IMPROVEMENT OF REPRODUCIBILITY IN CANCER CLASSIFICATION BASED ON PATHWAY MARKERS AND SUBNETWORK MARKERS

A Dissertation

by

NAVADON KHUNLERTGIT

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Byung-Jun Yoon
Committee Members,	Edward R. Dougherty
	Henry D. Pfister
	Ivan Ivanov
Head of Department,	Miroslav M. Begovic

December 2016

Major Subject: Electrical Engineering

ABSTRACT

Identification of robust biomarkers for cancer prognosis based on gene expression data is an important research problem in translational genomics. The high-dimensional and small-sample-size data setting makes the prediction of biomarkers very challenging. Biomarkers have been identified based solely on gene expression data in the early stage. However, very few of them are jointly shared among independent studies. To overcome this irreproducibility, the integrative approach has been proposed to identify better biomarkers by overlaying gene expression data with available biological knowledge and investigating genes at the modular level. These module-based markers jointly analyze the gene expression activities of closely associated genes; for example, those that belong to a common biological pathway or genes whose protein products form a subnetwork module in a protein-protein interaction network. Several studies have shown that modular biomarkers lead to more accurate and reproducible prognostic predictions than single-gene markers and also provide the better understanding of the disease mechanisms.

We propose novel methods for identifying modular markers which can be used to predict breast cancer prognosis. First, to improve identification of pathway markers, we propose using probabilistic pathway activity inference and relative expression analysis. Then, we propose a new method to identify subnetwork markers based on a message-passing clustering algorithm, and we further improve this method by incorporating topological attribute using association coefficients. Through extensive evaluations using multiple publicly available datasets, we demonstrate that all of the proposed methods can identify modular markers that are more reliable and reproducible across independent datasets compared to those identified by existing

methods, hence they have the potential to become more effective prognostic cancer classifiers.

DEDICATION

This dissertation is dedicated to my parents and teachers.

ACKNOWLEDGEMENTS

First and foremost, I would like to sincerely thank my advisor, Prof. Byung-Jun Yoon, for all of his guidance and continued support throughout these years. I feel grateful for his patience and encouragement during the time I struggled. I could not have gone this far without his vision and advice.

I would also like to thank Prof. Edward R. Dougherty, Prof. Henry D. Pfister, and Prof. Ivan Ivanov for serving on my doctoral committee and for their constructive comments.

I would like to express my appreciation to faculty and staff who have helped me all through my time at Texas A&M University. Thanks also go to my Thai friends, current and former members of the Genomic Signal Processing Lab for making my experience in Aggieland remarkable.

I would like to acknowledge the Royal Thai Government Scholarship. I truly appreciate the assistance from the Office of Educational Affairs of Royal Thai Embassy.

Last but not least, I would like to acknowledge and thank my beloved family for their patience, understanding, and love.

TABLE OF CONTENTS

Page

ABSTRACT ii

DEDICATION iv

ACKNOWLEDGEMENTS v

TABLE OF CONTENTS vi

LIST OF FIGURES viii

LIST OF TABLES x

1. INTRODUCTION 1

2. IDENTIFICATION OF ROBUST PATHWAY MARKERS FOR CANCER THROUGH RANK-BASED PATHWAY ACTIVITY INFERENCE 4

2.1 Introduction 4
2.2 Materials and Methods 5
2.2.1 Study Datasets 5
2.2.2 Gene Ranking 6
2.2.3 Pathway Activity Inference Based on Gene Ranking 6
2.2.4 Assessing the Discriminative Power of Pathway Markers . . . 8
2.2.5 Evaluation of the Classification Performance 8
2.3 Results and Discussion 10
2.3.1 Discriminative Power of the Pathway Markers Using the Proposed Method 10
2.3.2 Classification Performance of the Pathway Markers Using the Proposed Method 15
2.3.3 Reproducibility of the Pathway Markers Identified by Proposed Method 21
2.4 Conclusions 23

3. SIMULTANEOUS IDENTIFICATION OF ROBUST SYNERGISTIC SUBNETWORK MARKERS FOR EFFECTIVE CANCER PROGNOSIS . . . 25

3.1 Introduction 25

3.2 Materials and Methods 26
3.2.1 Datasets 26
3.2.2 The Affinity Propagation Algorithm: A Brief Overview 27
3.2.3 Computing the Similarity Between Genes 29
3.2.4 Post-Processing the Identified Gene Subnetworks 31
3.2.5 Probabilistic Inference of Subnetwork Activity 31
3.3 Results 32
3.3.1 Statistics of the Identified Subnetwork Markers 32
3.3.2 Computational Cost for Subnetwork Marker Identification . . 35
3.3.3 Discriminative Power of the Subnetwork Markers 36
3.3.4 Evaluating the Reproducibility of the Predicted Subnetwork Markers 42
3.4 Conclusions 45

4. INCORPORATING TOPOLOGICAL INFORMATION FOR PREDICTING ROBUST CANCER SUBNETWORK MARKERS IN HUMAN PROTEIN-PROTEIN INTERACTION NETWORK 47

4.1 Introduction 47
4.2 Materials and Methods 48
4.2.1 Datasets 48
4.2.2 Affinity Propagation-Based Subnetwork Identification 50
4.2.3 Incorporating Topological Information for Computing the Similarity Between Genes 51
4.2.4 Probabilistic Inference of Subnetwork Activity 54
4.2.5 Experimental Setup 54
4.3 Results 55
4.3.1 Statistics of the Subnetwork Markers 55
4.3.2 Discriminative Power of the Subnetwork Markers 59
4.3.3 Evaluating the Reproducibility of the Identified Subnetwork Markers 62
4.4 Conclusions 64

5. OVERALL CONCLUSIONS 67

REFERENCES 69

LIST OF FIGURES

FIGURE Page

2.1 Probabilistic inference of rank-based pathway activity [1]. 9

2.2 Discriminative power of pathway markers. Each dataset is used as published by their original studies [1]. 11

2.3 Discriminative power of pathway markers across different datasets. Each dataset is used as published by their original studies [1]. 12

2.4 Discriminative power of pathway markers identified on re-normalized datasets [1]. 14

2.5 Discriminative power of pathway markers across re-normalized datasets [1]. 16

2.6 Discriminative power of pathway markers identified on re-normalized USA dataset and evaluated on re-normalized Belgium dataset [1]. . . 17

2.7 Discriminative power of pathway markers identified on re-normalized Belgium dataset and evaluated on re-normalized USA dataset [1]. . . 18

2.8 Experimental setup for evaluating the classification performance [1]. . 19

2.9 Classification performance for within-dataset experiments [1]. 20

2.10 Classification performance for within-dataset experiments based on re-normalized datasets [1]. 20

2.11 Classification performance for cross-dataset experiments [1]. 22

2.12 Classification performance for cross-dataset experiments based on re-normalized datasets [1]. 23

2.13 Classification performance for re-normalized USA-Belgium cross-dataset experiments [1]. 23

2.14 Classification performance for re-normalized Belgium-USA cross-dataset experiments [1]. 24

3.1 Total CPU Time for identifying the top 50 subnetwork markers [2]. . 36

FIGURE Page

3.2 Discriminative power of the identified subnetwork markers from various methods [2]. 37

3.3 Comparison of discriminative power of subnetwork markers identified by the proposed method with and without post-processing step [2]. . 38

3.4 Discriminative power of subnetwork markers identified on USA and Netherlands datasets [2]. 40

3.5 Discriminative power of subnetwork markers identified on Belgium and Sweden datasets [2]. 41

3.6 Reproducibility of various subnetwork and pathway markers [2]. . . . 44

3.7 Classification performance [2]. 45

4.1 Discriminative power of subnetwork markers identified on GSE2034 by different methods [3]. 60

4.2 Discriminative power of subnetwork markers identified on GSE2034 and evaluated on NKI295 [3]. 61

4.3 Discriminative power of subnetwork markers identified on NKI295 by different methods [3]. 62

4.4 Discriminative power of subnetwork markers identified on NKI295 and evaluated on GSE2034 [3]. 63

4.5 Reproducibility of subnetwork markers identified by various methods [3]. 64

LIST OF TABLES

TABLE Page

3.1 Average size of the identified subnetwork markers [2]. 33

3.2 Total number of unique genes in the identified subnetwork markers [2]. 34

3.3 Total number of common genes between the top subnetwork markers identified using different α [2]. 34

3.4 Overlap between the top subnetwork markers identified on different datasets [2]. 34

4.1 The number of proteins and interactions for each PPI network [3]. . . 49

4.2 The number of proteins and interactions for each induced PPI network [3]. 49

4.3 The average size of top 50 highly discriminative subnetwork markers from GSE2034 and NKI295 [3]. 56

4.4 The number of unique genes in top 50 highly discriminative subnetwork markers from GSE2034 and NKI295 [3]. 58

4.5 Overlap between the top subnetwork markers identified on different gene expression datasets [3]. 59

4.6 The number of genes in top 50 highly discriminative subnetwork markers from *tve_p* method on GASOLINE categorized by their GO terms [3]. 66

1. INTRODUCTION

Advancement in microarray and sequencing technologies has enabled analysis of gene expressions on a genome-wide scale, leading to the translational and functional genomics research. This research aims to screen for phenotype-related genes (or gene combinations) and to utilize genomic signals to classify disease [4]. One of the important problems in this field is to find reliable diagnosis and prognosis of complex diseases based on genome-wide expression profiles. This problem regularly involves a small sample size of clinical data from patients with a vastly large number of genes that can be seen as the high-dimensional and small-sample-size feature selection issue. This problem generally involves heterogeneity across patients and samples. These statements make this problem practically difficult and very challenging.

Over the last decade, several research studies have been working on classifying type or state of human diseases, trying to identify disease-related genes – or "biomarkers" – from microarray gene expression data. The biomarkers identified solely from gene expression data have been shown to be good candidates for building classifiers for disease prediction. However, these gene-based biomarkers have limitations. For instance, two large-scale-dataset studies of breast cancer [5, 6] tried to find the gene markers that predict metastasis. These studies practically searched for key genes that show differential expression under distinct phenotype. Both of them identified approximately 70 gene markers with the accuracy rate of 60-70%. Surprisingly, they shared only 3 genes from 55 of possible genes that might share across two different microarray platforms [7]. Moreover, these gene-based markers yielded low classification performance across different datasets.

Many studies have attempted to improve prediction accuracy and reproducibility of the biomarkers. One of the possible ways is based on the assumption that the progression of complex diseases involves dysregulation of multiple genetic processes. Therefore, it might be beneficial to interpret multiple genes that are known to be in the similar biological pathways or genes whose protein products are functionally related (or belong to the same biological function) as a single attribute [8–14]. In order to do this, biological information sources compiled by experts, such as gene sets, pathways, and protein-protein interaction (PPI) network, have been utilized with gene expression profiles. This leads us to the concept of "integrative approach" where gene expression is analyzed and interpreted at modular level through data integration. The biomarker obtained from this approach is called "modular marker". This approach has a potential to yield better biomarkers in terms of accuracy and reproducibility. Moreover, it allows us to view the gene expression in function-organized fashion, which may facilitate the prediction of system-level features based on the markers.

The rest of this dissertation is organized as follows: In chapter 2, we propose a new strategy to identify pathway markers based on relative expression analysis and probabilistic inference. We evaluate the classification performance of obtained markers on independent datasets and different normalization methods. In chapter 3 and 4, to overcome the limited amount of available pathway information, we utilize PPI network to identify subnetwork markers. In chapter 3, we propose a new method for identifying subnetwork markers by adopting message-passing clustering algorithm. We demonstrate how to define inputs and employ the algorithm to this type of problem. We assess the performance of obtain subnetwork markers and compare to the markers identified by existing methods. In chapter 4, we improve its performance by using topological information. We study the impact from various association indices that we use to estimate the topological information on different independent datasets. Finally, we draw overall conclusions in chapter 5.

2. IDENTIFICATION OF ROBUST PATHWAY MARKERS FOR CANCER THROUGH RANK-BASED PATHWAY ACTIVITY INFERENCE*

2.1 Introduction

There has been significant amount of work on identifying markers and building classifiers that can be used to predict breast cancer metastasis. 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 these problems, several different approaches have been proposed so far. For example, a recent work by Geman et al. [15] 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 problems is the concept of modular marker through known biological data integration. 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 [11–14].

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. [9] 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

*Part of this chapter is reprinted from "Navadon Khunlertgit, and Byung-Jun Yoon. Identification of Robust Pathway Markers for Cancer through Rank-Based Pathway Activity Inference. *Advances in Bioinformatics*, vol. 2013, no. 618461, 2013. doi:10.1155/2013/618461" [1] © [2013] Navadon Khunlertgit and Byung-Jun Yoon.

al. [13] 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 chapter, 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 diseases classifiers. The proposed method is motivated by the relative gene expression analysis strategy proposed in [15,16] and it builds on the concept of probabilistic pathway activity inference proposed in [13,14]. In this chapter, 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 at [17].

2.2 Materials and Methods

2.2.1 Study Datasets

Six independent breast cancer microarray gene expression datasets have been used in this study: GSE2034 (USA) [6], NKI295 (Netherlands) [18], GSE7390 (Belgium) [19], GSE1456 (Stockholm) [20], GSE15852 [21], and GSE9574 [22]. The Netherlands dataset uses a custom Agilent chip and it has been obtained from the Stanford website [23]. All datasets have been profiled using the Affymetrix U133a platform and they have been downloaded from the Gene Expression Omnibus (GEO) website [24],

The above datasets have been used in our study both with and without renormalization. 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 pre-processing methods provided in the Bioconductor package [25]. We applied three popular normalization methods – RMA [26], GCRMA [27], and MAS5 [28] – with default setting.

The pathway data have been obtained from the Molecular Signatures Database (MSigDB) 3.0 Canonical Pathways [29]. This pathway dataset consists of 880 pathways, where 3,698 genes in these pathways intersect with all datasets.

2.2.2 Gene Ranking

In this study, we utilize "gene ranking" or the relative ordering of the genes based on their expression levels within each profile [15]. Consider a pathway that contains n member genes $\boldsymbol{\mathcal{G}} = \{g_1, g_2, ..., g_n\}$ after removing the genes that are not included in all datasets. Given a sample $\boldsymbol{x}_k = \{x_k^1, x_k^2, ..., x_k^n\}$ that contains the expression level of the member genes, the gene ranking $\boldsymbol{r}_k$ is defined as follows

$$\boldsymbol{r}_k = \{r_k^{i,j} | 1 \leq i < j \leq n\} \tag{2.1}$$

where

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

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

2.2.3 Pathway Activity Inference Based on Gene Ranking

To infer the pathway activity, we follow the strategy proposed in [13], where the activity level a_k of a given pathway in the k-th sample is predicted by aggregating the probabilistic evidence of all the member genes. The main difference between the strategy proposed in this study and the original strategy [13] 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

$$a_k = \sum_{1 \leq i < j \leq n} \lambda_{i,j} \left(r_k^{i,j} \right), \tag{2.2}$$

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 $\boldsymbol{r}_k$. The LLR $\lambda_{i,j}(r_k^{i,j})$ is defined as

$$\lambda_{i,j}(r_k^{i,j}) = \log \left[f_{i,j}^1(r_k^{i,j}) / f_{i,j}^2(r_k^{i,j}) \right], \tag{2.3}$$

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 [30] 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 (2.2). Although we selected the top 10% gene pairs for simplicity, this may not be 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 $f^1_{i,j}(r)$ and $f^2_{i,j}(r)$. For this reason, we normalize the original LLR $\lambda_{i,j}(r^{i,j}_k)$ as follows to decrease its sensitivity to small alterations in gene ranking

$$\hat{\lambda}_{i,j}(r^{i,j}_k) = \frac{\lambda_{i,j}(r^{i,j}_k) - \mu(\lambda_{i,j})}{\sigma(\lambda_{i,j})}, \tag{2.4}$$

where $\mu(\lambda_{i,j})$ and $\sigma(\lambda_{i,j})$ are the mean and standard deviation of $\lambda_{i,j}(r^{i,j}_k)$ across all $k = 1, \cdots, n$. Figure 2.1 illustrates the overall process.

2.2.4 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

$$t(\boldsymbol{a}) = \frac{\mu_1 - \mu_2}{\sqrt{\sigma_1/K_1 + \sigma_2/K_2}}, \tag{2.5}$$

where $\boldsymbol{a} = \{a_k\}$ 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 [12, 13].

2.2.5 Evaluation of the Classification Performance

In order to evaluate the classification performance, we use the Area Under ROC Curve (AUC). Many previous studies [11–14] 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 [31]. Given a classifier, let $x_1, x_2, ..., x_m$ be the output of the classifier for m positive samples,

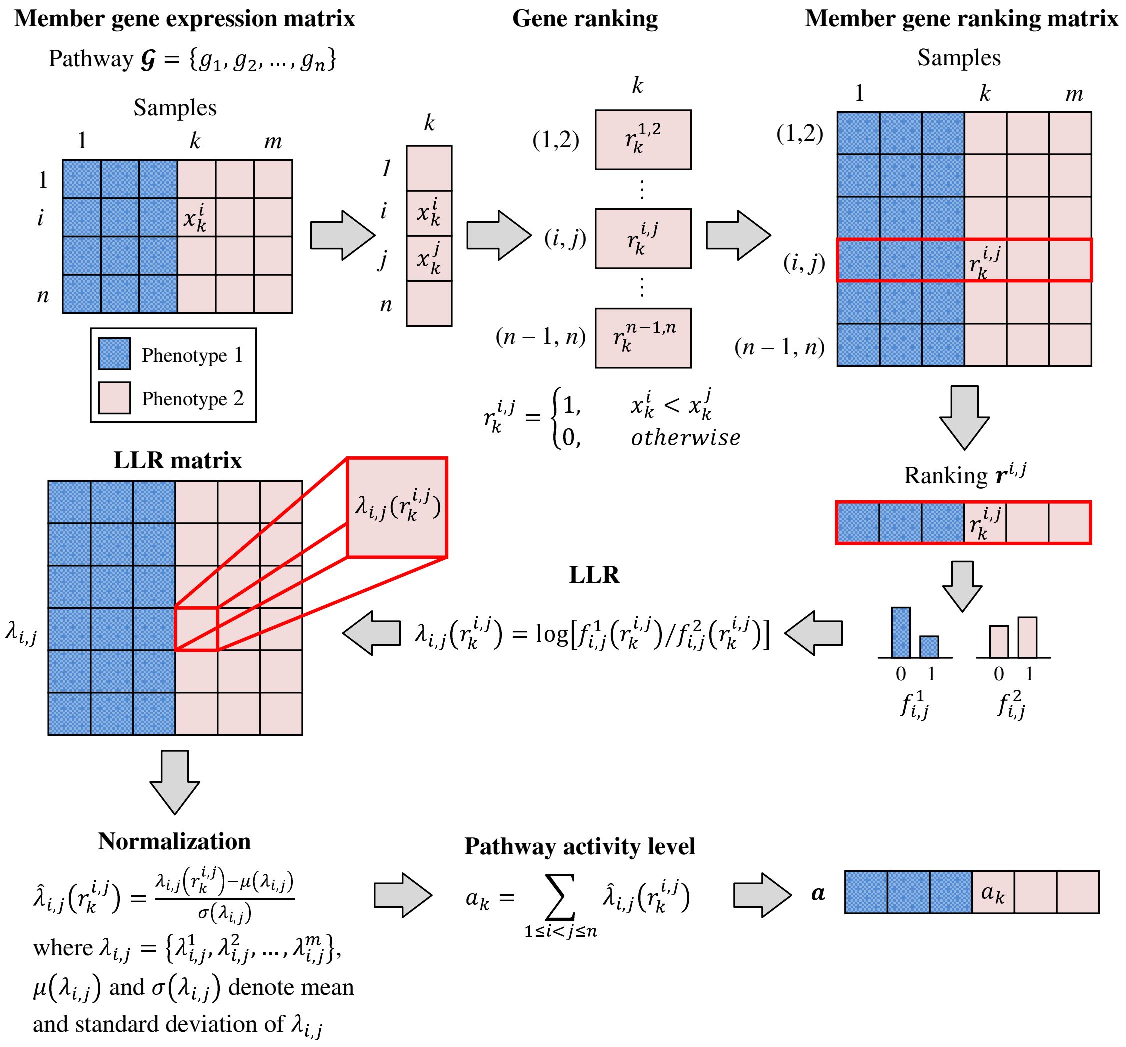

Member gene expression matrix
Pathway $\mathcal{G} = \{g_1, g_2, \dots, g_n\}$
Samples
1
k
m
1
i
n
x_k^i
Phenotype 1
Phenotype 2
k
1
i
j
n
x_k^i
x_k^j
Gene ranking
k
(1,2)
$r_k^{1,2}$
(i, j)
$r_k^{i,j}$
$(n-1, n)$
$r_k^{n-1,n}$
$r_k^{i,j} = \begin{cases} 1, & x_k^i < x_k^j \\ 0, & otherwise \end{cases}$
Member gene ranking matrix
Samples
1
k
m
(1,2)
(i, j)
$(n-1, n)$
$r_k^{i,j}$
Ranking $\boldsymbol{r}^{i,j}$
$r_k^{i,j}$
0
1
0
1
$f_{i,j}^1$
$f_{i,j}^2$
LLR
$\lambda_{i,j}(r_k^{i,j}) = \log[f_{i,j}^1(r_k^{i,j})/f_{i,j}^2(r_k^{i,j})]$
LLR matrix
$\lambda_{i,j}$
$\lambda_{i,j}(r_k^{i,j})$
Normalization
$\hat{\lambda}_{i,j}(r_k^{i,j}) = \frac{\lambda_{i,j}(r_k^{i,j}) - \mu(\lambda_{i,j})}{\sigma(\lambda_{i,j})}$
where $\lambda_{i,j} = \{\lambda_{i,j}^1, \lambda_{i,j}^2, \dots, \lambda_{i,j}^m\}$,
$\mu(\lambda_{i,j})$ and $\sigma(\lambda_{i,j})$ denote mean
and standard deviation of $\lambda_{i,j}$
Pathway activity level
$a_k = \sum_{1 \le i < j \le n} \hat{\lambda}_{i,j}(r_k^{i,j})$
$\boldsymbol{a}$
a_k

