Predicting distant metastasis in breast cancer using ensemble classifier based on context-specific miRNA regulation modules

Xionghui Zhou[1], Juan Liu [1,*], Jianghui Xiong[2]
[1]School of Computer, Wuhan University, Wuhan, P.R. China
[2]Bioinformatics Group and Data Coordination Center, State Key Lab of Space Medicine Fundamentals and Application, China Astronaut Research and Training Center, Beijing, P.R. China
[*]Correspondence should be addressed to J.L. (liujuan@whu.edu.cn).

Abstract—***Many methods based on building classifiers by selecting gene markers have been proposed to predict breast cancer patient's outcome. However, most of them suffer from the problem of poor robustness, which are mainly due to the fact that the overlap degree of gene markers derived by different methods is not high, leading that few of them are generalized and can be widely used for clinical practice. In this paper, we present a method based on context-specific miRNA regulation modules to predict distant metastasis in breast cancer. First, we describe the regulation activity of a miRNA on a specific context by using CoMi (Context-specific miRNA activity) score, based on which, several miRNAs regulate on the same context are regarded as a miRNA regulation module; then the discriminate regulation modules are selected and each is used to construct a classification model separately; finally, an ensemble classifier is established by combining all the models with a majority voting strategy. The evaluation experiment results show that our method performs better than previous works. In addition, the obtained discriminate modules show great stability across different data sets (with p-value of 1.119e-06).***

Keywords-Cancer prognosis; ensemble classifier; context specific miRNA activity; miRNA regulation network

I. INTRODUCTION

For cancer patients, it's important to evaluate an individual patient's risk of disease recurrence to ensure that he/she receives appropriate therapy. As to breast cancer, many signatures derived from gene expression profiles have been reported to can classify the cancer patients into different risk groups [1-4]. However, there are two main problems about these signatures. Firstly, the performance of classifier constructed by signatures obtained from one data set usually declines significantly in other datasets; secondly, there is little overlap among signatures from different data sets, which makes the signatures less convincible to physicians. For example, researches [3] and [2] separately identified a 70-gene signature and 76-gene signature from mRNA expression profiles to classify breast cancer samples into good-outcome and bad-outcome groups with accuracies between 60%-70% [5], but there are only one gene common in these two sets. Moreover, both of them have an accuracy of less than 55% in commutative dataset [6]. The reasons may be that tumor cells often have more 'passenger signals' than other types of cells, resulting in the most gene signatures are 'passengers' instead of 'drivers' [4]. Confronted with this situation, alternative features such as protein-protein interaction networks, pathways and GO terms have been used to predict cancer outcome [6-9], but the performances are still unsatisfactory, with an AUC less than 0.7 on independent datasets [7].

In our previous work [10], based on the hypothesis that miRNAs may be more stable than genes in cancer prognosis process, a new concept, Context-specific miRNA activity (CoMi activity), is introduced to predict breast cancer patients' outcome. The CoMi activity is described by CoMi score which is calculated by computing the statistical difference between the expression level of a miRNA's target genes and non-targets gene within a given gene set (context). The support vector machine (SVM) classifier constructed with the CoMi activity features has been proved to be more stable and has better performance than those with gene signatures [10].

In this work, with the hypothesis that several miRNAs may influence the outcome of cancer patients by regulating a certain biology process, we set the sub-networks of the CoMi activity network as modules (several miRNAs regulating one GO Term), and each module is used to construct a classifier to predict the outcome of breast cancer patients, and all the module classifiers are combined to an ensemble classifier by voting strategy. In order to validate the performance of our method, both five-fold cross validation and strict independent tests on other breast cancer datasets are taken to compare with other methods. In the meanwhile, the stability and the biological meaning of the discriminative sub-network are also investigated.

