Optimal separation of high dimensional transcriptome for complex multigenic traits

Aisharjya Sarkar
sarkar@cise.ufl.edu
University of Florida
Gainesville, FL, USA

Aaditya Singh
aaditya.kumar@iitkgp.ac.in
Indian Institute of Technology
Kharagpur, India

Richard Bailey
r.bailey@ufl.edu
University of Florida
Gainesville, FL, USA

Alin Dobra
adobra@cise.ufl.edu
University of Florida
Gainesville, FL, USA

Tamer Kahveci
tamer@cise.ufl.edu
University of Florida
Gainesville, FL, USA

ABSTRACT

The plight of navigating high-dimensional transcription datasets remains a persistent problem. This problem is further amplified for complex disorders, such as cancer, as these disorders are often multigenic traits with multiple subsets of genes collectively affecting the type, stage, and severity of the trait. We are often faced with a trade-off between reducing the dimensionality of our datasets and maintaining the integrity of our data. Almost exclusively, researchers apply techniques commonly known as dimensionality reduction to reduce the dimensions of the feature space to allow classifiers to work in more appropriately sized input spaces. As the number of dimensions is reduced, however, the ability to distinguish classes from one another reduces as well. Thus, to accomplish both tasks simultaneously for very high dimensional transcriptome for complex multigenic traits, we propose a new supervised technique, Class Separation Transformation (CST). CST accomplishes both tasks simultaneously by significantly reducing the dimensionality of the input space into a one-dimensional transformed space that provides optimal separation between the differing classes. We compare our method with existing state-of-the-art methods using both real and synthetic datasets, demonstrating that CST is the more accurate, robust, and scalable technique relative to existing methods. Code used in this paper is available on https://github.com/aisharjya/CST

KEYWORDS

class separation; data transformation; knowledge extraction; cancer omics

ACM Reference Format:
Aisharjya Sarkar, Aaditya Singh, Richard Bailey, Alin Dobra, and Tamer Kahveci. 2022. Optimal separation of high dimensional transcriptome for complex multigenic traits. In *13th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics (BCB '22), August 7–10, 2022, Northbrook, IL, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3535508.3545506

BCB '22, August 7–10, 2022, Northbrook, IL, USA

ACM ISBN 978-1-4503-9386-7/22/08…$15.00
https://doi.org/10.1145/3535508.3545506

1 INTRODUCTION

The art of extracting relevant information from extremely large datasets without compromising the class distinguishability in the data has proved itself to be a formidable task in bioinformatics. High throughput sequencing, seen in microarray gene expression profiling, often profiles thousands upon thousands of genes from numerous class types. This leads to immensely complex datasets in respect to both the feature space and class space. Complications in analysis of these complex datasets are further augmented through noise and perturbations in data measurements, making classification tasks even more challenging.

Studies involving gene expression analyses using microarray techniques almost always result in datasets large and complex enough to necessitate the application of dimensionality reduction (DR) techniques to allow for more efficient, but not necessarily as accurate, application of classifiers [17]. The primary goal of DR - preserving the information contained in the high dimensional data while reducing dimensionality - does not usually benefit class distinction ability [18], as different subsets of dimensions may play a role in encoding different characteristics of the data.

The complex nature of biological data amplifies this problem as each sample may simultaneously belong to many alternative classes according to different attributes, where differing subsets of genes play a role in determining the classes of samples for each attribute - but these corresponding gene subset-attribute correlations are often unknown a priori. DR techniques are all but necessary in the field of bioinformatics [12] due to the complexity and size of the data, but the risk of losing important information with respect to classification ability remains persistent [3]. Thus, the selection of DR techniques in classification tasks must be made with caution due to the nature of the datasets pertaining to each problem [14].

Relevant literature and shortcomings. Among the most well-known DR methods are mutual information-based feature selection (MI), principal component analysis (PCA), kernel principal component analysis (KPCA), and uniform manifold approximation (UMAP).

