Higher Order Generalized SVD based alignment-free method for inferring orthologous genes across species

Saket Choudhary
Computational Biology and Bioinformatics
Univeristy of California
Los Angeles, CA 900089
skchoudh@usc.edu

Abstract

The advent of next generation sequencing has made a plethora of biological data available. Comparative analysis of gene of gene expression datasets from multiple species can be used to enhance our fundamental understanding of biological mechanisms. Current methods rely on sequence information to infer conservation of functionality or *orthology*, but this information is *incomplete*. Gene expression datasets provide a sequence independent paradigm that can help separate the *conserved* from the *non-conserved*. However, the high-dimensionality of these datasets raises a need of an appropriate framework to narrow down our search space to genes that behave similarly across multiple species. We explore higher order generalized singular value decomposition (HOGSVD) to analyze large scale gene expression datasets tabulated as matrices with varying number of rows corresponding to the genes and fixed number of columns corresponding to the tissues across different species to identify genes with similar function.

1 Introduction

The availability of next generation sequencing datasets has made it possible to expand our fundamental understanding of biological mechanisms by looking for signatures that are shared among different species. For example, the liver in both human and mouse perform essentially the same molecular function. Next generation sequencing has made it possible to possible to profile such tissues across different species for their mRNA expression. Given the expression profiles of multiple tissues across different species, we want to ask the following this question: which genes in species A correspond to which genes in species B. The underlying motivation being, the expression values of different genes give the tissue its signature and at a molecular and even higher order level these tissues are essentially carrying out similar function. This is a loose definition of *orthologous* genes that we use without much context for the rest of the article.

2 Data

We obtained raw expression data from two public sequencing projects [Merkin et al., 2012, Brawand et al., 2011] and converted them to *gene-count* matrices.

Preprint. Work in progress.

	Tissue Gene	t_1	t_2	$\cdots$	t_m
X_1^{Human}	g_1				
	g_2				
	g_{n_1}				

	Tissue Gene	t_1	t_2	$\cdots$	t_m
X_2^{Mouse}	g'_1				
	g'_2				
	g'_3				
	g'_{n_2}				

Figure 1: mRNA expression data in this study can be represented as matrices, one for each species. The columns represent tissues while the rows represent genes. Each species has same number of tissues (columns), but different number of rows. Each species has its own set of genes. Our aim is to find those set of genes that behave similarly across species.