Figure 2.1: Probabilistic inference of rank-based pathway activity [1]. For a given pathway, we first compute the ranking of the members 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 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.

and $y_1, y_2, ..., y_n$ be the output for n negative samples. The AUC of the classifier can be computed as follows

$$A = \frac{1}{mn} \sum_{i=1}^{m} \sum_{j=1}^{n} I\left(x_i > y_j\right), \tag{2.6}$$

where

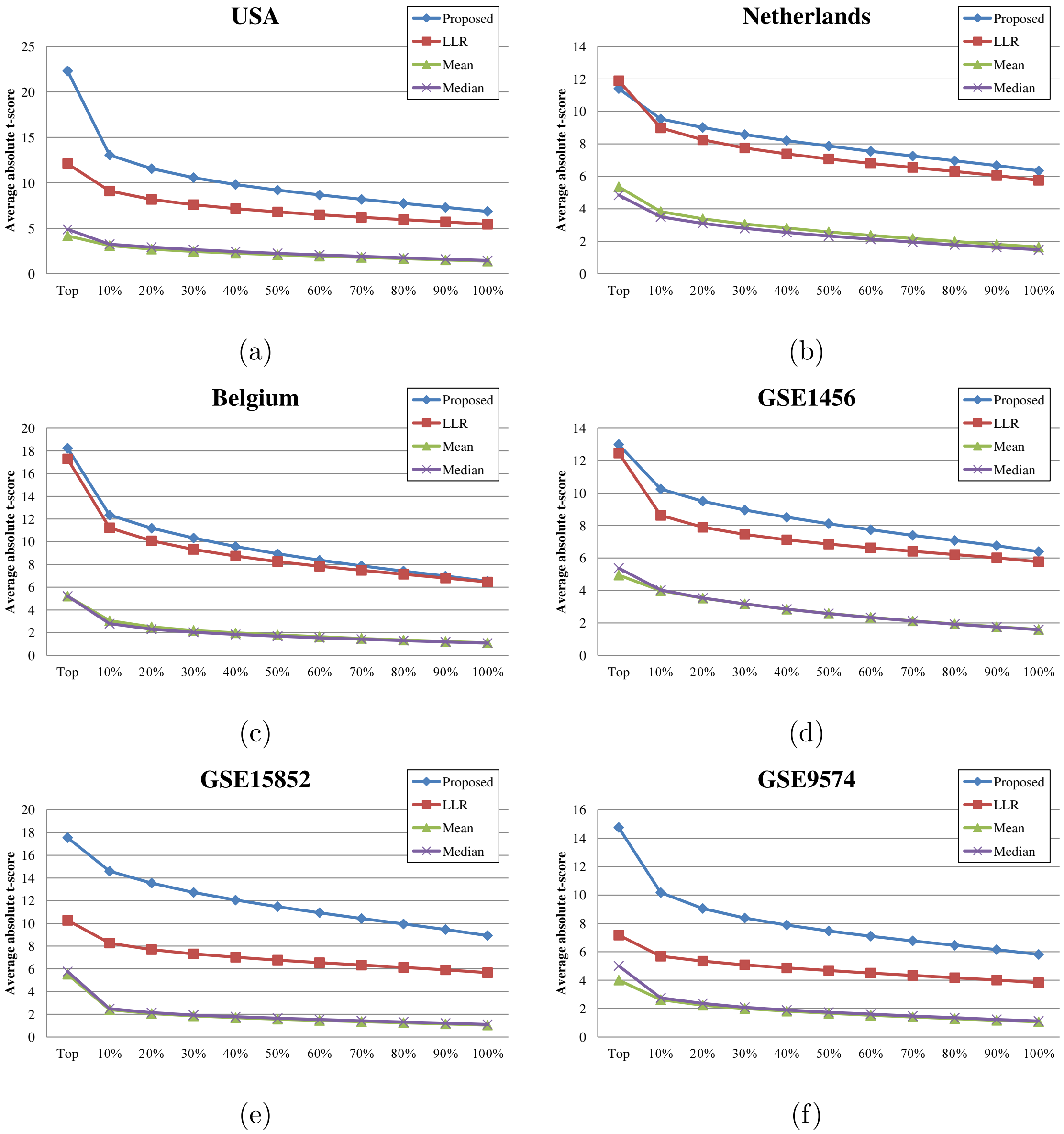
$$I(x_i > y_j) = \begin{cases} 1 & \text{if } x_i > y_j \\ 0 & \text{otherwise} \end{cases}$$

2.3 Results and Discussion

2.3.1 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 study, we first evaluated the discriminative power of the pathway markers following a similar set-up that was adopted in a number of previous studies [12,13]. For comparison, we also evaluated the performance of the mean and median-based schemes proposed in [9], and the original probabilistic pathway activity inference method (we refer to this method as the "LLR method" for simplicity) presented in [13]. As explained in 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 2.2 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 2.2, the



(a)

(b)

(c)

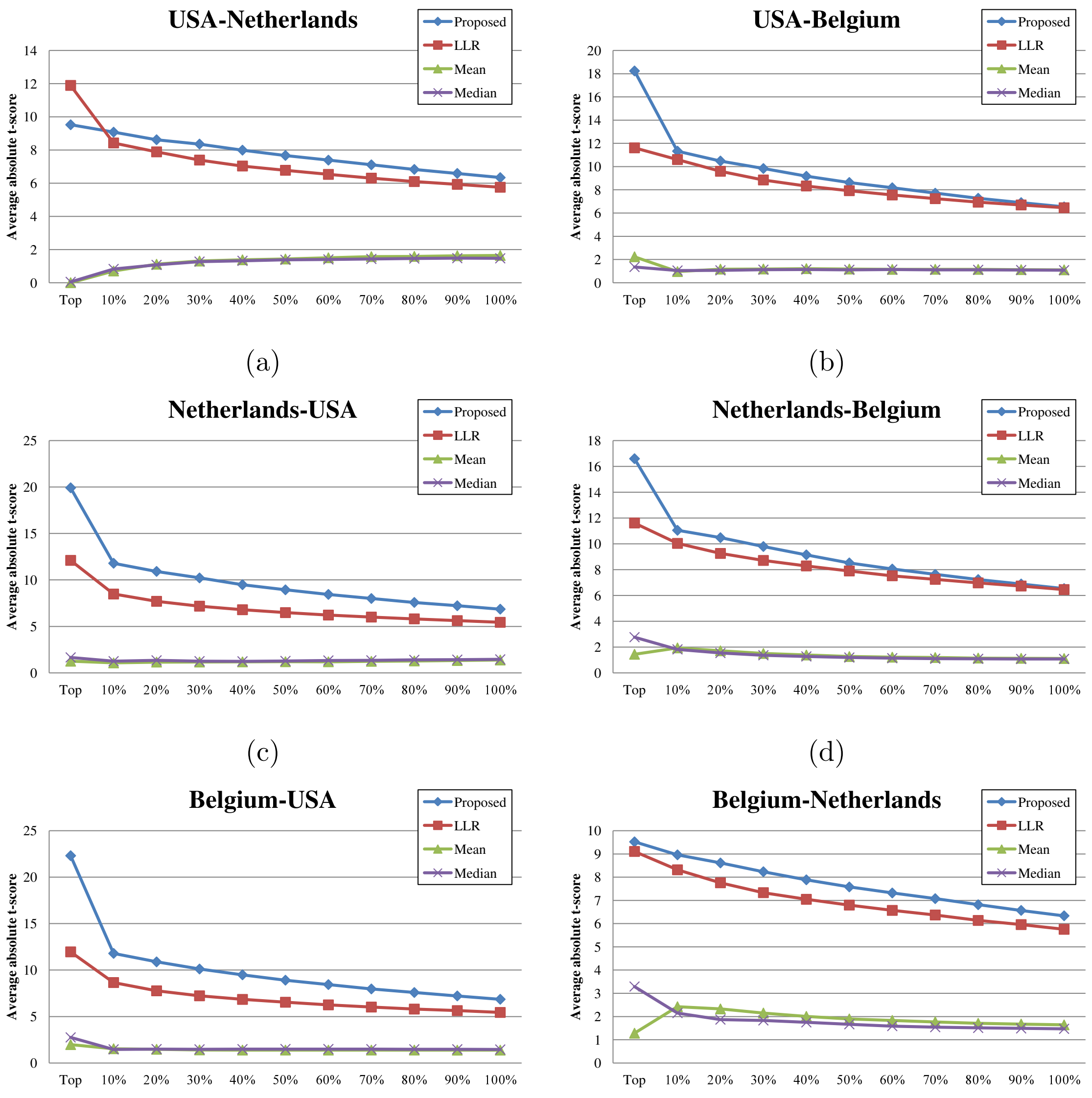
(d)

(e)

(f)

Figure 2.2: Discriminative power of pathway markers. Each dataset is used as published by their original studies [1]. We computed the mean absolute t-score of the top $P\%$ markers for each dataset without any further normalization.

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



(a) (b) (c) (d) (e) (f)

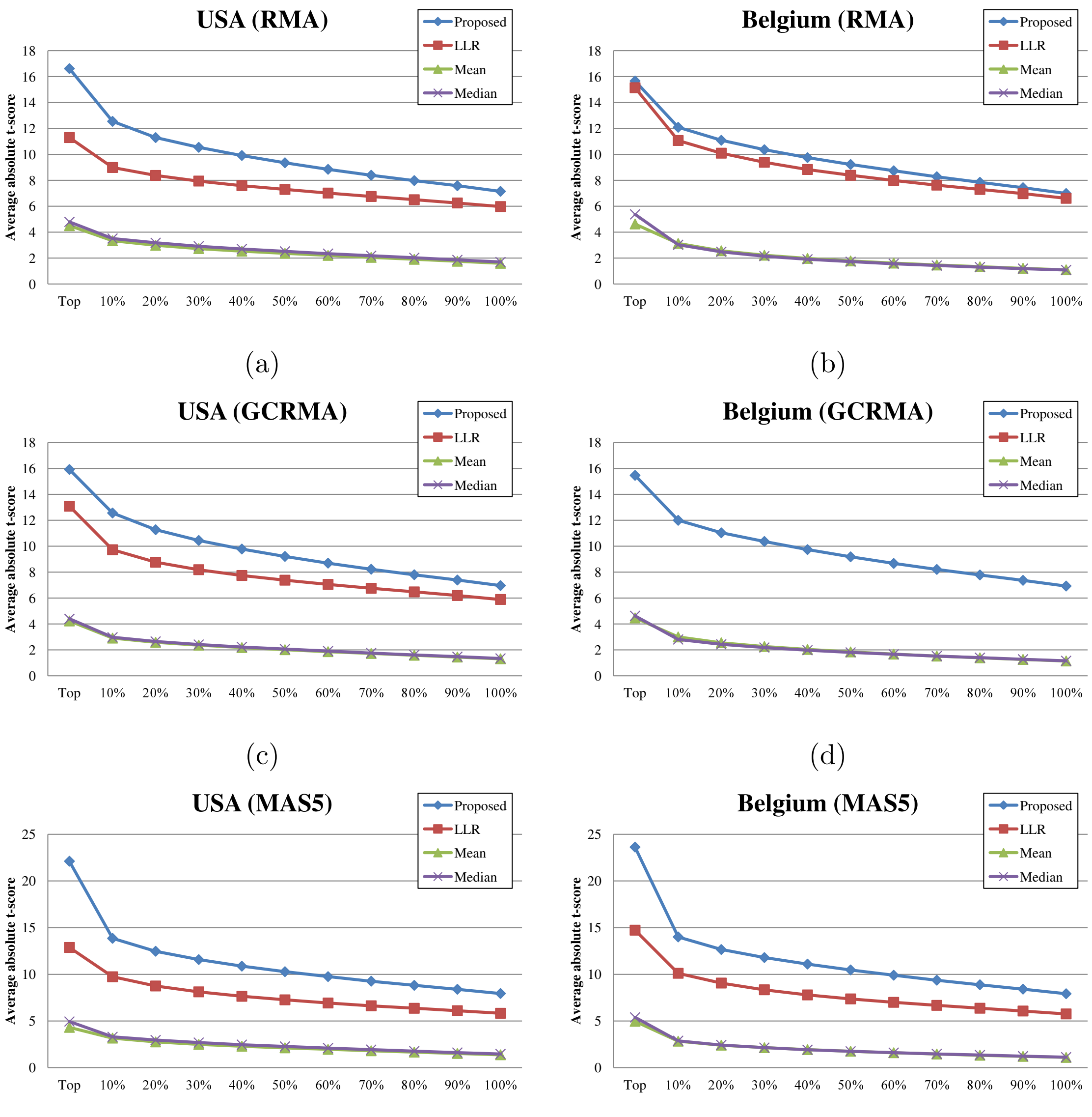
Figure 2.3: Discriminative power of pathway markers across different datasets. Each dataset is used as published by their original studies [1]. The pathway markers have been ranked and sorted using the first dataset and their discriminative power has been re-evaluated using the second dataset. As before, the mean absolute t-score was used for assessing the discriminative power.

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) then evaluated the discriminative power of the pathway markers. The results are summarized in Figure 2.4, 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 2.3, 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 2.3, 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 [13] 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 2.3 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 2.5, 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 2.6 and 2.7.

Another interesting observation is that the rank-based method can overcome



(a)

(b)

(c)

(d)

(e)

(f)

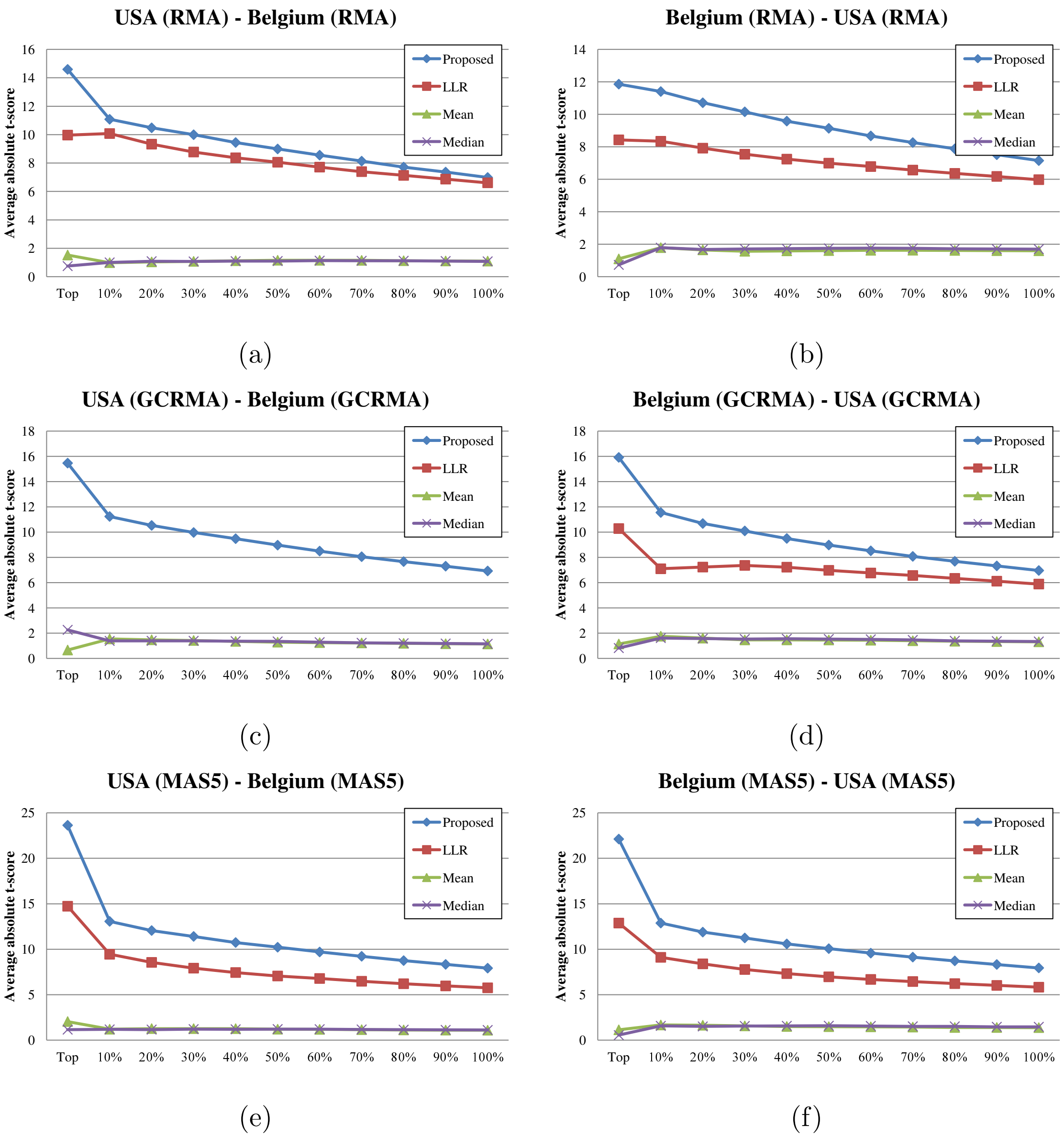
Figure 2.4: Discriminative power of pathway markers identified on re-normalized datasets [1]. We computed the mean absolute t-score of the top $P\%$ markers for the USA and the Belgium datasets after normalizing the raw data using three different normalization methods: RMA, GCRMA, and MAS5.

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 2.4 d, 2.5 c, 2.6 a, and 2.6 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.

