation (3) for each $D_i = \tilde{U}_i \Sigma_i P_i^T$ of Equation (6), and obtain

$$U_i \Sigma_i = B_i = D_i V^{-T} = \tilde{U}_i \tilde{Y} \tilde{W}.$$  \hspace{1cm} (15)

Following Equations (14) and (15), where $v_k = \tilde{w}_k^{-1}V\Sigma y_k$ corresponds to a minimum eigenvalue $\lambda_k=1$, and since $\tilde{Y}$ is orthonormal, we obtain

$$\tilde{W}^{-1} \Sigma_i (U_i^T \tilde{U}_i) \Sigma_i \tilde{W}^{-1} = \tilde{Y}^T (\tilde{U}_i^T \tilde{U}_j) \tilde{Y}$$

$$= \begin{bmatrix}
y_1^T (\tilde{U}_i^T \tilde{U}_j) \tilde{y}_1 & y_2^T (\tilde{U}_i^T \tilde{U}_j) \tilde{y}_1 & \cdots & 0 \\
y_1^T (\tilde{U}_i^T \tilde{U}_j) \tilde{y}_2 & y_2^T (\tilde{U}_i^T \tilde{U}_j) \tilde{y}_2 & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots \\
0 & 0 & \cdots & N^{-1}
\end{bmatrix}.$$  \hspace{1cm} (16)

with zeros in the $m$th row and the $n$th column of the matrix above everywhere except for the diagonal element. Thus, an eigenvalue of $S$ satisfies $\lambda_k=1$ if and only if the corresponding left basis vectors $m_{k,n}$ are orthonormal to all other vectors in $U_i$.

The corresponding higher-order generalized singular values are $\sigma_{k,n} = N^{-1/2} \tilde{w}_k > 0$. Thus $\sigma_{k,n}^2 = \tilde{w}_k > 0$ for all $i$ and $j$, and the corresponding right basis vector $v_k$ is of equal significance in all matrices $D_i$ and $D_j$.

**Corollary 1.** An eigenvalue of $S$ satisfies $\lambda_k=1$ if and only if the corresponding right basis vector $v_k$ is a generalized singular vector of all pairwise GSVD factorizations of the matrices $D_i$ and $D_j$.

**Proof.** From Equations (12) and (13), and since the pairwise quotients $A_i A_j^{-1}$ are similar to $(U_i^T \tilde{U}_j) (U_i^T \tilde{U}_j)^{-1}$ with the similarity transformation of $\tilde{V} \Sigma$ for all $i$ and $j$, it follows that an eigenvalue of $S$ satisfies $\lambda_k=1$ if and only if the corresponding right basis vector $v_k = \tilde{w}_k^{-1}V\Sigma y_k$ is also an eigenvector of each of the pairwise quotients $A_i A_j^{-1}$ of the matrices $A_i = D_i^T D_i$ with equal corresponding eigenvalues, or equivalently of all $S_j$ with all eigenvalues at their minimum of one,

$$S_j v_k = \frac{1}{2} (A_i A_j^{-1} + A_i A_j^{-1}) v_k = v_k.$$  \hspace{1cm} (17)

We prove (Theorems S1–S5 in Appendix S1) that in the case of $N=2$ matrices our definition of $V$ by using the eigensystem of $S_j$ leads algebraically to the GSVD, where an eigenvalue of $S_j$ equals its minimum of one if and only if the two corresponding generalized singular values are equal, such that the corresponding generalized singular vector of $v_k$ is of equal significance in both matrices $D_i$ and $D_j$. Thus, it follows that each of the right basis vectors $\{v_k\}$ span the common HO GSVD subspace is a generalized singular vector of all pairwise GSVD factorizations of the matrices $D_i$ and $D_j$ with equal corresponding generalized singular values for all $i$ and $j$. \hspace{1cm} \Box

Note that since the GSVD can be computed in a stable way [6], the common HO GSVD subspace we define (Theorem 3) can also be computed in a stable way by computing all pairwise GSVD factorizations of the matrices $D_i$ and $D_j$ (Corollary 1). It may also be possible to formulate the HO GSVD as a solution to an optimization problem, in analogy with existing variational formulations of the GSVD [33]. Such a formulation may lead to a more stable numerical algorithm for computing the HO GSVD, and possibly also to a higher-order general Gaus-Markov linear statistical model [34–36].