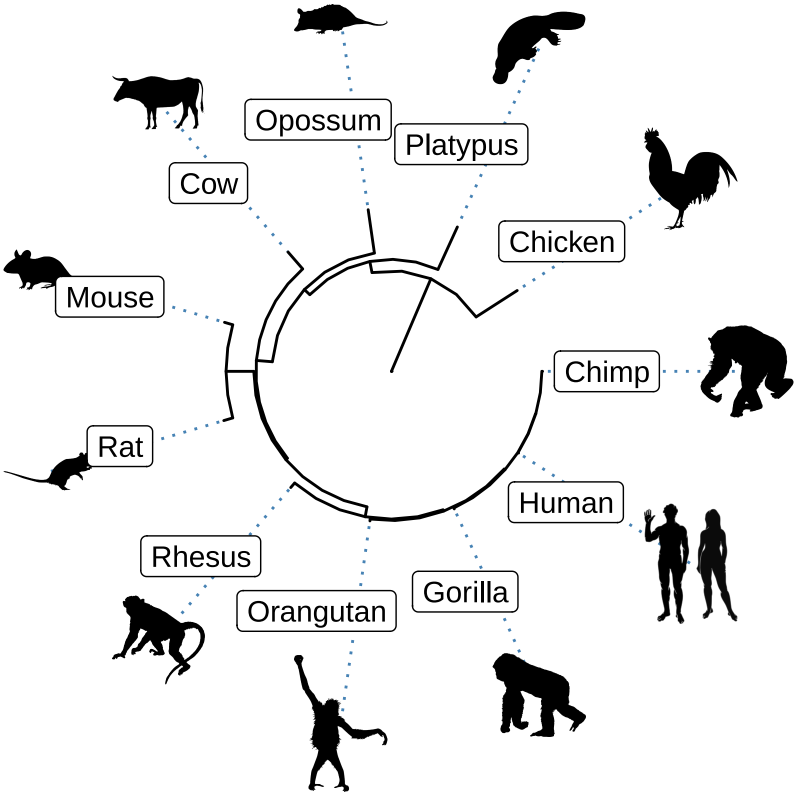


Figure 2: A phylogenetic tree with representative species used in our analysis. Longer branches indicate higher evolutionary distance. Species with smaller evolutionary distance are expected to share higher number of "orthologous" genes. Data obtained from [Brawand et al., 2011, Merkin et al., 2012]

3 Methods

The solution to our problem of interest can be thought of as a combination of dimensionality reduction and clustering problem. We want to first bring all the data matrices to a common subspace and then select the top features in this subspace to identify *orthologous* genes.

3.1 Singular Value Decomposition

The singular value decomposition of a matrix $\boldsymbol{X}^{(n \times m)}$ is

$$\boldsymbol{X} = \boldsymbol{U}\Sigma\boldsymbol{V}^{\mathrm{T}},\ \boldsymbol{X} : n \times m,\ \boldsymbol{U} : n \times n,\ \Sigma : n \times m,\ \boldsymbol{V} : m \times m.$$

where $\Sigma = \mathrm{diag}(\sigma_1, \sigma_2, \ldots,)$ such that

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k \geq \sigma_{k+1} = \cdots = \sigma_q = 0,\ q = \min(n, m)$$

and $\boldsymbol{U}, \boldsymbol{V}$ are unitary matrices, that is

$$\boldsymbol{U}^{\mathrm{T}}\boldsymbol{U} = \boldsymbol{I},\ \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V} = \boldsymbol{I}.$$

3.2 Generalized Singular Value Decomposition

Consider generalizing the SVD to two matrices, $\boldsymbol{X}_1^{(n_1 \times m)}$ and $\boldsymbol{X}_2^{(n_2 \times m)}$, with same number of columns m but different number of rows. The generalized singular value decomposition [Paige and Saunders, 1981] then finds the following decomposition:

$$\boldsymbol{X}_1 = \boldsymbol{U}_1\Sigma_1\boldsymbol{V}^{\mathrm{T}},\ \boldsymbol{X}_2 = \boldsymbol{U}_2\Sigma_2\boldsymbol{V}^{\mathrm{T}}.$$

such that,

$$\boldsymbol{U}_1 : n_1 \times m,\ \boldsymbol{U}_2 : n_2 \times m \text{ and } \Sigma_1, \Sigma_2, \boldsymbol{V} : m \times m.$$

where Σ_1, Σ_2 are both diagonal matrices such that,

$$\Sigma_1 = \mathrm{diag}(\sigma_{11}, \sigma_{12}, \ldots, \sigma_{1m}), \quad \text{and } \Sigma_2 = \mathrm{diag}(\sigma_{21}, \sigma_{22}, \ldots, \sigma_{2m}).$$

$\boldsymbol{U}_1$ and $\boldsymbol{U}_2$ are both unitary matrices. However, V is not necessarily orthonormal, though its rows all have unit norm,

$$\boldsymbol{U}_1^{\mathrm{T}}\boldsymbol{U}_1 = \boldsymbol{U}_2^{\mathrm{T}}\boldsymbol{U}_2 = \boldsymbol{I} \neq \boldsymbol{V}^{\mathrm{T}}\boldsymbol{V}$$

If $\boldsymbol{V} \equiv (\boldsymbol{v}_1, \boldsymbol{v}_2, \ldots \boldsymbol{v}_m)$, then $||\boldsymbol{v}_k|| = 1$, but not $\boldsymbol{v}_i\boldsymbol{v}_{j_{i \neq j}} \neq \boldsymbol{I}$

Also, the "respective singular values" of matrices $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$ are such that

$$\sigma_{1i}^2 + \sigma_{2i}^2 = 1\ \forall i \in [1, m]$$

The proof for 3.2 is given in [Paige and Saunders, 1981]. Intuitively, $\boldsymbol{V}$ provides a "shared" basis for representing both $\boldsymbol{X}_1$ and $\boldsymbol{X}_2$. The ratio of singular values σ_{1i}/σ_{2i} reflects the significance of the i^{th} basis vector $\boldsymbol{v}_i$ in both datasets. A value of the ratio close to one, $\sigma_{1i}/\sigma_{2i} = 1$, implies the two matrices equally share this basis.