2.3.2 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 five-fold cross-validation experiments, following a similar set-up used in previous studies [11–14]. We first performed the within-dataset experiments for each of the six datasets. First, a given dataset was randomly divided into five folds, where four folds ("training dataset") were used for constructing a Linear Discriminant Analysis (LDA) classifier and the remaining fold ("testing dataset") was used for evaluating its performance. To construct the classifier, the training dataset was again divided into three folds, where two folds ("marker-evaluation dataset") were used for evaluating the pathway markers and the remaining one fold ("feature-selection dataset") for feature selection. The entire training dataset was used for PDF/PMF estimation. The overall set-up is shown in Figure 2.8a.

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 a 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



(a)

(b)

(c)

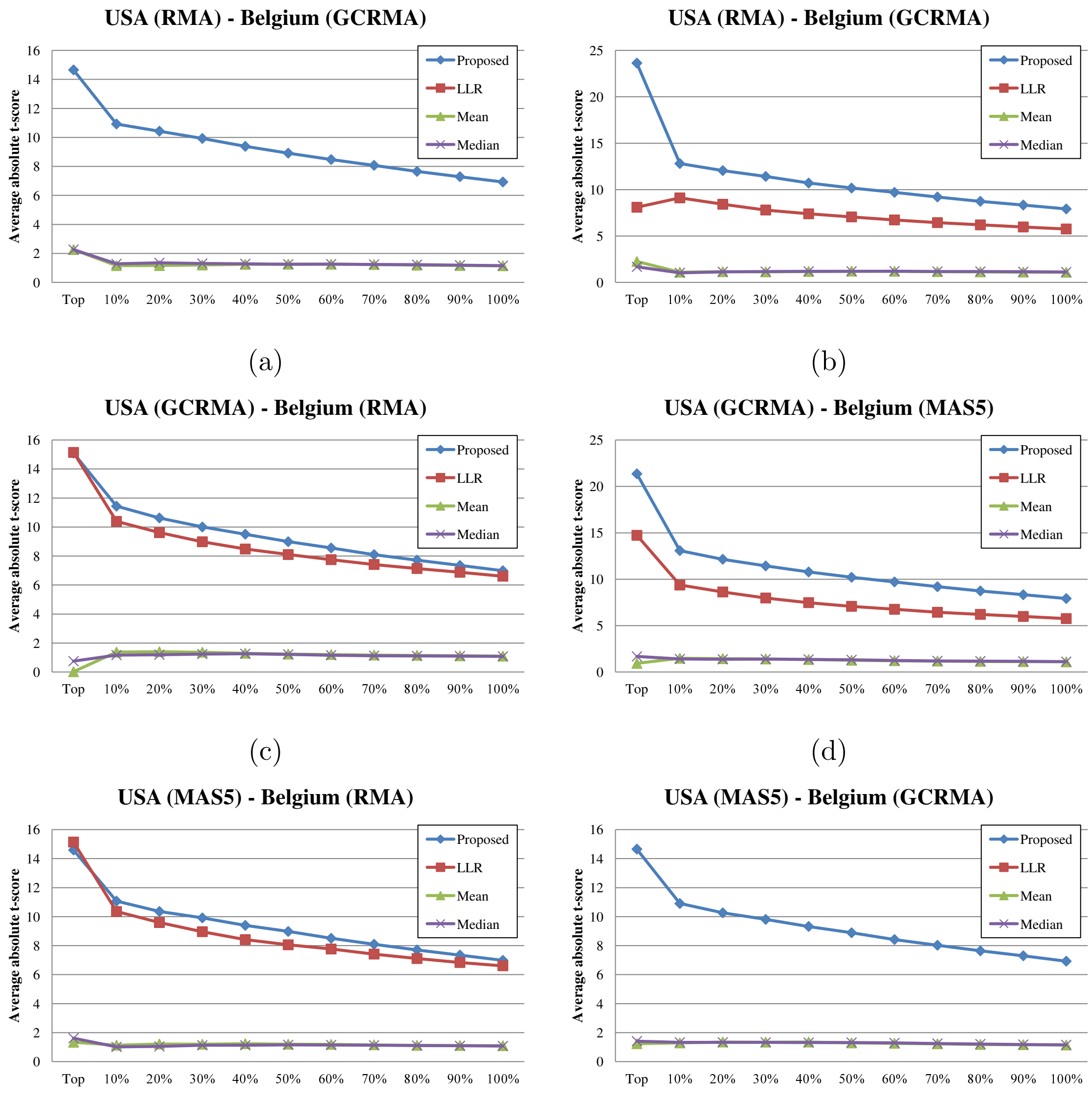
(d)

(e)

(f)

Figure 2.5: Discriminative power of pathway markers across re-normalized datasets [1]. The pathway markers have been ranked and sorted using the first dataset and their discriminative power has been re-evaluated using the second dataset. In all experiments, the datasets have been first normalized using the same normalization method.

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.



(a) (b)

(c) (d)

(e) (f)

Figure 2.6: Discriminative power of pathway markers identified on re-normalized USA dataset and evaluated on re-normalized Belgium dataset [1]. The pathway markers have been ranked and sorted using the USA dataset and their discriminative power has been re-evaluated using the Belgium dataset. In these experiments, the two datasets have been normalized using different normalization methods.

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

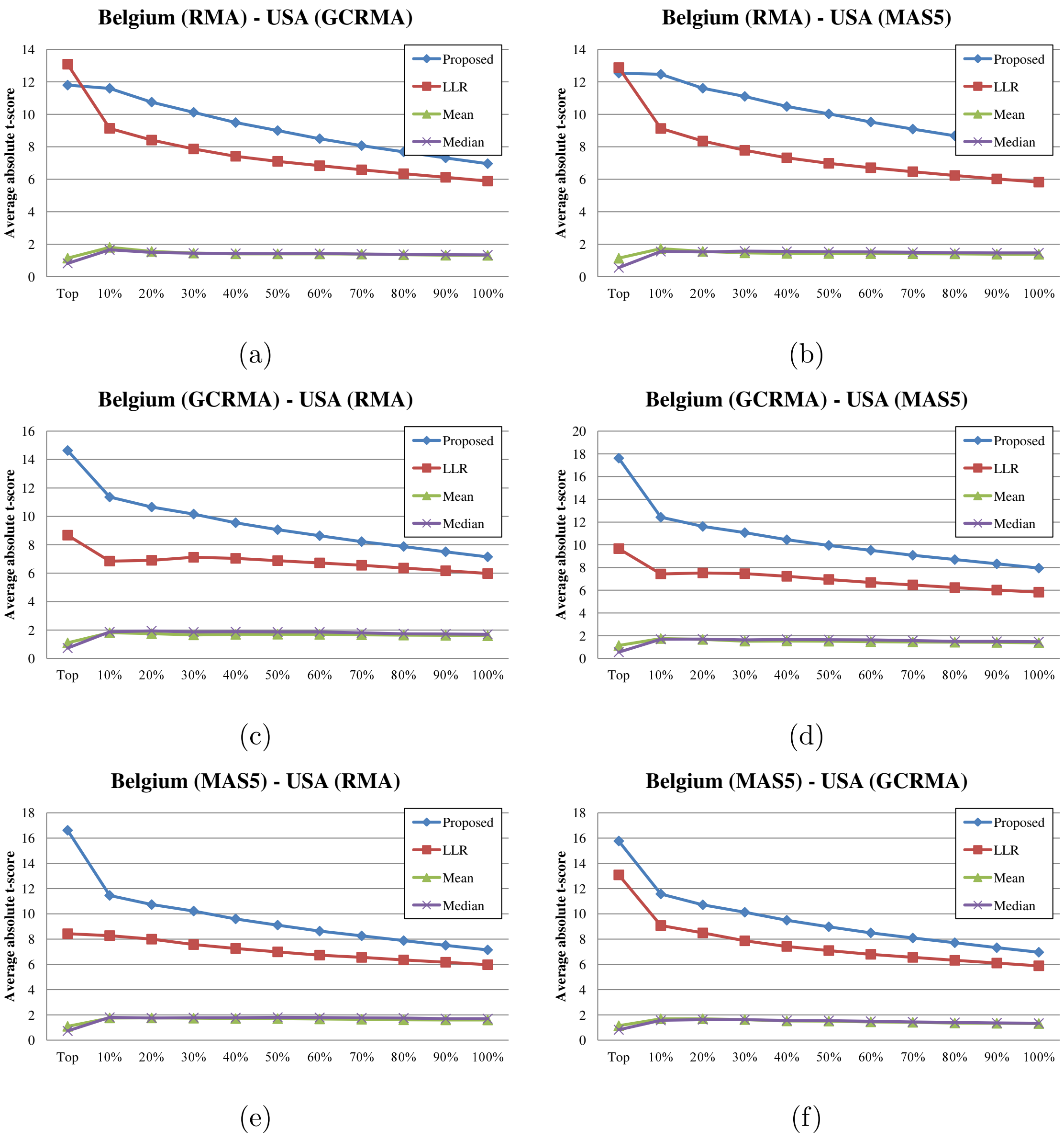
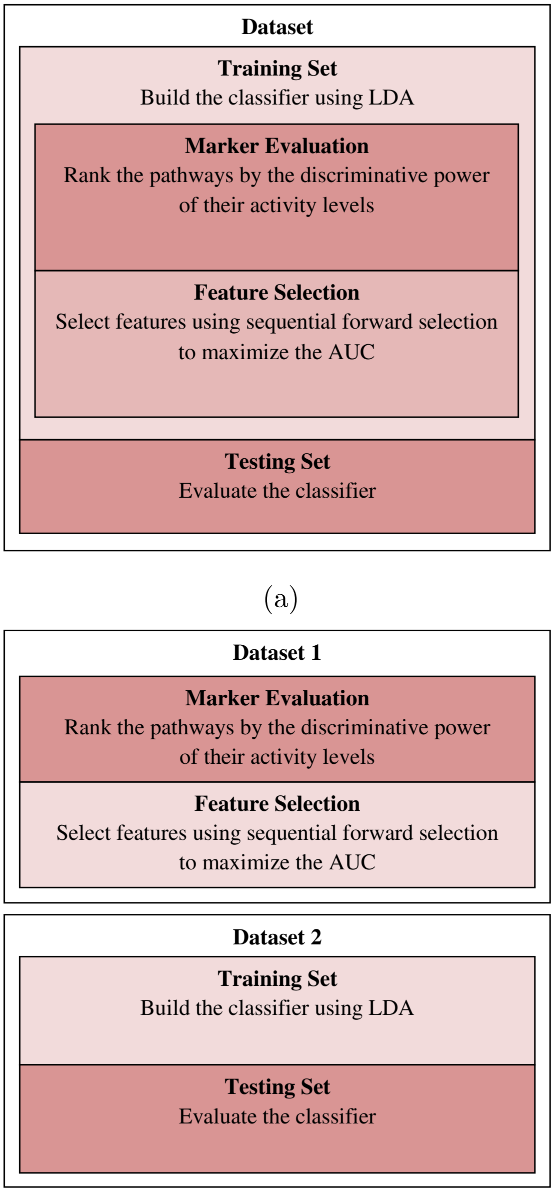


Figure 2.7: Discriminative power of pathway markers identified on re-normalized Belgium dataset and evaluated on re-normalized USA dataset [1]. The pathway markers have been ranked and sorted using the Belgium dataset and their discriminative power has been re-evaluated using the USA dataset. In these experiments, the two datasets have been normalized using different normalization methods.

100 random partitions to ensure reliable results, and we report the average AUC as the measure of overall classification performance.


Dataset
Training Set
Build the classifier using LDA
Marker Evaluation
Rank the pathways by the discriminative power of their activity levels
Feature Selection
Select features using sequential forward selection to maximize the AUC
Testing Set
Evaluate the classifier
Dataset 1
Marker Evaluation
Rank the pathways by the discriminative power of their activity levels
Feature Selection
Select features using sequential forward selection to maximize the AUC
Dataset 2
Training Set
Build the classifier using LDA
Testing Set
Evaluate the classifier

(a)

(b)

Figure 2.8: Experimental setup for evaluating the classification performance [1]. (a) The set-up for the within-dataset experiment. (b) The set-up for the cross-dataset experiment.

Figure 2.9 shows how the respective classifiers that use different pathway activity inference methods perform on different datasets. As we can see in Figure 2.9, among

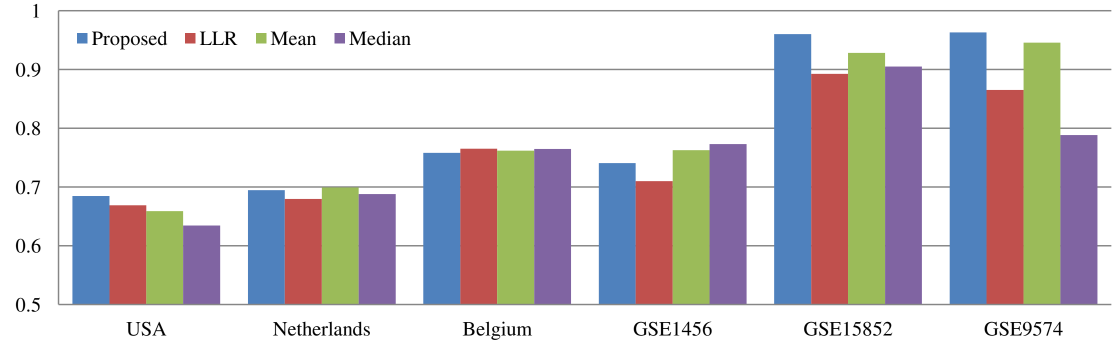


Figure 2.9: Classification performance for within-dataset experiments [1]. The bars show the classification performance (average AUC) of different pathway activity inference methods evaluated on various breast cancer datasets.

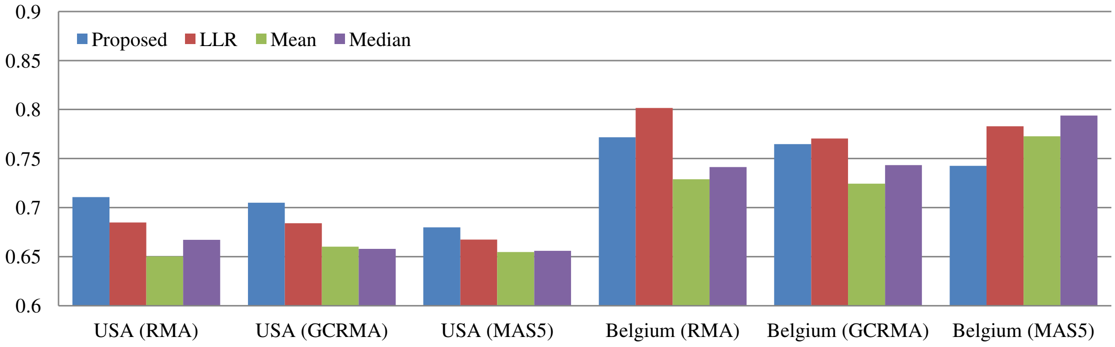


Figure 2.10: Classification performance for within-dataset experiments based on renormalized datasets [1]. We repeated the within-dataset classification experiments based on the USA and the Belgium datasets after normalizing the raw data using three different normalization methods: RMA, GCRMA, and MAS5.

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 dataset after normalizing the raw data using different normalization methods. These results are summarized in Figure 2.10. We can see from Figure 2.10 that the proposed method yields the best performance on the 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.

2.3.3 Reproducibility of the Pathway Markers Identified by Proposed Method

To assess the reproducibility of the pathway markers, we performed the following cross-dataset experiments based on a similar set-up that has been utilized in previous studies [11–14]. 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 three folds, where two folds were used for marker evaluation and the remaining fold was used for feature selection. The second dataset was randomly divided into five folds, where four folds 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 set-up is shown in Figure 2.8b. 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, Netherlands, and Belgium) among the six.

The results of the cross-dataset classification experiments are shown in Figure 2.11. As we can see from this figure, the proposed rank-based inference scheme typically outperforms other methods in terms of reproducibility. Furthermore, we can also observe that proposed method yields consistent classification performance across experiments, while the performance of other inference methods are much more sensitive on the choice of the dataset. Next, we repeated the cross-dataset classifi-

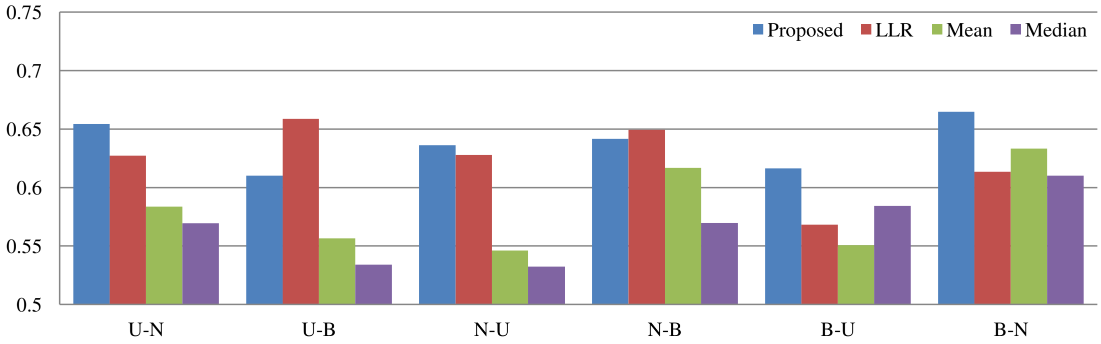


Figure 2.11: Classification performance for cross-dataset experiments [1]. The bars show the cross-dataset classification performance (average AUC) of different pathway activity inference methods. The first dataset was used for selecting the pathway markers and the second dataset was used for training and evaluation of the classifier. The three largest breast cancer datasets were used: USA (U), Netherlands (N), Belgium (B).

cation experiments based on the USA and the Belgium datasets after normalizing the raw data using RMA, GCRMA, and MAS5. As shown in Figure 2.12, 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 2.13 and 2.14. 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 [13] when used with GCRMA (see Figures 2.12, 2.13 and 2.14)

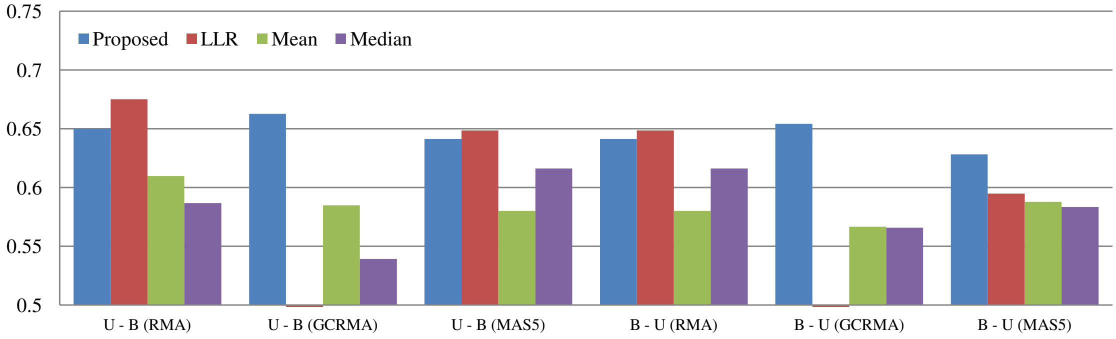


Figure 2.12: Classification performance for cross-dataset experiments based on re-normalized datasets [1]. We repeated the cross-dataset experiments based on the USA and the Belgium datasets after normalizing the raw data using same normalization method.