II. METHODS

The work flow of our method is shown in figure 1. First, for each miRNA, we compute its regulation activity on a specific context by CoMi score (we use each GO Biology Process term as a context in this work); then, based on the CoMi scores, each GO Term and its regulating miRNAs are grouped as a regulation module, a sub-network with star style; and then each selected module (no less than five miRNAs regulating one GO Term in this work) is used to construct a weak classifier to predict the patient's outcome; based on the prediction performance, some weak classifiers with distinguishing abilities (with an AUC no less than a threshold) are combined as the ensemble classifier with voting strategy.

978-1-4673-2560-8/12/$31.00 ©2012 IEEE

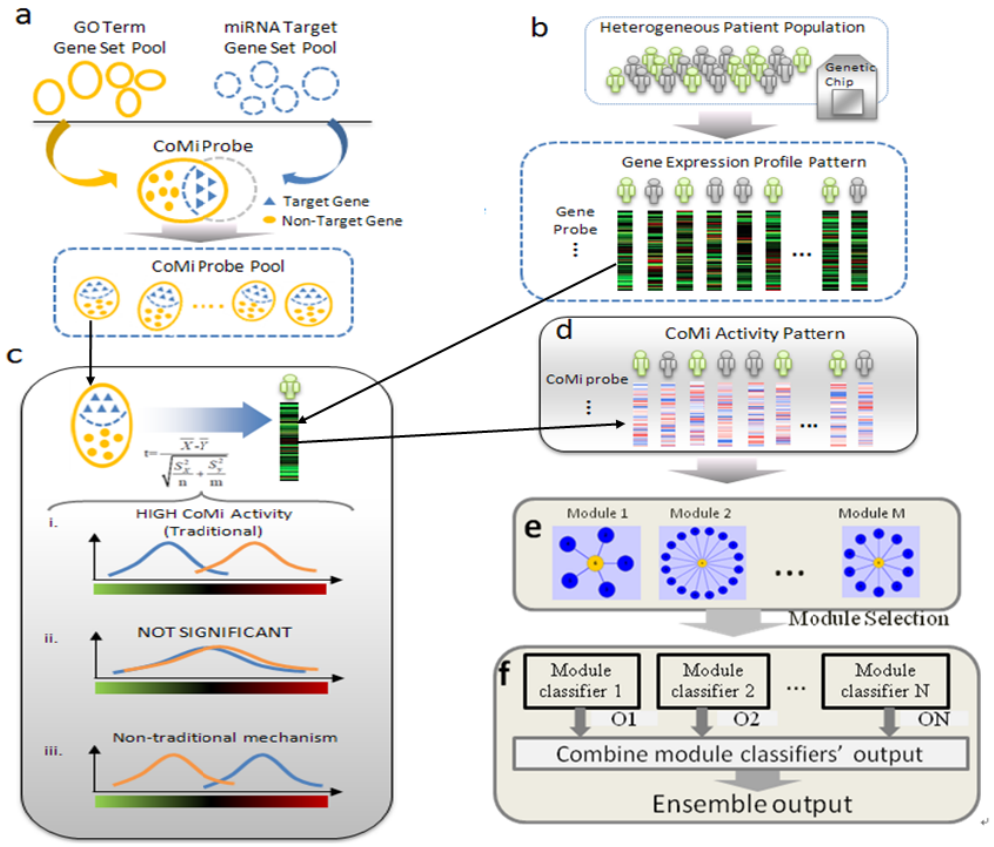

Figure 1. Workflow of our work

A. Computing CoMi activity of a miRNA on a context

In order to evaluate the regulating activity of a miRNA on a specific context (CoMi activity), we take the following steps as described in our previous work [10]:

(1) Collecting all miRNA target genes by using two miRNA target predicting tools: TargetScan [11] and RNA22 [12]. There are totally 680 different miRNAs in these two tools. We combined all the predicted targets by the union operation (the union set of the two tools are more suitable for CoMi calculating than other target sets [10]), and for each miRNA, we grouped all its targets into one set.

(2) Taking the category "Biological Process" in Gene Ontology into considered, we organize the genes in the given gene list (in this work, the gene list contains all the genes in the gene expression data) into different groups according to their GO term annotations (hereinafter, we shorten the Gene Ontology "Biological Process" term as GOBP term). All genes associated with one GOBP term form one GOBP gene set.

