Hindawi Publishing Corporation
Advances in Bioinformatics
Volume 2013, Article ID 618461, 8 pages
http://dx.doi.org/10.1155/2013/618461

Research Article

Identification of Robust Pathway Markers for Cancer through Rank-Based Pathway Activity Inference

Navadon Khunlertgit and Byung-Jun Yoon

Department of Electrical and Computer Engineering, Texas A&M University, College Station, TX 77843-3128, USA

Correspondence should be addressed to Byung-Jun Yoon; bjyoon@ece.tamu.edu

Received 30 November 2012; Accepted 19 January 2013

Academic Editor: Hazem Nounou

One important problem in translational genomics is the identification of reliable and reproducible markers that can be used to discriminate between different classes of a complex disease, such as cancer. The typical small sample setting makes the prediction of such markers very challenging, and various approaches have been proposed to address this problem. For example, it has been shown that pathway markers, which aggregate the gene activities in the same pathway, tend to be more robust than gene markers. Furthermore, the use of gene expression ranking has been demonstrated to be robust to batch effects and that it can lead to more interpretable results. In this paper, we propose an enhanced pathway activity inference method that uses gene ranking to predict the pathway activity in a probabilistic manner. The main focus of this work is on identifying robust pathway markers that can ultimately lead to robust classifiers with reproducible performance across datasets. Simulation results based on multiple breast cancer datasets show that the proposed inference method identifies better pathway markers that can predict breast cancer metastasis with higher accuracy. Moreover, the identified pathway markers can lead to better classifiers with more consistent classification performance across independent datasets.

1. Introduction

Advances in microarray and sequencing technologies have enabled the measurement of genome-wide expression profiles, which have spawned a large number of studies aiming to make accurate diagnosis and prognosis based on gene expression profiles [1–4]. For example, there has been significant amount of work on identifying markers and building classifiers that can be used to predict breast cancer metastasis [2, 4]. Many existing methods have directly employed gene expression data without any knowledge of the interrelations between genes. As a result, the predicted gene markers often lack interpretability and many of them are not reproducible in other independent datasets.

To overcome this problem, several different approaches have been proposed so far. For example, a recent work by Geman et al. [3] proposed an approach that utilizes the relative expression between genes, rather than their absolute expression values. It was shown that the resulting markers are easier to interpret, robust to chip-to-chip variations, and more reproducible across datasets. Another possible way to address the aforementioned problem is to interpret the gene expression data at a "modular" level through data integration [5–11]. These methods utilize additional data sources and prior knowledge—such as protein-protein interaction (PPI) data and pathway knowledge—to jointly analyze the expression of interrelated genes. This results in modular markers, such as *pathway markers* and *subnetwork markers*, which have been shown to improve the classification performance and also to be more reproducible across independent datasets [8–11]. In order to utilize pathway markers, we need to infer the pathway activity by integrating the gene expression data with pathway knowledge. For example, Guo et al. [6] used the mean or median expression value of the member genes (that belong to the same pathway) as the activity level of a given pathway. Recently, Su et al. [10] proposed a probabilistic pathway activity inference method that uses the log-likelihood ratio between different phenotypes based on the expression level of each member gene.

In this work, we propose an enhanced pathway activity inference method that utilizes the ranking of the member

genes to predict the pathway activity in a probabilistic manner. The immediate goal is to identify better pathway markers that are more reliable, more reproducible, and easier to interpret. Ultimately, we aim to utilize these markers to build accurate and robust disease classifiers. The proposed method is motivated by the relative gene expression analysis strategy proposed in [3, 12] and it builds on the concept of probabilistic pathway activity inference proposed in [10, 11]. In this study, we focus on predicting breast cancer metastasis and demonstrate that the proposed method outperforms existing methods. Preliminary results of this work have been originally presented in [13].

2. Materials and Methods

2.1. Study Datasets. Six independent breast cancer microarray gene expression datasets have been used in this study: GSE2034 (USA) [4], NKI295 (The Netherlands) [14], GSE7390 (Belgium) [15], GSE1456 (Stockholm) [16], GSE15852, and GSE9574. The Netherlands dataset uses a custom Agilent chip and it has been obtained from the Stanford website [17]. All datasets have been profiled using the Affymetrix U133a platform and they have been downloaded from the Gene Expression Omnibus (GEO) website [18].

The above datasets have been used in our study both with and without (re)normalization. To test the reproducibility of pathway markers, we selected the USA dataset and the Belgium dataset, both of which were obtained using the Affymetrix platform. The raw data for these two datasets have been normalized by utilizing the microarray preprocessing methods provided in the Bioconductor package [19]. We applied three popular normalization methods—RMA, GCRMA, and MAS5—with default setting.

The pathway data have been obtained from the MSigDB 3.0 Canonical Pathways [20]. This pathway dataset consists of 880 pathways, where 3,698 genes in these pathways intersect with all datasets.

