Received January 15, 2020, accepted January 31, 2020, date of publication February 11, 2020, date of current version February 19, 2020.

Digital Object Identifier 10.1109/ACCESS.2020.2973220

Classification of Cancers Based on a Comprehensive Pathway Activity Inferred by Genes and Their Interactions

PENG XU[1,2], GENG ZHAO[3], ZHENG KOU[1], GANG FANG[1], AND WENBIN LIU[1]

[1]Institute of Computational Science and Technology, Guangzhou University, Guangzhou 510006, China
[2]School of Computer Science of Information Technology, Qiannan Normal University for Nationalities, Duyun 558000, China
[3]Department of Physics and Electronic Information Engineering, Wenzhou University, Wenzhou 325035, China

Corresponding author: Wenbin Liu (wbliu6910@gzhu.edu.cn)

This work was supported in part by the National Key Research and Development Program of China under Grant 2019YFA0706402, and in part by the National Natural Science Foundation of China under Grant 61572367, Grant 61972107, and Grant 61972109).

ABSTRACT Cancers, a group of multifactorial complex diseases, are generally caused by mutation of multiple genes or dysregulation of gene interactions. Applying machine learning methods to microarray gene expression profiles for disease classification is a popular method to predict disease state or outcome. Traditional computational methods that detect genes differentially expressed between cancer and normal samples are ineffective in independent cohorts of patients. However, current methods consider pathways as simple gene sets and include pathway topological information but ignore significant individual genes and interactions between genes, which are essential to infer a more robust pathway activity. In this study, we proposed a novel approach to describe the activity of a pathway that incorporates both the differential expression degree of genes between the case and control and the interaction strength between genes. We applied the method to the classifications of seven cancers. Within-dataset experiments and cross-dataset experiments demonstrated that our novel method achieved robust and superior performance when compared to the five existing methods.

INDEX TERMS Classification, cancer, pathway activity.

I. INTRODUCTION

Analyses of genome-wide expression profiles can aid in understanding the mechanisms of biological processes, identifying biomarkers for cancers and designing therapeutic strategies [1]–[8]. One important challenge in clinical cancer research is accurately predicting disease states and treatment responses of a patient based on the expression of genes. An increasing number of disease markers have been identified through the analysis of genome-wide expression profiles [9]–[12]. One direct approach is to score each individual gene based on its power to discriminate samples between case and control [13]–[16]. However, the gene markers identified in one dataset usually share little overlap with those obtained in other datasets due to noise in microarray data and cellular heterogeneity within tissues. In addition, precise classification is also impeded by the so-called ''large p small n'' property, whereby the number of samples (or instances) is typically several orders of magnitude smaller than the number of genes (or features), making it difficult to extract reliable information from transcriptome profiles [17], [18]. All of these factors often lead to gene markers discovered in one dataset failing to be predictive of the same disease phenotype in other independent datasets.

As gene products are known to function coordinately in functional modules or signaling cascades, perturbed high-level functional modules may be more consistent with the disease state of interest than individual genes [19]. Thus, integrating gene expression data with available large protein-protein interaction (PPI) networks or known pathways may identify more reproducible biomarkers [20]–[26]. Network-level analyses can be categorized as PPI-based or pathway-based methods. Both approaches consist of three steps: first, search potential subnetworks or pathways and sort them according to their discriminative score; second, select feature subnetworks or pathways; finally, design a classifier

The associate editor coordinating the review of this manuscript and approving it for publication was Quan Zou.

according to the activity of the selected subnetworks or pathways. Chuang *et al*. proposed a method to search subnetwork markers based on mutual information or t-scores measuring the association between the marker's activity and class label [27]. Su *et al*. searched for the top discriminative linear paths using dynamic programming in a PPI network. The discriminative score of a path incorporated both the t-test statistics of the member genes and the correlation between their expression values [28]. The activity of a subnetwork was inferred by combining the normalized log-likelihood ratios (LLRs) of its member genes. In pathway-based analyses, the discriminative score of a pathway is defined as the t-test statistic score for the member genes. The main difference between these approaches lies in how they define pathway activity. For example, Guo *et al*. estimated the pathway activity using the mean or median of the gene expression values of the member genes [29]. In a PCA approach, Bild *et al*. used the first basis vector to weight the expression values of the member genes in a pathway [30]. Lee *et al*. proposed to infer the pathway activity by condition-responsive genes method [31]. Liu *et al*. proposed a directed random walk (DRW) to mine the topological importance of genes in a pathway network. The activity of a pathway was defined by the weighted expression values of the member genes [32]. They also extended this method to include both genomic and metabolic data [33]–[36]. Recently, this topological approach was applied to predict breast cancer survival outcomes [37].

Although previous methods have achieved great progress in cancer classification based on the activity of pathways or subnetworks, the activity of the pathway or subnetwork was defined as a simple summary of expression values of the member genes, which could not reflect the interactions between genes at the network level. However, it is the interaction between genes that shifts the direction of biological signaling cascades. In order to model their effect, Tarca *et al*. proposed a signaling pathway impact analysis (SPIA) method to model the impact of perturbed upstream genes on their downstream partners [38].

In this study, we proposed a method to quantify pathway activity using both genes and their interactions (PAGI). We first constructed a pathway expression profile matrix that includes both genes and their interactions. Then the first principle component of the expression profile matrix was calculated. Finally, the activity of the pathway was derived based on the product of the first component and their corresponding expression values. Both within-dataset experiments and cross-dataset experiments demonstrated that the proposed PAGI method was more accurate and more robust than the DRW, PAC, mean, median, and gene methods on datasets for seven different cancers.

