Multi-Classification of Cancer Samples Based on Co-Expression Analyses

1st Hongyang Jiang, 2nd Qiang Huang
Key Laboratory of Symbol Computation and Knowledge Engineering of Ministry of Education, College of Computer Science and Technology, Jilin University
Changchun, China
jianghy18@mails.jlu.edu.cn,
huangqiang18@mails.jlu.edu.cn

3rd Liang Chen
Department of Computer Science, College of Engineering, Key Laboratory of Intelligent ManufacturingTechnology of Ministry of Education, Shantou University
Shantou, China
chenliang@stu.edu.cn

4th Zhi Li
Department of Medical Oncology, Key Laboratory of Anticancer Drugs and Biotherapy of Liaoning Province, the First Hospital of China Medical University
Shenyang, China
zli@cmu.edu.cn

5th Ying Xu
Computational Systems Biology Lab,Department of Biochemistry and Molecular Biology and Institute of Bioinformatics, University of Georgia, Athens, USA Cancer Systems Biology Center, The China-Japan Union Hospital, Jilin University
Changchun, China
xyn@uga.edu

6th Huiyan Sun*
School of Artificial Intelligence, Cancer Systems Biology Center, The China-Japan Union Hospital, Jilin University
Changchun, China
huiyansun@jlu.edu.cn

7th Yi Chang
School of Artificial Intelligence, Jilin University
Changchun, China
yichang@jlu.edu.cn

***Abstract*—Cancer staging, grading and subtyping all represent important problems for precision diagnosis, treatment and mechanistic studies of cancer. The majority of the existing computational methods solve this problem via multi-classification of differential gene-expressions of cancer samples of specific classes (Stages, Grades and subtypes) vs. controls. However, the performance of such classification techniques is generally not satisfactory since the discerning power of differential expression patterns in such classifications is limited. We present here a multi-classification technique, based on co-expression patterns specific to individual subclasses in provided training data as co-expression patterns tend to be more conserved than differential expressions within each subclass. A challenge in implementing this strategy lies in how to effectively derive co-expression patterns in individual samples, which is solved through comparing co-expression patterns within a subclass and those in the subclass plus a new sample. Compared with the state-of-the-art gene expression-based classification methods, our method outperforms them in cancer staging, grading and subtyping of cancer samples from TCGA in almost all the measures used. In addition, the co-expressed genes computationally selected for classifications are biologically meaningful, which will prove important for diagnostic biomarker design, treatment plan selection and possibly mechanistic studies of cancer.**

***Index Terms*—cancer staging, grading and subtyping, transcriptomic data analyses, classification technique, specific co-expression**

I. Introduction

Cancer is a dynamic and complex disease. Different patients of the same cancer type may require distinct treatment strategies depending on the levels of the development, malignancy and other molecular characteristics of the disease [1], [2]. Cancer Stage has been used to measure the level of the disease development. For most of the cancer types, cancer tumors are classified into four Stages with Stage I being the earliest and Stage IV the most advanced [3].Cancer Grade is a parameter used to reflect the level of malignancy of a tumor, which is intended to capture how similar cancer cells of a tumor resemble stem cells, or their level of stem-ness. Cancer Grades range from well, moderately, poorly and un-differentiated, denoted as Grades I - IV with well differentiated cancers representing the least malignant and un-differentiated as the most malignant [4]. In addition, it has been well noted that cancers originated from the same cell type of the same organ may have substantially distinct phenotypes such as different growth rates, metastasis potential as well as unique exogenous signals to drive their cell division such as ER+, PR+ or triple negative breast cancers. Cancer subtypes have hence been introduced to capture commonalities within such a subset of the same cancer as well as differences between such subset and the other subsets [5]. It is noteworthy that in clinical practice, cancer staging, grading and subtyping have been predominantly done based on morphological information of a tumor possibly coupled with a limited number of markers [6]–[8].

