July 2019

Novel Strategies on Characterizing Biologically Specific Protein-protein Interaction Networks

Bi Zhao

University of South Florida, bizhao9@gmail.com

Follow this and additional works at: https://digitalcommons.usf.edu/etd

Part of the Bioinformatics Commons, Molecular Biology Commons, and the Systems Biology Commons

Scholar Commons Citation
Zhao, Bi, "Novel Strategies on Characterizing Biologically Specific Protein-protein Interaction Networks" (2019). USF Tampa Graduate Theses and Dissertations.
https://digitalcommons.usf.edu/etd/8428

This Dissertation is brought to you for free and open access by the USF Graduate Theses and Dissertations at Digital Commons @ University of South Florida. It has been accepted for inclusion in USF Tampa Graduate Theses and Dissertations by an authorized administrator of Digital Commons @ University of South Florida. For more information, please contact digitalcommons@usf.edu.
Novel Strategies on Characterizing Biologically Specific Protein-protein Interaction Networks

by

Bi Zhao

A dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy
Department of Cell Biology, Microbiology and Molecular Biology
College of Art and Science
University of South Florida

Major Professor: Bin Xue, Ph.D.
Robert Deschenes, Ph.D.
Vladimir Uversky, Ph.D.
Yu Sun, Ph.D.
Brant Burkhardt, Ph.D.

Date of Approval:
June 25, 2019

Keywords: Protein-protein Interaction, Gene Expression, MicroRNA:mRNA Prediction

Copyright © 2019, Bi Zhao
DEDICATION

This dissertation is dedicated to my brilliant and personable husband, Fei Guo, and my clever and pretty son, Lucas Z. Guo, who inspire and support me through this journey.
ACKNOWLEDGMENTS

First of all, I would like to specially thank my major professor, Dr. Bin Xue, who gives me an important guidance and support throughout my study and research. It is hard to imagine that I could complete my dissertation without his help. His spirit of research has a positive and lifelong impact on my professional development.

In addition, I would like to thank my committee members, Dr. Robert Deschenes, Dr. Vladimir Uversky, Dr. Yu Sun, and Dr. Brant Burkhardt for providing insightful comments and encouragement. Their guidance is invaluable for both during the writing of this dissertation and over the entire course of my time at University of South Florida.

Finally, I would like to thank my parents, Yang Zhao and Darong Wang, my grandparents, all my families and friends, all the faculty and staff members in the Department of Cell Biology, Microbiology and Molecular Biology for their help and support during these years.
# TABLE OF CONTENTS

LIST OF TABLES v

LIST OF FIGURES vii

ABSTRACT xi

CHAPTER 1 INTRODUCTION 1

1.1 Data Necessity and Availability 4

1.1.1 Data Sets Used in Designing Genetic Distance-based Method 5

1.1.2 MicroRNA:mRNA Interaction Predicted Databases 5

1.1.3 Protein-protein Interaction Databases 5

1.1.4 Data Sets Used in Rewiring Protein-protein Interaction Database 6

1.2 Contribution 6

1.3 Dissertation Overview 7

CHAPTER 2 DISTANCE-BASED STRATEGY TO STUDY GENE EXPRESSION 8

2.1 Note to Reader 8

2.2 Contributions and Acknowledgements 8

2.3 Introduction 9

2.4 How Many Differentially Expressed Genes 12

2.4.1 Materials and Methods 13

2.4.1.1 Microarray Data Sets 13

2.4.1.2 Bioconductor and DEG Analysis 13

2.4.1.3 Intra-group and Inter-group Distance 14

2.4.1.4 Standard Deviation of Distance Curve 15

2.4.2 Results 15

2.4.2.1 Inter-group or Intra-group Distance Curves 15

2.4.2.2 Selecting gpsDEGs 18

2.4.3 Discussion 22

2.5 Measuring the Inter-sample Heterogeneity by Dynamic PCA Bi-plot 23

2.5.1 Methods 24

2.5.2 Results 24

2.5.3 Discussion 26

2.6 Measuring Impacts of Heterogeneity on gpsDEGs Identification 27

2.6.1 Materials and Methods 28
LIST OF TABLES

Table 2.1 The summary of eight sets of Microarray data. 13

Table 2.2 Numbers of DEGs/gpsDEGs and the corresponding highest false discovery rate (FDR) adjusted p-values based on three different methods. 22

Table 2.3 The Intestinal microarray GEO data sets. 29

Table 3.1 Summary of the eleven datasets. 68

Table 3.2 Summary of threshold values of different individual predictors for both positive and negative predictions. 76

Table 3.3 Numbers of samples in each of the eleven newly-updated data sets. 82

Table 3.4 The 1st-step and 2nd-step threshold values for both true and false predictions in eleven DANN modules. 91

Table 3.5 Prediction performances of three meta-predictor processes using three different sets of techniques. 99

Table 3.6 F1 and Mathews Correlation Coefficient (MCC) in eleven training sets under multi-fold cross-validation. 102

Table 3.7 F1 and Mathews Correlation Coefficient (MCC) in the eleven independent test sets. 105

Table 3.8 Numbers of samples in each of the eleven newly-designed human data sets. 114

Table 3.9 The 1st-step and 2nd-step threshold values for both true and false predictions in eleven modules of newly-designed human microRNA:mRNA meta-predictor. 115

Table 3.10 F1 and Mathews Correlation Coefficient (MCC) in human microRNA:mRNA training sets under multi-fold cross-validation. 118

Table 3.11 F1 and Mathews Correlation Coefficient (MCC) in human microRNA:mRNA independent test sets. 121
Table 3.12 Prediction performance of the new strategy under five-fold cross-validation, in comparison with four component predictors, another four recently-designed predictors.

Table 3.13 Prediction performance of all nine predictors in the independent data set.

Table 4.1 The summary of protein-protein interaction databases.

Table 4.2 Summary of protein-protein interaction databases.

Table 4.3 Summary of tissue-specific gene expression data sets.

Table 4.4 Summary of tissue-specific protein expression data sets.

Table 4.5 Summary of tissue-specific microRNA expression data sets.

Table 4.6 The Summary of Genes with Multiple UniProt IDs.
LIST OF FIGURES

Figure 1.1 Overall description of the dissertation. 3

Figure 2.1 Expression levels of the top 2000 genes from GSE22606 data set ranked by (A) p-value and (B) FC. 16

Figure 2.2 The intra-group and inter-group distance curves as a function of the number of ranked genes in the GSE22606 data set. 18

Figure 2.3 Fluctuation of distance curve demonstrated by standard deviation in a sliding window of size 9 as a function of the number of ranked genes in the GSE22606 data set. 19

Figure 2.4 Fluctuation averaged over all the intra-group distance curves (gray dotted line) or all the inter-group distance curves (black) as a function of the number of ranked genes in every of the eight Microarray data sets (AH). 21

Figure 2.5 Dynamic PCA bi-plot of each sample for all the samples in the GSE47734 (A) and GSE47735 (B) data sets. 25

Figure 2.6 Distance between samples in the dynamics PCA bi-plot for samples in (A) GSE47734 and (B) GSE47735 data sets. 27

Figure 2.7 Distance curves of intra-group and inter-group from the GSE56337 data set. 32

Figure 2.8 PCA analysis of heterogeneous sample. 35

Figure 2.9 Impact of removing/regrouping heterogeneous samples. 37

Figure 2.10 Summary of KEGG pathways identified before and after removing/regrouping heterogeneous samples. 39

Figure 2.11 Distinct bypaths in KEGG pathways identified by removing/regrouping heterogeneous samples. 40

Figure 2.12 Comparison of gpsDEGs before/after regrouping or removing heterogeneous samples. 42
Figure 2.13  The standard deviation of replicates’ expression levels of samples from Affymetrix (upper), Agilent (middle), and Illumina (bottom) in site 1.

Figure 2.14  The variances of replicates’ expression levels of samples from Affymetrix (blue), Agilent (red), and Illumina (yellow) using (A) Pearson Correlation Coefficient, and (B) Student t-test.

Figure 2.15  The percentage of validated DEGs identified under different p-values using genetic-distance-based method (left) and fold change (FC) values (right) in new generated data sets.

Figure 2.16  The percentage of validated DEGs identified under different p-values using distance-based method and fold change (FC) values in new generated data sets.

Figure 2.17  The averaged intra- and inter- distance curves before and after removing or regrouping heterogeneous samples in new generated data sets.

Figure 2.18  The DEGs identified by regrouping/removing heterogeneous samples compared with original group method in new generated data sets.

Figure 3.1  Distribution of both true and false samples as a function of integrated confidence score in eleven data sets.

Figure 3.2  Infrastructure of the decision-tree-based meta-strategy. Query sequences normally refer to a pair of microRNA and mRNA sequences.

Figure 3.3  Prediction performances of three individual predictors.

Figure 3.4  Distribution of true and false samples as a function of their predictive scores for (A) miRanda, (B) PITA, and (C) miRDB.

Figure 3.5  Comparison of prediction performance between individual predictors and meta-predictor.

Figure 3.6  Infrastructure of decision-tree-based meta-predictor.

Figure 3.7  Information gain compared to the distribution of positive and negative samples in the D4 data set for (A) miRanda, (B) miRDB, (C) PITA, and (D) TargetScan.

Figure 3.8  Information gain compared to the distribution of positive and negative samples in four D3 series data sets and six D2 series data sets for (A) miRanda, (B) miRDB, (C) PITA, and (D) TargetScan. (Part1)

viii
Figure 3.8  (Continued) Information gain compared to the distribution of positive and negative samples in four D3 series data sets and six D2 series data sets for (A) miRanda, (B) miRDB, (C) PITA, and (D) TargetScan. (Part2)

Figure 3.8  (Continued) Information gain compared to the distribution of positive and negative samples in four D3 series data sets and six D2 series data sets for (A) miRanda, (B) miRDB, (C) PITA, and (D) TargetScan. (Part3)

Figure 3.9  ROC curves of individual predictors in the eleven newly-designed non-redundant data sets.

Figure 3.10  ROC curves of individual predictors in eleven newly-designed data set containing duplicate samples.

Figure 3.11  Overlap and coverage between individual predictors for (A) positive samples and (B) negative samples.

Figure 3.12  Comparison of prediction performance of different predictors in eleven newly-designed data sets under multi-fold cross validation.

Figure 3.13  Comparison of prediction performance of different predictors in the independent data sets that each is associated with one of the eleven newly-designed data sets.

Figure 3.14  Performance of eleven DANN modules in different subsets. (A) Performance of original trained DANN modules in different subsets.

Figure 3.15  Novel microRNA:mRNA interactions identified by mirTarDANN in the eleven newly-designed data sets.

Figure 3.16  Comparison of prediction performance of different predictors in the human microRNA:mRNA interaction training data sets under multi-fold cross-validation.

Figure 3.17  Comparison of prediction performance of different predictors in the human microRNA:mRNA interaction independent test sets that each is associated with one of the eleven newly-designed data sets.

Figure 3.18  Infrastructure of the new meta-strategy.

Figure 3.19  Prediction performance of four component predictors, including DisEMBL, IUPred, and ESpritz.

Figure 3.20  Comparison of balanced accuracy for twenty types of amino acids.
Figure 3.21 Comparison of per-sequence balanced accuracy among AUCpreD (filled circle), PONDR-FIT (open circle), MFDp2 (filled triangle), and this work. 132

Figure 4.1 The ER diagram for protein-protein interaction searching. 146

Figure 4.2 The comparison among eight protein-protein interaction database. 147

Figure 4.3 Analysis of mismatched IDs. 149

Figure 4.4 The false rate in different databases. 152

Figure 4.5 The distribution of proteins in different ranges of FP and FN rates in databases (A) APID; (B) IntAct; (C) MatrixDB; (D) PINA. 153

Figure 4.6 The tissue specific expressed mRNAs and microRNAs. 155

Figure 4.7 The interaction networks in-vitro by searching (A) NRIP1; (B) FGF1; (C) INS-IGF1-IGF2; (D) GATAD1. 156

Figure 4.8 The NRIP1 interaction networks in (A) brain; (B) lung; (C) pancreas; (D) placenta. 157

Figure 4.9 The FGF1 interaction networks in (A) brain; (B) lung; (C) pancreas; (D) placenta. 158

Figure 4.10 The INS-IGF1-IGF2 interaction networks in (A) brain; (B) lung; (C) pancreas; (D) placenta. 159

Figure 4.11 The GATAD1 interaction networks in (A) brain; (B) lung; (C) pancreas; (D) placenta. 160

Figure 4.12 The NRIP1 microRNA-regulated tissue-specific interaction networks in (A) brain; (B) lung; (C) pancreas; (D) placenta. 164

Figure 4.13 The FGF1 microRNA-regulated tissue-specific interaction networks in (A) brain; (B) lung; (C) pancreas; (D) placenta. 165

Figure 4.14 The INS-IGF1-IGF2 microRNA-regulated tissue-specific interaction networks in (A) brain; (B) lung; (C) pancreas; (D) placenta. 166

Figure 4.15 The GATAD1 microRNA-regulated tissue-specific interaction networks in (A) brain; (B) lung; (C) pancreas; (D) placenta. 167
ABSTRACT

The function, behavior, and environmental response of biological systems are essentially determined by the complex interaction and regulation of biomolecules inside the systems. Therefore, it is critical to characterize the inter-molecular interaction and regulation of biomolecules inside these systems. In this direction, many experimental techniques have been developed and these techniques have been used in many different model systems under various conditions. Consequently, a massive amount of data has been generated. These data cover multiple aspects of molecular interaction and regulation, such as protein-protein interaction, microRNA-RNA interaction, gene expression profiles, etc. While carrying rich information, these data may also contain significant levels of errors. In addition, decoding these data to get meaningful molecular mechanisms is still very challenging.

In this dissertation, our recent efforts on data cleaning and data mining were summarized in the following aspects: (1) Microarray gene expression data analysis. Traditional techniques for gene expression analysis are largely dependent on p-values calculated from statistical models. The accuracy and reproducibility of these techniques have serious concerns. Therefore, we designed a distance method based on the Euclidean Distance calculated from the expression levels of a set of pre-ranked genes in both control and treatment groups. The pre-ranked genes are ranked using fold-change. When more and more pre-ranked genes are included in the calculation, the distance is normally monotonically decreasing. Therefore, by selecting a specific cutoff value, a group of genes are identified. In practical, a standard deviation cutoff associated with a sliding window on the curve of distance as a function of number of pre-ranked genes was used to select the group of genes. This group of genes determine the genotypical and phenotypical differences between control samples and
treatment samples. By using data sets developed in the Microarray Quality Control project, the true-positive rate of the distance method was much higher than traditional methods.

(2) The DEGs identified from Microarray data analysis represent genes at the RNA level, but not at the protein level. It is well known that a high level of RNA may not necessarily result in a high level of protein because the translation process from RNA to protein is regulated by multiple factors. To name a few, microRNAs may interact with an mRNA to inhibit the translation or to degrade the mRNA, and thus lead to a low level of protein molecules. Therefore, by integrating RNA-level gene expression analysis and microRNA:mRNA interaction prediction, it is feasible to determine gene expression at protein level, at least for genes of which the translation is mainly regulated by microRNAs. For this purpose, a novel meta-strategy was developed for microRNA target prediction. This strategy improved prediction accuracy significantly, comparing to many other microRNA target predictors.

(3) Biology-specific and microRNA-regulated protein-protein interaction networks. Proteins frequently interact with each other to regulate various biological processes. Experimentally validated protein-protein interaction data have been deposited into databases. Nonetheless, the interaction data in these databases are not specific to tissue, cell, or biological conditions. Besides, the interaction data between databases are not consistent. Finally, many IDs in these databases are mislabeled. The fraction of mislabeled IDs in some databases is as high as 15%. In this project, an automatic pipeline was developed to remove mislabeled IDs, and multiple protein-protein interaction databases have been integrated. Thus, the protein-protein interaction information in new database is more accurate. Furthermore, gene expression analysis under specific biological conditions and microRNA regulated gene expression at the protein level are integrated into protein-protein interaction networks to generate biology-specific microRNA-regulated protein-protein interaction networks.
In summary, we developed novel strategy to identify differentially expressed genes at the RNA level, we also developed meta-strategy based high-accuracy miRNA target predictor to predict microRNA:mRNA base pairing and to identify mRNAs that may not be translated into proteins, we further removed mislabeled IDs from PPI databases and combine multiple PPI databases into a new database. By combining all the above-mentioned results together, we developed a novel strategy and pipeline to determine high-accuracy protein-protein interaction networks, as well as the rewiring of protein-protein interaction networks regulated by specific gene expression and miRNA-induced inhibition of protein synthesis.
CHAPTER 1

INTRODUCTION

The complex biological processes are based on the regulations of thousands of genes and the interactions of genes and the gene products [1]. Current studies of protein-protein interaction networks contain the analysis of both physical and functional interactions [2]. It can provide the information for the interpretation of biological relevant regulation and mechanisms [2]. However, the current protein-protein interaction databases are inconsistent in different sources [3]. The obvious biologically relevant reason for the inconsistent databases is the impact of the subtle inter-molecular interaction and regulation, such as tissue-specific/condition-specific gene expression and gene post-transcriptional regulation by microRNA [4]. Previous studies have suggested the integration of biological relevant information into the analysis of protein-protein interaction networks can improve the validity of the analysis result [5, 6]. Therefore, in this study, we developed several novel strategies to study the inter-molecular interaction and regulation with the consideration of the subtle impact.

Microarray are the most widely used techniques in the studies of gene expression profiles [7]. Identifying the differentially expressed genes (DEGs) is critical for Microarray data analysis [8]. However, the reliability of Microarray results is debatable because the identified DEGs lists are less reproducible even with the similar experiment [9]. Additionally, the statistical methods used in gene expression analysis increase the confusion in Microarray expression data analysis because of the variety of results from the analysis [10]. Moreover,
the pre-determined values used in those statistic models are arbitrary, and their correlation with biology is not clear [11].

In fact, the measured expression level of a gene is actually an averaged value of all replicates in the same group [12]. The presence of heterogeneous samples may affect the averaged value of all replicates, thus, masking or confounding the differences between samples in different groups [13]. Current methods used to determine heterogeneous samples require the pre-selected gene list [14]. The change of the pre-selected list can cause the determined heterogeneities significantly different [15]. Besides, the impact of heterogeneities on the results of gene expression were less studied [16]. Although previous studies have suggested the presence of heterogeneity can cause the bias of Microarray data analysis, there is no general pipeline used to estimate the impact of heterogeneities on the determination of DEGs [17].

The protein producing processes can be regulated by different biological mechanisms [2], one of which is regulated by microRNA. microRNA can bind to mRNA, leading to the degradation and/or translational inhibition of mRNAs [18]. microRNAs are small non-coding RNA, which plays a critical role on the whole genome-wide gene regulation [19, 20]. Around 2,000 miRNAs have been identified, of which 60% are involved in the regulation of genomes [21, 22]. Thus, determining the microRNA target pairs is critical for the study of the whole-genome regulation. Many computational methods have been developed to predict microRNA and microRNA target genes [23, 24, 25, 26]. However, current existing predictors were developed using the inconsistent microRNA and microRNA target datasets [27], resulting in a severely low prediction performance in applications. In addition, some machine-learning based predictors were developed using a great number of features, causing the intrinsic noise associated with the features in the predictors [28, 29].

Complex biological processes are usually regulated through protein-protein interaction networks [1]. Current protein-protein interaction databases are inconsistent from sources to
Figure 1.1. Overall description of the dissertation, including the designed strategies to identify differentially expressed genes (purple), predict microRNA and microRNA target genes (Yellow), and synthesize protein-protein interaction with tissue-specific expression information (Red).

sources [30, 31, 32]. Therefore, the most eager goal is to establish a consistent protein-protein interaction network. The validity of the analysis results can be improved by integrating the dynamic biological relevant processes into the protein-protein interactions to rewire the networks [5, 6]. However, the prior strategies used to analyze the protein-protein interactions were limited on inferring the tissue-specific protein-protein interaction networks.

In this dissertation, we focus on solving previous mentioned problems by developing different novel strategies, which means the final output of the synthesized strategy can provide the rewired protein-protein networks by including both gene expression analysis and the impacts of microRNA regulation on the post-transcriptional regulation. To achieve this goal, we work on three inter-related directions of researches and shown in Figure 1.1. Firstly,
we developed novel genetic-distance-based strategies in gene expression analysis to identify DEGs that determine the genotypic and phenotypic differences between samples. Additionally, the dynamic principle component analysis (PCA) bi-plot distance curve was applied to determine the heterogeneous samples and the integrated distance strategy was used to evaluate the impacts of heterogeneous samples on the determination of DEGs. Secondly, we developed decision-tree-based meta-strategy to improve the prediction performance of miRNA and miRNA target genes. The last but not the least, we synthesized the analysis of the gene expression and the regulation of miRNAs into protein-protein interaction analysis. We established a integrated protein-protein interaction data set and rewired the protein-protein interaction networks using tissue-specificity results.

1.1 Data Necessity and Availability

Data sets are crucial in different scientific fields because they are important for researchers to develop and test the new designed strategies. New data sets are continuously generated. In the meantime, there are many new experimental results included in the existing data set to extend the coverage of those data sets. A good data set is necessary for designing and verifying the strategies. To select the data sets, we need to consider several critical questions: (1) Whether the data set is public available; (2) What information is included in the data sets; (3) What are the differences between different data sets in the same field. After that, we can determine what data sets can be used to design the strategies, or what data sets can be combined to improve the existing data sets. Our aim is to design novel strategies to characterize the subtle inter-molecular regulation, by integrating miRNA interaction regulation and gene expression analysis. It is necessary for us to select the suitable data sets for the researches.
1.1.1 Data Sets Used in Designing Genetic Distance-based Method

To design genetic distance-based method to identify genotypically and phenotypically significantly differentially expressed genes (gpsDEGs), Eight Microarray data sets were extracted from the Gene Expression Omnibus (GEO) database. Those eight data sets are about acute lymphoblastic leukemia, breast cancer, colorectal cancer, kidney cancer, liver cancer, lung cancer, pancreatic cancer, and prostate cancer. Their corresponding GEO entries are GSE19315 [33], GSE53394 [34], GSE7259 [35], GSE65168 [36], GSE41806 [37], GSE6400 [38], GSE57728 [39] and GSE22606 [40].

1.1.2 MicroRNA:mRNA Interaction Predicted Databases

To build the infrastructure of microRNA:mRNA interaction predictor, we applied four different microRNA:mRNA interaction databases, MiRanda [18], MiRDB [41], PITA [42] and TargetScan [21], which include both true pairs and false pairs of human and mouse species, were used to establish the data sets. The two different experimental validated microRNA:mRNA interaction database, MiRTarBase [43] and DIANA-TarBase v7.0 [44] were applied to label the true pairs of microRNA:mRNA interactions.

1.1.3 Protein-protein Interaction Databases

Experimental validated Protein-protein interaction databases are extracted from eight different online databases: (1) BioGrid [45, 46, 47]; (2) APID [31]; (3) PINA [48]; (4) DIP [49, 50, 51]; (5) IntAct [52]; (6) MatrixDB [53, 54, 55]; (7) InnateDB [56]; and (8) HPRD [57, 58]. These databases were combined to establish the new comprehensive database.
1.1.4 Data Sets Used in Rewiring Protein-protein Interaction Database

Gene expression data sets were downloaded from Gene Expression Omnibus (GEO) databases. They are GSE7905 [59], GSE1166 [60] and GSE7307 [61], including up to 31 different tissues in Homo Sapiens. Protein expression data sets were extracted from Paxdb [62, 63] and ProteomicsDB [64], containing 17 different tissues. microRNA expression data were collected from database miRmine [65] and TissueAtlas [66]. These two database include microRNA expression levels in 11 different tissues.

All of the databases were applied to develop the novel strategies in different steps. We acknowledge with thanks the application of all the previous mentioned databases.

1.2 Contribution

The contributions of this thesis include

1. Designed the genetic distance-based strategy to identify genotypically and phenotypically significantly differentially expressed genes (gpsDEGs).

2. Applied Dynamic PCA bi-plot distance to determine heterogeneous samples.

3. Designed integrated distance-based method to estimate the impacts of heterogeneous samples on the identification of gpsDEGs.

4. Presented the decision-tree based meta-strategy and integrated with Artificial Neural network to develop microRNA:mRNA interaction predictors.

5. Established a comprehensive protein-protein interaction data set and rewired protein-protein interaction networks by using tissue-specificity expression profiles of genes, proteins, and microRNAs.
1.3 Dissertation Overview

In Chapter 2, we presented the genetic distance-based method for gpsDEGs identification. Based on the distance-based method, we extended the distance-based method on heterogeneous samples determination and impacts of heterogeneous samples measurement. In Chapter 3, we presented the novel strategy by integrating decision-tree-based meta-strategy and Artificial Neural network to improve the microRNA:mRNA interactions. In Chapter 4, the comprehensive protein-protein interaction database was established. We further rewired the interaction network using tissue-specificity information at gene, protein, and microRNA expression levels.
CHAPTER 2

GENETIC-DISTANCE-BASED STRATEGY TO IDENTIFY GENOTYPICALLY AND PHENOTYPICALLY SIGNIFICANTLY DIFFERENTIALLY EXPRESSED GENES

2.1 Note to Reader


2.2 Contributions and Acknowledgements

I would like to thank my adviser Dr. Bin Xue for his kind contributions in completion of these works. These works were supported by the start-up funding from the Department of Cell Biology, Microbiology and Molecular Biology, College of Arts and Sciences at the University of South Florida to Dr. Bin Xue.
2.3 Introduction

Microarray is a powerful technique that is widely used for the study of the gene expression profiles at the genome level [7]. The Microarray technique has been invented in the early 1980s [67]. Since then it has been widely used in the study of different cell lines, species, tissues, and under different conditions [8]. In the same Microarray experiment, different but related samples are used to ensure the reliability of the experimental results. The samples are usually assigned into different groups. For instances, the samples used as a reference group are called “control” samples, while the samples under different phenotypic status or various treatments are called “target” or “treatment” samples. In the same Microarray experiment, it can include more than one treatment group. In each group, normally multiple replicates are included. The gene expression levels may be different among those different groups. When the expression levels of genes are significantly different, those genes are selected as differentially expressed genes (DEGs), which are assumed to be the molecular driving force and/or the molecular biomarkers for the different phenotypes.

Many statistical methods have been developed and applied to identify the DEGs between samples [12, 68, 69, 70, 71]. The differences of expression level of a gene between different samples is shown by the ratio of the expression level of the gene in treatment group to the expression level of that gene in control group [10]. The absolute value of the ratio in base 2 logarithm is known as Fold-change (FC) [68]. FC is a critical value used to show the change of expression level of a gene between experimental groups [69, 71]. Normally, FC 1.5 or 2.0 is used as cutoff values to choose DEGs. Another statistic method used to identify DEGs is the student t-test [70]. It measures the statistical significance of the genes based on null hypothesis that the genes are not differentially expressed in control group and treatment groups [70]. The p-value of t-test is often used to select the DEGs [70]. Since the student t-test is very simple, there is another modified method, significance analysis of
microarrays (SAMs), designed to refine the results of t-test, and this method is widely used since then [70]. There are many new methods developed using different statistical models and algorithms, such as Rank product, Outliers Sums, Outlier Robust, Ranking Analysis, Mao and Sugihara’s method, and SPRING [72, 71, 73, 74, 75, 76].

In the recent study on Microarray Quality Control (MAQC) project, the researchers found the combination of non-stringent p-value from student t-test with FC ranking can significantly improve the efficiency of DEGs identification [10, 11, 77]. Nonetheless, it still has the challenge because it requires the pre-determined cutoff values [68]. In this direction, many strategies have been tested. For instances, the probability-based decision-tree was used to identify DEGs in breast cancer [76]. Additionally, the distance between the distribution of a gene’s expression profiles and the statistical significance associated with the distance was applied to select DEGs [76].

However, those methods have their limitations. For example, FC method needs the pre-determined cutoff values. Sometimes the selected DEGs is because of the noise, not the actual genes which have different expression levels [4, 69]. Student t-test is based on the null hypothesis, there is no significant difference between control group and treatment groups [12]. It assumes the expression levels of genes in one sample are normal distributed [12]. In fact, the distribution of genes in the same sample are not normal. The DEGs identified under assumption may be bias. Several newly-designed case specific methods are based on the specific condition, they are not very effective for the general conditions [9, 78, 79]. In a recent comprehensive comparison using data sets generated by the Microarray Quality Control (MAQC) project [9], the combination of non-stringent p-value from student t-test and FC ranking was shown to improve the efficiency of DEG identification significantly [10]. This study provides a generic strategy for the identification of DEGs, which has become rather prevailing in Microarray data analysis. Nonetheless, it is still challenged because this strategy needs the pre-determined cutoff values.
Taken together, all of those current existing methods may eventually lead to the high false-positive rates and low reproducibility in Microarray data analysis. Moreover, these “data analysis” based methods may not adequately address the biology behind the data of Microarray experiments. In our recent published paper, “How many differentially expressed genes: A perspective from the comparison of genotypic and phenotypic distances”, we stated the genetic-distance-based method can be used to identify the genotypically and phenotypically significantly differentially expressed genes (gpsDEGs) [80]. In the section 2.2, we will display the experimental materials and methods, experimental results and discussions of the genetic-distance-based method. In living species, each functional unit such as tissues and organs is made up of various types of cells [13, 81]. The cells may have different genetics and may be modulated by different intrinsic/extrinsic factors [13, 82]. Therefore, tissue-based samples may be very different from each other. This is known as tissue (or sample) heterogeneity [82, 83]. Due to the presence of tissue heterogeneity, the measured expression level of a gene is actually an averaged value over the heterogeneous subgroups of samples [82]. Consequently, the result may be biased when using these expression level of genes to identify DEGs and to analyze corresponding molecular mechanisms [14, 83].

Traditional strategy used to determine sample heterogeneity is principal component analysis (PCA) [84]. When applying the PCA bi-plot to identify heterogeneous samples, it needs a specific list of genes, the DEGs [84, 85]. However, as it has been well known, the identified DEGs may not include all the genes that are able to characterize the differences between samples [85]. Besides, different statistical strategies and cutoff values can be used for DEG identification. These alternatives will result in different sets of DEGs [85]. In this case, it is not clear if the obtained PCA bi-plot will be the same or not. Consequently, the reliability of obtained PCA bi-plots is of concern.

Another popular method used to determine the heterogeneous samples is Pearson correlation co-efficiency [86, 87, 88]. This method requires the whole batch of genes to estimate the
similarity between two samples [87]. Most of genes in the batch cannot represent differences between samples. These useless genes included in the analysis may mask the heterogeneous samples [86].

Another critical problem is to measure the impact of heterogeneous samples on the determination of DEGs. There is no general principle currently used to evaluate the impact of heterogeneous samples.

In our recent research published in IEEE proceeding, “Measuring the Inter-sample Heterogeneity by Dynamic PCA bi-plot”, we found the application of dynamic PCA bi-plot and its distance curves can provide additional information for the identification of heterogeneities [89]. We extended this result on murine intestinal cancer tissue to find strategies on how to deal with the heterogeneous samples. The manuscript “Distinct genetic responses of murine intestinal cancer tissues identified from existing microarray data by pre-processing inter-sample heterogeneity” was prepared. In section 2.3 and 2.4, the experimental details are performed.

To confirm those new designed strategies, we tested those strategies on the standard data set, the MAQC data set. The performances of those new strategies were evaluated on the data sets achieved from three different platforms. The results were shown in section 2.5.

2.4 How Many Differentially Expressed Genes: A Perspective from the Comparison of Genotypic and Phenotypic Distances

The paper “How many differentially expressed genes: A perspective from the comparison of genotypic and phenotypic distances” was published in Genomic in 2017. In this paper, we applied genetic distance-based method to analyze eight different cancer data sets from Gene Expression Omnibus (GEO) database. Here, we introduced the detailed experimental methods and results.
2.4.1 Materials and Methods

2.4.1.1 Microarray Data Sets

Eight sets of Microarray data sets were extracted from the GEO database. These eight different data sets were performed using Affymetrix platform and designed to evaluate the responses of different treatments on a specific type of carcinomatous cell line. The eight types of cancer are: acute lymphoblastic leukemia (GSE19315 [33]), breast cancer (GSE53394 [34]), colorectal cancer (GSE7259 [35]), kidney cancer (GSE65158 [36]), liver cancer (GSE41804 [37]), lung cancer (GSE6400 [38]), pancreatic cancer (GSE57728 [39]), and prostate cancer (GSE22606 [40]). The summary of those GEO data sets was shown in Table 2.1.