Although MI, PCA, KPCA, and UMAP are popular and widely used in analyzing gene expression microarray data [2, 4, 28], they can be misleading in many instances, leading to erroneous interpretations. First, these methods are sensitive to noise in experimental measurements as well as sampling bias [10]. Second, PCA and KPCA are both unsupervised techniques, meaning they do not consider the class label information for the samples, thus failing to capture interdependence between sampling groups and lose an element of class separability. Additionally, KPCA can further complicate the data in utilizing a kernel function by transforming a dataset into an even higher dimension than the input space. Furthermore, MI is supervised but considers independent features one at a time, thereby eliminating correlated features – genes in our context – as redundant, which is disadvantageous with cancer because groups of genes often work together in particular pathways for cancer initiation and progression. Finally, UMAP is supervised but distorts single cell gene expression trajectories due to the nature of utilizing topologies in a high dimension to transform into a low dimension [29], which thus does not preserve density and can lead to misleading data transformations.

MI, PCA, KPCA, and UMAP fail to satisfy the dual objective of accomplishing the adequate reduction of dimensions while preserving the separability of classes in the reduced feature space.

Our contributions. In this paper, we develop a novel supervised method for classification tasks, Class Separation Transformation (CST), that transforms a given dataset that contains class label information in a way such that the transformed dataset provides provably most optimal separation for classification. Briefly, given a dataset with m samples, n features, and class label information, CST transforms the n-dimensional feature space into a scalar value for each sample such that the transformed values provide maximum possible separation of the samples belonging to different classes.

Our experimental results demonstrate that CST is accurate, robust, and scalable. CST accomplishes the dual objective of transforming the data into a more ideal dimensional space for classifiers while also extracting a representation of the data that provides the most optimal separation for classification. Compared to MI, PCA, KPCA, and UMAP, application of CST on lung cancer, breast cancer, and colon cancer datasets always results in the highest accuracy on average for all six classifiers. Moreover, our experiments on these datasets suggest that our method can also be used to discover key biologically significant gene markers based on samples from different classes. [1]

2 METHODS

In this section, we describe our supervised method, *Class Separation Transformation* (CST), that transforms a given high-dimensional dataset into one-dimensional data. The goal is to find the most concise representation of the dataset such that the samples belonging to different classes are separated as much as possible. In the rest of this paper, we denote matrices in upper case bold, vectors in lower case bold, and elements in lower case. Matrices, vectors, and elements from the same matrix all use the same letter (e.g., $\mathbf{X}$, $\mathbf{x}$, x). We denote the transpose operation by the superscript $^\mathsf{T}$ and the identity matrix as $\mathbf{I}$. We denote all vectors as column vectors. In the following, first we formally define our problem statement as *Class Separation Analysis* (CSA). We then present our algorithm for transformation of feature space based on class information.

Let us consider an n-dimensional vector representing a sample as $\mathbf{x} \in \mathbb{R}^n$. Consider a dataset $\mathbf{X}$ consisting of m such samples, where, each sample is represented by $\mathbf{x}_i$, $(1 \le i \le m)$. Mathematically, $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_m] \in \mathbb{R}^{n \times m}$. Let us denote the class information for the samples belonging to τ classes with $\mathbf{x}_i \in \mathbf{X}$ by a vector $\mathbf{c} = [c_1, c_2, \ldots, c_m]$, where $c_i \in \{1, 2, \ldots, \tau\}$ denotes the class label of $\mathbf{x}_i$. Our goal is to transform each sample $\mathbf{x}_i \in \mathbf{X}$ into a scalar $y_i \in \mathbf{y}$, such that $\mathbf{y}$ maximizes the separation of the samples belonging to different classes.