(3) For each miRNA, we can divide all genes belonging to one GOBP into targets and non-targets by intersecting the miRNA's targets set and the GOBP gene set. By this way, we can get many miRNA-GOBP pairs, each standing for the miRNA and its regulating GOBP term. (Fig. 1a). To ensure the statistical significance of the regulation relationship of a miRNA on a specific GOBP term, we have done two-step preprocessing on the miRNA-GOBP pairs. First, we filtered out all miRNA-GOBP pairs where the number of miRNA targets is less than 10 (setting by experience); second, for each retained miRNA-GOBP pair, we calculated the significance of the miRNA target set by using hypergeometric cumulative distribution function, and only the miRNA-GOBP pairs with p-value smaller than 0.05 were considered.

(4) As in our earlier work [10], for each miRNA-GOBP pair, we define the CoMi activity of the miRNA on the specific GOBP context as the statistical difference by t-test between the gene expression levels of targets and non-targets. The t-score is used to evaluate the regulating activity of the miRNA on a specific context (GOBP term). (Fig. 1b, Fig1c).

Using the CoMi activity calculating method outlined above, we can get the CoMi activity patterns for all patients, which can be denoted as a data matrix, where columns represent patients and rows stand for the miRNA-GOBP pairs, and the matrix elements stand for the CoMi activity scores (Fig. 1d).

It should be noted that in miRNA-GOBP pairs may contain noises or be redundant. In order to reduce the noise, we filtered out 10% miRNA-GOBP pairs according the CoMi activity scores' variances across all the samples in GSE2034 (the miRNA-GOBP pairs with the 10% smallest variances was abandon); If two pairs have identical CoMi activity patterns in all samples, we only kept one pair to eliminate redundancy (This redundancy may be produced by the miRNA target prediction tools, for some miRNAs have the similar sequences, thus have same or similar target sets according to the prediction tools).

B. Module evaluation and ensemble classifier construction

The CoMi activity describes the activity of miRNA on a specific GOBP term. Since there may be several miRNAs to have regulation activity on a specific GOBP term, we call every GOBP term along with all of its regulating miRNAs as a module (Fig. 1e, only those with no less than five miRNAs are considered in this work). Since every module corresponds to one GOBP term, we simply name the module with its associated GOBP term.

With the hypothesis that several miRNAs usually co-regulate on a specific biological process, thus influence the breast cancer outcome, we think the single module should have the ability to distinguish the outcomes of different patients to some extent. Thus we evaluate the discrimination ability of each module by constructing a weak classifier, and only those with AUC no less than a threshold were considered to have the discrimination abilities and to be retained for further purpose (Fig. 1f).

The centroid classifier has been shown to perform well in analyzing microarray data where the feature number is usually much greater than the sample number [13]. Most importantly, the centroid classifier needs no tuning and is less overfitting while has similar or better performance than several other classifiers (such as SVM) [7]. So the centroid method was used to construct the weak classifier in this work.

The centriod classifier calculates the centriod of each class (A vector of mean expression levels of the features in each group), then each sample are classified into its nearest class by comparing its expression with the two centroids [7]. For each module, we used all the CoMi activity scores as features to construct module classifier and perform 5-fold cross validation on Wang dataset (GSE2034) [2], only the modules with AUC no less than a threshold were regarded to have discrimination ability and were selected.

A module classifier can only describe one aspect of the biology process in cancer metastasis, and all the discriminative modules together discover more comprehensive biology mechanism in cancer prognosis. Therefore, we established an ensemble classifier by combining all module classifiers, with the simple voting strategy (Fig. 1f). Finally, the ensemble classifier is used to predict the distinct metastasis of breast cancer patients.