<table>
<thead>
<tr>
<th>GEO Entry</th>
<th>Type of Cancer</th>
<th>Tissue/Cell Line</th>
<th>GPL</th>
<th>Year &amp; Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE19315</td>
<td>Acute Lymphoblastic leukemia</td>
<td>THP-1</td>
<td>GLP570</td>
<td>2010 [33]</td>
</tr>
<tr>
<td>GSE53394</td>
<td>Breast cancer</td>
<td>MCF-1</td>
<td>GPL96</td>
<td>2013 [34]</td>
</tr>
<tr>
<td>GSE7259</td>
<td>Colorectal cancer</td>
<td>Caco-2</td>
<td>GPL571</td>
<td>2007 [35]</td>
</tr>
<tr>
<td>GSE65168</td>
<td>Kidney cancer</td>
<td>RCC (786-O)</td>
<td>GPL6244</td>
<td>2015 [36]</td>
</tr>
<tr>
<td>GSE41804</td>
<td>Liver cancer</td>
<td>IL28B and TG/GG</td>
<td>GPL570</td>
<td>2012 [37]</td>
</tr>
<tr>
<td>GSE6400</td>
<td>Lung cancer</td>
<td>A549</td>
<td>GPL570</td>
<td>2006 [38]</td>
</tr>
<tr>
<td>GSE57728</td>
<td>Pancreatic cancer</td>
<td>AsPC1</td>
<td>GPL570</td>
<td>2014 [39]</td>
</tr>
<tr>
<td>GSE22606</td>
<td>Prostate cancer</td>
<td>LNCaP</td>
<td>GPL570</td>
<td>2010 [40]</td>
</tr>
</tbody>
</table>

2.4.1.2 Bioconductor and DEG Analysis

The CEL format files of those eight sets were pre-processing using RMA function in Affy package in R for background correction and normalization [90]. Afterwards, limma package was applied to calculate p-value, adjusted p-value and FC for all genes in the eight cancer data sets [91, 92]. The p-value cutoff 0.1 was used to select the preliminary list of genes and the selected gene list was ranked using FC [10].
### 2.4.1.3 Intra-group and Inter-group Distance

The Microarray data sets used in this study normally contain one control group and two or three treatment groups. In each group, it usually has three or four replicates. After achieving the ranked gene list, the scaled expression levels of the genes in different samples were used to calculate the genetic distance between samples by using the following equation:

\[
d_{j,j_2} = \frac{1}{N} \sqrt{\sum_{k=1}^{\infty} (E_{j,k} - E_{j_2,k})^2}
\]

where, \(d_{j,j_2}\) is the distance between sample \(j\) and \(j_2\). \(N\) is the total number of ranked genes in the calculation. \(E_{j,k}\) is the scaled expression level of the \(k\)-th gene in the \(j\)-th sample. The scaled expression level is the ratio of the expression level of a gene in a treatment group to the averaged expression level of the same gene in the control group.

In addition, all the samples in the \(i\)-th group can be characterized by an intra-group distance \(d_i\), which is calculated by:

\[
d_i = \frac{1}{N \times C(J, 2)} \sqrt{\sum_{j=1}^{J-1} \sum_{j_2=j+1}^{J} \sum_{k=1}^{N} (E_{i,j,k} - E_{i,j_2,k})^2}
\]

where, \(E_{i,j,k}\) is the scaled expression level of the \(k\)-th gene in the \(j\)-th sample that belongs to the \(i\)-th group. \(J\) is the total number of samples in the \(i\)-th group. \(j\) and \(j_2\) stand for two different samples in the \(i\)-th group. \(C(J, 2)\) is the number of pairwise combinations of all the \(J\) samples in the \(i\)-th group.

Similarly, the inter-group distance \(D_{i,i_2}\) can be calculated using:

\[
D_{i,i_2} = \frac{1}{N \times J \times J_2} \sqrt{\sum_{j=1}^{J} \sum_{j_2=1}^{J_2} \sum_{k=1}^{N} (E_{i,j,k} - E_{i_2,j_2,k})^2}
\]
Here, $i$ and $i2$ represent the $i$-th and $i2$-th groups, respectively. $E_{i,j,k}$ and $N$ have the same meaning as explained before. $J$ and $J2$ are the numbers of samples in the $i$-th and $i2$-th group, accordingly.

### 2.4.1.4 Standard Deviation of Distance Curve

Distance curves may have different scales of saturated values and different levels of fluctuations. The sliding window was taken to calculate the standard deviation (SD) of distance in the window to show the fluctuations of the curves. The smaller of the SD value indicates the fluctuation of distance becomes small and stable. Therefore, the genetic distance between samples becomes stable. For this reason, the SD value can be used to select an optimal set of DEGs which can represent the differences between samples. Since there are multiple distance curves calculated using different pairs of samples in the same group or in different groups, the averaged SD curves for inter-group and intra-group were finally used to select gpsDEGs.

### 2.4.2 Results

#### 2.4.2.1 Inter-group or Intra-group Distance Curves

The expression levels of top 2000 genes in data set GSE22606 were ranked by p-value and FC value, shown in Figure 2.1. In the Figure 2.1 (A), all the genes were ranked using p-value, while in Figure 2.1 (B), all the genes were ranked using FC. It is clear that using p-value to rank genes, some genes in the top-list with the low p-values may have the small FC values and some genes with high p-values may have the large FC values. Similarly, using FC value to rank genes, the top-list genes with large FC values may have the high p-values, while some genes with small FC values may have the small p-values. Since the top-list are usually selected as DEGs, it is clear that no matter using p-value or FC to select DEGs,
Figure 2.1. Expression levels of the top 2000 genes from GSE22606 data set ranked by (A) p-value and (B) FC. All of figures have two y-axis showing FC at left and p-value at right. All p-values were shown in log-scale. X-axis show the sequential order of the 2000 genes. In each figure, three panels, from top to bottom, are LNCaP cells treated by R1881, LNCaP cells treated with siRNA-mediated SRF silencing, and LNCaP cells treated by SRF silencing together with R1881.

There are pros and cons. The identified DEGs is very challenging. Anyhow, in Figure 2.1, the genes with lower p-value are normally have the large FC and the genes with lower bound of p-value are increasing as the FC decreasing. This is evidence to support the strategy that using non-stringent p-value and stringent FC to select DEGs.

The intra-group distances and inter-group distances were calculated using ranked genes. The result of those distances of data set GSE22606 were shown in Figure 2.2 (A). In data set GSE22606, there are four groups of samples. Thus, there are four intra-group distance curves and six inter-group distance curves. The intra-group distance curves are generally ended when they reached to their corresponding saturated values. Those intra-group curves can be roughly split into two segments: No.genes $< \sim 100$ or No.genes $> \sim 100$. That is, when the number of ranked genes is less than 100, the four intra-group distance curves
have their own patterns and separated from each other. The intra-group distance curves of both control group and SRF-silencing group increase at the beginning and then decrease gradually. The intra-group distance curves in other groups which are related with R1881 treatment increase slowly at the beginning and then drop back slightly. This observation indicates the treatment of R1881 may suppress the differences of gene expression profiles. When the number of ranked genes are more than 100, three intra-group distance curves come close to each other. They are control group, SRF-silencing group, and SRF-silencing group treated with R1881. One group are lower than other three curves, which is R1881 treatment group. This result indicates the samples with the treatment of R1881 has less differences on expression profiles.

For inter-group distance curves, we observed two different sets of distance curves. The first set includes the distance curves between the control group and the SRF-silencing group, and two groups related to R1881 treatment. In this set, the inter-distance curves increase rapidly when the number of ranked genes is less than 20 and then drop to the equilibrium values. The second set contains four distance curves of the rest four combinations of groups. In this set, all the distance curves decrease quickly when the number of ranked genes is less than \( \sim 200 \), and then decrease consistently but slowly when the number of ranked genes is larger than \( \sim 200 \). The distance curves in the first set are always lower than those in the second set.

In Figure 2.2 (B), it shows the interplay between the ranked genes and the ending behavior of calculated distance curves. The top 500 ranked genes were randomly shuffled to generate the new ranked gene lists, and the distance curves of new ranked gene lists were calculated. Compared the shuffled process distance curves with the original curves in Figure 2.2 (A) (black dashed line), the ending behaviors of those compared distance curves are very different. First, the highest distances calculated from randomly shuffled genes are much smaller than the original distance curve. Second, the monotonic decreasing of the original distance curve
is much steeper. Third, the distance curves of shuffled genes have larger fluctuations. Those observations indicate the distance curves calculated from the originally ranked genes are more effective to show the differences between samples.

### 2.4.2.2 Selecting gpsDEGs

The sliding windows along with the distance curves were selected and the standard deviation (SD) in each sliding window were calculated and shown in Figure 2.3. Since the SD values in each sliding window can characterize the fluctuation of curves and the stability of differences between samples, the SD curves of intra-group or inter-group distance curves are calculated. In Figure 2.3, the SD curves continuously decrease as the increased number of ranked genes included. The SD curves of intra-group distances and inter-group distances
Figure 2.3. Fluctuation of distance curve demonstrated by standard deviation in a sliding window of size 9 as a function of the number of ranked genes in the GSE22606 data set. Y-axis on the left shows the fluctuations of the intra-group distances, and the y-axis on the right-hand-side shows the fluctuations of the inter-group distances. Both y-axes are in log-scale. X-axis shows the number of ranked genes used in the calculation. Dotted line shows the SD cutoff of the inter-group distance curve at 0.01, while the dashed-line shows the SD threshold of the inter-group distance curve at 0.005.

are in two different classes. The intra-group SD curves drop quickly down to 0.01 at the first ∼50 genes, and then down to 0.005 at the first ∼150 genes, while the inter-group SD curves are higher than 0.01 when the number of ranked genes is less than ∼200. The SD value dropped to 0.005 when there are ∼400 ranked genes included in the calculation. Here, 0.01 and 0.005 are two cutoff values labeled in the Figure 2.3, indicating the fluctuation of the distance between groups. Therefore, the SD curves of distances can be used to display the stability of differences between samples in different groups, and be used to identify biologically significant DEGs which can represent the genotypic and phenotypic differences between groups.

The simplified SD curves which are calculated by averaging all the intra-group curves or all the inter-group curves of data set GSE22606 were shown in Figure 2.4, with another seven GEO data sets. In some data sets, the simplified intra-group SD curve and inter-group SD curve are well separated, such in lung cancer and prostate cancer data sets, indicating
the significantly differences between groups. However, the simplified intra-group SD curve and inter-group SD curve in liver cancer data set are very close to each other. Applying SD cutoff 0.01 to select genes from simplified inter-group SD curve, there will be a specific list of gpsDEGs can be identified from each data set. The number of gpsDEGs are changed from \( \sim 100 \) (GSE57728, pancreatic cancer) to \( \sim 190 \) (GSE19315, acute lymphoblastic leukemia) in those eight cancer related GEO data sets. These numbers of genes indicate the smallest of gpsDEGs that can represent the differences between samples in different groups. The upper-bounds of p-values associated with gpsDEGs in those eight sets are range from \( 10^{-5} \) to \( 10^{-2} \). When comparing the numbers of DEGs identified by FC method and the numbers of gpsDEGs identified by SD method, three categories of results can be seen: (1) The number of DEGs from FC method is much less than that of gpsDEGs from the SD method, such as in the cases of GSE53394 (breast cancer), GSE41804 (liver cancer), and GSE57728 (pancreatic cancer); (2) the number of DEGs from FC method is much more than that from the SD method, such as in the GSE6400 data set of lung cancer; (3) the numbers are comparable. Examples include GSE7259 (colorectal cancer), GSE65168 (kidney cancer), GSE19315 (leukemia), and GSE22606 (prostate cancer).

Clearly, the numbers of gpsDEGs identified using the genetic-distance-based method is more comparable among different studies. To demonstrate the reliability of the gpsDEGs identified using this new method, the corresponding highest adjusted p-value of DEGs identified using different methods, including FC1.5, FC2.0 and SD0.01 were compared and shown in Table 2.2. In most of data sets, the highest adjusted p-value of DEGs was less then 0.05, except in data set GSE41804, the highest adjusted p-value was 0.096 for all of the methods used to determine DEGs. In data set GSE57728, the new method identified 100 gpsDEGs which were more than 8 and 2 DEGs identified using FC cutoff 1.5 and 2.0. Although the increased number of DEGs identified, the highest adjusted p-value of the new method only slightly increased from \( 4e-6 \) to \( 6e-6 \).
Figure 2.4. Fluctuation averaged over all the intra-group distance curves (gray dotted line) or all the inter-group distance curves (black) as a function of the number of ranked genes in every of the eight Microarray data sets (AH). X-axis is the number of ranked genes. Y-axis on the left shows the standard deviation, while y-axis on the right shows the log-scaled p-value of each gene (gray). The horizontal dashed lines label SD cutoff at 0.01. Two vertical lines show the numbers of DEGs selected using FC cutoffs 1.5 (dash-dash-dash) and 2 (dot-dot-dash), respectively.
Table 2.2. Numbers of DEGs/gpsDEGs and the corresponding highest false discovery rate (FDR) adjusted p-values based on three different methods.

<table>
<thead>
<tr>
<th>GEO entry</th>
<th>Tissue</th>
<th>FC1.5 No.</th>
<th>FC1.5 p*</th>
<th>FC2.0 No.</th>
<th>FC2.0 p*</th>
<th>SD0.01 No.</th>
<th>SD0.01 p*</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE53394</td>
<td>Breast</td>
<td>39</td>
<td>0.021</td>
<td>23</td>
<td>0.021</td>
<td>111</td>
<td>0.048</td>
</tr>
<tr>
<td>GSE7259</td>
<td>Colorectal</td>
<td>310</td>
<td>0.027</td>
<td>153</td>
<td>0.027</td>
<td>106</td>
<td>0.027</td>
</tr>
<tr>
<td>GSE65168</td>
<td>Kidney</td>
<td>150</td>
<td>0.035</td>
<td>64</td>
<td>0.027</td>
<td>150</td>
<td>0.035</td>
</tr>
<tr>
<td>GSE19315</td>
<td>Leukemia</td>
<td>274</td>
<td>0.049</td>
<td>166</td>
<td>0.036</td>
<td>188</td>
<td>0.036</td>
</tr>
<tr>
<td>GSE41804</td>
<td>Liver</td>
<td>61</td>
<td>0.096</td>
<td>17</td>
<td>0.096</td>
<td>156</td>
<td>0.096</td>
</tr>
<tr>
<td>GSE6400</td>
<td>Lung</td>
<td>833</td>
<td>3e-5</td>
<td>483</td>
<td>2e-5</td>
<td>149</td>
<td>7e-6</td>
</tr>
<tr>
<td>GSE57728</td>
<td>Pancreatic</td>
<td>8</td>
<td>4e-5</td>
<td>2</td>
<td>6e-6</td>
<td>100</td>
<td>6e-5</td>
</tr>
<tr>
<td>GSE22606</td>
<td>Prostate</td>
<td>221</td>
<td>2e-5</td>
<td>102</td>
<td>9e-6</td>
<td>154</td>
<td>2e-5</td>
</tr>
</tbody>
</table>

N.B. “No.” shows the number of DEGs/gpsDEGs. “p∗” stands for the highest FDR adjusted p-value. FC1.5 and FC2.0 indicate that the FC cutoff values of 1.5 and 2.0 were used to select the final list of DEGs after all the genes were selected by a non-stringent p-value and then ranked by FC. SD0.01 shows that 0.01 is used as the cutoff value of SD. The calculation of SD was described in the method section.

2.4.3 Discussion

In the recent study, the researchers demonstrated that using the combination of non-stringent p-value and FC can improve the efficiency of DEGs identification [10]. In this method, a non-stringent p-value is applied to select a preliminary list of genes, and then the list of genes is ranked using FC. At last, the FC cutoff is used to select the final list of DEGs. This method has the advantages because the DEGs from this method are more reproducible and has less false-positive rate. Therefore, this method has been widely used in practice. As the increased applications of this method, some less-convenient aspects emerged. For example, when applying this method, it needs the appropriate selections of cutoff values. Although the FC cutoff is normally set as 1.5 or 2.0, it is rather arbitrary on determining those cutoff values. Moreover, the selection of those cutoff values is not clearly connected to the biological phenomenon. Hence, we developed a new strategy in this study to calculate the intra-group or inter-group distances using the scaled expression level of pre-ranked genes which were selected using above-mentioned protocol. The distance curve as a function of
the number of top ranked genes was made, and the SD of the distance curve within a sliding window was analyzed. The stability of intra-group and inter-group distances were measured using SD values. The inter-group SD values were also used to select a set of genes that can represent the genotypic and phenotypic differences between samples in different groups. In this way, the selection of SD cutoff values is more related to biology. By analyzing eight sets of Microarray data, we found the well tempered SD cutoff is 0.01. The false-discovery rate (FDR) is usually less than 0.05, except in data set GSE41804.

The comparison among methods FC1.5, FC2.0, and SD value 0.01 provides another rationale that using SD value to select DEGs is more suitable. When applying FC 1.5 to identify DEGs in data sets GSE53394 and GSE57728, about 40 and 10 DEGs were selected. Even though these lists of genes are informative for further analysis, it is more possible that there are many complex regulatory processes missed based on the small sets of genes. In another case, when using FC method to select DEGs in data set GSE6400, more than 800 DEGs were identified. Clearly, this list of genes are not specific for the further experimental analysis. When applying SD method, the number of gpsDEGs is normally range from 100 to 200 in these eight data sets. Since the number of identified genes is more compromise, this method may have multiple advantages in subsequent analysis.

2.5 Measuring the Inter-sample Heterogeneity by Dynamic PCA Bi-plot

One possible aspect that may affect the identification of DEGs is the presence of heterogeneous samples. The existing methods have the limitation because they require the pre-selected gene list. The changes of the list may cause the results very different. We published a paper “Measuring the Inter-sample Heterogeneity by Dynamic PCA Bi-plot” in IEEE Bioinformatics and Biomedicine conference. In this paper, we stated the dynamic PCA bi-plot as well as dynamic PCA bi-plot distance curves can provide additional information to determine heterogeneities in samples.
2.5.1 Methods

Two data sets were downloaded from the GEO database. In each data set, there were two experimental groups. In each group, there were three replicates. The data set GSE47734 was generated using Illumina Beadarray, while the data set GSE47735 was from Illumina RNAseq [82].

The raw data of those two data sets were pre-processed and analyzed using Limma package in BioConductor [14, 91, 92]. The non-stringent p-value cutoff (p=0.1) was used to pre-select a list of genes, and then the fold-change (FC) was used to rank the pre-selected list and make a new list, named preliminary ranked genes (PRGs) list. Various number of top PRGs were used to calculate the locations of the same sample in PCA bi-plot and the distance between samples in PCA bi-plot was used to display the function of the number of genes used in PCA analysis. Afterwards, there were two kinds of new plots generated, named dynamic PCA bi-plot and dynamic PCA bi-plot distance curve.

2.5.2 Results

Figure 2.5 presents the dynamic PCA bi-plots for all the samples in data sets GSE47734 and GSE47735. A point in the dynamic PCA bi-plot displays the location of the sample which was analyzed using the specific number of PRGs. The next point along with the previous point shows the location of the sample which was analyzed using all the PRGs used in the previous point and the next PRGs in the PRGs list. The samples in the same experiments move to the same direction in the dynamic PCA bi-plot. We also observed the large distance changes in all the panels. Normally, these large distance changes were located when the number of PRGs changes significantly.

The Euclidian Distances of each two samples in the same group in dynamic PCA bi-plot were calculated and shown in Figure 2.6. Generally, the inter-sample distance increases
Figure 2.5. Dynamic PCA bi-plot of each sample for all the samples in the GSE47734 (A) and GSE47735 (B) data sets. Each panel shows the dynamic PCA bi-plot of a sample using a continuously changed number of PRGs ranging from 2 to 350. A1 to A6 stand for samples 1 to 6 in the GSE47734 data set, while B1 to B6 represent samples 1 to 6 in the GSE47735 data set. Samples 1, 2 and 3 are in the control group, and samples 4, 5, and 6 come from the treatment group. X-axis and y-axis are the first (PC1) and second (PC2) principal components, respectively. Each symbol on the curve represents the location of the sample in the bi-plot under a specific number of PRGs. “S” stands for “start”, indicating the sample was analyzed using two PRGs. Consequently, the other end of the curve in each panel denotes the location of sample in bi-plot analyzed by using 350 PRGs.
steadily as the increasing number of PRGs used in PCA analysis and distance calculation. In Figure 2.6(A), the distance of d56 decreases when ~150 PRGs included in distance calculation. Most importantly, the number of PRGs influences the difference between distances significantly. For example, in Figure 2.6(A), when the number of PRGs is ~75 (FC=1.5), d45 and d56 are much larger than d46. This situation indicates the sample 5 is more different from sample 4 and 6, and should be identified as a heterogeneous sample. However, when the number of PRGs increases to ~150 (FC=1.2), d46 and d56 were close to each other, but d45 is much larger than d46 and d56. This result indicates the differences between sample 4 and sample 5 are very large, and the distances between those two samples to sample 6 are the same. This situation makes the determination of heterogeneous samples tricky. Similarly, we observed the situation in Figure 2.6(B), when using FC 2.0 and 1.5 to select PRGs and calculate distances. Furthermore, it was observed that the ending behavior of the dynamic PCA bi-plot distance curve can provide a unique perspective for the heterogeneous sample identification.

2.5.3 Discussion

As shown in the dynamic PCA bi-plot distance curve, the selection of different number of DEGs influenced the identification of heterogeneous samples critically. The distance curve of dynamic PCA bi-plot is a tool to display the dynamic changes of inter differences along with the number of selected genes used in the analysis. This curve can provide additional information for inter-sample heterogeneous sample identification and measurement, and can be used in other related studies. Since the selection of DEGs is still a challenge, it needs further studies of this distance curve when applying in practice.
Figure 2.6. Distance between samples in the dynamics PCA bi-plot for samples in (A) GSE47734 and (B) GSE47735 data sets. X-axis shows the number of PRGs used in PCA analysis with maximally 350 PRGs, and y-axis shows the distance between every two samples of the same group in the dynamic PCA bi-plot. Gray lines are intra-group distances in the control group, and blue line are intra-group distances in the treatment group. Dashed vertical lines show the numbers of PRGs selected by using FC cutoffs 2.0, 1.5, and 1.0, respectively, when the values are available.

2.6 Distinct Genetic Responses of Murine Intestinal Cancer Tissues Identified from Existing Microarray Data by Pre-processing Inter-sample

The samples in many Microarray expression are mixture of many heterogeneous subtypes, causing the analytic results to be biased [82]. It is necessary to determine the heterogeneous samples and then characterize the impacts of heterogeneous samples on gene expression analysis [81, 93, 94, 95, 96].

In this study, we extracted twenty-five Microarray data sets of murine intestinal cancer tissues from GEO database to characterize the heterogeneity of intestinal tissue samples. We selected intestine tissues because of a couple of reasons. First, Colorectal Cancer (CRC) is one of the three top causes of cancer-related death worldwide [97, 98]. The development of CRC normally accumulate genetic mutations and progress these mutations to colon adenoma or polyp in colonic tissue cells at the first stage [99]. The various genetic factors cause the presence of heterogeneity in intestinal tissue [99, 100]. Secondly, the colon has the
most densely packed microbial systems, accommodating hundreds or even more than 1,000 different species of bacteria, involving in different regulation and metabolism pathways [100, 101, 102, 103]. The dynamic composition and location of microbiota in colorectal tissue cause complex interplay of colorectal tissue with gut microbiota [104, 105]. The bacterial composition in colonic tissues in healthy persons and patients are significantly different [106, 107]. This fact indicates that there are significant systematic heterogeneities in the colonic and CRC systems [108]. Hence, the colorectal tissue has additional leverages on the heterogeneity. Since CRC is a combined result of dysregulation of gut microbiota and genetic mutations of multiple critical genes, it is a good model for the study of heterogeneities [109, 110, 111]. In this study, we applied genetic-distance-based method and dynamic PCA bi-plot based distance to characterize the heterogeneous samples. By removing or regrouping heterogeneous samples from gene expression data sets, we found there are several newly-identified DEGs which are associated with the intestinal abnormalities. We prepared the manuscript “Distinct genetic responses of murine intestinal cancer tissues identified from existing microarray data by preprocessing inter-sample”. The experimental materials and methods, results and discussion are shown in this section.

2.6.1 Materials and Methods

2.6.1.1 Microarray Datasets

Twenty-five Microarray experimental data sets were downloaded from NCBI GEO database (Gene Expression Omnibus (GEO)). The Microarray experiment associated with these data sets were designed to examine the influence of single gene deletion in mice model with intestinal diseases or intestinal cancer. The data sets information were shown in Table 2.3. These experiments were carried out using in-situ oligonucleotide platform Affymetrix, which contain more than 35,000 probes. All of these data sets include two groups of samples, with
Table 2.3. The Intestinal microarray GEO data sets.

<table>
<thead>
<tr>
<th>GSE ID</th>
<th>knockout/knockdown gene</th>
<th>Mice Model</th>
<th>GPL</th>
<th>reference</th>
</tr>
</thead>
<tbody>
<tr>
<td>GSE18393</td>
<td>Ihh</td>
<td>VillinCre</td>
<td>GPL6246</td>
<td>2010 [112]</td>
</tr>
<tr>
<td>GSE20968</td>
<td>Hnf4α</td>
<td>VillinCre</td>
<td>GPL1261</td>
<td>2010 [113]</td>
</tr>
<tr>
<td>GSE22416</td>
<td>GATA6</td>
<td>VillinCreERT2</td>
<td>GPL6246</td>
<td>2011 [114]</td>
</tr>
<tr>
<td>GSE27868</td>
<td>P53</td>
<td>VillinCre</td>
<td>GPL6246</td>
<td>2012 [115]</td>
</tr>
<tr>
<td>GSE39624</td>
<td>Klf5</td>
<td>VillinCreERT2</td>
<td>GPL6246</td>
<td>2013 [116]</td>
</tr>
<tr>
<td>GSE81920</td>
<td>N-cad</td>
<td>VillinCre</td>
<td>GPL1261</td>
<td>2017 [85]</td>
</tr>
<tr>
<td>GSE31771</td>
<td>E-cad</td>
<td>VillinCre</td>
<td>GPL6246</td>
<td>2012 [117]</td>
</tr>
<tr>
<td>GSE50188</td>
<td>HDAC3</td>
<td>HDAC3FF</td>
<td>GPL6246</td>
<td>2013 [118]</td>
</tr>
<tr>
<td>GSE60620</td>
<td>Dnmt1</td>
<td>APCmin/+</td>
<td>GPL6246</td>
<td>2015 [119]</td>
</tr>
<tr>
<td>GSE64643</td>
<td>MSi2</td>
<td>APCmin/+</td>
<td>GPL6246</td>
<td>2015 [120]</td>
</tr>
<tr>
<td>GSE23241</td>
<td>PPARγ</td>
<td>VillinCre</td>
<td>GPL1261</td>
<td>2011 [121]</td>
</tr>
<tr>
<td>GSE41507</td>
<td>Yap</td>
<td>VillinCre</td>
<td>GPL6246</td>
<td>2013 [122]</td>
</tr>
<tr>
<td>GSE58948</td>
<td>Nod2</td>
<td>B6</td>
<td>GPL6246</td>
<td>2014 [123]</td>
</tr>
<tr>
<td>GSE46821</td>
<td>PRDM5</td>
<td>APCmin/+</td>
<td>GPL6246</td>
<td>2014 [124]</td>
</tr>
<tr>
<td>GSE765</td>
<td>CFTR</td>
<td>C57BL/6</td>
<td>GPL81</td>
<td>2004 [125]</td>
</tr>
<tr>
<td>GSE82133</td>
<td>Tgfbr2</td>
<td>CDX2P-G19Cre</td>
<td>GPL6246</td>
<td>2016 [126]</td>
</tr>
<tr>
<td>GSE15419</td>
<td>Mthfd1</td>
<td>APCmin/+</td>
<td>GPL8321</td>
<td>2010 [127]</td>
</tr>
<tr>
<td>GSE30873</td>
<td>Caspase-8</td>
<td>VillinCre</td>
<td>GPL1261</td>
<td>2011 [128]</td>
</tr>
<tr>
<td>GSE21549</td>
<td>Bcl9/Bcl9l</td>
<td>VillinCre</td>
<td>GPL1261</td>
<td>2010 [129]</td>
</tr>
<tr>
<td>GSE11759</td>
<td>Hnf4α</td>
<td>VillinCre</td>
<td>GPL1261</td>
<td>2008 [130]</td>
</tr>
<tr>
<td>GSE35481</td>
<td>Sox9</td>
<td>WntCre</td>
<td>GPL1261</td>
<td>2003 [131]</td>
</tr>
<tr>
<td>GSE49471</td>
<td>GATA4&amp;GATA6</td>
<td>VillinCre</td>
<td>GPL6246</td>
<td>2014 [132]</td>
</tr>
<tr>
<td>GSE23040</td>
<td>Hnf1α</td>
<td>EllaCre</td>
<td>GPL1261</td>
<td>2011 [133]</td>
</tr>
<tr>
<td>GSE50189</td>
<td>HDAC3</td>
<td>HDAC3FF</td>
<td>GPL6246</td>
<td>2013 [134]</td>
</tr>
<tr>
<td>GSE56337</td>
<td>Atg7</td>
<td>APCmin/+</td>
<td>GPL6246</td>
<td>2016 [135]</td>
</tr>
</tbody>
</table>

each group containing three replicates. Samples in the treatment group are from mice with knockout genes, while samples in control group are from their littermates.

2.6.1.2 Statistical Analysis and Ranked Genes

After retrieving the data sets, BioConductor was used to analyze gene expression profiles. More specifically, Affy package was used to correct the background and normalize the original data, and limma package was used to calculate the p-values of all the genes [90, 92, 91]. After
using p-value cutoff 0.01 to select genes, all the selected genes were ranked using fold-change (FC). All the genes in this list are called pre-ranked genes (PRGs).

2.6.1.3 Genetic Distance-based Strategy to Identify gpsDEGs

Based on the ranked genes, a top list of genes were selected, and the expression levels of these genes in different samples were used to calculate the distances between samples by using the following equation [80]:

\[
d_{j,j'}^2 = \frac{1}{N} \sum_{k=1}^{\infty} (E_{j,k} - E_{j',k})^2
\]

Where, \(d_{j,j'}\) is the distance between samples \(j\) and \(j'\). \(N\) is the total number of ranked genes in the calculation. \(E_{j,k}\) is the scaled expression level of the \(k\)-th gene in the \(j\)-th sample. By using the different numbers of top genes in the list, the distance curve was created to illustrate the function of top ranked genes. Then, the standard deviation (SD) within the size of nine sliding window was calculated along with the distance curve. Finally, the SD cutoff of 0.01 was used to select another set of genes, which are called genotypically and phenotypically significantly DEGs (gpsDEGs) [80]. These gpsDEGs are associated with the fluctuation of genetic distance. When the fluctuation is small, the difference between samples becomes stable.

2.6.1.4 Analysis of PCA

The different numbers of top genes from PRG lists were used for PCA analysis. The locations of the same sample were determined along with the increased numbers of top genes included in PCA analysis. The dynamic PCA bi-plot distance curve was used to determine the heterogeneous samples. After selecting a set of pre-ranked genes, the expression levels of these genes in various samples were used for PCA bi-plot analysis. The PCA bi-plot
was used to demonstrate the similarity between samples and the distribution of samples in principal component space.

### 2.6.1.5 Association with Diseases

Identified gpsDEGs were mapped to KEGG pathway database to check if the genes are associated with intestinal abnormalities and intestinal cancer [136]. In addition, the human homolog of murine gpsDEGs was also inspected in the COSMIC database to see if any mutations of this gene is associated with CRC [137].

### 2.6.2 Results

#### 2.6.2.1 Distance-based Strategy to Determine Heterogeneities

The analysis of genetic distance curve per-pair from ranked gene expression profiles is either useful to identify gpsDEGs, or critical to characterize the heterogeneous samples [80]. To analyze the data set GSE56337, both intra-sample and inter-sample genetic distance curves were calculated and presented in Figure 2.7. Clearly, the distance curves have very different patterns, especially when the number of ranked genes was less than ∼50. The values of distances in treatment group (purple curves) are generally smaller than others. The distance curves were assigned to two subgroups: (1) The distance curve of d45 (intra-sample distance between sample 4 and sample 5); (2) The distance curves of d46 and d56. The distance curve d45 is obviously lower than d46 and d56, while the d46 and d56 are close to each other. This organization of distance curves in the same group further suggests that the sample 4 and sample 5 have the similar gene expression profile and can be set into the same subgroup. However, sample 6 is relatively far from sample 4 and 5. In this case, it should be set into a different subgroup. Gray curves are from samples in the control group. These samples have the similar conditions, d12 and d23 are comparable while d13 is lower.
Figure 2.7. Distance curves of intra-group and inter-group from the GSE56337 data set. The data set contains gene expression profiles of six samples, with three samples (samples 1, 2 and 3) in control group and the other three samples (samples 4, 5, and 6) in the treatment group. Therefore, there are three intra-group distance curves in each group (gray curves-control group, purple curves-treatment group) and nine inter-group distance curves (black curves). X-axis shows the number of ranked genes used in distance calculation, and y-axis shows the distance calculated from the expression levels of ranked genes. The inset shows the same plot with x-axis in the scale of base 10 logarithm. In the inset, four subgroups of the inter-group distance curves are highlighted by five colored blocks: blue-d26 (distance between samples 2 and 6); orange-d24 and d25; gray-d16 and d36; green-d14, d15, d34, and d35.

than others, indicating samples 1 and 3 can be organized into the same subgroup, but sample 2 should be assigned in a different subgroup. In the inset figure, the classification of inter-sample distance curves between groups is highlighted using different colors. The distance curve d26 is the highest, followed by d24 and d25, which compose the second subgroup. The third subgroup includes d16 and d36 and the rest forth subgroup curves with the lowest distance curves are in the fourth group. Taken together, the sample 2 and sample 6 are more heterogeneous than the other samples in the same group.
2.6.2.2 Determine Heterogeneities Using Dynamic PCA Bi-plot

The dynamic PCA bi-plot can be used to display the location of samples in data sets. The Euclidean distance curves of any two samples were calculated and shown in Figure 2.8(A). The PCA bi-plot distance curves of intra-samples increased as the number of PRGs included into the calculation. The distance curve between sample 1 and 3 is significantly lower than those between sample 1 and 2 and sample 2 and 3 in the control group, and the distance curve between sample 4 and 5 is lower than those between sample 4 and 6 and sample 5 and 6 in the treatment group. At the beginning, the PCA bi-plot distances are very crowd and increase quickly. As the increasing number of PRGs included into the calculation, the distance curves became stable. This result indicates when the number of PRGs is small, it is not sufficient to characterize the differences between samples. It needs more ranked genes included into the analysis of heterogeneity.