II. DATASETS

Table 1 lists the 22 microarray datasets for seven cancers downloaded from the NCBI Gene Expression Omnibus (GEO) database [39]. In these datasets, 9 datasets were studied in the within-dataset experiment and were used as training dataset in cross-dataset experiment: GSE10072 for lung cancer [40], GSE13911 for stomach cancer [41], GSE17856 for liver cancer [42], GSE5364 for thyroid cancer [42], GSE15641 and GSE17895 for kidney cancer [43], [44], GSE3494 and GSE1456 for breast cancer [34], [45], and GSE8511 for prostate cancer[46]. The other 13 datasets were used for validation in cross-dataset experiments. In the breast cancer datasets, patients died within 5 years were defined as negative samples, while the remaining patients were considered positive samples (patients with a survival time of 55 years without any reported events were excluded). The prostate cancer datasets contained three types of samples: Benign, PCA, and Mets, and we built two classifications to classify Benign and PCA samples as well as PCA and Mets samples. All pathway information was downloaded from the KEGG database [47].

TABLE 1. Cancer gene expression datasets.

GEO ID	Cancer	Sample size	Platform
GSE10072	Lung	58/49	GPL96
GSE19804		60/60	GPL570
GSE19188		91/65	GPL570
GSE13911	Stomach	38/31	GPL570
GSE19826		12/12	GPL570
GSE38940		24/24	GPL5936
GSE17856	Liver	43/44	GPL6480
GSE14520-1		64/60	GPL571
GSE14520-2		183/181	GPL571
GSE5364	Thyroid	35/16	GPL96
GSE33630		60/45	GPL570
GSE29265		29/20	GPL570
GSE15641	Kidney	69/23	GPL96
GSE17895		138/22	GPL9101
GSE36895		29/23	GPL570
GSE3494	Breast	178/58	GPL96
GSE1456		124/35	GPL96
GSE7390		35/163	GPL96
GSE8511	Prostate	16/12(Benign/PCA) 12/13(PCA/Mets)	GPL1708
GSE3325		6/7(Benign/PCA) 7/7(PCA/Mets)	GPL570
GSE32269		16/10(Benign/PCA) 10/8(PCA/Mets)	GPL96

III. METHODS

The pathway is a gene network that includes both genes and their interactions to fulfill some specific biological functions. Our motivation is that the activity of a pathway should reflect the following three factors: (1) the degree of the differential expression of genes between case and control group; (2) the correlation between a gene's expression and the class label (control, case, metastatic or non-metastatic); (3) their interaction strength between genes connected in a pathway. Based on

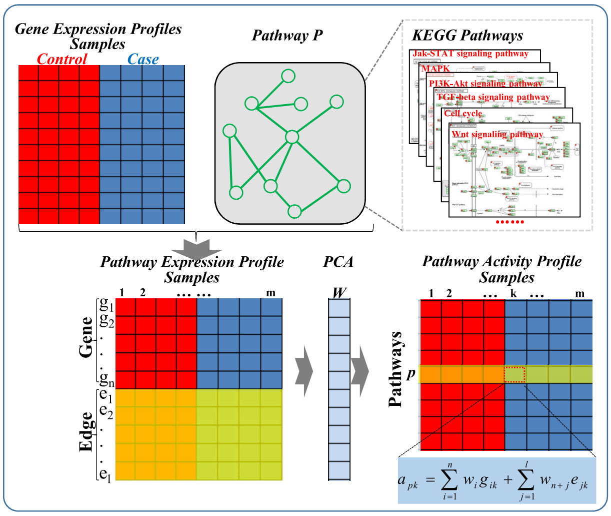


FIGURE 1. The workflow for inferring the pathway activity using genes and their interactions.

these considerations, we proposed a new way to infer the activity of a pathway. Fig. 1 displays the main workflow of the proposed method PAGI in this paper.

Given a pathway $P = \{G, E\}$ that includes genes $G = \{g_1, g_2, \ldots, g_n\}$ and interactions $E = \{e_1, e_2, \ldots, e_l\}$, we first constructed a new expression profile matrix including both genes and their corresponding interactions in a pathway network. The expression of a gene g_i in sample k is transformed as

$$z_{ik} = t_i^2 \left|\rho_i\right| g_{ik} \tag{1}$$

where t_i is the t-score of g_i calculated from a two-tailed t-test between two phenotypes, and ρ_i is the Pearson correlation coefficient between gene g_i and class label c. After this transformation, Z_{ij} actually represents a weighted expression of gene g_i in sample k which reflects both the differential expression degree of gene g_i and its correlation with the phenotype. The more differentially expressed, the larger Z_{ij}. And the larger its correlation with the phenotype, the larger Z_{ij}. Similarly, the expression profile of their interaction of gene pair g_i and g_j in sample k is defined as

$$e_{ijk} = \rho_{ij} \left|\beta_{ij}\right| \left(\frac{z_{ik} + z_{jk}}{2}\right) \tag{2}$$

where $\rho_{ij}$is the Pearson correlation coefficient between genes g_i and g_j, β_{ij} indicates the interaction type between gene g_i and g_j (1 for activation or -1 for inhibition). Obviously, the larger the interaction strength, the larger e_{ijk}. The expression profile of a pathway P can then be denoted by $a \cdot (n + 1) \times m$ matrix M_p, where rows represent the genes or their interactions and columns represent samples.

Secondly, we then applied the principal component analysis (PCA) on the matrix M_p to infer the activity score a_{pk} of pathway P in sample k as:

$$a_{pk} = \sum_{i=1}^{n} w_i z_{ik} + \sum_{j=1}^{l} w_{n+j} e_{jk} \tag{3}$$

where w_i and w_j are the corresponding component in the first eigenvector for gene i and interaction j respectively.