Without loss of generality, let us consider an n-dimensional unit vector $\mathbf{f}$. That is, $\mathbf{f} \in \mathbb{R}^{n \times 1}$, where $||\mathbf{f}|| = 1$. We transform an n-dimensional vector $\mathbf{x}$ using $\mathbf{f}$ into a value y as $y = \mathbf{f}^\mathsf{T}\mathbf{x}$ Thus, we have

$$\mathbf{y} = \mathbf{f}^\mathsf{T}\mathbf{X} \tag{1}$$

Consider two n-dimensional vectors $\mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}$ having class labels c_i and c_j, respectively, where $i \neq j$. We define an indicator function σ on a pair of vectors $\mathbf{x}_i$ and $\mathbf{x}_j$, that returns a value based on whether c_i and c_j are of the same class label or not. That is,

$$\sigma_{ij} = \begin{cases} +1, & \text{if } c_i = c_j \\ -1, & \text{otherwise} \end{cases}$$

Let us denote the pairwise distance between $\mathbf{x}_i$ and $\mathbf{x}_j$ transformed into y_i and y_j as $d(y_i, y_j)$. Consider the indices of all m samples as an ordered set $\mathcal{S} = \{1, 2, \ldots, m\}$. We compute the *class separation function* as the sum of the distance between all possible pairs of transformed values y_i and y_j in the same class minus those in different classes as

$$\mathcal{F} = \sum_{i<j,(i,j)\in\mathcal{S}\times\mathcal{S}} d(y_i, y_j)\sigma_{ij} \tag{2}$$

Notice that small values of this function promote to increase the gap between samples belonging to different classes and discourage that for those in the same class. Our goal is to identify the transformation vector $\mathbf{f}$ which minimize the function in Equation 2.

Definition 2.1. *Class Separation Analysis (CSA).* Given a dataset $\mathbf{X}$ consisting of m samples and classification vector $\mathbf{c}$ that denotes the class label of each sample, CSA seeks to transform $\mathbf{X}$ into a vector $\mathbf{y}$ such that the class separation function $\mathcal{F}$ in Equation 2 is minimized.

We compute d in Equation 2 as $d(y_i, y_j) = \left(\mathbf{f}^\mathsf{T}(\mathbf{x}_i - \mathbf{x}_j)\right)^2$ from Equation 1. Therefore, our objective function in Equation 2 becomes the following.

$$\begin{aligned} \mathcal{F}(\mathbf{f}) &= \sum_{i<j,(i,j)\in\mathcal{S}\times\mathcal{S}} \left(\mathbf{f}^\mathsf{T}(\mathbf{x}_i - \mathbf{x}_j)\right)^2 \sigma_{ij} \\ &= \mathbf{f}^\mathsf{T}\left(\sum_{i<j,(i,j)\in\mathcal{S}\times\mathcal{S}} \sigma_{ij}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\mathsf{T}\right)\mathbf{f} \\ &= \mathbf{f}^\mathsf{T}\mathbf{A}\mathbf{f} \end{aligned} \tag{3}$$

where,

$$\mathbf{A} = \sum_{i<j,(i,j)\in\mathcal{S}\times\mathcal{S}} \sigma_{ij}(\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\mathsf{T}.$$

[1] This paper is partially funded by NSF under award 2111679

Notice that $\mathbf{A}$ is a symmetric matrix. Minimizing $\mathcal{F}$ with respect to $\mathbf{f}$ without constraining the values of $\mathbf{f}$ produces $\mathbf{0}$ (i.e., n-dimensional vector with all values equal to zero) as an optimal solution regardless of the input data, which is clearly undesirable. To ensure that the solution is non-trivial, we constrain $\mathbf{f}$ to have unit norm. This leads to the following formulation of our optimization problem:

Problem 1. *Given a real symmetric matrix* $\mathbf{A}$*, minimize* $\mathcal{F}$ *such that*

$$\mathcal{F}(\mathbf{f}) = \underset{\mathbf{f}}{\operatorname{argmin}} \left\{ \mathbf{f}^\top \mathbf{A} \mathbf{f} \right\}$$

under the constraint $\mathbf{f}^\top \mathbf{f} = 1.$

Theorem 2.1. *The vector* $\mathbf{f}$ *that solves Problem 1 is the eigenvector of the symmetric matrix*

$$\mathbf{A} = \sum_{i<j}^{\mathcal{S}\times\mathcal{S}} (\mathbf{x}_i - \mathbf{x}_j)(\mathbf{x}_i - \mathbf{x}_j)^\top \sigma_{ij}$$

corresponding to the smallest eigenvalue.