The location of samples in the principal component space was also analyzed and presented in Figure 2.8(B). This figure shows the snapshots of distribution of all samples at five different numbers of ranked gens including 5, 8, 20, 95, and 166, which correspond to FC cutoffs 2.0 and 1.5, SD cutoffs 0.05, 0.01, and 0.005, respectively. Clearly, when there are only two DEGs, all the six samples are rather close to each other staying in the upper right quadrant. When the number of ranked genes reaches 95, all the samples in the treatment group move to the left, but all the samples in the control group move upward. When there are 166 ranked genes, samples in the treatment group move further up-left and are now in the middle of the left side of the plot, samples in the control group although moving left are still in the middle of the right half of the plot. Comparing the dynamic PCA distance curves between samples in the same group, it can also be seen the sample 2 in the control group is far away from the rest of two samples in the control group, and the sample 6 is generally far away from the
rest two samples in the treatment group, especially when there is small number of ranked genes.

An additional observation of Figure 2.8 is that when the number of ranked genes is small, the actual differences between samples may not be properly characterized. Therefore, to measure the differences between samples sufficiently, a larger number of ranked genes needs to be used. For this reason, FC cutoff 2.0, FC cutoff 1.5, and SD cutoff 0.05 may not be suitable standards to select DEGs. Actually, as demonstrated in previous published paper on Microarray data of eight different types of cancers, SD cutoff 0.01 is appropriate in the most cases for DEGs analysis. Based on the analysis of inter-sample genetic distance curves, dynamic PCA distance curves, and the distribution of samples in PCA bi-plot, it is clear that the sample 2 in control group and the sample 6 in treatment group are heterogeneous samples. The next question is whether and how those heterogeneous samples affect the identification of DEGs.

### 2.6.2.3 Regroup/Remove Heterogeneities to Identify gpsDEGs

To measure the impacts of heterogeneous samples on identification of gpsDEGs, we repeated the process of identifying gpsDEGs by removing/regrouping sample 2 and sample 6. The compared results were shown in Figure 2.9(A). The reason for choosing these two samples is that each of them is more heterogeneous than the others in each group. As shown is Figure 2.9(A), the averaged intra-group distance under group method 123-456 is the highest. This group method is actually the original group method of samples, which did not consider any pre-processing of sample heterogeneity. In addition, removing/regrouping either sample 2 or sample 6 can reduce the intra-group distance significantly. Nonetheless, removing both sample 2 and 6 (13-45) did not get the lowest averaged intra-group distance. However, group method 123-45 or 13-456 yielded the lowest averaged intra-distance. Figure 2.9(B) shows the inter-group distance curves. Clearly, removing/regrouping heterogeneous
Figure 2.8. PCA analysis of heterogeneous sample. (A) The intra PCA bi-plot distance curves for any two samples in control group (gray) or in treatment group (blue). X-axis displays the number of PRGs included in analysis and y-axis represents the dynamic PCA bi-plot distances between any two samples in the same group. (B) The distribution of samples in PCA bi-plot. The distances of six samples in the GSE56337 data set were calculated using the following numbers of ranked genes (N): 5, 8, 20, 95, 166. These numbers correspond to FC cutoff values 2 and 1.5, SD cutoff values 0.05, 0.01, and 0.005, respectively (see Method). X-axis is the first principal component, while y-axis shows the second principal component. Symbols in the same rectangle represent three samples in the control group (circle-sample 1, square-sample 2, star-sample 3), while symbols in the same eclipse represent three replicates in the treatment group (up triangle-sample 4, down triangle-sample 5, diamond-sample 6). Symbols with the same color indicate that their distributions in the PCA bi-plot are analyzed using the same number of ranked genes.
sample(s) reduce(s) the inter-group distances. However, the group method 13-45 still has rather high distance, especially when the number of ranked genes is large. The group method 123-45 is the only method leading to a lower distance. This is an expected result of the data shown in Figure 2.9(A). Figure 2.9(C) shows the SD values as a functions of the number of ranked genes under different removing/regrouping methods. In this figure, group method 13-45 has the lowest value of SD among all the other methods when the number of ranked genes is less than 10. Anyhow, when there are more ranked genes, the SD curves of all the methods converge quickly. Figure 2.9(D) presents the comparison of identified gpsDEGs before/after removing/regrouping samples 2 and 6. Clearly, removing/regrouping samples influence the lists of gpsDEGs significantly. When either sample 2 or sample 6 is regrouped from the original groups and set into a new group (13-2-456 and 123-45-6), each new list of gpsDEGs has \( \sim 70\% \) overlapped genes, \( \sim 30\% \) missed genes, and \( \sim 30\% \) newly-identified genes compared to the original list of gpsDEGs. Anyhow, if taking both samples 2 and 6 from their original groups and assigning them into two different groups (13-2-45-6), the number of overlapped genes became very low, but the number of missed and newly-identified genes were higher. When removing either sample 2 or sample 6 from the analysis (13-456 and 123-45), the number of overlapped genes are less than those in the first two situations, but the number of missed genes and newly-identified genes increased. If both samples 2 and 6 are removed from the analysis (13-45), the number of overlapped genes, missed genes and newly-identified genes became comparable to that in the first two situations.

### 2.6.2.4 Mapping Identified gpsDEGs into Pathways

Since removing/regrouping heterogeneous sample(s) has substantially changed the list of identified gpsDEGs, it becomes critical to check if the functions of gpsDEGs in different sets are the same. For this reason, the gpsDEGs associated KEGG pathways before/after removing/regrouping sample(s) were analyzed and presented in Figure 2.10. As shown in
Figure 2.9. Impact of removing/regrouping heterogeneous samples. In the original grouping method, samples 1, 2, and 3 were set in the control group, and samples 4, 5, and 6 were put in the treatment group, this grouping method is named after 123-456. Since samples 2 and 6 have larger distances from other samples in the sample group, any of them or both of them can be removed from the original group and/or put into new group(s). By notation, group method 13-2-456 indicates that sample 2 is removed from the control group and is put into a different group, 13-456 means that sample 2 is removed from the analysis resulting only five samples in the analysis, and so on so forth. After removing/regrouping sample(s), the corresponding new groups and samples can be used to identify ranked genes. The expression levels of the ranked genes can then be used to calculate (A) intra-group distances, (B) inter-group distances, and (C) SD values of inter-group distances (see Method). Panel (D) shows the change of identified gpsDEGs. The original grouping method 123-456 was set as the comparison method, and the gpsDEGs from this method were used as the comparison set of genes. The lists of gpsDEGs from all other regrouping methods, which were ranked along x-axis, were each compared to the comparison set of genes. All the genes overlapped in both lists are called “overlapped” (black bars), all the genes only contained in the comparison set of genes are named after “missed” (dark gray bars), and all the genes only included in the new list of gpsDEGs are addressed as “newly-identified” (gray bars).
Figure 2.10(A), all regrouping methods except 13-2-45-6 produced large overlap with group method 123-456 in the first ∼20 KEGG pathways. These pathways are mainly related to microbial infection and immune reaction. In the rest ∼30 KEGG pathways in which group method 123-456 identified only one or two gpsDEGs, other grouping methods can also reproduce the similar amount of gpsDEGs in the same pathway in most majority of the cases. Nonetheless, various regrouping methods also generated another ∼40 pathways, which are not identified by group method 123-456 as shown in Figure 2.10(B). These pathways may overlap between different regrouping methods, or be specific to a regrouping method. Since red and blue bars stand for pathways associated with intestinal diseases in either KEGG or COSMIC databases, it can be concluded that almost all the pathways in Figure 2.10(B) are associated with intestinal abnormalities.

Figure 2.11(A) presents another example where all the gpsDEGs were identified after removing/regrouping heterogeneous sample(s). The original grouping of samples w/o removing/regrouping heterogeneous samples (123-456) did not identify any gpsDEGs associated with the PI3K-Akt signaling pathway. After removing/regrouping heterogeneous samples, seven genes associated with this pathway were identified as gpsDEGs and shown in Figure 2.11(B). CD19, CD21, and FceR1a were up-regulated to activate the PI3K-Akt pathway. In the FceRI signaling pathway, MAP2K4 was up-regulated to activate multiple immune-response associated genes. In the HIF-1 signaling pathway, upon the activation of PI3K-Akt pathway, HIF1α will be activated, and then GADPH should be up-regulated. However, in the GSE56337 dataset, GADPH was down-regulated by −1.18. The reason is that HIF1α is also inhibited by PHD, which was upregulated by 0.65 in the experiment. In the FoxO signaling pathway, FoxO is inhibited by the activation of PI3K-Akt pathway, and therefore the downstream gene CCNB1 was also down-regulated.
Figure 2.10. Summary of KEGG pathways identified before and after removing/regrouping heterogeneous samples. The first column on the left shows the names of KEGG pathways, the first row shows different grouping methods that correspond to removing/regrouping of specific sample(s). Pathways highlighted in red are associated with intestinal diseases as annotated in the KEGG database. Pathways highlighted in blue indicate that the mutations on identified gpsDEGs in these pathways are related to intestine cancer as documented in the COSMIC database. Gray bars show the number of genes identified in each pathway using a specific grouping method. Solid-edged boxes over the gray bars indicate the number of genes overlapped with genes identified without removing heterogeneous samples (123-456).
Figure 2.11. Distinct bypaths in KEGG pathways identified by removing/regrouping heterogeneous samples. (A) Pathways related to RNA processing, and (B) pathways related to antigen reaction. Gray symbols are genes in the pathway but not identified as gpsDEGs by any method in this study. Blue symbols represent genes identified as gpsDEGs without removing/regrouping heterogeneous samples (123-456). Orange symbols denote identified gpsDEGs after removing/regrouping samples 2 and/or 6. The numbers next to blue/orange genes are the log(FC) values w/o removing heterogeneity (above the slash), and the log(FC) values after removing/regrouping heterogeneous samples (below the slash). For the simplicity of the figure, the information of exact grouping method of samples is not shown in this figure. In the case when a number is replaced by “∼”, the gene was not identified as a gpsDEG under that specific situation. Solid lines indicate direct interaction or regulation. Dashed lines indicate that DNA transcription is involved or multiple genes are along the path. Arrow head represents upregulation of the function, while oval arrow head indicates inhibition of the function. The solid-line rectangle contains the names of down-stream genes, while the dashed-line rectangles explain the functional roles of corresponding genes. The each rounded dashed-line rectangle shows a specific KEGG pathway.
2.6.2.5 Extended Analysis on Twenty-five GEO Datasets

We further extended the regrouping/removing method to twenty-five GEO datasets, the result shows a larger pool of gpsDEGs can be identified using regrouping/removing heterogeneous samples. Figure 2.12(A) illustrates the fraction of gpsDEGs changed after removing/regrouping heterogeneous samples. After regrouping or removing intra-sample(s), the gpsDEG are significantly changed. There are \( \sim 20\% \) new gpsDEGs newly identified and \( \sim 20\% \) gpsDEGs missing compared with the original experimental setup. If regrouping samples into four different groups, \( \sim 80\% \) new gpsDEGs can be identified, and the same percentage of gpsDEGs are missing.

Since removing/regrouping heterogeneous sample(s) has substantially changed the list of identified sets of gpsDGEs, we are so curious about whether the changed sets of gpsDEGs can suggest biological phenomenon. For this reason, the KEGG pathways associated with gpsDEGs before and after removing/regrouping sample(s) were analyzed and presented in Figure 2.12(B). After removing/regrouping sample(s), the newly identified gpsDEGs were analyzed together with KEGG pathways. The original group setup method was used as a control. The independent paired t-test was applied to compare a newly identified gpsDEGs and original gpsDEGs in the same pathway. From all twenty-five data sets, we found several pathways have significantly changed in more than eight GEO data sets. Figure 2.12(B) shows the pathways in which gpsDEGs are significantly changed. These pathways are highly associated with RNA processing and cancer developments.

2.6.3 Discussion

Tissue heterogeneity is a common phenomenon in molecular biology studies. In Microarray data analysis, PCA bi-plot based on identified gpsDEGs was frequently used to examine the heterogeneity between samples [138]. While the PCA analysis is successfully
Figure 2.12. Comparison of gpsDEGs before/after regrouping or removing heterogeneous samples. (A) Summary of gpsDEGs comparison after regrouping or removing intra-samples in twenty-five datasets. This figure shows the change of identified gpsDEGs. The original grouping method 123-456 was set as the comparison method, and the gpsDEGs from this method were used as the comparison set of genes. Regrouping/removing sample in control group is the sample 3 and in treatment group is the sample 6. X-axis represents the regrouping or removing methods, and y-axis represents the average fraction of gpsDEGs changed in different group methods. The fraction of all the genes overlapped in both lists are called “overlap_aver” (black star bars), the fraction of all the genes only contained in the comparison set of genes are named after “original_aver” (dark gray square bars), and all the genes only included in the new list of gpsDEGs are addressed as “new_aver” (black bars). (B) gpsDEGs associated pathways after regrouping or removing intra-heterogeneous samples. This figure represent the summary of changed gpsDEGs in different pathways through twenty-five pathways. Y-axis represents the name of pathways frequently changed through twenty-five datasets, along with x-axis is the number of GEO datasets significantly changed in that pathway.
and prevailing, several issues need to take into consideration when using PCA analysis because identification of gpsDEGs is challenged by effects of batch, dye, and so on. There are multiple methods using for identifying gpsDEGs. The most common procedure is to find the p-value based statistical model, selecting genes of which p-values are below a specific cutoff and FC are up a specific cutoff [10]. Clearly, p-value calculation may be influenced by many factors, and the selection of p-value and FC cutoffs may not be well justified. In most case, the cutoff p-value 0.01 with FC cutoffs 1.5 or 2.0 were used to select DEGs. The connection of these methods with biology is not well established yet. In addition, the weight of principal components and the distance in PCA bi-plot are not easy to interpret. Based on all of the reasons, it needs additional considerations to characterize heterogeneous samples when using PCA methods.

In our previous study on using genetic distance-based method to identify gpsDEGs, GEO Microarray data sets from eight different cell lines were analyzed. The advantage of using cell line based Microarray data is that the heterogeneity within cells of the same cell line should be marginal when identifying gpsDEGs. In that study, we noticed that the inter-group distances of most of the microarray data sets were less than 0.6 when the SD values became stable. For this reason, 0.6 can be used as an empirical threshold value to determine low-heterogeneity samples. In Figure 2.7, the sample 4 and 5 has the lowest intra-sample distance, this result indicated the sample 6 is more heterogeneous and should be regrouped or removed. By the same method for control group, the sample 2 can be regrouped and removed. In Figure 2.9, regroup 13-45, 123-45, 13-456 have low inter-sample heterogeneity, indicating removing/regrouping sample(s) in these three ways is able to remove heterogeneity efficiently. Consequently, the regrouped samples under these conditions can be used to identify gpsDEGs.

Even after removing/regrouping heterogeneous sample(s), the identified gpsDEGs using different regrouped sample(s) are not identical as shown in Figure 2.9, indicating there are
still “residue” heterogeneity in each group and influence the identification of gpsDEGs. Under this situation, since heterogeneity cannot be removed completely, an alternative strategy in DEG analysis is to combine all the gpsDEGs identified from different regrouping methods. In different GEO Microarray data sets, by combining newly-identified gpsDEGs using removing/regrouping methods, multiple critical DEGs have been identified and related with eight pathways, such as pathway in cancer, herpes simplex infection, HTLV-I infection, chemical carcinogenesis, Tueroulosis, protein digestion and absorption, Phagosome, and PI3K-Akt signaling pathways [139, 140, 141, 142, 143, 144, 145]. Two examples from different GEO data set GSE56337 were shown in Figure 2.11(A) and (B). Genetic distance-based method without removing heterogeneous samples identified dispersed DEGs in pathways, between which the regulation may not be rationalized easily. However, after removing/regrouping heterogeneous samples, additional gpsDEGs were identified, and the regulation among these genes can be explained well. For example, PI3K-Akt pathway plays critical roles in many processes related to intestine tumorigenesis and microbial infection [139]. However, gpsDEGs identified before removing/regrouping heterogeneous samples were not associated with these pathways. After removing heterogeneous samples, the gpsDEGs and their expression levels explained the regulatory process of this pathway very well.

While lacking a “golden standard” to evaluate the quality of identified gpsDEGs, the connections between genes and diseases were extracted from mutation-disease databases in this project to estimate the correlation between gpsDEGs and specific diseases. Although there are both pros and cons, the data in this study indicates that the information in the mutation-disease databases is helpful in identifying the functional roles of identified DEGs. For this reason, it is possible to use the information in the mutation-disease databases to further filter the identified DEGs in similar studies.

Genetic-distance-based method on the expression levels and dynamic PCA distance curves of pre-ranked genes between samples is able to characterize heterogeneous samples.
Removing/regrouping heterogeneous samples leads to the identification of new sets of gps-DEGs. Many of these newly-identified gpsDEGs are associated with phenotypic conditions of the samples. The combination of all the sets of gpsDEGs provides not only a large pool of DEGs, but also clear picture on the regulation of genes with signaling pathways.

2.7 Testing Distance-based Method on Microarray Quality Control Data Sets

In this chapter, we introduced the genetic distance-based method to determine genotypically and phenotypically significantly differentially expression genes (gpsDEGs) [80]. By analyzing eight different data sets, the genetic distance-based method is more stable and comparable to identify gpsDEGs among different Microarray data sets. The non-stringent p-value 0.1 and standard deviation (SD) 0.01 were used to identify gpsDEGs. The dynamic PCA bi-plot distance curves were then introduced to identify heterogeneous samples among replicates. This method overcome the issue caused by the current method, that is how many DEGs needs to be used to determine the heterogeneities, because the dynamic PCA bi-plot distance curves include continuous numbers of pre-ranked genes into the calculation and determine when the trends of PCA bi-plot distance curves became stable. To deal with the heterogeneous samples, we tested the impacts of the heterogeneous samples on the identification of gpsDEGs in Microarray data sets of intestinal disease. By removing or regrouping heterogeneous samples, there were a number of new gpsDEGs identified, which were closely associated with the development of the intestinal abnormalities.

In this section, we further optimized the dependent factors, p-value and SD value, of genetic-distance-based method to maximize the efficiency of Microarray data analysis. After that, we measured the impacts of heterogeneous samples on identification of gpsDEGs by removing or regrouping heterogeneous samples. The data sets from Microarray Quality Control (MAQC) project were used as the standard data sets to evaluate those methods [11, 146, 147, 148]. We chose the data set from MAQC project due to two reasons. First, different
platforms were used for a batch of Microarray data sets in this project. It provides a various of parallel experimental results. Secondly, these data sets have the experimental validated differentially expressed genes (DEGs), which can be used to validate DEGs identified using different statistical methods and genetic-distance based method. The experimental methods and results were shown in following.

2.7.1 Materials and Methods

2.7.1.1 Data Pre-processing

The Microarray Quality Control (MAQC) project data set GSE5350 [11, 146, 147, 148] was extracted from NCBI Gene Expression Omnibus (GEO) database. The samples from site1 using three different platforms, Affymetrix (AFX), Agilent one-color (AG) and Illumina (ILM) were used. Sample A in MAQC data set was stratene Universal Human reference RNA, sample B was the Ambion Human Brain reference RNA, sample C was the mixture of sample A and sample B with the ratio 3:1, and the sample D was the mixture sample A and sample B with the ratio 1:3. The DEGs were from the comparison of sample A and sample B, and were validated by TaqMan, the real-time Polymerase Chain Reaction (PCR) assay. The gene list from TaqMan assay was used as the validated set. Different preferred normalized methods were applied on these platforms. For instance, quantile normalization was used for platform AFX and ILM, and median-scaling was used for platform AG.

2.7.1.2 BioConductor and DEG Analysis

`Limma` package in R was used to calculate p-value and fold-change (FC) value of all genes in one expression profile [91, 92]. Then, different p-values were applied to select different preliminary lists of genes. These p-values are 0.0001, 0.001, 0.01, and 0.1. All of the lists
were ranked by FC as suggested from Benchmark study. Clearly, different sets of DEGs can be selected using different FC values in the preliminary ranked genes (PRGs) list.

2.7.1.3 Genetic Distance-based Strategy to Identify gpsDEGs

Based on PRGs lists, the top lists of genes were selected to calculate the genetic-distances between any two samples [80]. The formula is as follow:

\[
d_{j_1,j_2} = \frac{1}{N} \sqrt{\sum_{k=1}^{N} (E_{j_1,k} - E_{j_2,k})^2}
\]

Where, the \(d_{j_1,j_2}\) is the distance between samples \(j_1\) and \(j_2\). \(N\) is the number of top genes included in the calculated list. \(E_{j_1,k}\) is the scaled expression level of the \(k\)-th gene in the \(j_1\)-th samples and \(E_{j_2,k}\) is the scaled expression level of the \(k\)-th gene in the \(j_2\)-th samples.

The SD within the sliding window of size nine along with the distance changes was calculated. Each SD cutoff value, 0.001, 0.002, 0.003, 0.004, 0.005, 0.007, 0.01, 0.02, 0.03, 0.04, and 0.05, was applied to select a set of the gpsDEGs. These gpsDEGs were associated with the fluctuation of the genetic distances. Which means, the differences between samples became stable as the fluctuation of the genetic-distance curves was getting small.

2.7.1.4 Regroup/Remove Heterogeneities from Intra-groups

The determined heterogeneous samples were removed or regrouped accordingly. The methods used to deal heterogeneous samples can be set into four categories: (1) removing all heterogeneous samples; (2) regrouping all heterogeneous samples; (3) removing heterogeneous samples in each group individually; (4) regrouping heterogeneous samples in each group individually. The genetic distance-based method was applied on the refined data sets to identify gpsDEGs. These gpsDEGs were validated by comparing with TaqMan validated DEGs.
2.7.2 Results

2.7.2.1 Differences of Intra-samples Replicates

The Microarray data sets from MAQC project are generated using different platforms. In each data set, it usually has four sample groups, and each group has five replicates. Before using these Microarray data sets, we evaluated the quality and similarity of the expression levels of the replicates in the same sample. The standard deviation of each gene expression level from different replicates in the same sample was calculated and shown in Figure 2.13. Figure 2.13 (A)(B)(C)(D) shows the results of sample A, B, C, and D from platform AFX, respectively. The averaged scaled gene expression levels of replicates in the same group were normally distributed from 2 to 14. Comparing the standard deviation of gene expression levels among replicates with the averaged scaled gene expression level, the differences among the replicates in the same sample in platform AFX were not significantly. In the figures of AG platform (Figure 2.13 (E)(F)(G)(H)), the ranges of standard deviations of each sample were normally lower than 0.8. However, the SD of several genes were higher than 1.0. The averaged scaled expression levels of those genes were normally from 6 to 10. The expression levels of those genes from different replicates in the same sample were significantly different. However, the number of those genes was very small portion compared to the total numbers of genes. This result indicates the differences among replicates in the same sample in platform AG were not significant. In Figure 2.13 (I)(J)(K)(L) from platform ILM, the standard deviations of the gene expression levels among different replicates in each sample were not significantly different. Usually, the standard deviation values were lower than 0.8. Those results indicate the expression levels of replicates in the same sample were not significantly different in platform ILM.

The Pearson Correlation Coefficients (PCC) and the student t-test were calculated for each two replicates in the same sample from different platforms and shown in Figure 2.14 (A)
Figure 2.13. The standard deviation of replicates’ expression levels of samples from Affymetrix (upper), Agilent (middle), and Illumina (bottom) in site 1. X-axis displays the averaged expression levels of each gene, and y-axis displays the standard deviation of replicates’ expression level of each gene in the same sample. (A)-(D) display the results of sample A-D from Affymetrix, (E)-(H) display the results of sample A-D from Agilent, and (I)-(L) display the results of sample A-D from Illumina.
and (B), respectively. In Figure 2.14(A), the results of PCCs of any two replicates were all higher than \( \sim 0.97 \). These values are high enough to show the closely correlation coefficient between every two replicates in that sample. Figure 2.14(B) shows the student t-test results of any two replicates in the same sample. Although the highest t-score reached \( \sim 2.2 \), there was no significant differences between any two replicates in that sample because the degree of freedom was very large.

2.7.2.2 Determine the Dependent Factors: p-value and Standard Deviation (SD)

From the analysis of differences of intra-sample replicates, it demonstrated that there was no significant differences between replicates in the same sample. We formed the new expression data sets by randomly selecting two replicates from sample A and one replicate from sample C to form a new group A and similarly two replicates from sample B and one replicate from sample D to form a new group B. Since the sample C is the mixture of sample A and B in the ratio of 3:1 and the sample D is the mixture of sample A and B in the ratio of 1:3, the replicate from sample C in group A and the replicate from sample D in group B can be treated as heterogeneous samples.

At first, the dependent factors, p-value and SD value, from each newly formed data set were determined. Figure 2.15 shows the percentages of validated DEGs selected using genetic-distance-based method SD value (left) and FC (right). In AFX platform shown in Figure 2.15(A) and (B), when applying SD value to determine DEGs, the p-value cutoff 0.1 provided the highest validated DEGs percentage curve. Usually, the trends of true percentage DEG curves tend to be stable when SD cutoff is 0.01. Therefore, the best cutoff of SD value was 0.01. These values are supported by our previous publication [80]. When applying FC to determine DEGs, the validated DEG percentage curves using p-value 0.1 and 0.01 were very close to each other and higher than other curves. Therefore, the p-value used to select DEGs
Figure 2.14. The variances of replicates’ expression levels of samples from Affymetrix (blue), Agilent (red), and Illumina (yellow) using (A) Pearson Correlation Coefficient, and (B) Student t-test. X-axis shows the platform name and the sample name, and y-axis shows the values of PCC (A) and student t-test (B).
can be both 0.1 and 0.01. However, no matter which FC and p-value used to determine DEGs, the validated DEG percentages using FC method were significantly lower than those from genetic-distance-based method. In the platform AG and ILM, the similar results were observed. The true percentages of DEGs determined using genetic-distance-based method SD value were normally higher than FC method. For the genetic-distance-based method, the best p-value and SD values used to select gpsDEGs were 0.1 and 0.01, respectively. For the FC method, the best p-value and FC value were 0.01 and 2.0. Those results indicate the genetic-distance-based method SD value used to identify DEGs can provide higher true percentages of DEGs. All the procedures were repeated three times by randomly selected replicates from samples and the results were shown in Figure 2.16.

2.7.2.3 Removing or Regrouping Heterogeneous Samples

The sample C in group A and sample D in group B were treated as heterogeneous samples. The heterogeneous samples were removed or regrouped to form the new data sets, and the averaged intra-distance and inter-distance of the new formed data sets were calculated. Figure 2.17 (A) and (B) show the results of AFX platform. After removing or regrouping the heterogeneous samples, the averaged intra-distances were reduced significantly compared to the original group method, and the averaged inter-distances after removing the heterogeneous samples were kept similarly with the original group method. However, the averaged inter-distances were reduced by regrouping the heterogeneous samples. The similar results were observed in platform AG (Figure 2.17 (C) and (D)) and ILM (Figure 2.17 (E) and (F)).

2.7.2.4 Identification of New gpsDEGs After Removing or Regrouping Heterogeneous Samples

Figure 2.18 shows the changes of gpsDEGs after removing or regrouping the heterogeneous samples. In platform AFX shown in Figure 2.18 (A), after removing or regrouping
Figure 2.15. The percentage of validated DEGs identified under different p-values using genetic-distance-based method (left) and fold change (FC) values (right) in new generated data sets. (A) and (B) show the results of AFX, (C) and (D) show the results of AG, and (E) and (F) display the results of ILM. X-axis in left penal represents the different standard deviation used to select gpsDEGs, while in right penal displays the different FC cutoff values used to select DEGs. Y-axis shows the percentage of true DEGs compared with experimental validated data sets.
Figure 2.16. The percentage of validated DEGs identified under different p-values using distance-based method and fold change (FC) values in new generated data sets. (A1)-(A6) show the results of Affymetrix, (B1)-(B6) show the results of Agilent, and (C1)-(C6) display the results of Illumina. X-axis represents the different standard deviation used to select gpsDEGs (row 1, 3, and 5), and the different FC cutoff values used to select DEGs (row 2, 4, and 6). Y-axis shows the percentage of true DEGs compared with experimental validated data sets.
Figure 2.17. The averaged intra- and inter-distance curves before and after removing or regrouping heterogeneous samples in new generated data sets. X-axis displays the number of pre-ranked genes, which is calculated using p-value 0.1 and ranking by FC. Y-axis displays the intra-distances (A)(C)(E) and inter-distances (B)(D)(F). (A) and (B) display the results of AFX, (C) and (D) display the results of AG, and (E) and (F) show the results of ILM. The solid black lines display the intra- or inter-distances before removing or regrouping heterogeneous samples, and the different dash lines and grey solid lines display the intra- or inter-distance curves after removing or regrouping heterogeneous samples.
heterogeneous samples, there were many new gpsDEGs identified. By removing the heterogeneous samples, there were a batch of gpsDEGs identified, in which, ~10% gpsDEGs were validated by experiments. Compared with the original group method, only several gpsDEGs were missing. Additionally, by regrouping heterogeneous samples, there were also a batch of new gpsDEGs identified, among which, ~5% gpsDEGs were validated by experiments. However, there were ~10% validated gpsDEGs were missing after regrouping the heterogeneous samples. Similar results were also observed in platform AG and ILM shown in Figure 2.18 (B) and (C), respectively. These results indicate removing the heterogeneous samples from the data sets can identify more gpsDEGs, which were closely associated with the biological phenomenon. In addition, although regrouping the heterogeneous samples can provide additional validated gpsDEGs, many validated gpsDEGs were missing. The regrouping method needs to be further studied.

2.7.3 Discussion

The differences and the quality of Microarray expression profiles from three different platforms, Affymetrix (AFX), Agilent (AG), and Illumina (ILM) were analyzed. In the same expression data set, the standard deviations of replicates’ expression levels in the same sample of each gene were not significantly different. This indicates the differences among replicates were not significantly different. The analysis of the differences between any two replicates using Pearson Correlation Coefficient and student t-test shows that each two replicates in the same sample were not significantly different shown in Figure 2.14.

The dependent factors, p-value and SD values, are critical for the efficiency of genetic-distance-based method. It is necessary to find the best p-value and SD values for each data set. After forming the new data sets, p-value 0.1, 0.01, 0.001 and 0.0001 were used to generate pre-ranked gene lists. A series of SD values and FC values were used to select DEGs. The percentages of validated DEGs selected using SD values were significantly higher
Figure 2.18. The DEGs identified by regrouping/removing heterogeneous samples compared with original group method in new generated data sets. (A) shows the result of AFX, (B) shows the results of AG, and (C) shows the result of ILM. The blue bar displays the number of overlapped gpsDEGs before and after removing/regroups heterogeneous samples, and the light blue bar inside black bar display the number of overlapped gpsDEGs that are validated by experiments. The red bar displays the number of gpsDEGs only identified before removing/regrouping heterogeneous samples and the light red bar displays the number of gpsDEGs which were validated by experiments. The yellow bar displays the number of gpsDEGs only identified after removing/regrouping the heterogeneous samples. X-axis displays the group method, and y-axis displays the number of gpsDEG.
than those selected using FC values. We also found the best SD value to select gpsDEGs was 0.01, which was the same as our previous published paper [80]. These results suggest the genetic-distance-based method can be used to identify gpsDEGs, which are more associated with biological phenomenon.

The heterogeneous samples were regrouped or removed to determine the impacts of heterogeneous samples on the determination of gpsDEGs. From Figure 2.17, the intra-distances were reduced significantly after removing the heterogeneous samples, while the inter-distances were kept similar compared to the distances without removing the heterogeneous samples. This result suggests after removing the heterogeneous samples, the genetic differences within the same group can be reduced while the differences between groups were kept the same. In addition, after regrouping the heterogeneous samples, the intra-distances and inter-distances were both reduced. These results indicate regrouping the heterogeneous samples also can reduce the genetic differences within the group. However, the inter-distances were reduced after regrouping the heterogeneous samples. This observation indicates the regrouping heterogeneous sample method needs to be further considered.

The gpsDEGs before and after removing/regrouping heterogeneous samples are shown in Figure 2.18. The result shows after removing the heterogeneous samples, there were additional gpsDEGs can be identified, in which, there were \(\sim 10\%\) gpsDEGs can be validated by experiments. Therefore, removing the heterogeneous samples can provide a large pool of gpsDEGs which were associated with biological regulation.