2.2. Gene Ranking. In this study, we utilize "gene ranking" or the relative ordering of the genes based on their expression levels within each profile [3]. Consider a pathway that contains n member genes $\mathcal{G} = \{g_1, g_2, \ldots, g_n\}$ after removing the genes that are not included in all datasets. Given a sample $\mathbf{x}_k = \{x_k^1, x_k^2, \ldots, x_k^n\}$ that contains the expression level of the member genes, the gene ranking $\mathbf{r}_k$ is defined as follows:

$$\mathbf{r}_k = \left\{ r_k^{i,j} \mid 1 \le i < j \le n \right\}, \tag{1}$$

where

$$r_k^{i,j} = \begin{cases} 1, & \text{if } x_k^i < x_k^j, \\ 0, & \text{otherwise.} \end{cases} \tag{2}$$

The resulting gene ranking $\mathbf{r}_k$ is a binary vector representing the ordering of the member genes based on their expression values in the kth sample $\mathbf{x}_k$. To preserve the gene ranking in each sample, we do not employ any between-sample normalization.

2.3. Pathway Activity Inference Based on Gene Ranking. To infer the pathway activity, we follow the strategy proposed in [10], where the activity level a_k of a given pathway in the kth sample is predicted by aggregating the probabilistic evidence of all the member genes. The main difference between the strategy proposed in this work and the original strategy [10] is that we estimate the probabilistic evidence provided by each gene based on its ranking rather than its expression value. More specifically, the pathway activity level is given by

$$a_k = \sum_{1 \le i < j \le n} \lambda_{i,j}\left(r_k^{i,j}\right), \tag{3}$$

where $\lambda_{i,j}(r_k^{i,j})$ is the log-likelihood ratio (LLR) between the two phenotypes (i.e., class labels) for the ranking $\mathbf{r}_k$. The LLR $\lambda_{i,j}(r_k^{i,j})$ is defined as

$$\lambda_{i,j}\left(r_k^{i,j}\right) = \log\left[\frac{f_{i,j}^1\left(r_k^{i,j}\right)}{f_{i,j}^2\left(r_k^{i,j}\right)}\right], \tag{4}$$

where $f_{i,j}^1(r)$ is the conditional probability mass function (PMF) of the ranking of the expression level of gene g_i and gene g_j under phenotype 1 and $f_{i,j}^2(r)$ is the conditional PMF of the ranking of the expression level of gene g_i and gene g_j under phenotype 2.

In practice, the number of possible gene pairs $\binom{n}{2}$ may be too large when we have large pathways with many member genes (i.e., when n is large). To reduce the computational complexity, we prescreen the gene pairs based on the mutual information [21] as follows. For every gene pair (i, j), we first compute the mutual information between the ranking $r_k^{i,j}$ and the corresponding phenotype c_k. Then we select the top 10% gene pairs with the highest mutual information and use only these gene pairs for computing the pathway activity level defined in (3). Although we selected the top 10% gene pairs for simplicity, this may not be necessarily optimal and one may also think of other strategies for adaptively choosing this threshold.

In a practical setting, we may not have enough training data to reliably estimate the PMFs $f_{i,j}^1(r)$ and $f_{i,j}^2(r)$. For this reason, we normalize the original LLR $\lambda_{i,j}(r_k^{i,j})$ as follows to decrease its sensitivity to small alterations in gene ranking:

$$\widehat{\lambda}_{i,j}\left(r_k^{i,j}\right) = \frac{\lambda_{i,j}\left(r_k^{i,j}\right) - \mu\left(\lambda_{i,j}\right)}{\sigma\left(\lambda_{i,j}\right)}, \tag{5}$$

where $\mu(\lambda_{i,j})$ and $\sigma(\lambda_{i,j})$ are the mean and standard deviation of $\lambda_{i,j}(r_k^{i,j})$ across all $k = 1, \ldots, n$. Figure 1 illustrates the overall process.

2.4. Assessing the Discriminative Power of Pathway Markers. In order to assess the discriminative power of a pathway marker, we compute the t-test statistics score, which is given by

$$t(\mathbf{a}) = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1/K_1 + \sigma_2/K_2}}, \tag{6}$$