**Results**

**HO GSVD Comparison of Global mRNA Expression from Three Organisms**

Consider now the HO GSVD comparative analysis of global mRNA expression datasets from the $N=3$ organisms $S. pombe$, $S. cerevisiae$ and human (Section 2.1 in Appendix S1, Mathematica Notebooks S1 and S2, and Datasets S1, S2 and S3). The datasets are tabulated as matrices of $n=17$ columns each, corresponding to DNA microarray-measured mRNA expression from each organism at 17 time points equally spaced during approximately two cell-cycle periods. The underlying assumption is that there exists a one-to-one mapping among the 17 columns of the three matrices but not necessarily among their rows, which correspond to either $m_1=3167$-S. pombe genes, $m_2=4772$-S. cerevisiae genes or $m_3=13,068$-human genes. The HO GSVD of Equation (1) transforms the datasets from the organism-specific genes $17$-arrays spaces to the reduced spaces of the $17$-“arraylets,” i.e., left basis vectors $17$-“genelets,” i.e., right basis vectors, where the datasets $D_i$ are represented by the diagonal nonnegative matrices $\Sigma_i$, by using the organism-specific genes $17$-arraylets transformation matrices $U_i$ and the one shared $17$-genelets $17$-arrays transformation matrix $V^T$ (Figure 1).

Following Theorem 3, the approximately common HO GSVD subspace of the three datasets is spanned by the five genelets $k=1,\ldots,17$ that correspond to $1 \leq l_k \leq 2$. We find that these five genelets are approximately equally significant with $\sigma_{1,k} \approx \sigma_{2,k} \approx \sigma_{3,k} \approx 1:1:1$ in the $S. pombe$, $S. cerevisiae$ and human datasets, respectively (Figure 2 a and b). The five corresponding arraylets in each dataset are $v=0.33$-orthonormal to all other arraylets (Figure S3 in Appendix S1).

**Common HO GSVD Subspace Represents Similar Cell-Cycle Oscillations**

The expression variations across time of the five genelets that span the approximately common HO GSVD subspace fit normalized cosine functions of two periods, superimposed on time-invariant expression (Figure 2 c and d). Consistently, the corresponding organism-specific arraylets are enriched [39] in overexpressed or underexpressed organism-specific cell cycle-regulated genes, with $24$ of the $30 P$-values $<10^{-5}$ (Table 1 and Section 2.2 in Appendix S1). For example, the three $17$ arraylets, which correspond to the $0$-phase $17$ genelet, are enriched in overexpressed $G2$ $S. pombe$ genes, $G2/M$ and $M/G1$ $S. cerevisiae$ genes and $S$ and $G2$ human genes, respectively, representing the cell-cycle checkpoints in which the three cultures are initially synchronized.

Simultaneous sequence-independent reconstruction and classification of the three datasets in the common subspace outline cell-cycle progression in time and across the genes in the three organisms (Sections 2.3 and 2.4 in Appendix S1). Projecting the expression of the $17$ arrays of either organism from the corresponding five-dimensional arraylets subspace onto the two-dimensional subspace that approximates it (Figure S4 in Appendix S1), $\geq 50\%$ of the contributions of the arraylets add.
up, rather than cancel out (Figure 3 a–c). In these two-dimensional subspaces, the angular order of the arrays of either organism describes cell-cycle progression in time through approximately two cell-cycle periods, from the initial cell-cycle phase and back to that initial phase twice. Projecting the expression of the genes, ≥50% of the contributions of the five genelets add up in the overall expression of 343 of the 380 S. pombe genes classified as cell cycle-regulated, 554 of the 641 S. cerevisiae cell-cycle genes, and 632 of the 787 human cell-cycle genes (Figure 3 d–f). Simultaneous classification of the genes of either organism into cell-cycle phases according to their angular order in these two-dimensional subspaces is consistent with the classification of the arrays, and is in good agreement with the previous classifications of the genes (Figure 3 g–i). With all 3167 S. pombe, 4772 S. cerevisiae and 13,068 human genes sorted, the expression variations of the five arrays from each organism approximately fit one-period cosines, with the initial phase of each array (Figures S5, S6, S7 in Appendix S1) similar to that of its corresponding genelet (Figure 2). The global mRNA expression of each organism, reconstructed in the common HO GSVD subspace, approximately fits a traveling wave, oscillating across time and across the genes.