### 2.8 Conclusion

We designed the genetic-distance-based method to identify genotypically and phenotypically significantly differentially expression genes (gpsDEGs), and found the number of gpsDEGs identified through this method is more stable and comparable among different data sets with the relatively lower false discovery rate in eight different cancer cell lines.
Normally, the number of gpsDEGs identified by this method are limited in the range of 100 to 200. The number of gpsDEGs are more compromised and can bring multiple potential advantages in subsequent analysis.

In addition, the heterogeneous samples may affect the identification of gpsDEGs. Therefore, in our study, we introduced the dynamic PCA bi-plot distance curve to identify heterogeneities. The genetic-distance-based method integrated with removing or regrouping heterogeneous samples were applied to measure the impacts of heterogeneous samples on the determination of DEGs. In twenty-five intestinal diseases Microarray data sets, we found that removing or regrouping the heterogeneous samples provided a large pool of DEGs which were associated with the intestinal abnormalities and a clear picture on the regulation of genes in signaling pathways.

Furthermore, the Microarray data sets from MAQC project were used to optimize the dependent factors, p-value and SD, in genetic distance-based method. From the analysis, we found the most suitable p-value and SD value used to select gpsDEGs were p-value 0.1 and SD value 0.01, which support the results in our previous publication [80]. Additionally, the MAQC data sets were reformed to test whether removing or regrouping heterogeneous samples can find more gpsDEGs. By aligning the gpsDEGs with the experimental validated DEGs, we found removing the heterogeneous samples provide additional DEGs, which can be validated by experiment.
CHAPTER 3

IMPROVING PERFORMANCE OF MICRORNA AND MICRORNA TARGET PREDICTION

3.1 Note to Reader


3.2 Contributions and Acknowledgements

I would like to thank my adviser Dr. Bin Xue for his kind contributions in completion of these works. These works were supported by the start-up funding from the Department of
Cell Biology, Microbiology and Molecular Biology, College of Arts and Sciences at the University of South Florida to Dr. Bin Xue. I also would like to gratefully acknowledge the application of microRNA target predictors, miRanda, miRDB, PITA, TargetScan, RNA22, and ComiR, and disorder predictors, DisEMBL, IUPred, VSL2, ESpritz, PONDR-FIT, MFDp2, IUPred2A, AUCpreD, and DISOpred3 used in these publications.

3.3 Introduction

microRNAs are short non-coding RNA molecules with about 22 nucleotides [20]. Comparing to mRNA, which has around hundreds to thousands of nucleotides, microRNAs are very short [20]. The microRNA normally binds with the 3’UTR of mRNAs to inhibit the translation of mRNA [21, 22]. There are more than 2000 microRNAs in each mammalian genome, regulating about 60% of all the genes in the genome [21, 22]. Clearly, microRNAs are a critical family of gene expression regulators. However, a thorough understanding on function and mechanism of microRNAs is still elusive. The curated data sets of microRNA:mRNA interactions are very helpful to facilitate the studies of microRNA molecular biology. Nonetheless, microRNA:mRNA interaction databases are inconsistent from different sources [43, 44, 149]. Therefore, it is necessary to establish a new and comprehensive microRNA:mRNA interaction database, which can be used for the study of microRNA:mRNA regulation and to develop the microRNA:mRNA interaction predictors.

There are several well-designed large-scale microRNA:mRNA interaction databases, such as DIANA-TarBase [44], miR2Disease [150], miRTarBase [43], miRWalk [151], miRecords [152] and etc.. DIANA-TarBase can provide hundreds of thousands of high quality experimentally validated microRNA:mRNA interaction, and easily for users to identify positive or negative experiment results and search the experimental methodology and experimental conditions [44]. miR2Disease is a curated database, which provides the comprehensive resources of microRNA deregulation of different human diseases, including microRNA names,
disease types, and the relationships between microRNA:mRNA in diseases [152]. miRTarbase contains the experimentally validated microRNA:mRNA interactions [43]. The latest updated database includes the systematically Argonaute-microRNA:mRNA interaction from crosslinking and immunoprecipitation sequencing (CLIP-seq) dataset from 21 different independent studies [43]. miRWalk and miRecords provide not only experimentally validated microRNA:mRNA pairs in database, but also the predicted results of microRNA:mRNA pairs [151, 152].

In addition to study the potential correlations inside different types of data, machine-learning based methods are more and more prevailing and powerful [25, 47, 153, 154, 155]. Many computational tools have been developed to identify microRNA and microRNA binding targets [23, 24, 26, 25]. These tools are generally based on two categories of strategies: (1) Base pair of seed regions and 3’UTRs as well as their characters, including various sequential, structural, interaction, and evolutionary information [18, 156, 157], and (2) Machine-learning and deep-learning methods, such as hidden markov chain [24], support vector machine [158, 159], and regression models [160]. These techniques have improved the processes of microRNA and microRNA target genes prediction.

Normally, the machine-learning based methods are generally built by a specific procedure. Firstly, selected true samples and false samples and established the train and test sets. Secondly, chosen a suitable machine-learning based technique and develop the infrastructure to make a predictor. Thirdly, selected the input features and fed the features into the predictor. At last, optimized and validated the predictor on the independent test sets. Currently, there are many predictors which have been developed using this procedure.

MiRanda, MiRDB, PITA and TargetScan are four very popular predictors for microRNA and target mRNA interactions [18, 23, 26, 42, 153]. Those four predictors were designed using machine-learning based methods, and based on different sequential properties and algorithms. For instances, the original version of miRanda predicts microRNA targets based
on sequence match, free energy calculated from Vienna RNA package, and evolutionary conservation verified from sequence alignment [18]. Nonetheless, the newest version of miRanda has integrated an mirSVR score into the output prediction [18]. The mirSVR score is the output of “a support vector regression approach to model the degree of microRNA regulation given a set of numerical features representing the microRNA binding site and additional contextual information” [18]. The mirSVR score is actually a measure of the changes of logarithm-based expression levels of down-regulated mRNAs upon microRNA transfection [18]. The results in miRDB were predicted using mirTarget2, a support vector machine based predictor [19, 153]. The final score of miRDB ranges from 0 to 100 and shows the relative significance of predicted target genes [153, 159]. PITA predicts microRNA targets based on sequence match between mRNA and microRNA calculated from RNAduplex [161], and secondary structure of mRNA predicted using RNAFold [42, 161, 162]. The PITA prediction score is an energetic score showing the free-energy change upon the microRNA:mRNA binding [42]. The predictive results of PITA can be further filtered to keep conserved sequences by using phastCons, which is built on hidden Markov model [42, 163]. TargetScan was designed to search for the conserved sequence complementarity between the seed region of microRNA and the 3′UTR of mRNA [26]. The context score was used to evaluate the binding probability and the \( P_c \) score (Probability of Conserved Targeting) was used to evaluate the significance of the prediction [26, 164].

However, those predictors are facing the emerging challenges. For instance, many predictors were developed using small sets of features or small sets of training samples [44, 43]. When applying those predictors in a large sets, the results may be questioned. In addition, the accuracy of many predictors may not fit the requirements since many predictors were developed based on either high sensitivity or high specificity [44]. Moreover, the machining-learning based methods are less transparent [165]. How and why a query sample is determined to be positive or negative is elusive. All those issues may affect the application
of the computational predictors. Therefore, it needs the new methods to overcome these issues. There are multiple ways used to improve the prediction performance of computational predictors. These methods, especially the meta-strategy, show the high performance of predictions in various biological fields, including protein folding recognition [166], protein secondary structure prediction, protein intrinsic disorder [167, 168, 169], protein interaction [170], protein subcellular locations [105, 171], post-translational modification [172], promoter prediction [28], microRNA prediction [173, 174], nucleosome organization [175], mass-spectrometry analysis [176], and many others. The meta-strategy can combine those “orthogonal true predictions”, therefore, the final true prediction rate can be improved [165].

In this section, we first established a comprehensive murine microRNA:mRNA interaction database. We published paper “Consensus datasets of mouse miRNA-mRNA interactions from multiple online resources” in data brief. In section 3.2, we mainly introduced our new established eleven murine microRNA:mRNA data sets.

The decision-tree-based meta-strategy and multi-threshold sequential-voting technique was used to combine the predicted results from each component individual predictor. We published paper “Improving prediction accuracy using decision-tree-based meta-strategy and multi-threshold sequential-voting exemplified by miRNA target prediction” in Genomics. In section 3.3, we further talked about how to develop the novel meta-strategy and improve the microRNA:mRNA prediction. One of mouse eleven microRNA:mRNA interaction data set by combining miRanda, miRDB, and PITA databases was used to test and validate the effectiveness of the novel strategy.

Further applying the decision-tree-based multi-threshold sequential-voting technique on all of eleven murine microRNA:mRNA interaction data sets, the performance of the meta-strategy are not performed well in most of the data sets. These results indicate directly applying meta-strategy may not be efficient to improve the performance of predictors. Therefore, we further studied the indirect meta-strategy by integrating novel data analysis techniques,
including dual-threshold, significance voting, and encoding strategies into decision-tree and artificial neural network. We published paper “Significant improvement of miRNA target prediction accuracy in large datasets using meta-strategy based on comprehensive voting and artificial neural networks” in *BMC Genomics*. In section 3.4, we introduced the updated data sets, novel meta-strategy infrastructure, and the evaluations of the meta-strategy on the independent data sets. We also extended the finalized meta-predictor on human microRNA:mRNA interaction data sets. Since the human microRNA:mRNA data sets are significantly different compared with murine data sets, we retrained the model and the results are shown in section 3.5.

Furthermore, the similar model was extended in different biological field to test the prediction performances of those models in various biological fields, including microRNA structure, disorder protein, and binding region in disorder proteins. We published paper “Decision-tree Based Meta-Strategy Improved Accuracy of Disorder Prediction and Identified Novel Disordered Residues Inside Binding Motifs” on *International Journal of Molecular Sciences*. The methods and results are shown in section 3.6.

### 3.4 Consensus Datasets of Mouse MicroRNA:mRNA Interactions from Multiple Online Resources

The eleven mouse microRNA:mRNA data sets were established to contain as many as microRNA:mRNA interaction pairs. We published the paper, “Consensus datasets of mouse miRNA-mRNA interactions from multiple online resources” in *data brief* in 2017. In this paper, we stated the methods to combine the databases and the procedure to design confidence score for each microRNA:mRNA interaction pairs.
3.4.1 Materials, Methods, and Results

3.4.1.1 Databases

MirTarBase Release 6.0 [43] and TarBase v7.0 [44] are two experimental validated microRNA:mRNA interaction databases, which were used to download the mouse genome experimental validated microRNA:mRNA pairs. The mouse genome predicted microRNA:mRNA interactions, which are pre-assembled data sets, were downloaded from four individual databases, miRanda 2010 release [18], miRDB v5.0 [42], PITA v6 [21], and TargetScan 7.0 [177].

MirTarBase and TarBase are widely used experimental validated microRNA:mRNA interaction databases [44, 43], containing comprehensive microRNA:mRNA interactions from both dependent and independent sources, and can be downloaded without any restrictions [44, 43]. Therefore, we selected those two databases as the experimental validated microRNA:mRNA interactions. Both databases include ∼30 different types of experimental methods, including western blot, Cross-Linking Immunoprecipitation (CLIP), reporter assay, etc [44, 43].

miRanda, miRDB, PITA, and TargetScan are four of the most popular predicted microRNA:mRNA interaction databases. The reasons for choosing these four computational predictors are: (A) These predictors are based on the synthesized techniques and have the well-designed infrastructure. The default threshold values of positive prediction are < −1.0, > 80, < −10, and < −0.36, for miRanda, miRDB, PITA, and TargetScan, respectively. (2) These predictors were recently developed or updated, and have well-maintained webserver or applications. (3) These predictors provide pre-assembled predicted microRNA:mRNA interaction data set of mouse genome.
3.4.1.2 Eleven Data Sets

The asfore-mentioned predictors may make either a valid prediction or an invalid prediction for a pair of microRNA:mRNA. The valid prediction always needs a determinative conclusion of the prediction, while the invalid prediction has no meaningful output. Therefore, the predicted pairs from all those four predicted databases were filtered to keep only valid predictions. Since the valid predictions can be four-predictor-overlapped, three-predictor-overlapped, and two-predictor-overlapped, these valid predictions were saved in their corresponding data sets based on the number of predictors overlapped. The total number of data sets is 11, which is calculated by $C(4, 4) + C(4, 3) + C(4, 2) = 1 + 4 + 6 = 11$ and the summary of these 11 data sets was shown in Table 3.1. The microRNA:mRNA pairs which can be validly predicted by one predictor were not included in the data sets, because they are very limited. Afterwards, the microRNA:mRNA pairs in each data set were compared with experimental validated databases mirTarBase and TarBase. If microRNA:mRNA pairs can be found in either of those two databases, the microRNA:mRNA pairs were labeled as true samples, otherwise, the predicted scores of microRNA:mRNA pairs were compared with the default threshold values of each predictor. If all of the predicted scores showing the microRNA:mRNA pair was negative, this pair were labeled as a false sample. If not all of the predictors predicting the microRNA:mRNA pair was negative, this pair was removed from the data sets. The corresponding numbers of true and false samples of these eleven data sets are shown in Table 3.1.

3.4.1.3 Distribution of Confidence Score

All the microRNA:mRNA pairs have a specific number of predictive scores given by specific predictors. The confidence score of each microRNA:mRNA in each data set was calculated by procedure as follow. Assuming the $i$-th sample in this data set has $S$ predictive
Table 3.1. Summary of the eleven datasets.

<table>
<thead>
<tr>
<th>Dataset ID</th>
<th>Associated predictors</th>
<th>No. of true samples</th>
<th>No. of false samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>D4</td>
<td>miRanda.miRDB.PITA.TargetScan</td>
<td>1739</td>
<td>9870</td>
</tr>
<tr>
<td>D3-1</td>
<td>miRanda.miRDB.PITA</td>
<td>1423</td>
<td>11,062</td>
</tr>
<tr>
<td>D3-2</td>
<td>miRanda.miRDB.TargetScan</td>
<td>195</td>
<td>2667</td>
</tr>
<tr>
<td>D3-3</td>
<td>miRanda.PITA.TargetScan</td>
<td>2997</td>
<td>43,082</td>
</tr>
<tr>
<td>D3-4</td>
<td>miRDB.PITA.TargetScan</td>
<td>1990</td>
<td>24,088</td>
</tr>
<tr>
<td>D2-1</td>
<td>miRanda.miRDB</td>
<td>187</td>
<td>2722</td>
</tr>
<tr>
<td>D2-2</td>
<td>miRanda.PITA</td>
<td>3817</td>
<td>177,467</td>
</tr>
<tr>
<td>D2-3</td>
<td>miRanda.TargetScan</td>
<td>423</td>
<td>21,858</td>
</tr>
<tr>
<td>D2-4</td>
<td>miRDB.PITA</td>
<td>1713</td>
<td>14,264</td>
</tr>
<tr>
<td>D2-5</td>
<td>miRDB.TargetScan</td>
<td>1638</td>
<td>189,900</td>
</tr>
<tr>
<td>D2-6</td>
<td>PITA.TargetScan</td>
<td>6211</td>
<td>1340</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>22,333</td>
<td>498,320</td>
</tr>
</tbody>
</table>

scores $v^s_i$, $s = 1, ..., S$. $S$ is the total number of predictors associated with this data set.

The $z$-scores of all samples were calculated using the predicted scores of the $s$-th predictor in this data set. The scipy and numpy in python were used to calculate p-value of each pair. Afterwards, the $i$-th sample in the data set has $S$ p-values, $p^s_i$, $s = 1, ..., S$. Each of these p-values was then transformed to $d^s_i = 1/(1 + e^{-p^s_i})$. Finally, if the sample was a true sample, the integrated confidence score was calculated using $C_i = 1 - \pi^S_1 (1 - d^S_i)$; if the sample was a false sample, the integrated confidence $S$ score was determined by $C_i = 1 - \pi^S_1 d^S_i$. Figure 3.1 displays the distributions of integrated confidence scores of both true samples and false samples in the eleven data sets. It is obvious that the ranges of independent variables and dependent variable are different in different data sets. Since we want to show the differences among eleven data sets, we did not further normalize the distribution. One of subsets was refined and applied in our recent study to generate the new decision-tree-based algorithm to improve microRNA and microRNA targets prediction.
Figure 3.1. Distribution of both true and false samples as a function of integrated confidence score in eleven data sets. X-axis shows the integrated confidence score. Y-axis presents the fractions of true samples (solid lines) and false samples (dashed lines) in each interval of integrated confidence score.
3.4.2 Discussion

In this study, we built up eleven data sets, which contain the largest consistent data sets of microRNA:mRNA interaction pairs in mouse genome. Those data sets can be further refined to generate training and testing data sets and used to develop microRNA:mRNA target predictors. Since the microRNA:mRNA pairs in those data sets are consistent among different sources, those data sets can be used directly to search high-confidence microRNA:mRNA interaction pairs.

3.5 Improving Prediction Accuracy Using Decision-tree-based Meta-strategy and Multi-threshold Sequential-voting Exemplified by MicroRNA Target Prediction

In the paper “Improving prediction accuracy using decision-tree-based meta-strategy and multi-threshold sequential-voting exemplified by microRNA target prediction”, we developed a novel meta-strategy by modifying decision-tree-based technique to integrate multi-threshold and sequential-voting techniques. The prediction accuracy was improved in data set D3-1 (miRanda.miRDB.PITA) compared with each component individual predictor. Here, we introduced this proposed meta-strategy in details.

3.5.1 Materials and Methods

3.5.1.1 Data Set

The microRNA:mRNA predicted interaction pairs of mouse genome were from the databases miRanda, miRDB, and PITA. The interaction pairs of microRNA:mRNA from this combined data set were filtered by removing the redundant pairs. After that, the remaining microRNA:mRNA pairs were aligned with experimental validated databases to label the true microRNA:mRNA pairs. Otherwise, the pairs were labeled as ‘N’, which represent the
false microRNA:mRNA pairs. The finalized data set includes 1244 true and 1014 false microRNA:mRNA interaction pairs. This data set was named D2258, which was split into one training set and one independent test set. In training set, there were 1100 true pairs and 900 false pairs [178]. These interaction pairs were used to run five-fold cross-validation. The independent test set contained 144 true pairs and 114 false pairs. This data set was used to evaluate the predictor.

3.5.1.2 Performance Evaluation

The Receiver’s Operating Curve (ROC), accuracy \( ACC = \frac{TP+TN}{TP+FN+TN+FP} \), and sensitivity \( SENS = \frac{TP}{TP + FN} \) were used to evaluate the performances of individual predictors and the newly-designed meta-strategy. Here, TP, FN, TN, and FP represent True Positive, False Negative, True Negative and False Negative, respectively. We applied ACC and SENS to evaluate the newly-designed meta-strategy is because the newly-designed strategy has a significant improvement on the fraction of true positive predictions. Also, these two values are extremely critical for the real applications. When we evaluated the ACC and SENS in individual predictors, the default or most popular threshold values were used. They were \(-1.2\) for miRanda, 80 for miRDB, and \(-12\) for PITA. The threshold values used in meta-strategy were selected using error and try from the pre-selected candidate threshold values.

3.5.1.3 Meta-strategy

In this study, we used three machine-learning-associated predictors, including miRanda [18], miRDB [159] and PITA [42], as component predictors. Those three individual predictors are widely used in scientific research and well-maintained software packages or webservers. The predicted results of these three predictors were integrated to make meta-prediction using a decision-tree-based meta-strategy. The infrastructure of the meta-predictor was
shown in Figure 3.2. The predicted scores of those three component individual predictors were compared sequentially with their corresponding pre-selected threshold values. If the condition was satisfied, the condition that was associated with individual predicted results was assigned into the output of the meta-predictor, and then the execution of meta-predictor was terminated. If the condition was not satisfied, the execution of the meta-predictor was continued until the condition is satisfied. At the final step, there may exist the pairs that are not fit for any condition, those pairs were assigned into the meta-predictor as “null”. The meta-predictor applied different thresholds for true and false predictions. This is the reason to call the technique, multi-threshold. The sequential voting represents all the predicted results of individual predictors were compared with their corresponding threshold values one by one. One predictor has four threshold values, therefore, N individual predictors has 4N thresholds values. If each predictor has ten candidate threshold values, there will be $10^{4n}$ combinations of candidate values. Five-fold cross-validation were performed to evaluate and determine the prediction performances.

3.5.2 Results

3.5.2.1 The Performance of Individual Predictors

The ROC curves of miRanda, miRDB and PITA in data set D2288 were calculated and shown in Figure 3.3(A). The performance of miRanda and miRDB are similar, and generally better than the performance of PITA. Compared among these individual predictors, the PITA has the higher sensitivity to both miRanda and miRDB. However, the specificity of PITA is 0.4 lower than other two individual predictors. Based on this result, it indicates the necessity to calculate further information on the similarity or differences of these individual predictors. The Venn’s diagram of these individual predictors are shown in Figure 3.3(B). In Figure 3.3(B), it shows the comparison of true-positive predictions (left) and false-positive
Figure 3.2. Infrastructure of the decision-tree-based meta-strategy. Query sequences normally refer to a pair of microRNA and mRNA sequences. For each pair of query sequences, miRanda, miRDB, and PITA are used to make individual predictions. After having the predicted scores of all individual predictions, the procedure in the right-hand side of the figure will be followed to determine the output of the meta-strategy. \( S_{\text{miRanda}} \), \( S_{\text{miRDB}} \), and \( S_{\text{PITA}} \) are prediction scores of three individual predictors. \( T_{x,y,z} \) is the threshold value. \( x \) could be \( p \) or \( n \) standing for positive or negative, respectively. \( y \) represents each individual predictor. \( z \) has two values if the threshold has a lower bound and a higher bound. \( z \) is omitted if the threshold has a single value. The values of the thresholds can be found in Table 3.2. “Y” and “N” stand for “Yes” and “No”. “Pos” and “Neg” indicate positive prediction and negative prediction, respectively. When all the conditions in the workshop can’t be satisfied, the output of the meta-strategy is “null”. The chance of having “null” could be rare in simple application of meta-strategy.
predictions (right). It is obvious that the overlap of true-positive predictions of these three individual predictors is not two many compared to the number of true-positive predictions in each individual predictor. To be specific, miRanda has 30 unique true-positive prediction pairs that cannot be identified by PITA or miRDB, PITA has 89 unique true-positive predictions that can’t be predicted by miRDB and miRanda, and miRDB has 382 unique true-positive predictions that can’t be identified by miRanda and PITA. The total unique true samples from these three individual predictors is around 40%, while only around 12% of true-positive samples are overlapped among three individual predictors in data set D2258. The condition of true-negative samples are very different. There are a huge portion of true-negative prediction overlapped in at least two different predictors. There are 477 predictions which can be found among all three predictors. The overlap between miRanda and miRDB, miRanda and PITA, and miRDB and PITA, is 210, 201, and 22, respectively. The overlap of all true-negative samples is $\sim$90% of false samples. Additionally, miRanda can identify 86 unique true-negative predictions which cannot be identified by miRDB and PITA. PITA and miRDB can identify only 6 and 9 unique true-negative pairs, respectively.

3.5.2.2 Determining Threshold Values of Individual Predictors

The comparison of three individual predictors true-positive predictions indicates that combining the predictions can improve the overall true-positive predicted rates. The problem is how to integrate these predicted results of individual predictors. It is clear that simply integrated the results of individual predictors may not improve the performance of predicted accuracy if the distribution of samples in the space of prediction scores of true-positive and true-negative samples are very similar. For this reason, the distribution of each individual predictors as a function of predicted scores for both true samples and false samples in data set D2258 was analyzed and shown in Figure 3.4. The distribution of predicted scores of true samples and false samples is different. For miRanda predictor, when the prediction score
Figure 3.3. Prediction performances of three individual predictors. (A) ROC curves of miRanda, PITA and miRDB. (B) Overlap of predictive results among three individual predictor for true positive samples (left) and true negative samples (right).
is higher than $\sim 0.7$, the fraction of true samples is much higher than that of false samples. When the score is lower than $\sim 0.2$, the fraction of false samples is much higher than that of true samples. When the predicted score is between $\sim 0.2$ and $\sim 0.6$, the distribution of true samples and false samples are similar. For PITA predictor, when the predicted score is between $\sim 0.36$ and $\sim 0.56$, the fraction of true samples is higher than the fraction of false samples, while when the predicted scores is higher than $\sim 0.76$ or lower than $\sim 0.2$, the fraction of false samples is higher than the true samples. For miRDB predictor, if the predicted score is lower than $\sim 0.44$, the fraction of false samples is higher than the true samples. If the predicted score is higher than $\sim 0.62$, the fraction of true samples is higher than the false samples.

Since each predictor has its own threshold values for true samples and false samples, it is possible that using different threshold values for positive predictions and negative predictions can improve the predicted result. After choosing the threshold values of each predictor, the predicted results of individual predictors were compared to the threshold values sequentially the same as shown in Figure 3.2. In this way, the meta-predictor should have both high sensitivity and specificity. The threshold values of each individual predictors are shown on Table 3.2.

Table 3.2. Summary of threshold values of different individual predictors for both positive and negative predictions.

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Positive Prediction</th>
<th>Negative Prediction</th>
</tr>
</thead>
<tbody>
<tr>
<td>miRanda</td>
<td>$&gt;0.7$</td>
<td>$&lt;0.2$</td>
</tr>
<tr>
<td>PITA</td>
<td>$&gt;0.38$ and $&lt;0.58$</td>
<td>$&gt;0.37$ or $&gt;0.59$</td>
</tr>
<tr>
<td>miRDB</td>
<td>$&gt;0.66$ and $&lt;0.9$</td>
<td>$&lt;0.28$</td>
</tr>
</tbody>
</table>

3.5.2.3 Evaluate Meta-predictor

The performance of the newly-designed meta-predictor was shown in Figure 3.5. The comparison of individual predictors and the newly-designed meta-predictor in training set
under five-fold cross-validation shows that the meta-predictor has higher sensitivity and accuracy than other three individual predictors. The accuracy of individual predictors in training set is around 40%-60%, while the accuracy of meta-predictor is around 90%. For the comparison of sensitivity, the meta-predictor has around 87% sensitivity in training set, which is much higher than the sensitivities of any other three individual predictors. In the independent data sets, the meta-predictor has around 85% accuracy and around 80% sensitivity.

3.5.3 Discussion

Many studies have demonstrated the meta-strategy is one of the methods which can be used to efficiently improve the prediction accuracy. However, if the distribution of true samples and false samples predicted scores are similar, simply integrated multiple individual predictors may not improve the prediction performance significantly [174]. Non-linear transformations is one method which can be used to solve the distribution issues. However, it is hard to determine the useful non-linear transformation method. Machine-learning and deep learning methods were applied to build meta-predictors. They display the high performance
Figure 3.5. Comparison of prediction performance between individual predictors and meta-predictor. X-axis displays miRanda, PITA, miRDB are three individual predictors. DT and DT-indpt stand for the performance of decision-tree-based meta-strategy under five-fold cross-validation in training set and in independent test set, respectively. Grey bars represent accuracy ($\text{ACC} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{FN} + \text{TN} + \text{FP}}$), and black bars show sensitivity ($\text{SENS} = \frac{\text{TP}}{\text{TP} + \text{FN}}$). Here, TP, FN, TN, and FP stand for True Positive, False Negative, True Negative, and False Negative, respectively.
of the predicted result. However, those methods are less transparent. The developers may not clearly know why and how the results come.

In this study, we applied decision-tree-based technique, which is a popular machine-learning technique, to integrate multiple individual predictors. The infrastructure of this new meta-predictor was shown in Figure 3.2. The multi-threshold sequential voting were applied to integrate the predicted results from miRanda, PITA and miRDB, and the prediction performance of this meta-predictor is higher than other individual predictors in training set under five-fold cross-validation and independent test set.

Many individual predictors were designed to have high sensitivity of high specificity. In the real applications, both accuracy and sensitivity of the predictor are more critical. Meta-strategy is one of the method which can be used to solve the issues, by balancing the sensitivity and specificity. The reason is because some predictors have high sensitivity while some predictors have high specificity. Combining the results of individual predictors, the meta-predictor is able to have both high sensitivity and specificity.

In this study, only three individual predictors were applied to generate meta-predictor. Only one predictive score was compared in each step. It is possible to include more individual predictors into meta-predictor [168, 169]. In our previous projects on the prediction of disordered proteins and microRNA [169, 174], it has been demonstrated that meta-predictor built from different individual predictors may have different prediction performance, even if the numbers of individual predictors are the same [169, 174]. In addition, including more individual predictors into meta-predictor, the performance of meta-predictor can normally be improved [169, 174]. Therefore, we assumed it is possible to include the more individual predictors and consider more complicated situation into meta-predictors.
In our previous study, we found the integration of multi-threshold and sequential voting with decision-tree-based technique can improve the prediction performance of microRNA and mRNA interactions in data set D2258 [29]. However, when we applied this meta-strategy on other mouse microRNA:mRNA data sets, the performances based on the meta-strategy do not increase significantly in most of the data sets. These results indicate directly apply meta-strategy may not be efficient to improve the final predicted result [169, 174]. Further testing the multi-threshold and sequential-voting with decision-tree-based model, we found that if the more stringent threshold value was applied in the model, the higher confidence result was obtained and the fewer number of microRNA:mRNA true pairs was predicted [179]. If the less stringent threshold value was applied in the model, the more true pairs were predicted but the confidence of predicted pairs was low. The issue is how to balance the confidence and the number of true predictions. To solve this problem, we designed the upgraded strategy by (1) Updating the mouse microRNA:mRNA interaction data sets; (2) Modifying the multi-threshold sequential voting to dual-threshold two-step selection; and (3) Integrating with artificial neural network. In our recent published paper “Significant improvement of microRNA target prediction accuracy in large data sets using meta-strategy based on comprehensive voting and artificial neural networks”, we found by integrating dual-threshold threshold two-step significant voting decision-tree-based model with artificial neural network, the performances of newly-designed meta-predictors are better in most of eleven mouse microRNA:mRNA interaction data sets.
3.6.1 Materials and Methods

3.6.1.1 Data Sets

The eleven mouse microRNA:mRNA were updated to the latest version from each database using the same method in section 3.2. If the microRNA:mRNA pairs were found in all four databases, miRanda [18], miRDB [159], PITA [42] and TargetScan [21], these pairs were deposited into data set D4. If microRNA:mRNA pairs were only found in three databases, i.e. miRanda, miRDB, and PITA, the pairs were saved in data set D3-1. Similarly, D3-2 for pairs were only found in miRanda, MiRDB, and TargetScan; D3-3 for pairs were only found in miRanda, PITA, and TargetScan; and D3-4 for pairs were common in MiRDB, PITA, and TargetScan. If microRNA:mRNA pairs were only found in two databases, i.e. miRanda and miRDB, the pairs were kept in data set D2-1. Similarly, D2-2 for pairs were in miRanda and PITA, D2-3 for pairs were in miRanda and TargetScan, D2-4 for pairs were in miRDB and PITA, D2-5 for pairs were in miRDB and TargetScan, and D2-6 for pairs were in PITA and TargetScan. If the microRNA:mRNA pairs were only found in one database, the pairs were grouped in data sets D1-1, D1-2, D1-3 and D1-4, which included unique pairs in database miRanda, miRDB, PITA and TargetScan, respectively.

The microRNA:mRNA interaction pairs in different data sets were aligned with miRTarBase and TarBase to label the interaction pairs. If a sample was found in either of these two databases, it was assigned as a positive sample, otherwise, negative sample. If any two interaction pairs had less than 2% difference between all predicted scores, these pairs were treated as redundant pairs and removed from the data sets. Since the D1 series data sets only included one database, they were not used for further study. The finalized data sets are shown in Table 3.3.
Table 3.3. Numbers of samples in each of the eleven newly-updated data sets.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Associated predictors</th>
<th>No. of Original</th>
<th>No. of non-red.</th>
<th>No. of true</th>
<th>No. of False</th>
<th>No. of microRNA</th>
<th>No. of mRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>D4</td>
<td>miRanda.miRDB, PITA, TargetScan</td>
<td>45,517</td>
<td>22,446</td>
<td>1844</td>
<td>20,602</td>
<td>211</td>
<td>6747</td>
</tr>
<tr>
<td>D3-1</td>
<td>miRanda.miRDB, PITA</td>
<td>29,486</td>
<td>9271</td>
<td>1339</td>
<td>7932</td>
<td>212</td>
<td>4619</td>
</tr>
<tr>
<td>D3-2</td>
<td>miRanda.miRDB, TargetScan</td>
<td>7584</td>
<td>2478</td>
<td>198</td>
<td>2280</td>
<td>201</td>
<td>1097</td>
</tr>
<tr>
<td>D3-3</td>
<td>miRanda.PITA, TargetScan</td>
<td>107,813</td>
<td>5529</td>
<td>1220</td>
<td>4309</td>
<td>205</td>
<td>3541</td>
</tr>
<tr>
<td>D3-4</td>
<td>miRDB.PITA, TargetScan</td>
<td>66,384</td>
<td>2984</td>
<td>864</td>
<td>2120</td>
<td>323</td>
<td>1946</td>
</tr>
<tr>
<td>D2-1</td>
<td>miRanda.miRDB</td>
<td>5269</td>
<td>892</td>
<td>199</td>
<td>693</td>
<td>186</td>
<td>641</td>
</tr>
<tr>
<td>D2-2</td>
<td>miRanda.PITA</td>
<td>216,923</td>
<td>974</td>
<td>457</td>
<td>517</td>
<td>202</td>
<td>883</td>
</tr>
<tr>
<td>D2-3</td>
<td>miRanda.TargetScan</td>
<td>32,566</td>
<td>429</td>
<td>151</td>
<td>278</td>
<td>162</td>
<td>342</td>
</tr>
<tr>
<td>D2-4</td>
<td>miRDB.PITA</td>
<td>29,531</td>
<td>810</td>
<td>455</td>
<td>355</td>
<td>165</td>
<td>645</td>
</tr>
<tr>
<td>D2-5</td>
<td>miRDB.TargetScan</td>
<td>256,784</td>
<td>430</td>
<td>174</td>
<td>256</td>
<td>259</td>
<td>337</td>
</tr>
<tr>
<td>D2-6</td>
<td>PITA.TargetScan</td>
<td>384,944</td>
<td>363</td>
<td>179</td>
<td>184</td>
<td>175</td>
<td>288</td>
</tr>
</tbody>
</table>

3.6.1.2 Meta-predictor

The infrastructure of meta-predictor is shown in Figure 3.6. For the queried pairs of microRNA and mRNA sequences, the predictions are first made to obtain predicted scores from four individual predictors, including miRanda, miRDB, PITA and TargetScan. The predicted results are checked to see whether each individual predictor can provide the scored predictions. After that, the results will be put into the decision-tree-based Artificial neural network (DANN) to achieve the final meta-prediction. For example, if the queried microRNA:mRNA interaction pair has predicted scores from database miRanda and PITA, the predicted scores of the interaction pair will be put into the model DANN-2-2, which is trained from data set D2-2.