With the rapid availability and economical affordability of considerably more informative data such as transcriptomic data, it is only natural to consider cancer classification based

* To whom correspondence should be addressed

on such data, which could lead to biologically more meaningful and scientifically more objective classification of cancer samples, to improve cancer diagnosis and treatment. There have been published studies aiming to use gene-expression data to classify cancer samples, based on differential gene-expression patterns specific to certain groups of cancer samples [9]. However, the success has been very limited knowing that none of such gene-expression based approaches has been widely used clinically. A key issue, as we noted, is that differential expression patterns tend to be either not highly conserved among or not highly specific to cancer samples of the same Stage, Grade or subtype. In contrast, co-expression patterns tend to be considerably more conserved among and could be highly specific to cancer samples at the same developmental Stage or of a similar malignancy level. The reason is that such cancer samples tend to face the same or similar stress types and levels, and may have the same or related stress response programs available to them [10], hence giving rise to common sets of co-expressed genes to conduct the same/similar functional activities.

Based on such consideration, we have developed a co-expression based general classification framework for cancer staging, grading and subtyping. A technical challenge in accomplishing this comes from how to derive co-expression information from individual samples. Here we have adapted a recent algorithm by Chen et. al [11] to overcome this challenge. The key idea of the algorithm is: it infers two genes are co-expressed in a cancer sample with respect to a given reference dataset if the two genes are co-expressed in the reference data and adding the new sample to the reference set does not statistically weaken the level of co-expression between the two genes. In our multi-classification problem, we use multiple reference datasets for each Stage, Grade or subtype, and assign a new sample to a reference set whose co-expression patterns are most consistent with those of the reference data above some threshold.

To evaluate our method, we have compared it with HSIC-LASSO+K-means [12] and HSIC-LASSO+SVM, which are well known efficient multi-classification methods based on non-linear feature selection methods. The results show that our method considerably better in several evaluation metrics: accuracy, macro-precision and weighted F1.

II. Methods

We describe here a multi-class classification method for cancer samples based on the effect (or perturbation) due to inclusion of a new sample to the co-expression patterns among genes in each reference dataset among the given set of references. We use gene expression data of three cancer types from the TCGA database to demonstrate the same classification method can be used for staging, grading and subtyping. Specifically, lung adenocarcinoma (LUAD), with 59, 278, 124, 84 and 27 samples as normal and Stage I through IV caner tissues, respectively, is used for staging; kidney renal clear cell carcinoma (KIRC), having 72, 14, 229, 206 and 76 samples as normal and Grade I through IV samples, respectively, is for grading; and breast cancer (BRCA) is used for subtyping, which consists of 113, 437, 37 and 115 samples as normal breast tissues, ER+, HER2+ and triple negative breast cancer tissues, respectively. Normalized FPKM values are used in our analysis. Only the expressed genes are considered here [13].

A. Calculation of co-expressions between two genes and identification of specific co-expression

Consider a collection of K non-overlapping subsets of the TCGA samples of a cancer type, each representing a subset of samples in a particular Stage (or Grade, subtype), referred to as a group. K = 5 or 3, depending on if the problem is to Stage, Grade or subtype cancer samples based on the above information. For each pair of genes i and j in the $k-th$ group, we use the Spearman correlation between the expressions data of genes i and $j$across all samples in the group:

$$c_{ij}^{n_k} = 1 - \frac{6\sum_{1\le r\le n_k} d_r^2}{n_k(n_k^2-1)} \tag{1}$$

where n_k is the number of samples in the group, dr is the rank distance between the $r-th$ elements of genes i and j. Published studies have shown that Spearman correlation is the best for measuring co-expressions between two genes [14]. Through analyzing the co-expression network of each group, we found that the nodes degree distributions of various groups are totally different, which suggest that there should exist specific co-expression for each group.