3.3 Higher Order Generalized SVD (HOGSVD)

[Ponnapalli et al., 2011] extended GSVD for $N > 2$ matrices. Consider N matrices $\boldsymbol{X}_1, \boldsymbol{X}_2, \ldots \boldsymbol{X}_N$ such that they have same number of columns, but possibly different number of rows. Higher order GSVD performs the following decomposition:

$$
\begin{aligned}
\boldsymbol{X}_1^{(n_1 \times m)} &= \boldsymbol{U}_1 \Sigma_1 \boldsymbol{V}^T \\
\boldsymbol{X}_2^{(n_2 \times m)} &= \boldsymbol{U}_2 \Sigma_2 \boldsymbol{V}^T \\
&\vdots \\
\boldsymbol{X}_N^{(n_N \times m)} &= \boldsymbol{U}_N \Sigma_N \boldsymbol{V}^T
\end{aligned}
$$

As in the case of GSVD, $\boldsymbol{U}_i$ are orthonormal while $\boldsymbol{V}$ is not necessarily orthonormal. It's rows however have unit norm.

In order to solve for $\boldsymbol{V}$, we make use of the following relations:

$$
\begin{aligned}
\boldsymbol{A}_i &= \boldsymbol{X}_i^T \boldsymbol{X}_i, \\
S_{ij} &= \frac{1}{2}\left(A_i A_j^{-1} + A_j A_i^{-1}\right), \\
\boldsymbol{S} &\equiv \frac{1}{N(N-1)} \sum_{i=1}^{N} \sum_{j>i} \left(A_i A_j^{-1} + A_j A_i^{-1}\right) \\
\boldsymbol{S}\boldsymbol{V} &= \boldsymbol{V}\Lambda, \\
V &\equiv (v_1, \ldots v_m), \Lambda = \mathrm{diag}(\lambda_i).
\end{aligned}
$$

Thus, $\boldsymbol{V}$ can be obtained by eigen decomposition of matrix $\boldsymbol{S}$. Having obtained $\boldsymbol{S}$, we now solve for matrices $\boldsymbol{Z}_i$ to obtain $\boldsymbol{U}_i$:

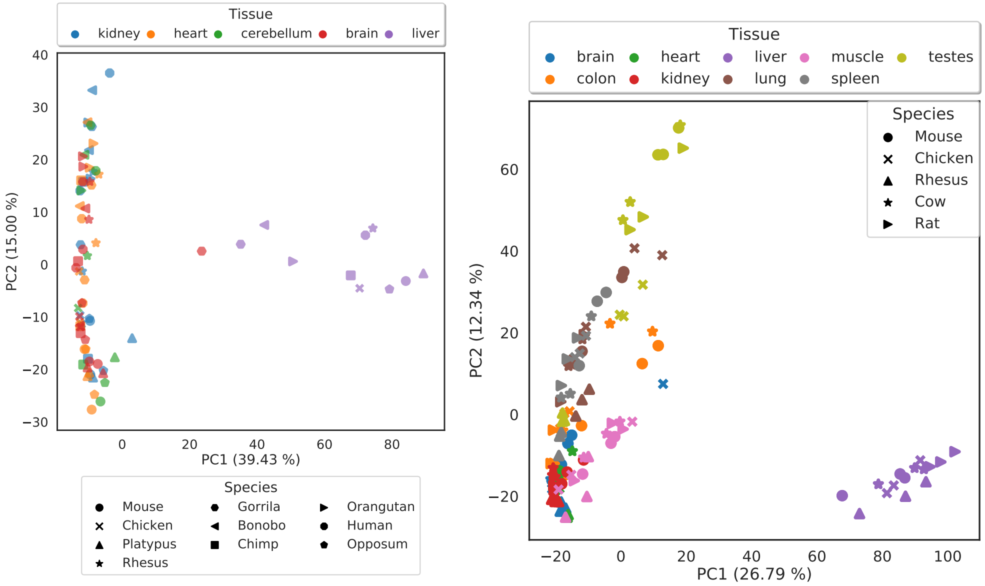
$$
\begin{aligned}
\boldsymbol{V}\boldsymbol{Z}_i &= \boldsymbol{A}_i^T \\
\boldsymbol{Z}_i &\equiv (z_{i1}, \ldots, z_{im}), i \in [1, N] \\
\sigma_{ik} &= ||z_{ik}||, \\
\Sigma_i &= \mathrm{diag}(\sigma_{ik}), \\
\boldsymbol{Z}_i &= \Sigma_i \boldsymbol{U}_i
\end{aligned}
$$