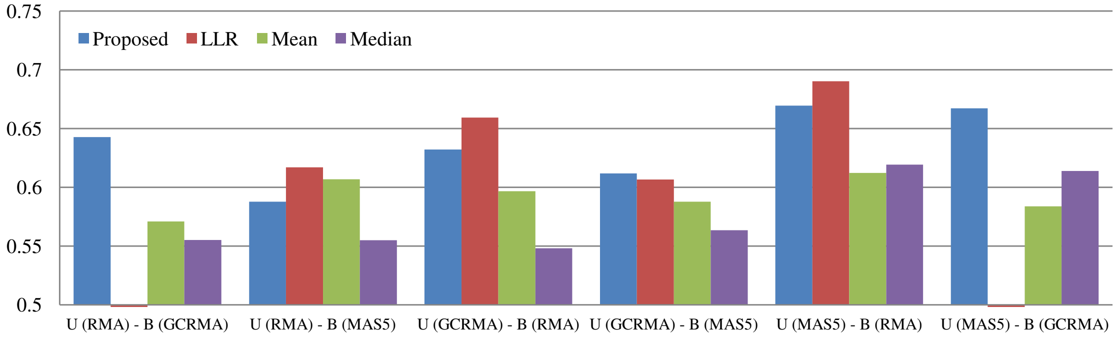


Figure 2.13: Classification performance for re-normalized USA-Belgium cross-dataset experiments [1]. The USA dataset was used to select the pathway markers and the Belgium dataset was for training and evaluating the classifier. In these experiments the two datasets have been normalized using different normalization methods.

2.4 Conclusions

In this chapter, 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

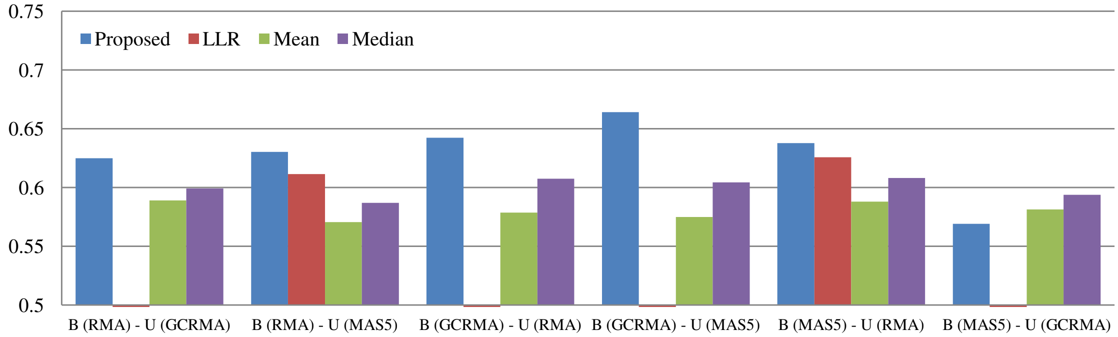


Figure 2.14: Classification performance for re-normalized Belgium-USA cross-dataset experiments [1]. The Belgium dataset was used to select the pathway markers and the USA dataset was for training and evaluating the classifier. In these experiments the two datasets have been normalized using different normalization methods.

strategies that have been recently proposed in the field: namely, the probabilistic pathway activity inference method [13] and the ranking-based relative gene expression analysis approach [15]. 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.

3. SIMULTANEOUS IDENTIFICATION OF ROBUST SYNERGISTIC SUBNETWORK MARKERS FOR EFFECTIVE CANCER PROGNOSIS[*]

3.1 Introduction

So far, it has been shown that pathway markers tend to be more effective and robust compared to traditional gene markers [1, 8, 9, 12, 13]. Unfortunately, the usefulness of pathway markers is practically limited by our incomplete pathway knowledge. In fact, currently known pathways cover only a relatively small number of genes, hence the reliance on pathway markers may result in excluding crucial genes that may play important roles in determining the phenotypes of interest.

The concept of subnetwork markers has been originally proposed to address the weakness of pathway markers [10, 11]. The main idea is to overlay the PPI network with the gene expression data to identify potential "subnetwork markers," which consist of discriminative genes whose protein products interact with each other, hence connected in the PPI network. Conceptually, we can find such subnetwork markers by identifying subnetwork regions that undergo significant differential expression across different phenotypes, and the detected subnetwork markers may potentially correspond to functional modules – such as signaling pathways or protein complexes – in the underlying biological network. PPI networks provide a much better gene coverage compared to the set of currently known pathways, hence this network-based approach can essentially overcome the major shortcoming of the pathway-based approach.

Until now, several different strategies have been proposed for identifying sub-

[*] Part of this chapter is reprinted from "Navadon Khunlertgit, and Byung-Jun Yoon. Simultaneous Identification of Robust Synergistic Subnetwork Markers for Effective Cancer Prognosis. *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2014, no. 1, p. 19, 2014. doi:10.1186/s13637-014-0019-9" [2] © [2014] Navadon Khunlertgit and Byung-Jun Yoon.

network markers. For example, Chuang et al. [11] proposed an efficient algorithm for finding subnetwork markers, where they first identify highly discriminative seed genes and then greedily grow the subnetworks around the seed genes to maximize the mutual information between the average z-score of the member genes and the class label. More recently, Su et al. [14] proposed a different strategy, where differentially expressed linear paths are found by dynamic programming and overlapping paths are combined to obtain discriminative subnetwork markers. Both studies [11, 14] have shown that subnetwork markers can lead to more accurate and robust classifiers, compared to pathway markers.

In this chapter, we propose a novel method for identifying effective subnetwork markers for predicting cancer prognosis. The proposed method is based on an efficient message-passing algorithm, called affinity propagation, which can be used to efficiently identify clusters of discriminative and synergistic genes whose protein products are either connected or closely located in the PPI network. Unlike previous subnetwork marker identification methods, the proposed method can simultaneous predict multiple subnetwork markers, which are mutually exclusive and have the potential to accurate predict cancer prognosis in a synergistic manner. Based on several independent breast cancer datasets, we demonstrate that the proposed method can identify better prognostic markers that have improved reproducibility and higher discriminative power compared to the markers identified by previous methods. Preliminary results of this work have been originally presented at [32].

3.2 Materials and Methods

3.2.1 Datasets

We obtained four independent breast cancer microarray gene expression datasets from previous studies, which we refer to as the USA dataset (GSE2034) [6], Netherlands dataset (NKI295) [18], Belgium dataset (GSE7390) [19], and Sweden dataset (GSE1456) [20], respectively. The USA, Belgium, Sweden dataset were profiled on the Affymetrix U133a platform and downloaded from the Gene Expression Omnibus (GEO) website [24]. The Netherlands dataset was profiled on a custom Agilent microarray platform and it was downloaded from the Stanford website [23]. The USA dataset contains the gene expression profiles of 286 breast cancer patients, the Netherlands dataset contains the profiles of 295 patients, the Belgium dataset contains the profiles of 198 patients, and the Sweden dataset contains the profiles obtained from 159 patients. In this study, gene expression profiles of the patients for whom metastasis had been detected within 5 years of surgery were labeled as "metastatic", while the remaining profiles were labeled as "non-metastatic." The USA, Netherlands, Belgium, and Sweden datasets respectively contain 106, 78, 35, and 35 metastatic profiles. The human protein-protein interaction network used in this study was obtained from a previous study on subnetwork marker identification by Chuang et al. [11], which consists of 11,203 proteins and 57,235 interactions. We overlaid the gene expression data in the four breast cancer datasets with this PPI network, by mapping each gene to the corresponding protein in the network. After removing the proteins that do not have corresponding genes in all four datasets, we obtained an induced network with 26,150 interactions among 4,936 proteins.

3.2.2 The Affinity Propagation Algorithm: A Brief Overview

In order to identify discriminative subnetwork markers, we apply *affinity propagation* [33], an efficient clustering algorithm based on a message-passing approach. In affinity propagation, real-valued messages are iteratively exchanged between data points until a good set of exemplars (i.e., representative data points) are identified. The data points are clustered around the exemplars that best represent them, which gives rise to clusters that consist of similar data points. During the message-passing process, two different types of messages are exchanged between data points: *responsibility* and *availability*. The responsibility $r(i,k)$ measures the suitability of the data point k to be an exemplar of the data point i, considering other potential exemplars. The availability $a(i,k)$ measures the appropriateness of choosing the data point k as the exemplar for the data point i, based on the choice of other data points. At each iteration, these messages are updated as follows:

$$r(i,k) \leftarrow s(i,k) - \max_{k' \text{ s.t. } k' \neq k} \left\{ a(i,k^{'}) + s(i,k^{'}) \right\} \tag{3.1}$$

$$a(i,k) \leftarrow \min \left\{ 0, r(k,k) + \sum_{i' \text{ s.t. } i' \notin \{i,k\}} \max \left\{ 0, r(i^{'},k) \right\} \right\}, \tag{3.2}$$

where $s(i,k)$ is the similarity between the data points i and k, used as the input of the clustering algorithm. This similarity $s(i,k)$ can be asymmetric. The self-availability is updated in a slightly different way, as shown below:

$$a(k,k) \leftarrow \sum_{i' \text{ s.t. } i' \neq k} \max \left\{ 0, r(i^{'},k) \right\}. \tag{3.3}$$

The data point k that maximizes the sum $a(i,k) + r(i,k)$ is chosen as the exemplar for the data point i, and the algorithm converges if the set of exemplars does not change further.

So far, affinity propagation has been applied to various applications – such as predicting genes from microarray data and clustering facial images – and it has been shown to effectively identify meaningful clusters of data points at a much lower computational cost than traditional clustering methods [33]. One important advantage of affinity propagation is that the number of clusters need not be specified in advance. This is especially useful in our current application, since we neither know how many functional modules are embedded in the biological network at hand nor how many of them are relevant to cancer prognosis, which makes it practically difficult to determine how many subnetwork markers we should look for.

3.2.3 Computing the Similarity Between Genes

In our proposed method, we use affinity propagation to identify clusters – or subnetworks – of discriminative and synergistic genes, whose protein products either interact with each other or are closely located in the PPI network. In order to use affinity propagation to identify the gene clusters, we first have to define the similarity $s(i,k)$ between gene g_i and g_k for all gene pairs. The characteristics of the final clusters – especially, their usefulness as potential subnetwork markers – will critically depend on how we define this similarity. For this reason, we take the following points into consideration when defining $s(i,k)$:

1. The proteins corresponding to the genes in the same cluster should have direct interaction or should be closely located in the PPI network.

2. Every gene in a potential subnetwork marker should have sufficient discriminative power to distinguish between the two class labels (metastatic vs. non-metastatic).

3. The discriminative power to distinguish between the two class labels should be increased by combining genes within the same cluster.

Based on these considerations, we define the similarity $s(i,k)$ as follows:

$$s(i,k) = t_k + \min\Big\{t_{ik} - t_i, t_{ik} - t_k\Big\} - \alpha\Big|t_i - t_k\Big| \tag{3.4}$$

if the shortest distance $d(i,k)$ between the protein products of the genes g_i and g_k in the PPI network satisfies $d(i,k) \leq 2$. Otherwise, we set the similarity to $s(i,k) = -\infty$. The discriminative power of a given gene is measured in terms of the t-test statistics score of the LLR between the two class labels, and t_i and t_k are the t-test scores of g_i and g_k, respectively. Similarly, t_{ik} is the t-test score of the combined LLRs of g_i and g_k which is computed by summing up the LLRs of the two genes. This term, t_{ik}, reflects the discriminative power of the gene pair (g_i, g_k) after combining them. The self-similarity was set to $s(k,k) = c$ for all k, where the constant c was chosen such that $s(i,k) \geq c$ for only 1% of all gene pairs (g_i, g_k). Uniform initialization of the self-similarity $s(k,k) = c$ guarantees that every gene in the dataset gets equal chance to be an exemplar at the beginning of the message-passing process.

As shown in (3.4), the similarity $s(i,k)$ between g_i and g_k is defined in an asymmetric way, where the the first term corresponds to the discriminative power of the gene g_k, the second term measures the improvement in discriminative power after combining the two genes g_i and g_k, and the last term corresponds to a penalty term for the difference between t_k and t_i. The parameter $\alpha \in [0,1]$ is used to control the penalty term. According to the above definition, gene g_i regards gene g_k as being "similar" to itself:

1. if g_k has high discriminative power (first term);
2. if combining the two genes increases the overall discriminative power;
3. if both genes have similar discriminative power.

The main reason underlying the asymmetric definition of the similarity $s(i,k)$ is to indicate the direction of similarity. Based on our asymmetric definition, the exemplars of the identified clusters tend to have higher discriminative power compared to other non-exemplars. Intuitively, the gene similarity defined in (3.4) will make the affinity propagation algorithm identify gene clusters that consist of highly discriminative genes that are synergistic to each other and whose protein products are closely located in the PPI network.

3.2.4 Post-Processing the Identified Gene Subnetworks

Although the affinity propagation algorithm can effectively identify subnetworks that consist of discriminative and synergistic genes, the clustering process does not completely rule genes with relatively lower discriminative power out of those subnetworks. As a result, the initial subnetworks that are predicted by affinity propagation may still contain genes with relatively lower discriminative power compared to other genes in the same subnetwork. In order to improve the overall discriminative power of the potential subnetwork markers, we post-processed the initial subnetworks as follows. First, we clustered the genes in a given subnetwork into k groups based on their t-test statistics scores using the k-means clustering algorithm, where k was chosen to be $k = \lfloor \log(\#\text{ of gene in considered subnetwork}) + 1 \rfloor$. After clustering, the genes in the group with the lowest average t-test score were removed from the subnetwork.

3.2.5 Probabilistic Inference of Subnetwork Activity

For estimating the activity level of a subnetwork based on the gene expression profile of a patient, we adopted the probabilistic pathway activity inference method introduced in [13]. Given a subnetwork (or a pathway) with n member genes $\boldsymbol{\mathcal{G}} = \{g_1, g_2, ..., g_n\}$ and the gene expression profile $\boldsymbol{x} = \{x^1, x^2, ..., x^n\}$ of a patient, where x^i is the expression level of the gene g_i, the activity level of the subnetwork is computed by:

$$A(\mathbf{x}) = \sum_{i=1}^{n} \lambda_i \left(x^i\right), \tag{3.5}$$

where $\lambda_i(x^i)$ is the log-likelihood ratio between the two class labels (in this study, metastatic vs. non-metastatic). This is given by

$$\lambda_i(x^i) = \log\left[f_i^1(x^i)/f_i^2(x^i)\right], \tag{3.6}$$

where $f_i^j(x^i)$ is the conditional probability density function (PDF) of x^i under phenotype j. We assume that the gene expression level of g_i under phenotype j follows a Gaussian distribution.

3.3 Results

3.3.1 Statistics of the Identified Subnetwork Markers

For each of the four dataset, we identified potential subnetwork markers using the proposed method and selected the top 50 markers based on their discriminative power, measured in terms of the t-test statistics score of the subnetwork activity. Three different values of α $(= 0.2, 0.5, 0.8)$ were used in our experiments to investigate the effect of the penalty term in (3.4) on the subnetwork marker identification result. Table 3.1 shows the average size of the top 50 subnetworks markers for each dataset and α. The last two columns in the table show the average size of the subnetwork markers identified using the method proposed by Chuang et al. [11], which we refer to as the "greedy" method, for simplicity. Two different values of r were used for this greedy method. This parameter r specifies the minimum improvement rate of the discriminative power of a subnetwork marker. The greedy method stops when extending the subnetwork marker by adding a neighboring gene does not improve the marker's discriminative power by at least the specified rate r. We tested the greedy method with $r = 0.05$ (or 5% minimum required improvement) which is the same as in [11]. We also tested the method with a lower rate $r = 0.001$ (or 0.1% minimum required improvement) in order to allow the greedy search to continue even if the improvement is not very significant and find out how a lower rate affects the subnetwork size and its discriminative power. As we can see from Table 3.1, the size of the network decreases as α gets larger. In fact, a large α tends to cluster only genes with similar discriminative power (i.e., genes with similar t-test scores), thereby yielding smaller subnetworks with fewer genes.

Table 3.1: Average size of the identified subnetwork markers [2].

Dataset	Proposed Method			greedy	
	$\alpha = 0.2$	$\alpha = 0.5$	$\alpha = 0.8$	$r = 0.05$	$r = 0.001$
USA	52.58	35.58	16.96	3.94	5.22
Netherlands	52.62	31.2	15.9	5.18	7.20
Belgium	37.64	20.2	12.3	4.12	5.48
Sweden	33.18	21.38	14.16	3.66	4.82

Similar trends can be also observed in Table 3.2, which shows the total number of unique genes in the top 50 subnetwork markers. As see can see in this table, a larger α results in a smaller number of unique genes in the top subnetwork markers, as each marker tends to get smaller.

Table 3.3 shows the total number of the common genes between the identified subnetworks using different α. We can see that around 77% of genes included in identified subnetworks using smaller α are also found in the subnetworks identified with larger α.

We examined the overlap between the subnetworks identified on different datasets, which is defined as the number of genes in the intersection divided by the number

Table 3.2: Total number of unique genes in the identified subnetwork markers [2].

Dataset	Proposed Method			greedy	
	$\alpha = 0.2$	$\alpha = 0.5$	$\alpha = 0.8$	$r = 0.05$	$r = 0.001$
USA	2,629	1,779	848	169	217
Netherlands	2,631	1,560	795	158	222
Belgium	1,916	1,010	615	113	149
Sweden	1,695	1,069	708	123	166

Table 3.3: Total number of common genes between the top subnetwork markers identified using different α [2].

Dataset	$\alpha = 0.2 \cap \alpha = 0.5$	$\alpha = 0.2 \cap \alpha = 0.8$	$\alpha = 0.5 \cap \alpha = 0.8$
USA	1,612	660	561
Netherlands	1,382	646	488
Belgium	767	454	372
Sweden	802	466	387

of genes in the union. As shown in Table 3.4, we can see that the average overlap is typically close to (or above) 20%, which is larger than the greedy method as well as the overlap reported in [11] (12.7%).

Table 3.4: Overlap between the top subnetwork markers identified on different datasets [2].

Dataset	Proposed Method ($\alpha = 0.5$)	greedy	
		$r = 0.05$	$r = 0.001$
USA - Netherlands	25.10%	8.28%	7.60%
USA - Belgium	19.04%	5.22%	6.09%
USA - Sweden	19.71%	5.32%	5.51%
Netherlands - Belgium	18.11%	8.84%	10.09%
Netherlands - Sweden	18.85%	7.92%	7.78%
Belgium - Sweden	17.13%	11.57%	11.31%

3.3.2 Computational Cost for Subnetwork Marker Identification

In order to evaluate the computational complexity of the proposed method, we computed the total CPU time that is needed for identifying the top 50 subnetwork markers on each dataset. We considered three different values of $\alpha\,(= 0.2, 0.5, 0.8)$ that were used in our simulations. For comparison, we also estimated the total CPU time for the greedy method that was previously proposed. It should be noted that the two methods take completely different approaches for identifying multiple markers. In our proposed method, all potential subnetwork markers (whose total number exceeds 50) are *simultaneously* identified, hence we need to rank the potential markers to select the top 50 markers with the highest discriminative power. As a result, for our proposed method, the total CPU time includes the time for calculating the similarity between genes, potential subnetwork marker identification through affinity propagation, and post-processing and ranking the subnetwork markers. On the other hand, for the greedy method, we measured the CPU time for calculating the discriminative power of the genes and iteratively searching for the top 50 markers. Since the greedy method finds one marker at a time, the search process needs to be repeated to find multiple markers. Figure 3.1 shows the total CPU time of the two methods for different parameters. All experiments were performed on a desktop computer with a 3.06 GHz Intel Core i3 CPU and 4GB 1333 MHz DDR3 memory. The results show that the proposed method is computationally more efficient for the given task as it can simultaneously identify all potential markers without repeating the search process multiple times. Unless one is interested in predicting only a few top markers, the proposed method provides a clear advantage over the previous greedy method. Figure 3.1 also shows that using different parameters do not affect the overall CPU time significantly.