For a new sample and the K co-expression values $(c_{ij}^1, c_{ij}^2, \ldots, c_{ij}^k)$ for each gene pair $(i,\ j)$ in K groups, by sorting their absolute value decreasingly to generate $(c_{ij}^{1\,\prime}, c_{ij}^{2\,\prime}, \ldots, c_{ij}^{k\,\prime})$, we use the difference between the maximum and the second largest as an index to select the specific co-expression feature. That is to say, for a genes pair i and j, if group p has the maximal co-expression value and group q ranked the second, and when their difference is big enough, pair i and j will be treated as the candidate specific co-expression feature for group p.

$$\Delta_{ij} = |c_{ij}^{1'} - c_{ij}^{2'}| \tag{2}$$

here $c_{ij}^{1'}$ is the maximal co-expression coefficient of gene i and gene j, $c_{ij}^{2'}$ ranked the second. Δ_{ij} is their absolute difference. For each group, we select most relevantly specific co-expression features accordingly by ordering the difference. Besides, in order to detect their biological function, we perform pathway enrichment analysis for specific co-expression genes by utilizing hypergeometric distribution.

B. Perturbation of co-expression patterns by including a new sample

It is noteworthy that if two genes are co-expressed across a set S of samples, their level of co-expression will not go down in an expanded set of S by including a new sample, in which the expression levels of two genes are highly consistent with those in S, otherwise the number will go down. This is the

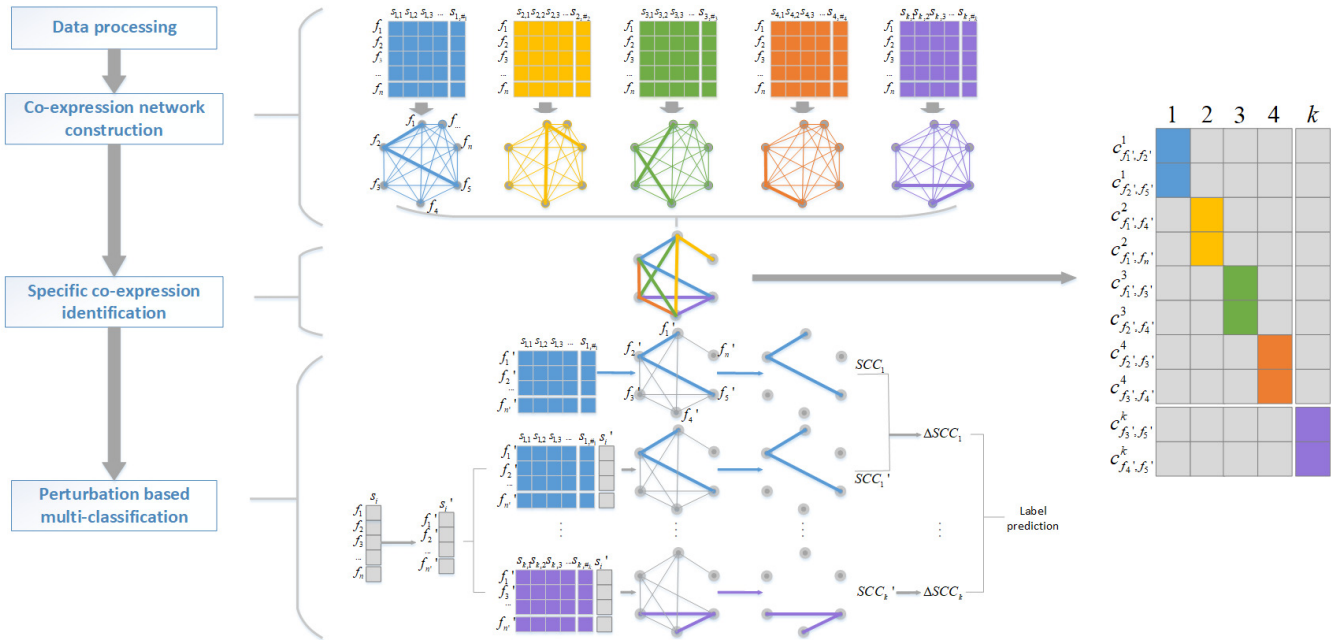

Fig. 1. **The overall workflow.** Take five-group classification as an example. Each group is represented as a $n * m$ gene expression matrix for n genes and m samples, with rows for genes and columns for samples. $s_{k,i}$ is the $i-th$ sample in group k. Each network represents a co-expression network, where a link in bold represents a pair of genes whose expressions are highly consistent between a group and a new sample. For a new sample s_j, we select the top P groups whose overall expression patterns are most consistent with the sample.

basis of the work by Chen et. al [15], and will also serve as a basis for our classification method.

