# COMPUTATIONAL APPROACHES FOR FINDING DISEASE RELATED GENES AND RNAS

by

NEGIN FRAIDOUNI

Presented to the Faculty of the Graduate School of
The University of Texas at Arlington in Partial Fulfillment
of the Requirements
for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON
August 2019

# ABSTRACT

## COMPUTATIONAL APPROACHES FOR FINDING DISEASE RELATED GENES AND RNAS

Negin Fraidouni, Ph.D.

The University of Texas at Arlington, 2019

Supervising Professors: Gergely Zaruba

Finding candidate genes that could cause specific diseases has been the subject of many studies. This is an important research task, however in the biological experimentation domain it can be very expensive and time consuming. So an alternative way is to find gene expression values from partial measurements and try to predict the rest. By using computational methods, we can statistically estimate these relationships faster and in a more efficient way, providing domain experts suggestions on what exploration of likely relationships they should be focusing. One common computational approach is to model the gene expression data as a matrix (where each row represents a gene and each column a subject); the entries of the matrix can then be mRNA measurements that show the extent of gene expressions. Since entries of the dataset are based on partial measurements, the dataset has missing values, and the problem is then to estimate the missing values and thus to recover the global matrix based on the known values. The main aim of this research is to investigate matrix completion methods for predicting gene expression values.

## TABLE OF CONTENTS

## LIST OF ILLUSTRATIONS

Figure Page

# LIST OF TABLES

Table Page

CHAPTER 1

# Introduction

## 1.1 Gene expression

In the center of all living cells, there is a nucleus and each nucleus contains 46 thread-like structures called chromosomes. Chromosomes are made of DNA that carry the genetic information. A gene is the basic unit of heredity and act as instructions to make molecules called proteins. The Human Genome Project has estimated that humans have between 20,000 and 25,000 genes.

Gene expression is a very sophisticated and highly regulated biological mechanism as it is responsible for the function of every living cell. Thousands of genes are expressed in every cell; every step of this expression process (making RNA from DNA and later making proteins from RNA) has a control point that determines which proteins are present in a particular cell and in what quantity. The amount of messenger RNA (mRNA) molecules in cells depends on the cell's function, so measuring the quantity of mRNA molecules can reflect gene expression. There are biological experimental methods to measure gene expression in biological samples. These methods have given researchers new opportunities to study the relationships between genes and diseases. Some example methods are Reporter gene [1], Microarray, and RNA sequencing [2, 3].

Generally speaking, a single gene is usually not responsible for regulating any complex process. If we consider the definition of a gene as DNA sequence that makes RNA or proteins, every process in the body would be the result of interactions between

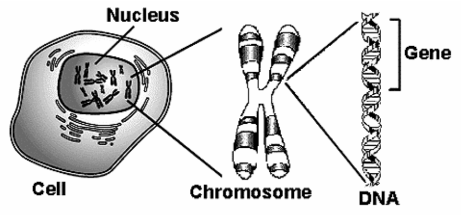

Figure 1.1. Structure of cell nucleus, chromosomes and genes.

these gene products. In other words, groups of genes must always work together in order to regulate any process in the body. Biologists believe that some proteins that have roles in the transcription of DNA molecules, named transcription factors, may have caused this and as a result the expression levels of genes are significantly correlated [4]. Recently the costs of data storage have reduced but the process of capturing data is still expensive and also the overall time needed for the whole process (preparing samples, measurements, storing data ...) is excessive. Computational methods relying on statistical forecasting can provide promising directions for these studies; thus there is a high demand for novel, efficient techniques and algorithms to predict gene expressions. Gene expression data can be stored in matrices in order to create a model, where the rows and columns correspond to different subjects and genes respectively. The entries of the matrix are the RNA levels found in the tissue samples of the different subjects. This model also makes the implicit assumption that similar expressions are present for people with similar diseases, making this matrix heavily over determined and thus considerably low-ranked. As we mentioned before, we can find a fraction of the data by partial measurements and predict the rest by computational approaches. Thus, the main goal is to complete or recover the matrix (i.e., to find the best representation for the missing values) when we can only observe a subset of the matrix's values, which is significantly less than the total size of the matrix. The most promising way to do this is to rely on the assumption that the resulting matrix has to be of low rank. By predicting missing values we could indeed predict gene expression patterns of which the most promising could be then investigated using biological experimentation.

## 1.2 MicroRNA (MiRNA)

MiRNAs are small (about 21 to 24 nucleotides), non-coding, single strand RNA molecules; they are involved in gene expression regulation. MiRNAs are found in most eukaryotes including those of humans, and they tend to bind target mRNA and prevent protein production [5]. MiRNA-directed gene expression regulation is a very active area of research. Hundreds of miRNAs have been discovered and the recent development of sequencing techniques and bioinformatics prediction methods significantly enhanced our information about miRNAs, including possible functions and regulatory targets [6–8]. MiRNAs have been found to be responsible for different processes including cell death, cell proliferation, neural patterning, immunity, fat metabolism, and hematopoietic differentiation [9]. Computational methods for finding genes regulated by miRNAs have suggested that all these examples only represent a few samples and thus they cannot describe the whole miRNA system [10]. Dysregulation of miRNAs have been shown to be the main reason of abnormal cell behavior and hence some human diseases. More and more miRNAs have been confirmed to be responsible for the development of human diseases [11, 12]. For instance studies confirmed that the miR-200 family has a strong association with breast cancer [13]; also, leukemia is one of the human cancers confirmed to be related to miR-15 and miR-16 dysregulation [14]. So recognizing miRNA-disease associations can help in diagnosing, treating, and preventing human diseases. However, it is prohibitive to find the associations one-by-one due to the significant amount of resources that have to be spent in performing such experiments. Meanwhile, known miRNA-disease associations are stored in databases like HMDD v.2.0 [15], dbDEMC [16] and miR2Disease [17] but there is a high demand for identifying new associations. Using computational methods to prioritize potential miRNAs for any specific miRNA-disease study could significantly reduce the time and financial resources needed for these experiments. Many computational methods that have been developed by scientists, are based on the assumption that similar miRNAs are likely to be related to similar diseases [18–20].

### 1.3 Long non-coding RNA(lncRNA)

Recent studies of transcriptomes have shown that a much greater part of the genome is transcribed than we knew and expected. The results of the transcription process are mostly non-protein coding RNAs [21] including Long non-coding RNAs(lncRNAs). LncRNAs are usually larger than 200 nucleotides and they are less expressed and more tissue-specific compared to protein-coding RNAs [22]. LncRNAs are thought to have almost 30,000 different types in humans and they consist the majority of the non-coding transcriptome. Although our knowledge about lncRNAs are very limited, they could have a very significant effect on regulation of transcription process [23, 24]. LncRNAs also can be responsible for cell differentiation, apoptosis and cell differentiation. Mutations and dysfunctions of lncRNAs are thought to be the reason of some human complex diseases like diabetes [25], neurodegeneration disease [26], cardiovascular diseases [27], colon cancer [28], prostate cancer [29], kidney cancer [30] and AIDS [31]. Computational methods can provide more efficient directions for these studies; so there is a high demand for novel, efficient techniques to predict lncRNA-disease association.

### 1.4 Matrix Completion Techniques

Probably the most often cited example for applied matrix completion is movie raters and recommenders for online streamers, like Netflix. The problem model contains an M*N matrix where each row represents a user and each column stands for a specific movie; each entry (m, n) then represent a rating that user m has for movie n. Although some of these values are provided by the users themselves, most entries are likely missing (as having the complete matrix would mean that all users have watched and explicitly rated all movies, and thus no prediction on how well a user would like a yet non-viewed movie would be needed). The goal is to predict missing values in order to make good movie recommendations for users (relying on the assumptions that similar users like similar movies and that these similarities are represented by similar ratings on the same movies). The goal of Matrix completion then is to find estimates for the missing values that result in the lowest rank matrix. The Matrix completion problem has been shown to be NP-hard but there are heuristic algorithms that can be used to recover the matrix with high probability [6].

### 1.5 Robust Principal Component Analysis

Principal component analysis (PCA) is a statistical tool that is mainly used for dimensionality reduction. Given a data matrix E , using PCA, we can find the most significant orthogonal vectors that show most variability in the data. For a

noise-free dataset, we can easily perform PCA using singular value decomposition (SVD). In the presence of noise, we can use another approach called robust PCA (RPCA) [33]. The presence of this noise is common in many applications such as image processing [34] and bioinformatics [35]. Robust PCA has the ability to recover a low rank matrix from sparse noise. Assume that our data matrix E is denoted by:

$$E = Y + S \tag{1.1}$$

Where $Y$ is a low-rank matrix capturing the noiseless data and S is a sparse matrix (i.e., for capturing noise, where most values are zero except for a small set of values that can be non-zero). The goal is then to estimate Y and S given some constraints on noise and data magnitude. For RPCA to work efficiently we need to know the location of the non-zero elements in S . We consider the problem of low rank matrix recovery using RPCA, where our goal is to recover $Y$ , given a matrix $E$ with missing values and knowing the location of those missing values (forming our matrix S ). Our RPCA problem can be solved by:

$$min_{Y,S}(\|Y\|_* + \lambda\|P_\Omega(S)\|_1) \tag{1.2}$$

such that: $E = Y + S$

where $\lambda$ is a parameter. There is a plethora of algorithms for solving convex optimization problems. One of the more robust iterative algorithms is known as the Alternating Direction Method of Multipliers (ADMM [36]).

## 1.6 Datasets

In chapter 2 and 3, we used the genomic data repositories of NCBI (National Center for Biotechnology Information). The GEO-NCBI (Gene Expression Omnibus) is a public repository of genomic data. GEO profiles show expression profiles for individual genes. We used the gene expression datasets of three studies that measured the mRNA levels of different genes in different subjects (the amount of mRNA levels show the gene expression values). We used the complete dataset as a reference and in each part of the experiment we removed some random portion (between 10% and 90%) of the data and then tried to recover the missing values. We then compared the original and the reconstructed matrix to measure how well the algorithm is performing.

The database we have used for chapter 4, contains data of the associations between human miRNA and disease from the Human microRNA Disease Database

(HMDD). The database includes about 579 miRNAs, 384 diseases and 10,381 experimentally confirmed associations between miRNAs and diseases. Using this data, we can construct matrix Y to capture the associations between miRNAs and diseases. Each row of matrix Y represents a different miRNA and each column represents a different disease. Based on the datasets, the elements of

matrix Y can only be either 0 or 1. If miRNA $m_i$ is associated with disease $d_j$, then $Y_{ij}$ is 1. $Y_{ij}$ Is 0 in the cases where there is no known association between $m_i$ and $d_j$(this does not mean that there is no relationship but merely that the relationship is unknown).

### 1.7 Methods

In chapter 2 we present the correlation based matrix completion model. As described earlier, there is strong evidence that genes are highly correlated and work in groups. In most biological processes we can find groups of genes that work together so if we measure gene expressions on a set of tissue samples, we should find groups of genes that are correlated to each other. Having this in mind, we can assume that our gene expression measurements can be a low rank matrix

so we can predict the missing values and complete the matrix. In neighborhood based approaches for collaborative filtering in recommendation systems, the main goal is to find similarities between neighbors. When there is a missing value, the system tries to make a prediction based on other users' ratings (for the movies); the more similar a user is to the one that has a missing value, the more impact his/her rating should have on the prediction. Our problem in gene expression prediction is very similar to neighborhood based approach of collaborating filtering, with the main task to find similarity (correlation) between genes. We will thus investigate using Pearson correlation coefficient and Cosine similarity to measure linear dependencies between genes. In chapter 3 we described how Robust Principal Component Analysis (RPCA) can be applied and used on NCBI-GEO biological data to find (artificially introduced) missing values and recover the datasets. After describing the RPCA approach, we presented the Alternating Direction Method of Multipliers (ADMM) algorithm.

We then described three well known algorithms that can be used when recovering low rank matrices and we compared the performances of the four approaches. To do this, we removed random elements from the datasets as represented by matrices and predicted them based on the assumption that genes have similar behaviors in similar conditions. Our study provides an insight for future work especially in biomedicine but also has implications to recommender systems. We found that ADMM approach outperforms the other three approaches, i.e., it predicted more accurate values. We hope that this study can open new opportunities to gene expression studies. As we stated earlier, gene expression experiments are very expensive and time consuming so using such computational methods can help biologists identify promising directions for studies based on partial measurements in gene expression experiments.

In chapter 4 we investigated a graph regularized matrix factorization approach for miRNA-disease association prediction. We assumed that similar miRNAs (functionally) tend to be related to similar diseases (phenotypically). We used miRNA functional similarity, disease semantic similarity, and known miRNA-disease associations form the HDMM v.2.0 database. To verify the accuracy of the GRMF method, we used five repetitions of 6-fold cross validation. We compared the result of the GRMF method with three state-of-the- art methods and concluded that GRMF outperforms the other three in terms of AUC. We also selected Breast Neoplasm as a case study in order to show the performance of GRMF for diseases which have no related miRNAs and based on the results, we could confirm all 50 miRNAs as identified by miR2Disease, dbDEMC and HDMM. As the second case study we chose Lymphoma to demonstrate the performance of GRMF and based on the results, we could confirm 45 miRNAs out of 50 as identified by dbDEMC, miR2Disease and experimental literature in PubMed. The GRMF method could provide an effective approach to study miRNA-disease associations. We also recognize that GRMF has some limitations which can be improved in future research. For example, the sequence information of miRNAs is used to measure miRNA similarity but some studies show that the structural information can be more effective. Furthermore, expression information of miRNAs could also be used to measure this similarity.

# REFERENCES

[1] Welsh S, Kay S.: Reporter gene expression for monitoring gene transfer Current Opinions in Biotechnology (1997) 617-622.

[2] Feng X, He X.: Inference on low rank data matrices with applications to microarray data. The Annuals of Applied Statistics (2010) 217-243.

[3] Wang Z, Gerstein M, Snyder M.: Rna-seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics (2009) 57-63.

[4] A. Wong A, Au W, Chen K.: Discovering high-order patterns of gene expression levels Journal of Computational Biology (2008) 625-637.

[5] MacFarlane L, Murphy R.: MicroRNA: Biogenesis, Function and Role in Cancer Current Genomics (11) (2010).

[6] Lai E, Tomancak P, Williams R, Rubin G.: Computational identification of Drosophila MicroRNA genes Genome Biology (4) (2003).

[7] Li S, Pan C, Lin W.: Bioinformatics discovery of microRNA precursor from human ESTs and introns BMC Genomics (7) (2006).

[8] Nam J, Shin K, Han J, Lee Y, Kim V, Zhang B.: Human microRNA prediction through a probabilistic co-learning model of sequence and structure Nucleic Acids Research (33) (2005) 3570-3581.

[9] He L, Hannon G.: MicroRNAs: small RNAs with a big role in gene regulation Nature Reviews Genetics (5) (2004) 522-531.

[10] Wahid F, Shehzad A, Khan T, Youngkim Y.: MicroRNAs: Synthesis, mechanism, function, and recent clinical trials Biochimica et Biophysica Acta (1803) no. 11, (2010) 1231-1243.

[11] Lynam-Lennon N, Maher S, Reynolds J.: The roles of microRNA in cancer and apoptosis Biological Reviews of the Cambridge Philosophical Society (84) (2009) 55-71.

[12] Meola N,Gennarino V, Banafi S.: MicroRNAs and genetic diseases Pathogenetics (2) no. 7 (2009).

[13] Lim Y, Wright J, Attema J, Gregory P, Bert A, Smith E.: Epigenetic modulation of the miR-200 family is associated with transition to a breast cancer stem-cell-like state Journal of Cell Science (126) (2013) 2256-2266.

[14] Callin G.: Frequent deletion and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia in Proceedings of the National Academy of Sciences of the United States of America (2002).

[15] Li Y.: HMDD v.2.0: a database for experimentally supported human microRNA and disease associations Nucleic Acids Research (42) (2013).

[16] Yang Z, Wu L, Wang A.: dbDEMC 2.0: updated database of differentially expressed miRNAs in human cancers Nucleic Acids Research (45) (2017).

[17] Jiang Q, Wang Y, Hao Y, Juan L, Teng M, Zhang X, Li M, Wamg G, Liu Y.: miR2Disease: a manually curated database for microRNA deregulation in human disease Nucleic Acids (2009).

[18] Chen X, Yan G.: Semi-supervised learning for potential human microRNA-disease associations inference Scientific Reports (40) no. 1 (2014).

[19] Chen H, Zhang Z.: Similarity-based methods for potential human microRNA-disease association prediction BMC Med Genomics (6) no. 12 (2013).

[20] Li J, Wu Z, Cheng F, Li W, Liu G, Tang Y.: Computational prediction of microRNA networks incorporating environmental toxicity and disease etiology Scientific Reports (4) (2014).

[21] Carninci P, Kasukawa T, Katayama S.: The transcriptional landscape of the mammalian genome. Science (309) (2005) 1559–63.

[22] Cabili MN, Trapnell C, Goff L.: Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Development (18) (2011) 1915-27.

[23] Chen X, Huang L.: LRSSLMDA: Laplacian Regularized Sparse Subspace Learning for MiRNA-Disease Association prediction. PLOS Computational Biology (13) (2017).

[24] You ZH, Huang ZA, Zhu Z, et al.: PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. PLOS Computational Biology (13) (2017).

[25] Li A, Zhang Z. : Role of long non-coding RNA in diabetes mellitus and its complications. Sheng Wu Gong Cheng Xue Bao(32) (2016) 284–91.

[26] Johnson R.: Long non-coding RNAs in Huntington's disease neurodegeneration. Neurobiological Diseases (46) (2012) 245–54.

[27] Busch A, Eken SM, Maegdefessel L.: Prospective and therapeutic screening value of non-coding RNA as biomarkers in cardiovascular disease. Annals of Translational Medicine (4) (2016).

[28] Di Cecilia S, Zhang F, Sancho-Medina A, et al.: RBM-AS1 is critical for self-renewal of colon cancer stem-like cells. Cancer Research (2016).

[29] Atala A. : Re: the long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. Journal of Urology (192) (2014).

[30] Fenner A. Kidney cancer: AR promotes RCC via lncRNA interaction. Nature Reviews Urology (13) (2016).

[31] Zhang Q, Chen CY, Yedavalli VS, et al. NEAT1 long noncoding RNA and paraspeckle bodies modulate HIV-1 posttranscriptional expression. American Society for Microbiology (4) (2013).

[32] Candes E, Recht B.: Exact matrix completion via convex optimization Applied and Computational Mathematics (2008).

[33] Lois B, Vaswani N.: Online matrix completion and online robust PCA in IEEE International Symposium on Information Theory (ISIT) (2015).

[34] Podosinnikova A, Setzer S, Hein M.: Robust PCA: optimization of the robust reconstruction error over the stiefel manifold in German Conference on Pattern Recognition(GCPR) (2015).

[35] Liu J, Wang J, Zheng C, Sha W, Mi J, Xu Y.: Robust PCA based method for discovering differentially expressed genes BMC Bioinformatics (2013).

[36] Zhong Y.: Alternating direction method of multipliers(ADMM) [Online]. Available: https://piazza-resources.s3.amazonaws.com.

CHAPTER 2

# A Correlation Based Matrix Completion Approach to Gene Expression Prediction

Negin Fraidouni, Gergely Zaruba

2017 IEEE 7th International Conference on Computational Advances in Bio and Medical Sciences (ICCABS), 19-21 Oct. 2017, Orlando, FL, USA.