Following from this theorem, we first compute the eigenvectors and eigenvalues of $\mathbf{A}$ to determine the principal directions and their significance, respectively. We compute eigenvalues and eigenvectors as solutions to the continuous optimization problem. Among all the eigenvectors, the one with the smallest eigenvalue gives the minimum value for the function $\mathcal{F}$ and thus solves Problem 1.

3 RESULTS AND DISCUSSION

In this section, we evaluate the performance of CST against existing methods using six classifiers, utilizing three datasets. The methods we will compare against come from two categories, that being linear (MI and PCA) and non-linear (KPCA and UMAP) dimensionality reduction algorithms. MI and UMAP are supervised techniques and PCA and KPCA are unsupervised techniques.

To evaluate the performance of MI, PCA, KPCA, UMAP, and CST in their class separability abilities, we use both supervised and unsupervised classifiers. These classifiers are Support Vector Machine (SVM), Gaussian Naïve Bayes (GNB), k-Nearest Neighbor (KNN), Random Forest (RF), Stochastic Gradient Descent (SGD), and Linear Discriminant Analysis (LDA). Evaluating performances using this wide range of classifiers ensures a holistic comparison between CST and its competing methods.

Cancer datasets. We use three cancer microarray datasets with varying characteristics in our experiments, namely, (i) lung cancer (GSE19804) [19, 20], (ii) breast cancer (GSE27562) [13] and (iii) colon cancer (GSE39582) [21] from the GEO database. Following is a brief summary of these datasets.

1. Lung cancer dataset. This is a study of transcriptional modulation in non-smoking female lung cancer in Taiwan. The dataset consists of 120 samples (60 paired samples from tumor and normal class tissues) with 54,675 probe sets from the Affymetrix chip. We obtain the dataset from NCBI (http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE19804).

2. Breast cancer dataset. This is a study on gene expression analysis of human peripheral blood mononuclear cells. The dataset consists of 162 samples with 54,675 probe sets from the Affymetrix chip. The samples contain 57 women with breast cancer diagnosis, 31 women with normal initial mammogram, 37 women with benign diagnosis, 15 breast cancer patients following surgery, 15 patients with gastrointestinal cancer, and 7 patients with brain tumor. In our final dataset, we consider 88 samples from two classes, namely, class 1 = normal (31 samples) and class 2 = malignant (57 samples). we obtain the dataset from NCBI (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE27562).

3. Colon cancer dataset. This study provides molecular classification of mRNA expression profiles of colon cancer (different subtypes) and control samples. The dataset consists of 566 samples in six classes with 54,675 probe sets from the Affymetrix chip. More specifically, it contains six colon cancer subtypes: class 1 = $\text{CIN}_{\text{Immune-Down}}$ (116 samples), class 2 = dMMR (104 samples), class 3 = KRASm (75 samples), class 4 = CSC (59 samples), class 5 = $\text{CIN}_{\text{WntUp}}$ (152 samples), and class 6 = $\text{CIN}_{\text{normL}}$ (60 samples). We use 176 samples belonging to two classes, namely, class 1 = $\text{CIN}_{\text{Immune-Down}}$ (116 samples) and class 6 = $\text{CIN}_{\text{normL}}$ (60 samples). we obtain the dataset from NCBI (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE39582)