For a new sample and the K co-expression values $(c_{ij}^{n_1}, c_{ij}^{n_2}, \ldots, c_{ij}^{n_k})$ for each gene pair (i, j) in K groups, define $c_{ij}^{n_k+1}$ as the new expression level of (i, j) in the expanded $k-th$ group by including the new sample. Hence $(c_{ij}^{n_k+1} - c_{ij}^{n_k})$ should be a very small value, close to zero, if the new sample has (i, j) expressions highly consistent with those in the kth group, or a negative value otherwise. Generally, if a sample intrinsically belongs a specific group, $(c_{ij}^{n_k+1} - c_{ij}^{n_k})$ values should be small for the majority of the (i, j) pairs, and should be large negative values for other groups. Based on this intuition, we formulate our classification problem as to identify a group among the given ones so the following is minimized as:

$$\arg \min_{1 \le k \le K} \sum_{ij=1} |c_{ij}^{n_k+1} - c_{ij}^{n_k}| \cdot \sqrt[3]{n_k^2} \tag{3}$$

where $\sqrt[3]{n_k^2}$ is a normalization factor aimed to minimize biases caused due to sample sizes of different groups, which is determined empirically. Figure 1 summarizes the overall idea of our method.

We have performed 10-fold cross-validation to validate the performance of the method. Each time, we randomly select 90% of the samples of each group as one of the reference dataset and test classification result on the remaining 10% samples to derive the average performance. The computer program for the whole algorithm is available on Github: https://github.com/JhyOnya/SCP

C. Evaluation

We have assessed our method by comparing it with two existing methods using three measures on three TCGA cancer datasets:

$$Accuracy = \sum_{i=1}^{k} TP_i \Big/ \sum_{i=1}^{k} \#_i \tag{4}$$

$$Macro - Precision = \frac{1}{k} \sum_{j=1}^{k} \frac{TP_j}{TP_j + FP_j} \tag{5}$$

$$weighted - F1 = \frac{1}{\sum_{i=1}^{n} \#_i} \sum_{i=1}^{n} F1 - score_i \times \#_i \tag{6}$$

where $\#_i$ is the sample number of $i-th$ group; and TP is for true positives, FP for false positives, FN for false negatives, and TN for true negatives.

D. Methods we have compared with

To the best of our knowledge, there is not published method for using co-expression based cancer classification method. Hence, we have compared two state-of-the-art methods: Hilbert-Schmidt independence criterion (HSIC)-Lasso is a kernel-based non-linear feature selection approach for multi-classification; and K-means is widely used for classification. SVM is a kernel-based two-class classification method and could be extended for multi-classification.

