IMPROVED FUNNEL-GSEA USING ADAPTIVE ELASTIC-NET PENALIZATION METHOD TO IDENTIFY SIGNIFICANT GENE SETS

NURUL NADZIRAH BINTI MOHD HASRI

A thesis submitted in fulfilment of the
requirements for the award of the degree of
Master of Philosophy

School of Computing
Faculty of Engineering
Universiti Teknologi Malaysia

OCTOBER 2021

DEDICATION

This thesis is dedicated to my beloved mother, father, and siblings, who give me love, strength, and helps. To my beloved friends especially my supervisors and AIBIG member, thank you so much for all support and help.

ACKNOWLEDGEMENT

This thesis would not have been possible without the help and support from many peoples. Firstly, I would like to thank my supervisors, Dr. Zuraini Binti Ali Shah, and Dr. Chan Weng Howe for their excellent supervision, guidance, knowledge, belief, encouragement, and motivation throughout my Master's journey. My thanks also go to my friends and lab mates of Artificial Intelligence and Bioinformatics Research Group (AIBIG) for their advice and support.

I also would like to express my sincere gratitude, appreciation, and thank my examiners for their advice and comments. It helps me in completing this thesis comprehensively and successfully.

At last, and most importantly, none of this would have been possible without the love and patience from my family and has been a constant give the love, concern, strength, and their sacrifices and understanding all these years during the journey of my Master study.

ABSTRACT

Gene set enrichment analysis (GSEA) is one of the methods in functional class scoring (FCS) categories for gene set analysis. GSEA is a popular method that was developed to identify, analyse and interpret set of genes or pathways from high-throughput transcriptomics experiments which are significantly enriched to help further analysis by biologist researchers. Many methods have been developed to enhance the original procedure of the GSEA. One of the evolutions of the GSEA method is the use of the elastic-net to reduce the effect of overlapping that reduces the statistical power and instability of the inference at the level of the gene set. However, elastic-net has limitations as it is inconsistent and bias in estimation. Thus, an ADaptive ELastic-NET in GSEA (ADELNET-GSEA) with an adaptive elastic-net was proposed to achieve a better result in identifying more gene sets that are informative and significant. The key part of the adaptive elastic-net is the weight parameter. It enables the adaptive elastic-net to perform different amounts of shrinkage to the different variables. Consequently, the ADELNET-GSEA is also consistent and unbiased in estimation. This research utilized the real dataset of Influenza A H3N2 time-course gene expression. It was found that the ADELNET-GSEA outperformed the previous GSEA method by identifying higher numbers of informative and significant gene sets to the immune response to human influenza infection. ADELNET-GSEA was able to identify the new gene sets, which were Spliceosome and Ubiquitin Mediated Proteolysis gene sets, related to the immune response for influenza. These findings have been validated through a word search strategy proven by previous researchers. Based on this result, this research brings benefits to the biological context validation and able to clarify the reliability of the improved method in identifying the significant gene sets.

ABSTRAK

Analisis pengayaan set gen (GSEA) adalah salah satu kaedah dalam kategori pemarkahan kelas kefungsian (FCS) untuk analisis set gen. GSEA adalah kaedah yang dikenali yang dibangunkan untuk mengenal pasti, menganalisis dan mentafsirkan kumpulan gen atau laluan dari eksperimen transkripomik hasil yang tinggi yang diperkaya secara signifikan untuk membantu analisis lebih lanjut oleh penyelidik biologi. Banyak kaedah telah dibangunkan untuk meningkatkan prosedur asal GSEA. Salah satu evolusi kaedah GSEA adalah penggunaan jaring elastik untuk mengurangkan kesan pertindihan yang mengurangkan kekuatan statistik dan ketidakstabilan inferens pada tahap kumpulan gen. Walau bagaimanapun, jaring elastik mempunyai batasan kerana ia tidak konsisten dan berat sebelah dalam perkiraan. Oleh itu, ADaptive ELastic-NET di GSEA (ADELNET-GSEA) dengan jaring elastik adaptif dicadangkan untuk mencapai hasil yang lebih baik dalam mengenal pasti lebih banyak kumpulan gen yang bermaklumat dan signifikan. Bahagian utama dari jaring elastik adaptif adalah parameter berat. Ini membolehkan jaring elastik adaptif untuk melakukan penyusutan jumlah yang berbeza terhadap pemboleh ubah yang berbeza. Oleh itu, ADELNET-GSEA juga konsisten dan tidak berat sebelah dalam perkiraan. Penyelidikan ini menggunakan kumpulan data sebenar ekspresi gen kursus masa Influenza A H3N2. Didapati bahawa ADELNET-GSEA mengungguli kaedah GSEA sebelumnya dengan mengenal pasti bilangan set gen yang berinformasi dan signifikan terhadap tindak balas imun terhadap jangkitan influenza manusia. ADELNET-GSEA dapat mengenal pasti kumpulan gen baru, yang terdiri daripada kumpulan gen *Spliceosome* dan *Ubiquitin Mediated Proteolysis* yang berkaitan dengan tindak balas imun terhadap influenza. Penemuan ini telah disahkan melalui strategi pencarian kata yang dibuktikan oleh penyelidik sebelumnya. Berdasarkan keputusan ini, penyelidikan ini membawa manfaat kepada pengesahan konteks biologi dan dapat menjelaskan kebolehpercayaan kaedah yang ditambah baik dalam mengenal pasti kumpulan gen yang signifikan.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