Pre-processing of the real dataset. We pre-processed each of the above cancer datasets in two steps as follows. First, for all the probes that map to a particular gene, we pick the probe with the highest differential expression. Second, we remove the genes that do not show significant variation in expression across all samples. To do that, for each gene, we first compute the mean across all samples for each of the two classes, say, c1_{avg} and c2_{avg}, respectively. Next, we calculate the relative difference for each gene as $\frac{|\text{c1}_{\text{avg}} - \text{c2}_{\text{avg}}|}{\min(\text{c1}_{\text{avg}}, \text{c2}_{\text{avg}})}$. We filter out the genes with a relative difference less than the average relative difference across all genes. After this pre-processing, the number of genes in lung, breast, and colon cancer datasets are 7092, 9161, and 7377, respectively.

3.1 Results on real dataset

In this section, we present our results on the three cancer datasets described above. We split each dataset into training and testing samples for five split ratios {0.1, 0.15, 0.2, 0.25, 0.3}, each indicating the proportion of test samples in the dataset. For each split ratio, we perform 10-fold cross validation and report the average BAC value.

Lung cancer. The results in figure 1 demonstrate that all the five methods have very high BAC values. However, our method yields the highest accuracy on average across all split ratios for each classifier. Our method's average performance across different split ratio sizes implies that as the size of the training dataset grows or shrinks, our method identifies the class separation more consistently than any other method can in either direction of the changing training dataset size, as well as relative to the chosen classifier.

Breast cancer. Seen in figure 2, our method's performance relative to competing methods with the breast cancer dataset is even better than seen in the lung cancer dataset, but follows a similar pattern seen in the lung cancer results, where CST yields the highest average BAC across all split rations and classifiers.

Colon cancer. Figure 3 presents the average BAC values across all split ratios for this dataset, which contains results consistent with the lung and breast cancer datasets. CST again has the highest average BAC across all split ratios for each classifier, but CST

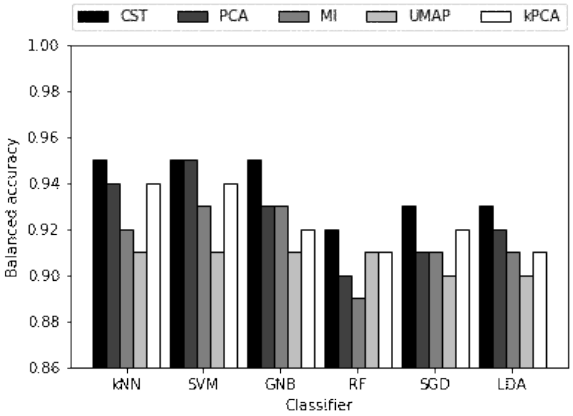

Figure 1: Lung cancer dataset average BAC across five split ratios $\{0.1, 0.15, 0.2, 0.25, 0.3\}$, each 10-fold cross validated, measured for each combination of a method and classifier.

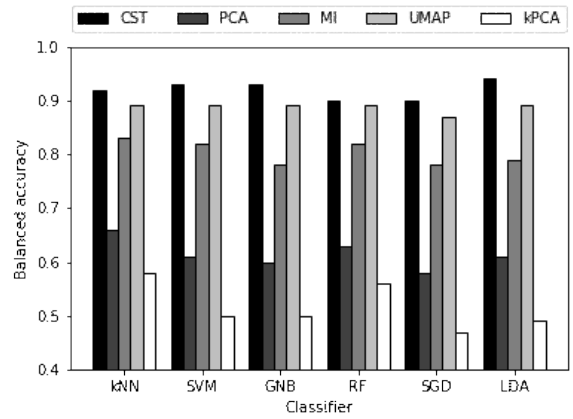

Figure 2: Breast cancer dataset average BAC across five split ratios $\{0.1, 0.15, 0.2, 0.25, 0.3\}$, each 10-fold cross validated, measured for each combination of a method and classifier.

doesn't outperform the other methods as much as seen in the lung and breast cancer datasets.