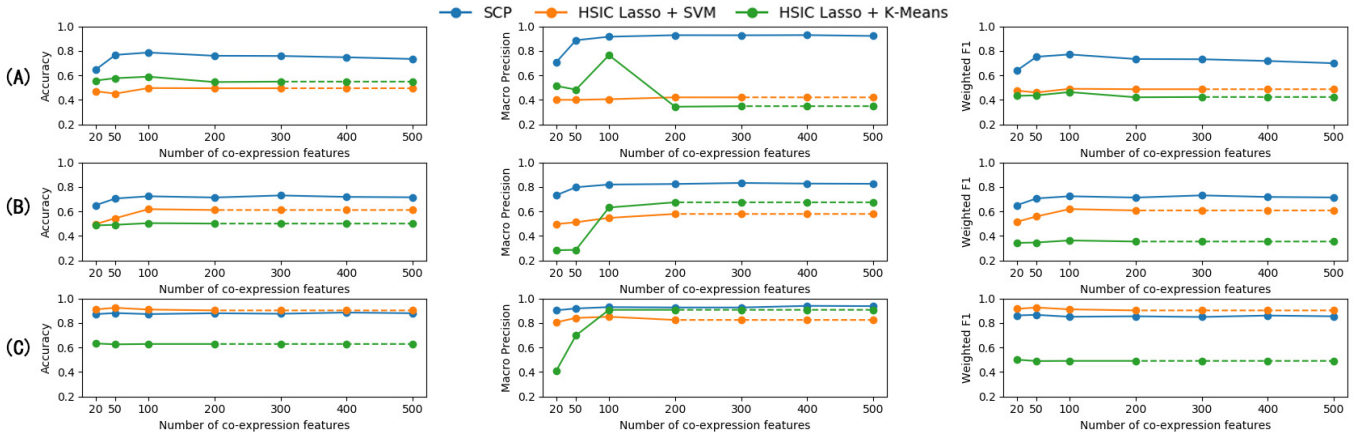

Fig. 2. Cancer staging, grading and subtyping performance by three methods: our method SCP, HSICLASSO+K-Means and HSICLASSO+SVM in (A) LUAD staging;(b) KIRC grading;(c) BRCA subtyping with selecting @ 20,50,100,200,300,400 500 specific co-expression. The dotted lines mean no updated results.

TABLE I
EVALUATION OF USING @100 SPECIFIC CO-EXPRESSION GENES IN EACH CATEGORY FOR LUAD STAGING AND KIRC GRADING.

DataSets		SCP @100			HSIC-LASSO + K-Means @100			HSIC-LASSO + SVM @100		
		precision	recall	f1-score	precision	recall	f1-score	precision	recall	f1-score
LUAD	Normal	**1.000**	0.967	**0.983**	0.868	**1.000**	0.929	0.903	0.949	0.926
	Stage I	**0.706**	**0.979**	**0.820**	0.554	0.971	0.706	0.613	0.626	0.619
	Stage II	0.894	**0.492**	**0.634**	**1.000**	0.008	0.016	0.262	0.266	0.264
	Stage III	**0.974**	**0.475**	**0.639**	0.400	0.071	0.121	0.250	0.250	0.250
	Stage IV	**1.000**	**0.633**	**0.776**	1.000	0.037	0.071	0.000	0.000	0.000
KIRC	Normal	**1.000**	**1.000**	**1.000**	0.973	0.986	0.979	0.986	0.972	0.979
	Grade I	**1.000**	**0.700**	**0.824**	0.000	0.000	0.000	0.100	0.071	0.083
	Grade II	**0.754**	0.613	**0.676**	0.436	**0.987**	0.605	0.654	0.603	0.627
	Grade III	0.619	**0.757**	**0.681**	**0.750**	0.015	0.029	0.566	0.583	0.574
	Grade IV	0.722	**0.713**	**0.717**	**1.000**	0.013	0.026	0.430	0.526	0.473

III. RESULTS

We have applied our method on three cancer multi-classification problems: cancer staging, grading and subtyping. Figure 2 summarizes the performance by our method and the two other methods.

A. A case study on human lung cancer staging

A total of 13,980 genes are used for co-expression analyses. We have performed seven experiments by selecting top N, with N= 20, 50, 100, 200, 300, 400, 500, co-expression values, respectively. Compared with HSIC Lasso + K-Means and HSIC Lasso + SVM, our method shows considerably better performance across all three metrics. Figure 2 show the performance when top 100 co-expression values are used. Further performance details are given in Table I. We can see that our method performs well in distinguishing between Stage II and Stage III samples, which are well known difficult to separate.