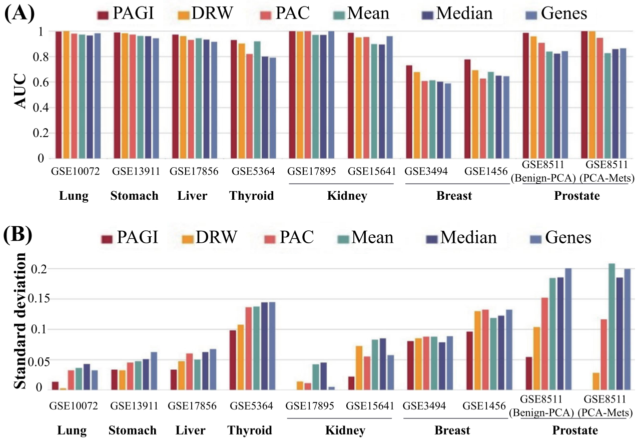


FIGURE 2. Classification performance and stability on within-datasets.

IV. RESULTS

In this section, we used the logistic regression model to evaluate the performance of gene method in [28], mean and median method in [29], PAC method in [30], DRW method in [31] and the proposed PAGI method. The average area under ROC curve (AUC) [48], [49] and the corresponding standard deviation (SD) by five-fold cross-validation [50], [51] over 1000 times were calculated for the six methods [52]–[58]. The experiment setting was the same as in [32], [35] for the DRW, PAC, mean, median, and gene methods. For the gene method, the top 50 discriminative gene markers were chosen as the candidate features in order to maintain an identical maximum number of features as in [28]. In cross-dataset experiments, the first dataset was used as the training set, and other independent datasets were used as the test set.

A. CLASSIFICATION PERFORMANCE ON WITHIN-DATASET EXPERIMENTS

Fig. 2. shows the average AUC and SD of the six methods on the 10 within-datasets. The average AUC of all the six methods were about more than 0.8 except on the two breast cancer datasets. First, PAGI achieved the largest AUC in all cancer datasets except the lung cancer dataset GSE10072 where it was slightly less than that of DRW. Especially, compared with other methods, PAGI sharply improved the AUC in four datasets GSE17895, GSE3494, GSE1456 and GSE8511. Secondly, the average SD of PAGI was the least except in three datasets GSE10072, GSE13911 and GSE3494 where it is the second least in the six methods. These two observations demonstrated that PAGI had the best overall classification performance and stability on within-datasets experiments.

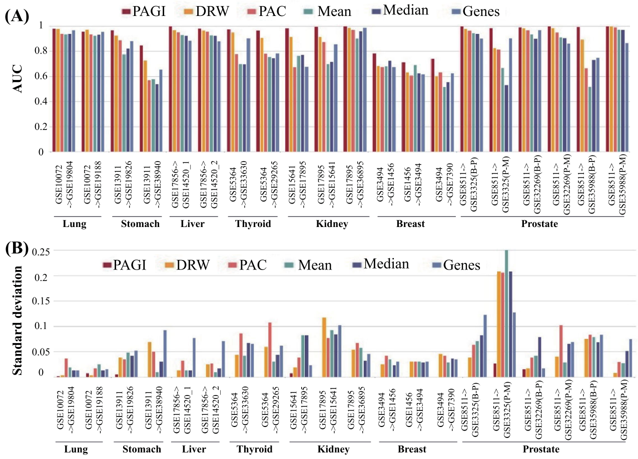


FIGURE 3. Classification performance and stability on cross-datasets.

B. CLASSIFICATION PERFORMANCE ON CROSS-DATASET EXPERIMENTS

To evaluate the generalization ability of the six methods, we carried out cross-dataset experiments using 18 additional independent datasets. Fig. 3 shows their average AUC and SD on these independent datasets. First, as expected, the average AUC for each method varied sharply in different independent datasets except the lung cancer and liver cancer datasets. Apart from the heterogeneity and noises inherent in these datasets, how to extract the internal characteristic is the key to deal with this reproducibility issue. Secondly, PAGI achieved the largest average AUC in all datasets except the independent lung cancer dataset GSE19188 where it was slightly less than that of DRW. Especially, PAGI sharply improved the AUC in at least one independent datasets in the six cancers except for lung cancer. Thirdly, the average SD of PAGI was the least except in one lung cancer dataset GSE19188 where it is the second least in the six methods. These observations demonstrated that PAGI also had the best overall classification performance and stability on cross-datasets experiments which were consistent with the with-datasets experiments. They indicated that the PAGI-based pathway activities were less sensitive to different cohorts of patients and microarray platforms and were more reliable in predicting clinical outcomes in practice. A potential reason may be that the PAGI incorporates both the importance of genes based on their differential expressions and topological interaction information to build the classifier [59].

C. ROBUSTNESS OF RISK-ACTIVE PATHWAYS

In cancer studies, many pathways, such as the MAPK signaling pathway, p53 signaling pathway, and pathway in cancer, have been found highly related to the development of various cancers [32], [60]. Table 2. lists 18 known cancer-related pathways which involve in various biological processes, including cell cycle, apoptosis, and senescence. The degree indicates the number of pathways connected with it in the whole pathway network. From the perspective of classification, the perturbation of gene expressions in these pathways should provide enough information.

TABLE 2. Cancer-related pathways studied by PAGI.