3.4 Identifying clusters

As described in the section on GSVD, σ_{ik} "measures" the significance of the k^{th} right eigenvector v_k and hence the ratio σ_{ik}/σ_{jk} measures the significance of v_k in X_i relative to its significance in X_j. For identifying orthologous genes we are interested in identifying all k^{th} eigengenes such that $\sigma_{ik}/\sigma_{jk} = 1$ for all $i, j \in [1, N]$. For each dataset $\boldsymbol{X}_i$, define projection matrices $\boldsymbol{P}_{\boldsymbol{X}_i}^{(n_1 \times m)} = \boldsymbol{X}_i^{(n_1 \times m)} V^{(m \times m)}$. If the eigengene v_k is such that $\sigma_{ik}/\sigma jk = 1 \ \forall i, j \in [1, N]$, extract genes with highest or lowest 10% projection values $\boldsymbol{P}_{\boldsymbol{X}}[:, k]$. To check if our shortlisted set of genes is statistically significanct, we use a hypergeometric test to asses the significance of overlap with the ground truth of orthologous genes.

3.5 Canonical Correlation Analysis

Consider two random variables $\boldsymbol{x} \in \mathcal{R}_x^D$ and $\boldsymbol{y} in \mathcal{R}_y^D$ such that we a set of n observations $S_x = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n\}$ and $S_y = \{\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n\}$ that are paired or coupled, *i.e.* for each $\boldsymbol{x}_i$, there is a corresponding set of observations $\boldsymbol{y}_i$. Canonical Correlation Analysis (CCA) aims at finding a projection direction $\boldsymbol{w}_x$ for $\boldsymbol{x}$ and $\boldsymbol{w}$ for $\boldsymbol{y}$ such that the correlation between the



(a) Sample SRP007412 [Brawand et al., 2011]

(b) Sample SRP016501 [Merkin et al., 2012]

Figure 3: PCA on expression matrices analyzed from two studies using a known list of orthologous genes. PCA fails to exhibit distinct clusters even with known ground truth.

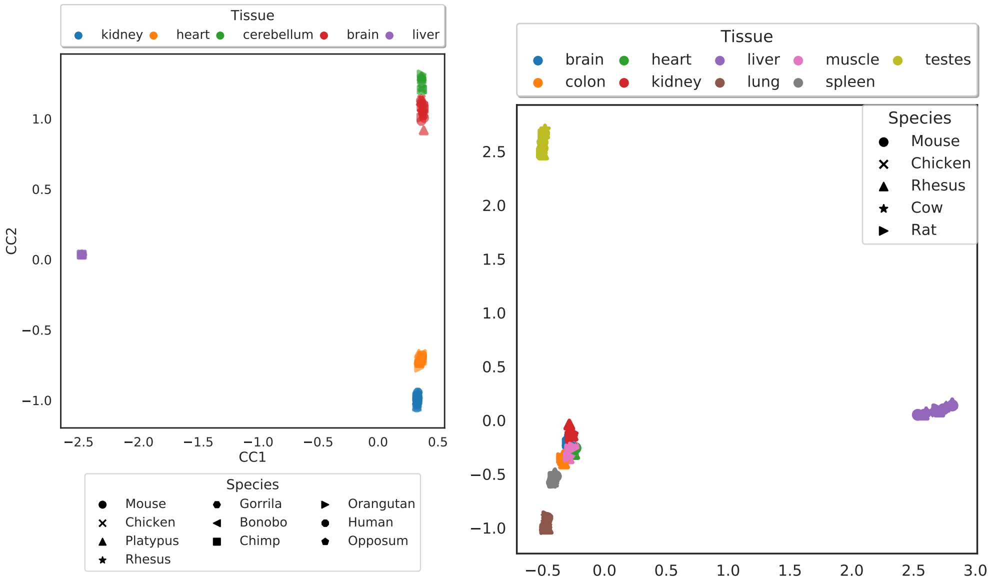
two vectors is maximizes. In other words, CCA solves the following maximization problem:

$$
\begin{aligned}
\rho &= \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \operatorname{corr}(S_x \boldsymbol{w}_x, S_y \boldsymbol{w}_y) \\
&= \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{\langle S_x \boldsymbol{w}_x, S_y \boldsymbol{w}_y \rangle}{||S_x \boldsymbol{w}_x|| ||S_y \boldsymbol{w}_y||} \\
&= \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{\mathbb{E}[\langle S_x \boldsymbol{w}_x, S_y \boldsymbol{w}_y \rangle]}{\sqrt{\mathbb{E}[\langle \boldsymbol{w}_x, x \rangle^2] \mathbb{E}[\langle \boldsymbol{w}_y, y \rangle^2]}} \\
&= \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{\boldsymbol{w}_x^{\mathrm{T}} \mathbb{E}[\boldsymbol{x}^{\mathrm{T}} \boldsymbol{y}] \boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x^{\mathrm{T}} \mathbb{E}[\boldsymbol{x}^{\mathrm{T}} \boldsymbol{x}] \boldsymbol{w}_x \boldsymbol{w}_y^{\mathrm{T}} \mathbb{E}[\boldsymbol{y}^{\mathrm{T}} \boldsymbol{y}] \boldsymbol{w}_y}} \\
&= \max_{\boldsymbol{w}_x, \boldsymbol{w}_y} \frac{\boldsymbol{w}_x^{\mathrm{T}} C_{xy} \boldsymbol{w}_y}{\sqrt{\boldsymbol{w}_x^{\mathrm{T}} C_{xx} \boldsymbol{w}_x \boldsymbol{w}_y^{\mathrm{T}} C_{yy} \boldsymbol{w}_y}}
\end{aligned}
$$

Here, we make use of CCA as an exploratory approach to visualize our clusters.

4 Results

We first demonstrate that if we know the list of "orthologous" genes across species and perform clustering using these features, we expect to see tight clusters. In our most basic analysis, we started off with Principle Component Analysis (PCA). However a näive PCA with a known set of orthologous genes did not reflect any clear clustering by tissue type among multiple organisms. (Figure 3). PCA represents a lower dimensional subspace (rotation) such that the average distance of this projection to its reconstruction is maximized. However, even if a clustering structure is present, it might not lie along the direction of maximum variation. In order to overcome this limitation of PCA, and given the context of expected correlation in our data, we used performed a CCA on the principal components. This has been demonstrated to be advantageous over just a näive CCA, since the PCA step acts as a denoising step [Soneson et al., 2010]. As seen in Figure 4, we see tight clusters organized by tissue type.



(a) Sample SRP007412 [Brawand et al., 2011]

(b) Sample SRP016501 [Merkin et al., 2012]

Figure 4: CCA on PCA projections of expression matrices in different species reveals a structure at tissue level with known ground truth of "orthologous" genes.