Several novel techniques were integrated in each DANN model, including dual-threshold, majority-voting, two-step selection, significant-voting, decision-tree and artificial neural network. Dual-threshold represents two different threshold values are used to determine positive
and negative predictions. Two-step selection represents the true positive and true negative result are decided by two independent processing. In the second step, if the pairs cannot be determined by their corresponding threshold values, their Euclidean distances from the corresponding threshold value will be calculated and used to determine the pairs are true or false predicted results. This is called significant-voting. The significant-voting usually come after the majority-voting technique. For instance, when the microRNA and mRNA pairs have four predicted scores, of which, two are positive predicted scores and two are negative predicted scores, the comparison of the number of predicted scores are the same and may not result in the useful conclusion. Under this condition, the comparison of the sum of Euclidean distances from true threshold values \(d_t\) and false threshold values \(d_f\) will be used to provide further information of the relative significance. After applying all the techniques, the predictive results from individual predictors will be encoded by six different categories. The encoded predicted results will be input into the artificial neural network, which is a fully-connected two-hidden layer structure.

The artificial neural network contains in input layer, the first and second hidden layers and the output layer. The unit of input layer is determined by the number of individual component predictors plus six, which is category dimension from the two-step significant voting decision-tree. The nodes of input layers can be 10, 9, and 8 for model DANN-4, DANN-3 and DANN-2, respectively. There are ten and twenty hidden units in the first and second layers, respectively. Since the interaction pairs have two categories, the output layer has two nodes, the true samples is [1,0] and the false samples is [0,1]. The activation function of the output layer is SoftMax \((O^T_i = exp^{O_i}/\sum exp^{O_i}, i = 1 \text{ and } 2)\).

3.6.1.3 Training and Validation

Each DANN predictor was trained using the corresponding data set. For each data set, 20% interaction pairs were randomly selected from the data set to compose an independent
Figure 3.6. Infrastructure of decision-tree-based meta-predictor. Queried miRNA:mRNA sequences are firstly fed into miRanda, miRDB, PITA, and TargetScan to get individual predictions. These individual predictions may be scored or unscored (null output). Based on the scored individual predictions, a specific module of decision tree based artificial neural networks (DANN) will be selected. For example, if only miRanda and miRDB have scored predictions, module DANN-2-1 will be selected. There are eleven modules in the pipeline, each module corresponds to one of the eleven data sets and uses scores from different predictors as follows, DANN-4: miRanda, miRDB, PITA, and TargetScan; DANN-3-1: miRanda, miRDB, and PITA; DANN-3-2: miRanda, miRDB, and TargetScan; DANN-3-3: miRanda and miRDB; DANN-2-1: miRanda and PITA; DANN-2-2: miRanda and TargetScan; DANN-2-3: miRanda and TargetScan; DANN-2-4: miRDB and PITA; DANN-2-5: miRDB and TargetScan; DANN-2-6: PITA and TargetScan. “Y” above an arrow and “N” along an arrow represent “Yes” and “No”. “T/F” inside a circle stands for positive (T) or negative (F) prediction. $N_{T1}$ is the number of predictors that make true prediction using the 1st-level true threshold values, and so on so forth for $N_{T2}$, $N_{F2}$, $b1$, and $b2$. “$c1$” and “$c2$” are the differences of the predictions score from their corresponding 1st-level threshold values. “$d_{T2}$” and “$d_{F2}$” are the Euclidean distances of prediction scores from their corresponding 2nd-level threshold values for true (T) predictions and false (F) predictions, respectively. The infrastructure of the 2-hidden-layer ANN is described in the text. There are in total eleven DANNs.
test set, the rest 80% interaction pairs were equally split into multiple subsets to run multi-
fold cross-validation. Based on the number of interaction pairs in each data set, five-fold
cross-validation or three-fold cross-validation was applied on the data set.

3.6.1.4 Pre-processing Individual Predictors

The four component individual predictors have different ranges: miRanda (-1.364, -0.1),
miRDB (50, 100), PITA (-43.24, 21.4), and TargetScan (-9.05, 0). The default threshold
values for positive predictions are <-1.0, >80, <-10 and <-0.36, respectively. The predicted
scores of miRDB were multiplied by -1 to inverse the scores. All of predicted scores from
each individual predictor were scaled into the range (-1, 1). Thus, the corresponding true
prediction threshold values for four predictors in their default settings are <-0.424, <-0.2,
<0.028, and <0.920, accordingly:

3.6.1.5 Information Gain

Information Gain (IG) was calculated using the scaled predicted scores from each indi-
vidual predictor. The formula is as follows:

\[ IG(x) = \sum_{i=1,2} p_i \log_2 p_i - \sum_{j=1,2} f_j(x) \sum_{k=1,2} p_{j,k} \log_2 p_{j,k} \]

Where, \( p_i \) is the fraction of positive \((i=1)\) or negative \((i=2)\) samples in the data set. \( x \) is
the threshold prediction score to split the data set into two groups, \( f_j(x) \) is the fraction of
samples with prediction score higher than the threshold \((j=1)\) or the fraction of samples
with prediction score lower than the threshold \((j=2)\), \( p_{j,k} \) refers to the fraction of positive
samples \((k=1)\) or negative samples \((k=2)\) in the \( j \)-th group.

The information gain as a function of the scaled predicted score, as well as the distribution
of true samples and false samples, for the component individual predictors in data set D4
was shown in Figure 3.7. Similar information of other ten data sets were shown in Figure 3.8. The curves of information gain are very different from the curves of distribution of true samples and false samples. The information gain can represent the association between the selected features with the target information. The spikes of curves indicate at those threshold values, the feature can provide the most information associated with the target. By taking into consideration, the true sample predicted scores are smaller in the scaled predicted scores, and the threshold values for positive predictions should be smaller than the threshold values for negative predictions. Therefore, the 1st-step threshold values should be more stringent than the 2nd-step threshold values. Multiple combinations of threshold values were systematically tested, and the final determined positive/negative threshold values of miRanda, miRDB, PITA, and TargetScan in the data set D4 are (-0.179/0.844, -0.702/0.059, -0.449/0.257, 0.896/0.982) as the 1st-step threshold values, and (0.067/0.199, -0.480/-0.269, -0.203/-0.107, 0.964/0.972) as the 2nd-step threshold values. The threshold values for other ten data sets were shown in Table 3.4.
Figure 3.7. Information gain compared to the distribution of positive and negative samples in the D4 data set for (A) miRanda, (B) miRDB, (C) PITA, and (D) TargetScan. X-axis shows the scaled prediction score, y-axis on the left shows the value of information gain, and y-axis on the right shows the distribution of positive samples (red dashed) and negative samples (cyan solid).
Figure 3.8. Information gain compared to the distribution of positive and negative samples in four D3 series data sets and six D2 series data sets for (A) miRanda, (B) miRDB, (C) PITA, and (D) TargetScan, when the prediction scores of these predictors are available. X-axis shows the scaled prediction score, y-axis on the left shows the value of information gain, and y-axis on the right shows the distribution of positive samples (red dashed) and negative samples (cyan solid).
Figure 3.8. (Continued)
Figure 3.8. (Continued)
Table 3.4. The 1st-step and 2nd-step threshold values for both true and false predictions in eleven DANN modules.

<table>
<thead>
<tr>
<th>Module</th>
<th>1st-step threshold</th>
<th>2nd-step threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>miRanda</td>
<td>miRDB</td>
</tr>
<tr>
<td>D4</td>
<td>(0.179)/0.844</td>
<td>(0.702)/0.059</td>
</tr>
<tr>
<td>D3-1</td>
<td>(0.329)/0.155</td>
<td>(0.476)/0.071</td>
</tr>
<tr>
<td>D3-2</td>
<td>0.570/0.798</td>
<td>0.013/0.247</td>
</tr>
<tr>
<td>D3-3</td>
<td>(0.135)/0.393</td>
<td>~~~~</td>
</tr>
<tr>
<td>D3-4</td>
<td>~~~~</td>
<td>(0.922)/(0.806)</td>
</tr>
<tr>
<td>D2-1</td>
<td>0.375/0.716</td>
<td>(0.624)/(0.447)</td>
</tr>
<tr>
<td>D2-2</td>
<td>(0.369)/0.403</td>
<td>~~~~</td>
</tr>
<tr>
<td>D2-3</td>
<td>(0.103)/0.592</td>
<td>~~~~</td>
</tr>
<tr>
<td>D2-4</td>
<td>~~~~</td>
<td>(0.888)/(0.247)</td>
</tr>
<tr>
<td>D2-5</td>
<td>~~~~</td>
<td>(0.664)/0.426</td>
</tr>
<tr>
<td>D2-6</td>
<td>~~~~</td>
<td>~~~~</td>
</tr>
</tbody>
</table>

N.b. Numbers above and below the slashes are for positive and negative predictions, respectively. Values inside parentheses are negative values. “~~~” indicates the threshold values are not applicable in that module.
3.6.1.6 Predictor Performance Evaluation

Sensitivity (Sens), Specificity (Spec), Accuracy (Acc), F1 score (F1), and Matthews Correlation Coefficient (MCC) were used to evaluate the performance of newly-designed meta-predictors in training and test sets. The results were compared with the component individual predictors, miRanda, miRDB, PITA, and TargetScan, and other two recently developed predictors, ComiR and Oliveira’s predictor.

3.6.2 Results

3.6.2.1 Performance of Individual Predictors in the Data Sets

The eleven data sets were formed based on the number of predicted results from the component individual predictors, miRanda, miRDB, PITA and TargetScan, to effectively compare the prediction performance and integrate the outputs of the component individual predictors. To evaluate the performances of four individual predictors in each of the subsets, the Receiver Operating Characteristic (ROC) curves of these individual predictors were applied and the results are shown in Figure 3.9. Overall, the performances of these four individual predictors in all of the eleven data sets are not very well and have the room for the improvement. In different data sets, the performances of individual predictors are various. When the sensitivity and specificity is under their default threshold values, the shape of ROC curves are different.

Using Area Under the ROC Curve (AUC) to evaluate the performances of each individual predictor in the specific data set, miRDB achieved the highest value in the data sets D4, D3-1, D3-2, D3-4, D2-1, D2-4 and D2-5, PITA obtained the highest value in the data sets D3-3, D2-2, and D2-6, and miRanda was ranked at the first position in data set D2-3. Applying sensitivity to evaluate the performance of these four individual predictors, miRDB outperformed other predictors in data sets D4, D3-1, D3-2, D3-4, and D2-1, PITA exceeded
others in data sets D3-3, D2-4, and D2-6, miRanda was better than other predictors in data sets D2-2, and TargetScan had the higher values than other predictors in data sets D2-3 and D2-5. It is clear that using different predicted scores of the individual predictors to form multiple smaller subsets may provide a novel start point for evaluating the performance of different predictors. It needs to be noted that the data sets used in Figure 3.9 are not redundant data sets. The distributions of predictors in data sets with redundant data may be very different. In Figure 3.10, the ROC curves of the eleven data sets with duplicate samples are displayed.

Figure 3.11(A) shows the values of overlap and coverage of different pairs of predictors for positive samples in the eleven data sets. Clearly, the values of pairwise overlap were mostly around or lower than 20%, except in the data sets D3-4, D2-5 and D2-6. In these three data sets, the values of pairwise overlap were still less than 40%. The values of pairwise coverage were normally more than 40%. In data sets D3-3, D2-2, and D2-3, the values of pairwise coverage reached to ~40%. In data sets D4, D3-1, D3-1, and D2-1, the value of coverage were ~60%. In data sets D3-4, D2-4, D2-5, and D2-6, the values reached to ~70%. Normally, the values of overall coverage of multiple predictors in the D4 and D3 series data sets were ~10% higher than the highest pairwise coverage in that data set. The overlap and coverage of negative samples are presented in Figure 3.11(B). The pairwise overlap for negative samples were normally between 40-50% in data sets D4, D3-1, D3-2, D3-3, D2-1, and D2-2, while the pairwise coverage values reached to ~90%. In the D3-4, D2-2, D2-3, and D2-4 data sets, the values of pairwise overlap were between 10-30%, and the values of pairwise coverage were normally around 80%. In the D2-5 and D2-6 data sets, the overlap is below 10% and the coverage was at ~60%. The values of overall coverage of multiple predictors were normally 5% higher than the highest pairwise coverage values. Based on this result, it is possible to improve the prediction accuracy by combining the different predictors,
Figure 3.9. ROC curves of individual predictors in the eleven newly-designed non-redundant data sets. The data sets from top to bottom and from left to right are: D4, D3-1, D3-2, D3-3, D3-4, D2-1, D2-2, D2-3, D2-4, D2-5, and D2-6, respectively. Samples in the D4 data set have prediction scores from four individual predictors, therefore, there are four ROC curves each for a predictor. Similarly, the D3 series data sets and D2 series data sets have three and two ROC curves, respectively. In each of the insets, x-axis shows the value of 1-specificity, while y-axis shows the values of sensitivity.
Figure 3.10. ROC curves of individual predictors in eleven newly-designed data set containing duplicate samples. The data sets from top to bottom and from left to right are: D4, D3-1, D3-2, D3-3, D3-4, D2-1, D2-2, D2-3, D2-4, D2-5, and D2-6, respectively. In each of the insets, x-axis shows the value of 1-specificity, while y-axis shows the values of sensitivity.
because the values of pairwise coverage are much higher than these of overlap in individual predictors.

3.6.2.2 Accuracy of Each Combined Processing

As the results of ROC curves, the prediction performances in each subsets may have the room to be improved. In our study, we designed novel meta-strategy to integrate a number of novel techniques. The dual-threshold value and two-step significant voting were integrated into decision-tree-based techniques to make the first prediction. The ANN was combined after that to integrate the outputs from the first step and make the final meta-prediction. The two steps are essentially separated from each other. That means, there are actually two categories of techniques used in the newly-designed meta-predictor: (1) Decision-tree (incl. dual-threshold value and significant-voting); (2) ANN. Therefore, there are three different processes of these techniques: (1) individual predictors + decision tree (C-I); (2) individual predictors + ANN (C-II); and (3) individual predictors + decision tree + ANN (mirTarDANN).

We tested the performances of these three processes in the eleven data sets under multi-fold cross-validation. The results are presented in Table 3.5. In the first process, the performances of decision-tree-based strategy have limitations to improve the prediction performance. The sensitivity and specificity were not usually balanced in data sets. In the second process, the performances of ANN model had the significant improvement in accuracy for the eleven data sets. However, the sensitivity was very low. In the third process, the decision-tree-based model was combined with ANN, the prediction performances were improved significantly in most of eleven data sets. We named the model MirTarDANN, and this model has the advantages to obtain higher sensitivity in most data sets, with the relatively higher accuracy. The novel model MirTarDANN improved the balances between sensitivity and specificity. Therefore, the third process was used as the formal model to
Figure 3.11. Overlap and coverage between individual predictors for (A) positive samples and (B) negative samples. Grey bars are the values of overlap between two predictors (pairwise overlap). Each cap indicates the value of overlap for a specific pair of predictors. Dark grey bars stand for the values of coverage between two predictors (pairwise coverage), with caps each for a specific pair of predictors. Black bars show values of all-inclusive coverage, which are calculated from all predictors in that data set. Apparently, only D4 and D3 series data sets have the all-inclusive coverage.
predict microRNA:mRNA interaction. The MirTarDANN includes eleven models, DANN-4, four DANN-3-x and six DANN-2-x, which were trained and validated based on their corresponding data sets.
Table 3.5. Prediction performances of three meta-predictor processes using three different sets of techniques.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>C-I</th>
<th></th>
<th></th>
<th>C-II</th>
<th></th>
<th></th>
<th></th>
<th>mirTarDANN</th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Sens</td>
<td>Spec</td>
<td>Acc</td>
<td>Sens</td>
<td>Spec</td>
<td>Acc</td>
<td>Sens</td>
<td>Spec</td>
<td>Acc</td>
<td>Sens</td>
<td>Spec</td>
</tr>
<tr>
<td>D4</td>
<td>0.240±0.120</td>
<td>0.780±0.128</td>
<td>0.735±0.108</td>
<td>0.357±0.093</td>
<td>0.733±0.087</td>
<td>0.703±0.072</td>
<td>0.568±0.049</td>
<td>0.555±0.031</td>
<td>0.556±0.025</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3-1</td>
<td>0.475±0.063</td>
<td>0.602±0.036</td>
<td>0.584±0.022</td>
<td>0.297±0.032</td>
<td>0.789±0.044</td>
<td>0.716±0.031</td>
<td>0.602±0.048</td>
<td>0.530±0.041</td>
<td>0.541±0.032</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3-2</td>
<td>0.671±0.097</td>
<td>0.235±0.108</td>
<td>0.268±0.094</td>
<td>0.433±0.128</td>
<td>0.690±0.041</td>
<td>0.669±0.036</td>
<td>0.511±0.176</td>
<td>0.605±0.111</td>
<td>0.593±0.092</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3-3</td>
<td>0.087±0.018</td>
<td>0.693±0.042</td>
<td>0.557±0.028</td>
<td>0.693±0.052</td>
<td>0.704±0.069</td>
<td>0.701±0.047</td>
<td>0.660±0.037</td>
<td>0.730±0.056</td>
<td>0.715±0.037</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D3-4</td>
<td>0.383±0.058</td>
<td>0.434±0.088</td>
<td>0.420±0.050</td>
<td>0.622±0.061</td>
<td>0.752±0.030</td>
<td>0.713±0.012</td>
<td>0.650±0.051</td>
<td>0.722±0.051</td>
<td>0.700±0.026</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2-1</td>
<td>0.700±0.073</td>
<td>0.221±0.040</td>
<td>0.334±0.029</td>
<td>0.402±0.065</td>
<td>0.653±0.040</td>
<td>0.593±0.013</td>
<td>0.312±0.066</td>
<td>0.722±0.078</td>
<td>0.624±0.037</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2-2</td>
<td>0.519±0.049</td>
<td>0.444±0.040</td>
<td>0.480±0.023</td>
<td>0.668±0.202</td>
<td>0.728±0.017</td>
<td>0.701±0.094</td>
<td>0.685±0.102</td>
<td>0.796±0.020</td>
<td>0.745±0.038</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2-3</td>
<td>0.306±0.137</td>
<td>0.307±0.062</td>
<td>0.312±0.035</td>
<td>0.527±0.055</td>
<td>0.849±0.019</td>
<td>0.738±0.042</td>
<td>0.558±0.047</td>
<td>0.849±0.026</td>
<td>0.749±0.017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2-4</td>
<td>0.398±0.024</td>
<td>0.659±0.038</td>
<td>0.509±0.004</td>
<td>0.777±0.081</td>
<td>0.393±0.182</td>
<td>0.608±0.050</td>
<td>0.745±0.023</td>
<td>0.540±0.029</td>
<td>0.749±0.017</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2-5</td>
<td>0.388±0.175</td>
<td>0.188±0.174</td>
<td>0.276±0.044</td>
<td>0.700±0.079</td>
<td>0.837±0.060</td>
<td>0.779±0.011</td>
<td>0.721±0.086</td>
<td>0.849±0.022</td>
<td>0.797±0.050</td>
<td></td>
<td></td>
</tr>
<tr>
<td>D2-6</td>
<td>0.700±0.155</td>
<td>0.255±0.032</td>
<td>0.466±0.063</td>
<td>0.678±0.065</td>
<td>0.724±0.048</td>
<td>0.700±0.038</td>
<td>0.622±0.074</td>
<td>0.769±0.060</td>
<td>0.693±0.017</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

N.b. C-I integrates individual predictors with decision-tree-based model. C-II uses ANN to integrate the individual predictors. mirTarDANN combines individual predictors, decision-tree-based model, and ANN. The performance is measured by sensitivity (Sens), specificity (Spec), and accuracy (Acc) under multi-fold cross-validation. The bold values are the highest among three processes of meta-predictors in the same subsets.
3.6.2.3 Performance of Newly-designed MirTarDANN

The performances of component individual predictors, ComiR, and the newly-designed meta-predictor, mirTarDANN were compared using accuracy, sensitivity, and F1 score in their corresponding training sets under multi-fold cross-validation. In Figure 3.12, it was clear that the mirTarDANN had the highest accuracy in data sets D3-3, D3-4, D2-2, D2-3, D2-4, D2-5 and D2-6, and the matched accuracy in data sets D2-1. The accuracy of mirTarDANN was lower than other predictors in data sets D4, D3-1 and D3-2. When comparing the sensitivity of mirTarDANN with other predictors, the mirTarDANN outperformed in almost all the eleven data sets, except data sets D3-2 and D2-1. In these two data sets, the sensitivity of these four predictors was very low but the accuracy of them was high. It may be because of the high specificity. The sensitivity of mirTarDANN was higher than other individual predictors in these two data sets. ComiR, a meta-predictor, achieved reasonably high or highest values on both sensitivity and accuracy. The comparison of F1 score of all the predictors shows the mirTarDANN achieved the significant improvement in data sets D3-3, D3-4 and another five D2 data sets (except D2-1), and equaled or similar scores in data sets D4, D3-1 and D3-2. The F1 score of ComiR beaten other predictors in data set D2-1. Overall, the averaged accuracy and sensitivity of mirTarDANN in the eleven datasets under multi-fold cross-validation were 59.5% and 59.0%, which were 3% and 9% higher than the predictor at the second position, ComiR. The averaged F1 score and MCC values of mirTarDANN in all the eleven data sets were 0.492 and 0.287, which were higher than 0.329 and 0.101 in ComiR, respectively. The detailed F1 scores and MCCs are shown in Table 3.6.
Figure 3.12. Comparison of prediction performance of different predictors in eleven newly-designed data sets under multi-fold cross validation. X-axis shows the eleven newly-designed data sets, while y-axis shows (A) accuracy, (B) sensitivity, and (C) F1 score, respectively. Error bars are standard deviation from multi-fold cross validation. In the D4 data set, the performance of mirTarDANN was compared with four individual predictors and ComiR. In each of the D3 series data sets, mirTarDANN was compared to ComiR, and three out of four individual predictors. In each of the D2 series data sets, mirTarDANN was compared to ComiR and two out of four individual predictors. When calculating the accuracy of individual predictors, their default cutoff values were used. For ComiR, a false discovery rate of 5% was recommended by the developer to determine the cutoff. Therefore, based on the calculations of 50 randomly selected miRNAs and their targets in the data sets, 0.82 was used as the cutoff of ComiR.
Table 3.6. F1 and Mathews Correlation Coefficient (MCC) in eleven training sets under multi-fold cross-validation.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>mRanda F1 (MCC)</th>
<th>mRDB F1 (MCC)</th>
<th>PITA F1 (MCC)</th>
<th>TargetScan F1 (MCC)</th>
<th>ComiR F1 (MCC)</th>
<th>miTarDANN F1 (MCC)</th>
</tr>
</thead>
<tbody>
<tr>
<td>D4</td>
<td>0.125±0.006 (0.001)±0.010</td>
<td>0.158±0.010 (0.003)±0.018</td>
<td>0.113±0.011 (0.005)±0.021</td>
<td>0.130±0.005 (0.001)±0.007</td>
<td>0.136±0.009 (0.002)±0.007</td>
<td>0.175±0.007 (0.008)±0.011</td>
</tr>
<tr>
<td>D3-1</td>
<td>0.198±0.015 (0.016)±0.019</td>
<td>0.248±0.006 (0.056)±0.007</td>
<td>0.176±0.031 (0.044)±0.040</td>
<td>~</td>
<td>~</td>
<td>0.272±0.006 (0.086)±0.016</td>
</tr>
<tr>
<td>D3-2</td>
<td>0.115±0.042 (0.004)±0.044</td>
<td>0.142±0.038 (0.021)±0.062</td>
<td>~</td>
<td>~</td>
<td>0.085±0.034 (0.068)±0.037</td>
<td>0.168±0.025 (0.069)±0.046</td>
</tr>
<tr>
<td>D3-3</td>
<td>0.131±0.012 (0.094)±0.014</td>
<td>~</td>
<td>~</td>
<td>0.206±0.027 (0.098)±0.047</td>
<td>0.151±0.011 (0.229)±0.019</td>
<td>0.325±0.022 (0.074)±0.022</td>
</tr>
<tr>
<td>D3-4</td>
<td>~</td>
<td>~</td>
<td>0.406±0.039 (0.098)±0.037</td>
<td>0.348±0.006 (0.017)±0.029</td>
<td>0.276±0.031 (0.252)±0.053</td>
<td>0.394±0.017 (0.138)±0.036</td>
</tr>
<tr>
<td>D2-1</td>
<td>0.239±0.044 (0.013)±0.047</td>
<td>0.324±0.026 (0.053)±0.019</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
</tr>
<tr>
<td>D2-2</td>
<td>0.223±0.060 (0.133)±0.057</td>
<td>~</td>
<td>~</td>
<td>0.368±0.040 (0.088)±0.061</td>
<td>~</td>
<td>~</td>
</tr>
<tr>
<td>D2-3</td>
<td>0.213±0.070 (0.094)±0.049</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>0.249±0.012 (0.393)±0.044</td>
</tr>
<tr>
<td>D2-4</td>
<td>~</td>
<td>~</td>
<td>0.521±0.046 (0.090)±0.058</td>
<td>0.501±0.023 (0.084)±0.049</td>
<td>~</td>
<td>~</td>
</tr>
<tr>
<td>D2-5</td>
<td>~</td>
<td>~</td>
<td>0.428±0.059 (0.043)±0.125</td>
<td>~</td>
<td>~</td>
<td>0.403±0.068 (0.481)±0.085</td>
</tr>
<tr>
<td>D2-6</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>0.527±0.016 (0.099)±0.042</td>
<td>0.474±0.032 (0.282)±0.082</td>
</tr>
</tbody>
</table>

N.b. The bold values are the highest value in that category. “∼” indicates the value of predictor is not applicable in that data set.
Figure 3.13 shows the comparison of accuracy, sensitivity, and F1 score among all component individual predictors, ComiR and mirTarDANN in their corresponding independent test sets. Overall, the trends of performances in independent test sets were similar to those in training sets. mirTarDANN had the best accuracy in almost all the data sets except data sets D4, D3-1, and D3-2. The sensitivity of mirTarDANN beaten other predictors in data sets D3-1, D3-3, D3-4, D2-2, and D2-4, while falling behind ComiR in data sets D4, D3-2, D2-1 and D2-3 and other individual predictors in data sets D2-1, D2-5, and D2-6. For the comparison of F1 scores, the mirTarDANN were only lower than ComiR in data sets D2-1 and D2-3. The averaged accuracy, sensitivity, F1 score, and MCC of mirTarDANN in all eleven data sets were 65.1%, 50.6%, 0.455, and 0.276, compared to 56.2%, 51.4%, 0.315, and 0.199 from ComiR. The detailed F1 scores and MCCs are displayed in Table 3.7.
Figure 3.13. Comparison of prediction performance of different predictors in the independent data sets that each is associated with one of the eleven newly-designed data sets. X-axis shows the data sets, while the y-axis shows (A) accuracy, (B) sensitivity, and (C) F1 score, respectively. Only mirTarDANN has error bars since mirTarDANN has multiple sets of parameters optimized under multi-fold cross-validation.
Table 3.7. F1 and Mathews Correlation Coefficient (MCC) in the eleven independent test sets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>miRanda</th>
<th>miRDB</th>
<th>PITA</th>
<th>TargetScan</th>
<th>ComiR</th>
<th>mirTarDANN</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F1</td>
<td>MCC</td>
<td>F1</td>
<td>MCC</td>
<td>F1</td>
<td>MCC</td>
</tr>
<tr>
<td>D4T</td>
<td>0.130</td>
<td>0.016</td>
<td>0.151</td>
<td>0.038</td>
<td>0.135</td>
<td>0.018</td>
</tr>
<tr>
<td>D3-1T</td>
<td>0.209</td>
<td>0.050</td>
<td>0.245</td>
<td>0.078</td>
<td>0.161</td>
<td>0.048</td>
</tr>
<tr>
<td>D3-2T</td>
<td>0.093</td>
<td>(0.044)</td>
<td>0.125</td>
<td>(0.017)</td>
<td>0.105</td>
<td>(0.042)</td>
</tr>
<tr>
<td>D3-3T</td>
<td>0.116</td>
<td>(0.108)</td>
<td>0.203</td>
<td>(0.090)</td>
<td>0.165</td>
<td>(0.193)</td>
</tr>
<tr>
<td>D3-4T</td>
<td>~</td>
<td>~</td>
<td>0.356</td>
<td>0.023</td>
<td>0.328</td>
<td>(0.039)</td>
</tr>
<tr>
<td>D2-1T</td>
<td>0.158</td>
<td>(0.065)</td>
<td><strong>0.340</strong></td>
<td><strong>0.145</strong></td>
<td>~</td>
<td>~</td>
</tr>
<tr>
<td>D2-2T</td>
<td>0.298</td>
<td>(0.078)</td>
<td>~</td>
<td>~</td>
<td>0.299</td>
<td>(0.153)</td>
</tr>
<tr>
<td>D2-3T</td>
<td>0.160</td>
<td>(0.142)</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
</tr>
<tr>
<td>D2-4T</td>
<td>~</td>
<td>~</td>
<td>0.463</td>
<td>0.046</td>
<td>0.451</td>
<td>(0.108)</td>
</tr>
<tr>
<td>D2-5T</td>
<td>~</td>
<td>~</td>
<td>0.530</td>
<td>0.110</td>
<td>0.451</td>
<td>(0.450)</td>
</tr>
<tr>
<td>D2-6T</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>0.478</td>
<td>0.051</td>
</tr>
</tbody>
</table>

N.b. The bold values are the highest value in that category. “~” indicates the value of predictor is not applicable in that data set.
3.6.2.4 Cross-subset Performance of MirTarDANN

The newly-designed meta-predictor, mirTarDANN, is composed of eleven sub-modules. Each sub-module is trained and validated using its corresponding training sets and independent test sets. The modules trained by the subsets which have less number of predicted scores are possible to make prediction for samples in subsets with more predicted scores. For example, the module DANN-2-1 can be used to predict samples in data set D4, D3-1 and D3-2. Samples in data set D4 can be predicted using all of the modules. Samples in data set D3-1 (including miRanda, miRDB, and PITA predicted scores) can be predicted using modules DANN-2-1, DANN-2-2, and DANN-2-4. The prediction performance of the trained DANN modules in the non-corresponding independent test sets are called cross-subsets accuracy. Figure 3.14 displays the performance of DANN modules in different subsets. Clearly, mirTarDANN has lower performance in cross-subset shown in Figure 3.14(A). It is expected that mirTarDANN has the lower performance in cross-subset since all the modules in mirTarDANN were trained using their corresponding data sets. Further evaluating the influence of different subsets on the performance of the modules, we merged all the associated subsets and re-trained the corresponding DANN modules using merged-subsets. For example, the data set D4 and D3-1 were merged to one subsets *D3-1, and the DANN-3-1 module was re-trained using merged-subsets *D3-1. The new re-trained module DANN-3-1 was evaluated using its original independent test sets and merge-subset independent test sets to determine whether the re-trained module using merged-subsets had better performances or not. The similar analysis was performed for modules DANN-3-x and DANN-2-x and the results are shown in Figure 3.14(B) and (C), accordingly. In Figure 3.14(B), the performance of each re-trained DANN module in its corresponding merged-subset was analyzed. In Figure 3.14(C), the performance of each re-trained DANN in its original associated subsets was analyzed. Clearly, even if a new DANN module was re-trained in the corresponding merged-subset,
Figure 3.14. Performance of eleven DANN modules in different subsets. (A) Performance of original trained DANN modules in different subsets. (B) Performance of DANN modules re-trained in merged subsets. (C) Performance of re-trained DANN modules in original subsets. X-axis shows the prediction accuracy, while y-axis shows the sensitivity. Each symbol stands for the performance of a specific subset predicted by specific DANN module. Larger symbols are performance of eleven original DANN modules trained in their corresponding subsets, while small symbols show performance of original DANN modules in non-corresponding subsets or performance of re-trained DANN modules in various subsets. The subsets are represented by the colors filled in the symbols as follows: D4:black, D3-1:red, D3-2:green, D3-3:yellow, D3-4:blue, D2-1:pink, D2-2:cyan, D2-3:grey, D2-4:dark red, D2-5:dark green, D2-6:dark yellow. DANN modules are denoted by the shape of symbols as follows (All DANN-3-x have dark edges, while all DANN-2x have red edges): DANN-4:star, DANN-3-1:up triangle, DANN-3-2:down triangle, DANN-3-3:square, DANN-3-4:diamond, DANN-2-1:up triangle, DANN-2-2:down triangle, DANN-2-3:square, DANN-2-4:diamond, DANN-2-5:hex, DANN-2-6:circle

the performance of this re-trained module was still less satisfactory in the merged-subset, as well as in the original associated subsets.