NO.	Pathway Name	Degree
1	MAPK signaling pathway	69
2	Adherens junction	36
3	Pathway in cancer	31
4	ECM-receptor interaction	26
5	Tight junction	22
6	Adipocytokine signaling pathway	19
7	Regulation of actin cytoskeleton	18
8	p53 signaling pathway	17
9	Calcium signaling pathway	14
10	Endocytosis	13
11	PPAR signaling pathway	12
12	Progesterone-mediated oocyte maturation	10
13	Proteasome	10
14	Focal adhesion	8
15	Wnt signaling pathway	8
16	Insulin signaling pathway	4
17	Axon guidance	3
18	RNA transport	3

In this paper, the proposed PAGI only used one of them as a feature to build classifier. The performance of PAGI was actually the best AUC obtained by one of the 18 pathways. The best feature pathway for different cancer datasets might be different. For example on within-datasets, insulin signaling pathway was the best pathway for lung cancer, RNA transport for stomach cancer, adipocytokine signaling pathway for liver cancer, p53 signaling pathway for thyroid, MAPK signaling pathway and regulation of actin cytoskeleton for the two kidney cancer datasets respectively, MAPK signaling pathway and regulation of actin cytoskeleton for the two breast cancer datasets respectively, and MAPK signaling pathway for prostate cancer. On cross-datasets, the best pathway for different cancer datasets was also different.

For a given dataset, we found that the results of PAGI by most of the 18 pathways were very close to the best performance on both within-datasets and cross-datasets. Fig. 4. shows the average AUC of PAGI by the five pathways with the largest degree in Table 2. The close performance by these pathways demonstrated that the proposed activity score could provide highly discriminative information for cancer classification only by one pathway. This indicates that the newly proposed pathway activity might capture more of the essential features of various cancers than that used by mean, median, PAC and DRW methods whose best performance was derived from a selected pathway set.

MAPK signaling pathway is an important known cancer pathway connected with 69 other pathways. Our results showed that the average AUC by this pathway was the largest both in the within-datasets and cross-datasets of the seven cancers. That is, we could reach a relative satisfactory classification result by the MAPK pathway without feature selection. Apart from it connects with many important pathways,

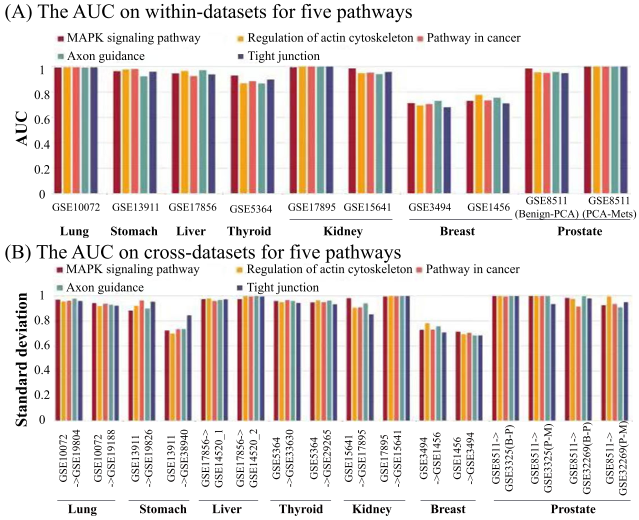


FIGURE 4. The average AUC of PAGI by the five pathways on within-datasets (a) and cross-datasets (b).

another reason may be MAPK signaling pathway shared many important genes with other pathways.

D. SIGNIFICANT DIFFERENCE OF GENES AND THEIR INTERACTIONS BETWEEN TWO PHENOTYPES

To mine the important information for medical diagnosis, we further analyzed the significant difference of genes and their interactions in pathway between two phenotypes. We acquired 10(42) genes and 56(361) interactions with significant difference between the two phenotypes in the ‘‘Benign-PCA’’ case (the ‘‘PCA–Mets’’ case).

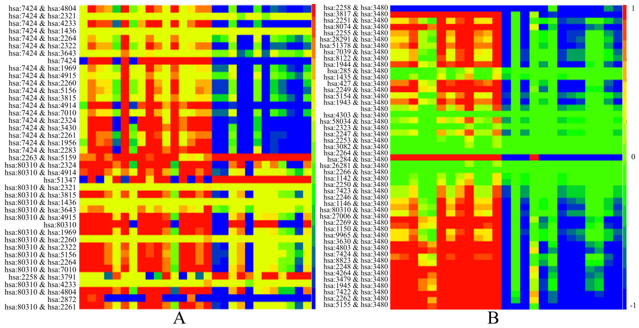


FIGURE 5. Heat maps of top difference genes and interactions in two prostate cancer cases.

Fig.5 show the heat maps (A and B) of the significant difference of genes and interactions in the ‘‘Benign-PCA’’ case and the ‘‘PCA–Mets’’ case. On one side these individual genes such as NGFR and FLT1 have small difference between two phenotypes, but their interactions with other genes had significant difference between two phenotypes. On other side, some individual genes such as NGF and FGFR2 had small difference between two phenotypes, but their interactions had outstanding difference when they interacted with each other.

V. CONCLUSION

How to accurately discriminate various cancers is a crucial issue for clinical treatment. As genes are corporately interacted with each other to fulfill specific biological functions, the activities of pathways become a potential feature for cancer classification. In this paper, we proposed a novel method to describe pathway activity which incorporates both the genes' activity and their interactions. Specifically, in order to extract the essential features for a disease state, we first transformed the expression of a gene to a weighted activity based on their differential expression degree between case and control and their correlation with the phenotype. Then we defined the activity of a gene pair in a pathway by their interaction strength. Finally, the activity score of a pathway for a sample was calculated as an arithmetic weighted activity of genes and gene pairs by the first eigenvector of PCA on the expression profile matrix of the pathway.