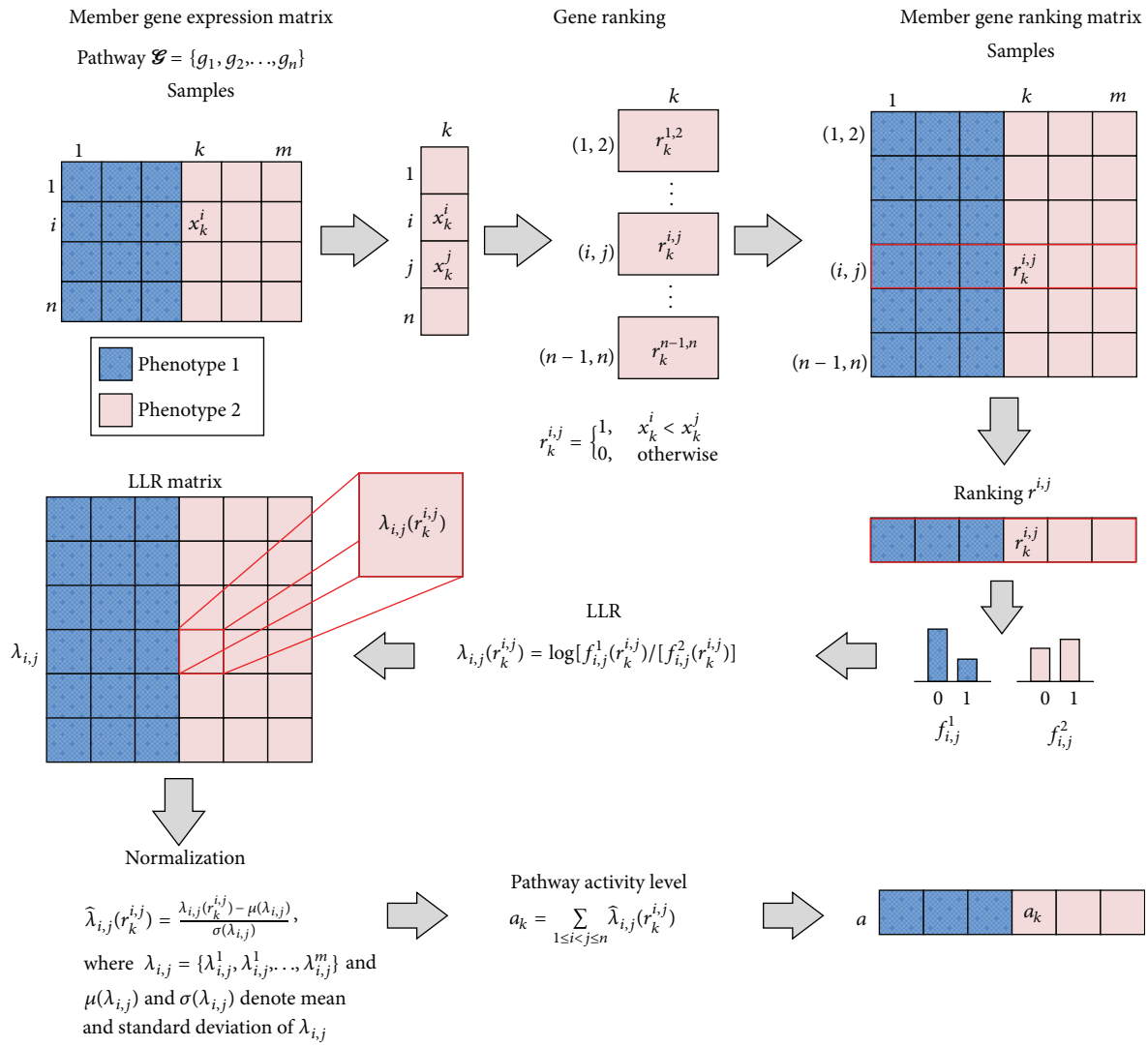

FIGURE 1: Probabilistic inference of rank-based pathway activity. For a given pathway, we first compute the ranking of the member genes for each individual sample in the dataset. Then we estimate the conditional probability mass function (PMF) of the gene ranking under each phenotype. Next, we transform the gene ranking into log-likelihood ratios (LLRs) based on the estimated PMFs and normalize the LLR matrix. Finally, the pathway activity level is inferred by aggregating the normalized LLRs of the member genes.

where $\mathbf{a} = \{a_k\}$ is the set of inferred pathway activity levels for a given pathway, μ_ℓ and σ_ℓ represent the mean and the standard deviation of the pathway activity levels for samples with phenotype $\ell \in \{1, 2\}$, respectively, and K_ℓ represents the number of samples in the dataset with phenotype ℓ. This measure has been widely used in previous studies to evaluate the performance of pathway markers [9, 10].

2.5. Evaluation of the Classification Performance. In order to evaluate the classification performance, we use the AUC (Area under ROC Curve). Many previous studies [8–11] have utilized AUC due to its ability to summarize the efficacy of a classification method over the entire range of specificity and sensitivity. We compute the AUC based on the method proposed in [22]. Given a classifier, let $x_1, x_2, \ldots, x_m$ be the output of the classifier for m positive samples and let $y_1, y_2, \ldots, y_n$ be the output for n negative samples. The AUC of the classifier can be computed as follows:

$$A = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} I\left(x_i > y_j\right), \tag{7}$$

where

$$I\left(x_i > y_j\right) = \begin{cases} 1, & \text{if } x_i > y_j, \\ 0, & \text{otherwise.} \end{cases} \tag{8}$$

3. Results and Discussion

3.1. Discriminative Power of the Pathway Markers Using the Proposed Method. In order to assess the performance of the rank-based pathway activity inference method proposed in this paper, we first evaluated the discriminative power of the pathway markers following a similar setup that was adopted

in a number of previous studies [9, 10]. For comparison, we also evaluated the performance of the mean and median-based schemes proposed in [6] and the original probabilistic pathway activity inference method (we refer to this method as the "LLR method" for simplicity) presented in [10]. As explained in Materials and Methods, the discriminative power of a pathway marker was measured based on the absolute t-test score of the inferred pathway activity level. Then the pathway markers were sorted according to their t-score, in a descending order.