Table 1. Arraylets or left basis vectors.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Arraylet</th>
<th>Overexpression</th>
<th>P-value</th>
<th>Underexpression</th>
<th>P-value</th>
</tr>
</thead>
<tbody>
<tr>
<td>S. pombe</td>
<td>13</td>
<td>G2</td>
<td>2.4 x 10^{-10}</td>
<td>G1</td>
<td>1.0 x 10^{-15}</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>M</td>
<td>2.2 x 10^{-21}</td>
<td>G2</td>
<td>1.3 x 10^{-9}</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>M</td>
<td>4.1 x 10^{-13}</td>
<td>S</td>
<td>1.6 x 10^{-17}</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>G2</td>
<td>5.2 x 10^{-18}</td>
<td>G1</td>
<td>1.2 x 10^{-26}</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>G2</td>
<td>2.4 x 10^{-10}</td>
<td>S</td>
<td>5.3 x 10^{-35}</td>
</tr>
<tr>
<td>S. cerevisiae</td>
<td>13</td>
<td>S/G2</td>
<td>4.3 x 10^{-15}</td>
<td>M/G1</td>
<td>1.4 x 10^{-9}</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>M/G1</td>
<td>4.9 x 10^{-26}</td>
<td>G2/M</td>
<td>2.2 x 10^{-12}</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>G1</td>
<td>7.7 x 10^{-17}</td>
<td>S</td>
<td>1.3 x 10^{-8}</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>G2/M</td>
<td>2.3 x 10^{-38}</td>
<td>G1</td>
<td>2.0 x 10^{-32}</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>G2/M</td>
<td>2.3 x 10^{-41}</td>
<td>G1</td>
<td>2.6 x 10^{-40}</td>
</tr>
<tr>
<td>Human</td>
<td>13</td>
<td>G1/S</td>
<td>1.1 x 10^{-33}</td>
<td>G2</td>
<td>2.4 x 10^{-44}</td>
</tr>
<tr>
<td></td>
<td>14</td>
<td>M/G1</td>
<td>5.7 x 10^{-3}</td>
<td>G2</td>
<td>4.7 x 10^{-2}</td>
</tr>
<tr>
<td></td>
<td>15</td>
<td>G2</td>
<td>9.8 x 10^{-24}</td>
<td>None</td>
<td>1.4 x 10^{-1}</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>G1/S</td>
<td>9.8 x 10^{-13}</td>
<td>G2</td>
<td>4.1 x 10^{-4}</td>
</tr>
<tr>
<td></td>
<td>17</td>
<td>G2</td>
<td>9.3 x 10^{-33}</td>
<td>M/G1</td>
<td>2.7 x 10^{-2}</td>
</tr>
</tbody>
</table>

Probabilistic significance of the enrichment of the analysets, i.e., HO GSVD patterns of expression variation across the S. pombe, S. cerevisiae and human genes, that span the common HO GSVD subspace in each dataset, in over- or underexpressed cell cycle-regulated genes. The P-value of each enrichment is calculated as described [39] (Section 2.2 in Appendix S1) assuming hypergeometric distribution of the annotations (Datasets S1, S2, S3) among the genes, including the m = 100 genes most over- or underexpressed in each arraylet.