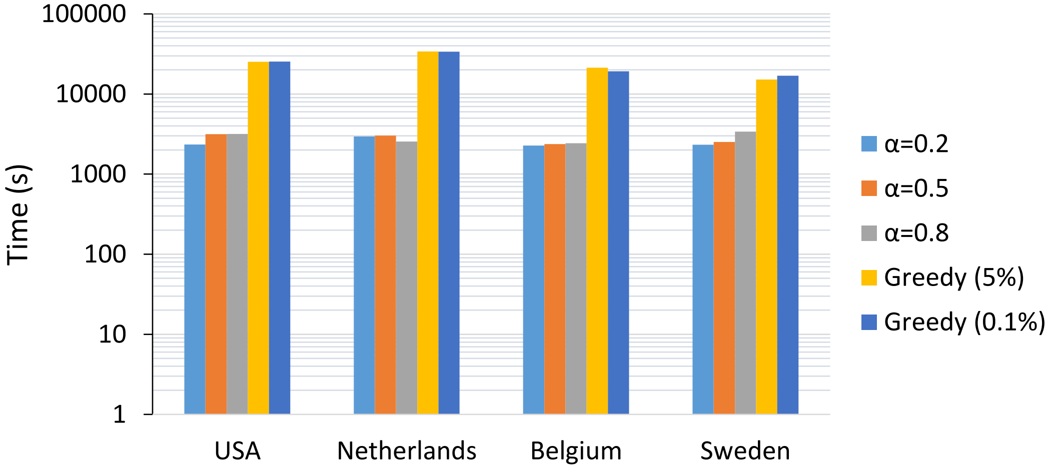
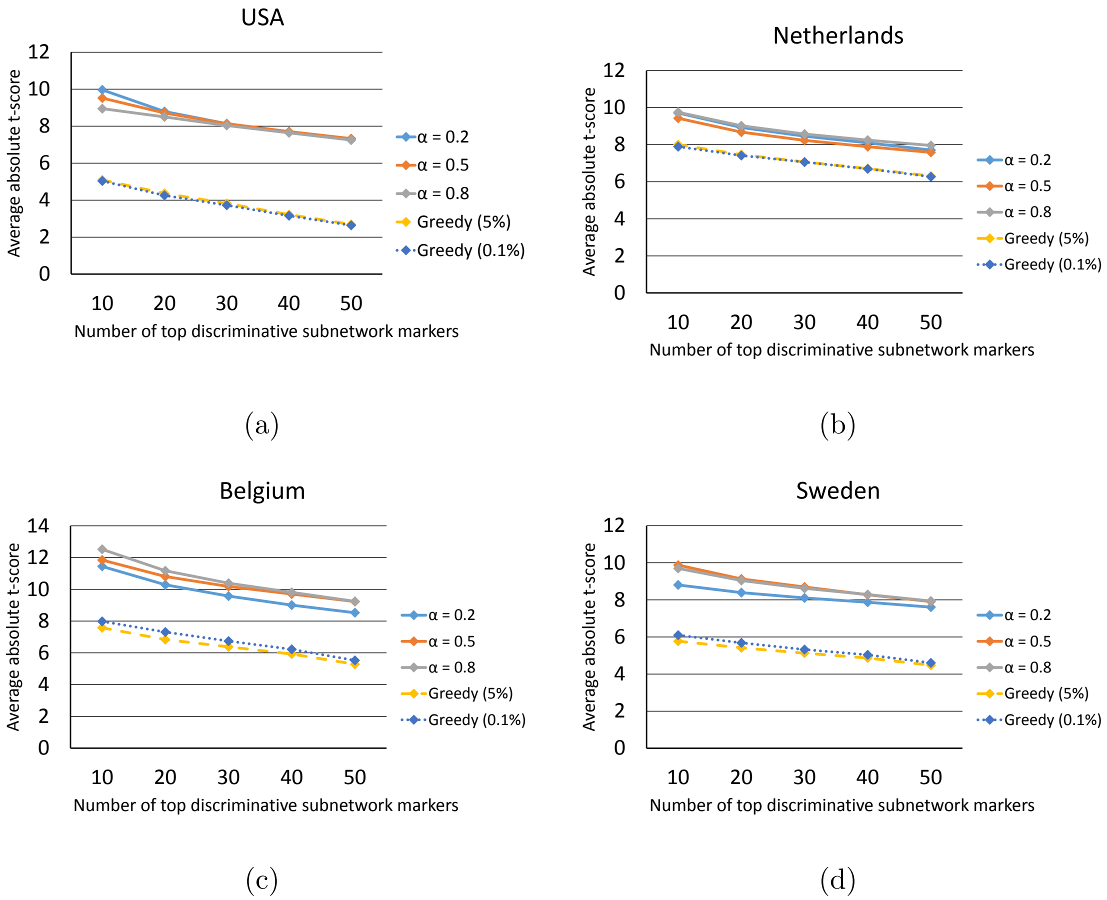


Figure 3.1: Total CPU Time for identifying the top 50 subnetwork markers [2]. We evaluated the computational complexity of the proposed method by estimating the total CPU time needed for identifying the top 50 subnetwork markers in a given dataset. We compared our method with the previously proposed greedy method for a number of different parameters.

3.3.3 Discriminative Power of the Subnetwork Markers

We evaluated the discriminative power of the predicted subnetwork markers by following a similar procedure as in previous studies [12, 13]. For each subnetwork marker identified using the proposed method, we first inferred its activity level for the gene expression profile of each patient, and then computed the t-test score of the the inferred subnetwork activity level. Next, we sorted the subnetwork markers according to their absolute t-test score in descending order. We then computed the average absolute t-test score of the top $K = 10, 20, 30, 40, 50$ subnetwork markers, as shown in Figure 3.2.

The horizontal axis in Figure 3.2 corresponds to K and the vertical axis corresponds to the mean absolute t-test score of the top K subnetwork markers. We



(a)

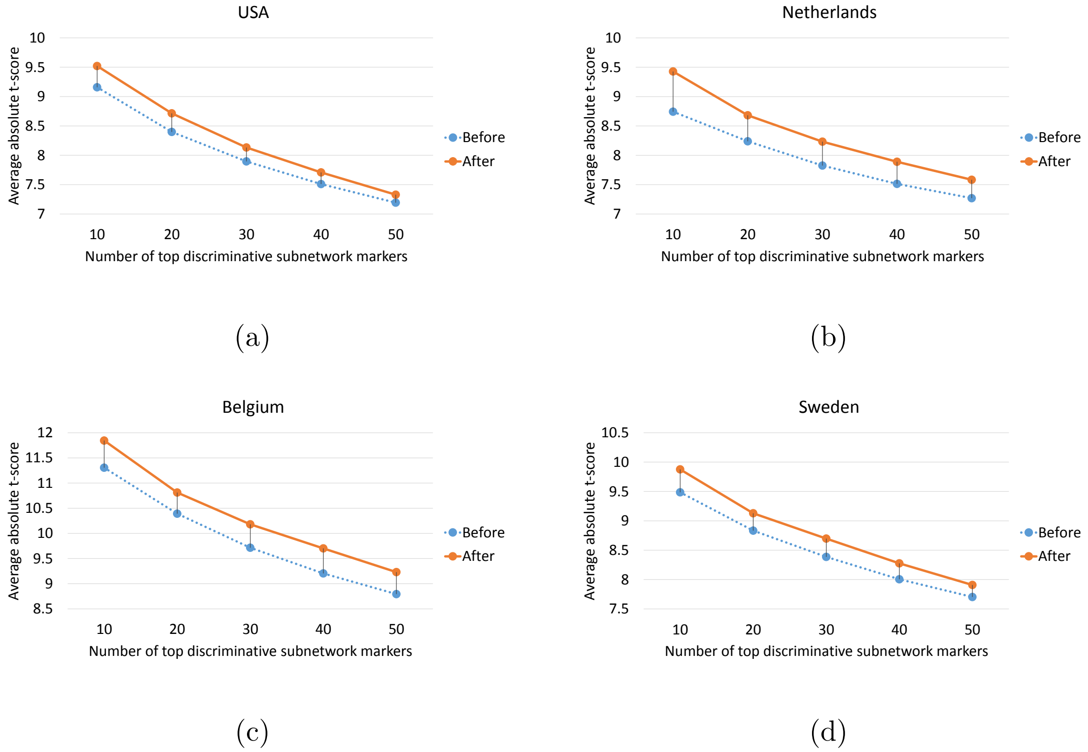
(b)

(c)

(d)

Figure 3.2: Discriminative power of the identified subnetwork markers from various methods [2]. We computed the mean absolute t-test statistics score of the top K=10, 20, 30, 40, and 50 subnetwork markers identified by different methods.

compared the discriminative power of the subnetwork markers predicted by the proposed method with the discriminative power of the subnetworks predicted by the greedy method proposed in [11]. The activity level of these subnetworks (identified by the greedy method) was inferred based on the same scheme that was originally used in [11]. As we can see from Figure 3.2, the proposed method typically finds subnetwork markers with comparable or slightly higher discriminative power compared to the previous greedy method, although both methods work very well. In this



(a) (b)

(c) (d)

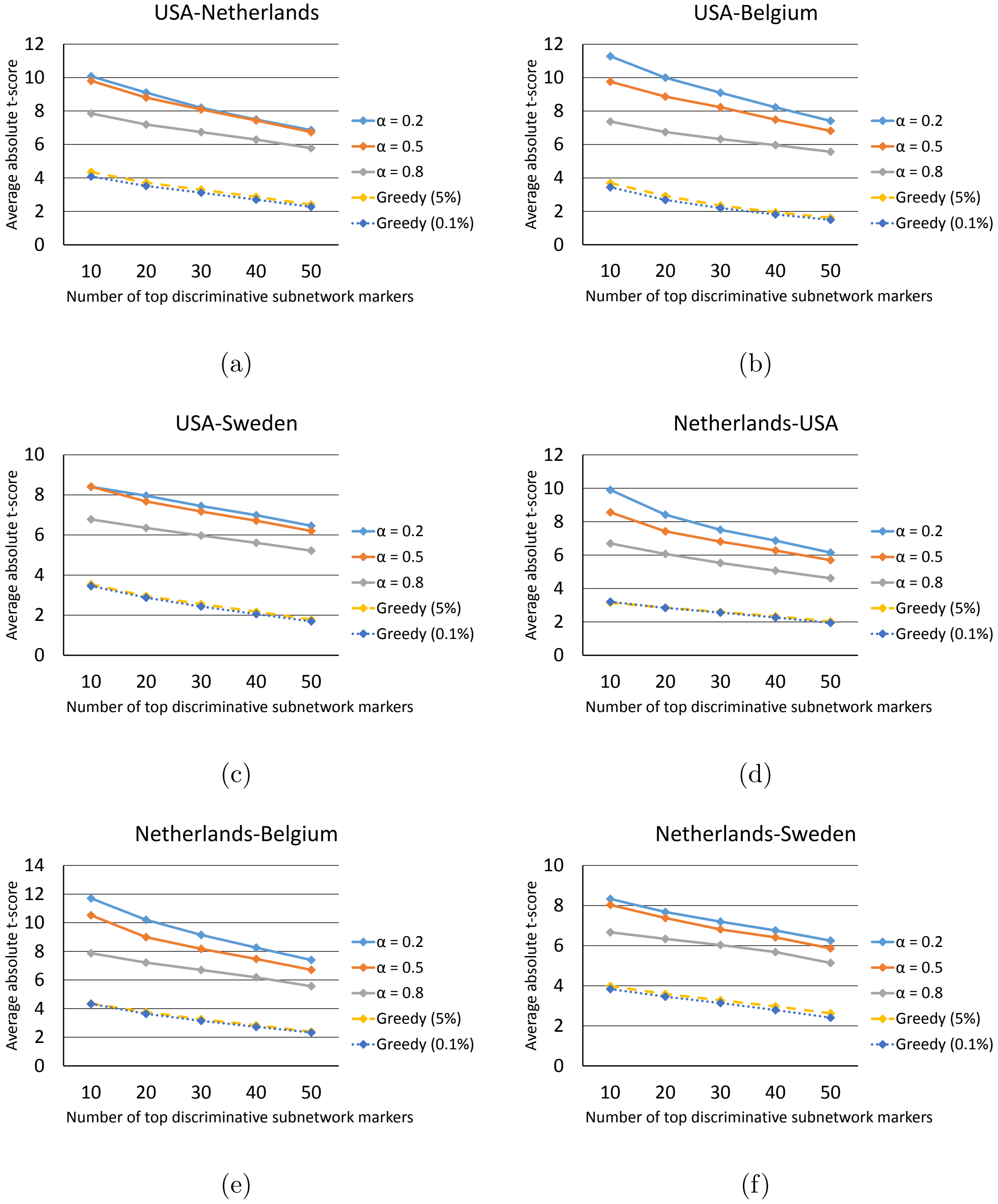
Figure 3.3: Comparison of discriminative power of subnetwork markers identified by the proposed method with and without post-processing step [2]. We computed the mean absolute t-score of the top K=10, 20, 30, 40, and 50 markers for all datasets identified by the proposed method with and without post-processing step.

experiment, the parameter α did not significantly affect the average discriminative power of the subnetwork markers identified by the proposed method.

We also investigated the impact of the post-processing step by comparing the discriminative power of the subnetwork markers before and after post-processing. Figure 3.3 shows the results obtained using $\alpha = 0.5$. We can see that the discriminative power of the top 50 subnetwork markers improves as a result of the post-processing step, during which we remove the genes that have relatively lower discriminative power.

Next, to test the reproducibility of the subnetwork markers identified by the proposed method, we performed cross-dataset experiments as follows. First, we identified subnetwork markers using the proposed method on one of the datasets and ranked the markers based on their absolute t-test statistics score. After ranking the subnetwork markers, we re-evaluated the discriminative power of the top 50 markers on a different dataset. This experiment allows us to find out how much discriminative power is retained by the top predicted markers in a different, and independent, dataset. The cross-dataset experiments are shown in Figure 3.4 and Figure 3.5, where we can see that the markers identified by the proposed method remain highly discriminative across datasets. This is in clear contrast to the subnetwork markers identified by the greedy method [11], for which we can typically observe a sharp decrease in discriminative power when applied to an independent dataset that was not used for predicting the markers. Interestingly, we can also see that the proposed method finds effective markers that retain high discriminative power even on an independent gene expression dataset profiled on a different microarray platform. For example, in Figure 3.4a, the subnetwork markers were first identified using the USA dataset profiled on an Affymetrix chip, and then evaluated on the Netherlands dataset profiled on a custom Agilent chip. Figure 3.4a shows that the markers predicted by the proposed method using the first dataset can also effectively discriminate between the two class labels based on the gene expression profiles in the second dataset. Similar trends can also be observed in Figures 3.4d, 3.4e, 3.4f, 3.5b, and 3.5e.

One interesting observation we can make from these figures is that a smaller α tends to yield subnetwork markers that retain their discriminative power relatively better across independent datasets. This observation makes intuitively sense, since a larger α tends to penalize genes with different discriminative power thereby giving rise to relatively smaller subnetwork markers that mostly consist of a few highly



(a)

(b)

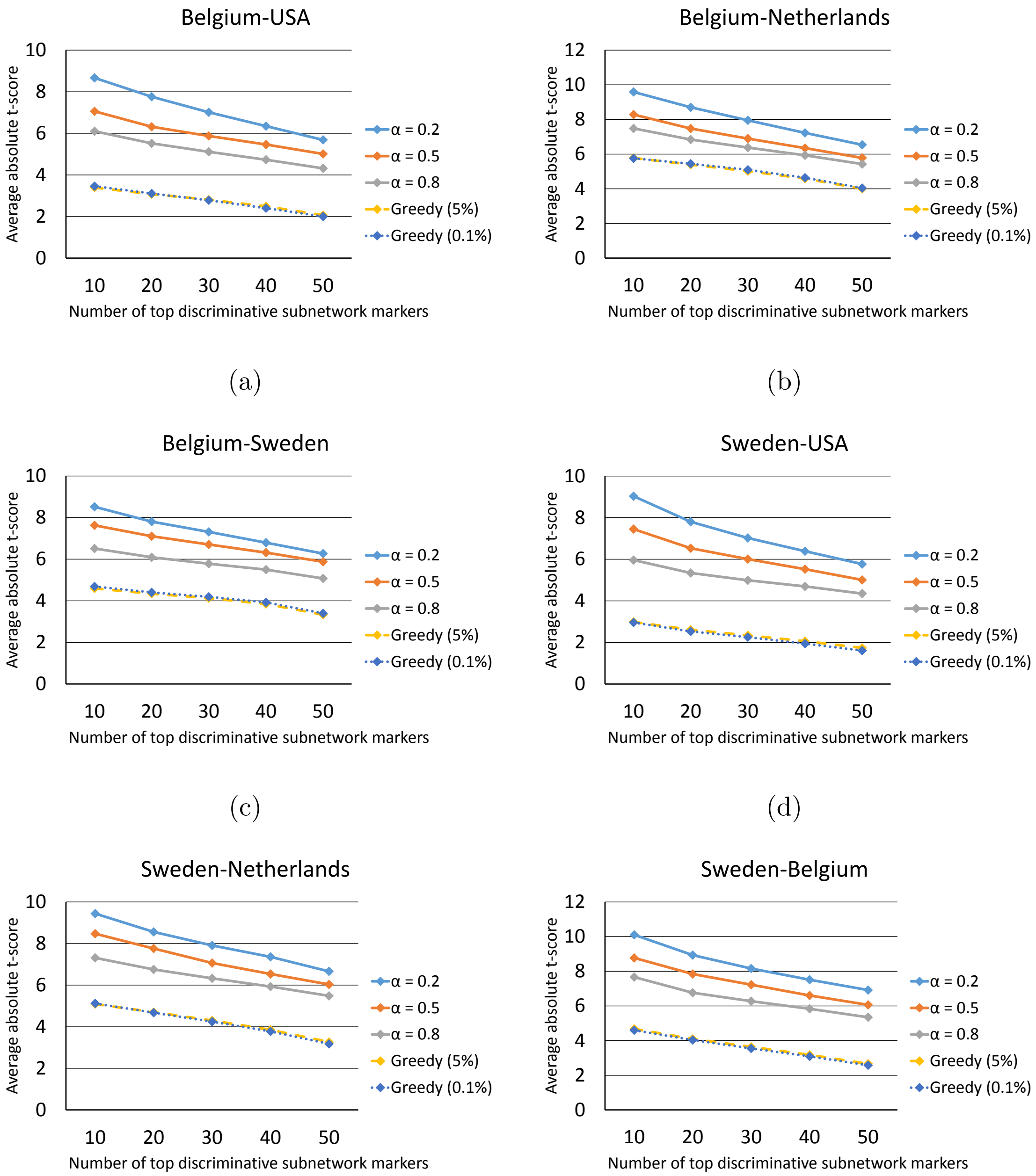
(c)

(d)

(e)

(f)

Figure 3.4: Discriminative power of subnetwork markers identified on USA and Netherlands datasets [2]. We computed the mean absolute t-score of the top K=10, 20, 30, 40, and 50 markers for all datasets. The markers were identified using the first dataset and their discriminative power was evaluated on the second dataset.



(a)

(b)

(c)

(d)

(e)

(f)

Figure 3.5: Discriminative power of subnetwork markers identified on Belgium and Sweden datasets [2]. We computed the mean absolute t-score of the top K=10, 20, 30, 40, and 50 markers for all datasets. The markers were identified using the first dataset and their discriminative power was evaluated on the second dataset.

discriminative genes that may not be necessarily synergistic. This increases the risk of overfitting the data, thereby degrading the effectiveness of the predicted markers on other independent datasets.

3.3.4 Evaluating the Reproducibility of the Predicted Subnetwork Markers

In order to evaluate the efficacy of the predicted subnetwork markers in cancer prognosis, we performed five-fold cross-validation experiments based on a similar set-up that has been commonly used in previous studies [1, 11–14].