3.6.2.5 Diseases Association of Identified microRNA:mRNA

The newly-designed meta-predictor, mirTarDANN, can identify 298 experimental validated microRNA:mRNA pairs which cannot be identified by other component individual predictors. In those 298 new-identified pairs, 126 pairs can be identified using ComiR. The mirTarDANN can identify 172 novel microRNA:mRNA pairs in total. Figure 3.15(A) shows the result.
many of these 172 novel microRNA:mRNA validated pairs are from the same data set. For example, there are 43, 28, 24, 17, 16, and 14 interactions from data sets D3-3, D2-2, D3-4, D3-1, D4 and D2-4, respectively. The rest 30 pairs are from the other five data sets. There are 167 mRNAs and 82 microRNAs involved in the microRNA:mRNA interactions.

Among the 167 mRNAs, there are multiple mRNAs belonging to the same gene families, such as members in the AKAP, EIF, MED, SLC, TMEM, USP, ZBTB, ZFP, and ZMAT families. Several genes, such as ACO1, SMARCA2, SUMF1, TOX4, and ZMAT3 can be regulated by more than one microRNA. ACO1 is a cytoplasmic aconitate hydratase, which can bind with iron and serve as an aconitase to generate the isocitrate in the citric acid cycle when the cellular iron levels are high [180]. When the iron levels are low, ACO1 can bind with iron-responsive elements to act as a RNA-binding protein [181]. In the novel microRNA:mRNA interactions, we found the ACO1 can be regulated by mmu-miR-339 and mmu-miR-10B. In addition, the mRNA SMARCA2 acts as an ATP-dependent helicase and a global transcription activator. It works as a member for SWI/SNF chromatin remodelling complexes and npBAF complex, which is important for the neural stem cells development [182, 183, 184]. In our novel identified microRNA:mRNA interactions, the SMARCA2 was regulated by mmu-miR-33 and mmu-miR-466A. SUMF1 is a formylglycine-generating enzyme, which can cause the multiple sulfatase deficiency [185, 186]. SUMF1 is bound by mmu-miR-743B and mmu-miR-488 in the novel microRNA:mRNA predicted pairs. TOX4 is a component of PTW/PP1 phosphotase complex, which regulates chromatin structures. The interaction partners of TOX4 were mmu-miR-221 and mmu-miR-376C from the true positive predictions [187]. ZMAT3 is a Target of P53 and plays a critical roles in P53 dependent regulatory pathways [188, 189]. mirTarDANN found the additional regulated microRNA, mmu-miR-764 and mmu-miR-292A.

Among the 82 microRNAs, 36 microRNAs can regulate multiple mRNAs, out of which, mmu-miR-129, mmu-miR-340, mmu-miR-362, and mmu-miR-9 can regulate 12, 9, 8, and
8 mRNAs. These 37 downstream mRNAs are involved in over thirty signalling pathways, including MARK signalling pathway, cancer-related pathways, hippo signalling pathway, Ras signalling pathway, RNA transport pathway, spliceosome pathway, etc. Among those 37 downstream genes, PDGFRA and MYLK are involved in 18 and 8 signalling pathways, respectively. In Figure 3.15(B), there are 16 microRNAs which can regulate more than three novel-predicted mRNAs. The regulated signaling pathways and the number of genes that involved in each pathway of the 36 microRNAs are shown in Figure 3.15(C).

3.6.3 Discussion

Identifying microRNA target genes is critical for interpreting microRNA functions and for comprehending the regulation of gene expression. Although various experimental techniques are available, computational predictors are still in many cases preferred to provide fast preliminary results for further experimental validations, especially when the computational prediction provides additional information on microRNA:mRNA interactions. For these reasons, many microRNA target predictors have been designed. Nonetheless, many of these predictors are confronting emerging issues. First, almost all the predictors were optimized using small data sets that may only account for a small portion of the currently known microRNA:mRNA interactions. Therefore, the performances of these predictors in large data sets are not systematically tested. Second, since the predictors are developed using different small data sets, comparing the performances between predictors and comparing the pros-and-cons of various strategies used for different predictors are elusive. It is worth to note that the prediction accuracy is greatly affected by the data set. In the eleven data sets used in this project, all the duplicate samples were removed by the difference of their prediction scores. Since PITA and TargetScan produce narrow distribution for both positive and negative samples as a function of prediction score, both PITA and TargetScan may lose more samples compared to other predictors when a constant value is used as the cut-off for
Figure 3.15. Novel microRNA:mRNA interactions identified by mirTarDANN in the eleven newly-designed data sets. (A) Number of novel microRNA:mRNA interactions identified by mirTarDANN in the eleven newly-designed data sets. Grey bars show the number of novel interactions compared to four individual predictors, including miRanda, miRDB, PITA, and TargetScan. Dashed grey bars show the comparison with four individual predictors and ComiR. (B) List of microRNAs and the corresponding number of mRNAs that can be regulated by each of these microRNA. Only microRNAs having more than three mRNA targets are shown in the figure. (C) KEGG pathways containing three or more genes, which are found in the novel microRNA:mRNA interactions.
identifying duplicates. Consequently, the prediction accuracy of PITA and TargetScan may be underestimated in the eleven data sets (see Figure 3.10). Third, many predictors may not have sufficiently high enough prediction performance, especially these predictors may have high specificity but low sensitivity. All these issues bring in concerns on the application of these predictors in biological studies.

Meta-strategy has the capability to combine the results from different individual predictors and can improve the prediction performances of predictors, as well as balance sensitivity and specificity. However, the integrated outputs of individual predictors may not be improved by directly applying meta-strategy under specific cases. Therefore, in this section, a various set of techniques, including dual-threshold values, two-step significance voting, information gain, decision-tree, and deep neural networks, were used to integrate the outputs of individual predictors. We selected those individual predictors are based on multiple reasons. First, these predictors were well-designed and well-maintained, and widely used in this field. Second, the analysis of overlap and coverage of these predictors displayed that many predictors have their specific predictions. The usage of majority voting is critically based on the measurement of overlap. Meanwhile, for significant-voting based strategy, both overlap and coverage play critical roles for the newly-designed predictors. Most importantly, the majority-voting has its advantages on selecting the true-positive predictions which have high confidences, and significant-voting is usually the complementary of majority-voting to select additional true-positive predictions. In this way, the overall prediction performance can be improved. Dual-threshold value and two-step voting strategies were used to maximize the efficency of significant-voting. The selecting can be separated into two processes. In the first-process, the thresholds values are very stricter than those of single predictors. Therefore, the selected true-positive and true-negative predictions from individual predictors have high confidences. In the second process, the less-stringent threshold values were applied to identify additional true-positive predictions.
Another critical factor for the success of mirTarDANN is to split a large data set into eleven smaller predictor-specific data sets based on the prediction scores of individual predictors. This strategy makes it easy to compare the performance, as well as the overlap and coverage, of individual predictors and other predictors. This strategy also provides another protocol to improve prediction accuracy further by selecting predictors based on the prediction scores of four individual predictors. For example, if a microRNA:mRNA pair has miRanda, miRDB, and TargetScan scores (or in other words, this pair is similar to samples in the D3-2 data set), ComiR can be selected to make prediction to ensure the highest sensitivity, but miRanda should be used to ensure a higher specificity.

The new meta-strategy improved the prediction performance significantly, as demonstrated by multiple evaluation criteria (including, sensitivity, specificity, accuracy, F1 score, and MCC) under multi-fold cross-validation and in independent data sets as shown in Figures 3.12 and 3.13. More specifically, the prediction accuracy of the newly developed predictor is at least nine percentage points higher than other predictors, the F1 and MCC values are also improved by at least 40 percent. In addition, this newly proposed strategy identified over 170 microRNA:mRNA interactions, which cannot be identified by individual predictors or ComiR. Therefore, this predictor is able to find novel microRNA:mRNA interactions.

3.7 MirTarDANN Improve Human MicroRNA:mRNA Prediction Performance

In the paper “Significant improvement of microRNA target prediction accuracy in large data sets using meta-strategy based on comprehensive voting and artificial neural networks”, we found the integration of decision-tree-based dual-threshold value and significant-voting with ANN can improve the prediction performance of microRNA:mRNA interactions significantly. The same meta-predictor was re-trained using human microRNA:mRNA interaction databases. In this section, we introduced the data sets and the prediction performance of the new-trained modules in human microRNA:mRNA interactions.
3.7.1 Materials and Methods

The same procedure was applied to build eleven predictor-specific human microRNA:mRNA interaction data sets. The detailed information of these eleven data sets was shown in Table 3.8. To simplify the future data analysis, the score of miRDB was multiply by -1, and the score of all of the component predictors in each data set was normalized to range (-1,1). To determine the threshold values of the component predictors in each corresponding data set, the information gain and distribution of true samples and false samples were calculated. Multiple thresholds were selected and tested using different combinations. Finally, the threshold values of two-step were selected. For example, the final threshold values in data set D4 are (-0.772/0.359), (-0.407/0.016), (0.189/0.421), and (0.652/0.786) for the 1st-step, and (-0.047/0.237), (-0.321/-0.101), (0.239/0.263), and (0.710/0.750) for the 2nd-step for component predictors miRanda, miRDB, PITA, and TargetScan, respectively. Similarly, the threshold values for other ten data sets were selected and shown in Table 3.9. The performances of newly-designed meta-predictor were evaluated using Sensitivity (Sens), Specificity (Spec), Accuracy (Acc), F1 score (F1), and Matthews Correlation Coefficient (MCC) under multi-fold cross-validation in training sets and in independent test sets.
Table 3.8. Numbers of samples in each of the eleven newly-designed human data sets.

<table>
<thead>
<tr>
<th>Datasets</th>
<th>Associated predictors</th>
<th>No. of Original</th>
<th>No. of non-red.</th>
<th>No. of true</th>
<th>No. of False</th>
<th>No. of microRNA</th>
<th>No. of mRNA</th>
</tr>
</thead>
<tbody>
<tr>
<td>D4</td>
<td>miRanda.miRDB. PITA. TargetScan</td>
<td>5839</td>
<td>3748</td>
<td>1601</td>
<td>2047</td>
<td>27</td>
<td>2653</td>
</tr>
<tr>
<td>D3-1</td>
<td>miRanda.miRDB. PITA</td>
<td>7180</td>
<td>3221</td>
<td>1687</td>
<td>1534</td>
<td>26</td>
<td>1954</td>
</tr>
<tr>
<td>D3-2</td>
<td>miRanda.miRDB. TargetScan</td>
<td>1356</td>
<td>842</td>
<td>342</td>
<td>500</td>
<td>26</td>
<td>651</td>
</tr>
<tr>
<td>D3-3</td>
<td>miRanda.PITA. TargetScan</td>
<td>221,014</td>
<td>1145</td>
<td>855</td>
<td>290</td>
<td>202</td>
<td>1014</td>
</tr>
<tr>
<td>D3-4</td>
<td>miRDB.PITA. TargetScan</td>
<td>20,134</td>
<td>2386</td>
<td>775</td>
<td>1612</td>
<td>107</td>
<td>1991</td>
</tr>
<tr>
<td>D2-1</td>
<td>miRanda.miRDB</td>
<td>1718</td>
<td>474</td>
<td>229</td>
<td>245</td>
<td>25</td>
<td>402</td>
</tr>
<tr>
<td>D2-2</td>
<td>miRanda.PITA</td>
<td>262,508</td>
<td>934</td>
<td>717</td>
<td>217</td>
<td>190</td>
<td>801</td>
</tr>
<tr>
<td>D2-3</td>
<td>miRanda.TargetScan</td>
<td>82,628</td>
<td>902</td>
<td>671</td>
<td>231</td>
<td>202</td>
<td>693</td>
</tr>
<tr>
<td>D2-4</td>
<td>miRDB.PITA</td>
<td>5582</td>
<td>620</td>
<td>282</td>
<td>338</td>
<td>69</td>
<td>556</td>
</tr>
<tr>
<td>D2-5</td>
<td>miRDB.TargetScan</td>
<td>20,832</td>
<td>902</td>
<td>520</td>
<td>381</td>
<td>111</td>
<td>834</td>
</tr>
<tr>
<td>D2-6</td>
<td>PITA.TargetScan</td>
<td>289,730</td>
<td>462</td>
<td>315</td>
<td>147</td>
<td>188</td>
<td>402</td>
</tr>
</tbody>
</table>
Table 3.9. The 1st-step and 2nd-step threshold values for both true and false predictions in eleven modules of newly-designed human microRNA:mRNA meta-predictor.

<table>
<thead>
<tr>
<th>Module</th>
<th>1st-step threshold</th>
<th>2nd-step threshold</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>miRanda (0.772)/0.359</td>
<td>miRanda (0.047)/0.237</td>
</tr>
<tr>
<td></td>
<td>miRDB (0.407)/0.016</td>
<td>miRDB (0.321)/0.101</td>
</tr>
<tr>
<td></td>
<td>PITA 0.189/0.421</td>
<td>TargetScan 0.239/0.263</td>
</tr>
<tr>
<td></td>
<td>TargetScan 0.652/0.7858</td>
<td>PITA 0.710/0.750</td>
</tr>
<tr>
<td>D4</td>
<td>~~~</td>
<td>~~~</td>
</tr>
<tr>
<td>D3-1</td>
<td>(0.772)/(0.041)</td>
<td>(0.530)/(0.205)</td>
</tr>
<tr>
<td></td>
<td>(0.508)/0.151</td>
<td>(0.111)/(0.041)</td>
</tr>
<tr>
<td></td>
<td>(0.073)/0.247</td>
<td>0.091/0.163</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>~~~</td>
</tr>
<tr>
<td>D3-2</td>
<td>(0.069)/0.459</td>
<td>(0.379)/0.736</td>
</tr>
<tr>
<td></td>
<td>(0.427)/(0.015)</td>
<td>0.083/0.245</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>(0.377)/(0.095)</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>0.544/0.654</td>
</tr>
<tr>
<td>D3-3</td>
<td>(0.550)/0.269</td>
<td>(0.381)/0.610</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>0.091/0.163</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>~~~</td>
</tr>
<tr>
<td>D3-4</td>
<td>~~~</td>
<td>0.229/0.738</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>(0.193)/0.427</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>0.826/0.878</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>~~~</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>0.365/0.423</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>(0.151)/(0.115)</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>0.844/0.858</td>
</tr>
<tr>
<td>D2-1</td>
<td>(0.365)/0.233</td>
<td>(0.321)/0.233</td>
</tr>
<tr>
<td></td>
<td>(0.441)/(0.007)</td>
<td>(0.361)/(0.237)</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>~~~</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>~~~</td>
</tr>
<tr>
<td>D2-2</td>
<td>(0.582)/0.359</td>
<td>(0.309)/(0.215)</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>~~~</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>0.325/0.389</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>~~~</td>
</tr>
<tr>
<td>D2-3</td>
<td>(0.550)/0.427</td>
<td>~~~</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>~~~</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>0.868/0.938</td>
</tr>
<tr>
<td></td>
<td>(0.694)/(0.666)</td>
<td>~~~</td>
</tr>
<tr>
<td></td>
<td>(0.329)/(0.321)</td>
<td>~~~</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>~~~</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>0.896/0.918</td>
</tr>
<tr>
<td>D2-4</td>
<td>~~~</td>
<td>(0.764)/(0.580)</td>
</tr>
<tr>
<td></td>
<td>(0.460)/(0.329)</td>
<td>~~~</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>~~~</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>(0.694)/(0.666)</td>
</tr>
<tr>
<td></td>
<td>(0.329)/(0.321)</td>
<td>~~~</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>~~~</td>
</tr>
<tr>
<td>D2-5</td>
<td>~~~</td>
<td>(0.472)/0.786</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>0.718/0.874</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>0.157/0.580</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>~~~</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>0.780/0.830</td>
</tr>
<tr>
<td>D2-6</td>
<td>~~~</td>
<td>~~~</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>(0.596)/0.596</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>0.868/0.960</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>~~~</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>(0.207)/0.472</td>
</tr>
<tr>
<td></td>
<td>~~~</td>
<td>0.900/0.922</td>
</tr>
</tbody>
</table>

N.b. Numbers above and below the slashes are for positive and negative predictions, respectively. Values inside parentheses are negative values. “~~~” indicates the threshold values are not applicable in that module.
3.7.2 Results

The accuracy, sensitivity and F1 score of individual predictors, miRanda, miRDB, PITA, and TargetScan, and newly-designed meta-predictor under multi-fold cross-validation were displayed in Figure 3.16. Comparing the accuracy of all the predictors in eleven human microRNA:mRNA data sets, the newly-designed meta-predictor outperformed other predictors in data sets in almost all of the data sets, except data set D2-1. Comparing the sensitivity of all the predictors, the newly-designed meta-predictor had higher sensitivities in data sets D3-1, D3-3, D3-4, and all the D2 series data sets. In data sets D4 and D3-2, the sensitivity of miRDB was higher than any other individual predictors and the newly-designed meta-predictors. Comparing the F1 score of all the predictors, the newly-designed meta-predictor beaten other predictors in data sets D3-1, D3-3, D3-4, D2-2, D2-3, D2-4, D2-5, and D2-6, while miRDB’s F1 score was at the first rank in data sets D4, D3-2, and D2-1. The averaged accuracy of newly-designed meta-predictor among eleven data sets under multi-fold cross-validation was 67.1%, which was 12% higher than miRDB, the predictor at the second position for accuracy, and the averaged sensitivity was 69.0%, which was 13% higher than PITA, the predictor at the second position for sensitivity. The F1 and MCC values of newly-designed meta-predictor as an average of all the eleven data sets were 0.670 and 0.277, which was higher than F1 score 0.495 from PITA and MCC value 0.189 from TargetScan, respectively. The F1 score and MCC score for all the eleven data sets are shown in Table 3.10.
Figure 3.16. Comparison of prediction performance of different predictors in the human microRNA:mRNA interaction training data sets under multi-fold cross-validation. X-axis shows the data sets, while the y-axis shows (A) accuracy, (B) sensitivity, and (C) F1 score, respectively.
### Table 3.10. F1 and Mathews Correlation Coefficient (MCC) in human microRNA:mRNA training sets under multi-fold cross-validation.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>miRanda F1, MCC</th>
<th>miRDB F1, MCC</th>
<th>PITA F1, MCC</th>
<th>TargetScan F1, MCC</th>
<th>new meta-predictor F1, MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>D4</td>
<td>0.391±0.024, 0.076±0.030</td>
<td>0.518±0.021, 0.178±0.035</td>
<td>0.410±0.010, 0.040±0.032</td>
<td>0.370±0.022, 0.083±0.042</td>
<td>0.490±0.023, 0.183±0.017</td>
</tr>
<tr>
<td>D3-1</td>
<td>0.396±0.015, 0.058±0.034</td>
<td>0.513±0.025, 0.136±0.031</td>
<td>0.505±0.024, 0.163±0.036</td>
<td>~</td>
<td>~</td>
</tr>
<tr>
<td>D3-2</td>
<td>0.335±0.074, 0.049±0.078</td>
<td>0.475±0.054, 0.151±0.103</td>
<td>~</td>
<td>~</td>
<td>~</td>
</tr>
<tr>
<td>D3-3</td>
<td>0.367±0.052, 0.066/0.046</td>
<td>~</td>
<td>~</td>
<td>0.546±0.035, 0.050±0.057</td>
<td>0.445±0.018, 0.163±0.060</td>
</tr>
<tr>
<td>D3-4</td>
<td>~</td>
<td>~</td>
<td>0.364±0.029, 0.023±0.028</td>
<td>0.362±0.020, 0.066±0.035</td>
<td>0.251±0.022, 0.185±0.023</td>
</tr>
<tr>
<td>D2-1</td>
<td>0.386±0.067, 0.057±0.102</td>
<td>0.525±0.015, 0.142±0.044</td>
<td>~</td>
<td>~</td>
<td>~</td>
</tr>
<tr>
<td>D2-2</td>
<td>0.407±0.023, 0.071±0.033</td>
<td>~</td>
<td>~</td>
<td>0.510±0.029, 0.053±0.025</td>
<td>~</td>
</tr>
<tr>
<td>D2-3</td>
<td>0.417±0.015, 0.044±0.028</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
</tr>
<tr>
<td>D2-4</td>
<td>~</td>
<td>~</td>
<td>0.407±0.083, 0.089±0.070</td>
<td>0.497±0.018, 0.019±0.007</td>
<td>~</td>
</tr>
<tr>
<td>D2-5</td>
<td>~</td>
<td>~</td>
<td>0.468±0.036, 0.064±0.032</td>
<td>~</td>
<td>~</td>
</tr>
<tr>
<td>D2-6</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
</tr>
</tbody>
</table>

N.b. The bold values are the highest value in that category. “~” indicates the value of predictor is not applicable in that data set.
In Figure 3.17, it displays the accuracy, sensitivity and F1 score of four individual predictors and the newly-designed meta-predictor. Normally, the newly-designed meta-predictor had the highest accuracy in almost all of the independent test sets, except D4, D3-2, D2-1 and D2-4. In the terms of sensitivity, meta-predictor outperformed other predictors in D4, D3-1, D3-3, and all the D2 series independent test sets. The averaged accuracy and sensitivity of meta-predictor in independent test sets were 64.9% and 63.0%, which were around 7% and 20% higher than miRDB, the predictor at the second position. In addition, the F1 score was 0.626 compared to 0.469 from PITA, and the MCC value was 0.246 compared to 0.212 from TargetScan. The F1 score and MCC score for all the eleven independent test sets are shown in Table 3.11.
Figure 3.17. Comparison of prediction performance of different predictors in the human microRNA:mRNA interaction independent test sets that each is associated with one of the eleven newly-designed data sets. X-axis shows the data sets, while the y-axis shows (A) accuracy, (B) sensitivity, and (C) F1 score, respectively. Only newly-designed meta-predictor has error bars since it has multiple sets of parameters optimized under multi-fold cross-validation.
Table 3.11. F1 and Mathews Correlation Coefficient (MCC) in human microRNA:mRNA independent test sets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>miRanda F1</th>
<th>miRanda MCC</th>
<th>miRDB F1</th>
<th>miRDB MCC</th>
<th>PITA F1</th>
<th>PITA MCC</th>
<th>TargetScan F1</th>
<th>TargetScan MCC</th>
<th>mirTarDANN F1</th>
<th>mirTarDANN MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>D4T</td>
<td>0.399</td>
<td>0.072</td>
<td>0.494</td>
<td><strong>0.153</strong></td>
<td>0.437</td>
<td>0.084</td>
<td>0.368</td>
<td>0.100</td>
<td><strong>0.499±0.011</strong></td>
<td><strong>0.138±0.021</strong></td>
</tr>
<tr>
<td>D3-1T</td>
<td>0.480</td>
<td>0.148</td>
<td>0.483</td>
<td>0.084</td>
<td>0.500</td>
<td>0.172</td>
<td>~</td>
<td>~</td>
<td><strong>0.648±0.011</strong></td>
<td><strong>0.216±0.018</strong></td>
</tr>
<tr>
<td>D3-2T</td>
<td>0.182</td>
<td>0.147</td>
<td><strong>0.470</strong></td>
<td><strong>0.196</strong></td>
<td>~</td>
<td>~</td>
<td>0.316</td>
<td>0.069</td>
<td>0.387±0.109</td>
<td>0.148±0.042</td>
</tr>
<tr>
<td>D3-3T</td>
<td>0.393</td>
<td>0.083</td>
<td>~</td>
<td>~</td>
<td>0.538</td>
<td>0.032</td>
<td>0.443</td>
<td><strong>0.204</strong></td>
<td><strong>0.688±0.016</strong></td>
<td>0.118±0.060</td>
</tr>
<tr>
<td>D3-4T</td>
<td>~</td>
<td>~</td>
<td>0.302</td>
<td>0.067</td>
<td>0.361</td>
<td>0.061</td>
<td>0.226</td>
<td>0.225</td>
<td><strong>0.476±0.014</strong></td>
<td><strong>0.246±0.017</strong></td>
</tr>
<tr>
<td>D2-1T</td>
<td>0.277</td>
<td>0.048</td>
<td>0.481</td>
<td>0.118</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>0.473±0.059</td>
<td>0.052±0.019</td>
</tr>
<tr>
<td>D2-2T</td>
<td>0.339</td>
<td>0.036</td>
<td>~</td>
<td>~</td>
<td>0.507</td>
<td>0.015</td>
<td>~</td>
<td>~</td>
<td><strong>0.865±0.011</strong></td>
<td><strong>0.473±0.028</strong></td>
</tr>
<tr>
<td>D2-3T</td>
<td>0.402</td>
<td>0.025</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>0.538</td>
<td><strong>0.285</strong></td>
<td><strong>0.758±0.025</strong></td>
<td>0.264±0.032</td>
</tr>
<tr>
<td>D2-4T</td>
<td>~</td>
<td>~</td>
<td>0.484</td>
<td>0.164</td>
<td>0.400</td>
<td>0.035</td>
<td>~</td>
<td>~</td>
<td><strong>0.511±0.026</strong></td>
<td><strong>0.168±0.036</strong></td>
</tr>
<tr>
<td>D2-5T</td>
<td>~</td>
<td>~</td>
<td>0.530</td>
<td>0.169</td>
<td>~</td>
<td>~</td>
<td>0.358</td>
<td>0.443</td>
<td><strong>0.783±0.008</strong></td>
<td><strong>0.482±0.023</strong></td>
</tr>
<tr>
<td>D2-6T</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>~</td>
<td>0.542</td>
<td>0.065</td>
<td>0.613</td>
<td>0.158</td>
<td><strong>0.819±0.010</strong></td>
<td><strong>0.409±0.046</strong></td>
</tr>
</tbody>
</table>

N.b. The bold values are the highest value in that category. “~” indicates the value of predictor is not applicable in that data set.
3.7.3 Discussion

We applied human microRNA:mRNA interaction data sets to train and validate the newly-designed meta-predictor, mirTarDANN, and found the improvement of microRNA and mRNA interaction prediction in these human data sets. Specifically, the prediction sensitivity and accuracy are significantly improved compared with other component individual predictors. The F1 score and MCC value of the newly-designed meta-predictor are also improved in most of the data sets. These results have the similar trends as in the mouse data sets. It indicates the newly-designed meta-predictor has the advantage to balance the sensitivity and specificity to obtain the higher evaluation results (e.g. F1 and MCC). Moreover, it indicates the meta-strategy can be extended to renovate other exist predictors.

3.8 Decision-tree Based Meta-Strategy Improved Accuracy of Disorder Prediction and Identified Novel Disordered Residues Inside Binding Motifs

We extended the newly-designed meta-strategy with artificial neural networks to other biological fields, such as microRNA prediction, intrinsic disorder protein prediction, disorder protein binding region prediction, etc., and found this newly-designed strategy has its own advantages and disadvantages. We published paper “Decision-tree Based Meta-Strategy Improved Accuracy of Disorder Prediction and Identified Novel Disordered Residues Inside Binding Motifs” on International Journal of Molecular Sciences. In this paper, we found the application of newly-designed meta-strategy can improve the prediction performance of disorder protein prediction. The results were shown as following.
3.8.1 Materials and Methods

3.8.1.1 Databases

The disorder protein data set was established by combining both proteins from DisProt 7.0 [190] database and PDB (Protein DataBank) [191]. DisProt contains more than 800 protein sequences, which have the identified Intrinsically disordered amino acid (IDAAs) or intrinsically disordered regions (IDRs) [190]. Those IDAAs/IDRs were identified by various experiments, including X-ray, NMR, circular dichroism (CD) spectrometry proteolysis, and etc. [190]. The annotations of proteins in DisProt were used to label the sequences. The high-resolution structured proteins were collected from PDB database, which have less than 2.5 angstrom resolution and more than 30 residues length [191]. The sequences which have abnormal residues or more than 30 missing residues were removed from list. After that, randomly select 20% of sequences from the PDB database for further analysis. The missing residues in PDB sequences were labeled as disordered residues, otherwise, the residues were treated as structured residues. The sequences from both DisProt database and PDB database are merged and filtered using CD-HIT [192] to remove the sequences which the identify is more than 30%. The final obtained data set included 312 protein sequences containing 30,140 disorder residues and 75,945 structured residues. All the sequences in CASP10 [193] were extracted and aligned with the previous mentioned 312 sequences. The identify lower than 30% were used as the second independent test set. There were 35 sequences in the second independent test set.

3.8.1.2 Infrastructure of Meta-strategy

Figure 3.18 displays the infrastructure of meta-strategy. All the sequences in the established data sets were predicted using DisEMBL [194], IUPred [195], ESpritz [196], and VSL2 [197]. The reasons to select those predictors are shown as follow: (1) Different strategies were
applied to generate the predictors. For instances, IUPred applies knowledge-based potential interaction, DisEMBL is based on artificial neural networks, ESPrritz applied bidirectional recursive neural network (BRNN) and was trained separately on N-terminal, C-terminal, and the general sequences, and VSL2 uses neural networks on different lengths of sequences DisEMBL [194, 195, 196, 197]. (2) All of these predictors obtain higher predicted accuracy relatively. (3) These predictors have well-maintained web-servers or released implementations for a large scale sequences prediction. The predicted results of those four individual predictors were used as input to feed into a decision-tree-based artificial neural network (DBann) to make the final prediction. DBann has two separated processes. In the first process, the techniques, dual-threshold, significant-voting, two-step selection are included. The output of the first processed will add six category dimensions to input. The second process of DBann is two-hidden layer fully-connected artificial neural network. The input of the second process has ten dimensions. The output layer’s activation function is Soft matrix, which can effectively transform the data into two dimensions. Therefore, the final output of DBann will have two nodes.

3.8.1.3 Train and Validate Meta-strategy

The 312 sequences data set were separated into two subsets, training and test sets. The training set was established by randomly selecting 80% residues of those sequences, and the rest 20% residues were used as independent test set. The ratios of true residues (disordered residues) to false residues (structured residues) in those separated two data sets were roughly the same. Five-fold cross-validation was applied on training set to generate model. In brief, four of five subsets were used to train the predictor and the rest one subset was used to evaluate the overfitting and validate the final prediction performance. The final predicted performance was the average of the results from the five-fold processes. At last, the generated predictor was evaluated using the independent test set.
Figure 3.18. Infrastructure of the new meta-strategy. “NT” and “NF” are the numbers of predictors making true prediction and false prediction, respectively. “a₁” and “a₂” are the differences of prediction score from the 1st-step and the 2nd-step threshold values, respectively. The letter subscripts represent DisEMBL(D), IUPred(I), VSL2(V), and ESpritz(E), accordingly. “dT” and “dF” are Euclidean distances of prediction scores from their corresponding threshold values for true predictions and false prediction, accordingly.
The generated model was evaluated using Sensitivity (Sens), Specificity (Spec), Accuracy (Acc), balanced Accuracy (Acc-b, the average of sensitivity and specificity), F1 score (F1), Matthews Correlation Coefficient (Mcc), Area Under ROC curve (AUC_ROC), and Area Under precision-recall curve (AUC_PR) in either training set or independent test sets. The performance of the new designed meta-predictor was compared with four individual disordered predictors, including IUPred [194], DisEMBL [195], ESpritz [196], and VSL2 [197], as well as another four recently developed predictors, including PONDR-FIT [169], MFDp2 [198], IUPred2A [199], and AUCpreD [200].

### 3.8.1.4 Information Gain

Information gain (IG) is a novel technique, which is usually used to select threshold value for decision-tree-based model. The peak value of IG represents the selected feature that can provide the highest information related with target under that threshold value. The IG was calculated using the followed formula.

\[
IG(x) = \sum_{i=1,2} p_i \log_2 p_i - \sum_{j=1,2} f_j(x) \sum_{k=1,2} p_{j,k} \log_2 p_{j,k}
\]

Where, \( p_i \) represents the fraction of true samples (\( p_1 \)) and false samples (\( p_2 \)). \( x \) is the scores used to split the data set into positive predicted or negative predicted groups. \( f_j(x) \) is the fraction of samples with prediction score higher than the threshold (\( j_1 \)) or the fraction of samples with prediction score lower than the threshold (\( j_2 \)). \( p_{j,k} \) is the fraction of true positive samples \( p_{j,1} \) and true negative samples \( p_{j,2} \). \( j \) represents the \( j \)-th group.