We studied the performance of the new proposed method PAGI on datasets of seven cancers, which included 10 within-dataset experiments and 20 cross-dataset experiments. Results on these datasets demonstrated that the proposed PAGI performed better and was more robust than the other five methods. Furthermore, the proposed PAGI could achieve the best performance by using only one pathway while the other methods might need to select the best pathway set. Results on the 18 known cancer-related pathways showed that the performance of most pathways was very close to the best performance. This indicated that the proposed PAGI was even robust on many cancer-related pathways. Additionally, we found that the proposed PAGI could achieve a satisfactory performance for all datasets by the MAPK signaling pathway. Although PAGI had above advantages, we believe there is still room to study pathway activity more effectively, by employing new generation machine learning [61]–[67] and computational intelligence algorithms [68]–[73].

REFERENCES

[1] Z. Liao, D. Li, X. Wang, L. Li, and Q. Zou, ‘‘Cancer diagnosis through IsomiR expression with machine learning method,’’ *Current Bioinf.*, vol. 13, no. 1, pp. 57–63, Feb. 2018.

[2] L. J. Van 't Veer, ‘‘Gene expression profiling predicts clinical outcome of breast cancer,’’ *Nature*, vol. 415, no. 6871, pp. 530–536, Jan. 2002.

[3] S. Ramaswamy, K. N. Ross, E. S. Lander, and T. R. Golub, ‘‘A molecular signature of metastasis in primary solid tumors,’’ *Nature Genet.*, vol. 33, no. 1, pp. 49–54, Jan. 2003.

[4] T. R. Golub, D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Loh, J. R. Downing, M. A. Caligiuri, C. D. Bloomfield, and E. S. Lander, ‘‘Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring,’’ *Science*, vol. 286, no. 5439, pp. 531–537, Oct. 1999.

[5] A. A. Alizadeh, ‘‘Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling,’’ *Nature*, vol. 403, no. 6769, pp. 503–511, Feb. 2000.

[6] L. Jiang, Y. Xiao, Y. Ding, J. Tang, and F. Guo, ‘‘Discovering cancer subtypes via an accurate fusion strategy on multiple profile data,’’ *Frontiers Genet.*, vol. 10, p. 20, Feb. 2019.

[7] G. I. Lambrou, M. Sdraka, and D. Koutsouris, ‘‘The 'gene cube': A novel approach to three-dimensional clustering of gene expression data,’’ *Current Bioinf.*, vol. 14, no. 8, pp. 721–727, 2019.

[8] B. Liu, X. Gao, and H. Zhang, ‘‘BioSeq-Analysis2.0: An updated platform for analyzing DNA, RNA, and protein sequences at sequence level and residue level based on machine learning approaches,’’ *Nucleic Acids Res.*, vol. 47, no. 20, p. e127, 2019.

[9] W. Tang, S. Wan, Z. Yang, A. E. Teschendorff, and Q. Zou, ‘‘Tumor origin detection with tissue-specific miRNA and DNA methylation markers,’’ *Bioinformatics*, vol. 34, no. 3, pp. 398–406, Feb. 2018.

[10] T. Wang, H. Ke, X. Zheng, K. Wang, A. K. Sangaiah, and A. Liu, "Big data cleaning based on mobile edge computing in industrial sensor-cloud," *IEEE Trans. Ind. Informat.*, vol. 16, no. 2, pp. 1321–1329, Feb. 2020.

[11] L. Li, J. Li, J. Sun, P. Yi, C. Li, Z. Zhou, M. Xin, J. Sheng, L. Shuai, Z. Li, D. Ling, X. He, F. Zheng, G. Liu, and Y. Tang, "Role of phospholipase d inhibitor in regulating expression of senescence-related phospholipase D gene in postharvest longan fruit," *Current Bioinf.*, vol. 14, no. 7, pp. 649–657, Sep. 2019.

[12] B. Liu and K. Li, "IPromoter-2L2.0: Identifying promoters and their types by combining smoothing cutting window algorithm and sequence-based features," *Mol. Therapy-Nucleic Acids*, vol. 18, pp. 80–87, Dec. 2019.

[13] L. Ein-Dor, I. Kela, G. Getz, D. Givol, and E. Domany, "Outcome signature genes in breast cancer: Is there a unique set?" *Bioinformatics*, vol. 21, no. 2, pp. 171–178, Jan. 2005.

[14] W. F. Symmans, J. Liu, D. M. Knowles, and G. Inghirami, "Breast cancer heterogeneity: Evaluation of clonality in primary and metastatic lesions," *Hum. Pathol.*, vol. 26, no. 2, pp. 210–216, Feb. 1995.

[15] S. A. Tomlins, "Recurrent fusion of TMPRSS2 and ETS transcription factor genes in prostate cancer," *Science*, vol. 310, no. 5748, pp. 644–648, Oct. 2005.

[16] V. K. Mootha, "PGC-1 alpha-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes," *Nature Genet.*, vol. 34, no. 3, pp. 267–273, Jul. 2003.

[17] B. Peng, D. Zhu, B. P. Ander, X. Zhang, F. Xue, F. R. Sharp, and X. Yang, "An integrative framework for Bayesian variable selection with informative priors for identifying genes and pathways," *PLoS ONE*, vol. 8, no. 7, Jul. 2013, Art. no. e67672.

[18] T. Wang, H. Luo, X. Zheng, M. Xie, and Technology, "Crowdsourcing mechanism for trust evaluation in CPCS based on intelligent mobile edge computing," *ACM Trans. Intell. Syst.*, vol. 10, no. 6, p. 62, 2019.