Figure 2 shows the discriminative power of the pathway markers on the six datasets using different activity inference methods. On each dataset, we computed the mean absolute t-test statistics score of the top P% pathways for each of the four pathway activity inference methods. The x-axis corresponds to the proportion (P%) of the top pathway markers that were considered and the y-axis shows the mean absolute t-test score for these pathway markers. As we can see from Figure 2, the proposed method clearly improves the discriminative power of the pathway markers on all six datasets that we considered in this study. In order to investigate the effect of normalization on the discriminative power of the pathway activity inference methods, we repeated this experiment using the USA and the Belgium datasets, where we first normalized the raw data using three different normalization methods (RMA, GCRMA, and MAS5) and then evaluated the discriminative power of the pathway markers. The results are summarized in Figure S1 (see Supplementary Material available online at http://dx.doi.org/10.1155/2013/618461), where we can see that the proposed rank-based scheme is not very sensitive to the choice of the normalization method and performs consistently well in all cases.

Next, we investigated how the top pathway markers identified on a specific dataset perform in other independent datasets. We first ranked the pathway markers based on their mean absolute t-test statistics score in one of the datasets and then estimated the discriminative power of the top P% markers on a different dataset. These results are shown in Figure 3, where the first dataset is used for ranking the markers and the second dataset is used for assessing the discriminative power. As we can see from Figure 3, the pathway markers identified using the mean- and the median-based schemes do not retain their discriminative power very well in other datasets. Both the LLR method [10] and the proposed rank-based inference method perform well across different datasets, where the proposed method clearly outperforms the previous LLR method. It is interesting to see that the discriminative power of the markers is retained even when we consider datasets that are obtained using different platforms. For example, USA/Belgium datasets are profiled on the U133a platform and The Netherlands dataset is profiled on a custom Agilent chip, but Figure 3 shows that pathway markers identified using the proposed method retain their discriminative power across these datasets. As before, we repeated these experiments after normalizing the datasets using different normalization methods. The results are depicted in Figure S2, where we can see that the proposed method works very well, regardless of the normalization method that was used. Interestingly, this is also true even when the first dataset and the second dataset are normalized using different methods, as shown in Figures S3 and S4.

Another interesting observation is that the rank-based method can overcome one of the limitations of the previous LLR method. For example, normalization of the Belgium dataset using GCRMA results makes the LLR method fail, as some of the genes loose variability and some of the LLR values become infinite. We can see this issue in Figures S1(d), S2(c), S3(a), and S3(f). However, this limitation is easily overcome by the proposed method through the use of gene ranking and the preselection of informative gene pairs based on mutual information.

3.2. Classification Performance of the Pathway Markers Using the Proposed Method. Next, we evaluated the classification performance of the proposed rank-based pathway activity inference method. For this purpose, we performed fivefold cross validation experiments, following a similar setup used in previous studies [8–11]. We first performed the within-dataset experiments for each of the six datasets. First, a given dataset was randomly divided into fivefolds, where fourfolds (training dataset) were used for constructing an LDA (Linear Discriminant Analysis) classifier and the remaining fold (testing dataset) was used for evaluating its performance. To construct the classifier, the training dataset was again divided into threefolds, where twofolds (marker-evaluation dataset) were used for evaluating the pathway markers and the remaining onefold (feature-selection dataset) for feature selection. The entire training dataset was used for PDF/PMF estimation. The overall setup is shown in Figure 4(a).

In order to build the classifier, we first evaluated the discriminative power of each pathway on the marker-evaluation dataset. The pathways were sorted according to their absolute t-test statistics score in a descending order and the top 50 pathways were selected as potential features. Initially, we started with an LDA-based classifier with a single feature (i.e., the pathway marker that is on the top of the list) and continued to expand the feature set by considering additional pathway markers in the list. The classifier was trained using the marker-evaluation dataset and its performance was assessed on the feature-selection dataset by measuring the AUC. Pathway markers were added to the feature set only when they increased the AUC. Finally, the performance of the classifier with the optimal feature set was evaluated by computing the AUC on the testing dataset. The above process was repeated for 100 random partitions to ensure reliable results, and we report the average AUC as the measure of overall classification performance.