III. EXPERIMENTS

The motivation of this work is to find robust and biological meaningful modules across different data sets, as well as constructing a high performance classifier to predict distant metastasis in breast cancer. Therefore, we first compared the classification performance of the proposed ensemble classifier with other representative methods by using five-fold cross validation in GSE2034 and independent tests on the other four data sets; and then we investigated the stability of discriminate modules by investigating the overlap of discriminate modules selected from different data sets and did case studies on their biological meanings to see whether they can reveal some metastasis mechanism of breast cancer.

A. Datasets and preprocessing

We used five breast cancer datasets from NCBI GEO to evaluate our method: GSE2034 (Wang dataset) [4], GSE7390 [14], GSE11121 [15], GSE4922 [16] and GSE6532 [17]. All the data sets are based on Affymetrix HG-U133A microarray, and normalized by the original authors, using MAS5, except GSE4922, which using RMA. Among the five datasets, the first four contains both ER+ and ER- cancer patients, while the last one contains only ER+ patients. In addition, the patients in the first three datasets are lymph-node-negative and the patients in last two are lymph-node-negative or lymph-node-positive. All the five datasets, listed in Table 1, are divided into two groups, the good outcome and the poor outcome, according that the distant metastasis occurs five years after or not. Patients who are censored within five years or received adjuvant treatment were removed from the data.

TABLE I. BREAST CANCER DATASETS

Data	Bad outcome	Good outcome	Total	Removed samples
GSE2034	93	183	276	10
GSE4922	30	103	133	156
GSE6532	23	77	100	227
GSE7390	36	154	190	8
GSE11121	28	154	182	18

Samples were removed if they are censored before the 5-year or received adjuvant treatment.

B. Representative classifiers for comparison purpose

We selected three representative methods for breast cancer distant metastasis prediction in this work: the 76-gene classifier [2], the 70-gene classifier [3] and the gene set statistics classifier [7]. The 70-gene and 76-gene classifiers are the most famous methods used to predict breast cancer distant metastasis, and the gene set statistics classifier has been reported to have the same prediction performance, yet be more stable than previous works [7].

The 76-gene classifier is an ER specific classification method proposed by Wang *et al.* [2]. For each sample to be classified, a relapse score is calculated by using the weighted linear combination of the 76 genes' expression levels, and then the sample is assigned to one of two outcome groups according to its relapse score.

The 70-gene classifier proposed by van de Vijver *et al.* [3] is similar to the centroid classifier. In which, the centroid value for each class is first calculated based on the 70 gene signatures, then the samples are classified into poor outcome groups and good outcome groups according to their Pearson correlations with the centroids of the positive and negative metastasis classes.

In gene set statistics classifier [7], the pre-specified gene sets were downloaded from the Molecular Signatures Database (MSigDB) [18] and were evaluated by a statistical method, by which, the optimal feature set was derived from

the gene sets and were used to construct the centroid classifier for predicting breast cancer distinct metastasis within 5 years. There are several statistics methods used to evaluate the gene sets in the original publication [7], including Set Centroid (mean), Set Median, PCA and t-test. In the experiments, we use Set Centroid and Set median statistics methods, for they are reported to perform better than other methods [7]. In order to choose the optimal feature set, the features were first ranked by the distances between the two centroids of the two classes, and then the centroid classifiers were constructed based on the top ranked features with number ranging from 1 to 200 and evaluated respectively, the feature set with the highest cross validation AUC was selected as the optimal feature set for the classifier construction.

C. Performance measures

Because the poor outcome patients and good outcome patients in the dataset are seriously unbalanced (For example, there are 154 good outcome patients, but only 36 bad outcome patients in GSE7390), so Matthews Correlation Coefficient (MCC), which is the most informative measure method when the distribution of the classes in a dataset is highly unbalanced [19], was applied as the main evaluation standard in our study.

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}}$$

Where TP is true positive, TN is true negative, FP is false positive and FN is false negative.

The receiver operating character (ROC) curve is a plot of the sensitivity versus (1-specifivity) for a binary classifier at varying threshold from 0 to 1 (the threshold is the probability of sample belongs to bad outcome group). Area under the receiver operating characteristic curve (AUC) is a popular measure method for binary classifier, so it was used as the other important measure standard in our work.