[19] Y. Wu, H. Huang, Q. Wu, A. Liu, and T. Wang, "A risk defense method based on microscopic state prediction with partial information observations in social networks," *J. Parallel Distrib. Comput.*, vol. 131, pp. 189–199, Sep. 2019.

[20] T. Wang, Y. Mei, W. Jia, X. Zheng, G. Wang, and M. Xie, "Edge-based differential privacy computing for sensor–cloud systems," *J. Parallel Distrib. Comput.*, vol. 136, pp. 75–85, Feb. 2020.

[21] Y. Wu, H. Huang, N. Wu, Y. Wang, M. Z. Alam Bhuiyan, and T. Wang, "An incentive-based protection and recovery strategy for secure big data in social networks," *Inf. Sci.*, vol. 508, pp. 79–91, Jan. 2020.

[22] X. Zeng, Y. Lin, Y. He, L. Lv, X. Min, and A. Rodriguez-Paton, "Deep collaborative filtering for prediction of disease genes," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, to be published, doi: 10.1109/tcbb.2019.2907536.

[23] X. Zhang, Q. Zou, A. Rodriguez-Paton, and X. Zeng, "Meta-path methods for prioritizing candidate disease miRNAs," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 16, no. 1, pp. 283–291, Jan. 2019.

[24] X. Zeng, Y. Zhong, W. Lin, and Q. Zou, "Predicting disease-associated circular RNAs using deep forests combined with positive-unlabeled learning methods," *Briefings Bioinf.*, to be published, doi: 10.1093/bib/bbz080.

[25] B. Wang, "Early stage identification of Alzheimer's disease using a two-stage ensemble classifier," *Current Bioinf.*, vol. 13, no. 5, pp. 529–535, 2018.

[26] B. Liu, "BioSeq-Analysis: A platform for DNA, RNA and protein sequence analysis based on machine learning approaches," *Briefings Bioinf.*, vol. 20, no. 4, pp. 1280–1294, Jul. 2019.

[27] H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Mol. Syst. Biol.*, vol. 3, no. 1, p. 140, Oct. 2007.

[28] J. Su, B.-J. Yoon, and E. R. Dougherty, "Accurate and reliable cancer classification based on probabilistic inference of pathway activity," *PLoS ONE*, vol. 4, no. 12, p. e8161, Dec. 2009.

[29] Z. Guo, "Towards precise classification of cancers based on robust gene functional expression profiles," *BMC Bioinf.*, vol. 6, p. 58, Mar. 2005.

[30] A. H. Bild, G. Yao, J. T. Chang, Q. Wang, A. Potti, D. Chasse, M.-B. Joshi, D. Harpole, J. M. Lancaster, A. Berchuck, J. A. Olson, J. R. Marks, H. K. Dressman, M. West, and J. R. Nevins, "Oncogenic pathway signatures in human cancers as a guide to targeted therapies," *Nature*, vol. 439, no. 7074, pp. 353–357, Jan. 2006.

[31] E. Lee, H.-Y. Chuang, J.-W. Kim, T. Ideker, and D. Lee, "Inferring pathway activity toward precise disease classification," *PLoS Comput Biol*, vol. 4, no. 11, Nov. 2008, Art. no. e1000217.

[32] W. Liu, C. Li, Y. Xu, H. Yang, Q. Yao, J. Han, D. Shang, C. Zhang, F. Su, X. Li, Y. Xiao, F. Zhang, M. Dai, and X. Li, "Topologically inferring risk-active pathways toward precise cancer classification by directed random walk," *Bioinformatics*, vol. 29, no. 17, pp. 2169–2177, Sep. 2013.

[33] N. M. Evers, S. Wang, J. H. J. Van Den Berg, R. Houtman, D. Melchers, L. H. J. De Haan, A. G. H. Ederveen, J. P. Groten, and I. M. C. M. Rietjens, "Identification of coregulators influenced by estrogen receptor subtype specific binding of the ER antagonists 4-hydroxytamoxifen and fulvestrant," *Chem.-Biol. Interact.*, vol. 220, pp. 222–230, Sep. 2014.

[34] L. D. Miller, "An expression signature for p53 status in human breast cancer predicts mutation status, transcriptional effects, and patient survival," *Proc. Nat. Acad. Sci. USA*, vol. 102, no. 38, pp. 13550–13555, Sep. 2005.

[35] W. Liu, "Topologically inferring pathway activity toward precise cancer classification via integrating genomic and metabolomic data: Prostate cancer as a case," *Sci. Rep.*, vol. 5, Aug. 2015, Art. no. 13192.

[36] X. Zeng, S. Zhu, X. Liu, Y. Zhou, R. Nussinov, and F. Cheng, "DeepDR: A network-based deep learning approach to in silico drug repositioning," *Bioinformatics*, vol. 35, no. 24, pp. 5191–5198, Dec. 2019, doi: 10.1093/bioinformatics/btz418.

[37] W. Liu, W. Wang, G. Tian, W. Xie, L. Lei, J. Liu, W. Huang, L. Xu, and E. Li, "Topologically inferring pathway activity for precise survival outcome prediction: Breast cancer as a case," *Mol. BioSyst.*, vol. 13, no. 3, pp. 537–548, Jan. 2017.

[38] A. L. Tarca, S. Draghici, P. Khatri, S. S. Hassan, P. Mittal, J.-S. Kim, C. J. Kim, J. P. Kusanovic, and R. Romero, "A novel signaling pathway impact analysis," *Bioinformatics*, vol. 25, no. 1, pp. 75–82, Jan. 2009.