doi:10.1371/journal.pone.0028072.t001
Note also that simultaneous reconstruction in the common HO GSVD subspace removes the experimental artifacts and batch effects, which are dissimilar, from the three datasets. Consider, for example, the second genelet. With $\sigma_{1,2} : \sigma_{2,2} : \sigma_{3,2} \sim 1 : 8 : 3$ in the S. pombe, S. cerevisiae and human datasets, respectively, this genelet is almost exclusive to the S. cerevisiae dataset. This genelet is anticorrelated with a time decaying pattern of expression (Figure 2a). Consistently, the corresponding S. cerevisiae-specific arraylet is enriched in underexpressed S. cerevisiae genes that were classified as up-regulated by the S. cerevisiae synchronizing agent, the $z$-factor pheromone, with the $P$-value $<10^{-40}$. Reconstruction in the common subspace effectively removes this S. cerevisiae-

**Figure 3. Common HO GSVD subspace represents similar cell-cycle oscillations.** (a–c) S. pombe, S. cerevisiae and human array expression, projected from the five-dimensional common HO GSVD subspace onto the two-dimensional subspace that approximates it (Sections 2.3 and 2.4 in Appendix S1). The arrays are color-coded according to their previous cell-cycle classification [15–18]. The arrows describe the projections of the $k=13, \ldots, 17$ arraylets of each dataset. The dashed unit and half-unit circles outline 100% and 50% of added-up (rather than canceled-out) contributions of these five arraylets to the overall projected expression. (d–f) Expression of 380, 641 and 787 cell cycle-regulated genes of S. pombe, S. cerevisiae and human cell-cycle programs. The arrows describe the projections of the $k=13, \ldots, 17$ shared genelets and organism-specific arraylets that span the common HO GSVD subspace and represent cell-cycle checkpoints or transitions from one phase to the next. doi:10.1371/journal.pone.0028072.g003
approximately exclusive pattern of expression variation from the three datasets.

Simultaneous HO GSVD Classification of Homologous Genes of Different Cell-Cycle Peak Times

Notably, in the simultaneous sequence-independent classification of the genes of the three organisms in the common subspace, genes of significantly different cell-cycle peak times [19] but highly conserved sequences [20,21] are correctly classified (Section 2.5 in Appendix S1).

For example, consider the G2 S. pombe gene BFR1 (Figure 4a), which belongs to the evolutionarily highly conserved ATP-binding cassette (ABC) transporter superfamily [22]. The closest homologs of BFR1 in our S. pombe, S. cerevisiae and human datasets are the S. cerevisiae genes SNQ2, PDR5, PDR15 and PDR10 (Table S1a in Appendix S1). The expression of SNQ2 and PDR5 is known to peak at the S/G2 and G2/M cell-cycle phases, respectively [17]. However, sequence similarity does not imply similar cell-cycle peak times, and PDR15 and PDR10, the closest homologs of PDR5, are induced during stationary phase [23], which has been
hypothesized to occur in G1, before the Cdc28-defined cell-cycle arrest [24]. Consistently, we find PDR15 and PDR10 at the M/G1 to G1 transition, antipodal to (i.e., half a cell-cycle period apart from) SNQ2 and PDR5, which are projected onto S/G2 and G2/M, respectively (Figure 4f). We also find the transcription factor PDR1 at S/G2, its known cell-cycle peak time, adjacent to SNQ2 and PDR5, which it positively regulates and might be regulated by, and antipodal to PDR15, which it negatively regulates [25–28].