Figure 5 shows how the respective classifiers that use different pathway activity inference methods perform on different datasets. As we can see in Figure 5, among the four inference methods, the proposed rank-based scheme typically yields the best average performance across these datasets. We also performed similar experiments based on the USA and the Belgium datasets after normalizing the raw data using different normalization methods. These results are summarized in Figure S5. We can see from Figure S5 that the proposed method yields the best performance on the

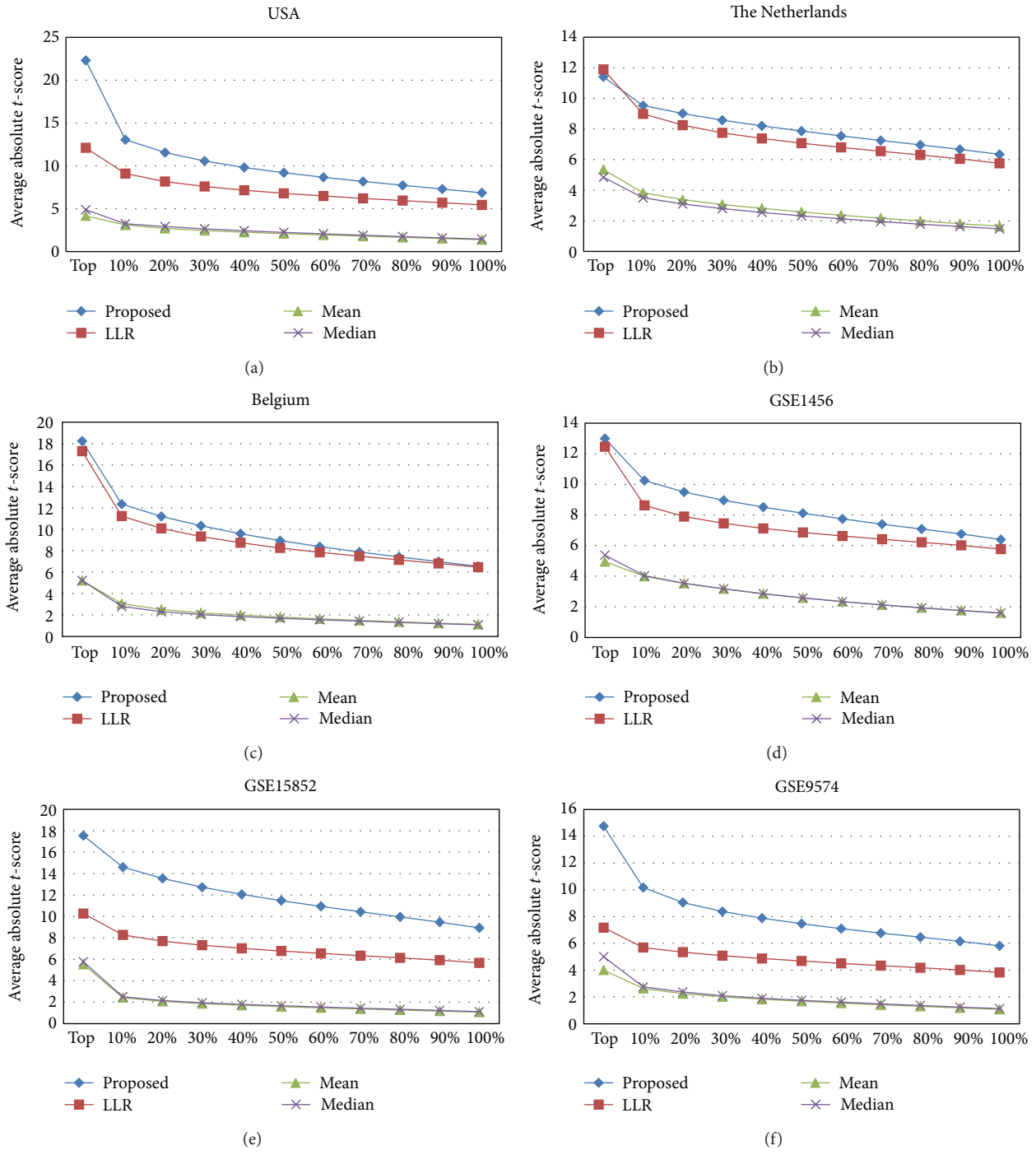

FIGURE 2: Discriminative power of pathway markers. We computed the mean absolute t-score of the top P% markers for each dataset without any further normalization.

USA dataset for all three normalization methods. On the Belgium dataset, the proposed method yields good consistent performance that is not very sensitive to the normalization method.

3.3. Reproducibility of the Pathway Markers Identified by the Proposed Method. To assess the reproducibility of the pathway markers, we performed the following cross-dataset experiments based on a similar setup that has been utilized in previous studies [8–11]. In this experiment, we used one of the breast cancer datasets for selecting the best pathway markers (i.e., only for feature selection) and a different dataset for building the classifier (using the selected pathways) and evaluating the performance of the resulting classifier. More specifically, we proceeded as follows. The first dataset was first divided into threefolds, where twofolds were used for