[39] R. Edgar, M. Domrachev, and A. E. Lash, "Gene expression omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Res.*, vol. 30, pp. 207–210, Jan. 2002.

[40] M. T. Landi, T. Dracheva, M. Rotunno, J. D. Figueroa, H. Liu, A. Dasgupta, F. E. Mann, J. Fukuoka, M. Hames, A. W. Bergen, S. E. Murphy, P. Yang, A. C. Pesatori, D. Consonni, P. A. Bertazzi, S. Wacholder, J. H. Shih, N. E. Caporaso, and J. Jen, "Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival," *PLoS ONE*, vol. 3, no. 2, p. e1651, Feb. 2008.

[41] M. D'Errico, E. D. Rinaldis, M. F. Blasi, V. Viti, M. Falchetti, A. Calcagnile, F. Sera, C. Saieva, L. Ottini, D. Palli, F. Palombo, A. Giuliani, and E. Dogliotti, "Genome-wide expression profile of sporadic gastric cancers with microsatellite instability," *Eur. J. Cancer*, vol. 45, no. 3, pp. 461–469, Feb. 2009.

[42] M. Tsuchiya, J. S. Parker, H. Kono, M. Matsuda, H. Fujii, and I. Rusyn, "Gene expression in nontumoral liver tissue and recurrence-free survival in hepatitis C virus-positive hepatocellular carcinoma," *Mol. Cancer*, vol. 9, no. 1, p. 74, 2010.

[43] G. L. Dalgliesh *et al.*, "Systematic sequencing of renal carcinoma reveals inactivation of histone modifying genes," *Nature*, vol. 463, no. 7279, pp. 360–363, Jan. 2010.

[44] J. Jones, "Gene signatures of progression and metastasis in renal cell cancer," *Clin. Cancer Res.*, vol. 11, no. 16, pp. 5730–5739, Aug. 2005.

[45] Y. Pawitan, "Gene expression profiling spares early breast cancer patients from adjuvant therapy: Derived and validated in two population-based cohorts," *Breast Cancer Res.*, vol. 7, no. 6, pp. R953–R964, 2005.

[46] L. M. Poisson, A. Sreekumar, A. M. Chinnaiyan, and D. Ghosh, "Pathway-directed weighted testing procedures for the integrative analysis of gene expression and metabolomic data," *Genomics*, vol. 99, no. 5, pp. 265–274, May 2012.

[47] M. Kanehisa and S. Goto, "KEGG: Kyoto encyclopedia of genes and genomes," *Nucleic Acids Res.*, vol. 28, no. 1, pp. 27–30, Jan. 2000.

[48] H. Yang, "Identification of secretory proteins in *mycobacterium tuberculosis* using pseudo amino acid composition," *Biomed. Res. Int.*, vol. 2016, Aug. 2016, Art. no. 5413903.

[49] H. Tang, W. Chen, and H. Lin, "Identification of immunoglobulins using Chou's pseudo amino acid composition with feature selection technique," *Mol. BioSyst.*, vol. 12, no. 4, pp. 1269–1275, Feb. 2016.

[50] H. Yang, "A comparison and assessment of computational method for identifying recombination hotspots in *Saccharomyces cerevisiae*," *Brief Bioinf.*, Oct. 2019, Art. no. bbz123.

[51] H. Lv, Z. M. Zhang, S. H. Li, J. X. Tan, W. Chen, and H. Lin, "Evaluation of different computational methods on 5-methylcytosine sites identification," *Briefings Bioinf.*, Jun. 2019, Art. no. bbz048.

[52] L. Wei, P. Xing, J. Zeng, J. Chen, R. Su, and F. Guo, "Improved prediction of protein-protein interactions using novel negative samples, features, and an ensemble classifier," *Artif. Intell. Med.*, vol. 83, pp. 67–74, Nov. 2017.

[53] L. Wei, S. Wan, J. Guo, and K. K. L. Wong, "A novel hierarchical selective ensemble classifier with bioinformatics application," *Artif. Intell. Med.*, vol. 83, pp. 82–90, Nov. 2017.

[54] Y. Shen, J. Tang, and F. Guo, "Identification of protein subcellular localization via integrating evolutionary and physicochemical information into Chou's general PseAAC," *J. Theor. Biol.*, vol. 462, pp. 230–239, Feb. 2019.

[55] Y. Ding, J. Tang, and F. Guo, "Identification of drug-side effect association via multiple information integration with centered kernel alignment," *Neurocomputing*, vol. 325, pp. 211–224, Jan. 2019.

[56] L. Xu, G. Liang, C. Liao, G.-D. Chen, and C.-C. Chang, "k-Skip-n-Gram-RF: A random forest based method for Alzheimer's disease protein identification," *Frontiers Genet., Original Res.*, vol. 10, no. 33, Feb. 2019, pp. 1–7.

[57] L. Xu, G. Liang, S. Shi, and C. Liao, "SeqSVM: A sequence-based support vector machine method for identifying antioxidant proteins," *Int. J. Mol. Sci.*, vol. 19, no. 6, p. 1773, Jun. 2018.

[58] J.-X. Tan, S.-H. Li, Z.-M. Zhang, C.-X. Chen, W. Chen, H. Tang, and H. Lin, "Identification of hormone binding proteins based on machine learning methods," *Math. Biosci. Eng.*, vol. 16, no. 4, pp. 2466–2480, 2019.

[59] T. Wang, H. Luo, W. Jia, A. Liu, and M. Xie, "MTES: An intelligent trust evaluation scheme in sensor-cloud-enabled industrial Internet of Things," *IEEE Trans. Ind. Informat.*, vol. 16, no. 3, pp. 2054–2062, Mar. 2020.