We have analyzed the biological pathways enriched by genes achieving the top 100 co-expression values. We not: in Stage I, protein metabolic process (p-value = 1.34E-03) and purine metabolism (p-value = 5.14E-03) are significantly enriched. In Stage II, positive regulation of cell proliferation (p-value = 4.01E-04) and axon guidance (p-value = 2.87E-04) are significantly enriched. In Stage III, cytoskeleton organization and biogenesis (p-value = 1.88E-03) and MYC active pathway (p-value = 8.05E-04) are considerably enriched. In Stage IV, establishment of cellular localization (p-value = 4.72E-03) and epithelial to mesenchymal transition (EMT) (p-value = 5.95E-04) are significantly enriched, which is well known associated with cancer metastasis. Clearly all these enriched pathways at each Stage are highly consistent with our knowledge about cancer biology. Specifically, in the beginning, cancer cells prepare and accumulate enough nutrients and start cell proliferation; as a cancer develops, cancer cells adjust their cell shape by organizing cytoskeleton preparing for invasion; and finally activate the EMT for metastasis.

B. A case study on human kidney cancer grading

A total of 13,789 genes are used for co-expression analyses. Like in the previous case, top N, N= 20, 50,100, 200, 300, 400, 500) co-expression values from each group are used for cancer grading. Figure 2 shows the performance by our method and two other methods for the case of N = 100. Clearly, our method outperforms considerably the other methods, with more detailed information shown in Table I.

Pathway enrichment analyses are also conducted over the above selected genes for each Grade. We note: for the case when N=100, Grade I samples significantly enrich metabolic processes (p-value = 5.95E-04). Grade II samples enrich DNA modification (p-value = 8.17E-04). Transcription (p-value = 8.42E-04) and cell proliferation (p-value = 1.08E-14) are significantly enriched by samples of Grade III and Grade IV, respectively. These Grade specific co-expressed genes reflect the cancer cell dedifferentiation level and characteristics as cell dedifferentiation generally has a compatible rate of cell division. In the opposite, the higher the degree of dedifferentiation, the higher chance for cells to divide. Our enrichment results capture these characteristics.

C. A case study on human breast cancer subtyping

Breast cancer is known to fall into distinct subtypes, including ER+, HER2+ and triple-negative breast cancers. Understanding molecular level differences among these subtypes could contribute to mechanistic studies of cancer. A similar set of performance is done using the three programs using N = 100.

Pathway enrichment analyses are also conducted among the genes whose pair exhibit high level of co-expressions among samples of each subtype. We noted: (1) ER+ breast cancer samples tend to enrich glycosaminoglycan metabolic processes, such as keratin sulfate (p-value = 1.14E-08), epidermis development (p-value = 4.74E-06) and ECM receptor interaction (p-value = 2.76E-08). HER2+ breast cancers generally enrich immune system processes (p-value = 2.75E-04), lipoprotein metabolism (p-value = 5.14E-04) and signaling by FGFR1 fusion pathways (p-value = 3.15E-04). Triple negative breast cancers are found to enrich cation homostasis related pathways (p-value = 9.95E-06), MYC active pathway (p-value = 5.37E-05), gap junction (p-value = 1.31E-04), TGF-beta signaling pathway(p-value = 4.75E-04) and cell proliferation (p-value = 3.92E-04), all of which indicate that this group of breast cancers is more stressed associated with pH balance, and faster cell proliferation and cell migration. Further studies will be conducted to gain detailed understanding about the differences among these three subtypes of breast cancers.

IV. CONCLUSION

We presented here a multi-classification method for cancer staging, grading and subtyping based on co-expression patterns unique to each groups of cancers, including Stages, Grades and subtypes. Our preliminary results strongly indicate the power of co-expression based cancer classification, which warrants further development and application of the approach. The initial analyses associated with each group, particularly breast cancer subtypes, have revealed exciting information about the unique biology of each subtype, particularly triple negative breast cancer. All these point to future directions of further analyses of such co-expression based cancer classification.

ACKNOWLEDGMENT

The authors thank funding support from the National Natural Science Foundation of China (61902144, 61972174, 61572227), Guangdong Premier Key-Discipline Enhancement Scheme (Grant 2016GDYSZDXK036), and Guangdong Key-Project for Applied Fundamental Research (Grant 2018KZDXM076), the Development Project of Jilin Province of China (No. 20180414012GH), Jilin Provincial Key Laboratory of Big Date Intelligent Computing (No. 20180622002JC), the STU Scientific Research Foundation for Talents (NTF19032).