## 2.1 Abstract

Finding candidate genes that could cause specific diseases has been the subject of many studies. This is an important research task, however in the biological experimentation domain it can be very expensive and time consuming. So an alternative way is to find gene expression values from partial measurements and try to predict the rest. By using computational methods, we can statistically estimate these relationships faster and in a more efficient way, providing domain experts suggestions on what exploration of likely relationships they should be focusing. One common computational approach is to model the gene expression data as a matrix (where each row represents a gene and each column a subject); the entries of the matrix can then be mRNA measurements that show the extent of gene expressions. Since entries of the dataset are based on partial measurements, the dataset has missing values, and the problem is then to estimate the missing values and thus to recover the global matrix based on the known values. In this paper, we present a correlation based approach to the matrix completion task (CMC) and discuss its functionality. The CMC based algorithm is then compared to a state-of-the-art nuclear-norm minimization iterative algorithm. Our results show that the CMC based algorithm significantly outperforms the iterative algorithm and even shows a better tendency when the amount of missing values grow. We argue that the CMC based algorithm can provide good estimates for missing values, possibly guiding time consuming biological gene expression profiling as to which values should be confirmed. Although our primary focus is on gene expression prediction, the strategy discussed is applicable to any highly correlated dataset where missing values need to be estimated/recovered.

## 2.2 Introduction

Gene expression is a very sophisticated and highly regulated biological mechanism as it is responsible for the function of every living cell. Thousands of genes are expressed in every cell; every step of this expression process (making RNA from DNA and later making proteins from RNA) has a control point that determines which proteins are present in a particular cell and in what quantity. The amount of messenger RNA (mRNA) molecules in cells depends on the cell's function, so measuring the quantity of mRNA molecules can reflect gene expression. There are biological experimental methods to measure gene expression in biological samples. These methods have given researchers new opportunities to study the relationships between genes and diseases. Some example methods are Reporter gene [1], Microarray, and RNA sequencing [2, 3]. Recently the costs of data storage have reduced but the process of capturing data is still expensive and also the overall time needed for the whole process (preparing samples, measurements, storing data,...) is excessive. Computational methods relying on statistical forecasting can provide promising directions for these studies; thus there is a high demand for novel, efficient techniques and algorithms to predict gene expressions.

This paper discusses a Correlation based Matrix Completion algorithm (CMC) with which such forecasting can be done. More precisely, we describe how the CMC algorithm can be applied to gene expression data so missing values can be predicted computationally. We use two approaches to calculate correlation between genes, the Pearson Correlation Coefficient (PCC) and the Cosine Similarity (CS). We then compare the result of these two methods to a state-of-the-art iterative matrix rank minimization algorithm [4] and show their performance advantages.

The rest of the paper is organized as follows. Section 2 describes the background information and related work with an emphasis on the low rank matrix completion approach of Kapur et al. [4]. Section 3 describes the dataset and the CMC approach. In Section 4, we present our computational experiment comparing the three approaches. Finally, section 5 concludes the paper.

### 2.3 Background

In this section we will describe gene expression data and the ideas behind matrix completion. Then we will provide an overview of similar works; finally we will provide a short overview of the nuclear norm minimization based matrix completion method of Kapur et al. [4].

Genes make proteins while proteins regulate all cell functions. Generally speaking, a single gene is usually not responsible for regulating any complex process. If we consider the definition of a gene as DNA sequence that makes RNA or proteins, every process in the body would be the result of interactions between these gene products. In other words, groups of genes must always work together in order to regulate any process in the body. Biologists believe that some proteins that have roles in the transcription of DNA molecules, named transcription factors, may have caused this and as a result the expression levels of genes are significantly correlated [5].

Gene expression data can be stored in matrices in order to create a model, where the rows and columns correspond to different subjects and genes respectively. The entries of the matrix are the RNA levels found in the tissue samples of the different subjects. This model also makes the implicit assumption that similar expressions are present for people with similar diseases, making this matrix heavily overdetermined and thus considerably low-ranked. As we mentioned before, we can find a fraction of the data by partial measurements and predict the rest by computational approaches. Thus, the main goal is to *complete* or *recover* the matrix (i.e., to find the best representation for the missing values) when we can only observe a subset of the matrix's values, which is significantly less than the total size of the matrix. The most promising way to do this is to rely on the assumption that the resulting matrix has to be of low rank. By predicting missing values we could indeed predict gene expression patterns of which the most promising could be then investigated using biological experimentation.

Probably the most often cited example for applied matrix completion is movie raters and recommenders for online streamers, like Netflix. The problem model contains an M*N matrix where each row represents a user and each column stands for a specific movie; each entry(m,n) then represent a rating that user-m has for movie-n. Although some of these values are provided by the users themselves, most entries are likely missing (as having the complete matrix would mean that all users have watched and explicitly rated all movies, and thus no prediction on how well a user would like a yet non-viewed movie would be needed).The goal is to predict missing values in order to make good movie recommendations for users (relying on the assumptions that similar users like similar movies and that these similarities are represented by similar ratings on the same movies). The goal of Matrix completion then is to find estimates for the missing values that result in the lowest rank matrix. The Matrix completion problem has been shown to be NP-hard but there are heuristic algorithms that can be used to *recover* the matrix with high probability [6].

### 2.3.1 Related work

Researchers always look for more effective approaches to extract relevant information from biological data. There is a vast and diverse amount of biological data available in online repositories and effective algorithms to make use of these data are highly sought after [7]. In the previous decade several machine learning algorithms have been developed or improved, causing a significant advancement in many aspects of finding gene relations (as they do not need tedious biological experimentation). There are studies that are based on developing tools for prioritizing disease genes; not surprisingly, they mostly use machine learning techniques. This is useful when we know which genes are responsible for which diseases but we do not actually know which ones play a more important role in causing that specific disease [8].

In some other studies researchers model the available biological data as a low rank matrix and then try to complete the matrix and find the missing values. Kapur et al. [4] showed that a known gene expression matrix can be artificially recovered using convex optimization methods. They show the applicability of their solution by removing some of the known values from matrix-contained biological datasets and then predict these values. By employing this recovery method, they can calculate an error between the predicted and actual value; the assumption then is that the algorithm will perform similarly well on a matrix with true unknown data. Natarajan et al. [21] used characteristics (or features) for diseases (extracted using methods like text mining) and genes to find a better prediction for gene-disease relations.

Collaborative filtering methods have been mainly developed to be used for recommendation systems. The neighborhood based collaborative filtering method [10] is based on collecting information about users' similarities so that this information can be used to make a prediction about movies that users will probably enjoy watching. Some component algorithms that are useful for collaborative filtering are K- Nearest Neighbors [11] and Pearson Correlation [12]. For example, El Alami et al. [13] employed Pearson correlation combined with a Jaccard similarity index for the purpose of similarity measures between neighbors.

### 2.3.2 Nuclear Norm Minimization Based Matrix Completion

In this subsection we will be briefly review the low-rank matrix recovery approach proposed in [4]. This low-rank minimization method is an iterative method, where better and better candidate solutions are found based on an error norm (the rank of the reconstructed matrix) and a gradient that helps calculating the next candidate solution. To be able to keep track which elements are missing we define $\Omega$; if an element $x_{i,j}$ of matrix X is known then $(i, j) \in \Omega$. The objective here is to:

$$min(rank(Y))$$

where

$$y_{(i,j)} = x_{(i,j)} \quad \forall (i, j) \in \Omega$$

The rank minimization problem is known to be NP-hard, to make the problem tractable, Kapur et al. minimize the nuclear norm instead of the rank. More precisely, the following minimization problem is solved (including a soft threshold operation):

$$minimise(\tau \left\|X\right\|_* + \frac{1}{2} \left\|X\right\|_F)$$

where is is the nuclear norm, $\tau$ is a threshold parameter with a recommended $\tau = 5 * \sqrt{M * N}$ (where M and N are matrix X's number of rows and columns respectively).

Matrix $X$ is iteratively reconstructed by applying a shrinkage method. In each iteration matrix $Y$ is deconstructed into its singular representation, the singular values that are smaller than $\tau$ are set to zero; and then matrix $Y$ is reconstructed from the new singular values and the original singular vectors. This process is repeated until the difference between original matrix and reconstructed matrix is less than a threshold tolerance value($\epsilon$): $\epsilon = \frac{\left\|P_\Omega(X^k - M)\right\|_F}{\|P_\Omega(M)\|_F}$. The pseudo code provided in Algorithm-1 presents pseudo code for the nuclear norm minimization based matrix completion approach.

**Algorithm 1: Nuclear Norm Minimization Based Matrix Completion**

```
Y = shrink (X, $\tau$):
    (U,$\Sigma$,V) = SVD(X)
    for every singular value $\sigma$ in $\Sigma$
        if $\sigma$ < $\tau$:
           $\sigma$ = 0
    $Y = U\ \Sigma\ V*$

M = minimization_MC(X, $\Omega$)
    $\delta$ = 1.2 * m * n / |$\Omega$|
    for each row i of X:
        for each column j of X:
            if (i,j) in $\Omega$ :
                P$\Omega$(X) = X
             else
                P$\Omega$(X) = 0
    $Y^0$ = 0
    k = 1
    while error < $\epsilon$:
        $X^k$ = shrink ( $Y^{k-1}$ , $\tau$)
        $Y^k$ =  $Y^{k-1}$ + P$\Omega$(M - $X^k$) * $\delta$
        k++
```

## 2.4 Correlation based matrix completion Model

In this section we will be describing the *neighborhood approach* of collaborative filtering. Then we will explain Pearson correlation coefficient, Cosine similarity, and Correlation based matrix completion.

As described earlier, there is strong evidence that genes are highly correlated and work in groups. In most biological processes we can find groups of genes that work together so if we measure gene expressions on a set of tissue samples, we should find groups of genes that are correlated to each other. Having this in mind, we can assume that our gene expression measurements can be a low rank matrix so we can predict the missing values and complete the matrix. In neighborhood based approaches for collaborative filtering in recommendation systems, the main goal is to find similarities between neighbors. When there is a missing value, the system tries to make a prediction based on other users' ratings (for the movies); the more similar a user is to the one that has a missing value, the more impact his/her rating should have on the prediction. Our problem in gene expression prediction is very similar to neighborhood based approach of collaborating filtering, with the main task to find similarity (correlation) between genes. We will thus investigate using Pearson correlation coefficient and Cosine similarity to measure linear dependencies between genes. A more detailed description of the Pearson correlation is provided in the next section.

### 2.4.1 Pearson Correlation Coefficient

A common measure of correlation between two variables is the Pearson Correlation Coefficient (PCC), aka, Pearson Product-Moment Correlation Coefficient (PPMCC). The PCC: $r$, indicates the strength of a linear association between two variables. Suppose that we have two datasets X and Y so that $X = \{x_1, ..., x_n\}$and$Y = \{y_1, ..., y_n\}$, we can calculate the PCC as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

or using a shorthand notation:

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

The PCC can take on values from the range: $[-1, 1]$. A value of 0 indicates no association between two variables. A positive value shows a positive association meaning that the values for two variables increase and decrease together. On the other hand a negative value shows a negative association and it means when the value of one variable increases, the value of the other one decreases and vice versa. A PCC closer to +1 or -1 means a stronger association (correlation).

### 2.4.2 Cosine Similarity

Cosine similarity measures the similarity between two vectors by calculating the cosine of the angle between them. This metric measures orientation and not magnitude so it can be seen as comparison on a normalized space that only the angle between two vectors matters and not their magnitude. One of the reasons for the popularity of cosine similarity is that it is very efficient to evaluate, especially for sparse vectors, as only the non-zero dimensions need to be considered.The cosine similarity can be calculated as shown below:

$$cos\theta = \frac{\vec{x}.\vec{y}}{\|\vec{x}\|.\|\vec{y}\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\sqrt{\sum_{i=1}^{n} y_i^2}}$$

with the dot product:

$$\vec{x}.\vec{y} = \|\vec{x}\|.\|\vec{y}\|.cos\theta$$

The resulting similarity ranges from -1 (meaning exactly opposite) to 1 (meaning exactly the same), with 0 indicating orthogonality, and in-between values indicating intermediate similarity or dissimilarity. The cosine similarity is independent of the two vectors' magnitudes.

### 2.4.3 Correlation based Matrix Completion

In this section we will describe the correlation based method used to complete partial matrices. Let us denote a known low-rank matrix as $X$. All of the values of $X$ are assumed to be known; we acknowledge that this is not the case in real-life but for the sake of the denotation we will assume that this matrix actually exists (and thus we can use it as the ground truth). The problem then becomes how to recreate matrix $X$ (or rather estimate matrix $X$) when some of its values are missing, erased or unknown; We will refer to the matrix with the missing values as $X^{'}$. Finally, we will refer to the reconstructed matrices as $X_P$ and $X_C$ when using PCC and CS in the reconstruction process, respectively. The ultimate goal is for $X_P$ and $X_C$ to be as close to $X$ as possible (see section on error calculation).

To use the notation from [4] we define a set of indices $\Omega$ such that if the matrix element $x_{i,j}$ is known then $(i,j) \in \Omega$.

We will calculate a PCC value and a CS value for every pair of rows. More precisely if $(i,j) \notin \Omega$ then we will calculate all $PCC_k$ and $CS_k$ values between row $i$ and all other rows $k$ $(k \neq i)$. To do this, we will have to skip using a value in column $j$ if either $(i,j) \notin \Omega$ or $(k,j) \notin \Omega$. Thus for each pair of rows $i$ and $k$ we will have two scalars $PCC(i,k)$ and $CS(i,k)$ describing how similar they are in their non-missing values. To estimate a missing value at $(i,j)$ then we will calculate:

$$X_{P_{i,j}} = \frac{\sum_{k=1}^{N} PCC^{'}(i,j,k) * x^{''}_{k,j}}{\sum_{k=1}^{N} y(k,j)} \tag{2.1}$$

$$X_{C_{i,j}} = \frac{\sum_{k=1}^{N} CS^{'}(i,j,k) * x^{''}_{k,j}}{\sum_{k=1}^{N} y(k,j)} \tag{2.2}$$

where

$$x^{''}_{k,j} = \begin{cases} x^{'}_{k,j} & (k,j) \in \Omega \\ 0 & (k,j) \notin \Omega \end{cases} \tag{2.3}$$

$$PCC^{'}_{i,j,k} = \begin{cases} PCC(i,k) & (k,j) \in \Omega \\ 0 & (k,j) \notin \Omega \end{cases} \tag{2.4}$$

$$CS^{'}_{i,j,k} = \begin{cases} CS(i,k) & (k,j) \in \Omega \\ 0 & (k,j) \notin \Omega \end{cases} \tag{2.5}$$

$$y(k,j) = \begin{cases} 1 & (k,j) \in \Omega \\ 0 & (k,j) \notin \Omega \end{cases} \tag{2.6}$$

Essentially what Equation 2.1 represents is a mean value over all known values in the column weighted by how similar two rows are based on PCC values and also Equation 3.2 represents a mean value over all known values in the column weighted based on CS values. To help normalize the mean, $\sum_k y(k,j)$ is the number of rows that do contain a value in column j. The pseudo code provided in Algorithm-2 presents an algorithmic view on the correlation based matrix completion calculation as described above.

## 2.5 Evaluation

In this section we will be describing error calculation and performance of PCC based approach and CS based approach compared to nuclear norm minimization based matrix completion.

**Algorithm 2: Correlation based Matrix Completion**

```
X_P,X_C = CMC(X, Ω)
    for each row i of X
        for each row k of X
            if i != k :
                calculate PCC(k,i)
                calculate CS(k,i)

    for each row i of X
        for each column j of X
            if (i,j) in Ω :
                x_p(i,j) = x(i,j)
                x_c(i,j) = x(i,j)
            else if (i,j) not in Ω :
                x_p(i,j) = 0
                x_c(i,j) = 0
                y = 0
                for each row k of X
                    if (k,j) in Ω :
                        x_p(i,j) += x_p(k,j)*PCC(i,k)
                        x_c(i,j) += x_c(k,j)*CS(i,k)
                        y ++
                x_p(i,j) /=  y
                x_c(i,j) /=  y
```

### 2.5.1 Error Calculation

In order to determine how well the matrix completion algorithm works, we are going to calculate a scalar error that increases with an increase in the difference between the original matrix $X$ and the reconstructed matrix $X'$ (i.e., matrices $X_P$ and $X_C$). More precisely, we define the relative error (RE) as the Frobenius Norm of the difference of the original and reconstructed matrices normalized with respect to the Frobenius Norm of the original matrix:

$$RE = \frac{\|X - X'\|_F}{\|X\|_F} \tag{2.7}$$

The Frobenius norm is defined as the square root of the sum of the squares of all values in the matrix.

We will also use mean squared error (MSE) to evaluate the performance of the aforementioned methods; MSE can be an important metric when measuring the performance of a predictor. The MSE between N-element vectors x and y is defined as:

$$MSE = \frac{\sum_{i=1}^{N}(x_i - y_i)^2}{N} \tag{2.8}$$

### 2.5.2 Datasets

In our experiment we will use the genomic data repositories of **NCBI** (National Center for Biotechnology Information). The GEO-NCBI (Gene Expression Omnibus) is a public repository of genomic data. GEO profiles show expression profiles for individual genes. We will use the gene expression datasets of three studies that measured the mRNA levels of different genes in different subjects (the amount of mRNA levels show the gene expression values).

1. **Bladder cancer study** : "Combination of a novel gene expression signature with a clinical nomogram improves the prediction of survival in high-risk bladder cancer". In this study, Riester et al. evaluated microarrayed data from 93 patients with bladder cancer that had cystectomy to determine gene expression patterns. This dataset has 54675 rows and 93 columns [14].

2. **Leukemia study** : "Identification of genes with abnormal expression changes in acute myeloid leukemia". In this study, Stirewalt et al. compared gene expression profiles between normal hematopoietic cells from 38 healthy donors and leukemic blasts from 26 leukemia patients. This data set has 22283 rows and 64 columns [15].

3. **Lung cancer study** : "Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival". In this study, Landi et

al. performed gene expression analysis on 135 fresh frozen tissue samples of adenocarcinoma and non-involved lung tissue from current, former and never smokers, with biochemically validated smoking information.This data set has 22283 rows and 107 columns [16].

### 2.5.3 Performance

We used the complete dataset as a reference and in each part of the experiment we removed some random portion (between 10% and 90%) of the data and then tried to recover the missing values. We then compared the original and the reconstructed matrix to measure how well the algorithm is performing.

In order to compare the low-rank nuclear iterative matrix completion algorithm of [4] to the proposed approach, we had to implement both approaches and feed the same data to both of them. For nuclear norm minimization algorithm, we used parameters $\tau = 5\sqrt{MN}$ and $\delta = \frac{1.2MN}{|\Omega|}$ as stated in [4]. We used Python with the *numpy, scipy* and *sklearn* packages for implementation and visualization of the data. Each datapoint in our figures is an average over 10 separate experiments where different entries where randomly removed from the original matrix.

One important question is, how these algorithms perform at varying degrees of missing information. We made this a factor in our experimentation, i.e., we varied the percentage of data removal in the matrix. More precisely, we varied the proportion of missing data (to all values) in our experiments from 10% to 90% with 10% increments (nine measurement values each) percentage of the values.