In summary, when compared to competing methods, CST improves both the accuracy of the classifiers and significantly reduces the cost of classification since it feeds fewer dimensions (only 1 dimension) to the classifiers. The competing methods each have advantages and disadvantages - in how they can reduce the dimensionality of data to aid in classification - that are significantly influenced by the shape and properties of the data. Under certain circumstances, whether it be class data existing along different manifolds or certain attributes having significantly overlapping statistical dependencies, the competing methods have the capacity to perform well. However, their performances are inconsistent with respect to the dataset, since their performances rely on certain characteristics being met in the data, as well as the chosen classifier. Our method proves to be superior because it consistently results in the highest average BAC results no matter the dataset nor the

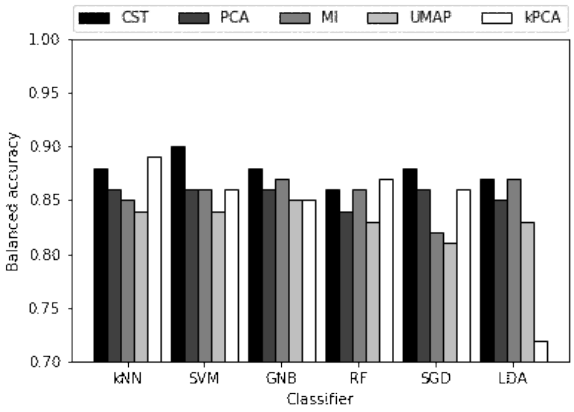

Figure 3: Colon cancer dataset average BAC across five split ratios $\{0.1, 0.15, 0.2, 0.25, 0.3\}$, each 10-fold cross validated, measured for each combination of a method and classifier.

classifier. There is no clear second place winner among the four competing method. In the case of breast cancer data, UMAP outperforms the remaining methods, while for lung cancer data and colon cancer, PCA outperforms the remaining methods. When we take the average BAC of CST for all the three cancer datasets and all split ratios, we observe that the classifiers that perform the best with CST are SVM with the highest BAC (0.93), followed by GNB (0.92) and kKNN (0.91).

3.2 Biological significance.

To investigate the biological implications of CST, we analyze the extent to which our method identifies genes that overlap with genes already known and associated with particular cancers. We will perform this analysis with the lung cancer dataset already used in this paper. We use the elements within the transformation vector $\mathbf{f}$ as the coefficient weights associated with the corresponding genes. In this manner, we consider the top 50 genes (weights) that contribute most to the objective function $\mathcal{F}$ we wish to minimize. It is important to note that groups of proteins that perform similar functions are often connected to each other in protein-protein interaction (PPI) networks [23]. Cancer initiation, progression, and severity propagate through these PPI network pathways [7]. Therefore, we input the selected genes from our transformation vector $\mathbf{f}$ to generate the corresponding PPI network using the STRING database [27].

In performing analysis using the STRING database, the number of edges and average node degree are 148 and 6.3, respectively, and the PPI enrichment p-value is $< 1.0e{-}16$. This signifies that the proteins have more interactions among themselves compared to a random set of proteins of similar size drawn from the genome and are therefore biologically connected as a group. Next, we consider the functional enrichment analysis from the STRING database that provides the list of publications significantly enriched in the number of genes we identify. Table 1 shows six publications related to lung cancer, the observed gene count in those publications among our selection of genes, and the corresponding false discovery rate (FDR).

PMID	Citation	OGC	FDR
31681566	[1]	13	5.43E-13
30556321	[15]	6	4.61E-06
31106044	[16]	6	1.23E-06
28410204	[31]	6	4.64E-06
15217521	[9]	5	0.00078
26124566	[8]	5	0.0042

Table 1: PubMed ID, observed gene count (OGC), false discovery rate (FDR) of lung cancer related publications extracted from STRING functional enrichment analysis.