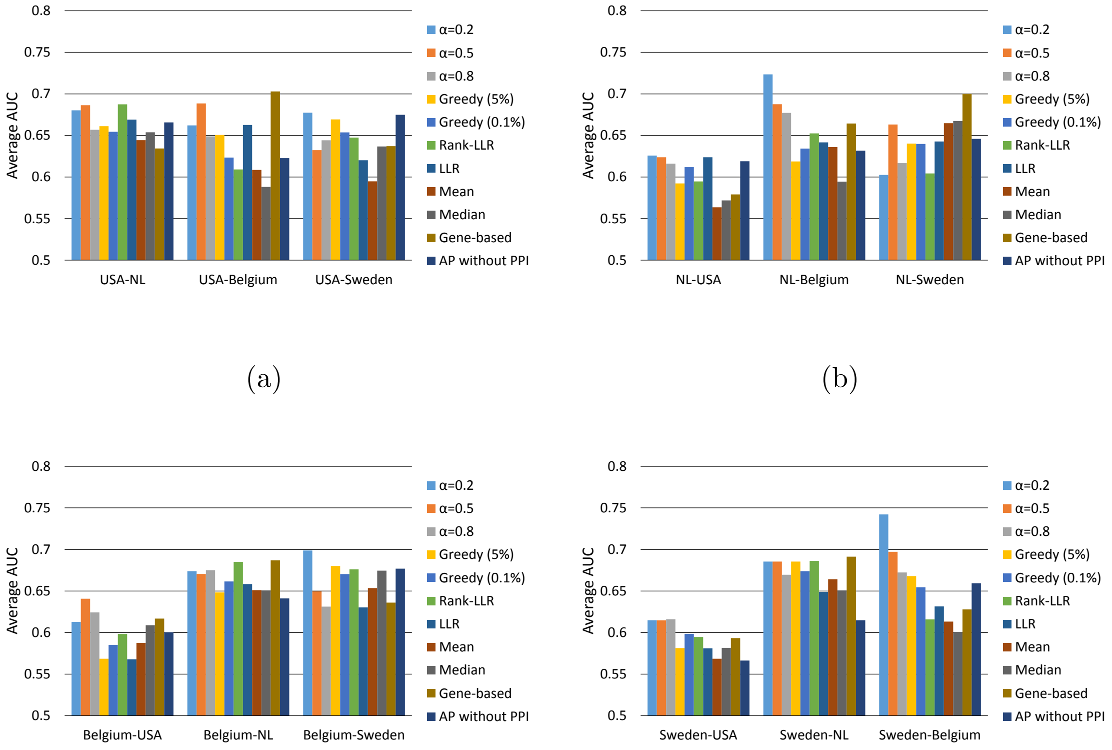
Considering that our ultimate goal is to identify effective subnetwork markers that can be used for building robust classifiers that can accurately predict breast cancer prognosis, it is important to verify whether the predicted markers can actually lead to better classifiers whose performance can be reproduced on independent datasets. For this purpose, we performed the following cross-dataset experiments.

First of all, we selected one of the four breast cancer datasets just for identifying the potential subnetwork markers and selecting the optimal feature set (i.e., the set of markers to be used for building the classifier). To select the optimal set of features, we randomly divided the chosen dataset into three folds, where two folds (marker-evaluation set) were used for evaluating the discriminative power of the subnetwork markers and the remaining one fold (feature-selection set) was used for selecting the features to be used in the classifier. We used the entire set for estimating the class conditional probability density functions that are needed for the pathway activity inference [13].

We evaluated the discriminative power of all potential subnetwork markers based on the marker-evaluation set, selected the top 50 markers, and sorted them according to their absolute t-test score in descending order. Initially, we built a classifier based on LDA, where only the top subnetwork marker was included in the feature set. The classifier was trained on the marker-evaluation set and its classification performance was assessed by measuring the AUC on the feature-selection set. Subsequently, we added the next best subnetwork marker to the feature set, re-trained and re-evaluated the classifier, and kept the subnetwork marker only if the AUC increased. We repeated this process for the top 50 subnetwork markers.

Next, we chose a different dataset to train an LDA classifier (using the markers selected from the first dataset) and evaluate its performance. For this, the second dataset was randomly divided into five folds, where four folds were used for training (without reselecting the features) and the rest was used for computing the AUC. The entire process was repeated for 100 random partitions and we report the average AUC as the performance measure. Similar experiments have been performed to evaluate the classification performance of previous methods, including the greedy subnetwork marker identification method [11] as well as a number of pathway-based classification methods: Rank-LLR [1], LLR [13], Mean, and Median [9]. Each method uses a different way to infer the pathway activity level based on the expression levels of its gene members. For example, Mean (or Median) method uses the mean (or median) expression value of the member genes that belong to the same pathway. LLR and Rank-LLR both utilize the log-likelihood ratio between different phenotypes based on the expression level of each member gene. For pathway markers, we selected the top 50 pathways among the 880 pathways in the C2 curated gene sets in MsigDB [29]. Figure 3.6 summarizes the classification performance of different methods, where we can clearly see that the proposed method leads to more reliable classifiers with much more consistent performance across different breast cancer datasets.

Finally, we also performed within-dataset experiments to investigate the performance of the proposed method, and compare it with previous subnetwork and pathway-based methods. In these experiments, the classifiers were trained and eval-



(a)

(b)

(c)

(d)

Figure 3.6: Reproducibility of various subnetwork and pathway markers [2]. In order to evaluate the reproducibility of various modular markers, we used the first dataset to identify potential markers and select the optimal set of features and the second dataset to train the classifier (using the selected features) and evaluate its performance.

uated on different folds of the same dataset, where a similar five-fold cross-validation set-up was used as before. We first selected a dataset and then randomly divided it into five folds. Four out of the five folds were used as a training set for building the classifier. The remaining one fold was used as a test set for evaluating the classification performance. The subnetwork markers were identified using the entire dataset, and not just the four fold training set, due to the high computational burden

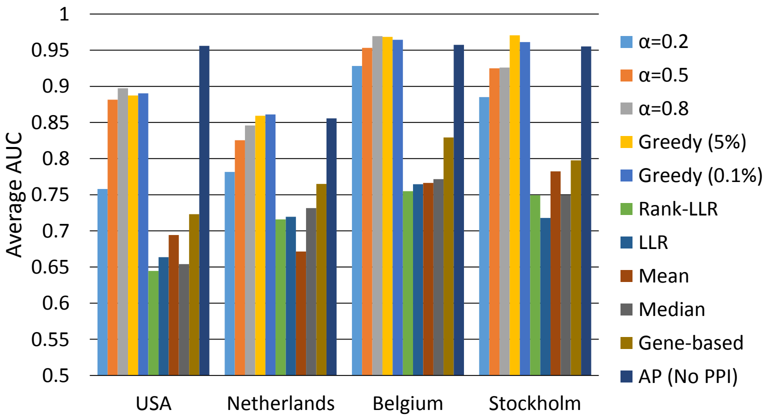


Figure 3.7: Classification performance [2]. We performed cross-validation experiments to evaluate the classification performance of several subnetwork and pathway marker identification methods. The marker were identified using the entire dataset. The classifiers were trained and evaluated on different folds of the same dataset.

for re-identifying the subnetwork markers every time for a large number of random partitions. The results are depicted in Figure 3.7. We can see that classifiers based on subnetwork markers performed significantly better compared to those based on pathway markers. The main reason for this significant performance improvement is the substantially increased coverage of genes, which was the main motivation for identifying subnetwork markers and using them for cancer classification. The proposed subnetwork marker identification method and the greedy method performed both well in the within-dataset experiments, although our proposed method outperformed the greedy method in terms of robustness and reproducibility across different datasets as we have shown before.

3.4 Conclusions

In this chapter, we proposed a novel method for identifying robust and synergistic subnetwork markers that can be used to accurately predict breast cancer prognosis.

Our proposed method utilizes an efficient message-passing algorithm called affinity propagation [33] to identify gene subnetworks that consist of discriminative and synergistic genes whose protein products are known to interact with each other or to be closely located in the protein-protein interaction network. The proposed method allows us to simultaneously identify multiple mutually exclusive subnetwork markers that have the potential to synergistically improve the prediction of breast cancer prognosis. Extensive evaluation based on four large-scale breast cancer datasets demonstrates that the proposed method can predict effective subnetwork markers with high discriminative power and reproducible performance across independent datasets. Furthermore, the predicted markers can be used to construct robust cancer classifiers that can yield more consistent classification performance across datasets compared to other existing methods.

4. INCORPORATING TOPOLOGICAL INFORMATION FOR PREDICTING ROBUST CANCER SUBNETWORK MARKERS IN HUMAN PROTEIN-PROTEIN INTERACTION NETWORK*

4.1 Introduction

In the previous chapter, we utilized a message-passing clustering algorithm to identify subnetwork markers with high-accuracy disease prediction. The method is capable to simultaneously predict multiple non-overlapping subnetwork markers which may lead to cover more genes with lower computational cost compared to the existing methods.

With these advantages, we adopt our previous message-passing based approach while incorporating the topological information from the PPI network to identify the potential functional modules–or subnetworks. Initially, we adopt widely made assumptions that densely connected subnetworks may likely be potential functional modules and that proteins that are not directly connected but interact with similar sets of other proteins may share similar functionalities. We employ association indices to estimate the topological information.

Association indices have been shown to be one of powerful tools for measuring similarity between genes [34]. For example, Jaccard index has been successfully used to measure neighborhood similarity for clustering and constructing Power Graph in the work of Royer et al. [35].

In this chapter, we propose a novel method for incorporating PPI network topological information to enhance identification of subnetwork markers for predicting

*Part of this chapter is reprinted from "Navadon Khunlertgit, and Byung-Jun Yoon. Incorporating topological information for predicting robust cancer subnetwork markers in human protein-protein interaction network. *BMC Bioinformatics*, vol. 17, no. 13, p. 351, 2016. doi:10.1186/s12859-016-1224-1" [3] © [2016] Navadon Khunlertgit and Byung-Jun Yoon.

cancer prognosis. We utilize various association coefficients to estimate the topological similarity and also apply different approaches to integrate into our previous message-passing based method. We assess the identified subnetwork markers and evaluate their discriminative power and their classification performance through experiments using publicly available independent breast cancer gene expression datasets and PPI networks.

4.2 Materials and Methods

4.2.1 Datasets

In this study, we obtained two independent breast cancer microarray gene expression datasets from the public domain, which we refer to as GSE2034 [6] and NKI295 [18]. GSE2034 was profiled on the Affymetrix U133a platform (GPL96) and downloaded from the GEO database [36]. NKI295 was profiled on Agilent Hu25K platform and downloaded from the supplement information from Chang et al. [23]. We used both datasets as published by their original studies. GSE2034 contains expression profiles of 286 breast cancer patients, NKI295 contains expression profiles of 295 patients. For 108 patients in GSE2034 and 78 patients in NKI295, metastasis had been detected within 5 years of surgery. We labeled them as “metastatic”, while the remainder was labeled as “non-metastatic”.

Four publicly available human PPI networks were used in this study which we refer to as Chuang, HPRD, GASOLINE, and BioGRID. Chuang was obtained from a previous study by Chuang et al. [11]. HPRD was downloaded from the Human Protein Reference Database Release 9 [37]. GASOLINE was obtained from the work of Micale et al. [38]. It was derived from STRING database [39] considering only experimentally verified protein interactions. BioGRID was downloaded from the Biological General Repository for Interaction Datasets version 3.4.134 (Homo Sapiens) [40]. We did not combine all the PPI networks because they were compiled based on different criteria and domain of interest.

Table 4.1 shows the number of unique proteins and interactions for each PPI network. BioGRID contains the largest number of interactions while HPRD contains the largest number of proteins.

Table 4.1: The number of proteins and interactions for each PPI network [3].

PPI network	Number of unique proteins	Number of interactions
Chuang	11,203	57,235
HPRD	30,047	41,327
GASOLINE	9,556	53,859
BioGRID	20,364	315,507

We overlaid the gene microarray datasets with each PPI network by mapping each gene to its corresponding protein in the network. After removing the proteins that do not have corresponding genes in both gene expression datasets, we obtained an induced networks with the statistics shown in Table 4.2. After data integration, the numbers of proteins are quite similar to each other. BioGRID still contains the largest number of interactions while the others contain approximately the same.

Table 4.2: The number of proteins and interactions for each induced PPI network [3].

PPI network	Number of unique proteins	Number of interactions
Chuang	5,293	26,773
HPRD	4,762	18,684
GASOLINE	4,277	22,253
BioGRID	5,697	99,426

4.2.2 Affinity Propagation-Based Subnetwork Identification

We adopt the subnetwork identification procedure from our previous study [2], where we utilized a message-passing clustering algorithm, called affinity propagation, to cluster genes whose protein products interact with each other or are closely located in PPI network. The input of this clustering algorithm is the measure of similarity between genes. We originally defined the similarity of genes based entirely on the discriminative power to distinguish between the two class labels as follows:

$$s_{DP}(i,k) = t_k + \min\{t_{ik} - t_i, t_{ik} - t_k\} - \alpha|t_i - t_k| \tag{4.1}$$

where t_i, and t_k are t-test statistics score of the log-likelihood ratio (LLR) between metastatic and non-metastatic samples of genes i, and k, respectively. t_{ik} is the t-test score of the summation of the LLRs of genes i, and k.

The LLR, λ, of gene i, $\lambda(x_i)$, is based on probabilistic inference strategy proposed in [13] and it is computed by

$$\lambda(x_i) = \log\left[f^1(x_i)/f^2(x_i)\right], \tag{4.2}$$

where x_i is the expression level of the gene i and $f^j(x_i)$ is the conditional Gaussian probability density function of x_i under phenotype j.

The last term is the penalty term measured by the difference between discriminative power of considering genes. The parameter, α, is defined between $[0,1]$ to control this term. It is shown in our previous work [2] that the size of the network decreases as α gets larger. It is because a larger α tends to cluster genes with similar discriminative power. As a result of that, it yields small subnetworks with fewer genes.

The Equation 4.1 is based on original assumptions that when considering similarity between two genes, the gene itself should have high discriminative power, combining both genes as subnetwork should increase the overall discriminative power, and both genes should have similar discriminative power.

4.2.3 Incorporating Topological Information for Computing the Similarity Between Genes

With the assumption that the proteins corresponding to the genes in the same subnetwork should have common topological attributes, we consider two following points:

- Densely connected subnetworks may likely be potential functional modules.
- Proteins that are not directly connected but interact with similar sets of other proteins may share similar functionalities.

Based on these considerations, we incorporate the topological information of proteins in the PPI network by measuring their association coefficient–or topological similarity.

We measure topological attribute using different types of association coefficients. Let N_i and N_k be the neighborhood binary vectors of protein i and k. We define the topological similarity between proteins i and k, $s_T(i,k)$, based on different similarity indexes as follows:

1. Jaccard index: We define topological similarity, $s_{T_J}(i,k)$, as

$$s_{T_J}(i,k) = \frac{|N_i \cap N_k|}{|N_i \cup N_k|} \tag{4.3}$$

Jaccard index is widely used to quantify the similarity

2. Kulczyński index: This measure, $s_{T_K}(i,k)$, represents the average proportion of the number of common neighbors to the total number of neighbors of each protein. It is given by

$$s_{T_K}(i,k) = \frac{1}{2}(\frac{|N_i \cap N_k|}{|N_i|} + \frac{|N_i \cap N_k|}{|N_k|}) \tag{4.4}$$

3. Tversky index: We define topological similarity based on Tversky index, $s_{T_T}(i,k)$, as

$$s_{T_T}(i,k) = \frac{|N_i \cap N_k|}{|N_i \cap N_k| + a_{T_T}|N_i - N_k| + b_{T_T}|N_k - N_i|} \tag{4.5}$$

   In order to indicate the direction of similarity (asymmetric similarity), we let $a_{T_T} = 1$ and $b_{T_T} = 0$. This asymmetric definition lets the exemplars of the identified clusters be more densely connected than other non-exemplars. We can rewrite the equation as followings

$$s_{T_T}(i,k) = \frac{|N_i \cap N_k|}{|N_i|} \tag{4.6}$$

   Tversky index can be viewed as a general form of Tanimoto coefficient (Jaccard index) when $a_{T_T} = 1$ and $b_{T_T} = 1$, and Dice coefficient when $a_{T_T} = 0.5$ and $b_{T_T} = 0.5$.

We do not include other similarity indices whose results are in the same order (no alteration in the ranks) because they give the same output when applying affinity propagation. For example, Dice coefficient, $\frac{(2 \cdot |N_i \cap N_k|)}{|N_i| + |N_k|}$, and Jaccard index share similar results in terms of ranking. Ochiai index (or Cosine index), $\frac{|N_i \cap N_k|}{\sqrt{|N_i| \cdot |N_k|}}$, and Geometric index, $\frac{|N_i \cap N_k|^2}{|N_i| \cdot |N_k|}$ provide the same ranks as of Kulczyński index.

As we focus on retrieving topological information from the PPI network, we do not make use of the number of common non-neighbor proteins $|\neg N_i \cap \neg N_k|$ in this study.

Finally, we add the topological similarity, (4.3), (4.4), and (4.6), to the computation of similarity between genes i and k, $s(i,k)$, in two different ways.

1. Similarity between genes i and k, $s(i,k)$, as a product of the topological similarity $s_T(i,k)$ and the discriminative power based similarity $s_{DP}(i,k)$. We define as:

$$s(i,k) = s_T(i,k) \cdot s_{DP}(i,k) \tag{4.7}$$

2. Similarity between genes i and k, $s(i,k)$, as a combination of the topological similarity $s_T(i,k)$ and the discriminative power based similarity $s_{DP}(i,k)$. We first scale the discriminative power based similarity $s_{DP}(i,k)$ into the range $[0,1]$ as same as topological similarity's by

$$\hat{s}_{DP}(i,k) = \frac{s_{DP}(i,k) - \min(s_{DP})}{\max(s_{DP}) - \min(s_{DP})} \tag{4.8}$$

where s_{DP} is the set of all discriminative power based similarity of all gene pairs. Then, we combine them as follows

$$s(i,k) = \beta(s_T(i,k)) + (1-\beta)(\hat{s}_{DP}(i,k)) \tag{4.9}$$

where $\beta = [0,1]$ is used to control the magnitude between each similarity. Topological similarity, $s_T(i,k)$, has more effects as β increases. It should be noted that $s(i,k)$ can be viewed as the summation of topological similarity and discriminative power based similarity when $\beta = 0.5$.

We use the same setting for preference as in [2]. The self-similarity is set to $s(k,k) = c$ for all k, where $s(i,k) \leq c$ for only 1% of all gene pairs (g_i, g_k) to guarantee that every gene gets equal chance to be an exemplar at the initial stage of clustering process.

4.2.4 Probabilistic Inference of Subnetwork Activity

To estimate the modular—or subnetwork—activity of identified subnetwork, we employ the probabilistic inference method proposed in [13] which is the aggregation of the LLRs of all member genes to represent the activity level of the subnetwork markers, $A(\boldsymbol{\mathcal{G}})$. It is computed by

$$A(\boldsymbol{\mathcal{G}}) = \sum_{i=1}^{n} \lambda\left(x_i\right), \tag{4.10}$$

where x_i is the expression level of the gene g_i in the subnetwork $\boldsymbol{\mathcal{G}} = \{g_1, g_2, ..., g_n\}$. This inference method can be viewed as the aggregation of the probabilistic evidence of the expression level of genes in the subnetworks.