Figure 2.1 shows the aggregated performance of the three approaches in Bladder cancer, Leukemia and Lung cancer studies. The vertical axis in plots A, C and E represents the relative error (RE) as described in Equation 2.7 and the vertical axis in plots B, D and F represents MSE as described in equation 2.8. The horizontal axis in all plots (A - F) shows the percentage of the missing values (the factor). The green line depicts the performance of the nuclear norm matrix completion approach, while the dotted red line depicts the PCC based correlation approach and the dotted blue line shows the CS based correlation approach detailed earlier. We can observe that the PCC based correlation approach not only consistently outperforms the minimization based approach in all three settings, but also has a different trend. While the green line shows a progressively increasing error, the error presented by the red curve depicts a decreasing tendency. The CS based approach also shows better results compared to nuclear norm minimization. When the proportion of missing elements is increased, the RE and MSE of PCC based CMC grew much slower than that of the nuclear norm minimization approach. The difference between PCC based CMC and CS based CMC

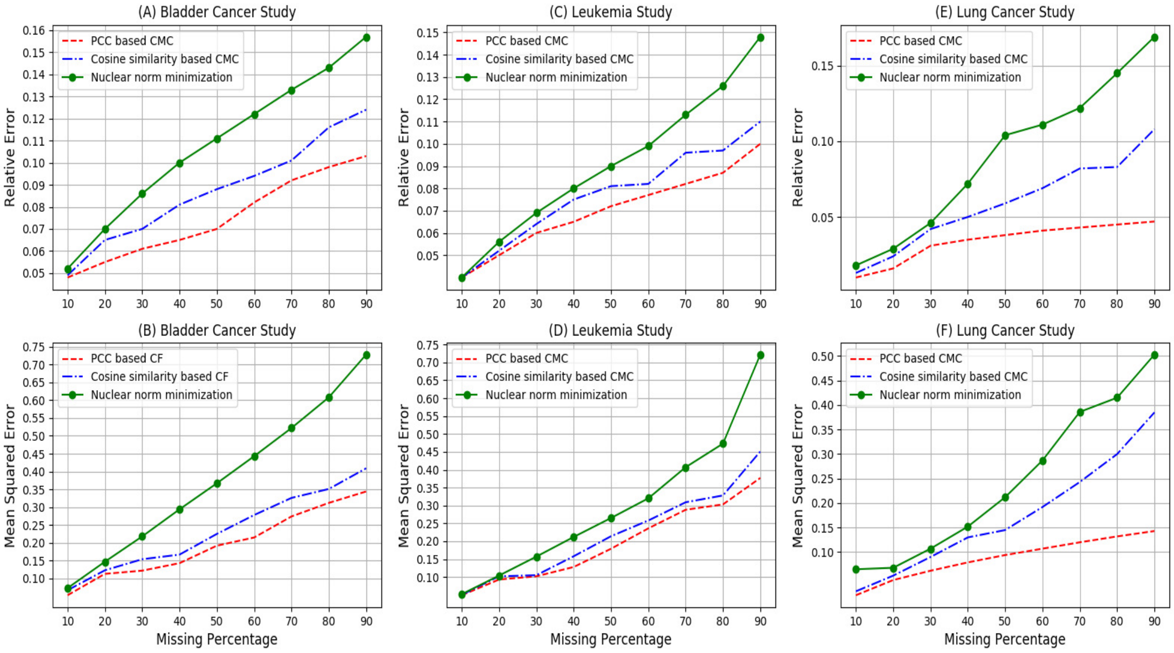

Figure 2.1. Comparison of matrix recovery methods on 3 studies from NCBI-GEO.

|  | **Dataset 1** |  |  | **Dataset 2** |  |  |
|---|---|---|---|---|---|---|
| **Method** | 30% Unknowns | 60% Unknowns | 90% Unknowns | 30% Unknowns | 60% Unknowns | 90% Unknowns |
| K = 50 | 0.321 | 0.452 | 0.524 | 0.354 | 0.386 | 0.498 |
| K = 100 | 0.152 | 0.232 | 0.387 | 0.232 | 0.264 | 0.358 |
| K = 150 | 0.095 | 0.151 | 0.212 | 0.098 | 0.136 | 0.274 |
| K = 250 | 0.115 | 0.145 | 0.259 | 0.112 | 0.124 | 0.255 |
| K = 400 | 0.121 | 0.215 | 0.309 | 0.235 | 0.265 | 0.365 |
| K = 550 | 0.132 | 0.265 | 0.354 | 0.322 | 0.398 | 0.471 |
| K = 700 | 0.354 | 0.422 | 0.521 | 0.458 | 0.487 | 0.548 |
| PCC-based-CMC | 0.0421 | 0.061 | 0.079 | 0.05 | 0.069 | 0.082 |

Table 2.1. The relative error of PCC based method versus KNN for 450K array methylation datasets.

could be explained by the Pearson correlation coefficient being better in capturing the correlation between genes compared to cosine similarity.

We also applied the PCC-based method to two 450K methylation datasets [17, 18] to predict missing values and to compare performance with KNN, which is the most used method for predicting missing values in methylation data. Table 1 shows the relative error of predicted values versus actual ones for 30%, 60% and 90% missing entries with the K values ranging from 50 to 700. When the number of neighbors is around 150-250, the relative error of KNN is the least. As Table 1 shows, the PCC-based approach outperforms KNN in the both methylation datasets for the all experiments with 30%, 60% and 90% data removal. Also KNN is very slow in case of large datasets.

## 2.6 Conclusions

In this paper we investigated how Pearson Correlation Coefficient and Cosine Similarity could be applied and used on NCBI-GEO biological data to find (artificially introduced) missing values in the datasets. After describing the matrix completion algorithms, we compared the performances of the two approaches to that of a recent nuclear-norm minimization based approach. To do this, we removed random elements from the datasets as represented by matrices and predicted them based on the assumption that subjects have similar tendencies; more precisely that characteristics of genes where genes work in groups for any process in body are similar. Our study provides an insight for future work especially in bio-medicine as well as recommender systems. We found the correlation based approaches to outperform the low nuclear-rank matrix completion approach, i.e., it predicted more accurate values. We have also found that Pearson correlation coefficient provides more accurate reconstructions when compared to cosine similarity when used on gene databases. We hope that this study can open new opportunities to gene expression studies. As we states earlier, gene expression experiments are very expensive and time consuming so biologists can perform partial measurements in gene expression studies and find the rest in a more fast and efficient way using computational approaches like the ones we employed in this paper.

## REFERENCES

[1] Welsh, S., Kay, S.: Reporter gene expression for monitoring gene transfer. Current Opinions in Biotechnology (8) (1997) 617–22.

[2] Wang, Z., Gerstein, M., Snyder, M.: Rna-seq: a revolutionary tool for transcriptomics. Nature Reviews Genetics (10) (2009) 57–63.

[3] Feng, X., He, X.: Inference on low rank data matrices with applications to microarray data. The Annuals of Applied Statistics (2010) 217–243.

[4] Kapur, A., Marwah, K., Alterovitz, G.: Gene expression prediction using low-rank matrix completion. BMC Bioinformatics (3) (2016) 1634–1654.

[5] Wong, A., Au, W.H., Chan, K.: Discovering high-order patterns of gene expression levels. Journal of Computational Biology (15) (2008) 625–637.

[6] Candes, E., Recht, B.: Exact matrix completion via convex optimization. Applied and Computational Mathematics (2008).

[7] Gligorijević, V., Pržulj, N.: Computational Methods for Integration of Biological Data. Springer International Publishing, Cham (2016) 137–178.

[8] Bromberg, Y.: Chapter 15: Disease gene prioritization. PLoS Computational Biology (9) (2013).

[9] Natarajan, N., Dhillon, I.S.: Inductive matrix completion for predicting gene-disease association. Bioinformatics (30) (2014) 60–68.

[10] Alqadah,F., Reddy, C. , Hu, J.: Biclustering neighborhood-based collaborative filtering method for top-n recommender systems Springer-Verlag London,(2014).

[11] Park, Y., Park, S., Lee, S. ,Jung, W.: Fast Collaborative Filtering with a k-nearest neighbor graph BIGCOMP,(2014), 92–95.

[12] Ekstrand, M., Riedl, J.T. , Konstan, J.: Collaborative Filtering Recommender Systems Foundations and Trends in Human–Computer Interaction,(4) (2011) 81–173.

[13] El Madani El Alami,Y., Nfaoui, E. H. , El Beqqali, O.: Improving Neighborhood-Based Collaborative Filtering by A Heuristic Approach and An Adjusted Similarity Measure Proceedings of the International Conference on Big Data, Cloud and Applications, Tetuan, Morocco,(2015).

[14] Riester M, Taylor JM, Feifer A, Koppie T et al: Combination of a novel gene expression signature with a clinical nomogram improves the prediction of survival in high-risk bladder cancer. Clinical Cancer Res (2012) 1; 18(5): 1323-33 PMID: 22228636.

[15] Stirewalt DL, Meshinchi S, Kopecky KJ, Fan W et al.: Identification of genes with abnormal expression changes in acute myeloid leukemia. Genes Chromosomes Cancer (2008); 47(1): 8-20 PMID: 17910043.

[16] Landi MT, Dracheva T, Rotunno M, Figueroa JD et al.: Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival. PLoS One (2008) 20; 3(2): e1651 PMID: 18297132.

[17] Larsson C, Ali MA, Pandzic T, Lindroth AM et al. : Loss of DIP2C in RKO cells stimulates changes in DNA methylation and epithelial-mesenchymal transition. BMC Cancer 2017 Jul 17; 17(1): 487 PMID: 28716088.

[18] Hammamieh R, Chakraborty N, Gautam A, Muhie S et al.: Whole-genome DNA methylation status associated with clinical PTSD measures of OIF/OEF veterans. Transl Psychiatry 2017 Jul 11; 7(7): e1169 PMID: 28696412.

CHAPTER 3

# A Robust Principal Component Analysis via Alternating Direction Method of Multipliers to Gene-Expression Prediction

Negin Fraidouni, Gergely Zaruba

2017 International Conference on Computational Science and Computational Intelligence (CSCI), 14-16 Dec. 2017, Las Vegas, NV, USA.

## 3.1 Abstract

Gene expression is the main process responsible for the function of every living cell. Thousands of genes expressed in a specific cell determine what that cell can do. Gene expression values can be measured by measuring the amount of messenger RNA (mRNA) molecules. There are biological methods to measure gene expression in biological samples so researchers can find genes responsible for each disease. Some example methods are Reporter gene, Microarray, and RNA sequencing. These methods however are very costly and time consuming. Computational methods have the potential to help these studies by identifying reliable directions using prediction techniques on incomplete data; so novel and efficient techniques and algorithms to predict gene expressions are in high demand. In this paper, we describe a method to recover gene expression dataset based on robust principal component analysis (RPCA). We treat the differentially expressed genes as sparse noise $S$ and non-differentially expressed genes as low-rank matrix $Y$. We show how $S$ and $Y$ can be recovered from gene expression data using RPCA. We also used existing implementations of three other iterative optimization based matrix completion methods to provide a comparative analysis of their performances. We show that this approach consistently outperforms the other methods with reaching improvement factors beyond 7.9 in measured mean squared error.

## 3.2 Introduction

Gene expression is the process by which the instructions contained in a gene are used in the synthesis of a functional product (i.e., RNA or proteins). Proteins are responsible for regulating all cell functions. We can measure the gene expression level by measuring the amount of RNA that is generated inside the cell. We can capture the gene expression data in matrices, where the rows and columns correspond to different genes and samples respectively; the entries of the matrix are the RNA levels found (in the given tissue sample of the given subject). It is generally believed that groups of genes work together so every process in the body would be the result of interactions between gene products. This means that the expression levels of genes should be highly correlated and the aforementioned matrix should be considerably low-ranked. Since gene expression measurements are highly expensive and time consuming, an alternative way is to find gene expression values from partial measurements and try to predict the rest based on computational approaches. Thus, a main goal here could be to find the best prediction for the missing values, when we can only observe a small subset of the matrix's values. Fortunately, there are many promising approaches to this matrix completion problem. In this paper we will investigate iterative solutions based on complex relaxation formulations to the matrix completion problem; more precisely, we will focus on a promising combination of Robust Principal Component Analysis (RPCA) and an efficient solver and show the benefits of this combined approach.

### 3.2.1 Low Rank Matrix Completion

Low rank matrix completion can be achieved by nuclear norm based non-convex optimization. One way of solving non-convex optimization problems is to use convex approximations instead of the original problem. Let $E \in R^{n_1 \times n_2}$ be a low rank matrix. The location of the known values can be encoded in $\Omega$, where $(i, j) \in \Omega$ if the value at indices (i,j) is known. We can define a function $P_\Omega(X)$ that returns a matrix where values in $\Omega$ are the same as the input matrix while it set the others to zero:

$$P_{\Omega}(X)_{i,j} \begin{cases} X_{i,j} & (i,j) \in \Omega \\ 0 & (i,j) \notin \Omega \end{cases} \tag{3.1}$$

To estimate the missing values we usually assume that the matrix is low rank. Essentially we are then trying to assign values to the missing entries to minimize the rank of the matrix. Rank minimization problem is NP-hard and a more computationally feasible version to rank minimization is minimizing the sum of singular values of data matrix $X$, as the sum of singular values has a direct relationship to the rank (the sum of the singular values is also known as nuclear norm or trace norm). Thus to find a good approximation, we can minimize the nuclear norm of matrix $X$ instead of minimizing its rank:

$$min(\|X\|_*) \tag{3.2}$$

$$\text{such that: } P_{\Omega}(X) = P_{\Omega}(E)$$

where:

$\|X\|_* = \sum_{i=1}^{r} \sigma_i$ , is the nuclear norm of matrix $X$, with $\sigma_i$ denoting the $i^{th}$ nonzero singular value of $X$.

There are many efficient iterative algorithms to solve problem 3.2. One way is to apply the singular value thresholding algorithm (SVT) [1]:

$$min(\tau \left\|X\right\|_* + \frac{1}{2} \left\|X\right\|_F) \tag{3.3}$$

$$\text{such that: } P_{\Omega}(X) = P_{\Omega}(E)$$

where:

- $\left\|X\right\|_F = \sqrt{\sum_{i=1}^{M}\sum_{j=1}^{N} (x_{ij})^2}$, is the Frobenius norm of matrix $X$

- $\tau > 0$, is a threshold parameter.

The Frobenius norm is used as a regularization in the above function, trying to control the magnitude of the values in the recovered matrix.

The rest of the paper is organized as follows. Section 2 describes the background information and related work with an emphasis on the robust PCA approach. Section 3 describes the alternating direction method of multipliers (ADMM) approach for solving convex optimization problems. Section 4 describes the convex optimization formulation of the ADMM matrix completion method. In section 5 we describe three competitive methods that used RPCA. In Section 6, we provide the datasets that we used and present our computational experiments comparing it with three RPCA based approaches. Section 7 concludes the paper.

### 3.3 Background

Principal component analysis (PCA) is a statistical tool that is mainly used for dimensionality reduction. Given a data matrix $E$, using PCA, we can find the most significant orthogonal vectors that show most variability in the data. For a noise-free dataset, we can easily perform PCA using singular value decomposition (SVD). In the presence of noise, we can use another approach called robust PCA (RPCA) [2]. The presence of this noise is common in many applications such as image processing [3] and bioinformatics [4]. Robust PCA has the ability to recover a low rank matrix from sparse noise. Assume that our data matrix $E$ is denoted by:

$$E = Y + S \tag{3.4}$$

where $Y$ is a low-rank matrix capturing the noiseless data and $S$ is a sparse matrix (i.e., for capturing noise, where most values are zero except for a small set of values that can be non zero). The goal is then to estimate $Y$ and $S$ given some constraints on noise and data magnitude. For RPCA to work efficiently we need to know the location of the non-zero elements in $S$.

In this paper we consider the problem of low rank matrix recovery using RPCA, where our goal is to recover $Y$, given a matrix $E$ with missing values and knowing the location of those missing values (forming our matrix $S$). Our RPCA problem can be solved by :

$$min_{Y,S}(\|Y\|_* + \lambda\|P_\Omega(S)\|_1) \tag{3.5}$$

$$\text{such that: } E = Y + S$$

Where $\lambda$ is a parameter. There is a plethora of algorithms for solving convex optimization problems. One of the more robust iterative algorithms is known as the Alternating Direction Method of Multipliers (ADMM). Before we dwell more into solving our optimization problem, in the next section we will revisit ADMM; then in section 4 we will investigate the solution to problem 3.5 in detail.

### 3.4 The Alternating Direction Method of Multipliers

The Alternating Direction Method of Multipliers (ADMM) is an iterative convex optimization solver. It has the robustness of the *Method of Multipliers* but can also decompose the search space into smaller pieces, each of which is then easier to handle. ADMM solves convex optimization problems of the following form:

$$min_{x,z} f(x) + g(z)$$

$$\text{such that: } Ax + Bz = c$$

The augmented Lagrangian function, for a parameter $\rho > 0$ with the Lagrangian multiplier $m$, can then be defined as:

$$\begin{aligned} L_\rho(x, z, m) = f(x) + g(z) + m^T(Ax + Bz - c) + \\ \frac{\rho}{2}\|Ax + Bz - c\|_F^2 \end{aligned} \tag{3.6}$$

With these definitions we can now perform an iterative algorithm, where at iteration k+1:

$$x^{k+1} = arg_x min L_\rho(x^k, z^k, m^k)$$

$$z^{k+1} = arg_z min L_\rho(x^{k+1}, z^k, m^k)$$

$$m^{k+1} = m^k + \rho(Ax^{k+1} + Bz^{k+1} - c)$$

The algorithm is performed until a predefined convergence criteria is met [5].

## 3.5 Convex optimization formulation

Let us now continue the exploration of problem 3.5 by applying the ADMM optimization to it. The resulting ADMM formulation of problem 3.5 is:

$$\begin{gathered} L_\rho(Y, S, m) = \|Y\|_* + \lambda\|P_\Omega(S)\|_1 + \\ m^T(P_\Omega(E - Y - S)) + \frac{\rho}{2}\|P_\Omega(E - Y - S)\|_F^2 \end{gathered} \tag{3.7}$$

where $m$ is the Lagrangian multiplier and $\rho > 0$ is the penalty parameter. At iteration k=1, each variable can then be calculated as:

$$Y^{k+1} = arg_{_Y} min L_\rho(Y^k, S^k, m^k)$$

$$S^{k+1} = arg_{_S} min L_\rho(Y^{k+1}, S^k, m^k)$$

$$m^{k+1} = m^k + \rho(E - Y^{k+1} - S^{k+1})$$

### 3.5.1 Updating Y:

In each iteration $k+1$ we can find $Y$ through:

$$\begin{aligned} Y^{k+1} = min_Y \|Y\|_* + m^T(P_\Omega(E - Y^k - S^k)) \\ + \frac{\rho}{2}\|P_\Omega(E - Y^k - S^k)\|_F^2 \end{aligned} \tag{3.8}$$

which is the solution of:

$$min_Y \|Y\|_* + \frac{\rho}{2}\|P_\Omega(Y^k + S^k - E) - \frac{m}{\rho}\|_F^2 \tag{3.9}$$

we can use a soft thresholding operation (from [6,7]) to solve problem 3.9. This way in each iteration we can update $Y$ through:

$$\begin{aligned} Y^{k+1} = shrink(A^k, \frac{1}{\rho}) \\ A^k = (E - S^k + \frac{m}{\rho^k}) \end{aligned} \tag{3.10}$$

where shrink is a soft thresholding operator:

$$
\begin{aligned}
shrink(A,b) := \sum_{i=1}^{r} u_i, max(\sigma_i - b, 0)v_i^T \\
A = \sum_{i=1}^{r}(u_i \sigma_i v_i^T)
\end{aligned}
\tag{3.11}
$$

here $\sigma_i$ are the singular values and $u_i, v_i$ are the singular vectors of matrix $A$. The value of $\frac{1}{\rho}$ determines the amount by which the singular values of matrix $A$ are decreased.