REFERENCES

[1] L. Hood, and M. Flores, "A personal view on systems medicine and the emergence of proactive P4 medicine: predictive, preventive, personalized and participatory," New biotechnology, vol. 29, no. 6, pp. 613-624, 2012.

[2] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," Nature reviews genetics, vol. 12, no. 1, pp. 56, 2011.

[3] M. B. Amin, F. L. Greene, S. B. Edge, C. C. Compton, J. E. Gershenwald, R. K. Brookland, L. Meyer, D. M. Gress, D. R. Byrd, and D. P. Winchester, "The Eighth Edition AJCC Cancer Staging Manual: Continuing to build a bridge from a population-based to a more "personalized" approach to cancer staging," CA: a cancer journal for clinicians, vol. 67, no. 2, pp. 93-99, 2017.

[4] L. Ludemann, W. Grieger, R. Wurm, M. Budzisch, B. Hamm, and C. Zimmer, "Comparison of dynamic contrast-enhanced MRI with WHO tumor grading for gliomas," Eur Radiol, vol. 11, no. 7, pp. 1231-41, 2001.

[5] T. Sorlie, R. Tibshirani, J. Parker, T. Hastie, J. S. Marron, A. Nobel, S. Deng, H. Johnsen, R. Pesich, S. Geisler, J. Demeter, C. M. Perou, P. E. Lonning, P. O. Brown, A. L. Borresen-Dale, and D. Botstein, "Repeated observation of breast tumor subtypes in independent gene expression data sets," Proc Natl Acad Sci U S A, vol. 100, no. 14, pp. 8418-23, Jul 8, 2003.

[6] L. Huang, H. Fernandes, H. Zia, P. Tavassoli, H. Rennert, D. Pisapia, M. Imielinski, A. Sboner, M. A. Rubin, and M. Kluk, "The cancer precision medicine knowledge base for structured clinical-Grade mutations and interpretations," Journal of the American Medical Informatics Association, vol. 24, no. 3, pp. 513-519, 2017.

[7] D. P. Cook, and B. C. Vanderhyden, "Ovarian cancer and the evolution of subtype classifications using transcriptional profiling," Biology of reproduction, 2019.

[8] Y. Xu, J. Cui, and D. Puett, Cancer bioinformatics: Springer, 2014.

[9] G. Victo Sudha, and R. Cyril, "Review On Feature Selection Techniques And The Impact Of Svm For Cancer Classification Using Gene Expression Profile," International Journal of Computer Science & Engineering Survey, vol. 2, no. 3, pp. 16-27, 2011.

[10] H. Sun, C. Zhang, S. Cao, T. Sheng, N. Dong, and Y. Xu, "Fenton reactions drive nucleotide and ATP syntheses in cancer," Journal of molecular cell biology, vol. 10, no. 5, pp. 448-459, 2018.

[11] X. Liu, Y. Wang, H. Ji, K. Aihara, and L. Chen, "Personalized characterization of diseases using sample-specific networks," Nucleic acids research, vol. 44, no. 22, pp. e164-e164, 2016.

[12] M. Yamada, W. Jitkrittum, L. Sigal, E. P. Xing, and M. Sugiyama, "High-dimensional feature selection by feature-wise kernelized lasso," Neural computation, vol. 26, no. 1, pp. 185-207, 2014.

[13] H. Sun, L. Chen, S. Cao, Y. Liang, and Y. Xu, "Warburg Effects in Cancer and Normal Proliferating Cells: Two Tales of the Same Name," Genomics, proteomics & bioinformatics, 2019.

[14] R. Anglani, T. M. Creanza, V. C. Liuzzi, A. Piepoli, A. Panza, A. Andriulli, and N. Ancona, "Loss of connectivity in cancer co-expression networks," PloS one, vol. 9, no. 1, pp. e87075, 2014.