4.2.5 Experimental Setup

We identified subnetwork markers incorporating three different strategies to measure topological similarity which we referred to as Jaccard-based, Kulczyński-based, and Tversky-based. As mentioned previously, we used two different approaches to integrate topological similarity to measure similarity between genes: 1) Product of topological and discriminative power based similarity, namely, “product-based approach”, and 2) Linear combination of topological and discriminative power based similarity, namely, “linear-combination-based approach”. In the latter approach, we used three different values of $\beta(= 0.25, 0.5, 0.75)$ to investigate the impact of topological similarity to the subnetwork identification. In fact, we can also setup the experiments the other way around to find the optimal the value of β for each data.

After computing similarity between genes and applying affinity propagation-based subnetwork identification, all output clusters were ranked based on the t-test statistics score of their activity level. Then we selected the top 50 clusters with high discriminative power as the potential subnetwork markers for assessing their classification performance.

We repeated these processes to both gene expression datasets and all four PPI networks.

4.3 Results

For comparison, we also evaluated the method proposed in [11], and [2] which we refer to as the 'greedy' method, and the 'AP-based' method, respectively. We applied the greedy method with 5% minimum required improvement which is the same setting as originally published in [11]. In the AP-based method, we set the magnitude of the penalty term, α, to 0.5 by reason shown in [2] that it yields high and consistent classification performance as of smaller α with the smaller size of identified subnetworks compared to larger α.

For simplicity in displaying Tables and Figures in this section, we abbreviate Jaccard-based, Kulczyński-based, and Tversky-based to *jac*, *kul*, and *tve*, respectively. The suffixes, *_p*, and *_lc* are appended to indicate product-based approach, and linear-combination-based approach, respectively.

4.3.1 Statistics of the Subnetwork Markers

Table 4.3 shows the average size of top 50 highly discriminative subnetwork markers identified by each method on GSE2034 and NKI295. Each column shows the results for each PPI network. The average size of markers identified by product-based and linear-combination-based approach is similar to the original AP-based method. We can clearly see that the average size of top markers identified by the proposed method and AP-based is larger than the greedy-based.

As we can see from Table 4.3, the average size of top 50 highly discriminative

Table 4.3: The average size of top 50 highly discriminative subnetwork markers from GSE2034 and NKI295 [3].

Gene expression dataset = GSE2034					
Methods		Chuang	HPRD	GASOLINE	BioGRID
Greedy		3.1	3.26	3.54	3.66
AP-based		36.28	35.78	34.18	38.78
jac_p		18.06	19.94	19.58	29
kul_p		21.16	25.32	22.48	36.28
tve_p		34.48	45.26	45.98	61.8
jac_lc	$\beta = 0.25$	18.3	21.36	23.14	34
	$\beta = 0.5$	15.08	15.38	16.44	24.24
	$\beta = 0.75$	13.28	16.34	13.44	19.18
kul_lc	$\beta = 0.25$	24	30.14	28.68	39.02
	$\beta = 0.5$	18.98	22.86	24.18	38.32
	$\beta = 0.75$	16.06	19.12	20.84	30.98
tve_lc	$\beta = 0.25$	34.1	46.58	43.44	53.54
	$\beta = 0.5$	28.98	43.8	45.5	71.24
	$\beta = 0.75$	22.92	44.78	46.32	82.66
Gene expression dataset = NKI295					
Methods		Chuang	HPRD	GASOLINE	Biogrid
Greedy		4.12	3.68	4.46	4.42
AP-based		31.34	30.32	28.78	34.66
jac_p		14.62	16	18.94	27.72
kul_p		12.3	22.5	26.9	33.34
tve_p		28.22	42.24	49.9	57.1
jac_lc	$\beta = 0.25$	15.14	16.8	19.66	30.06
	$\beta = 0.5$	13.38	12.44	13.68	22.66
	$\beta = 0.75$	11.54	12.88	10.78	17.98
kul_lc	$\beta = 0.25$	14.8	24.6	27.06	39.26
	$\beta = 0.5$	15.9	18.5	23.28	33.96
	$\beta = 0.75$	13.7	17.12	17.22	27.14
tve_lc	$\beta = 0.25$	30.76	41.78	48.66	52.44
	$\beta = 0.5$	27.26	41.62	50.7	72.88
	$\beta = 0.75$	18.52	43.22	48.24	81.42

subnetwork markers increases as the PPI network with larger number of interactions and unique proteins is used. This trend can be clearly seen when BioGRID is employed. Among product-based approach group, Tversky-based similarity, *tve_p*, yields larger subnetworks. In linear-combination-based approach, we can see that the average size decreases as β increases in most cases. However, we cannot see this trend distinctly in Tversky-based, *tve_lc*. The main reason is that Tversky-based similarity mostly provides higher similarity index compared with the others as it is designed to indicate the direction of the similarity. For instance, when a gene shares all of its neighbors with another gene ($|N_i \cap N_k| = |N_i|$), it returns the maximum similarity ($s_{T_T}(i,k) = 1$), whereas the other topological similarities yield lower because they depend on the number of neighbors the both genes.

As defined in Equation 4.9, the clustering process relies more on topological information as β gets larger. Therefore, in this case, more genes tend to be clustered into the same subnetwork.

We can see the similar trends for the number of unique genes in top 50 discriminative subnetwork markers as shown in Table 4.4. We can also clearly see that the top markers identified by the proposed method and AP-based cover more genes than the greedy-based. The larger unique genes covered show that the proposed method may increase the chance to discover genes that are not known to be related to the disease. This also means the higher probability of identifying new subnetwork and pathway.

Next, we studied the overlap between the top 50 highly discriminative subnetwork markers identified on different gene expression datasets. The proposed method yield larger overlap when comparing to all of the previous methods as shown in Table 4.5. Again, similar trends as in Table 4.3 can also be observed here. The larger overlaps show that more of common genes are covered and shared among identified subnetworks from independent dataset from different platforms. This may lead us to more robust classifiers, we demonstrate the robustness by providing classification

Table 4.4: The number of unique genes in top 50 highly discriminative subnetwork markers from GSE2034 and NKI295 [3].

Gene expression dataset = GSE2034					
Methods		Chuang	HPRD	GASOLINE	Biogrid
Greedy		130	121	140	139
AP-based		1814	1789	1709	1939
jac_p		903	997	979	1450
kul_p		1058	1266	1124	1814
tve_p		1724	2263	2299	3090
jac_lc	$\beta = 0.25$	915	1068	1157	1700
	$\beta = 0.5$	754	769	822	1212
	$\beta = 0.75$	664	817	672	959
kul_lc	$\beta = 0.25$	1200	1507	1434	1951
	$\beta = 0.5$	949	1143	1209	1916
	$\beta = 0.75$	803	956	1042	1549
tve_lc	$\beta = 0.25$	1705	2329	2172	2677
	$\beta = 0.5$	1449	2190	2275	3562
	$\beta = 0.75$	1146	2239	2316	4133
Gene expression dataset = NKI295					
Methods		Chuang	HPRD	GASOLINE	Biogrid
Greedy		114	110	118	150
AP-based		1567	1516	1439	1733
jac_p		731	800	947	1386
kul_p		615	1125	1345	1667
tve_p		1411	2112	2495	2855
jac_lc	$\beta = 0.25$	757	840	983	1503
	$\beta = 0.5$	669	622	684	1133
	$\beta = 0.75$	577	644	539	899
kul_lc	$\beta = 0.25$	740	1230	1353	1963
	$\beta = 0.5$	795	925	1164	1698
	$\beta = 0.75$	685	856	861	1357
tve_lc	$\beta = 0.25$	1538	2089	2433	2622
	$\beta = 0.5$	1363	2081	2535	3644
	$\beta = 0.75$	926	2161	2412	4071

performance charts showing that the experimental results from the proposed method are consistent in the next section.

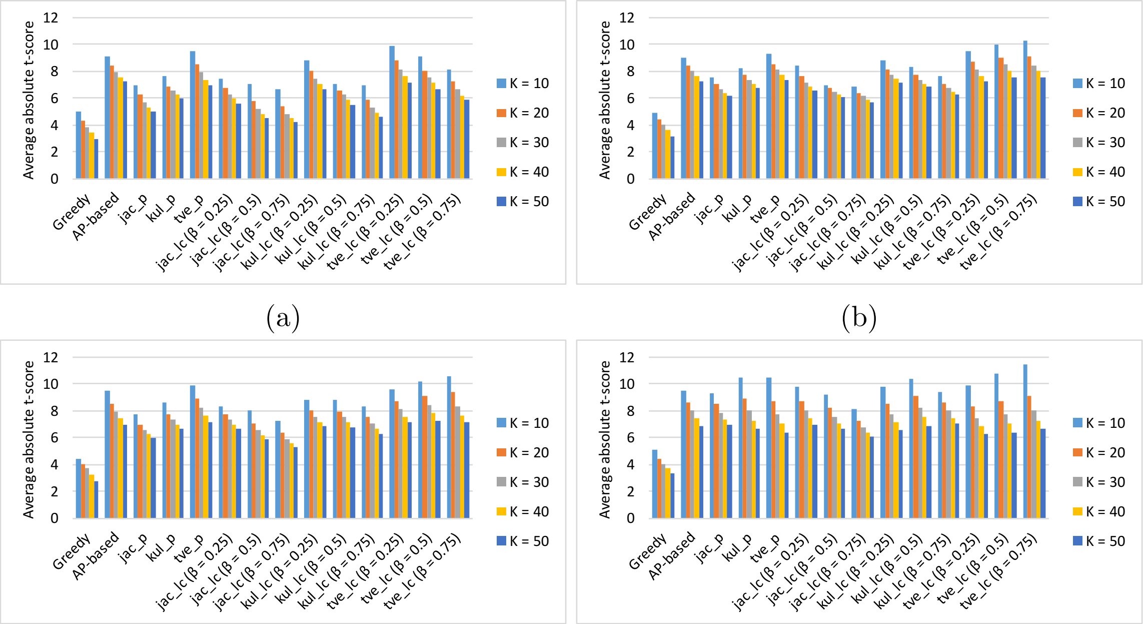
Additionally, we analyzed enriched functions of the genes in the subnetwork markers using Panther [41], a web-based system designed to facilitate analysis of large numbers of genes and provide comprehensive function information which includes up-to-date comprehensive Gene Ontology (GO) annotations (GO database version 1.2, released 2016-05-20 with 44,588 total annotations). An example of the enrichment analysis of the top 50 highly discriminative subnetworks identified using *tve_p* method on GASOLINE is shown in Table 4.6. We can see that the genes in identified subnetworks from different gene expression datasets also share common GO terms.

4.3.2 Discriminative Power of the Subnetwork Markers

We evaluated the discriminative power of the subnetwork markers based on the same procedure as previously used in these studies [1, 12–14]. We computed the t-test score of the inferred subnetwork activity level. And then we sorted the

Table 4.5: Overlap between the top subnetwork markers identified on different gene expression datasets [3].

Methods		Chuang	HPRD	GASOLINE	Biogrid
Greedy		5.63%	4.05%	4.88%	3.96%
AP-based		24.90%	28.70%	27.71%	23.89%
jac_p		37.89%	29.28%	32.01%	31.97%
kul_p		15.38%	27.52%	26.49%	28.26%
tve_p		25.80%	44.15%	50.57%	42.33%
jac_lc	$\beta = 0.25$	39.10%	22.54%	26.55%	30.20%
	$\beta = 0.5$	53.51%	26.68%	26.87%	37.94%
	$\beta = 0.75$	54.55%	31.74%	26.67%	40.12%
kul_lc	$\beta = 0.25$	12.73%	24.47%	27.90%	28.50%
	$\beta = 0.5$	39.86%	28.29%	31.18%	33.26%
	$\beta = 0.75$	50.61%	35.53%	31.42%	40.73%
tve_lc	$\beta = 0.25$	27.53%	44.47%	46.75%	36.57%
	$\beta = 0.5$	32.14%	43.47%	52.41%	54.90%
	$\beta = 0.75$	32.99%	50.94%	57.71%	69.05%



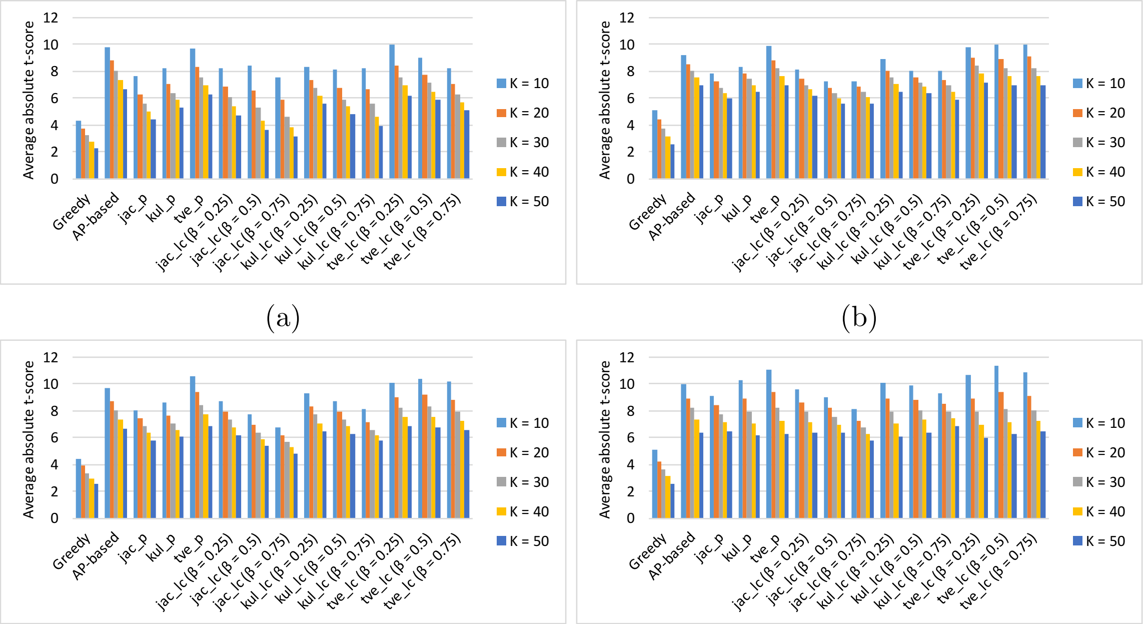
(a) (b)

(c) (d)

Figure 4.1: Discriminative power of subnetwork markers identified on GSE2034 by different methods [3]. We computed the average absolute t-test score of the top K=10, 20, 30, 40, and 50 subnetwork markers identified on GSE2034 by various methods for the following PPI datasets: (a) Chuang, (b) HPRD, (c) GASOLINE, and (d) BioGRID.

absolute value in descending order. The average absolute t-test score of the top $K = 10, 20, 30, 40, 50$ subnetwork markers is shown in Figure 4.1. We can see that the discriminative power of subnetwork markers identified by product-based approach, and linear-combination-based approach are considerably higher than the result of the greedy method. Among product-based approach group, Tversky-based yields the highest in most of the results.

We also assessed how the subnetwork markers identified on specific gene expression dataset perform in another independent dataset. We sorted the subnetwork markers based on their t-test score of the inferred subnetwork activity level on one dataset and we reevaluated the discriminative power on the other dataset. As shown



(a)

(b)

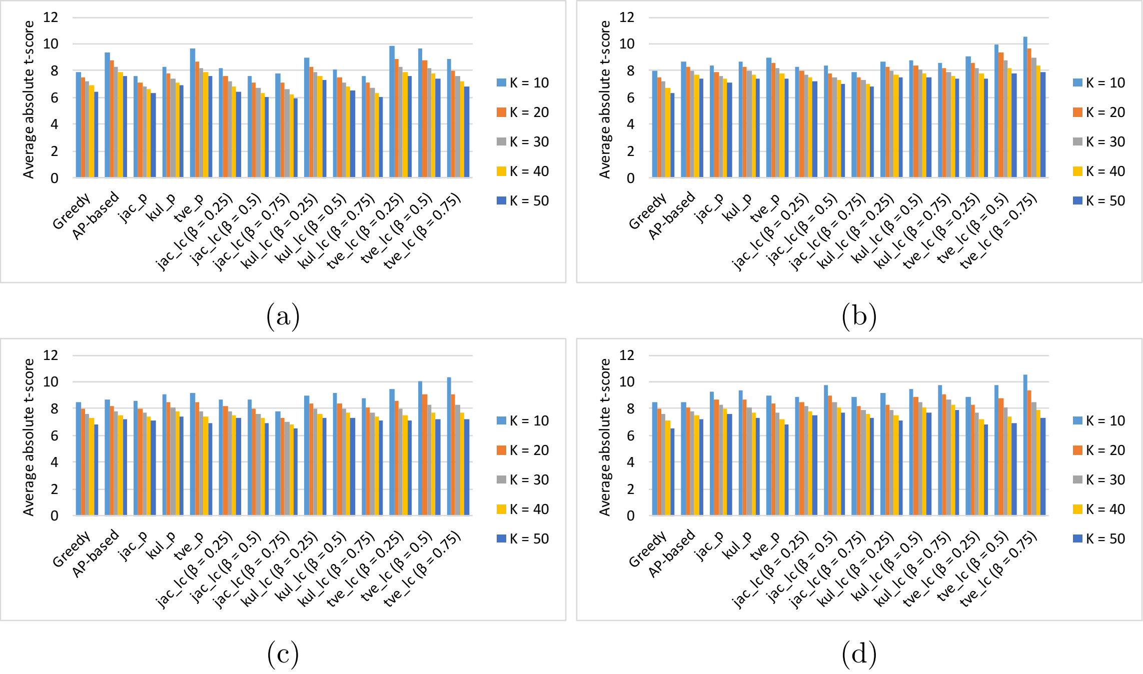
(c)

(d)

Figure 4.2: Discriminative power of subnetwork markers identified on GSE2034 and evaluated on NKI295 [3]. The markers were identified and ranked on GSE2034 and their discriminative power was evaluated on NKI295. We computed the mean absolute t-score of the top K=10, 20, 30, 40, and 50 markers by different methods for the following PPI datasets: (a) Chuang, (b) HPRD, (c) GASOLINE, and (d) BioGRID.

in Figure 4.2, we can see that the trends of discriminative power of subnetwork markers across different gene expression datasets are similar to those observed in Figure 4.1. The analysis of discriminative power of the subnetwork markers identified on NKI295 data also shows a similar trend (Figures 4.3 and 4.4).

About the impact of different PPI networks, the PPI network with larger number of interactions tends to yield the higher discriminative power. One of the reasons may be that it contains more topological information which may help to measure the similarity between genes. As intuitively expected, we can see that BioGRID is advantageous to the other PPI networks because it contains the largest number of



(a) (b) (c) (d)

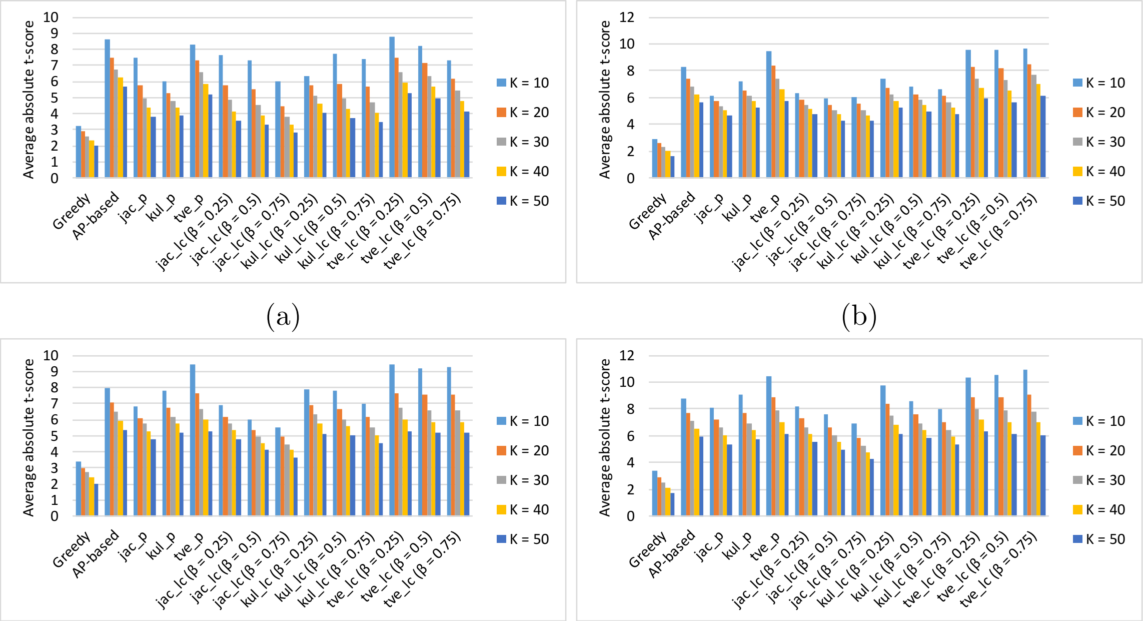
Figure 4.3: Discriminative power of subnetwork markers identified on NKI295 by different methods [3]. We evalutated the discriminative power of the top K=10, 20, 30, 40, and 50 subnetwork markers identified on NKI295 by varied methods for the following PPI datasets: (a) Chuang, (b) HPRD, (c) GASOLINE, and (d) BioGRID.

interactions (as shown in Figures 4.1(d) and 4.3(d)).