### 3.5.2 Updating S:

In iteration $k+1$ we can find $S$ via:

$$
\begin{aligned}
S^{k+1} = min_S \lambda \| P_\Omega(S^k) \|_1 + m^T(P_\Omega(E - Y^{k+1} - S^k)) \\
+\frac{\rho}{2} \| P_\Omega(E - Y^{k+1} - S^k) \|_F^2
\end{aligned}
\tag{3.12}
$$

which is the solution of:

$$
min_S \lambda \| P_\Omega(S^k) \|_1 + \frac{\rho}{2} \| P_\Omega(Y^{k+1} + S^k - E) - \frac{m}{\rho} \|_F^2 \tag{3.13}
$$

We can use a shrinkage operator to solve problem 3.13:

$$
\begin{cases}
S_{ij} = H_{\frac{\lambda}{\rho}}(E - Y^{k+1} + \frac{m}{\rho}) & (i,j) \in \Omega \\
\\
S_{ij} = 0, & (i,j) \notin \Omega
\end{cases}
\tag{3.14}
$$

where $H_{\frac{\lambda}{\rho}}$ is the shrinkage operator discussed in [8] as calculated via:

$$H_\sigma(S_{ij}) = \begin{cases} S_{ij} - \sigma, & S_{ij} > \sigma \\ S_{ij} + \sigma, & S_{ij} < -\sigma \\ 0, & Otherwise \end{cases} \tag{3.15}$$

to be able to solve the problem efficiently, we can assume that $S$ is zero at the indices that represent unknown values [9]. Combining all the above descriptions, we provide a pseudo-code for solving problem 3.5 in Algorithm 1.

**Algorithm 1: Solving problem 3.5 via ADMM**

```
Input: $E$, $\rho$ , $\lambda$, $\epsilon$
$m_0 = Y_0 = S_0 = 0$
while $\| E - Y^{k+1} - S^{k+1} \|_F > \epsilon$:

    Updating Y:
    $Y^{k+1} = arg_Y \ min \ L_\rho(Y^k, \ S^k, m^k)$
    $Y^{k+1} = shrink \ ((E - S^k + \frac{m}{\rho}), \rho^{-1})$:
        $(U, \Sigma, V) = SVD(E - S^k + \frac{m}{\rho^k})$
        for every singular value $\sigma$ in $\Sigma$
            if $\sigma$ < $\rho^{-1}$:
               $\sigma$ = 0
        $Y^{k+1} = U \ \Sigma \ V^T$
    Updating S:
    $S^{k+1} = arg_S \ min \ L_\rho(Y^{k+1}, \ S^k, m^k)$
    for each row i of S:
        for each column j of S:
            if (i,j) in $\Omega$ :
```

```
            $S_{ij} \;=\; H_{\frac{\lambda}{\rho}}(E - Y^{k+1} + \frac{m}{\rho})$
        else
            $S_{ij}$ = 0
    Updating m:
    $m^{k+1} \;=\; m^k \; + \rho \; (E - Y^{k+1} \; - \; S^{k+1})$

Output: $Y^{k+1}$ $and$ $S^{k+1}$
```

### 3.6 Competitive methods

To evaluate and compare the ADMM approach, we measure the recovery of the gene expression data matrix to the recovery provided by three similar approaches:

**1. Singular Value Thresholding algorithm (SVT)**: The singular value thresholding algorithm is a baseline for the matrix completion task proposed in [6] and it solves the robust PCA relaxation of:

$$min_{A,E}\|A\|_* + \lambda\|E\|_1 + \frac{1}{2\tau}\|A + E\|_F^2 \tag{3.16}$$

$$\text{subject to: } A + E = D$$

A pseudo-code to SVT is provided in Algorithm 2, where:

$$\begin{cases} US_\tau[S]V^T = argmin_x \tau\|X\|_* + \frac{1}{2}\|X - W\|_F^2 \\ S_\tau[W] = argmin_x \tau\|X\|_1 + \frac{1}{2}\|X - W\|_F^2 \\ USV^* \text{ is the SVD of W.} \end{cases} \tag{3.17}$$

**Algorithm 2: Singular value thresholding algorithm(SVT)**

```
Input: Observation matrix $D$, $\tau$, $\lambda$
while not converged do:
    $(U, S, V) \;=\; svd(Y_k)$
    $A_{k+1} \;=\; US_{\tau}[S]V^*$
    $E_{k+1} \;=\; S_{\lambda\tau}[Y_k]$
    $y_{k+1} \;=\; y_k \;+\; \sigma_k(D - A_{k+1} - E_{k+1})$
end while
Output: $A \;=\; A^{k+1}$ , $E \;=\; E^{k+1}$
```

**2. Exact Augmented Lagrangian Multiplier (EALM)**: This method was proposed in [10], the formulation of the problem is:

$$\begin{aligned} f(X) &= \|A\|_* + \lambda\|E\|_1 \\ h(X) &= D - A - E \end{aligned} \tag{3.18}$$

and the Lagrangian function of the problem is:

$$\begin{aligned} L(A, E, y, \mu) = \|A\|_* + \lambda\|E\|_1 + \\ y^T(D - A - E) + \frac{\mu}{2}\|D - A - E\|_F^2 \end{aligned} \tag{3.19}$$

A pseudo-code representation of ELAM is shown in Algorithm 3.

**Algorithm 3: Exact ALM algorithm(ELAM)**

```
Input: Observation matrix $D$, $\lambda$
while not converged do:
    $(A_{K+1}, E_{K+1}) \;=\; arg\ min_{A,E}\ L(A, E, y_k, \mu_k)$
    while not converged do:
        $U, S, V \;=\; svd(D - E^j_{k+1} + \frac{y_k}{\mu_k})$
        $A^{j+1}_{k+1} \;=\; US_{\frac{1}{\mu_k}}[S]V^T$
        $E_{k+1}^{j+1} \;=\; S_{\frac{\lambda}{\mu_k}}[D - A_{k+1}^{j+1} + \frac{y_k}{\mu_k}]$
        j++
    end while
    $y_{k+1} = y_k + \;\mu_k(D - A_{k+1} - E_{k+1})$
    k++
end while
Output: $A \;=\; A_{k+1}$ , $E \;=\; E_{k+1}$
```

**3. Inexact Augmented Lagrangian Multiplier (IALM)**: This method was described in detail in [9], the formulation of the problem is:

$$\begin{gathered} min_A \|A\|_* \\ \text{subject to: } A + E = D \\ \text{and } \pi_\Omega(E) = 0 \end{gathered} \tag{3.20}$$

where $\pi_\Omega$ is a linear operator that keeps all the entries in $\Omega$ unchanged and sets others ($\notin \Omega$) to zeros. This simply means that the unknown values of D will be set to zeros. The Lagrangian function of the problem is:

$$\begin{aligned} L(A, E, y, \mu) = \|A\|_* + y^T(D - A - E)+ \\ +\frac{\mu}{2}\|D - A - E\|_F^2 \end{aligned} \tag{3.21}$$

A pseudo-code for the ILAM method is provided in Algorithm 4.

**Algorithm 4: Inexact ALM algorithm(ILAM)**

```
Input: Observation matrix $D$, $(i,j) \in \Omega$
while not converged do:
    $(A_{K+1}, E_{K+1}) \;=\; arg\ min_{A,E}\ L(A, E, y_k, \mu_k)$
    $U, S, V \;=\; svd(D - E_k + \frac{y_k}{\mu_k})$
    $A_{k+1} \;=\; US_{\frac{1}{\mu_k}}[S]V^T$
    $E_{k+1} \;=\; \pi_{\not\Omega}(D - A_{k+1} + \frac{y_k}{\mu_k})$
    $\pi_{\Omega}(E) \;=\; 0$
    $y_{k+1} = y_k \;+\; \mu_k(D - A_{k+1} - E_{k+1})$
    k++
end while
Output: $A \;=\; A^{k+1}$ , $E \;=\; E^{k+1}$
```

## 3.7 Evaluation

In this section we evaluate the accuracy and effectiveness of the ADMM algorithm as it applies to biomedical data matrix completion. All of the following experiments were performed using Python 2.7 and Matlab(2016) on an Intel Core i7 PC running Windows 10 with 16GB main memory.

### 3.7.1 Error Calculation

In order to determine how well the matrix completion algorithm works, we are going to start with a known matrix $X$, remove a random portion of it (i.e., simulating missing entries), and then trying to reconstruct the matrix($X'$). We define the relative error (RE) as the Frobenius Norm of the difference of the original and reconstructed matrices normalized with respect to the Frobenius Norm of the original matrix:

$$RE = \frac{\|X - X'\|_F}{\|X\|_F} \tag{3.22}$$

We will also use mean squared error (MSE) to evaluate the performance of the methods; MSE can be an important metric when measuring the performance of a predictor. The MSE between $X$ and $X'$ (where $X$, $X' \in R^{n_1 \times n_2}$) is:

$$MSE = \frac{\|X - X'\|_F}{n_1 * n_2} \tag{3.23}$$

### 3.7.2 Datasets

In our experiment we will use the genomic data repositories of the **NCBI** (National Center for Biotechnology Information). The NCBI-GEO (Gene Expression Omnibus) is a public repository of genomic data. GEO profiles show expression values for individual genes. We will use the following gene expression datasets that measured the mRNA levels of different genes in different subjects (the amount of mRNA levels show the extent of gene expressions).

1. **Autism study**: ”Autism and increased paternal age related changes in global levels of gene expression regulation”. In this study, Alter et al. compared gene expression profiles from peripheral blood lymphocytes of children with autism (n=82) and controls(n=64). This data set has 54613 rows and 146 columns [11].

2. **Psoriasis study**: ”Shrinking the Psoriasis Assessment Gap: Early Gene-Expression Profiling Accurately Predicts Response to Long-Term Treatment”. In this study, Skin biopsy samples (n=170) were collected at baseline for RNA extraction and microarray analysis from 85 patients with moderate-to-severe psoriasis without receiving active psoriasis therapy. This data set has 54675 rows and 170 columns [12].

3. **Dementia study**: ”Variations in the progranulin gene affect global gene expression in frontotemporal lobar degeneration”. In this study postmortem brain samples were isolated from normal controls, FTLD-U patients with progranulin gene

mutations and FTLD-U patients without progranulin gene mutations. This data set has 22277 rows and 56 columns (i.e., 56 subjects) [13].

4. **Lung cancer study**: "Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival". In this study, Landi et al. performed gene expression analysis on fresh frozen tissue samples of adenocarcinoma and non-involved lung tissue from current, former and never smokers, with biochemically validated smoking information. This data set has 22283 rows and 107 columns [14].

### 3.7.3 Performance

We used the complete datasets as starting points for all experiments and removed a random set of values from the data matrices. The resulting incomplete matrices were then used as inputs to the algorithms to recover the missing values. We compared the original and the reconstructed matrices to measure how well the algorithms are performing. One important question is, how the algorithms perform at varying degrees of missing information. We made this a factor in our experiment. More precisely, we varied the proportion of missing data in our experiments from 10% to 90% with 10% increments (nine measurement sets each). Each data-point in our figures is an average over ten separate experiments where different entries were randomly removed from the original matrix.

Figure 3.1 shows the aggregated performance of the four approaches in Autism, Psoriasis, Dementia and Lung cancer studies. The vertical axes in plots A, C, E, and G represents the relative error (RE) as described in Equation 3.22 and the vertical axes in plots B, D, F, and H represents MSE as described in Equation 3.23. The horizontal axes in all plots (A - H) shows the percentage of the missing values (the experiment's factor). The black lines depict the performance of the SVT approach, the dotted blue lines represent the ELAM approach, the green lines show the ILAM approach, and the dotted red lines depict the performance of the featured ADMM approach. We can observe that the ADMM approach not only consistently outperforms the other three approaches in all four settings, but when it comes to the MSE, it also has a different trend. While the black, blue and green lines show progressively increasing errors, the errors presented by the red curves depict a decreasing acceleration tendency. When the proportion of missing elements is increased, the MSE of the ADMM approach grows much slower than that of the other three.

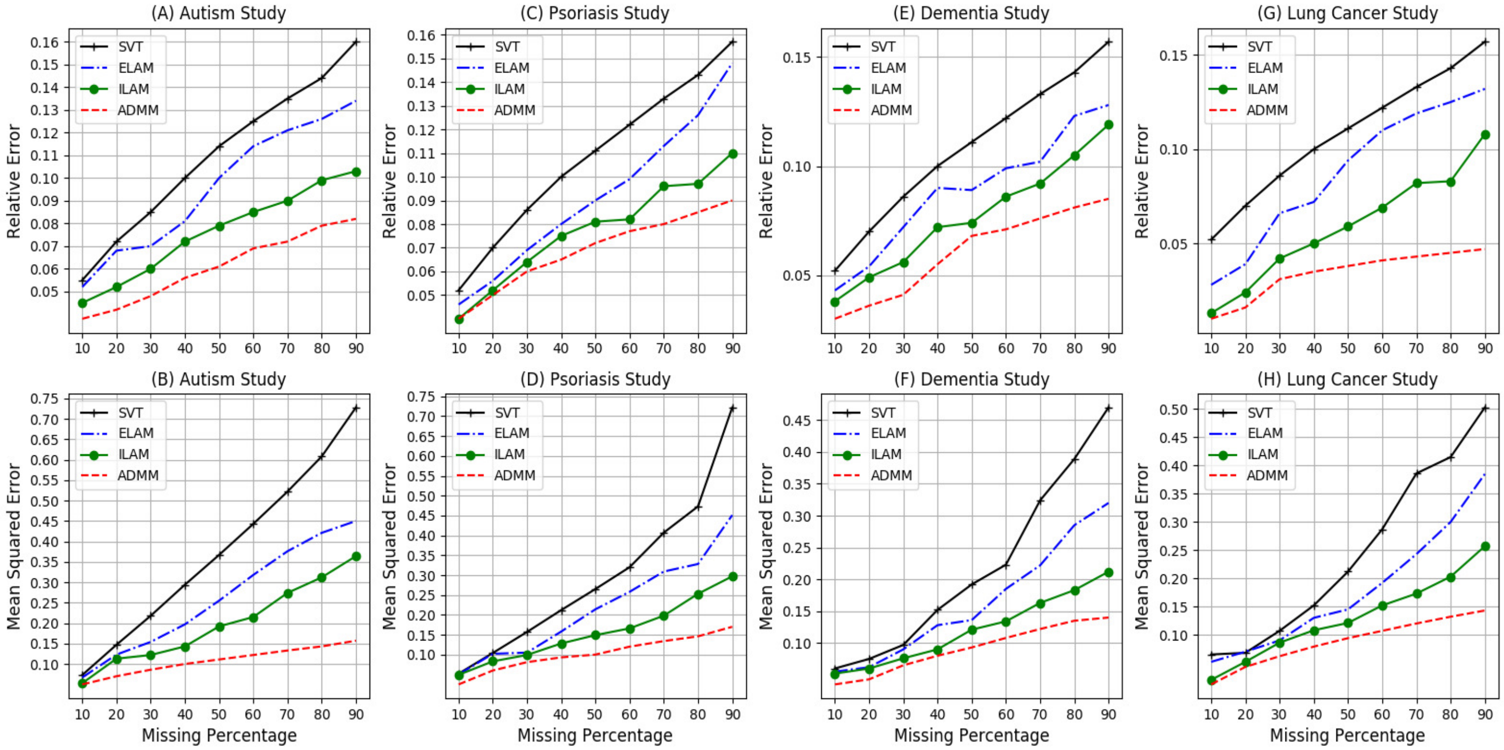

Figure 3.1. Comparison of matrix recovery methods on 4 different studies from NCBI-GEO.

Based on our results, The ADMM approach has higher accuracy especially in the cases where the matrix has more missing values. In the case of 90% missing values for relative error, in the best case, ADMM outperforms ILAM, ELAM, and SVT by a factor of 5.7, 6.5, and 7 respectively. In the worst case ADMM outperforms ILAM, ELAM, and SVT by a factor of 2, 3.5, and 4.5 respectively. When looking at MSE, ADMM outperforms the other three approaches by as much as a factor of 5.7, 6.6, and 7.9 respectively (for the same order as previously) and in the worst case we get an improvement factor of 3.4, 5.7, and 7 respectively.

## 3.8 Conclusion

In this paper we described how Robust Principal Component Analysis(RPCA) can be applied and used on NCBI-GEO biological data to find (artificially introduced) missing values and recover the datasets. After describing the RPCA approach, we presented the Alternating Direction Method of Multipliers(ADMM) algorithm. We then described three well known algorithms that can be used when recovering low rank matrices and we compared the performances of the four approaches. To do this, we removed random elements from the datasets as represented by matrices and predicted them based on the assumption that genes have similar behaviors in similar conditions. Our study provides an insight for future work especially in bio-medicine but also has implications to recommender systems. We found that ADMM approach outperforms the other three approaches, i.e., it predicted more accurate values. We hope that this study can open new opportunities to gene expression studies. As we stated earlier, gene expression experiments are very expensive and time consuming so using such computational methods can help biologists identify promising directions for studies based on partial measurements in gene expression experiments.

# REFERENCES

[1] A. Kapur ,K. Marwah and G. Alterovitz, *Gene expression prediction using low-rank matrix completion*, BMC Bioinformatics, 2016.

[2] B. Lois and N. Vaswani, *Online matrix completion and online robust PCA*, IEEE International Symposium on Information Theory (ISIT), 2015.

[3] A. Podosinnikova, S. Setzer, M. Hein, *Robust PCA: optimization of the robust reconstruction error over the stiefel manifold*, German Conference on Pattern Recognition(GCPR), 2015.

[4] J. Liu, Y. Wang, C. Zheng, W. Sha, J. Mi and Y. Xu, *Robust PCA based method for discovering differentially expressed genes*, BMC Bioinformatics, 2013.

[5] Y. Zhong, *Alternating direction method of multipliers(ADMM)*, https://piazza-resources.s3.amazonaws.com.

[6] J. Cai, E. Candes and Z. Shen, *A singular value thresholding algorithm for matrix completion*, SIAM Journal of Optimization, 20(4), 1956-1982, 2008.

[7] E. Candes and B. Recht, *Exact matrix completion via convex optimization*, Foundations of Computational Mathematics, 9(6), 717-772, 2009.

[8] I. Daubechies, M. Dferis and C. De Mol, *An iterative algorithm for linear inverse problems with a sparsity constraint*, Communications on Pure & Applied Mathematics & Programming, 55(1), 293-318, 1992.

[9] Z. Lin, M. Chen and Y. Ma, *Linearized Alternating Direction Method with Adaptive Penalty for Low Rank Representation*, Conference on Neural Information Processing Systems(Nips), 2011.

[10] Z. Lin, M. Chen and Y. Ma, *The augmented lagrange multiplier method for exact recovery of corrupted low-rank matrices*, Mathematical Programming, 2010.

[11] MD. Alter ,R. Kharkar , KE. Ramsey , DW. Craig, et al.,*Autism and increased paternal age related changes in global levels of gene expression regulation*, PLoS One, 6(2), PMID: 21379579, 2011.

[12] J. Correa da Rosa,J. Kim ,S. Tian ,LE . Tomalin, et al.,*Shrinking the psoriasis assessment gap: early gene-expression profiling accurately predicts response to long-term treatment*, Journal of Investigative Dermatology, 137(2), 305-312. PMID: 27667537, 2017.

[13] AS. Chen-Plotkin ,F. Geser ,JB. Plotkin , CM . Clark, et al., *Variations in the progranulin gene affect global gene expression in frontotemporal lobar degeneration*, Human Molecular Genetics, 17(10), 1349-62, PMID: 18223198, 2008.