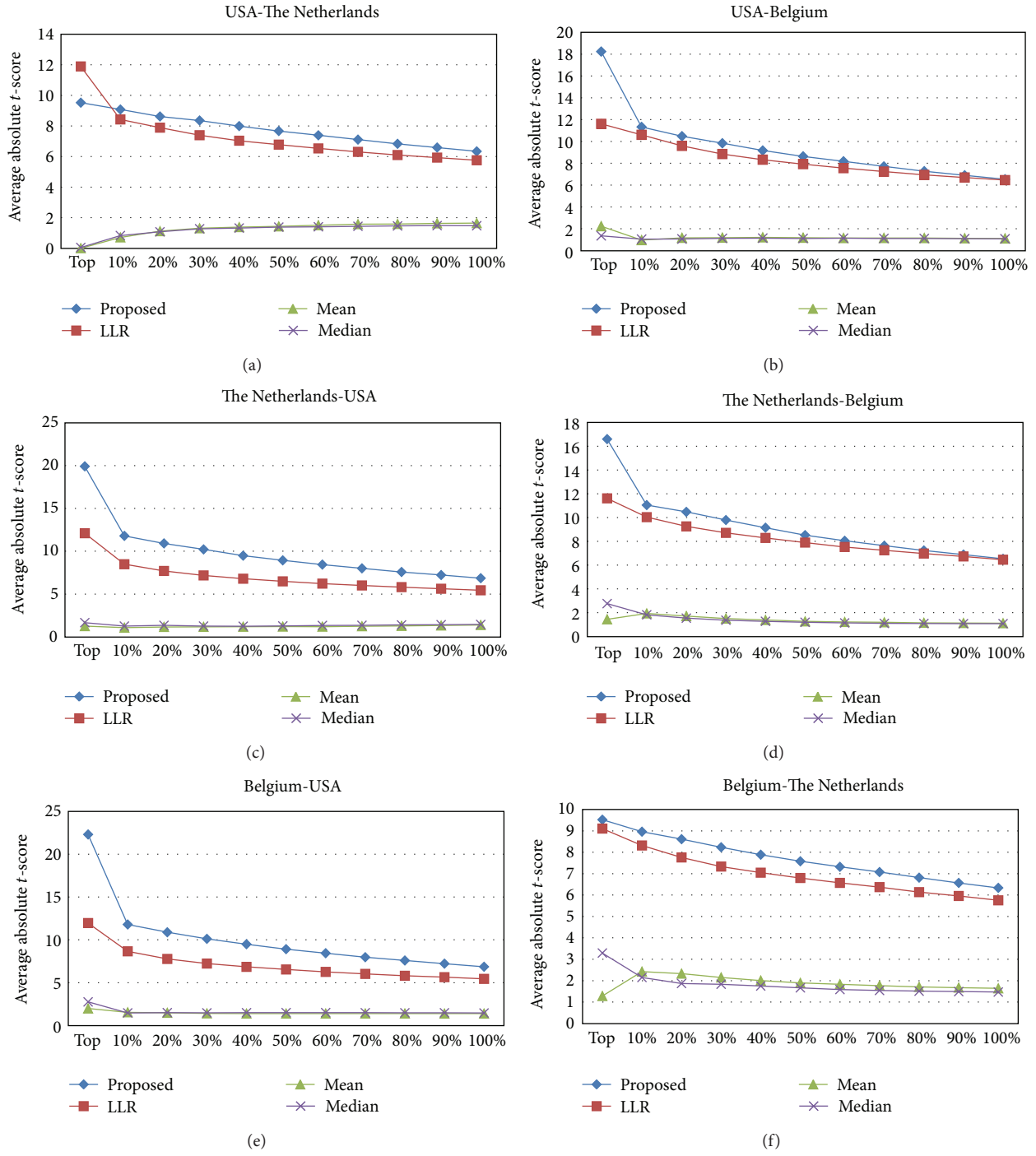

FIGURE 3: Discriminative power of pathway markers across different datasets. The pathway markers have been ranked and sorted using the first dataset, and their discriminative power has been reevaluated using the second dataset. As before, the mean absolute t-score was used for assessing the discriminative power.

marker evaluation and the remaining fold was used for feature selection. The second dataset was randomly divided into fivefolds, where fourfolds were used to train the LDA classifier, using the features selected from the first dataset, and the remaining fold was used to evaluate the classification performance. The overall setup is shown in Figure 4(b). To obtain reliable results, we repeated this experiment for 100 random partitions (of the second dataset) and report the average AUC as the performance metric. For these experiments, we used the three largest breast cancer datasets (USA, The Netherlands, and Belgium) among the six.

The results of the cross-dataset classification experiments are shown in Figure 6. As we can see from this figure, the proposed rank-based inference scheme typically outperforms