Species	# Candidate Orthologs	# Known Orthologs	# Fraction identified	p-value
Opposum	5732	6054	0.40	$< 1e-16$
Mouse	6444	6054	0.39	$< 1e-16$
Chicken	6218	6054	0.39	$< 1e-16$
Platypus	5955	6054	0.39	$< 1e-16$
Rhesus	5705	6054	0.39	$< 1e-16$
Gorrila	6227	6054	0.35	$< 1e-16$
Bonobo	6591	6054	0.33	$4.87e-10$
Chimp	6386	6054	0.39	$< 1e-16$
Orangutan	5921	6054	0.38	$< 1e-16$
Human	5611	6054	0.34	$< 1e-16$

Table 1: Identifying orthologous genes and their overlap with ground truth

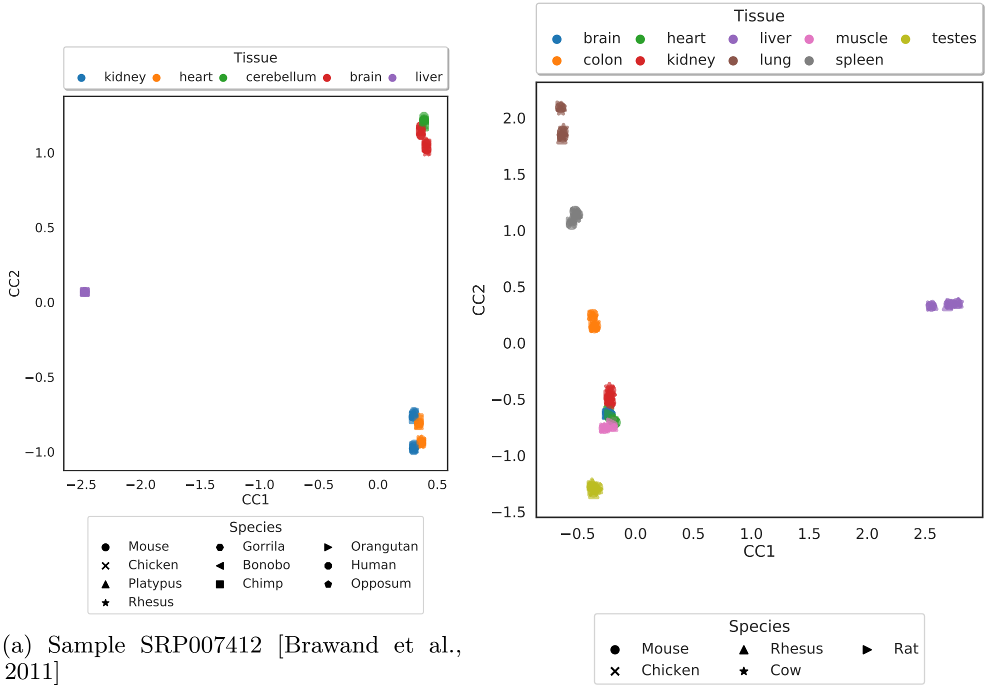
Using HOGSVD and our clustering approach, we achieve a tight clustering that resembles that of the ground truth. (Figure 5). The list of genes identified using the clustering approach shows a significant overlap with the ground truth of "orthologous" genes. (Table 1).

References

David Brawand, Magali Soumillon, Anamaria Necsulea, Philippe Julien, Gábor Csárdi, Patrick Harrigan, Manuela Weier, Angélica Liechti, Ayinuer Aximu-Petri, Martin Kircher, et al. The evolution of gene expression levels in mammalian organs. *Nature*, 478(7369): 343, 2011.

Jason Merkin, Caitlin Russell, Ping Chen, and Christopher B Burge. Evolutionary dynamics of gene and isoform regulation in mammalian tissues. *Science*, 338(6114):1593–1599, 2012.

Christopher C Paige and Michael A Saunders. Towards a generalized singular value decomposition. *SIAM Journal on Numerical Analysis*, 18(3):398–405, 1981.



(a) Sample SRP007412 [Brawand et al., 2011]

(b) Sample SRP016501 [Merkin et al., 2012]

Figure 5: CCA on PCA projections of genes shorted listed using our clustering algorithm after performing HOGSVD.

Sri Priya Ponnapalli, Michael A Saunders, Charles F Van Loan, and Orly Alter. A higher-order generalized singular value decomposition for comparison of global mrna expression from multiple organisms. *PloS one*, 6(12):e28072, 2011.

Charlotte Soneson, Henrik Lilljebjörn, Thoas Fioretos, and Magnus Fontes. Integrative analysis of gene expression and copy number alterations using canonical correlation analysis. *BMC bioinformatics*, 11(1):191, 2010.