### 3.8.2 Results

Figure 3.19 (A) displays the Receiver Operating Characteristic (ROC) curves of these six component predictors. The Area Under the Curve (AUC) of predictors IUPred (I),
Figure 3.19. Prediction performance of four component predictors, including DisEMBL, IUPred, and ESPritez. (A) ROC curves of four component predictors. The ROC curves were obtained by using the default settings of these predictors. (B) The pairwise overlap (gray bars) and coverage (dashed bars) for true positive predictions (upper panel) and true negative predictions (lower panel) between each pair of predictors. Axis shows pairs of predictors as follows: D-DisEMBL, I-IUPred, V-VSL2, and E-ESPritez. All-coverage on x-axis stands for the maximum coverage all predictors.

DisEMBL (D), ESPritez (E), and VSL2 (V) were 0.78, 0.82, 0.84, 0.88, respectively. Figure 3.19 (B) displays the overlap and coverage between any two component predictors in the true samples (disordered residues) and false samples (structured residues). The overlap represents the ratio of true-positive or true-negative predicted samples from both predictors to the total number of positive samples or negative samples, respectively. The overlap between any two component predictors was normally in range 30% to 50%, while the coverage between any two component predictors was normally in range ~60% to ~80% in true samples. For the false samples, the overlap was usually in range ~70% to ~90%, and the coverage was normally higher than ~90%. These comparison values indicated the theoretical probability to improve the prediction performance by combining those four predictors.
The information gain obtained threshold values for each predictor are integrated into the meta-strategy to train the model. These threshold values are around 0.5, 0.52, 0.64, 0.26, respectively. The performance of new meta-predictor was evaluated using Sensitivity (Sens), Specificity (Spec), Accuracy (Acc), balanced Accuracy (Acc-b, the average of sensitivity and specificity), F1 score (F1), Matthews Correlation Coefficient (Mcc), Area Under ROC curve (AUC_ROC), and Area Under precision-recall curve (AUC_PR). It was obvious that the performance of our new designed predictor was better than other component predictors. Table 3.12 shows the performance of this new predictor and other predictors in training set. The Acc, Balance-Acc, F1, Mcc, ROC_AUC, and PR_AUC of the new meta-strategy achieved 84.2%, 83.1%, 0.635, 0.744, 0.899, and 0.788, respectively, and was ranked as the first place compared with other eight predictors. The performance of the new meta-strategy and other six component predictors were evaluated in independent test set. The result are shown in Table 3.13. Combining the results from Table 3.12 and 3.13, it was obvious to conclude that the overall levels and trends of predictors were similar.

The performance of new meta-predictor for twenty types of amino acids was compared with other eight different disorder predictors. The balance-Acc was shown in Figure 3.20. The new meta-predictor had the highest balance-Acc in twenty types of amino acids, and had the similar first ranked place with MFDp2 for residues P and Q in training set. For the residue C, N and Y, the newly-designed meta-strategy was ranked lower than the recently-designed predictor, MFDp2.
Table 3.12. Prediction performance of the new strategy under five-fold cross-validation, in comparison with four component predictors, another four recently-designed predictors.

<table>
<thead>
<tr>
<th></th>
<th>DisEMBL</th>
<th>IUPred</th>
<th>VSL2</th>
<th>Espritz</th>
<th>PONDR-FIT</th>
<th>MFDp2</th>
<th>IUPred2</th>
<th>AUCpreD</th>
<th>This Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sens</td>
<td>0.440±0.008</td>
<td>0.650±0.003</td>
<td><strong>0.817±0.004</strong></td>
<td>0.514±0.009</td>
<td>0.713±0.004</td>
<td>0.777±0.004</td>
<td>0.640±0.003</td>
<td>0.592±0.006</td>
<td><strong>0.807±0.012</strong></td>
</tr>
<tr>
<td>Spec</td>
<td>0.914±0.002</td>
<td>0.874±0.004</td>
<td>0.736±0.003</td>
<td><strong>0.939±0.002</strong></td>
<td>0.859±0.004</td>
<td>0.859±0.004</td>
<td>0.877±0.004</td>
<td>0.909±0.002</td>
<td><strong>0.856±0.007</strong></td>
</tr>
<tr>
<td>Acc</td>
<td>0.779±0.003</td>
<td>0.810±0.003</td>
<td>0.759±0.002</td>
<td>0.818±0.004</td>
<td>0.817±0.004</td>
<td>0.836±0.003</td>
<td>0.810±0.003</td>
<td>0.819±0.003</td>
<td><strong>0.842±0.003</strong></td>
</tr>
<tr>
<td>Acc-b</td>
<td>0.677±0.006</td>
<td>0.762±0.002</td>
<td>0.776±0.002</td>
<td>0.726±0.004</td>
<td>0.786±0.003</td>
<td>0.818±0.003</td>
<td>0.759±0.002</td>
<td>0.751±0.003</td>
<td><strong>0.831±0.004</strong></td>
</tr>
<tr>
<td>MCC</td>
<td>0.410±0.007</td>
<td>0.529±0.006</td>
<td>0.504±0.004</td>
<td>0.521±0.007</td>
<td>0.561±0.007</td>
<td>0.614±0.006</td>
<td>0.526±0.006</td>
<td>0.535±0.006</td>
<td><strong>0.635±0.006</strong></td>
</tr>
<tr>
<td>F1</td>
<td>0.531±0.006</td>
<td>0.660±0.003</td>
<td>0.658±0.003</td>
<td>0.616±0.006</td>
<td>0.689±0.004</td>
<td>0.729±0.003</td>
<td>0.657±0.004</td>
<td>0.651±0.005</td>
<td><strong>0.744±0.004</strong></td>
</tr>
<tr>
<td>AUC_ROC</td>
<td>0.776±0.004</td>
<td>0.823±0.001</td>
<td>0.841±0.003</td>
<td>0.886±0.003</td>
<td>0.857±0.003</td>
<td>0.879±0.002</td>
<td>0.822±0.001</td>
<td>0.869±0.003</td>
<td><strong>0.899±0.004</strong></td>
</tr>
<tr>
<td>AUC_PR</td>
<td>0.607±0.007</td>
<td>0.675±0.007</td>
<td>0.656±0.020</td>
<td>0.752±0.006</td>
<td>0.696±0.004</td>
<td>0.629±0.006</td>
<td>0.657±0.004</td>
<td>0.716±0.007</td>
<td><strong>0.788±0.010</strong></td>
</tr>
</tbody>
</table>

N.b. The measures of predictor performance include: sensitivity (Sens), Specificity (Spec), accuracy (Acc), balanced accuracy (Acc-b), Matthews Correlation (MCC), Area under ROC Curve (AUC_ROC) and Area Under Precision-Recall Curve (AUC_PR). The highest value in each measures is in bold.
Table 3.13. Prediction performance of all nine predictors in the independent data set.

<table>
<thead>
<tr>
<th></th>
<th>DisEMBL</th>
<th>IUPred</th>
<th>VSL2</th>
<th>Espritz</th>
<th>PONDR-FIT</th>
<th>MFDp2</th>
<th>IUPred2</th>
<th>AUCpreD</th>
<th>This Work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sens</td>
<td>0.454</td>
<td>0.656</td>
<td><strong>0.820</strong></td>
<td>0.529</td>
<td>0.728</td>
<td>0.780</td>
<td>0.647</td>
<td>0.609</td>
<td>0.811±0.007</td>
</tr>
<tr>
<td>Spec</td>
<td>0.915</td>
<td>0.872</td>
<td>0.735</td>
<td><strong>0.932</strong></td>
<td>0.856</td>
<td>0.857</td>
<td>0.870</td>
<td>0.908</td>
<td>0.856±0.006</td>
</tr>
<tr>
<td>Acc</td>
<td>0.784</td>
<td>0.811</td>
<td>0.759</td>
<td>0.818</td>
<td><strong>0.820</strong></td>
<td>0.835</td>
<td>0.811</td>
<td>0.823</td>
<td><strong>0.844±0.003</strong></td>
</tr>
<tr>
<td>Acc-b</td>
<td>0.684</td>
<td>0.764</td>
<td>0.777</td>
<td>0.792</td>
<td>0.819</td>
<td>0.761</td>
<td>0.759</td>
<td></td>
<td><strong>0.834±0.001</strong></td>
</tr>
<tr>
<td>MCC</td>
<td>0.424</td>
<td>0.532</td>
<td>0.507</td>
<td>0.521</td>
<td>0.569</td>
<td>0.615</td>
<td>0.530</td>
<td>0.530</td>
<td><strong>0.639±0.003</strong></td>
</tr>
<tr>
<td>F1</td>
<td>0.544</td>
<td>0.663</td>
<td>0.659</td>
<td>0.622</td>
<td>0.696</td>
<td>0.729</td>
<td>0.660</td>
<td>0.660</td>
<td><strong>0.747±0.002</strong></td>
</tr>
<tr>
<td>AUC_ROC</td>
<td>0.779</td>
<td>0.824</td>
<td>0.841</td>
<td>0.888</td>
<td>0.857</td>
<td>0.880</td>
<td>0.822</td>
<td>0.872</td>
<td><strong>0.900±0.002</strong></td>
</tr>
<tr>
<td>AUC_PR</td>
<td>0.617</td>
<td>0.673</td>
<td>0.642</td>
<td>0.754</td>
<td>0.695</td>
<td>0.622</td>
<td>0.672</td>
<td>0.720</td>
<td><strong>0.789±0.005</strong></td>
</tr>
</tbody>
</table>

N.b. The new strategy was optimized five times under five-fold cross-validation. The performance was also tested in the independent test data set five times. The highest value in each measures is in bold.
Figure 3.20. Comparison of balanced accuracy for twenty types of amino acids. X-axis shows amino acid types in the alphabetic order, while the y-axis shows the value of balanced accuracy. For each type of amino acid, the predictors from left to right are: DisEMBL, IUPred, VSL2, ESpritz, PONDR-FIT, MFDp2, IUPred2A, and AUCpre.

The per-sequence accuracy of newly-designed meta-predictor and other three predictors, PONDR-FIT, MFDp2, and AUCpreD in CASP10, were analyzed and shown in Figure 3.21 (A). All the symbols above the diagonal line represent sequences with higher accuracy when predicted using PONDR-FIT, MFDp2, or AUCpreD. The symbols in the dash circle display the prediction accuracy of all the predictors are not very well. The symbols in the dash box represent another group of sequences, which the accuracy of the newly-designed meta-predictor is higher than other predictors. Overall, PONDR-FIT was ranked at the first place for per-sequence prediction in CASP10. The newly-designed meta-predictor and AUCpreD had the similar prediction performances on per-sequence prediction. Since the newly-designed meta-predictor made a very low-accuracy prediction on some sequences, the analysis of the possible reason is necessary. The analysis of the relationship between balanced accuracy and the fraction of experimentally validated IDAAs per-sequence was shown in Figure 3.21(B). The fraction of IDAAs is critical for the performance of the newly-designed meta-predictor.
Figure 3.21. (A) Comparison of per-sequence balanced accuracy among AUCpreD (filled circle), PONDR-FIT (open circle), MFDp2 (filled triangle), and this work on sequences in the CASP10 test data set. The reasons for selecting these predictors are: (1) they are developed in recent years; (2) they have higher performance on some of the accuracy measures; (3) for simplicity of visualization, only four predictors were selected. X-axis shows the per-sequence balanced accuracy of this work, and y-axis shows the per-sequence accuracy of the other three predictors. (B) Per-sequence balanced accuracy of this work (y-axis) as a function of the fraction of experimentally validated intrinsically disordered (IDAAs) (x-axis). The size of the symbol is proportional to the length of the sequence.
3.8.3 Discussion

This newly-designed meta-strategy by integrating several novel techniques, such as dual-threshold, two-step voting, decision-tree, and artificial neural networks, has its several advantages: (1) integrating lower-accuracy predictors is able to produce higher-accuracy output; (2) the improvement of prediction performance of meta-strategy is significant and impressive, compared to individual predictors and other state-of-the-art predictors, including deep-learning based predictors; (3) the meta-strategy has well-balanced results for sensitivity and specificity, and therefore, is able to achieve higher values on other evaluation quantities, such as F1, MCC, etc.; (4) the meta-strategy provides novel ideas on the renovation of existing predictors.

3.9 Conclusion

To improve the microRNA:mRNA prediction performance, the most critical work is to build the comprehensive data sets, which contain as many as microRNA:mRNA pairs. To achieve this goal, four microRNA:mRNA predicted databases, miRanda, miRDB, PITA, and TargetScan, and two microRNA:mRNA experimental validated databases, were integrated to build eleven predicted score specific data sets. In each data set, we labeled the true sample pairs and false sample pairs, and provided the interaction confidence scores for each pair. These eleven mouse microRNA:mRNA interaction data sets include all of the microRNAs which are released from database miRBase. The samples in these eleven data sets can be used directly to analyze the microRNA:mRNA interactions. In addition, these eleven data sets can be further refined as training and testing data sets, which can be used to develop the new microRNA:mRNA interaction predictors.

Further analyzing the performances of the component individual predictors, the number of overlapped microRNA:mRNA interactions of any two individual predictors in each data set
is significantly fewer than the number of overall true positive microRNA:mRNA interaction pairs. This result indicates the prediction performance of microRNA:mRNA interactions can be improved if the suitable method used to integrate the predicted scores of the component individual predictors. The multi-threshold sequential voting was applied to integrate the predicted scores from MiRanda, miRDB and PITA, and we found this meta-strategy can improve the prediction performance of microRNA:mRNA significantly compared with the component individual predictors. Furthermore, we changed the order of individual predictors when applying sequential voting, and found the performances of those meta-predictors are not changed significantly. This result suggests the order of the predictors included are not critical for the meta-strategy. However, when applying this meta-strategy on other ten microRNA:mRNA interaction data sets, the performances of this meta-strategy are not improved in most of those data sets. The results indicate directly applying meta-strategy may not work properly to improve the prediction performances of microRNA:microRNA targets, it needs other techniques integrated to improve the final result.

Therefore, the decision-tree-based dual-threshold value and significant-voting were integrated with Artificial neural network to generate the meta-strategy. This newly-designed meta-strategy are component of two processes. In the first process, the queried microRNA:mRNA pairs go through the four component individual predictors and obtained the predicted scores from each predictors. The programming will check the number of scores obtained from these four predictors and then determining which decision-tree-based artificial neural network (DANN) modules will be applied. In the second process, the predicted scores from individual predictors will go into the DANN modules. In each DANN module, the input data will firstly be processed by decision-tree-based dual-threshold significant-voting. The input data will be encoded and extended by six more dimensions. The data with the new dimension will be used as input to go through the two-hidden layer ANN and obtain the final outputs. The performances of the newly-designed model (mirTarDANN) were evaluated
using Sensitivity, Specificity, Accuracy, F1 scores, and MCC values. The results of eleven microRNA:mRNA interaction data sets display the great improvement of mirTarDANN on the prediction sensitivity, accuracy, F1 scores and MCC values. These results suggest that the newly-designed meta-strategy has its advantage on balancing the sensitivity and specificity to obtain the better other prediction evaluated values (e.g. F1 score and MCC). It is possible to extend the newly-designed meta-strategy on other biological data fields.

We extended the newly-designed meta-strategy on human microRNA:mRNA interaction data sets, and found the great improvement of prediction performances, such as sensitivity, F1 scores, and MCC values. However, the application of this newly-designed meta-strategy on microRNA structure prediction, the prediction performance can be improved, but it is lower than the recent developed meta-predictor published in 2016 [174]. The applications of the newly-designed meta-strategy on disorder prediction and disorder binding prediction display improvement. We published paper “Decision-tree Based Meta-Strategy Improved Accuracy of Disorder Prediction and Identified Novel Disordered Residues Inside Binding Motifs” in *International Journal of Molecular Science*. In addition, the newly-designed meta-strategy applied on the prediction of binding regions in disorder proteins shows the improved prediction performances in both training and testing data set. However, when applying the generated model on the DisBind data set, the result is not improved and is similar with other disorder binding predictors. Taken all of the results, it suggested that the meta-strategy is possible to extend to various fields. It can balance the sensitivity and specificity to obtain the better other evaluated scores. However, it has its own limitations, which need to be further modified and optimized when applied on various fields.
CHAPTER 4

DESIGN SYNTHESIZED STRATEGIES TO CHARACTERIZE PROTEIN-PROTEIN INTERACTION NETWORKS UNDER SPECIFIC CONDITIONS

4.1 Introduction

The complex biological processes are based on the precise regulations of thousands of genes and the interactions of genes and the gene products [201, 202]. Protein-protein interactions are involved in many biological processes, including signaling transduction [203], formation of protein complex [204], protein carrier [205], and functional kinase [206], etc.. The study of protein-protein interaction can provide information to predict the function of unknown proteins in networks [207, 208, 209], and improve the understanding of protein complexes and pathways characterization [210, 211]. Currently, there are many experimental techniques used to study protein-protein interactions which are generating a huge amount of protein-protein interaction data [212]. However, current methods have several issues. For instance, the method used to identify protein-protein interaction is performed under the static physiological conditions [213, 214]. The change of the conditions may cause the identified protein-protein interaction pairs to behave differently. Even some experiments were performed under the same condition, the identified interaction pairs from these experiments are significantly different [213, 214]. Ruling out systematic error, one possible explanation is that it may be caused by the subtle inter-molecular interactions and regulations.
There are many protein-protein interaction databases that have been established in recent years [31, 47, 52]. They were formed through different biological aspects, such as protein structure, subcellular localization, and disease associations [59, 215]. Table 4.1 displays the summary of twenty-four different validated protein-protein interaction databases. Some protein-protein interaction databases also extended the interaction pairs by including the high confident predicted protein-protein interaction pairs [216]. However, those protein-protein interaction databases have the same issue because many interactions are missing (false negative rate) and many are false positive [3]. Furthermore, the protein-protein interactions from different sources are inconsistent [3]. The inconsistent protein-protein interaction networks used to study the regulatory systems may lead to unreliable or contradictory results when applied to analysis the biological processes [3]. Therefore, it is necessary to establish a comprehensive protein-protein interaction database, which contains as many as validated protein-protein interaction pairs.

Another critical issue for the protein-protein interaction is inadequate consideration of subtle inter-molecular interactions. Including the information concerning cellular processes is essential for protein-protein interaction analysis. Previous studies have shown that rewiring protein-protein interaction networks by integrating biological factors, such as cellular localization, gene alternative splicing, and protein post-translational modification, can improve the validity of the interaction maps and exclude the interactions which have weak relationships to the conditions [4, 5, 6].

The human body is made up of different tissues and cell types, in which a various set of genes and proteins are expressed [241, 242, 243, 244]. Different interaction pairs are present across different tissues and cell types [245, 246]. Hence, the gene and protein expression levels are critical for understanding the protein-protein interactions. However, current experimental methods used to identify protein-protein interactions are performed under the same condition and cannot be repeated in different tissues [247, 248]. There is a need to
Table 4.1. The summary of protein-protein interaction databases.

<table>
<thead>
<tr>
<th>Name</th>
<th>Type</th>
<th>Update</th>
<th>Websites</th>
<th>Database</th>
<th>Exp. tp.&amp;meth.</th>
<th>Ref.</th>
</tr>
</thead>
<tbody>
<tr>
<td>APIID</td>
<td>Integrated</td>
<td>2016</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>[31]</td>
</tr>
<tr>
<td>ComPPI</td>
<td>Integrated</td>
<td>2018</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>[217]</td>
</tr>
<tr>
<td>ConsensusPathDB</td>
<td>Integrated</td>
<td>2013</td>
<td>Yes</td>
<td>Part</td>
<td>No</td>
<td>[218, 219, 220]</td>
</tr>
<tr>
<td>HIPPIE</td>
<td>Integrated</td>
<td>2017</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>[221, 222, 223]</td>
</tr>
<tr>
<td>IID</td>
<td>Integrated</td>
<td>2019</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>[3, 224]</td>
</tr>
<tr>
<td>InterMitoBase</td>
<td>Integrated</td>
<td>2011</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>[225]</td>
</tr>
<tr>
<td>PINA</td>
<td>Integrated</td>
<td>2011</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>[48]</td>
</tr>
<tr>
<td>PRIMOS</td>
<td>Integrated</td>
<td>2013</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>[226]</td>
</tr>
<tr>
<td>SCAN</td>
<td>Integrated</td>
<td>2011</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>[227]</td>
</tr>
<tr>
<td>BioGrid</td>
<td>Integrated</td>
<td>2017</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>[45, 46, 47]</td>
</tr>
<tr>
<td>CORUM</td>
<td>Integrated</td>
<td>2018</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>[228, 229]</td>
</tr>
<tr>
<td>BIND</td>
<td>Single</td>
<td>2000</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>[230, 231]</td>
</tr>
<tr>
<td>BindingDB</td>
<td>Single</td>
<td>2007</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>[232, 233]</td>
</tr>
<tr>
<td>DIP</td>
<td>Single</td>
<td>2015</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>[49, 50, 51]</td>
</tr>
<tr>
<td>DisBind</td>
<td>Single</td>
<td>2017</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>[234]</td>
</tr>
<tr>
<td>HPRD</td>
<td>Single</td>
<td>2009</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>[57, 58]</td>
</tr>
<tr>
<td>InnateDB</td>
<td>Single</td>
<td>2013</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>[56]</td>
</tr>
<tr>
<td>IntAct</td>
<td>Single</td>
<td>2015</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>[52]</td>
</tr>
<tr>
<td>LMPID</td>
<td>Single</td>
<td>2015</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>[235]</td>
</tr>
<tr>
<td>MatrixDB</td>
<td>Single</td>
<td>2015</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>[53, 54, 55]</td>
</tr>
<tr>
<td>MINT</td>
<td>Single</td>
<td>2013</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>[236]</td>
</tr>
<tr>
<td>Reactome</td>
<td>Single</td>
<td>2017</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
<td>[237]</td>
</tr>
<tr>
<td>SPIKE</td>
<td>Single</td>
<td>2010</td>
<td>Yes</td>
<td>Yes</td>
<td>No</td>
<td>[238, 239]</td>
</tr>
<tr>
<td>TIMBAL v2.0</td>
<td>Single</td>
<td>2013</td>
<td>Yes</td>
<td>No</td>
<td>No</td>
<td>[240]</td>
</tr>
</tbody>
</table>

N.B. Exp. tp. & meth. is short for experiment types and methods.

Refine the protein-protein interaction networks by using the tissue-specific genes and protein expression analysis. The common methods used to rewire protein-protein interaction networks in specific tissues are based on the analysis of expression data, including the analysis of both gene expression profiles and protein expression profiles [249, 250]. When applying gene and protein expression analysis to rewire the protein-protein networks, there is a condition that some genes are expressed in the tissue but their corresponding proteins are not detected or lowly expressed in that tissue. One possible explanation of this condition is that the mRNAs are interacting with microRNAs and inhibited by the post-transcriptional regulation of microRNAs [251]. The microRNAs also have tissue-specificity, therefore the expression
of microRNAs may also affect the analysis of finalized protein-protein interaction networks. In our research, we integrated gene, microRNA, and protein expression information into the protein-protein interaction database to rewire the protein-protein interaction networks.

In this section, the integrated protein-protein interaction data set were established by combining eight different protein-protein interaction databases, including BioGrid [45, 46, 47], HPRD [57, 58], DIP [49, 50, 51], IntAct [52], InnateDB [56], MatrixDB [53, 54, 55], PINA [48], and APID [31]. We also provided the confidence scores for each interaction pairs. After that, the gene, protein and microRNA tissue-specific expression information were integrated into the protein-protein interaction networks to rewire the networks. Here, we will introduce the materials and methods used in this section, and the results.

4.2 Materials and Methods

4.2.1 Database Sources

The experimental validated protein-protein interaction pairs were downloaded from eight different databases, including BioGrid (version 3.5.165) [45, 46, 47], DIP (2015) [49, 50, 51], HPRD (release 9.0) [57, 58], IntAct (2015) [52], APID (2016) [31], InnateDB (2013) [56], MatrixDB (2015) [53, 54, 55], and PINA (2011) [48]. APID and PINA are integrated PPI databases. DIP, IntAct, InnateDB and MatrixDB are members of International Molecular Exchange Consortium (IMEx) [252]. BioGrid is the observer for IMEx. HPRD includes both protein interaction information as well as the phosphorylation data belonging to around 26 different types [57, 58]. There are three reasons that those databases have been selected. (1) These databases are well-maintained and have the PPI databases released to download; (2) These databases provide either gene symbol IDs or UniProt IDs; (3) These databases include the experimental methods and experimental types that are used to identify protein-protein interactions. The summary of those PPI databases is listed in Table 4.2.
### Table 4.2. Summary of protein-protein interaction databases.

<table>
<thead>
<tr>
<th>Database</th>
<th>Last Update</th>
<th>Total ID</th>
<th>No. Interactions</th>
<th>No. Genes</th>
<th>No. Uniprots</th>
</tr>
</thead>
<tbody>
<tr>
<td>BioGrid</td>
<td>2017</td>
<td>17299</td>
<td>318489</td>
<td>17298</td>
<td>16193</td>
</tr>
<tr>
<td>HPRD</td>
<td>2009</td>
<td>9678</td>
<td>39204</td>
<td>9616</td>
<td>9676</td>
</tr>
<tr>
<td>DIP</td>
<td>2015</td>
<td>3262</td>
<td>5552</td>
<td>3157</td>
<td>3170</td>
</tr>
<tr>
<td>IntAct</td>
<td>2015</td>
<td>23606</td>
<td>194555</td>
<td>15757</td>
<td>20414</td>
</tr>
<tr>
<td>InnateDB</td>
<td>2013</td>
<td>5747</td>
<td>13704</td>
<td>5724</td>
<td>5559</td>
</tr>
<tr>
<td>MatrixDB</td>
<td>2015</td>
<td>14428</td>
<td>78621</td>
<td>11760</td>
<td>14428</td>
</tr>
<tr>
<td>PINA</td>
<td>2011</td>
<td>17109</td>
<td>166776</td>
<td>16172</td>
<td>17109</td>
</tr>
<tr>
<td>APID</td>
<td>2016</td>
<td>19297</td>
<td>319273</td>
<td>16778</td>
<td>18651</td>
</tr>
</tbody>
</table>

### 4.2.2 Gene Expression

Human gene expression data sets were downloaded from Gene Expression Omnibus (GEO) databases. These data sets are GSE1133 [60], GSE7905 [59], and GSE7307 [61]. These data sets were designed for the study of the distribution of gene expression in different human and mouse tissues. GSE1133 was designed to study the tissue-specific patterns of mRNA. This data set included up to 79 human tissues and 61 mouse tissues gene expression profile [60]. GSE7905 was designed to study the comprehensive function of human tissue-specific expressed genes, and had up to 31 human tissues [59]. GSE7307 was established to study the differences of gene expression level between human normal and diseased tissues. There were 90 distinct tissue types, including both normal and diseased tissues [61]. Taken together, those data sets provide up to \(~30\) critical tissues in Homo Sapiens (Table 4.3). In each data set, the disease samples were removed and the *affy* package in R was applied to normalize the data set [90]. The probe IDs were mapped to gene symbol IDs using the platform information provided by each data set. In each tissue, the replicates were averaged and the averaged expression levels of gene were ranked by percentiles.
Table 4.3. Summary of tissue-specific gene expression data sets.

<table>
<thead>
<tr>
<th>Tissues</th>
<th>GSE1133</th>
<th>GSE7905</th>
<th>GSE7307</th>
</tr>
</thead>
<tbody>
<tr>
<td>adipose tissue</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>adrenal gland</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>amygdala</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bone</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>bone marrow</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>brain</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>hindgut</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>dorsal root ganglia</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>eye</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>heart</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>hypothalamus</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>kidney</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>liver</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>lung</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>lymph nodes</td>
<td>X</td>
<td></td>
<td></td>
</tr>
<tr>
<td>mammary gland</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>midgut</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>ovary</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>pancreas</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>pituitary gland</td>
<td>X</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>placenta</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>prostate</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>salivary gland</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>skeletal muscle</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>small intestine</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>spleen</td>
<td></td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>stomach</td>
<td></td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>testis</td>
<td>X</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>uterus</td>
<td>X</td>
<td>X</td>
<td></td>
</tr>
</tbody>
</table>

4.2.3 Protein Expression

Protein expression data sets were downloaded from PaxDb [62, 63] and ProteomicsDB [64]. The database PaxDb includes protein expression information from 414 data sets in 53 organisms. The integrated results of PaxDb included up to ∼32 tissues [62, 63]. ProteomicsDB was established using 16,857 liquid chromatography tandem-mass-spectrometry
Table 4.4. Summary of tissue-specific protein expression data sets.

<table>
<thead>
<tr>
<th>Tissues</th>
<th>PaxDb</th>
<th>ProteomicsDB</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Colon</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Esophagus</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Gall bladder</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Heart</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Kidney</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Liver</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Lung</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Pancreas</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Placenta</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Platelet</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Prostate gland</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Rectum</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Saliva</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Skin</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Testis</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Uterus</td>
<td>X</td>
<td>X</td>
</tr>
</tbody>
</table>

experiment results of human tissues, cell lines, body fluids and other biological post-translational modification studies [64]. We selected those two protein expression databases for these several reasons: (1) those databases are well-established and updated; (2) those databases provide public available expressed data; (3) The Human Protein Atlas not used is because the PaxDb included the information from this database [253]. Integrated those two human protein expression databases, there are \( \sim 17 \) tissues included and shown in Table 4.4. The protein expression values higher than 0 or “NA” were treated as tissue specific expressed proteins. The Entrez Gene IDs were mapped to UniProt IDs and gene symbol IDs using UniProt database released on 2018-05.

4.2.4 MicroRNA Expression

The data of microRNA expression data were collected from database miRmine [65] and TissueAtlas [66]. These two databases are the only current released databases from mi-
croRNA expression data. MiRmine was established by high-quality microRNA sequencing data sets from NCBI-SRA. It includes 16 different types of human tissues and bio fluids, and 24 different types of cell-lines [65]. TissueAtlas was designed to study the distribution of microRNA expression in different human tissues. There are 16 human normal tissues microRNA expression data included [66]. Integrated both microRNA expressed databases, there are 11 tissues found and shown in Table 4.5. The expression level of each microRNA was averaged between expression levels of the replicates in the same tissue and the averaged expression levels of microRNA were ranked by percentile for the analysis.

<table>
<thead>
<tr>
<th>Tissues</th>
<th>miRmine</th>
<th>TissueAtlas</th>
</tr>
</thead>
<tbody>
<tr>
<td>Brain</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Breast</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Heart</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Lung</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Liver</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Ovary</td>
<td></td>
<td>X</td>
</tr>
<tr>
<td>Placenta</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Pancreas</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Skeletal muscle</td>
<td>X</td>
<td></td>
</tr>
<tr>
<td>Testis</td>
<td>X</td>
<td>X</td>
</tr>
<tr>
<td>Thymus</td>
<td></td>
<td>X</td>
</tr>
</tbody>
</table>

### 4.2.5 Mapping Protein IDs

In different protein-protein interaction databases, the UniProt IDs or gene symbol IDs were assigned differently. To make them uniform, the UniProt IDs or gene symbol IDs were mapped using UniProt databases which was downloaded on 2018-05. For instance, if the major searching ID was UniProt IDs and the gene symbol IDs were mismatched with the UniProt released database, the gene symbol IDs were changed to the gene symbol IDs the same with UniProt database. Similarly, if the major searching ID was gene symbol IDs and
the UniProt IDs were mismatched with the UniProt released database, the UniProt IDs were changed to the UniProt IDs the same with the UniProt database.

### 4.2.6 Generating Protein-protein Interaction Data Set

All the protein IDs from different databases were merged together to form protein ID list. The protein in this list includes UniProt ID, gene symbol ID, protein name, and the names of protein-protein interaction databases which have this protein. After sorting the protein ID list with those eight protein-protein interaction databases, the interaction pairs were merged together and formatted to UniProt A, gene symbol A, Protein name A, UniProt B, gene symbol B, protein name B, the names of databases which have the interaction pair, and the confidence score of the interaction pair. The confidence score was calculated through two steps. First, the score of interaction pairs from each database was calculated. For example, the score of the interaction pair was calculated by the experimental methods, experiment types and the number of the experiments performed in each database. Then, the score from each database of the interaction pair was normalized in a range of 0 to 1. The integrated score (confidence score) was calculated using the formula as follows:

\[
S_i = \sum e_j p_{i,j}
\]

Where \(S_i\) is the confidence score of interaction pair \(i\) in the eight integrated protein-protein interaction databases. \(e_j\) is the ratio between number of interaction pairs in database \(j\) and the total number of integrated protein-protein interaction database. \(p_{i,j}\) is the normalized confidence score for the \(i\)-th interaction pair in database \(j\).
4.2.7 Tissue-specific Protein-protein Interactions

The protein-protein interactions were further refined by tissue-specific expression at both RNA and protein levels. Most importantly, the tissue-specific microRNAs were also used to refine protein-protein interaction networks.