Further inspection of the genes in the transformation vector reveals that there exist several studies in literature that provide evidence that these genes play a role in lung cancer. For example, many studies show that IL6 acts as a driver gene that is as a powerful pro-inflammatory cytokine essential for inflammatory acute phase response induced by tissue damage, thereby playing a pivotal role in lung cancer [6, 11, 22, 24–26]. Studies also show that dysregulation of lowly expressed AGER significantly reduces the survival time of patients with lung cancer [5, 30]. Our method is able to identify AGER and IL6 as key lung cancer genes through cheap calculations on transcriptional data, as opposed to the more costly techniques the aforementioned studies used. This further demonstrates CST's value and ability to identify biologically significant genes that play an important role in the initiation, progression, and/or severity of cancer.

4 CONCLUSION

Expressing high-dimensional transcriptome to study complex multigenic traits with the smallest number of features while maintaining a particular knowledge representation within the data is of utmost importance. Here, we developed a novel method, Class separation transformation. CST optimally solves this dual objective by reducing the dimensionality of the input space to a single dimension while concurrently providing the most meaningful representation of the data in that one dimension with respect to optimal class distinction. On the lung, breast, and colon datasets, we demonstrated that CST had the best average BAC performance across all split ratios and classifiers when compared to the competing methods of PCA, MI, UMAP, and KPCA. From our experimental results, we have shown that CST is the most accurate, robust, and scalable technique relative to its competing methods.

REFERENCES

[1] Firoz Ahmed. Integrated network analysis reveals foxm1 and mybl2 as key regulators of cell proliferation in non-small cell lung cancer. *Frontiers in Oncology*, 9:1011, 2019.

[2] Becht, E., McInnes, L., and J. et al Healy. Dimensionality reduction for visualizing single-cell data using umap. *Nature Biotechnology*, 37(1):38–44, 2019.

[3] H. B. Borges and J. C. Nievola. Comparing the dimensionality reduction methods in gene expression databases. *Expert Syst. Appl*, 39(12):10780–95, 2012.

[4] Alex et al. Diaz-Papkovich. A review of umap in population genetics. *Journal of human genetics*, 11(1):85–91, 2021.

[5] Junjie Fu and et al. Discovery of gene regulation pattern in lung cancer by gene expression profiling using human tissues. *Genomics data*, 3:112–115, 2015.

[6] José J Fuster and et al. The good, the bad, and the ugly of interleukin-6 signaling. *The EMBO journal*, 33(13):1425–1427, 2014.

[7] Isaac S Harris and et al. Glutathione and thioredoxin antioxidant pathways synergize to drive cancer initiation and progression. *Cancer cell*, 27(2):211–222, 2015.

[8] Atif Noorul Hasan, Mohammad Wakil Ahmad, Inamul Hasan Madar, B Leena Grace, and Tarique Noorul Hasan. An in silico analytical study of lung cancer and smokers datasets from gene expression omnibus (GEO) for prediction of differentially expressed genes. *BMC Bioinformatics*, 11(5)(229-35), 2015.

[9] Hongying Jiang, Youping Deng, Huann-Sheng Chen, Lin Tao, Qiuying Sha, Jun Chen, Chung-Jui Tsai, and Shuanglin Zhang. Joint analysis of two microarray gene-expression data sets to select lung adenocarcinoma marker genes. *BMC Bioinformatics*, 5(81), 2004.

[10] Tomokazu Konishi. Principal component analysis for designed experiments. *BMC bioinformatics*, 16(S18):S7, 2015.

[11] Manfred Kopf and et al. Impaired immune and acute-phase responses in interleukin-6-deficient mice. *Nature*, 368(6469):339–342, 1994.

[12] N. Fadhel L. B. Romdhane and B. Ayeb. An efficient approach for building customer profiles from business data. *Expert Syst. Appl*, 37(2):1573–1585, 2010.

[13] Heather G LaBreche and et al. Integrating factor analysis and a transgenic mouse model to reveal a peripheral blood predictor of breast tumors. *BMC medical genomics*, 4(1):61, 2011.