In addition, sensitivity (SN), specificity (SP) and accuracy (ACC) were also including in our work.

IV. RESULTS AND DISCUSSION

A. Performances of our ensemble classifier

In our work, we use GSE2034 as the training set to construct the classifier. From all of the 14122 miRNA-GOBP pairs, 347 candidate modules with no less than five miRNAs are obtained. Among of them, 55 modules are selected as discriminative modules according to the criterion that AUC is not less than 0.60, which was set based on our experiment results (data not shown).

We then constructed the ensemble classifier based on the 55 selected modules. To evaluate the performance of the classifier, we did ten times five-fold cross validation on the training set GSE2034. In the meanwhile, we also did independent tests on GSE7390, GSE11121, GSE4922 and GSE6532 respectively. The results are listed in table 2.

From this table, we can see that our ensemble classifier performs well and stably both on the training and the independent test sets. Most importantly, the AUCs of our classifier on most test sets are higher than 0.70, whereas previous genes or gene sets methods, to our best knowledge, can only reach AUC on independent tests less than 0.70 [9, 24].

TABLE II. THE CLASSIFICATION PERFORMANCE OF ENSEMBLE CLASSIFIER

	ACC	SN	SP	AUC	MCC
GSE2034	0.67	0.67	0.66	0.74	0.32
GSE7390	0.64	0.62	0.73	0.74	0.28
GSE11121	0.78	0.83	0.50	0.71	0.29
GSE4922	0.76	0.69	0.88	0.69	0.24
GSE6532	0.73	0.80	0.51	0.75	0.29

B. Comparing results with other methods

The AUC and MCC scores of our ensemble classifier and other four classifiers (There are two version of gene set statistics classifier as described above: Set-Median and Set-centroid) on GSE2034 and other four independent test sets are shown in Figure 2 and Figure 3 respectively (the performance details of the four other methods are not shown).

From figure 2 we can see that Set-Median and Set-centroid methods based on gene sets can achieve better AUC performances than two gene signature classifiers, which is consistent with the original report [7]. While our ensemble classifier is more robust than the others, and achieves the best AUC performances on most data sets, except that it performs not as well as two gene sets methods on GSE11121 data set.

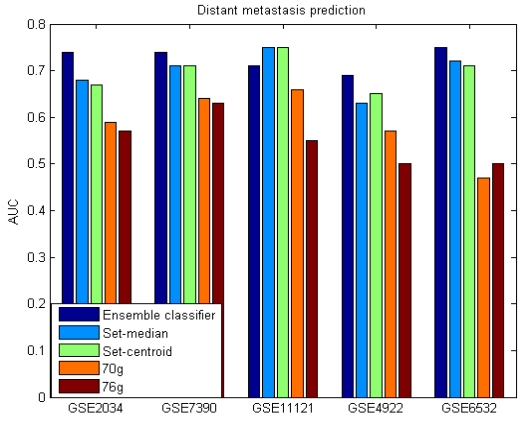

Figure 2. The AUC of the five classifiers on the five datasets.

Since all the four independent datasets are all seriously unbalanced, MCC is more suitable to evaluate the classifier's ability of handling unbalanced data sets, as described in section III. The MCC comparing results shown in figure 3, from which we can see that the other four classifiers, especially the gene signature classifiers, perform worse on the last two datasets than on the first three datasets.

This may due to the fact that last two datasets contain lymph-node-negative and lymph-node-positive patients, while the first three contains only lymph-node-negative patients. However, our ensemble classifier shows a robust performance in these two datasets. Even in GSE4922, our classifier reaches the AUC of 0.69 (MCC of 0.24), which is much better than the other methods. This phenomenon demonstrates that our method is less sensitive to datasets and is actually very stable, whereas other four classifiers vary dramatically on different datasets.