[14] MT. Landi , T. Dracheva , M. Rotunno ,JD. Figueroa , et al.: *Gene expression signature of cigarette smoking and its role in lung adenocarcinoma development and survival*, PLoS One, 3(2), PMID: 18297132, 2008.

CHAPTER 4

# Graph Regularized Matrix Factorization for MiRNA-Disease Association Prediction

Negin Fraidouni, Gergely Zaruba

International Conference on Bioinformatics and Computational Biology
(BIOCOMP) Jul 30-Aug 02 2018, Las Vegas, NV USA.

## 4.1 Abstract

MicroRNAs (miRNAs) are a small, non-coding class of RNAs; they are involved in the development and progression of many human diseases. Although many miRNA-disease associations have already been discovered, there are many more which are still unknown. Unfortunately, experimental verification of miRNA-disease associations is very expensive and time consuming. So, computational methods and bioinformatics algorithms can be applied to help scientists pinpoint the most likely associations for more experimental verification, thus making such future discoveries less time and energy consuming. In this paper we investigate the Graph Regularized Matrix Factorization (GRMF) method for miRNA-disease prediction. This method combines miRNA functional similarity, disease semantic similarity, and known miRNA-disease associations to determine the likelihood of unknown miRNA-disease associations. Using 6-fold cross validation, we show that the GRMF method can reach a mean AUC (area under the curve) of 0.91, outperforming three state-of-the-art methods. To test the performance of GRMF for diseases with no known associations, we used Breast Neoplasm, removing all related miRNAs; the 50 predicted miRNAs by GRMF was verified by the databases: HMDD v.2.0, dbDEMC, and miR2Disease. For another case study, we used Lymphoma using known associations from HMDD v.2.0; 45 out of 50 (90%) of the GRMF predicted miRNAs were verified by dbDEMC, miR2Disease and PubMed literature. Therefore, we believe that GRMF could be an effective method to predict miRNA-disease associations.

## 4.2 Introduction

miRNAs are small (about 21 to 24 nucleotides), non-coding, single strand RNA molecules; they are involved in gene expression regulation. miRNAs are found in most eukaryotes including those of humans, and they tend to bind target mRNA and prevent protein production [1]. miRNA-directed gene expression regulation is a very active area of research. Hundreds of miRNAs have been discovered and the recent development of sequencing techniques and bioinformatics prediction methods significantly enhanced our information about miRNAs, including possible functions and regulatory targets [2–4]. miRNAs have been found to be responsible for different processes including cell death, cell proliferation, neural patterning, immunity, fat metabolism, and hematopoietic differentiation [5]. Computational methods for finding genes regulated by miRNAs have suggested that all these examples only represent a few samples and thus they can not describe the whole miRNA system [6].

Dysregulation of miRNAs have been shown to be the main reason of abnormal cell behavior and hence some human diseases. More and more miRNAs have been confirmed to be responsible for the development of human diseases [7, 8]. For instance studies confirmed that the miR-200 family has a strong association with breast cancer [9]; also, leukemia is one of the human cancers confirmed to be related to miR-15 and miR-16 dysregulation [10]. So recognizing miRNA-disease associations can help in diagnosing, treating, and preventing human diseases. However, it is prohibitive to find the associations one-by-one due to the significant amount of resources that have to be spent in performing such experiments. Meanwhile, known miRNA-disease associations are stored in databases like HMDD v.2.0 [11], dbDEMC [12] and miR2Disease [13] but there is a high demand for identifying new associations. Using computational methods to prioritize potential miRNAs for any specific miRNA-disease study could significantly reduce the time and financial resources needed for these experiments. Many computational methods that have been developed by scientists, are based on the assumption that similar miRNAs are likely to be related to similar diseases [14–16].

In this paper we will investigate how the Graph Regularized Matrix Factorization (GRMF) method could be used to discover miRNA-disease relationships. The rest of the paper is organized as follows: Section 2 describes the GRMF method in detail. In Section 3 we describe our GRMF experiments and compare our result with three state-of-the-art methods for predicting miRNA-disease associations. Finally, Section 4 concludes the paper.

## 4.3 Methods

In this section, we first describe the datasets used, then we provide a description of Graph Regularized Matrix Factorization as it applies to miRNA-disease associations.

### 4.3.1 Human miRNA-Disease Association

The database we have used for our study contains data of the associations between human miRNA and disease from the Human microRNA Disease Database (HMDD [11]). The database includes about 579 miRNAs, 384 diseases and 10,381 experimentally confirmed associations between miRNAs and diseases. Using this data, we can construct matrix $Y$ to capture the associations between miRNAs and diseases. Each row of matrix $Y$ represents a different miRNA and each column represents a different disease. Based on the datasets, the elements of matrix $Y$ can only be either 0 or 1. If miRNA $m_i$ is associated with disease $d_j$, then $Y_{ij}$ is 1. $Y_{ij}$ is 0 in the cases where there is no known association between $m_i$ and $d_j$ (this does not mean that there is no relationship but merely that the relationship is unknown).

### 4.3.2 miRNA Functional Similarity

miRNA functional similarity scores are were calculated under the assumption that if two miRNAs are functionally similar, they are more likely to be related to phenotypically similar diseases. Wang et al. developed a method called MISIM [17] for measuring the similarity between two different miRNAs. MISIM has 4 main steps: first, diseases associated with two miRNAs are recognized, denoted as $d_1$ and $d_2$. In the next step the semantic values of diseases were calculated. In the third step the semantic similarity were calculated between every pair of the diseases between $d_1$ and $d_2$. And finally the functional similarity of two miRNA were calculated based on the semantic similarity of $d1$ and $d2$. We downloaded the scores from: http://www.cuilab.cn/files/images/cuilab/misim.zip

### 4.3.3 Disease Semantic Similarity

For the purpose of calculating disease semantic similarities, diseases can be described as a Directed Acyclic Graph (DAG). Each disease represents a node in the graph while edges represent relationship between diseases. Disease $d_1$ can be described as $DAG(d_1) = (d_1, S_{d_1}, E_{d_1})$ where $S_{d_1}$ is the set of all nodes including all ancestors of node $d_1$ including $d_1$ itself and $E_{d_1}$ is the set of all corresponding links, this includes all direct edges from parents to child nodes. So the contribution of disease $d$ in disease $d_1$ can be calculated as:

$$S_{d_1}(d) = \begin{cases} 1 & d_1 = d \\ max\{\Delta * S_{d_1}(d') | d' \in \text{children of } d\} & d_1 \neq d \end{cases} \tag{4.1}$$

Here $\Delta$ is the contribution factor for all connection links from disease $d$ to disease $d'$. The contribution of disease $d_1$ to its own semantic value is 1 so the farther the nodes from disease $d_1$, the less effect they have on the $d_1$'s semantic value. Thus, the value of $\Delta$ should be between 0 and 1. Wang et al. suggested that when the value of $\Delta = 0.5$ then the results show better correlation with the expression similarity [17]. The semantic value(SV) of disease $d_1$ can be described as:

$$SV(d_1) = \sum_{d \in D_{d_1}} S_{d_1}(d) \tag{4.2}$$

If two diseases have much in common in the DAG then their similarity value would become larger. The semantic similarity between diseases $d_1$ and $d_2$ can be calculated as:

$$SV(d_1, d_2) = \frac{\sum_{d \in (S_{d_1} \cap S_{d_2})} (S_{d1}(d) + S_{d2}(d))}{SV(d_1) + SV(d_2)} \tag{4.3}$$

where $S_{d_1}(d)$ is the semantic value of disease $d$ related to disease $d_1$ and $S_{d_2}(d)$ is the semantic value of disease $d$ related to disease $d_2$. What equation 5.3 calculates, is the semantic similarity between two different diseases based on their location in DAG and the common links in their ancestors.

### 4.3.4 Weighted K-nearest Known Neighbors

Our miRNA-disease association matrix $Y \in \mathbb{R}^{n*m}$ has $n$ rows representing miRNAs and $m$ columns representing diseases. Matrix $Y$ is a sparse matrix and most of it's values are zero although many of these zeros are unknown interactions that could potentially be true. Our aim is to replace zeros with a continuous value between 0 and 1; in the preprocessing step, we use the weighted k-nearest known neighbor algorithm to estimate an association likelihood based on the known associations. Algorithm 1 describes the process in detail. In a nutshell, first we calculate the weighted average of the k nearest neighbors to miRNA $m_i$, then we calculate the weighted average of the k nearest neighbors to disease $d_j$; in the final step we replace the entries in $Y$ that are 0 by the average likelihood of $m_i$ and $d_j$.

**Algorithm 1: Weighted K-nearest Known Neighbors**

```
Input: $Y$, $S^m$, $S^d$, $k$, $\eta$
$Y_m = Y_d = 0$
for $q = 1$ to $n$:
    $knn$ = k-nearest neighbors of row $q$ from $S^m$
    for $i > k$:
        $w_i = \eta^{i-1} * S^m(q, knn_i)$
    end for
    $P_q = \sum_{i=1}^{k} S^m(q, knn_i)$
    $Y_m(q) = \frac{1}{P_q} \sum_{i=1}^{k} w_i Y (knn_i)$
end for

for $r = 1$ to $m$:
    $dnn$ = k-nearest neighbors of row $r$ from $S^d$
    for $j > k$:
        $w_j = \eta^{j-1} * S^d(r, dnn_j)$
    end for
    $P_r = \sum_{j=1}^{k} S^d(r, dnn_j)$
    $Y_d(r) = \frac{1}{P_r} \sum_{j=1}^{k} w_i Y (dnn_j)$
end for

$Y = max(Y, \frac{Y_m+Y_d}{2})$
Output: $Y$
```

### 4.3.5 Graph Regularized Matrix Factorization

A linear approximation of our miRNA-disease association matrix $Y \in \mathbb{R}^{n*m}$ can be shown by $Y \approx W.H^T$ where $W \in \mathbb{R}^{n*f}$, $H \in \mathbb{R}^{m*f}$ and $f$ is the number of latent features in $W$ (in miRNAs) and $H$ (in diseases).

Given a data matrix $Y$, the choice of $W$ and $H$ have to be such to minimize the reconstruction error between $Y$ and $WH^T$. Among the various error functions that have been proposed [18], the most widely used is the squared error or euclidean distance with respect to the Frobenius norm. So the problem can be written as:

$$min_{W,H} \|Y - WH^T\|_F^2 \tag{4.4}$$

The objective function in Eq. 4.4 is convex in $W$ only or $H$ only, but it is not convex in both of them together. For the aim of preventing overfitting, we can add linear and graph regularization terms. Linear regularization term minimizes norms of both $W$ and $H$ while graph regularization terms minimize the distance between latent feature vectors of two neighbor miRNAs and diseases. So our objective function becomes:

$$\begin{aligned} min_{_{W,H}} &\|Y - WH^T\|_F^2 \\ &+\lambda_a(\|W\|_F^2 + \|H\|_F^2) \\ +\lambda_b \sum_{i=1}^{n} \sum_{q=1}^{n} &(S_{i,q}^m)\|w_i - w_q\|^2 \\ +\lambda_c \sum_{j=1}^{m} \sum_{r=1}^{m} &(S_{j,r}^d)\|h_j - h_r\|^2 \end{aligned} \tag{4.5}$$

where $\lambda_a$, where $\lambda_b$ and where $\lambda_c$ are all positive parameters, $w_i$ and $w_q$ are $i^{th}$ and $q^{th}$ row of $W$, $h_j$ and $h_r$ are $j^{th}$ and $r^{th}$ row of $H$. We can rewrite Eq. 4.5 as:

$$\begin{aligned} min_{W,H} &\|Y - WH^T\|_F^2 \\ +\lambda_a&(\|W\|_F^2 + \|H\|_F^2) \\ +\lambda_b&\mathrm{Tr}(W^T \mathcal{L}_b W) \\ +\lambda_c&\mathrm{Tr}(H^T \mathcal{L}_c H) \end{aligned} \tag{4.6}$$

where Tr is the trace of a matrix, $\mathcal{L}_b = D^b - S^m$ and $\mathcal{L}_c = D^c - S^d$ are the graph Laplacians for $S^m$ and $S^d$ respectively and $D^b_{ii} = \sum_q S^m_{iq}$ and $D^c_{jj} = \sum_r S^d_{jr}$ are diagonal matrices. (We refer the reader to [19] for more details on obtaining Eq. 4.6 from Eq. 4.5.)

We provide a pseudocode for the Graph Regularized Matrix Factorization (GRMF) in Algorithm 2. We use singular value decomposition (SVD) to obtain $U \in \mathbb{R}^{n*f}$, $\Sigma \in \mathbb{R}^{f*f}$ and $V \in \mathbb{R}^{m*f}$ from $Y$. Then we initialize $W$ and $H$ as $W = U\sqrt{\Sigma}$ and $H = V\sqrt{\Sigma}$. We used alternating least squares to update $W$ and $H$ in each iteration. If we denote the objective function of Eq. 4.6 as J, we set $\frac{\partial J}{\partial W} = 0$ and $\frac{\partial J}{\partial H} = 0$ so we can update $W$ and $H$ through:

$$W = (YH - \lambda_b \mathcal{L}_b W)(H^T H + \lambda_a I_k)^{-1} \tag{4.7}$$

$$H = (Y^T W - \lambda_c \mathcal{L}_c H)(W^T W + \lambda_a I_k)^{-1} \tag{4.8}$$

**Algorithm 2: Graph Regularized Matrix Factorization (GRMF)**

```
Input: $Y$, $S^m$, $S^d$, $f$, $\lambda_a$, $\lambda_b$, $\lambda_c$
```

```
$U \ \Sigma \ V^T$ = $SVD(Y, f)$
$W \ = \ U\sqrt{\Sigma}$
$H \ = \ V\sqrt{\Sigma}$
$\mathcal{L}_b \ = \ D^b \ - \ S^m$
$\mathcal{L}_c \ = \ D^c \ - \ S^d$
while not converged:
    $W \ = \ (YH - \lambda_b \mathcal{L}_b W)(H^T H + \lambda_a I_k)^{-1}$
    $H \ = \ (Y^T W - \lambda_c \mathcal{L}_c H)(W^T W + \lambda_a I_k)^{-1}$
end while
$YY \ = \ WH$
Output: $YY$
```

## 4.4 Evaluation

To evaluate the proposed GRMF-based method we compared its performance to three state-of-the-art miRNA-disease prediction methods.

### 4.4.1 Competitive Methods

#### 4.4.1.1 RLSMDA

Chen et al. [14] developed the method of Regularized Least Squares for MiRNA-Disease Association (RLSMDA) to find miRNAs associated with different diseases using a semi-supervised learning method. RLSMDA is designed using a continuous classification function to reflect the probability with which each miRNA is associated with a specific disease. RLSMDA can predict miRNAs related to diseases that have no known associated miRNA and it does not need negative miRNA-disease associations. However the ways of combining classifiers in different spaces and also the choice of parameters can affect the prediction performance of this method.

#### 4.4.1.2 NetCBI

In this study Chen et al. [15] constructed the miRNA-disease association network (NetCBI) using a representation of for bipartite graphs, where the nodes correspond to either diseases or miRNAs, and edges correspond to the associations between them. The main idea behind NetCBI is that if a given miRNA is related to a disease, other miRNAs that are similar to it, will be chosen and recommended to be related to that disease as well. Also if a miRNA is related to a disease, that miRNA will also be likely classified to be related to similar diseases.

#### 4.4.1.3 NBI

Li et al. [16] developed a computational method (NBI) to predict new miRNA-disease associations by integrating environmental factor (EF) similarity and disease phenotypic similarity. More precisely, in NBI, three comprehensive bipartite networks are constructed, i.e., the EF-disease, the EF-miRNA, and the miRNA-disease associations. This method uses known associations to obtain predicted candidates. The miRNAs that are related to EFs, average their resources to all of their neighbors and thus they distribute the associations to every miRNA neighbor.

### 4.4.2 Performance

We plotted Receiver Operating Characteristics curve (ROC) and used Area Under the ROC curve (AUC) as the main metric for evaluating their performance. The area under ROC curve is calculated as an index of the prediction power of the GRMF method. The value of AUC is between 0 and 1 and higher amounts shows more prediction power. If the value is equal to 0.5, it means the performance is equal to a random prediction.

For an specific disease $d_1$, all the known $d_1$ related miRNAs are defined as labeled nodes and the remaining miRNAs (on which there is no relevance information) are defined as unlabeled nodes. Given a threshold $\delta$, if the result prediction of a labeled node is greater than $\delta$, then the node is identified as positive sample. If the result prediction of a an unlabeled node is less than $\delta$, then the node is a considered as a negative sample. To plot a ROC curve we calculated the true positive rates (TPR or sensitivity) and false positive rates (FPR, 1-specificity) through:

$$\begin{aligned} TPR &= \frac{TP}{TP + FN} \\ FPR &= \frac{FP}{TN + FP} \end{aligned} \tag{4.9}$$

where:

- TP : Number of correctly identified positive samples.
- TN : Number of correctly identified negative samples.
- FP : Number of misidentified positive samples.
- FN : Number of misidentified negative samples.

Sensitivity means the percentage of the positive samples that are correctly identified among all the positives and specificity means the percentage of the negative samples correctly identified among all negatives.

We conducted five repetitions of a 6-fold cross validation for each of the methods. The 6-fold cross validation is implemented using the known miRNA-disease association in the HMDD V2.0 database. So in each repetition, we divided our association matrix $Y$ in to six parts and each parts, one-by-one , was left out as the test set while we used the remaining five parts as the training set. All of unknown miRNA-disease association pairs can be seen as candidate samples. After applying GMRF, scores of the test samples were compared with the all scores of the candidates samples. In order to make the validation more accurate, we repeated this process 5 times. Figure 4.1 depicts the ROC curve and the calculated AUC of each fold in 6-fold cross validation for GRMF method.

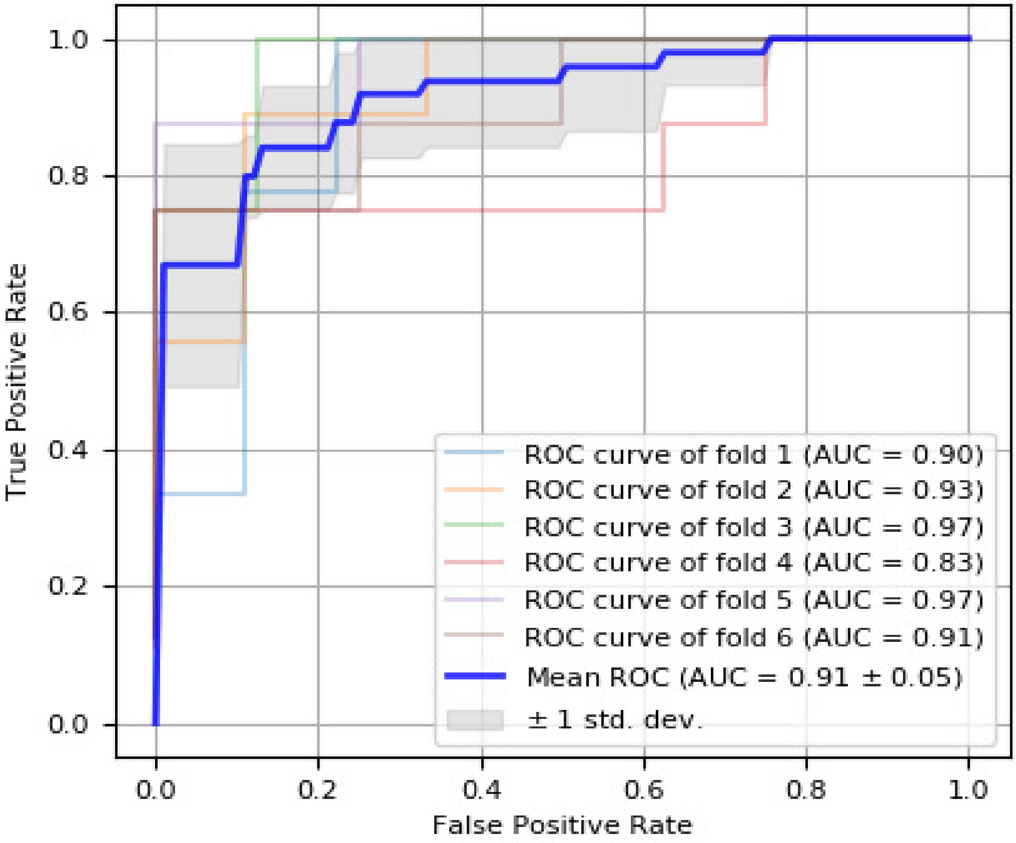

Figure 4.1. Performance of each fold in 6-fold cross validation for GRMF method.