[14] S. Liang, A. J. Ma, S. Yang, Y. Wang, and Q. Ma. A review of matched-pairs feature selection methods for gene expression data analysis. *Comput. Struct. Biotechnol.*, 16(1):88–97, 2018.

[15] Wei Liu, Songyun Ouyang, Zhigang Zhou, Meng Wang, Tingting Wang, Yu Qi, Chunling Zhao, Kuisheng Chen, and Liping Dai. Identification of genes associated with cancer progression and prognosis in lung adenocarcinoma: Analyses based on microarray from oncomine and the cancer genome atlas databases. *Molecular Genetics & Genomic Medicine*, 7(2):e00528, 2019.

[16] Yu Liu, Deyao Xie, Zhifeng He, and Liangcheng Zheng. Integrated analysis reveals five potential cerna biomarkers in human lung adenocarcinoma. *PeerJ*, 7:e6694, 2019.

[17] Q. Lu and X. Qiao. Sparse fisher's linear discriminant analysis for partially labeled data. *Stat. Anal. Data Mining*, 11(1):17–31, 2018.

[18] Qiyi Lu and Xingye Qiao. Sparse fisher's linear discriminant analysis for partially labeled data. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 11(1):17–31, 2018.

[19] Tzu-Pin Lu and et al. Identification of a novel biomarker, sema5a, for non–small cell lung carcinoma in nonsmoking women. *Cancer Epidemiology and Prevention Biomarkers*, 19(10):2590–2597, 2010.

[20] Tzu-Pin Lu and et al. Identification of regulatory snps associated with genetic modifications in lung adenocarcinoma. *BMC research notes*, 8(1):1–11, 2015.

[21] Laetitia Marisa and et al. Gene expression classification of colon cancer into molecular subtypes: characterization, validation, and prognostic value. *PLoS Med*, 10(5):e1001453, 2013.

[22] Hiroyuki Ogawa and et al. Interleukin-6 blockade attenuates lung cancer tissue construction integrated by cancer stem cells. *Scientific reports*, 7(1):1–13, 2017.

[23] Clara Pizzuti and et al. Algorithms and tools for protein–protein interaction networks clustering, with a special focus on population-based stochastic methods. *Bioinformatics*, 30(10):1343–1352, 2014.

[24] Zhaoxia Qu and et al. Interleukin-6 prevents the initiation but enhances the progression of lung cancer. *Cancer research*, 75(16):3209–3215, 2015.

[25] Shawn J Rice and et al. Advanced nsclc patients with high il-6 levels have altered peripheral t cell population and signaling. *Lung Cancer*, 131:58–61, 2019.

[26] Estela Maria Silva and et al. High systemic il-6 is associated with worse prognosis in patients with non-small cell lung cancer. *PloS one*, 12(7):e0181125, 2017.

[27] Damian Szklarczyk and et al. The string database in 2017: quality-controlled protein–protein association networks, made broadly accessible. *Nucleic acids research*, p. gkw937, 2016.

[28] F.W. Townes, Hicks, S.C., and M.J. et al. Aryee. Feature selection and dimension reduction for single-cell rna-seq based on a multinomial model. *Genome Biology*, 20(1):295, 2019.

[29] Yan et al. Wu. Visualizing and interpreting single-cell gene expression datasets with similarity weighted nonnegative embedding. *Cell Systems*, 7(6):656–66, 2018.

[30] Weiguo Zhang and et al. Spp1 and ager as potential prognostic biomarkers for lung adenocarcinoma. *Oncology letters*, 15(5):7028–7036, 2018.

[31] Dan Zhou, Weiwei Tang, Xinli Liu, Han-Xiang An, and Yun Zhang. Clinical verification of plasma messenger RNA as novel noninvasive biomarker identified through bioinformatics analysis for lung cancer. *Oncotarget*, 8(27):43978–43989, 2017.