4.3.3 Evaluating the Reproducibility of the Identified Subnetwork Markers

In order to evaluate the reproducibility of subnetwork markers, we performed five-fold cross-validation experiments based on a similar set-up that has been commonly used in previous studies [1, 2, 11–14], where the entire process was repeated for 100 random partitions.

We identified potential subnetwork markers and selected the top 50 subnetworks as a feature set for the classifier on one gene expression dataset. After that, we built the LDA classifiers based on the selected features and evaluated the accuracy on the other dataset. The classification performance assessed by the AUC is shown in



(a)

(b)

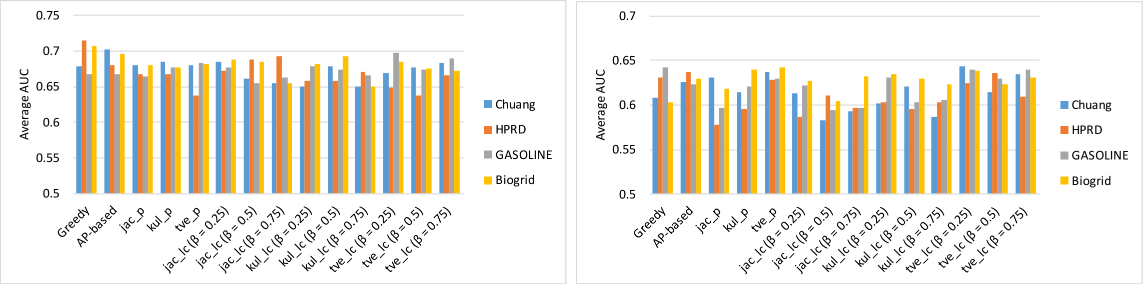
(c)

(d)

Figure 4.4: Discriminative power of subnetwork markers identified on NKI295 and evaluated on GSE2034 [3]. The markers were identified and sorted on NKI295 and their discriminative power was evaluated on GSE2034. We repeated the cross-dataset discriminative power assessment for the following PPI datasets: (a) Chuang, (b) HPRD, (c) GASOLINE, and (d) BioGRID.

Figure 4.5. We can see that both product-based approach and linear-combination based approach yield consistently high performance across different gene expression datasets and PPI networks.

In this study, we use the term, ‘reproducibility’ in the sense of the ability to identify common discriminative genes or subnetworks across different independent datasets. Therefore, using these subnetworks as biomarkers for disease classification may lead to consistent performance. Furthermore, in terms of reproducibility in practical usage, the AP-based methods, including our proposed methods, cost less computation time compared to the greedy algorithm as shown in [2].



(a) (b)

Figure 4.5: Reproducibility of subnetwork markers identified by various methods [3]. The bars show the cross-dataset classification performance (average AUC) of different methods. (a) GSE2034 was used for identifying the potential markers and NKI295 was used for training and evaluating the classifier, (b) We repeated as NKI295 was used for identifying the markers and GSE2034 was used for training and evaluation of the classifier.

4.4 Conclusions

In this chapter, we propose a novel method that incorporates topological information to identify subnetwork markers that can be used in cancer prognosis prediction. We demonstrate how widely used association coefficients, such as Jaccard index, Kulczyński index, and Tversky index can be utilized to measure topological similarity. Also, we show how to integrate these measures by two different approaches, product-based, and linear-combination based.

Based on our experimental results, Tversky-based strategy is most suitable to measure similarity between genes when the direction of interaction is involved. It yields consistently high discriminative power across different datasets. Furthermore, utilizing the larger PPI network with larger number of unique proteins and interactions, such as BioGRID, may lead to the better subnetwork identification with higher classification performance.

The proposed method considerably increases the coverage of genes and also the overlap of genes when identified across different independent datasets. Through extensive evaluations using various independent breast cancer gene expression datasets and PPI networks, the experimental results show that our method leads to the identification of robust and reproducible subnetwork markers that may lead to better cancer classification.

Table 4.6: The number of genes in top 50 highly discriminative subnetwork markers from *tve_p* method on GASOLINE categorized by their GO terms [3].

Ontology: Molecular function			
GO term	GO id	GSE2034	NKI295
transporter activity	GO:0005215	240	251
translation regulator activity	GO:0045182	37	41
protein binding transcription factor activity	GO:0000988	35	42
enzyme regulator activity	GO:0030234	193	205
catalytic activity	GO:0003824	1146	1221
channel regulator activity	GO:0016247	5	6
receptor activity	GO:0004872	346	370
nucleic acid binding transcription factor activity	GO:0001071	307	316
antioxidant activity	GO:0016209	8	6
structural molecule activity	GO:0005198	226	260
binding	GO:0005488	1237	1330

Ontology: Cellular component			
GO term	GO id	GSE2034	NKI295
synapse	GO:0045202	15	15
cell junction	GO:0030054	13	11
membrane	GO:0016020	288	290
macromolecular complex	GO:0032991	213	214
extracellular matrix	GO:0031012	50	58
cell part	GO:0044464	765	794
organelle	GO:0043226	411	441
extracellular region	GO:0005576	151	153

Ontology: Biological process			
GO term	GO id	GSE2034	NKI295
cellular component organization or biogenesis	GO:0071840	278	309
cellular process	GO:0009987	1559	1679
localization	GO:0051179	536	577
apoptotic process	GO:0006915	174	194
reproduction	GO:0000003	104	118
biological regulation	GO:0065007	886	933
response to stimulus	GO:0050896	547	593
developmental process	GO:0032502	634	692
rhythmic process	GO:0048511	3	1
multicellular organismal process	GO:0032501	393	413
locomotion	GO:0040011	20	24
biological adhesion	GO:0022610	127	147
metabolic process	GO:0008152	1773	1876
growth	GO:0040007	1	3
immune system process	GO:0002376	314	342

5. OVERALL CONCLUSIONS

In this dissertation, we proposed several methods for utilizing known biological knowledge with gene expression data to identify disease-related modular markers for predicting breast cancer metastasis. Our aim is to improve accuracy, robustness, and reproducibility of the biomarkers.

In chapter 2, we proposed an improved pathway activity inference scheme – called rank-based pathway activity inference – by integrating two effective strategies which are the probabilistic pathway activity inference and the ranking-based relative gene expression analysis approach. We showed that the proposed method can identify more reproducible pathway markers with higher discriminative power. It can lead to better classifiers that yield more consistent performance across datasets from different platforms and normalization techniques.

In chapter 3, we proposed a new method for identifying subnetwork markers by utilizing a message-passing clustering algorithm, called affinity propagation. We also proposed a strategy to measure the similarity between genes based on gene expression data and its overlaid PPI networks in order to provide the input to the clustering algorithm. We demonstrated that the proposed method can simultaneously predict non-overlapping subnetwork markers with high discriminative power and reproducible performance across independent datasets.

In chapter 4, we enhanced the affinity propagation-based subnetwork identification proposed in chapter 3 by incorporating topological information. We utilized various association coefficients and suggested multiple approaches to integrating topological information into the previously proposed method. We showed that the proposed method increases the coverage of genes and identifies more common genes across different gene expression datasets and PPI networks. The proposed method leads to the identification of robust and reproducible subnetwork markers.

All of the proposed methods in this dissertation are shown to improve the identification of potential modular biomarkers associated with breast cancer metastasis based on biological knowledge, such as pathway information or PPI networks. This improvement may lead to better cancer classification.

REFERENCES

[1] N. Khunlertgit and B.-J. Yoon, “Identification of robust pathway markers for cancer through rank-based pathway activity inference,” *Advances in Bioinformatics*, vol. 2013, no. 618461, 2013.

[2] ——, “Simultaneous identification of robust synergistic subnetwork markers for effective cancer prognosis,” *EURASIP Journal on Bioinformatics and Systems Biology*, vol. 2014, no. 1, p. 19, 2014.

[3] ——, “Incorporating topological information for predicting robust cancer subnetwork markers in human protein-protein interaction network,” *BMC Bioinformatics*, vol. 17, no. 13, p. 351, 2016.

[4] E. R. Dougherty, A. Datta, and C. Sima, “Research issues in genomic signal processing,” *IEEE Signal Processing Magazine*, vol. 22, no. 6, pp. 46–68, Nov 2005.

[5] L. J. van't Veer, H. Dai, M. J. van de Vijver, Y. D. He, A. A. Hart, M. Mao, H. L. Peterse, der Kooy van, M. J. Marton, A. T. Witteveen, G. J. Schreiber, R. M. Kerkhoven, C. Roberts, P. S. Linsley, R. Bernards, and S. H. Friend, “Gene expression profiling predicts clinical outcome of breast cancer,” *Nature*, vol. 415, pp. 530–536, 2002.

[6] Y. Wang, J. G. Klijn, Y. Zhang, A. M. Sieuwerts, M. P. Look, F. Yang, D. Talantov, M. Timmermans, M. M. van Gelder, J. Yu, T. Jatkoe, E. M. Berns, D. Atkins, and J. A. Foekens, “Gene-expression profiles to predict distant metastasis of lymph-node-negative primary breast cancer,” *Lancet*, vol. 365, pp. 671–679, 2005.

[7] L. Ein-Dor, O. Zuk, and E. Domany, “Thousands of samples are needed to generate a robust gene list for predicting outcome in cancer,” *Proceedings of the National Academy of Sciences of the United States of America*, vol. 103, no. 15, pp. 5923–5928, 2006.

[8] 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, pp. 13 544–13 549, 2005.

[9] Z. Guo, T. Zhang, X. Li, Q. Wang, J. Xu, H. Yu, J. Zhu, H. Wang, C. Wang, E. J. Topol, Q. Wang, and S. Rao, “Towards precise classification of cancers based on robust gene functional expression profiles,” *BMC Bioinformatics*, vol. 6, p. 58, 2005.

[10] C. Auffray, “Protein subnetwork markers improve prediction of cancer outcome,” *Mol Syst Biol*, vol. 3, p. 141, 2007.

[11] 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.

[12] 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, p. e1000217, 2008.

[13] 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, 12/07 2009.

[14] ——, “Identification of diagnostic subnetwork markers for cancer in human protein-protein interaction network,” *BMC Bioinformatics*, vol. 11, no. 6, pp.

1–13, 2010.

[15] D. Geman, C. d'Avignon, D. Q. Naiman, and R. L. Winslow, "Classifying gene expression profiles from pairwise mrna comparisons," *Stat Appl Genet Mol Biol.*, vol. 3:Article19, 2004.

[16] J. A. Eddy, L. Hood, N. D. Price, and D. Geman, "Identifying tightly regulated and variably expressed networks by differential rank conservation (dirac)," *PLoS Comput Biol*, vol. 6, no. 5, p. e1000792, 05/27 2010.

[17] N. Khunlertgit and B.-J. Yoon, "Finding robust pathway markers for cancer classification," in *IEEE International Workshop on Genomic Signal Processing and Statistics (GENSIPS)*, Dec. 2012.

[18] M. J. van de Vijver, Y. D. He, L. J. van 't Veer, H. Dai, A. A. Hart, D. W. Voskuil, G. J. Schreiber, J. L. Peterse, C. Roberts, M. J. Marton, M. Parrish, D. Atsma, A. Witteveen, A. Glas, L. Delahaye, der Velde van, H. Bartelink, S. Rodenhuis, E. T. Rutgers, S. H. Friend, and R. Bernards, "A gene-expression signature as a predictor of survival in breast cancer," *N Engl J Med*, vol. 347, no. 25, pp. 1999–2009, 12/19; 2011/09 2002.

[19] C. Desmedt, F. Piette, S. Loi, Y. Wang, F. Lallemand, B. Haibe-Kains, G. Viale, M. Delorenzi, Y. Zhang, M. S. d'Assignies, J. Bergh, R. Lidereau, P. Ellis, A. L. Harris, J. G. Klijn, J. A. Foekens, F. Cardoso, M. J. Piccart, M. Buyse, and C. Sotiriou, "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.

[20] Y. Pawitan, J. Bjöhle, L. Amler, A.-L. Borg, S. Egyhazi, P. Hall, X. Han, L. Holmberg, F. Huang, S. Klaar, E. T. Liu, L. Miller, H. Nordgren, A. Ploner,

K. Sandelin, P. M. Shaw, J. Smeds, L. Skoog, S. Wedrén, and J. Bergh, “Gene expression profiling spares early breast cancer patients from adjuvant therapy: Derived and validated in two population-based cohorts,” *Breast Cancer Research*, vol. 7, no. 6, pp. R953–R964, 2005.

[21] I. B. Pau Ni, Z. Zakaria, R. Muhammad, N. Abdullah, N. Ibrahim, N. Aina Emran, N. Hisham Abdullah, and S. N. A. Syed Hussain, “Gene expression patterns distinguish breast carcinomas from normal breast tissues: The malaysian context,” *Pathol Res Pract*, vol. 206, no. 4, pp. 223–8, Apr 2010.

[22] A. Tripathi, C. King, A. de la Morenas, V. K. Perry, B. Burke, G. A. Antoine, E. F. Hirsch, M. Kavanah, J. Mendez, M. Stone, N. P. Gerry, M. E. Lenburg, and C. L. Rosenberg, “Gene expression abnormalities in histologically normal breast epithelium of breast cancer patients,” *Int J Cancer*, vol. 122, no. 7, pp. 1557–66, Apr 2008.

[23] H. Y. Chang, D. S. A. Nuyten, J. B. Sneddon, T. Hastie, R. Tibshirani, T. Sørlie, H. Dai, Y. D. He, L. J. van’t Veer, H. Bartelink, M. van de Rijn, P. O. Brown, and M. J. van de Vijver, “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.

[24] 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.

[25] R. C. Gentleman, V. J. Carey, D. M. Bates, and others, “Bioconductor: Open software development for computational biology and bioinformatics,” *Genome Biology*, vol. 5, p. R80, 2004.

[26] R. A. Irizarry, B. Hobbs, F. Collin, Y. D. BeazerBarclay, K. J. Antonellis, U. Scherf, and T. P. Speed, "Exploration, normalization, and summaries of high density oligonucleotide array probe level data," *Biostatistics*, vol. 4, no. 2, pp. 249–264, 2003.

[27] Z. Wu, R. A. Irizarry, R. Gentleman, F. Martinez-Murillo, and F. Spencer, "A model-based background adjustment for oligonucleotide expression arrays," *Journal of the American Statistical Association*, vol. 99, no. 468, pp. 909–917, 2004.

[28] S. D. Pepper, E. K. Saunders, L. E. Edwards, C. L. Wilson, and C. J. Miller, "The utility of mas5 expression summary and detection call algorithms," *BMC Bioinformatics*, vol. 8, no. 1, pp. 1–12, 2007.

[29] 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.

[30] T. M. Cover and J. A. Thomas, *Elements of information theory*, New York, NY, USA, 2006.

[31] T. Fawcett, "An introduction to roc analysis," *Patt Recog Letters*, vol. 27, pp. 861–874, 2006.

[32] N. Khunlertgit and B. J. Yoon, "Finding robust subnetwork markers that improve cross-dataset performance of cancer classification," in *IEEE Global Conference on Signal and Information Processing (GlobalSIP)*, Dec 2013, pp. 94–94.

[33] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, no. 5814, pp. 972–976, 2007.

[34] J. I. Fuxman Bass, A. Diallo, J. Nelson, J. M. Soto, C. L. Myers, and A. J. M. Walhout, "Using networks to measure similarity between genes: Association index selection," *Nat Methods*, vol. 10, no. 12, pp. 1169–76, Dec 2013.

[35] L. Royer, M. Reimann, B. Andreopoulos, and M. Schroeder, "Unraveling protein networks with power graph analysis," *PLoS Comput Biol*, vol. 4, no. 7, p. e1000108, 2008.

[36] T. Barrett, S. E. Wilhite, P. Ledoux, C. Evangelista, I. F. Kim, M. Tomashevsky, K. A. Marshall, K. H. Phillippy, P. M. Sherman, M. Holko, A. Yefanov, H. Lee, N. Zhang, C. L. Robertson, N. Serova, S. Davis, and A. Soboleva, "Ncbi geo: Archive for functional genomics data sets–update," *Nucleic Acids Res*, vol. 41, no. Database issue, pp. D991–5, Jan 2013.

[37] T. S. Keshava Prasad, R. Goel, K. Kandasamy, S. Keerthikumar, S. Kumar, S. Mathivanan, D. Telikicherla, R. Raju, B. Shafreen, A. Venugopal, L. Balakrishnan, A. Marimuthu, S. Banerjee, D. S. Somanathan, A. Sebastian, S. Rani, S. Ray, C. J. Harrys Kishore, S. Kanth, M. Ahmed, M. K. Kashyap, R. Mohmood, Y. L. Ramachandra, V. Krishna, B. A. Rahiman, S. Mohan, P. Ranganathan, S. Ramabadran, R. Chaerkady, and A. Pandey, "Human protein reference database–2009 update," *Nucleic Acids Res*, vol. 37, no. Database issue, pp. D767–72, Jan 2009.

[38] G. Micale, A. Pulvirenti, R. Giugno, and A. Ferro, "Gasoline: A greedy and stochastic algorithm for optimal local multiple alignment of interaction networks," *PLoS One*, vol. 9, no. 6, p. e98750, 2014.

[39] D. Szklarczyk, A. Franceschini, S. Wyder, K. Forslund, D. Heller, J. Huerta-Cepas, M. Simonovic, A. Roth, A. Santos, K. P. Tsafou, M. Kuhn, P. Bork, L. J. Jensen, and C. von Mering, "String v10: Protein-protein interaction networks,

integrated over the tree of life," *Nucleic Acids Res*, vol. 43, no. Database issue, pp. D447–52, Jan 2015.

[40] A. Chatr-Aryamontri, B.-J. Breitkreutz, R. Oughtred, L. Boucher, S. Heinicke, D. Chen, C. Stark, A. Breitkreutz, N. Kolas, L. O'Donnell, T. Reguly, J. Nixon, L. Ramage, A. Winter, A. Sellam, C. Chang, J. Hirschman, C. Theesfeld, J. Rust, M. S. Livstone, K. Dolinski, and M. Tyers, "The biogrid interaction database: 2015 update," *Nucleic Acids Res*, vol. 43, no. Database issue, pp. D470–8, Jan 2015.

[41] H. Mi, S. Poudel, A. Muruganujan, J. T. Casagrande, and P. D. Thomas, "Panther version 10: Expanded protein families and functions, and analysis tools," *Nucleic Acids Res*, vol. 44, no. D1, pp. D336–42, Jan 2016.