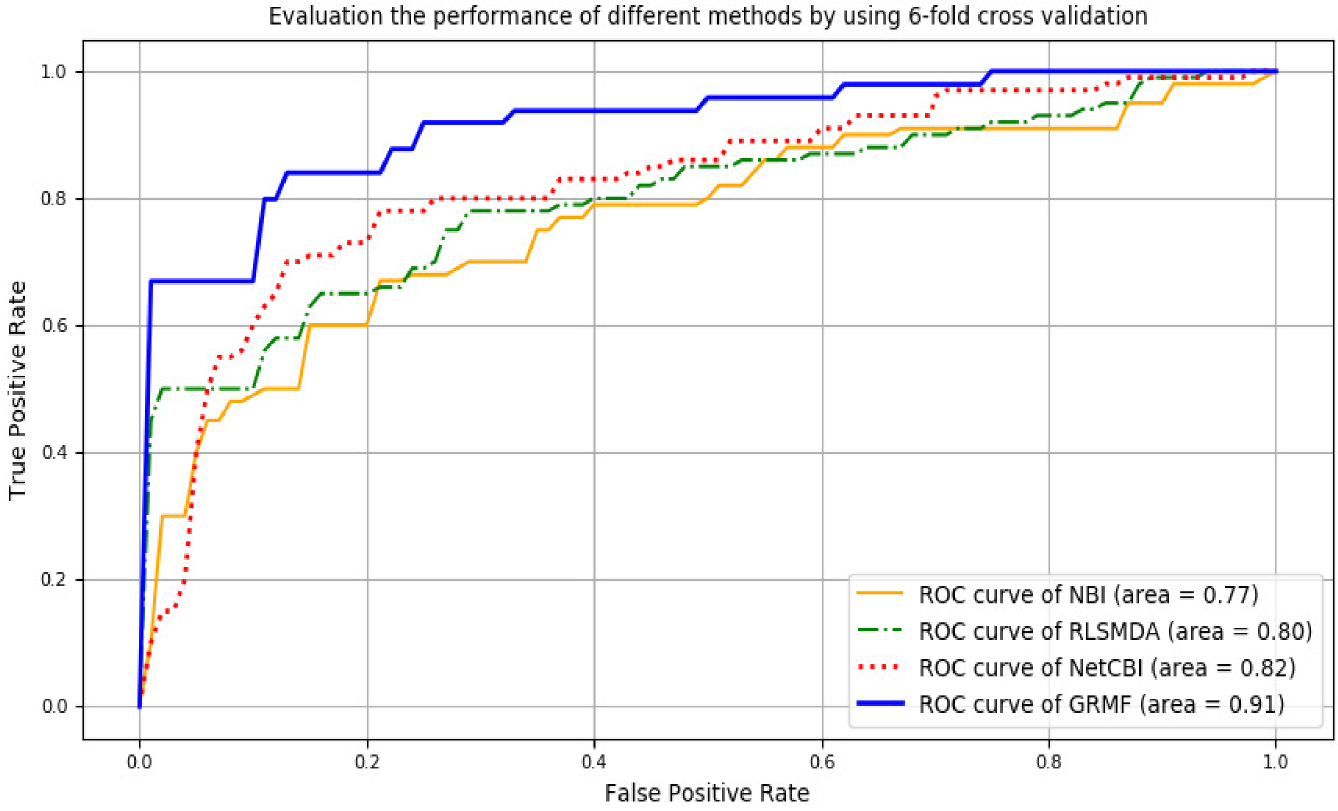

Figure 4.2. Comparison of 4 different studies on miRNA-disease association prediction.

We compare the performance of GRMF approach with three state-of-the-art methods for miRNA-disease association prediction. Figure 4.2 shows the prediction performance of GRMF, NBI [16], RLSMDA [14] and NetCBI [15]. The GRMF achieves the AUC value of 0.91 compared with other methods: NBI: 0.77, RLSMDA: 0.80 and NetCBI: 0.82 and outperforms the other three.

### 4.4.3 Case studies

In order to demonstrate the performance of GMRF, we evaluated the prediction ability of GMRF for miRNAs related to Breast Neoplasm and Lymphoma. Two miRNA-disease datasets (dbDEMC [12] and miR2Disease [13]) and previous PubMed studies were used to confirm the correctness of the prediction.

#### 4.4.3.1 Breast Neoplasm

Breast Neoplasm (BN or breast cancer) is the second most common cancer in American women; dysregulation of miRNAs play an important role in this disease [22]. We used BN in order to show the performance of GRMF for diseases which have no related miRNAs. The total number of miRNAs related to BN was 202 so we removed all 202 related miRNAs in our dataset to ensure that only the information from other diseases would be used to predict the related miRNAs to BN. We then ranked the predicted scores for all candidate miRNAs so the top 50 miRNAs was selected and they are shown in Table 4.1 Based on our results, we could confirm all 50 miRNAs by miR2Disease, dbDEMC and HDMM.

#### 4.4.3.2 Lymphoma

Lymphoma is recognized as the fifth most common cancer type and is cancer of lymphatic system (Blood B and T cells). It includes Hodgkin Lymphoma (HL) and Non-Hodgkin Lymphoma (NHL) [20]. B-cells Lymphoma is the most common type of NHL in the United states and worldwide. Because Lymphoma can be derived form B-cells at different stages of cell cycle, miRNAs can be both target genes and specific markers [21]. For the second case study, we chose Lymphoma and the results are summarized in Table 4.2 We could confirm 45 miRNAs out of 50 as shown by by dbDEMC, miR2Disease and experimental literature in PubMed.

Table 4.1. Prediction of the top 50 predicted miRNAs associated with Breast Neoplasm based on the known associations in HDMM V.2.0 database.

| **miRNA** | **Evidence** | **miRNA** | **Evidence** |
|---|---|---|---|
| hsa-mir-10b | dbDEMC; miR2Disease; HMDD | hsa-mir-210 | dbDEMC; miR2Disease; HMDD |
| hsa-mir-372 | dbDEMC | hsa-mir-516a | HMDD |
| hsa-mir-134 | dbDEMC | hsa-mir-146b | dbDEMC; miR2Disease; HMDD |
| hsa-mir-143 | dbDEMC; miR2Disease; HMDD | hsa-mir-192 | dbDEMC |
| hsa-let-7e | dbDEMC; HMDD | hsa-mir-183 | dbDEMC; HMDD |
| hsa-mir-1 | dbDEMC; HMDD | hsa-mir-320a | HMDD |
| hsa-mir-499a | HMDD | hsa-mir-499a | HMDD |
| hsa-mir-150 | dbDEMC | hsa-mir-182 | dbDEMC; miR2Disease; HMDD |
| hsa-let-7i | dbDEMC; miR2Disease; HMDD | hsa-mir-152 | dbDEMC; miR2Disease; HMDD |
| hsa-mir-137 | dbDEMC; HMDD | hsa-mir-221 | dbDEMC; miR2Disease; HMDD |
| hsa-mir-19b | dbDEMC; HMDD | hsa-mir-184 | dbDEMC |
| hsa-mir-125a | dbDEMC; miR2Disease; HMDD | hsa-mir-32 | dbDEMC |
| hsa-mir-214 | dbDEMC; HMDD | hsa-mir-325 | dbDEMC |
| hsa-mir-302b | dbDEMC; HMDD | hsa-mir-30b | dbDEMC; HMDD |
| hsa-mir-219 | dbDEMC; HMDD | hsa-mir-106a | dbDEMC |
| hsa-mir-204 | dbDEMC; miR2Disease; HMDD | hsa-mir-205 | dbDEMC; miR2Disease; HMDD |
| hsa-mir-20b | HMDD | hsa-mir-181c | dbDEMC |
| hsa-mir-101 | dbDEMC; miR2Disease; HMDD | hsa-let-7b | dbDEMC; HMDD |
| hsa-mir-302c | dbDEMC; HMDD | hsa-mir-212 | dbDEMC |
| hsa-mir-20a | miR2Disease; HMDD | hsa-mir-506 | HMDD |
| hsa-let-7d | dbDEMC; miR2Disease; HMDD | hsa-mir-140 | dbDEMC; HMDD |
| hsa-mir-25 | dbDEMC; HMDD | hsa-mir-708 | HMDD |
| hsa-mir-195 | dbDEMC; miR2Disease; HMDD | hsa-mir-433 | dbDEMC |
| hsa-mir-107 | dbDEMC; HMDD | hsa-mir-153 | dbDEMC; HMDD |
| hsa-mir-187 | dbDEMC; HMDD | hsa-mir-141 | dbDEMC; miR2Disease; HMDD |

## 4.5 Conclusion

Finding the molecular mechanism of diseases can help exploring disease pathogenesis and finding effective treatments. miRNAs as a class of non-coding RNAs are responsible for regulating gene expression so they can cause various diseases [23, 24]. Some computational approaches have been proposed to capture miRNA-disease association [14–16]. However, these methods have limitations.

In this paper we presented a Graph Regularized Matrix Factorization method (GRMF) to predict miRNA-disease associations based on the assumption that similar miRNAs (functionally) tend to be related to similar diseases (phenotypically). We used miRNA functional similarity, disease semantic similarity, and known miRNA-disease associations form the HDMM v.2.0 database. To verify the accuracy of the GRMF method, we used five repetitions of 6-fold cross validation. We compared the result of the GRMF method with three state-of-the-art methods and concluded that GRMF outperforms the other three in terms of AUC.

We selected Breast Neoplasm as a case study in order to show the performance of GRMF for diseases which have no related miRNAs and based on the results, we could confirm all 50 miRNAs as identified by miR2Disease, dbDEMC and HDMM. As the second case study we chose Lymphoma to demonstrate the performance of GRMF and based on the results, we could confirm 45 miRNAs out of 50 as identified by dbDEMC, miR2Disease and experimental literature in PubMed. The GRMF method could provide an effective approach to study miRNA-disease associations. We also recognize that GRMF has some limitations which can be improved in future research. For example, the sequence information of miRNAs is used to measure miRNA similarity but some studies show that the structural information can be more effective [25, 26]. Furthermore, expression information of miRNAs could also be used to measure this similarity.

Table 4.2. Prediction of the top 50 predicted miRNAs associated with Lymphoma.

| **miRNA** | **Evidence** | **miRNA** | **Evidence** |
|---|---|---|---|
| hsa-mir-141 | dbDEMC | hsa-mir-182 | dbDEMC |
| hsa-mir-10b | dbDEMC | hsa-let-7e | dbDEMC; miR2Disease |
| hsa-mir-30a | dbDEMC | hsa-mir-335 | dbDEMC |
| hsa-mir-193b | PMID:22235305 | hsa-mir-183 | dbDEMC |
| hsa-mir-151a | Unconfirmed | hsa-mir-148a | dbDEMC |
| hsa-mir-106a | dbDEMC; miR2Disease | hsa-mir-34a | dbDEMC |
| hsa-mir-221 | dbDEMC | hsa-mir-9 | dbDEMC |
| hsa-mir-7 | dbDEMC | hsa-mir-125b | PMID:23527180 |
| hsa-mir-195 | dbDEMC | hsa-mir-429 | Unconfirmed |
| hsa-mir-214 | dbDEMC | hsa-mir-100 | dbDEMC |
| hsa-mir-29b | dbDEMC | hsa-mir-132 | dbDEMC |
| hsa-mir-219b | dbDEMC | hsa-let-7i | dbDEMC |
| hsa-mir-196a | dbDEMC | hsa-mir-205 | dbDEMC |
| hsa-mir-378a | Unconfirmed | hsa-mir-192 | dbDEMC |
| hsa-mir-191 | dbDEMC | hsa-mir-223 | dbDEMC |
| hsa-mir-133a | dbDEMC | hsa-mir-30e | dbDEMC |
| hsa-mir-103a | Unconfirmed | hsa-mir-145 | dbDEMC; miR2Disease |
| hsa-mir-146b | PMID:24931464 | hsa-mir-194 | dbDEMC |
| hsa-mir-30C | dbDEMC | hsa-mir-15b | dbDEMC |
| hsa-mir-34b | dbDEMC | hsa-mir-142 | Unconfirmed |
| hsa-mir-152 | dbDEMC | hsa-mir-30d | dbDEMC |
| hsa-mir-26b | dbDEMC | hsa-mir-143 | dbDEMC; miR2Disease |
| hsa-mir-338 | dbDEMC | hsa-mir-22 | dbDEMC |
| hsa-mir-29a | dbDEMC | hsa-mir-96 | dbDEMC |
| hsa-mir-27b | dbDEMC | hsa-mir-222 | dbDEMC |

# REFERENCES

[1] L. MacFarlane and P. R. Murphy, *MicroRNA*: Biogenesis, Function and Role in Cancer, Current Genomics, vol.11, no.7, 2010.

[2] E. C. Lai, P. Tomancak, R. W. Williams, G. M. Rubin, *Computational identification of Drosophila MicroRNA genes*, Genome Biology, 4, 2003.

[3] J. W. Nam, K. R. Shin, J. Han, Y. Lee, V. N. Kim, B. T. Zhang, *Human microRNA prediction through a probabilistic co-learning model of sequence and structure*, Nucleic Acids Research., 33, 3570-3581, 2005.

[4] S. C. Li, C. Y. Pan, W. C. Lin, *Bioinformatics discovery of microRNA precursor from human ESTs and introns*,BMC Genomics, 7, 2006.

[5] L. He, G. J. Hannon, *MicroRNAs: small RNAs with a big role in gene regulation*, Nature Reviews Genetics, 5, 522-531, 2004.

[6] F. Wahid, A. Shehzad, T. Khan and Y. YoungKim, *MicroRNAs: Synthesis, mechanism, function, and recent clinical trials*, Biochimica et Biophysica Acta Vol.1803, no.11, 1231-1243, 2010.

[7] N. Lynam-Lennon, S. G. Maher and J. V. Reynolds, *The roles of microRNA in cancer and apoptosis.*, Biological Reviews of the Cambridge Philosophical Society vol.84, 55-71, 2009.

[8] N. Meola, V. A. Gennarino and S. Banafi, *MicroRNAs and genetic diseases.*, Pathogenetics vol.2, no.7, 2009.

[9] Y. Y. Lim, J. A. Wright, J. L. Attema, P. A. Gregory, A. G. Bert, E. Smith, et al., *Epigenetic modulation of the miR-200 family is associated with transition to a breast cancer stem-cell-like state.*,Journal of Cell Science vol.126, 2256-2266, 2013.

[10] G. A. Callin, et al., *Frequent deletion and down-regulation of micro-RNA genes miR15 and miR16 at 13q14 in chronic lymphocytic leukemia.*,Proceedings of the National Academy of Sciences of the United States of America vol.199, 15524-15529, 2002. Res 37:D98-104, 2009.

[11] Y. Li, et al. *HMDD v.2.0: a database for experimentally supported human microRNA and disease associations.* Nucleic Acids Research. 42(D1):D1070-D1074, 2013.

[12] Z. Yang, L. Wu, A. Wang, et al. *dbDEMC 2.0: updated database of differentially expressed miRNAs in human cancers.* Nucleic Acids Research. 45(D1):D812-D818, 2017.

[13] Q. Jiang, Y. Wang, Y. Hao, L. Juan, M. Teng, X. Zhang, M. Li, G. Wang and Y. Liu *Jiang Q., Wang Y., Hao Y., Juan L., Teng M., Zhang X., Li M., Wang G., Liu Y., miR2Disease: a manually curated database for microRNA deregulation in human disease.* Nucleic Acids, 2009.

[14] X. Chen and G. Y. Yan*Semi-supervised learning for potential human microRNA-disease associations inference*, Scientific Reports, vol.40, no.1, 2014.

[15] H. Chen, Z. Zhang, *Similarity-based methods for potential human microRNA-disease association prediction.*, BMC Med Genomics vol.6 no.12, 2013.

[16] J. Li, Z. Wu, F. Cheng, W. Li, G. Liu and Y. Tang *Computational prediction of microRNA networks incorporating environmental toxicity and disease etiology*, Scientific Reports, vol.4, 2014.

[17] D. Wang, J. Wang, M. Lu, F. Song and Q. Cui, *Inferring the human microRNA functional similarity and functional network based on microRNA-associated disease*, Bioinformatics, vol.26, no.13, pp 1644-1650, 2010.

[18] P. Hoyer, *Non-negative Matrix Factorization with sparseness constraints*, Journal of machine learning research, vol.5, pp 1457-1469, 2004.

[19] Q. Gu, J. Zhou and C. Ding *Collaborative filtering: Weighted Nonnegative Matrix Factorization incorporating user and item graphs*, SIAM International Conference of Data Mining, pp 199-210, 2010.

[20] M. Fernandez-Mercado, L. Manterola and C. Lawrie. *MicroRNAs in Lymphoma: Regulatory Role and Biomarker Potential.* Current Genomics 16(5):349-358, 2015.

[21] F. Jardin and M. Figeac. *MicroRNAs in lymphoma, from diagnosis to target therapy.* Current Opinion in Oncology. 25(5):480–486, 2013.

[22] W. Wang and Y. Luo *MicroRNAs in breast cancer: oncogene and tumor suppressors with clinical potential.* Journal of Zhejiang University Science B. 16(1), 2015.

[23] T. Ideker and R. Sharan, *Protein networks in disease.* Genome research vol.18, no.4, 644-652, 2008.

[24] L. Hood, J. R. Heath, M. E. Phelps and B. Lin *Systems biology and new technologies enable predictive and preventative medicine.* Science vol.306, no.5956, 640-643, 2004.

[25] W. Lan, Q. F. Chen, T. S. Li, C. G. Yuan, S. Mann and B. S. Chen , *Identification of important positions within miRNAs by integrating sequential and structural features.* Current Protein & Peptide Science vol.15, no.6, 591-597, 2014.

[26] X. Zeng, X. Zhang and Q. Zou *Integrative approaches for predicting microRNA function and prioritizing disease-related microRNA using biological interaction networks.* Brief Bioinform, 2016.

CHAPTER 5

# A Matrix Completion Approach for Predicting lncRNA-disease association

Negin Fraidouni, Gergely Zaruba

Manuscript submitted to:

International Conference on Bioinformatics and Computational Biology (BIOCOMP) Jul 29-Aug 01, 2019, Las Vegas, NV USA.