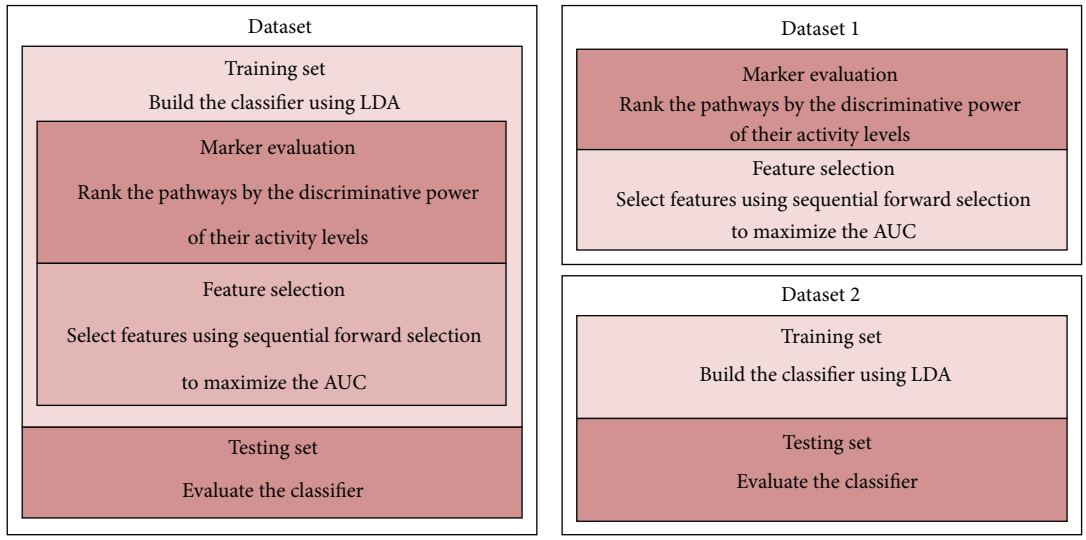

Figure 4: Experimental setup for evaluating the classification performance. (a) The setup for the within-dataset experiment. (b) The setup for the cross-dataset experiment.

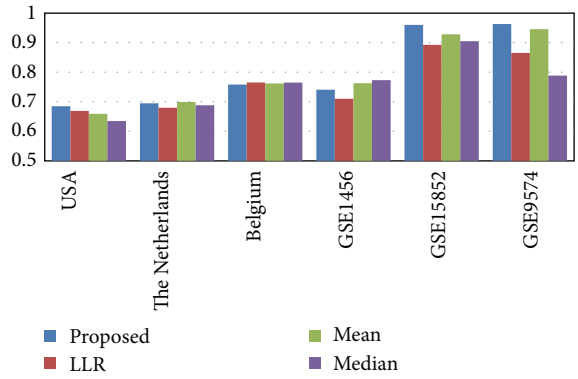

Figure 5: Classification performance for within-dataset experiments. The bars show the classification performance (average AUC) of different pathway activity inference methods evaluated on various breast cancer datasets.

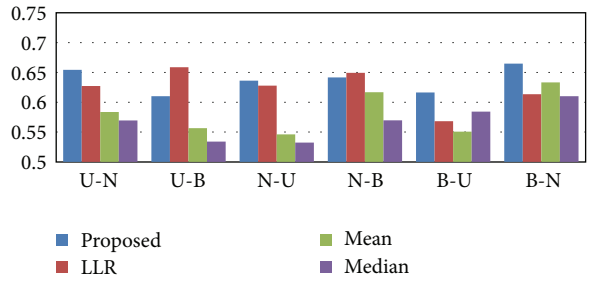

Figure 6: Classification performance for cross-dataset experiments. The bars show the cross-dataset classification performance (average AUC) of different pathway activity inference methods. The first dataset was used for selecting the pathway markers and the second dataset was used for training and evaluation of the classifier. The three largest breast cancer datasets were used: USA (U), The Netherlands (N), and Belgium (B).

other methods in terms of reproducibility. Furthermore, we can also observe that the proposed method yields consistent classification performance across experiments, while the performance of other inference methods is much more sensitive on the choice of the dataset. Next, we repeated the cross-dataset classification experiments based on the USA and the Belgium datasets after normalizing the raw data using RMA, GCRMA, and MAS5. As shown in Figure 7, the proposed method yields consistently good performance, regardless of the normalization method that was used.

Finally, we performed additional cross-dataset experiments after normalizing the USA and the Belgium datasets using different normalization methods. These results are summarized in Figures S6 and S7. We can see that the proposed pathway activity inference scheme is relatively robust to "normalization mismatch." Moreover, these results also show that the proposed scheme overcomes the problem of the previous LLR-based scheme [10] when used with GCRMA (see Figures 7, S6, and S7).

4. Conclusions

In this work, we proposed an improved pathway activity inference scheme, which can be used for finding more robust and reproducible pathway markers for predicting breast cancer metastasis. The proposed method integrates two effective strategies that have been recently proposed in the field: namely, the probabilistic pathway activity inference method [10] and the ranking-based relative gene expression analysis approach [3]. Experimental results based on several breast cancer gene expression datasets show that our proposed inference method identifies better pathway markers that have higher discriminative power, are more reproducible, and can lead to better classifiers that yield more consistent performance across independent datasets.

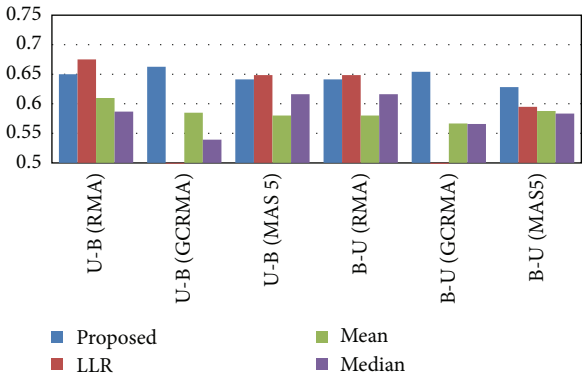

FIGURE 7: Classification performance for cross-dataset experiments. We repeated the cross-dataset experiments based on the USA and the Belgium datasets after normalizing the raw data using different normalization methods.