[60] K.-Q. Liu, Z.-P. Liu, J.-K. Hao, L. Chen, and X.-M. Zhao, "Identifying dysregulated pathways in cancers from pathway interaction networks," *BMC Bioinf.*, vol. 13, no. 1, p. 126, 2012.

[61] T. Song, A. Rodriguez-Paton, P. Zheng, and X. Zeng, "Spiking neural P systems with colored spikes," *IEEE Trans. Cogn. Develop. Syst.*, vol. 10, no. 4, pp. 1106–1115, Dec. 2018.

[62] F. G. C. Cabarle, H. N. Adorna, M. Jiang, and X. Zeng, "Spiking neural P systems with scheduled synapses," *IEEE Trans. Nanobiosci.*, vol. 16, no. 8, pp. 792–801, Dec. 2017.

[63] X. Chen, M. J. Pérez-Jiménez, L. Valencia-Cabrera, B. Wang, and X. Zeng, "Computing with viruses," *Theor. Comput. Sci.*, vol. 623, pp. 146–159, Apr. 2016.

[64] L. Wei, M. Liao, Y. Gao, R. Ji, Z. He, and Q. Zou, "Improved and promising identification of human MicroRNAs by incorporating a high-quality negative set," *IEEE/ACM Trans. Comput. Biol. Bioinf.*, vol. 11, no. 1, pp. 192–201, Jan. 2014.

[65] L. Xu, G. Liang, C. Liao, G.-D. Chen, and C.-C. Chang, "An efficient classifier for Alzheimer's disease genes identification," *Molecules*, vol. 23, no. 12, p. 3140, Nov. 2018.

[66] X.-X. He, A.-Y. Guo, C.-R. Xu, Y. Chang, G.-Y. Xiang, J. Gong, Z.-L. Dan, D.-A. Tian, J.-Z. Liao, and J.-S. Lin, "Bioinformatics analysis identifies miR-221 as a core regulator in hepatocellular carcinoma and its silencing suppresses tumor properties," *Oncol. Rep.*, vol. 32, no. 3, pp. 1200–1210, Sep. 2014.

[67] B. Liu, C. Li, and K. Yan, "DeepSVM-fold: Protein fold recognition by combining Support Vector Machines and pairwise sequence similarity scores generated by deep learning networks," *Briefings Bioinf.*, to be published, doi: 10.1093/bib/bbz098.

[68] H. Xu, W. Zeng, X. Zeng, and G. G. Yen, "An evolutionary algorithm based on minkowski distance for many-objective optimization," *IEEE Trans. Cybern.*, vol. 49, no. 11, pp. 3968–3979, Nov. 2019.

[69] X. Zeng, W. Wang, C. Chen, and G. G. Yen, "A consensus community-based particle swarm optimization for dynamic community detection," *IEEE Trans. Cybern.*, to be published, doi: 10.1109/tcyb.2019.2938895.

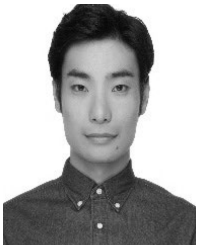
[70] L. Xu, G. Liang, L. Wang, and C. Liao, "A novel hybrid sequence-based model for identifying anticancer peptides," *Genes*, vol. 9, no. 3, p. 158, Mar. 2018.

[71] H. Ding, W. Yang, H. Tang, P.-M. Feng, J. Huang, W. Chen, and H. Lin, "PHYPred: A tool for identifying bacteriophage enzymes and hydrolases," *Virol. Sinica*, vol. 31, no. 4, pp. 350–352, Aug. 2016.

[72] B. Liu, Y. Zhu, and K. Yan, "Fold-LTR-TCP: Protein fold recognition based on triadic closure principle," *Briefings Bioinf.*, to be published, doi: 10.1093/bib/bbz139.

[73] B. Liu and Y. Zhu, "ProtDec-LTR3.0: Protein remote homology detection by incorporating profile-based features into learning to rank," *IEEE Access*, vol. 7, pp. 102499–102507, 2019.

PENG XU received the Ph.D. degree in biomedical engineering from Southeast University. He currently holds a postdoctoral position at the Institute of Computing Science and Technology, Guangzhou University, China. He is also a volunteer Teacher for supporting Qiannan Normal University for Nationalities, China. His research interests include bioinformatics and data mining.



GENG ZHAO received the M.Sc. degree in data mining and pattern recognition from Wenzhou University. He is an Artificial Intelligent Expert with Huawei. His research interests mainly include machine learning algorithms and nature language processing.

ZHENG KOU received the Ph.D. degree in bioinformatics from the Huazhong University of Science and Technology. He is currently a Professor with the Institute of Computing Science and Technology, Guangzhou University, China. His research interests include bio-inspired computing, machine learning, and bioinformatics.

GANG FANG received the Ph.D. degree in system analysis and integration from the Huazhong University of Science and Technology. He is currently a Professor with the Institute of Computing Science and Technology, Guangzhou University, China. His research interests include bio-inspired computing, machine learning, and bioinformatics.

WENBIN LIU received the Ph.D. degree in systems engineering from the Huazhong University of Science and Technology, in 2004. He is currently a Professor with Guangzhou University and Wenzhou University. In 2004, he joined the College of Mathematics, Physics, and Electronic Information Engineering with Wenzhou University. In 2007, he was a Visiting Scholar with the Institute for Systems Biology and Texas A&M University, in 2006 and 2013, respectively. His work was supported by four NSFC grants and other funds from Zhejiang province. His research interests include bioinformatics, pattern recognition, and data mining.