## 5.1 Abstract

The most part of the human genome is known as Long non-coding RNAs (lncRNAs) which have been thought to be responsible for many developmental processes and diseases. Finding the potential functions of lncRNAs is very important for further study of human complex diseases. Computational methods can be effective in order to make predictions based on the known information [1]. In this study we describe the LncRNA-disease association prediction method (LDAPM) and show the its performance compared to three state-of-the-art methods. To do this, we employ the ILNCSIM method [16] to compute functional similarities of lncRNAs. In next step we measure the semantic similarities of diseases. Then we extract feature vectors for lncRNAs and diseases and finally we recover the lncRNA-disease association matrix and find new potential associations.

We used three lncRNA-disease association datasets from LncRNADisease database. For dataset 1, the LDAPM obtained the AUC of 0.88 which is significantly higher than the AUC of other methods which are 0.60, 0.68 and 0.75 for RWRLncD, RWRH and LRLSLDA respectively. For dataset 2, the LDAPM obtained the AUC of 0.81 which is significantly higher than the AUC of other methods which are 0.65, 0.60 and 0.62 for RWRLncD, RWRH and LRLSLDA respectively. Likewise for dataset 3, the LDAPM obtained the AUC of 0.83 which is significantly higher than the AUC of other methods which are 0.69, 0.64 and 0.63 for RWRLncD, RWRH and LRLSLDA respectively.

## 5.2 Introduction

Recent studies of transcriptomes have shown that a much greater part of the genome is transcribed than we knew and expected. The results of the transcription process are mostly non-protein coding RNAs [2] including Long non-coding RNAs(lncRNAs). LncRNAs are usually larger than 200 nucleotides and they are less expressed and more tissue-specific compared to protein-coding RNAs [3]. LncRNAs are thought to have almost 30,000 different types in humans and they consist the majority of the non-coding transcriptome. Although our knowledge about lncRNAs are very limited, they could have a very significant effect on regulation of transcription process [4, 5]. LncRNAs also can be responsible for cell differentiation, apoptosis and cell differentiation. Mutations and dysfunctions of lncRNAs are thought to be the reason of some human complex diseases like diabetes [6], neurodegeneration disease [7], cardiovascular diseases [8], colon cancer [9], prostate cancer [10], kidney cancer [11] and AIDS [12]. Computational methods can provide more efficient directions for these studies; so there is a high demand for novel, efficient techniques to predict lncRNA-disease association.

Computational methods that have been proposed before belong to three different groups. First group consists of methods that use machine learning models to find lncRNA-disease association. An example of this approach is a study by Lan et al. which they used multiple data sources and employed a SVM classifier to find new lncRNA-disease association [13]. Second groups take advantage of this assumption that functionally similar lncRNAs can be related to phenotypically similar diseases and vice versa. An example of this approach is a study by Zhang et al. which they applied a propagation algorithm on a constructed network and combined all information from proteins, lncRNAs and diseases [14]. Methods in third group utilize biological information on lncRNAs in order to find lncRNA-disease association. The examples of biological information are tissue specificity, genome location and expression profile. An example of this approach is a study by Chen et al. which they proposed LRLSLDA, a semi supervised learning method to find associations lncRNAs and diseases by using Laplacian regularized least squares [15]. Despite previous studies on lncRNAs, there is still a high demand for new methods to predict lncRNA-disease associations more accurately.

This paper discusses a matrix completion algorithm in order to predict unknown lncRNA-disease association. We describe how the this algorithm can be applied to lncRNA-disease data so missing values can be predicted computationally. We compare the result of this method to some state-of-the-arts approaches and show their performance advantages.

The rest of the paper is organized as follows. Section 2 describes the LDAPM approach in details. Section 3 describes the data set and the competitive approaches. In Section 4, we present our computational experiment comparing the three approaches. Finally, section 5 concludes the paper.

## 5.3 Methods

LncRNA-disease association prediction method (LDAPM) has four steps. In step 1, we employ the ILNCSIM method [16] to compute functional similarities of lncRNAs. In step 2, we measure the semantic similarities of diseases. In step 3, we extract feature vectors for lncRNAs and diseases and in the last step we recover the lncRNA-disease association matrix and find new potential associations. Here we first introduce the lncRNA-disease association data and then we show how LDAPM can be apply to lncRNA-disease association data in order to predict new associations.

### 5.3.1 LncRNA-Disease Association Data

In order to create a model, known lncRNA-disease association data can be stored in a matrix $A \in R^{m*n}$, where each row corresponds to a different lncRNA and each column corresponds to a different disease and $m$ and $n$ are the numbers of lncRNAs and diseases respectively. The entries of the matrix $A$ can be either 0 or 1. If the entry $A_{i,j}$ is 1 it means there is an association between lncRNA $i$ and disease $j$ and if the entry is 0, it means there is no known relationship between those. The main goal here is to *complete* or *recover* the matrix (i.e., finding the best prediction for the unknown relations). The most promising way to do this is to rely on the assumption that the resulting matrix has to be of low rank. By predicting missing values we could indeed predict gene expression patterns of which the most promising could be then investigated using biological experimentation.

### 5.3.2 LncRNA Functional Similarity

For measuring lncRNA functional similarity scores, Huang et al. developed a method called ILNCSIM [16]. Scores were calculated based on this assumption that two functionally similar lncRNAs are more likely to be related to functionally similar diseases. ILNCSIM consists of two steps. In step 1, ILNCSIM finds the common ancestors of each pairs of the diseases and then based on their directed acyclic graph (DAG), it calculates their functional similarities. In step 2, for each pair of lncRNAs, the lncRNA functional similarity was calculated using the semantic similarities of all diseases that are related to these two lncRNAs.

### 5.3.3 Disease Semantic Similarity

Diseases can be described as a Directed Acyclic Graph (DAG)in order to calculate disease semantic similarities. Each node in the graph represents a disease and edges of the graph show the relationship between diseases. Disease $d_1$ can be described as $DAG(d_1) = (d_1, S_{d_1}, E_{d_1})$ where $S_{d_1}$ is the set of all nodes including all ancestors of node $d_1$ including $d_1$ itself and $E_{d_1}$ is the set of all corresponding links, this includes all direct edges from parents to child nodes. So the contribution of disease $d$ in disease $d_1$ can be calculated as:

$$S_{d_1}(d) = \begin{cases} 1 & d_1 = d \\ max\{\Delta * S_{d_1}(d')|d' \in \text{children of } d\} & d_1 \neq d \end{cases} \tag{5.1}$$

Where $\Delta$ is the contribution factor for all links from disease $d$ to disease $d'$. The contribution of disease $d_1$ to its own semantic value is 1 so the farther the nodes from disease $d_1$, the less effect they have on the $d_1$'s semantic value. So, the value of $\Delta$ should be between 0 and 1. Wang et al. suggested that when the value of $\Delta = 0.5$ then the results show better correlation with the expression similarity [17]. The semantic value(SV) of disease $d_1$ can be described as:

$$SV(d_1) = \sum_{d \in D_{d_1}} S_{d_1}(d) \tag{5.2}$$

When two diseases are much more similar in the DAG, their similarity value would become larger. The semantic similarity between diseases $d_1$ and $d_2$ can be calculated as:

$$SV(d_1, d_2) = \frac{\sum_{d \in (S_{d_1} \cap S_{d_2})} (S_{d1}(d) + S_{d2}(d))}{SV(d_1) + SV(d_2)} \tag{5.3}$$

where $S_{d_1}(d)$ is the semantic value of disease $d$ related to disease $d_1$ and $S_{d_2}(d)$ is the semantic value of disease $d$ related to disease $d_2$. What equation 5.3 calculates, is the semantic similarity between two different diseases based on their location in DAG and the common links in their ancestors [18].

### 5.3.4 Feature Extraction for LncRNAs and Diseases

For the aim of extracting primary features for both lncRNAs and diseases, we use singular value decomposition (SVT) to perform PCA. If we denote the lncRNA similarity matrix as $R$ and disease similarity matrix as $D$, and because $R$ and $D$ are symmetric, if we perform SVT on $R$ and $D$, we have:

$$R = U_r \Sigma_r U_r^* \tag{5.4}$$

$$D = U_d \Sigma_d U_d^* \tag{5.5}$$

where $U_d$ and $U_r$ are unitary matrices and $\Sigma$ is a diagonal matrix with non-negative values in descending order on the diagonal. For finding the most significant singular values of matrices $R$ and $D$, we use the energy approach explained in [19]. Based on this study, the energy energy of a matrix $A$ is defined as:

$$E(A) = \|A\|_F = \sum_{i=1}^{r} \sigma_i \tag{5.6}$$

where $\sigma_i$ are the singular values of matrix $A$. In the same way, the energy of a k-rank approximation $U_a \Sigma_a V_a^*$ is:

$$E(U_a \Sigma_a V_a^*) = \|U_a \Sigma_a V_a^*\|_F \tag{5.7}$$

so the percentage of the energy occupied by the k-rank approximation with respect to the overall energy of $A$ becomes:

$$P = \frac{\|U_a \Sigma_a V_a^*\|_F}{\|A\|_F} = \frac{\sum_{i=1}^{k} \sigma_{ai}}{\sum_{i=1}^{r} \sigma_i} \tag{5.8}$$

where $\sigma_{ai}$ is the $i$th non-zero diagonal elements of $\Sigma_a$. Then we can find the proper parameters $P_r$ and $P_d$ for lncRNAs and diseases respectively:

$$p_r = arg_k min(\frac{\sum_{i=1}^{k} \sigma_{ii}}{\sum_{j=1}^{r} \sigma_{jj}} \geq \alpha_r) \tag{5.9}$$

$$p_d = arg_k min(\frac{\sum_{i=1}^{k} \sigma_{ii}}{\sum_{j=1}^{d} \sigma_{jj}} \geq \alpha_d) \tag{5.10}$$

where $\alpha_r$ and $\alpha_d$ are parameters. After we find the proper value for both $p_r$ and $p_d$, we construct feature matrices using top singular vectors corresponding to the $p_r$ and $p_d$ singular values for both lncRNAs and diseases. Our feature matrices become:

$$R = (V_{r1}, V_{r2}, ..., V_{pr})$$

$$D = (V_{d1}, V_{d2}, ..., V_{pd})$$

### 5.3.5 LncRNA-Disease Association Matrix Completion

The location of the known values can be encoded in $\Omega$, where $(i, j) \in \Omega$ if the value at indices (i,j) is known. We can define a function $P_\Omega(A)$ that returns a matrix where values in $\Omega$ are the same as the input matrix while it set the others to zero:

$$P_\Omega(A)_{i,j} \begin{cases} A_{i,j} & (i,j) \in \Omega \\ 0 & (i,j) \notin \Omega \end{cases} \tag{5.11}$$

We formulate the LDAPM problem based on inductive matrix completion (IMC) [20]. IMC formulation combines side information associated with rows and columns (in our case, lncRNAs and diseases respectively). The goal here is to complete data matrix $A$ using matrices $R$ and $D$ so we can consider the following IMC problem:

$$A = RMD^T \tag{5.12}$$

where $R$ is the feature matrix for lncRNAs, $D$ is the feature matrix for diseases and $M$ is an unknown matrix. The process is illustrated in Figure 5.1. The problem here is to find a low rank matrix $M$ based on the known values of lncRNA disease association matrix $A$. We can denote $M$ as $M = WH^T$, where $W \in \mathbb{R}^{p_r \times k}$ and $H \in \mathbb{R}^{p_d \times k}$ and $k$ is small. In order to relax the low rank constraint of $M$, we replace it with nuclear norm of $M = WH^T$ which we can rewrite as $\frac{1}{2}(\|W\|_F^2 + \|H\|_F^2)$ [21]. We can obtain $W$ and $H$ by solving the following optimization problem:

$$min_{W,H} \sum_{(i,j)} L(A_{i,j}, r_i^T WH^T d_j) + \frac{\lambda}{2}(\|W\|_F^2 + \|H\|_F^2) \tag{5.13}$$

where $\lambda$ is regularization parameter and $L$ is the loss function and is: $L(x, y) = (x - y)^2$ so the problem becomes:

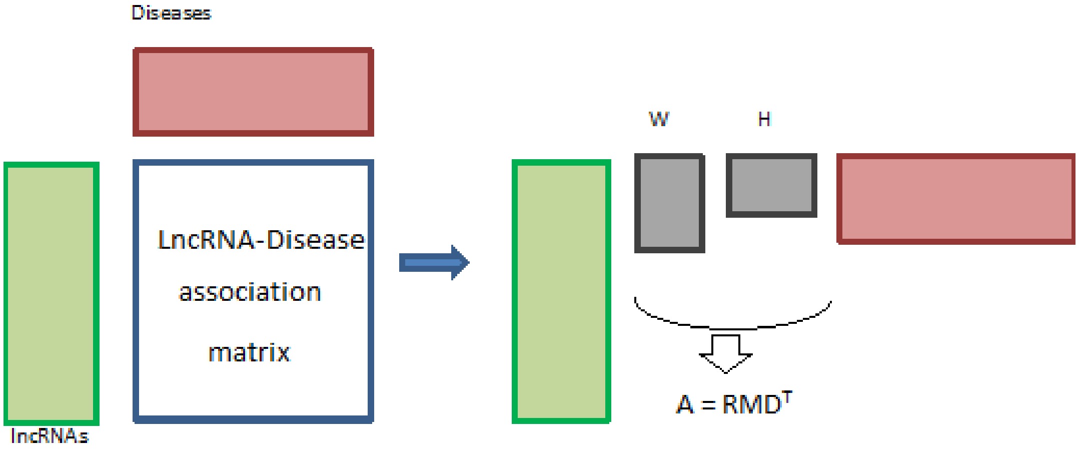

Figure 5.1. The process of LDAPM using lncRNA and disease features..

$$min_{W,H} \sum_{(i,j)} (A_{i,j} - r_i^T W H^T d_j)^2 + \frac{\lambda}{2}(\|W\|_F^2 + \|H\|_F^2) \tag{5.14}$$

Here an entry $A_{i,j}$ is modeled as $r_i^T M d_j$ where $M$ can be recovered by solving Eq.5.14. So $M$ becomes the result of multiplication of matrices $W$ and $H$. Eq.5.14 can be solved using Algorithm 1.

**Algorithm 1: LncRNA-Disease Association Matrix Completion**

```
Input: A
Output: W, H
Extract lncRNA features and store it in matrix R
Extract disease features and store it in matrix D
Initialize W and H with random numbers such that
the constraint should be met:
$W_{i,j} \geq 0$
$H_{i,j} \geq 0$
while not converged do:
    $H^{jk} = H^k \frac{(D^T A^T RW)_{jk}}{(D^T DHW^T R^T RW + \lambda.H)_{jk}}$
    $W^{ik} = W^k \frac{(R^T A^T DH)_{ik}}{(R^T RWH^T D^T DH + \lambda.W)_{ik}}$
end while
return W, H
```

## 5.4 Evaluation

In this section we will be first describing lncRNA-disease datasets and then we compare the performance of LDAPM approach to three state-of-the-art methods to see how well each methods perform.

### 5.4.1 Datasets

We used three lncRNA-disease association datasets from LncRNADisease database. First dataset contains 256 lncRNAs and 189 diseases with the total of 685 known association. Second dataset contains 145 lncRNAs and 176 diseases with the total of 293 known associations. Last dataset contains 156 lncRNAs and 189 diseases with the total of 351 known associations. then we removed the repeating data and also those that were not belong to human beings. The final statistics of our datasets is summarized in table 5.1.

| | **Number of lncRNAs** | **Number of diseases** | **Number of known associations** |
|---|---|---|---|
| **Dataset 1** | 285 | 226 | 621 |
| **Dataset 2** | 112 | 150 | 276 |
| **Dataset 3** | 131 | 169 | 319 |

Table 5.1. The statistics of datasets after removing repeating and non-human data.

### 5.4.2 Parameters $P_r$ and $P_d$

$\alpha_r$ and $\alpha_d$ are the number of feature vectors for lncRNAs and diseases respectively so choosing the proper amount for them is important. To find the proper amount for $\alpha_r$ and $\alpha_d$, we measure the AUC of LDAPM when $0.1 \leq \alpha_r, \alpha_d \leq 0.9$. We see that when $0.6 \leq \alpha_r \leq 0.8$, the performance becomes much stronger but above 0.8, the AUC decreases rapidly. For $\alpha_d$, the AUC is maximum when $\alpha_d = 0.6$. Based on these result we choose $\alpha_r = 0.7$ and $\alpha_d = 0.6$ as default.

### 5.4.3 Convergence

Matrix $M$ is reconstructed iteratively until the error in the convergence of the known associations is lower than a threshold:

$$\frac{\left\|P_\Omega(W_{k+1}H_{k+1}^T - M_k)\right\|_F}{\||P_\Omega(M_k)\|_F} \leq \epsilon \tag{5.15}$$

In our implementation, we set the threshold to $10^{-6}$. Also we set the upper limit of the number of iterations to 500.

### 5.4.4 Competitive Methods

We compare the performance of LDAPM with three state-of-the-art methods on the same three datasets that we mentioned before.

1. **LRLSLDA**: The first method is LRLSLDA proposed by Chen et al. [15]. They developed a model of Laplacian Regularized Least Squares, a semi-supervised learning method for predicting LncRNA–Disease Association (LRLSLDA). LRLSLDA prioritizes lncRNAs for any specific disease by integrating known lncRNA-disease association obtained from the LncRNADisease database, disease similarity network and lncRNA similarity network.

2. **RWRlncD**: The second method was proposed by Sun et al. [22], in which they proposed a global network-based computational framework to infer potential human lncRNA-disease associations. To do this, they implemented the random walk on a lncRNA functional similarity network. They evaluated the performance of RWRlncD by experimentally verified lncRNA-disease associations, based on leave-one-out cross-validation.

3. **RWRH**: The third method proposed by Li ae al. [23], in which they used the OMIM database in order to construct a heterogeneous network. This was done by connecting the phenotype network and gene network using the phenotype–gene association. They extended the random walk with restart algorithm to the heterogeneous network. RWRH prioritizes the genes and phenotypes and they used leave-one-out cross-validation to evaluate the ability of gene–phenotype association prediction.

### 5.4.5 Performance

We plotted Receiver Operating Characteristics curve (ROC) and used Area Under the ROC curve (AUC) as the main metric for evaluating their performance. The area under ROC curve is calculated as an index of the prediction power of the LDAPM method. The value of AUC is between 0 and 1 and higher amounts shows more prediction power. If the value is equal to 0.5, it means the performance is equal to a random prediction.

For an specific disease $d_1$, all the known $d_1$ related lncRNAs are defined as labeled nodes and the remaining lncRNAs (on which there is no relevance information) are defined as unlabeled nodes. Given a threshold $\delta$, if the result prediction of a labeled node is greater than $\delta$, then the node is identified as positive sample. If the result prediction of a an unlabeled node is less than $\delta$, then the node is a considered as a negative sample. To plot a ROC curve we calculated the true positive rates (TPR or sensitivity) and false positive rates (FPR, 1-specificity) through:

$$\begin{aligned} TPR &= \frac{TP}{TP+FN} \\ FPR &= \frac{FP}{TN+FP} \end{aligned} \tag{5.16}$$

where:

- TP : Number of correctly identified positive samples.
- TN : Number of correctly identified negative samples.
- FP : Number of misidentified positive samples.
- FN : Number of misidentified negative samples.

Sensitivity means the percentage of the positive samples that are correctly identified among all the positives and specificity means the percentage of the negative samples correctly identified among all negatives.