4.2.8 Protein-protein Interaction Searching Pipeline

We designed the new searching pipeline to search the combined protein-protein interaction data set. Figure 4.1 shows the infrastructure of the new searching pipeline. The queried ID can be UniProt IDs, gene IDs or protein names of the molecules. After obtaining the queried ID, the program will search all the corresponding UniProt IDs in the new generated protein hash table, then use all the corresponding UniProt IDs to search protein-protein interactions. The program will return the first output including all the interaction pairs. After that, the mRNA, microRNA and protein tissue-specific expression information will be used to filter the first returned networks and then generate the tissue-specific interaction tables.

4.3 Results

4.3.1 Comparison of Eight Protein-protein Interaction Databases

Figure 4.2 shows the comparison of eight protein-protein interaction databases. Clearly, there were a batch of unique protein-protein interactions which can be found in each database. The database BioGrid, APID, PINA, and IntAct each provides more than ten thousands of unique protein-protein interaction pairs. The database InnateDB and MatrixDB have thousands of unique protein-protein interaction pairs. This result indicates if the protein-protein interactions were analyzed based on only one database, the obtained results were biased.
Figure 4.1. The ER diagram for protein-protein interaction searching. The input queried protein ID can be gene symbol ID, UniProt ID and protein name. In the Part I, the input ID will be formatted by protein corresponding ID list and the outputs of the Part I are the corresponding UniProt IDs. In the Part II, the corresponding UniProt IDs will be searched using Protein-protein interaction list. The outputs of Part II is the protein-protein interaction pairs. “↔” indicates the proteins in these two squares can interact with each other. In Part III, the protein-protein interactions will be rewired by tissue-specific expressed proteins, gene, and microRNAs.
Figure 4.2. The comparison among eight protein-protein interaction database. The black words inside each circle display the number of unique protein-protein interactions included in that database. The percentage after each database represents the ratio of protein-protein interactions in that databases to all those eight databases.
4.3.2 Impacts of Mismatch IDs in Protein-protein Interaction Database

The databases, BioGrid, HPRD, IntAct, InnateDB, APID, and PINA, can provide both UniProt IDs and gene symbol IDs. After checking these IDs from each database with UniProt database, we found the gene symbol IDs of proteins from databases were mismatched. The number of mismatched IDs are 1891, 551, 82, 149, 1090, and 744 in the database BioGrid, HPRD, IntAct, InnateDB, APID, and PINA, respectively. Figure 4.3(A) shows the amount of proteins in each database compared with the number of mismatched IDs. The percentages of mismatched IDs are shown in Figure 4.3(B). In Figure 4.3(B), it shows the percentages of mismatched IDs in database BioGrid was more than 10%. In databases HPRD and APID, the percentages of mismatched IDs were higher than 5%. Combining the mismatched IDs from those six databases, the total number of mismatched IDs from those six databases was 3592. Compared to the total number of IDs from those six databases, the percentage of mismatched IDs was ∼14%.

The mismatched IDs in those six databases were filtered with UniProt database. The database DIP did not provide the gene symbol IDs. The gene symbol IDs of DIP database were obtained using UniProt database. After that, the filtered protein IDs from database APID, BioGrid, DIP, HPRD, IntAct, InnateDB, MatrixDB, and PINA were merged together to form the Protein ID list. Figure 4.3(C) shows the number of proteins from each database. After combining proteins from all eight filtered database, the total number of proteins was 29,481. The filtered protein list was used to filter the protein-protein interaction pairs from each database and merged those interaction pairs to form comprehensive protein-protein interaction data sets. Figure 4.3(D) displays the number of filtered interactions pairs from each database, and after combining those database, the total number of interaction pairs was 535,717.
Figure 4.3. Analysis of mismatched IDs. (A) The number of mismatched IDs compared to the total number of IDs from each database. (B) The percentage of mismatched IDs from each database. (C) The number of proteins after filtering with UniProt database. (D) The number of interactions after filtering with filtered protein list. X-axis displays the name of databases, and y-axis in (A) (C) (D) shows the number of proteins and in (B) represents the percentage of mismatched IDs.
Table 4.6. The Summary of Genes with Multiple UniProt IDs.

<table>
<thead>
<tr>
<th>Database</th>
<th>No. genes more than one UniProt ID</th>
<th>No. FN genes</th>
<th>No. FP genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>HPRD</td>
<td>81</td>
<td>65</td>
<td>10</td>
</tr>
<tr>
<td>IntAct</td>
<td>3862</td>
<td>3450</td>
<td>414</td>
</tr>
<tr>
<td>MatrixDB</td>
<td>1891</td>
<td>1724</td>
<td>242</td>
</tr>
<tr>
<td>PINA</td>
<td>565</td>
<td>432</td>
<td>96</td>
</tr>
<tr>
<td>APID</td>
<td>1711</td>
<td>608</td>
<td>1116</td>
</tr>
</tbody>
</table>

4.3.3 False Positive or False Negative Protein-protein Interactions

The number of UniProt IDs in each database are normally higher than the number of gene symbol IDs. This result indicates the same gene symbol IDs have two or more UniProt IDs. Table 4.6 displays the number of gene symbols that have multiple UniProt IDs in each database. Therefore, it is possible that the UniProt IDs with the same gene symbols cause false positive (FP) or false negative (FN) results when searching protein-protein interactions. For instance, if the gene symbol has multiple UniProt IDs which represent different proteins, using this gene symbol ID to search PPIs, the obtained protein-protein interaction networks may have FP issues because of the obtained protein-protein interaction networks including protein-protein interactions of those multiple UniProt IDs. If the gene symbol has multiple UniProt IDs which represent the same protein, using one of the UniProt IDs to search protein-protein interactions, the obtained protein-protein interaction networks may have FN issues since the protein-protein interactions only contain interactions of the queried UniProt ID.

Here, we set out to overcome a challenge, that is how to determine if the UniProt IDs are the same or unsame proteins. The proteins with the same gene symbol were aligned using Blast-2.9.0. Under different threshold values of sequence coverage and identity, different FP and FN rates were achieved. Figure4.4(A) displays the false rate in each database, which was calculated by the sum of FP rate and FN rate. The threshold value of the lowest false rate was selected as a cutoff in that database to determine whether proteins are the same or
not. From Figure 4.4(A), the cutoffs of coverage and identity are 0.8 and 0.9 in APID; 0.5 and 0.9 in HPRD; 0.3 and 0.8 in IntAct; 0.3 and 0.7 in MatrixDB; and 0.3 and 0.8 in PINA. After that, the FP rate and FN rate caused in each database were calculated and shown in Figure 4.4(B). Figure 4.4(B) shows the databases of IntAct, MatrixDB and APID can cause around or more than 10% false rate.

We calculated the percentage of proteins in differing range of FP rate or FN rate range, the results of which are shown in Figure 4.5 (A) APID; (B) IntAct; (C) MatrixDB; (D) PINA. Many proteins had a higher distribution when FP rate or FN rate was in the range of 0-10% and 90-100%. The percentage of proteins is more than 10% when FN rate is 0-10% in database IntAct.

4.3.4 **Tissue-specific Protein-protein Interactions**

To assign the interaction pairs into the different tissues, we chose the pairs in which both proteins or genes were expressed in the tissue. The genes in Microarray data sets which are higher than 30, 40, 50, 60, 70, and 80 percentile were selected as tissue specific expressed genes. The proteins with the expressed value larger than 0 were selected as tissue specific expressed proteins. The comparison of tissue specific expressed proteins and tissue specific expressed genes shows there are many genes that can express in the tissue but the corresponding proteins cannot be found in the same tissue. The Figure 4.6(A) displays the number of genes that can only be found in specific tissues. The number of the genes is reduced as the increased percentile is used to select tissue specific genes. This result indicates there is a positive relation between expressed genes and unexpressed proteins in the specific tissues. The translational processes of those genes are regulated by many biological processes. One possible explanation of these observations is the post-transcriptional regulation of microRNA on mRNA. Figure 4.6(B) shows the association between the number of microRNA and mRNA interaction pairs in the tissue and the expression level of either microRNAs or mRNAs in
Figure 4.4. The false rate in different databases. (A) The false rate of each database using different threshold values of coverage and identity from sequence alignment. X-axis represents the combination of coverage and identity, and y-axis represents the false rate based on the coverage and identity. (B) The FP and FN percentage in each database using coverage and identity which result in the lowest false rate. X-axis displays the names of databases, and y-axis displays the percentage of FP, FN and all false protein-protein interactions.
Figure 4.5. The distribution of proteins in different ranges of FP and FN rates in databases (A) APID; (B) IntAct; (C) MatrixDB; (D) PINA. X-axis displays the different ranges of FP and FN rate, and y-axis displays the percentages of proteins in those ranges.
the same tissue when using the percentile 30, 40, 50, 60, 70 and 80 to select tissue specific microRNAs or mRNAs. It was also observed that as the percentile value used to select either microRNAs or mRNAs in the specific tissue was increased, the number of microRNAs and mRNAs interaction pairs were reduced.

4.3.5 Impacts of Tissue-specificity on Protein-protein Interactions

The protein-protein interaction networks were rewired when one or several proteins are missing in the networks. Figure 4.7 displays the interaction networks of the four genes by using gene name as keyword to search for protein-protein interactions (A) NRIP1, (B) FGF1, (C) INS-IGF1-IGF2, and (D) GATAD1. The protein-protein interaction networks were further refined by integrating the tissue-specific gene expression at both RNA and protein levels. Figure 4.8, 4.9, 4.10, and 4.11 display the interaction networks of these four genes in four different tissues which are (A) brain; (B) lung; (C) pancreas; and (D) placenta. After being rewired, the networks in different tissues become further different. For instance, the protein-protein network for NRIP1 has 72 nodes and 440 binary interactions before rewiring as shown in Figure 4.8. However, there are only 25, 26, 24, 24 nodes in brain, lung, pancreas, and placenta, respectively. Figure 4.9 shows the networks for FGF1, which 32 nodes with 91 interaction pairs before rewiring. After rewiring the networks based on differentially expressed genes at RNA and protein levels, 15, 19, 16, and 18 nodes left in the brain, lung, pancreas, and placenta. In Figure 4.10, the interaction networks of INS-IGF1-IGF2 were rewired in tissue brain, lung, pancreas and placenta. The number of interaction nodes dropped from 74 to 27, 30, 39, and 36 nodes in brain, lung, pancreas, and placenta, respectively. In Figure 4.11, the number of GATAD1 interaction nodes in-vitro was 51. After refining the interaction networks, the number of interaction nodes are 13, 16, 14, and 13 in brain, lung, pancreas, and placenta, respectively.
Figure 4.6. The tissue specific expressed mRNAs and microRNAs. (A) The number of genes that expressed in specific tissues while the corresponding proteins are not found in those tissues. X-axis displays the names of tissues, y-axis displays the number of mRNAs which are expressed in that tissue while the corresponding proteins are not expressed, and z-axis displays the percentiles that used to select tissue specific expressed genes. (B) The number of microRNA and mRNA interaction pairs found in specific tissues. X-axis represents the names of tissues, y-axis represents the number of microRNA and mRNA interaction pairs that can be found in that tissue, and z-axis represents the percentiles that were used to select either mRNA or microRNA tissue specific expressed genes.
Figure 4.7. The interaction networks in vitro by searching (A) NRIP1; (B) FGF1; (C) INS-IGF1-IGF2; (D) GATAD1. The light blue hexagons represent the interaction partners.
Figure 4.8. The NRIP1 interaction networks in (A) brain; (B) lung; (C) pancreas; (D) placenta. The light blue hexagons represent the molecule was found in-vitro. The green hexagons display the molecule was expressed in the tissue at the RNA level but not at the protein level. The blue hexagons represent the molecule can express in the tissue at both RNA and protein levels. The grey lines display the in-vitro interaction and the red solid lines represent the binary interaction in that tissue.
Figure 4.9. The FGF1 interaction networks in (A) brain; (B) lung; (C) pancreas; (D) placenta. The light blue hexagons represent the molecule was found in-vitro. The green hexagons display the molecule was expressed in the tissue at the RNA level but not at protein level. The blue hexagons represent the molecule can express in the tissue at both RNA and protein levels. The grey lines display the in-vitro interaction and the red solid lines represent the binary interaction in that tissue.
Figure 4.10. The INS-IGF1-IGF2 interaction networks in (A) brain; (B) lung; (C) pancreas; (D) placenta. The light blue hexagons represent the molecule was found in-vitro. The green hexagons display the molecule was expressed in the tissue at the RNA level but not at protein level. The blue hexagons represent the molecule can express in the tissue at both RNA and protein levels. The grey lines display the in-vitro interaction and the red solid lines represent the binary interaction in that tissue.
Figure 4.11. The GATAD1 interaction networks in (A) brain; (B) lung; (C) pancreas; (D) placenta. The light blue hexagons represent the molecule was found in-vitro. The green hexagons display the molecule was found expressed in the tissue at the RNA level but not at protein level. The blue hexagons represent the molecule can express in the tissue at both RNA and protein levels. The grey lines display the in-vitro interaction and the red solid lines represent the binary interaction in that tissue.
The impact of microRNA on protein-protein interaction rewiring was analyzed by adding microRNAs, which targets mRNAs that have high RNA level but low protein level. The interaction nodes without the annotation of tissue-specific information were removed from the networks, and the mRNA interaction microRNAs were added to the networks to further analyze. Figure 4.12 shows the microRNA-regulated tissue-specific interaction networks for NRIP1. In brain tissue (Figure 4.12(A)), 21 genes are expressed at the protein level, and four genes have RNA-level expression but not at the protein level. These 4 genes were JUN, NR3C1, NRIP1, and CIB1, and are predicted by our predictor to interact with microRNAs. In Figure 4.12(B), the lung-specific interaction network containing 74 protein-protein interaction pairs was presented. Out of 26 genes, 21 are highly expressed at both the RNA and protein levels. The rest 5 genes, of which the proteins are not expressed in lung, were NR3C1, TNFRSF14, JUN, THRA, and CARM1. From the microRNA interaction analysis, it is clear that those 5 genes interact with microRNA in lung tissue. Figure 4.12(C) displays the interaction network of NRIP1 in pancreas, in which eighty protein-protein interaction pairs were found. Out of 24 genes, 22 genes were highly expressed at both RNA and the protein levels. The rest 2 genes, which are not expressed at the protein level in pancreas, were TNFSF14 and THRA. In Figure 4.12(D), the placenta-specific interaction network was presented. The interaction network was the same with the pancreas-specific interaction network in pancreas tissue. In the previous research, it has stated that JUN N-terminal kinase can phosphorylate the APP Thr668 residue, which is critical to cause the aggregation of APP [254]. At the normal state, JUN is not expressed in brain and cannot interact with APP. Figure 4.12(A) displays the low protein-level expression of JUN, which may be because of the interaction with microRNAs. The change of the state may cause the development of Alzheimer disease. In lung (Figure 4.12(B)), JUN also has low protein-level expression. The previous study has demonstrated that the low protein-level expression of JUN can improve the function of
epithelial cells [255]. The interactions between microRNAs and JUN can down-regulate the protein-level expression of JUN [256] to improve the function of lung epithelial cells.

Figure 4.13 displays the microRNA-regulated tissue-specific interaction networks for FGF1. In brain (Figure 4.13(A)), 14 genes are expressed at both RNA and the protein levels, and one gene has RNA-level expression but not has protein-level expression. This gene was TGFBR3, which can be predicted to interact with a great number of microRNAs. The lung-specific interaction network containing 91 interaction pairs was displayed in Figure 4.13(B). Out of 19 genes, 6 genes are highly expressed at the RNA level but not at the protein level. These six genes were FGFR1, FGFBP1, GPC1, FGF1, TGFBR3, and SDC2. Through the microRNA interaction analysis, it is obvious that those six genes interact with microRNAs in lung. Figure 4.13(C) displays the pancreas-specific interaction network of FGF1. FGF1 is highly expressed at the RNA level but not at the protein level. In Figure 4.13(D), 3 genes have the RNA-level expression but not at the protein level. These three genes were FGFR1, SYT1, and FGFBP1, which affect 16 binary interactions in placenta-specific interaction network. Recently, studies have suggested the amplification of FGFR1 is associated with the development of lung squamous cell carcinoma [257]. In Figure 4.13(B), FGFR1 was predicted to interact with microRNAs and the low protein-level expression of FGFR1 may significantly affect the interaction network. Additionally, the protein-level expression of FGF1R was also undetected in placenta (Figure 4.13(D)). Previous study has stated that FGF1R is critical in trophectoderm development and blastocyst implantation [258]. The interactions of FGF1R and microRNAs may affect the function of placenta during the pregnancy.

Figure 4.14 shows the microRNA-regulated interaction networks in different tissues for genes INS, IGF1, and IGF2. Figure 4.14(A) presents the brain-specific interaction network containing 57 interaction pairs, of which 21 interaction pairs were affected by 5 genes, IGF1, IGFBP3, IGFBP6, ADAMTSL4, and LRP2. These five genes, which were expressed at the
RNA level but not at the protein level, were predicted to interact with microRNAs. In lung (Figure 4.14(B)), 30 genes have both RNA-level and protein-level expression, out of which, 8 genes were detected highly expressed at the RNA level but not at the protein level. These 8 genes were IGF1, IGFBP2, IGFBP4, IGFBP5, IGFBP6, ADAMTS14, SOX4, and PTEN, which affect more than 200 binary interactions in lung-specific interaction network. Figure 4.14(C) displays the pancreas-specific interaction network. Out of 40 genes, 4 genes are highly expressed at the RNA level but not at the protein level. They were ADAMTS14, IGF1, IGF2 and SOX4, which were predicted to interact with microRNAs. In Figure 4.14(D), 8 genes have high RNA-level expression but are not expressed at protein level. These eight genes were SOX4, PTEN, MAPT, WISP2, IGFBP4, IGFBP5, IGFBP6 and IGF1, which are predicted to interaction with microRNAs, and affect 36 binary interactions in placenta-specific interaction network. Previous study has demonstrated that the higher circulating levels of IGFs may increase the risk of pancreas cancer [259]. In Figure 4.14(C), we observed the interactions of microRNAs and genes IGF1 and IGF2 can affect the IGF1 and IGF2 protein-level expression. By this way, the protein-level expression of IGF families keep at the normal condition.

In Figure 4.15 displays the microRNA-regulated tissue-specific interaction networks in various tissues for GATAD1. In brain (Figure 4.15(A)), ARID4A only has RNA-level expression. However, the protein-level expression was undetected in brain. From microRNA interaction analysis, it is clear that ARID4A interacts with a batch of microRNA in brain. In lung (Figure 4.15(B)), 5 genes were detected at the RNA level but not at the protein level. They were ING1, ARID4A, RBFOX1, RBFOX2, and ZNF281, which are predicted to interact with microRNAs and affect 17 interaction pairs in lung. Figure 4.15 (C) displays the pancreas-specific interaction network for GATAD1. Out of 14 genes, 4 genes have the RNA-level expression but not at the protein level. The four genes were NPAT, ING1, GATAD1, and ZNF281, which affect 13 interaction pairs. In Figure 4.15(D), the GATAD1
placenta-specific interaction network was presented. Out of 13 genes, 3 genes, which were ING1, GATAD1, and ZNF281, were expressed at the RNA level but not at the protein level. These three genes were predicted to interact with microRNAs and affect 19 binary interactions in placenta. Previous study has shown that GATAD1 is very critical for the third-trimester placentas [260]. In placenta (Figure 4.15(D)), the interaction of GATAD1 with microRNAs may improve the probability to cause the preeclampsia placentas.

Figure 4.12. The NRIP1 microRNA-regulated tissue-specific interaction networks in (A) brain; (B) lung; (C) pancreas; (D) placenta. The green hexagons display the molecules which are expressed in the tissue at the RNA level but not at the protein level. The blue hexagons represent the molecules that express in the tissue at both protein and RNA level. The diamonds display the microRNAs that also express in the tissue. The red diamonds display the microRNAs which are predicted to interact with mRNAs and the pink diamonds represent the microRNAs that interact with more than two mRNAs. The red solid lines indicate the affected protein-protein interaction networks.
Figure 4.13. The FGF1 microRNA-regulated tissue-specific interaction networks in (A) brain; (B) lung; (C) pancreas; (D) placenta. The green hexagons display the molecules which are expressed in the tissue at the RNA level but not at the protein level. The blue hexagons represent the molecules that express in the tissue at both protein and RNA level. The diamonds display the microRNAs that also express in the tissue. The red diamonds display the microRNAs which are predicted to interact with mRNAs and the pink diamonds represent the microRNAs that interact with more than two mRNAs. The red solid lines indicate the affected protein-protein interaction networks.
Figure 4.14. The INS-IGF1-IGF2 microRNA-regulated tissue-specific interaction networks in (A) brain; (B) lung; (C) pancreas; (D) placenta. The green hexagons display the molecules which are expressed in the tissue at the RNA level but not at the protein level. The blue hexagons represent the molecules that express in the tissue at both protein and RNA levels. The diamonds display the microRNAs that also express in the tissue. The red diamonds display which are predicted to interact with mRNAs and the pink diamonds represent the microRNAs that interact with more than two mRNAs. The red solid lines indicate the affected protein-protein interaction networks.
Figure 4.15. The GATAD1 microRNA-regulated tissue-specific interaction networks in (A) brain; (B) lung; (C) pancreas; (D) placenta. The green hexagons display the molecules which are expressed in the tissue at the RNA level but not at the protein level. The blue hexagons represent the molecules that express in the tissue at both protein and RNA level. The diamonds display the microRNAs that also express in the tissue. The red diamonds display the microRNA which are predicted to interact with mRNAs and the pink diamonds represent the microRNAs that interact with more than two mRNAs. The red solid lines indicate the affected protein-protein interaction networks.
4.4 Discussion

Current protein-protein interaction databases are not consistent from source to source. There are many protein-protein interaction databases established in recent years. For example, HIPPIE database includes the information about tissues, subcellular localizations, and disease associations. TissueNet database contains human protein-protein interactions [244]. ComPPI database includes the human sub-cellular protein-protein interactions and other model organism protein-protein interactions [217]. There are several other protein-protein interaction online sources, such as STRING [216], ConsensusPathDB [218], and iRefWeb [261]. Taking all the databases and online sources together, there are up to 150,000 human interactions being validated through experiments. In addition, STRING also provides the predicted interaction pairs.

However, the existing protein-protein interaction databases are limited by several issues. Firstly, the interactions in each database are inconsistent. Different interaction databases used for analysis can lead to very different analysis results. In our study, the eight different databases were compared and the result was as shown in Figure 4.2. There are a batch of unique interactions in each individual database. Therefore, the eight different databases were combined to form a new protein-protein interaction data set. Since the integrated protein-protein interaction data set contains the interaction pairs from eight different public available databases, it provides more comprehensive information.

Secondly, many proteins have the mismatched UniProt IDs and gene IDs in some individual databases. Therefore, when combining the protein-protein interaction databases and developing the search pipeline for the new combined data set, it is necessary to filter and cluster the protein IDs. Figure 4.3 (A) and (B) display the mismatched IDs and the percentage of mismatched IDs in each individual database. The percentage of total mismatched IDs to the total number of IDs from the six database was $\sim$14%. Therefore, it is necessary to filter
the IDs before integrating the individual protein-protein interaction database. The UniProt database released on 2018-05 was used to filter the protein IDs. Afterwards, these individual protein-protein interaction databases were integrated. There are around 30,000 human proteins with about 550,000 protein-protein interactions included in the newly-integrated database. In Figure 4.3 (C) and (D), the components of each database in the new combined data set were shown.

Thirdly, several genes have more than one UniProt IDs in some individual databases. These genes with multiple UniProt IDs can cause false negative or false positive search results in these databases. In Figure 4.4 and 4.5, we applied protein sequence alignment to determine whether the UniProt IDs with the same gene symbol are the same protein or not in these individual databases. Afterwards, the false positive and false negative rates in these individual databases were further analyzed. In databases IntAct and MatrixDB, more than ∼10% false protein-protein interactions were found to be false positive or false negative results. In our integrated protein-protein interaction data set, we designed a protein ID hash table to cluster the protein IDs based on sequence alignment result. The UniProt IDs of the same protein will be assigned in the same table. The UniProt IDs in the same table will be used to search protein-protein interactions in the newly-integrated data set. In this way, the false negative and false positive results can be removed.

Since the protein-protein interaction networks are various under different biological conditions, we used tissue-specific protein and mRNA expression level to refine the protein-protein interaction networks. In addition, microRNA tissue-specific expression was also applied to interpret the regulation in specific tissues. Figure 4.8, 4.9, 4.10, and 4.11 display the examples of protein-protein interaction networks in different tissues. The impact of microRNA regulation in different tissues are shown in Figure 4.12, 4.13, 4.14, and 4.15. It is obvious that the regulation of microRNA and the expression level of mRNAs and proteins have sig-
nificant impacts on the final protein-protein interaction networks and are closely associated with the development of diseases.

4.5 Conclusion

In this study, eight different protein-protein interaction databases were combined to establish a new protein-protein interaction data set. The newly established data set is more comprehensive because it includes the interactions from different sources. In addition, we filtered IDs using the database released by UniProt, and checked the association of proteins with sequence alignment. The false positive searching and false negative searching were reduced significantly with this method. Furthermore, the newly designed pipeline including both expression information and the regulation of microRNA interactions can be used to discover the new regulatory mechanisms.


[28] Paul Ryvkin, Yuk Yee Leung, Lyle H. Ungar, Brian D. Gregory, and Li-San Wang. Using machine learning and high-throughput rna sequencing to classify the precursors


[56] Karin Breuer, Amir K. Foroushani, Matthew R. Laird, Carol Chen, Anastasia Sribnaia, Raymond Lo, Geoffrey L. Winsor, Robert E. W. Hancock, Fiona S. L. Brinkman, and


[110] T Moradi, RJ Delfino, SR Bergstrom, ESH Yu, HO Adami, and J Yuen. Cancer risk among scandinavian immigrants in the us and scandinavian residents compared with


[133] Carine R. Lussier, Francois Brial, Sbastien A. B. Roy, Marie-Jose Langlois, Elena F. Verdu, Nathalie Rivard, Nathalie Perreault, and Francois Boudreau. Loss of hepatocyte-


David Israeli, Ellen Tessler, Ygal Haupt, Adi Elkeles, Sylvia Wilder, Robert Amson, Adam Telerman, and Moshe Oren. A novel p53-inducible gene, pag608, encodes a nu-


[209] Peter Uetz, Loïc Giot, Gerard Cagney, Traci A. Mansfield, Richard S. Judson, James R. Knight, Daniel Lockshon, Vaibhav Narayan, Maithreyan Srinivasan, Pascale Pochart, Alia Qureshi-Emili, Ying Li, Brian Godwin, Diana Conover, Theodore Kalbfleisch, Govindan Vijayadamodar, Meijia Yang, Mark Johnston, Stanley Fields,


[222] Yih-Chii Hwang, Chiao-Feng Lin, Otto Valladares, John Malamon, Pavel P. Kuksa, Qi Zheng, Brian D. Gregory, and Li-San Wang. HIPPIE: a high-throughput identifica-


[253] Mathias Uhlén, Linn Fagerberg, Björn M. Hallström, Cecilia Lindskog, Per Oksvold, Adil Mardinoglu, Åsa Sivertsson, Caroline Kampf, Evelina Sjöstedt, Anna Asplund, In-


[258] Agata Kurowski, Andrei Molotkov, and Philippe Soriano. Fgfr1 regulates tropho-
toderm development and facilitates blastocyst implantation. Developmental Biology,


[260] Xiaoling Ma, Jinping Li, Brian Brost, Wenjun Cheng, and Shi-Wen Jiang. Decreased
expression and dna methylation levels of gatad1 in preeclamptic placentas. Cellular

[261] Brian Turner, Sabry Razick, Andrei L. Turinsky, James Vlasblom, Edgard K. Crowdy,
Emerson Cho, Kyle Morrison, Ian M. Donaldson, and Shoshana J. Wodak. iRefWeb:
interactive analysis of consolidated protein interaction data and their supporting evi-
APPENDIX A

COPYRIGHT PERMISSIONS

Below are permissions for the use of materials in Chapter 2 and 3.

Permissions

As a general rule, permission should be sought from the rights holder to reproduce any substantial part of a copyrighted work. This includes any text, illustrations, charts, tables, photographs, or other material from previously published sources. Obtaining permission to re-use content published by Elsevier is simple. Follow the guide below for a quick and easy route to permission.

Permission guidelines

For further guidelines about obtaining permission, please review our Frequently Asked Questions below:

- When is permission required?
- When is permission not required?
- From whom do I need permission?
- How do I obtain permission to use photographs or illustrations?
- Do I need to obtain permission to use material posted on a website?
- What rights does Elsevier require when requesting permission?
- How do I obtain permission from another publisher?
- What is Rightslink?
- What should I do if I am not able to locate the copyright owner?
- What is Elsevier’s policy on using patient photographs?
- Can I obtain permission from a Reproduction Rights Organization (RRO)?
- Is Elsevier an STM signatory publisher?
- Do I need to request permission to re-use work from another STM publisher?
- Do I need to request permission to text mine Elsevier content?
- Can I post my article on ResearchGate without violating copyright?
- Can I post on ArXiv?

Yes. Authors can include their articles in full or in part in a thesis or dissertation for non-commercial purposes.
Copyright and License

- Copyright on any open access article in a journal published by BMC is retained by the author(s).
- Authors grant BMC a license to publish the article and identify itself as the original publisher.
- Authors also grant any third party the right to use the article freely as long as its integrity is maintained and its original authors, citation details and publisher are identified.
- The Creative Commons Attribution License 4.0 formalizes these and other terms and conditions of publishing articles.
- In accordance with our Open Data policy, the Creative Commons CC0 1.0 Public Domain Dedication waiver applies to all published data in BMC open access articles.

Reprints and permissions

Reprint service

Reprint services are available for those requiring professional quality reproductions of articles.

Open access articles

The open access articles published in BMC's journals are made available under the Creative Commons Attribution (CC-BY) license, which means they are accessible online without any restrictions and can be reused in any way, subject only to proper attribution (which, in an academic context, usually means citation).

The re-use rights are enshrined in our license agreement. Include the right for anyone to reproduce printed copies themselves, without formal permission or payment of permission fees. As a courtesy, however, anyone wishing to reproduce large quantities of an open access article (500+) should inform the copyright holder and we suggest a contribution in support of open access publication.

Can I use an open access figure or table?

Modified on Thu, 7 Mar, 2019 at 11:44 PM

Reproduction of figures or tables from any open access article is permitted free of charge and without formal written permission from the publisher or the copyright holder, provided that:

- the figure/table is original
- the publisher is duly identified as the original publisher
- the proper attribution of authorship and the correct citation details are given as acknowledgment

Can I use an open access article?

Modified on Sun, 25 Nov, 2018 at 12:24 PM

Most open access articles are made available under the Creative Commons Attribution (CC-BY) license, which means they are accessible online without any restrictions and can be reused in any way, subject to proper attribution (which, in an academic context, usually means citation).

Please do check the license type of any article and how it can be used by accessing the Copyright information at the bottom of the article.

For more information about copyright and licensing for each of our open access brands, please see:

BMC  SPRIEGLOPEN  NATURE RESEARCH  PALGRAVE MACMILLAN  SPRINGER
The IEEE does not require individuals working on a thesis to obtain a formal reuse license, however, you may print out this statement to be used as a permission grant:

Requirements to be followed when using any portion (e.g., figure, graph, table, or textual material) of an IEEE copyrighted paper in a thesis:

1) In the case of textual material (e.g., using short quotes or referring to the work within these papers) users must give full credit to the original source (author, paper, publication) followed by the IEEE copyright line © [Year of original publication] IEEE.
2) In the case of illustrations or tabular material, we require that the copyright line © [Year of original publication] IEEE appear prominently with each reprinted figure and/or table.
3) If a substantial portion of the original paper is to be used, and if you are not the senior author, also obtain the senior author’s approval.

Requirements to be followed when using an entire IEEE copyrighted paper in a thesis:

1) The following IEEE copyright/credit notice should be placed prominently in the references: © [year of original publication] IEEE. Reprinted, with permission, from [author names], paper title, IEEE publication title, and month/year of publication.
2) Only the accepted version of an IEEE copyrighted paper can be used when posting the paper or your thesis on-line.
3) In placing the thesis on the author's university website, please display the following message in a prominent place on the website: In reference to IEEE copyrighted material which is used with permission in this thesis, the IEEE does not endorse any of university/educational entity's name goes here's products or services. Internal or personal use of this material is permitted. If interested in reprinting/republishing IEEE copyrighted material for advertising or promotional purposes or for creating new collective works for resale or redistribution, please go to http://www.ieee.org/publications_standards/publications/rights/rights_link.html to learn how to obtain a License from RightsLink.

If applicable, University Microfilms and/or ProQuest Library, or the Archives of Canada may supply single copies of the dissertation.
Copyrights

Copyright and Licensing

For all articles published in MDPI Journals, copyright is retained by the authors. Articles are licensed under an open access Creative Commons CC BY 4.0 license, meaning that anyone may download and read the paper for free. In addition, the article may be reused and quoted provided that the original published version is cited. These conditions allow for maximum use and exposure of the work, while ensuring that the authors receive proper credit.

In exceptional circumstances articles may be licensed differently. If you have specific condition (such as one linked to funding) that does not allow this license, please mention this to the editorial office of the journal at submission. Exceptions will be granted at the discretion of the publisher.