GSEA	-	Gene Set Enrichment Analysis
RNA-seq	-	Ribonucleic Acid Sequence
DNA	-	Deoxyribonucleic acid
SGA	-	Single Gene Analysis
GSA	-	Gene Set Analysis
IGA	-	Individual Gene Analysis
ORA	-	Over-Representation Analysis
FCS	-	Functional Class Scoring
PT	-	Pathway Topology
MSigDB	-	Molecular Signatures Database
KEGG	-	Kyoto Encyclopaedia of Genes and Genomes
PAGE	-	Parametric Analysis of Gene set Enrichment
GAGE	-	Generally Applicable Gene-set Enrichment
CAMERA	-	Correlation Adjusted Mean Rank
FUNNEL-GSEA	-	Functional Elastic-net regression in Gene Set Enrichment Analysis
FPCA	-	Functional Principal Component Analysis
LASSO	-	Least Absolute Shrinkage and Selection Operator
SCAD	-	Smoothly Clipped Absolute Deviation
ADELNET-GSEA	-	Adaptive Elastic-net in Gene Set Enrichment Analysis
GEO	-	Gene Expression Omnibus
C	-	Cytosine
A	-	Adenine
G	-	Guanine
T	-	Thymine
RNA	-	Ribonucleic Acid
tRNA	-	Transfer Ribonucleic Acid
rRNA	-	Ribosomal Ribonucleic Acid
mRNA	-	Messenger Ribonucleic Acid

GO	-	Gene Ontology
EA	-	Enrichment Analysis
maSigFun	-	Microarray Significant Functional
NP	-	Nonparametric Test Statistic
TcGSA	-	Time-course Gene Set Analysis
LLCT	-	Longitudinal Linear Combination Test
SAM-GS	-	Significance Analysis of Microarray to Gene-Set Analysis
QuSAGE	-	Quantitative Set Analysis of Gene Expression
FDR	-	False Discovery Rate
ES	-	Enrichment Score
MES	-	Maximum Enrichment Score
SAM	-	Significance Analysis of Microarray
VIF	-	Variance Inflation Factor
MWU	-	Mann-Whitney U test
FN	-	False Negative
FP	-	False Positive
ROC	-	Receiver Operating Characteristic
Adaptive LASSO	-	Adaptive Least Absolute Shrinkage and Selection Operator
Elastic SCAD	-	Elastic Smoothly Clipped Absolute Deviation
MSA-Enet	-	Multi-step Adaptive Elastic-net
GCV	-	Generalized Cross Validation
AIC	-	Akaike Information Criterion
BIC	-	Bayesian Information Criterion
PMID	-	PubMed Unique Identifiers
pdf	-	Portable Document Format
RAM	-	Random-access Memory
FWER	-	Familywise Error Rate
IQR	-	Inter-quantile Range
FDA	-	Functional Data Analysis
PCA	-	Principal Component Analysis
EVs	-	Extracellular Vesicle

IAVs	-	Influenza A Viruses
DCs	-	Dendritic Cells
HAI	-	Hemagglutination Antibody Inhibition
PML	-	Promyelocytic Leukaemia Protein

LIST OF SYMBOLS

ϵ	-	Random noise
y	-	Respond vector
X	-	Gene matrix
β^*	-	Vector
F_i	-	F-statistic or F-value
RSS_i^0	-	Residual sum of squares under the null hypotheses
RSS_i^1	-	Residual sum of squares under alternative hypotheses
$\hat{\mu}_i$	-	Mean expression of the temporal sample
$\hat{\xi}_{il}$	-	Principal Component Score
$\hat{\phi}_l^k(t)$,	-	the lth eigen-function that obtain from FPCA method
$\hat{a}_i$	-	Weight parameter in adaptive elastic-net
λ	-	Tuning parameter

LIST OF APPENDICES

CHAPTER 1

INTRODUCTION

1.1 Overview

Over of the last few decades, the field of molecular biology has targeted and focused on studying biological systems at the molecular level which provided richer information. Then the comprehension of the genes and their function has been assisted by microarray experiments (Tusher *et al.*, 2001; Beadling and Smith, 2002; Xie *et al.*, 2007; Nan *et al.*, 2012; Mathur *et al.*, 2018). Across several of the clinical condition and experimental, RNA-seq and DNA microarray has made simultaneous expression level profiling of thousands of genes that can be widely accessible by researchers.