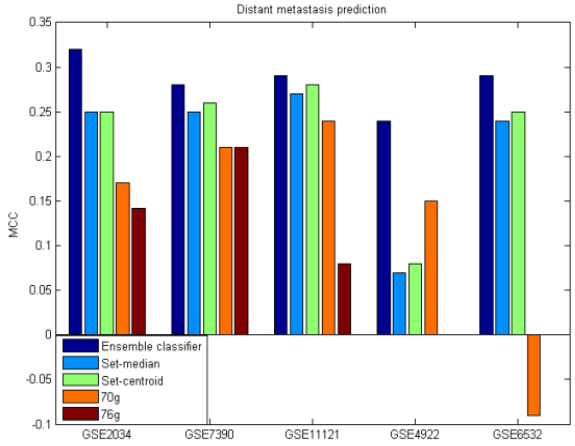

Figure 3. The MCC of the five classifiers on the five datasets.

In all, we can draw a conclusion that our ensemble classifier is super to other methods, not only in the performance, but also in the stability on different data sets.

C. *The stability of the discriminative modules*

As described before, the main challenge of cancer prognosis study is that signatures derived from different datasets are not stable. For an example, there are only one common gene in 76-gene and 70-gene signatures [6], which make the classifiers lack of robustness and less convincible to the clinicians.

Different with previous researches, we consider the GOBP and its regulating miRNAs as a whole by the module. One module can reflect one aspect of the biological process of the cancer, thus should be stable in different datasets.

We have already filtered 55 out of 347 modules based on GS2034 data set. To check out the stability of our selected modules, we first merged all other four data sets into one (the merged data set) to ensure that there are enough samples in both outcome groups; then we did the same process as in GS2034 to select out 97 discriminative modules. By investigating the discriminative modules from GS2034 and the merged data sets, we found that there are 31 modules in common (54.39% in GS2034, and 31.96% in the merged dataset) with p-value 1.119e-06 (hypergeometric cumulative distribution function). That is to say, the discriminative modules obtained from different datasets are great stable by our method, thus can be used in different data sets.

D. *Case studies of the biology meaning of the modules*

One advantage of our CoMi activity is that it can not only discover the role of miRNA, but also reveal which biology process the miRNA take part in. Thus we have checked whether the selected modules can uncover some hidden biological mechanism in cancer metastasis and found that many selected modules are actually cancer-related. Some prognosis markers related GO terms reported by other works [4], such as cell death, cell adhesion, DNA repair are also found in our modules. Moreover, there are some novel biology processes rarely reported in literature are also discovered by our modules.

For an example, in the “autophagy” module found by our method (fig. 4), we found that all miRNAs are cancer related by literature works. hsa-miR-34a is known to be a famous tumor suppressor [20], hsa-miR-34b is reported to be significantly associated with the presence of breast cancer’s metastases [20], hsa-miR-9 is reported to be related to vascular invasion and lymph node metastasis in breast cancer [21], hsa-miR-373 is a activator of metastasis of breast cancer [21], and hsa-miR-503 seems to be a putative tumor suppressor [22]. The biological process “autophagy” is becoming a hot topic these days, for it is a cellular degradation pathway for the clearance of damaged or superfluous proteins and organelles. It has relation with cancer by providing a protective function to limit tumor necrosis and inflammation, and to mitigate genome damage in tumor cells in response to metabolic stress [23]. This discriminative module may indicate that five cancer related miRNAs may work together on autophagy to influence the metastasis risk of breast cancer patients. It is noteworthy that hsa-miR-34 family (hsa-miR-34a, hsa-miR-34c) and hsa-miR-9* has been reported to involve in the biology process autophagy [24], which also proves the validity of our method.