### 5.5 Results

We conducted five repetitions of a 5-fold cross validation for each of the methods. The 5-fold cross validation is implemented using the known lncRNA-disease association in the LncRNADisease database. So in each repetition, we divided our association matrix $A$ in to 5 parts and each parts, one-by-one , was left out as the test set while we used the remaining four parts as the training set. All of unknown lncRNA-disease association pairs can be seen as candidate samples. After applying LDAPM, scores of the test samples were compared with the all scores of the candidates samples. In order to make the validation more accurate, we repeated this process 5 times. Figures 3, 4 and 5 depicts the ROC curve and the calculated AUC of different methods in dataset 1, 2 and 3 respectively. We set $\lambda = 0.2$ in the optimization problem 5.14. We use the best value for parameters obtained by cross-validation for LDAPM method.

As it shows in figure 5.2, we can see that for dataset 1, the LDAPM obtained the AUC of 0.88 which is significantly higher than the AUC of other methods which are 0.60, 0.68 and 0.75 for RWRLncD, RWRH and LRLSLDA respectively. for dataset 2 (figure 5.3), the LDAPM obtained the AUC of 0.81 which is significantly higher than the AUC of other methods which are 0.65, 0.60 and 0.62 for RWRLncD, RWRH and LRLSLDA respectively. Likewise for dataset 3 (figure 5.4), the LDAPM obtained the AUC of 0.83 which is significantly higher than the AUC of other methods which are 0.69, 0.64 and 0.63 for RWRLncD, RWRH and LRLSLDA respectively. These results suggest that the LDAPM outperforms other methods and can achieve more accurate associations.

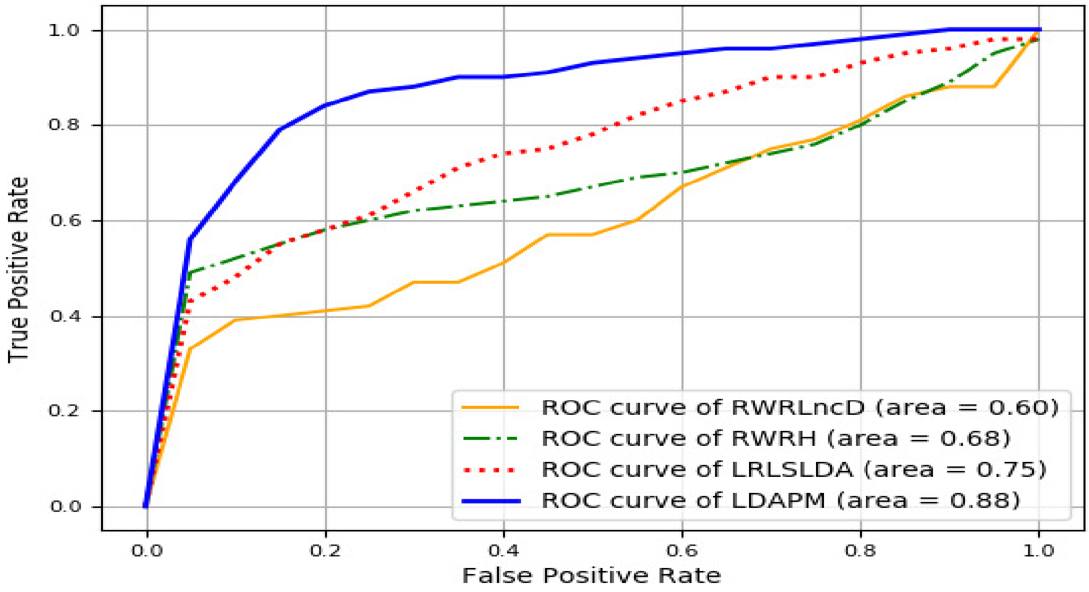

Figure 5.2. Evaluation the performance of different methods on Dataset 1.

## 5.6 Conclusion

Finding the molecular mechanism of diseases can help exploring disease prevention, prognosis, diagnosis and also effective treatments. LncRNAs as a class of non-coding RNAs are responsible for regulating gene expression so they can cause various diseases [24–26]. Some computational approaches have been proposed to capture lncRNA-disease association. However most of these methods have limitations in finding accurate associations.

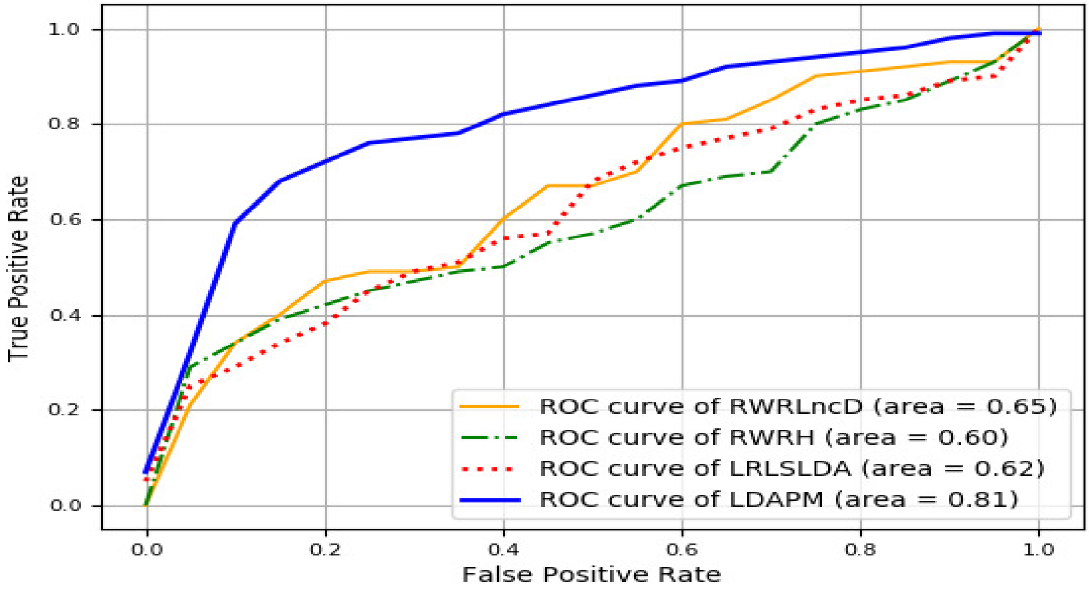

Figure 5.3. Evaluation the performance of different methods on Dataset 2.

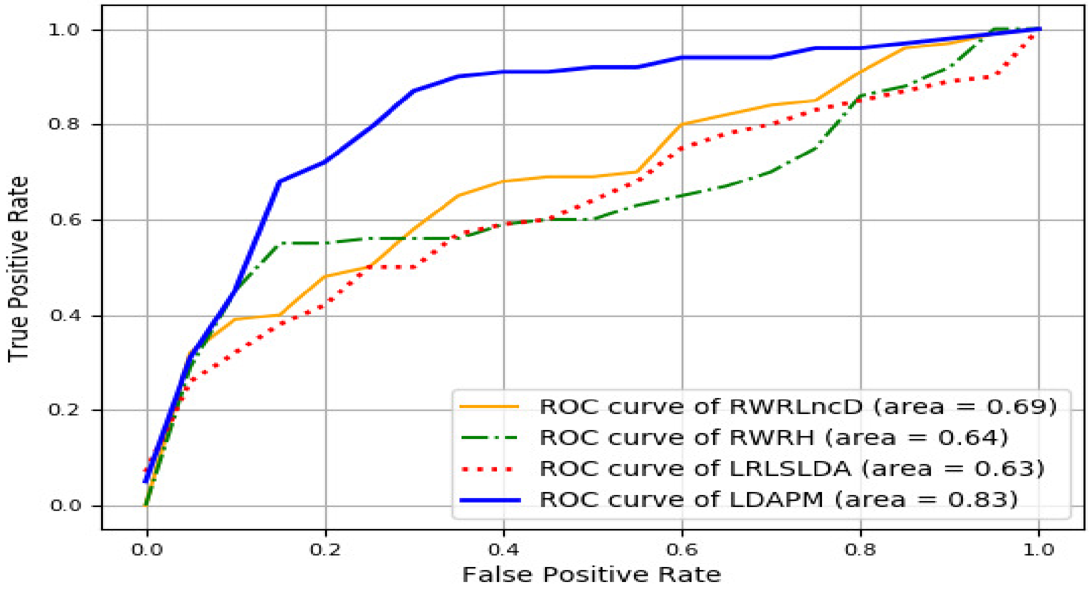

Figure 5.4. Evaluation the performance of different methods on Dataset 3.

In this paper we investigated LDAPM approach to predict lncRNA-disease associations based on this assumption that similar diseases tend to be related to functionally similar lncRNAs. We used three datasets from LncRNADisease dataset and calculated lncRNA similarity and disease similarity. We also extracted lncRNA and disease features using PCA. The idea behind the LDAPM is to find a low rank matrix that can integrate lncRNA and disease features to recover the lncRNA-disease association matrix. We compared the performance of LDAPM to three state-of-the-art-method and saw that the LDAPM outperforms other approaches in case of finding more accurate associations.

LncRNAs are regulate gene expression so mutation and dysfunction of lncRNAs can lead to several diseases. Computational methods have been proposed to find the relationship between lncRNAs and diseases. We hope that this study can open new opportunities to lncRNA-disease studies. lncRNA-disease experiments are very expensive and time consuming so computational methods can help biologists in this manner.

# REFERENCES

[1] Ge M, Li A, Wang M. A Bipartite Network-based Method for Prediction of Long Non-coding RNA–protein Interactions. Genomics, Proteomics and Bioinformatics (14) (2016) 62-71.

[2] Carninci P, Kasukawa T, Katayama S.: The transcriptional landscape of the mammalian genome. Science (309) (2005) 1559–63.

[3] Cabili MN, Trapnell C, Goff L.: Integrative annotation of human large intergenic noncoding RNAs reveals global properties and specific subclasses. Genes Development (18) (2011) 1915-27.

[4] Chen X, Huang L.: LRSSLMDA: Laplacian Regularized Sparse Subspace Learning for MiRNA-Disease Association prediction. PLOS Computational Biology (13) (2017).

[5] You ZH, Huang ZA, Zhu Z, et al.: PBMDA: a novel and effective path-based computational model for miRNA-disease association prediction. PLOS Computational Biology (13) (2017).

[6] Li A, Zhang Z. : Role of long non-coding RNA in diabetes mellitus and its complications. Sheng Wu Gong Cheng Xue Bao(32) (2016) 284–91.

[7] Johnson R.: Long non-coding RNAs in Huntington's disease neurodegeneration. Neurobiological Diseases (46) (2012) 245–54.

[8] Busch A, Eken SM, Maegdefessel L.: Prospective and therapeutic screening value of non-coding RNA as biomarkers in cardiovascular disease. Annals of Translational Medicine (4) (2016).

[9] Di Cecilia S, Zhang F, Sancho-Medina A, et al.: RBM-AS1 is critical for self-renewal of colon cancer stem-like cells. Cancer Research (2016).

[10] Atala A. : Re: the long noncoding RNA SChLAP1 promotes aggressive prostate cancer and antagonizes the SWI/SNF complex. Journal of Urology (192) (2014).

[11] Fenner A. Kidney cancer: AR promotes RCC via lncRNA interaction. Nature Reviews Urology (13) (2016).

[12] Zhang Q, Chen CY, Yedavalli VS, et al. NEAT1 long noncoding RNA and paraspeckle bodies modulate HIV-1 posttranscriptional expression. American Society for Microbiology (4) (2013).

[13] Lan W, Li M, Zhao K, Liu J, Wu FX, Pan Y, Wang J LDAP: a web server for lncRNA-disease association prediction. Bioinformatics (3) (2017) 458-460.

[14] Zhang J, Zhang Z, Chen Z, Deng L. Integrating Multiple Heterogeneous Networks for Novel LncRNA-Disease Association Inference. IEEE/ACM Transactions on Computational Biology and Bioinformatics (2)(2019) 396-406.

[15] Chen X, Yan G. Novel human lncRNA-disease association inference based on lncRNA expression profiles. Bioinformatics (29) (2013) 2617-2624.

[16] Huang Y, Chen X, You Z, Huang D, Chan K. ILNCSIM: improved lncRNA functional similarity calculation model Oncotarget (18) (2016) 25902-25914.

[17] Wang D, Wang J, Lu M, Song F and Cui Q. Inferring the human microRNA functional similarity and functional network based on microRNA-associated disease Bioinformatics (26) (2010) 1644-1650.

[18] Fraidouni N, Zaruba G. Graph Regularized Matrix Factorization for MiRNA-Disease Association Prediction BIOCOMP (2018) 10-18.

[19] Ji H, Yu W, Li Y. A Rank Revealing Randomized Singular Value Decomposition (R3SVD) Algorithm for Lowrank Matrix Approximations arXiv preprint arXiv: 1605.08134.

[20] Jain P, Dhillon I.S. Provable Inductive Matrix Completion arXiv preprint arXiv: 1306.0626.

[21] Natarajan N, Dhillon I. Inductive matrix completion for predicting gene–disease associations. Bioinformatics(30) (2014) 60-68.

[22] Sun J, Shi H, Wang Z, Zhang C, Liu L, Wang L, He W, Hao D, Liu S, Zhou M. Inferring novel lncRNA–disease associations based on a random walk model of a lncRNA functional similarity network. Molecular Biosystems (8) (2014).

[23] Li Y, Patra J. Genome-wide inferring gene–phenotype relationship by walking on the heterogeneous network Bioinformatics (26) (2010) 1219-1224.

[24] Gutschner T, Hammerle M, Eissmann M, Hsu J, Kim Y, Hung G, Revenko A, Arun G, Stentrup M, Gross M, Zornig M, MacLeod AR, Spector DL, Diederichs S. The noncoding RNA MALAT1 is a critical regulator of the metastasis phenotype of lung cancer cells. Cancer Research (73) (2013) 1180-1189.

[25] Lorenzen JM, Thum T. Long noncoding RNAs in kidney and cardiovascular diseases. Nature Reviews Nephrology (12) (2016) 360-373.

[26] Jia M, Jiang L, Wang YD, Huang JZ, Yu M, Xue HZ. lincRNA-p21 inhibits invasion and metastasis of hepatocellular carcinoma through notch signaling-induced epithelial-mesenchymal transition. Hepatology Research (46) (2016) 1137-1144.

# CHAPTER 6

# Conclusion

## 6.1 First Project

In the first project, we investigated how Pearson Correlation Coefficient and Cosine Similarity could be applied and used on NCBI-GEO biological data to find (artificially introduced) missing values in the datasets. The GEO-NCBI (Gene Expression Omnibus) is a public repository of genomic data. GEO profiles show expression profiles for individual genes. We used the gene expression datasets of three studies that measured the mRNA levels of different genes in different subjects (the amount of mRNA levels show the gene expression values). We used the data information of Bladder cancer, Leukemia and Lung cancer study. We modeled the gene expression data as a matrix (where each row represents a gene and each column a subject); the entries of the matrix can then be mRNA measurements that show the extent of gene expressions. Since entries of the dataset are based on partial measurements, the dataset has missing values, and the problem is then to estimate the missing values and thus to recover the global matrix based on the known values.

If we measure gene expressions on a set of tissue samples, we should find groups of genes that are correlated to each other. Having this in mind, we can assume that our gene expression measurements can be a low rank matrix so we can predict the missing values and complete the matrix. In neighborhood based approaches for collaborative filtering in recommendation systems, the main goal is to find similarities between neighbors. When there is a missing value, the system tries to make a prediction based on other users' ratings (for the movies); the more similar a user is to the one that has a missing value, the more impact his/her rating should have on the prediction. Our problem in gene expression prediction is very similar to neighborhood based approach of collaborating filtering, with the main task to find similarity (correlation) between genes. We used two approaches to calculate correlation between genes, the Pearson Correlation Coefficient (PCC) and the Cosine Similarity (CS).

We then compared the performances of the two approaches to that of a recent nuclear-norm minimization based approach. To do this, we removed random elements from the datasets as represented by matrices and predicted them based on the assumption that subjects have similar tendencies; more precisely that characteristics of genes where genes work in groups for any process in body are similar. Our study provides an insight for future work especially in bio-medicine as well as recommender systems. We found the correlation based approaches to outperform the low nuclear-rank matrix completion approach, i.e., it predicted more accurate values. We have also found that Pearson correlation coefficient provides more accurate reconstructions when compared to cosine similarity when used on gene databases.

## 6.2 Second Project

In this project we described how Robust Principal Component Analysis (RPCA) can be applied and used on NCBI-GEO biological data to find (artificially introduced) missing values and recover the datasets. For a noise-free dataset, we can easily perform PCA and find the most significant orthogonal vectors by using singular value decomposition (SVD). In the presence of noise, we can use another approach called robust PCA (RPCA). The presence of this noise is common in many applications such as image processing and bioinformatics. Robust PCA has the ability to recover a low rank matrix from sparse noise. We used the Alternating Direction Method of Multipliers (ADMM) to solve the objective function. We then described three well known algorithms that can be used when recovering low rank matrices and we compared the performances of the four approaches. To do this, we removed random elements from the datasets as represented by matrices and predicted them based on the assumption that genes have similar behaviors in similar conditions. Our study provides an insight for future work especially in bio-medicine but also has implications to recommender systems. We found that ADMM approach outperforms the other three approaches, i.e., it predicted more accurate values.

## 6.3 Third Project

Finding the molecular mechanism of diseases can help exploring disease pathogenesis and finding effective treatments. MiRNAs as a class of non-coding RNAs are responsible for regulating gene expression so they can cause various diseases. Some computational approaches have been proposed to capture miRNA- disease association. However, these methods have limitations. In this paper we presented a Graph Regularized Matrix Factorization method (GRMF) to predict miRNA-disease associations based on the assumption that similar miRNAs (functionally) tend to be related to similar diseases (phenotypically). We used miRNA functional similarity, disease semantic similarity, and known miRNA-disease associations form the HDMM v.2.0 database. To verify the accuracy of the GRMF method, we used five repetitions of 6-fold cross validation. We compared the result of the GRMF method with three state-of-the-art methods and concluded that GRMF outperforms the other three in terms of AUC. We selected Breast Neoplasm as a case study in order to show the performance of GRMF for diseases which have no related miRNAs and based on the results, we could confirm all 50 miRNAs as identified by miR2Disease, dbDEMC and HDMM. As the second case study we chose Lymphoma to demonstrate the performance of GRMF and based on the results, we could confirm 45 miRNAs out of 50 as identified by dbDEMC, miR2Disease and experimental literature in PubMed. The GRMF method could provide an effective approach to study miRNA-disease associations. We also recognize that GRMF has some limitations which can be improved in future research. For example, the sequence information of miRNAs is used to measure miRNA similarity but some studies show that the structural information can be more effective. Furthermore, expression information of miRNAs could also be used to measure this similarity.

## 6.4 Fourth Project

LncRNAs as a class of non-coding RNAs are responsible for regulating gene expression so they can cause various diseases. Some computational approaches have been proposed to capture lncRNA-disease association. However most of these methods have limitations in finding accurate associations.

In this paper we investigated LDAPM approach to predict lncRNA-disease associations based on this assumption that similar diseases tend to be related to functionally similar lncRNAs. We used three datasets from LncRNADisease dataset and calculated lncRNA similarity and disease similarity. We also extracted lncRNA and disease features using PCA. The idea behind the LDAPM is to find a low rank matrix that can integrate lncRNA and disease features to recover the lncRNA-disease association matrix. We compared the performance of LDAPM to three state-of-the-art-method and saw that the LDAPM outperforms other approaches in case of finding more accurate associations.

LncRNAs are regulate gene expression so mutation and dysfunction of lncRNAs can lead to several diseases. Computational methods have been proposed to find the relationship between lncRNAs and diseases.

I hope that this study can open new opportunities to gene-disease studies. gene-disease experiments are very expensive and time consuming so computational methods can help biologists in this manner.