Another example is the S. cerevisiae phospholipase B-encoding gene PLB1 [29], which peaks at the cell-cycle phase M/G1 [30]. Its closest homolog in our S. cerevisiae dataset, PLB3, also peaks at M/G1 [17] (Figure 4e). However, among the closest S. pombe and human homologs of PLB1 (Table S1e in Appendix S1), we find the S. pombe genes SPA1777.00c and SPA1786.02, which expressions peak at the almost antipodal S. pombe cell-cycle phases S and G2, respectively [19] (Figure 4e). As a third example, consider the S. pombe G1 B-type cyclin-encoding gene CIG2 [31,32] (Table S1e in Appendix S1). Its closest S. pombe homolog, CDC13, peaks at M [19] (Figure 4e). The closest human homologs of CIG2, the cyclins CCNA2 and CCNB2, peak at G2 and G2/M, respectively (Figure 4g). However, while periodicity in mRNA abundance levels through the cell cycle is highly conserved among members of the cyclin family, the cell-cycle peak times are not necessarily conserved [1]: The closest homologs of CIG2 in our S. cerevisiae dataset, are the G2/M promoter-encoding genes CLB1.2 and CLB5.4, which expressions peak at G2/M and S respectively, and CLB5, which encodes a DNA synthesis promoter, and peaks at G1 (Figure 4f).

Discussion

We mathematically defined a higher-order GSVD (HO GSVD) for two or more large-scale matrices with different row dimensions and the same column dimension. We proved that our new HO GSVD extends to higher orders almost all of the mathematical properties of the GSVD: The eigenvalues of $S$ are always greater than or equal to one, and an eigenvalue of one corresponds to a right basis vector of equal significance in all matrices, and to a left basis vector in each matrix factorization that is orthogonal to all other left basis vectors in that factorization. We therefore mathematically defined, in analogy with the GSVD, the common HO GSVD subspace of the $N \geq 2$ matrices to be the subspace spanned by the right basis vectors that correspond to the eigenvalues of $S$ that equal one.

The only property that does not extend to higher orders in general is the complete column-wise orthogonality of the normalized left basis vectors in each factorization. Recent research showed that several higher-order generalizations are possible for a given matrix decomposition, each preserving some but not all of the properties of the matrix decomposition [12–14]. The HO GSVD has the interesting property of preserving the exactness and diagonality of the matrix GSVD and, in special cases, also partial or even complete column-wise orthogonality. That is, all $N$ matrix factorizations in Equation (1) are exact, all $N$ matrices $\Sigma_i$ are diagonal, and when one or more of the eigenvalues of $S$ equal one, the corresponding left basis vectors in each factorization are orthogonal to all other left basis vectors in that factorization.

The complete column-wise orthogonality of the matrix GSVD [5] enables its stable computation [6]. We showed that each of the right basis vectors that span the common HO GSVD subspace is a generalized singular vector of all pairwise GSVD factorizations of the matrices $D_i$ and $D_j$ with equal corresponding generalized singular values for all $i$ and $j$. Since the GSVD can be computed in a stable way, the common HO GSVD subspace can also be computed in a stable way by computing all pairwise GSVD factorizations of the matrices $D_i$ and $D_j$. That is, the common HO GSVD subspace exists also for $N$ matrices $D_i$ that are not all of full column rank. This also means that the common HO GSVD subspace can be formulated as a solution to an optimization problem, in analogy with existing variational formulations of the GSVD [33].

It would be ideal if our procedure reduced to the stable computation of the matrix GSVD when $N = 2$. To achieve this ideal, we would need to find a procedure that allows a computation of the HO GSVD, not just the common HO GSVD subspace, for $N$ matrices $D_i$ that are not all of full column rank. A formulation of the HO GSVD, not just the common HO GSVD subspace, as a solution to an optimization problem may lead to a stable numerical algorithm for computing the HO GSVD. Such a formulation may also lead to a higher-order general Gauss-Markov linear statistical model [34–36].

It was shown that the GSVD provides a mathematical framework for sequence-independent comparative modeling of DNA microarray data from two organisms, where the mathematical variables and operations represent experimental or biological reality [7,9]. The variables, subspaces of significant patterns that are common to both or exclusive to either one of the datasets, correlate with cellular programs that are conserved in both or unique to either one of the organisms, respectively. The operation of reconstruction in the subspaces common to both datasets outlines the biological similarity in the regulation of the cellular programs that are conserved across the species. Reconstruction in the common and exclusive subspaces of either dataset outlines the differential regulation of the conserved relative to the unique programs in the corresponding organism. Recent experimental results [9] verify a computationally predicted genome-wide mode of regulation [10,11], and demonstrate that GSVD modeling of DNA microarray data can be used to correctly predict previously unknown cellular mechanisms.