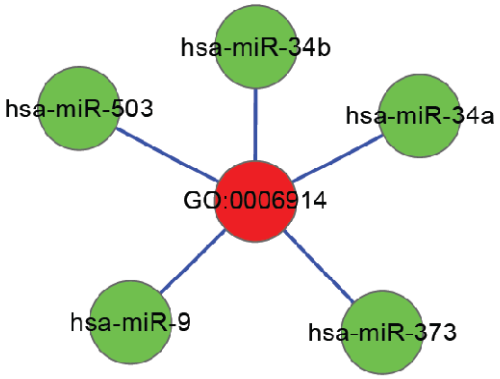

Figure 4. Module ‘autophagy’(GO:0006914)

It is interesting that the CoMi scores of any single miRNA-GOBP pairs in most of the modules are not significantly different in different groups (data not shown) while the modules collecting several pairs together do show significant discriminating abilities. This phenomenon demonstrates that our method focuses on selecting features

which cooperate together to be relevant to specific biological process instead of selecting features which are singly correlated with the cancer outcomes, thus it may select the "drivers" rather than the "passengers" of the biological process, so that the constructed classifier show great robustness across different data sets.

V. CONCLUSION

With the hypothesis that several miRNAs regulating on a biology process may co-influence the outcome of breast cancer, we proposed to detect context specific miRNA regulation modules to construct the ensemble classifier for the prediction of distant metastasis in breast cancer in this paper. Based on the GSE2034 data set, 55 out of 347 modules were selected as discriminative modules, each was used to construct a module classifier respectively, and all module classifiers were used to establish an ensemble classifier based on majority voting strategy. The comparing experiment results show that the ensemble classifier performs better than the representative methods. In addition to this, the selected modules are very stable across different data sets, which leads that the classifier is very robust and can be used for practical purpose. The biology analysis also demonstrates that the selected modules can discover hidden mechanism in breast cancer's metastasis.

ACKNOWLEDGMENT

This work was supported by the grants from the National Science Foundation of China (60970063), the program for New Century Excellent Talents in Universities (NCET-10-0644), the Ph.D. Programs Foundation of Ministry of Education of China (20090141110026) and the Fundamental Research Funds for the Central Universities (6081007).

REFERENCES

[1] T. Sorlie, C. M. Perou, R. Tibshirani, T. Aas, S. Geisler, H. Johnsen, et al, "Gene expression patterns of breast carcinomas distinguish tumor subclasses with clinical implications," *Proc Natl Acad Sci U S A,* vol. 98, pp. 10869-74, Sep 11 2001.

[2] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, et al, "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *Lancet,* vol. 365, pp. 671-9, Feb 19-25 2005.

[3] M. J. van de Vijver, Y. D. He, L. J. van't Veer, H. Dai, A. A. Hart, D. W. Voskuil, et al, "A gene-expression signature as a predictor of survival in breast cancer," *N Engl J Med,* vol. 347, pp. 1999-2009, Dec 19 2002.

[4] J. Li, A. E. Lenferink, Y. Deng, C. Collins, Q. Cui, E. O. Purisima,et al, "Identification of high-quality cancer prognostic markers and metastasis network modules," *Nat Commun,* vol. 1, p. 34, 2010.

[5] 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, p. 140, 2007.

[6] W. K. Lim, E. Lyashenko, and A. Califano, "Master regulators used as breast cancer metastasis classifier," *Pac Symp Biocomput,* pp. 504-15, 2009.

[7] G. Abraham, A. Kowalczyk, S. Loi, I. Haviv, and J. Zobel, "Prediction of breast cancer prognosis using gene set statistics provides signature stability and biological context," *BMC Bioinformatics,* vol. 11, p. 277, 2010.

[8] M. H. van Vliet, C. N. Klijn, L. F. Wessels, and M. J. Reinders, "Module-based outcome prediction using breast cancer compendia," *PLoS One,* vol. 2, p. e1047, 2007.

[9] I. W. Taylor, R. Linding, D. Warde-Farley, Y. Liu, C. Pesquita, D. Faria, et al, "Dynamic modularity in protein interaction networks predicts breast cancer outcome," *Nat Biotechnol,* vol. 27, pp. 199-204, Feb 2009.

[10] X.H. Zhou, J. Liu, C.N. Liu, Rayner S., F.J. Liang, J.F. Ju, et al, "Context-Specific miRNA Regulation Network Predicts Cancer Prognosis," in *Proceeding of IEEE international conference on system biology*, 2011, pp. 225-243.