Acknowledgment

N. Khunlertgit has been supported by a scholarship from the Royal Thai Government.

References

[1] M. West, C. Blanchette, H. Dressman et al., "Predicting the clinical status of human breast cancer by using gene expression profiles," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 98, no. 20, pp. 11462–11467, 2001.

[2] L. J. Van't Veer, H. Dai, M. J. Van de Vijver et al., "Gene expression profiling predicts clinical outcome of breast cancer," *Nature*, vol. 415, no. 6871, pp. 530–536, 2002.

[3] D. Geman, C. D'Avignon, D. Q. Naiman, and R. L. Winslow, "Classifying gene expression profiles from pairwise mRNA comparisons," *Statistical Applications in Genetics and Molecular Biology*, vol. 3, no. 1, article 19, 2004.

[4] Y. Wang, J. G. M. Klijn, Y. Zhang et al., "Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer," *The Lancet*, vol. 365, no. 9460, pp. 671–679, 2005.

[5] L. Tian, S. A. Greenberg, S. W. Kong, J. Altschuler, I. S. Kohane, and P. J. Park, "Discovering statistically significant pathways in expression profiling studies," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 38, pp. 13544–13549, 2005.

[6] Z. Guo, T. Zhang, X. Li et al., "Towards precise classification of cancers based on robust gene functional expression profiles," *BMC Bioinformatics*, vol. 6, article 58, 2005.

[7] C. Auffray, "Protein subnetwork markers improve prediction of cancer outcome," *Molecular Systems Biology*, vol. 3, article 141, 2007.

[8] H. Y. Chuang, E. Lee, Y. T. Liu, D. Lee, and T. Ideker, "Network-based classification of breast cancer metastasis," *Molecular Systems Biology*, vol. 3, article 140, 2007.

[9] E. Lee, H. Y. Chuang, J. W. Kim, T. Ideker, and D. Lee, "Inferring pathway activity toward precise disease classification," *PLoS Computational Biology*, vol. 4, no. 11, Article ID e1000217, 2008.

[10] 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, Article ID e8161, 2009.

[11] J. Su, B. J. Yoon, and E. R. Dougherty, "Identification of diagnostic subnetwork markers for cancer in human protein-protein interaction network," *BMC Bioinformatics*, vol. 11, no. 6, article 8, 2010.

[12] J. A. Eddy, L. Hood, N. D. Price, and D. Geman, "Identifying tightly regulated and variably expressed networks by Differential Rank Conservation (DIRAC)," *PLoS Computational Biology*, vol. 6, no. 5, Article ID e1000792, 2010.

[13] N. Khunlertgit and B. J. Yoon, "Finding robust pathway markers for cancer classification," in *Proceedings of the IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS '12)*, 2012.

[14] M. J. Van De Vijver, Y. D. He, L. J. Van 'T Veer et al., "A gene-expression signature as a predictor of survival in breast cancer," *New England Journal of Medicine*, vol. 347, no. 25, pp. 1999–2009, 2002.

[15] C. Desmedt, F. Piette, S. Loi et al., "Strong time dependence of the 76-gene prognostic signature for node-negative breast cancer patients in the TRANSBIG multicenter independent validation series," *Clinical Cancer Research*, vol. 13, no. 11, pp. 3207–3214, 2007.

[16] Y. Pawitan, J. Bjohle, L. Amler, and A. L. Borg, "Gene expression profiling spares early breast cancer patients from adjuvant therapy: derived and validated in two population-based cohorts," *Breast Cancer Research*, vol. 7, pp. R953–R964, 2005.

[17] H. Y. Chang, D. S. A. Nuyten, J. B. Sneddon et al., "Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival," *Proceedings of the National Academy of Sciences of the United States of America*, vol. 102, no. 10, pp. 3738–3743, 2005.

[18] R. Edgar, M. Domrachev, and A. E. Lash, "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository," *Nucleic Acids Research*, vol. 30, no. 1, pp. 207–210, 2002.

[19] R. C. Gentleman, V. J. Carey, D. M. Bates et al., "Bioconductor: open software development for computational biology and bioinformatics," *Genome Biology*, vol. 5, no. 10, p. R80, 2004.

[20] A. Liberzon, A. Subramanian, R. Pinchback, H. Thorvaldsdóttir, P. Tamayo, and J. P. Mesirov, "Molecular signatures database (MSigDB) 3.0," *Bioinformatics*, vol. 27, no. 12, pp. 1739–1740, 2011.

[21] T. M. Cover and J. A. Thomas, *Elements of Information Theory*, Wiley Interscience, New York, NY, USA, 2006.

[22] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters*, vol. 27, no. 8, pp. 861–874, 2006.