Here we showed, comparing global cell-cycle mRNA expression from the three disparate organisms S. pombe, S. cerevisiae and human, that the HO GSVD provides a sequence-independent comparative framework for two or more genomic datasets, where the variables and operations represent biological reality. The approximately common HO GSVD subspace represents the cell-cycle mRNA expression oscillations, which are similar among the datasets. Simultaneous reconstruction in the common subspace removes the experimental artifacts, which are dissimilar, from the datasets. In the simultaneous sequence-independent classification of the genes of the three organisms in this common subspace, genes of highly conserved sequences but significantly different cell-cycle peak times are correctly classified.

Additional possible applications of our HO GSVD in biotechnology include comparison of multiple genomic datasets, each corresponding to (i) the same experiment repeated multiple times using different experimental protocols, to separate the biological signal that is similar in all datasets from the dissimilar experimental artifacts; (ii) one of multiple types of genomic information, such as DNA copy number, DNA methylation and mRNA expression, collected from the same set of samples, e.g., tumor samples, to elucidate the molecular composition of the overall biological signal in these samples; (iii) one of multiple chromosomes of the same organism, to illustrate the relation, if any, between these chromosomes in terms of their, e.g., mRNA expression in a given set of samples; and (iv) one of multiple interacting organisms, e.g., in an ecosystem, to illuminate the exchange of biological information in these interactions.
Supporting Information

Appendix S1 A PDF format file, readable by Adobe Acrobat Reader.

Mathematica Notebook S1 Higher-order generalized singular value decomposition (HO GSVD) of global mRNA expression datasets from three different organisms. A Mathematica 5.2 code file, executable by Mathematica 5.2 and readable by Mathematica Player, freely available at http://www.wolfram.com/products/player/.

Mathematica Notebook S2 HO GSVD of global mRNA expression datasets from three different organisms. A PDF format file, readable by Adobe Acrobat Reader.

Dataset S1 S. pombe global mRNA expression. A tab-delimted text format file, readable by both Mathematica and Microsoft Excel, reproducing the relative mRNA expression levels of m3 = 3167 S. pombe gene clones at n = 17 time points during about two cell-cycle periods from Rustici et al. [15] or Oliva et al. [16].

Dataset S2 S. cerevisiae global mRNA expression. A tab-delimted text format file, readable by both Mathematica and Microsoft Excel, reproducing the relative mRNA expression levels of m2 = 4772 S. cerevisiae open reading frames (ORFs), or genes, at n = 17 time points during about two cell-cycle periods, including cell-cycle classifications, from Spellman et al. [17].

Dataset S3 Human global mRNA expression. A tab-delimted text format file, readable by both Mathematica and Microsoft Excel, reproducing the relative mRNA expression levels of m3 = 13,068 human genes at n = 17 time points during about two cell-cycle periods, including cell-cycle classifications, from Whitfield et al. [18].

Acknowledgments

We thank G. H. Golub for introducing us to matrix and tensor computations, and the American Institute of Mathematics in Palo Alto and Stanford University for hosting the 2004 Workshop on Tensor Decompositions and the 2006 Workshop on Algorithms for Modern Massive Data Sets, respectively, where some of this work was done. We also thank C. H. Lee for technical assistance, R. A. Horn for helpful discussions of matrix analysis and careful reading of the manuscript, and L. De Lathauwer and A. Gofkhe for helpful comments.

Author Contributions

Conceived and designed the experiments: OA. Performed the experiments: SPP OA. Analyzed the data: SPP OA. Contributed reagents/materials/analysis tools: SPP OA. Wrote the paper: SPP MAS CFVL OA. Proved mathematical theorems: SPP MAS CFVL OA.

References