[11] A. Grimson, K. K. Farh, W. K. Johnston, P. Garrett-Engele, L. P. Lim, and D. P. Bartel, "MicroRNA targeting specificity in mammals: determinants beyond seed pairing," *Mol Cell,* vol. 27, pp. 91-105, Jul 6 2007.

[12] K. C. Miranda, T. Huynh, Y. Tay, Y. S. Ang, W. L. Tam, A. M. Thomson,et al, "A pattern-based method for the identification of MicroRNA binding sites and their corresponding heteroduplexes," *Cell,* vol. 126, pp. 1203-17, Sep 22 2006.

[13] C. S. a. A. K. J. Bedo, "An Efficient Alternative to SVM Based Recursive Feature Elimination with Applications in Natural Language Processing and Bioinformatics " *Lecture Notes in Computer Science,* vol. 4304/2006, pp. 170-180, 2006.

[14] C. Desmedt, F. Piette, S. Loi, Y. Wang, F. Lallemand, B. Haibe-Kains, et al, "Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series," *Clin Cancer Res,* vol. 13, pp. 3207-14, Jun 1 2007.

[15] M. Schmidt, D. Bohm, C. von Torne, E. Steiner, A. Puhl, H. Pilch, H. A. Lehr, J. G. Hengstler, H. Kolbl, and M. Gehrmann, "The humoral immune system has a key prognostic impact in node-negative breast cancer," *Cancer Res,* vol. 68, pp. 5405-13, Jul 1 2008.

[16] A. V. Ivshina, J. George, O. Senko, B. Mow, T. C. Putti, J. Smeds, T. Lindahl, et al, "Genetic reclassification of histologic grade delineates new clinical subtypes of breast cancer," *Cancer Res,* vol. 66, pp. 10292-301, Nov 1 2006.

[17] S. Loi, B. Haibe-Kains, C. Desmedt, F. Lallemand, A. M. Tutt, C. Gillet, et al, "Definition of clinically distinct molecular subtypes in estrogen receptor-positive breast carcinomas through genomic grade," *J Clin Oncol,* vol. 25, pp. 1239-46, Apr 1 2007.

[18] A. Subramanian, P. Tamayo, V. K. Mootha, S. Mukherjee, B. L. Ebert, M. A. Gillette, et al, "Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles," *Proc Natl Acad Sci U S A,* vol. 102, pp. 15545-50, Oct 25 2005.

[19] L. Shi, G. Campbell, W. D. Jones, F. Campagne, Z. Wen, S. J. Walker, et al, "The MicroArray Quality Control (MAQC)-II study of common practices for the development and validation of microarray-based predictive models," *Nat Biotechnol,* vol. 28, pp. 827-38, Aug 2010.

[20] Y. Li, D. Kong, Z. Wang, and F. H. Sarkar, "Regulation of microRNAs by natural agents: an emerging field in chemoprevention and chemotherapy research," *Pharm Res,* vol. 27, pp. 1027-41, Jun 2010.

[21] M. S. Nicoloso, R. Spizzo, M. Shimizu, S. Rossi, and G. A. Calin, "MicroRNAs--the micro steering wheel of tumour metastases," *Nat Rev Cancer,* vol. 9, pp. 293-302, Apr 2009.

[22] Q. Jiang, M. G. Feng, and Y. Y. Mo, "Systematic validation of predicted microRNAs for cyclin D1," *BMC Cancer,* vol. 9, p. 194, 2009.

[23] R. Mathew, V. Karantza-Wadsworth, and E. White, "Role of autophagy in cancer," *Nat Rev Cancer,* vol. 7, pp. 961-7, Dec 2007.

[24] L. L. Fu, X. Wen, J. K. Bao, and B. Liu, "MicroRNA-modulated autophagic signaling networks in cancer," *Int J Biochem Cell Biol,* vol. 44, pp. 733-6, May 2012.