Microarray data was used in many areas such as cancer classification in order to build the powerful classifier, cancer diagnosis, providing more comprehensive understanding for complex disease, discovering and finding the hidden taxonomies (Piatetsky-Shapiro and Tamayo, 2003), data normalization (Quackenbush, 2002), identify biomarkers (Takamiya *et al.*, 2021) and other. One of the objectives of microarray analysis is to identify the constant differential expression pattern between two classes of samples. However, it needs to go through a critical data preparation step in biological function analyses as shown in Figure 1.1 to get microarray data. Such experiments generate a very large number of data that lead to difficult analyses, especially without great gene annotation.

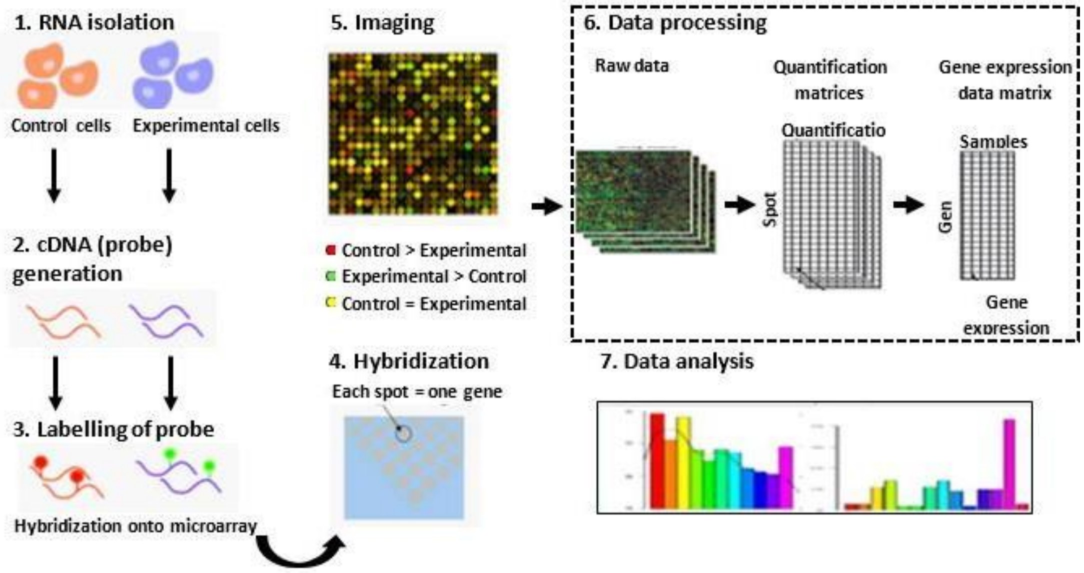


Figure 1.1 Overview of the microarray data preparation step

Another kind of microarray data is time-course gene expression data that also known as time-series data has gained more popularity in interpretation studies in recent years (Zhang *et al.*, 2011; Wu and Wu, 2013; Hejblum *et al.*, 2015; Khodayari Moez *et al.*, 2019). The time-course gene expression data is different from microarray data that usually use before. In which, it is the static experiment that captures only the expression value. Meanwhile, the time-course gene expression data capture the expression value over several time points in a given biological process. This enables the specialists and biologists to study the gene expression pattern over time points in order to monitor the dynamic behaviors of the genes (L. Wang *et al.*, 2007; Wu and Wu, 2013). Microarray advancements have made it conceivable to measure the gene expression values of all the genes.

In order to obtain significant outcomes, it is important to interpret these data sets accurately. Various inferential and statistical methods have been developed and available to extract useful information and detecting significant genes from these data sets in the past decade. For example, ErmineJ (Lee *et al.*, 2005), DAVID (Dennis *et al.*, 2003), and GeneMerge (Castillo-Davis and Hartl, 2003). Then for single time-course gene analysis is maSigPro (Conesa *et al.*, 2006), ANOVA based method model (Park *et al.*, 2003), and EDGE (Storey *et al.*, 2005). All these methods are known as single gene analysis (SGA) or individual gene analysis (IGA) (Nam and Kim, 2008).

It discovers differently expressed genes by evaluating every single gene. However, a usual microarray data has a dimensional limitation, where this data has a large number of genes and a frequently a small number of samples. This causes the interpretation expression level profile to remain a key challenge.

Thus, the concept of this area moved from the differential expression of single or individual genes to sets of biologically related genes, known as gene set analysis (GSA). This area divides into groups of analyses, which are network-based analysis and pathway-based analysis. The term "pathway-based analysis" has been used widely in the literature (Green and Karp, 2006; Khatri *et al.*, 2012) and is also known as gene set analysis. However, the term "gene set" is used in this thesis. Gene sets or pathways are ordinarily grouped by genes that share some of the basic or common biological properties such as having a common function, same metabolic pathway, or existence of the binding motif. Figure 1.2 shows the difference between single gene analysis and gene set analysis.

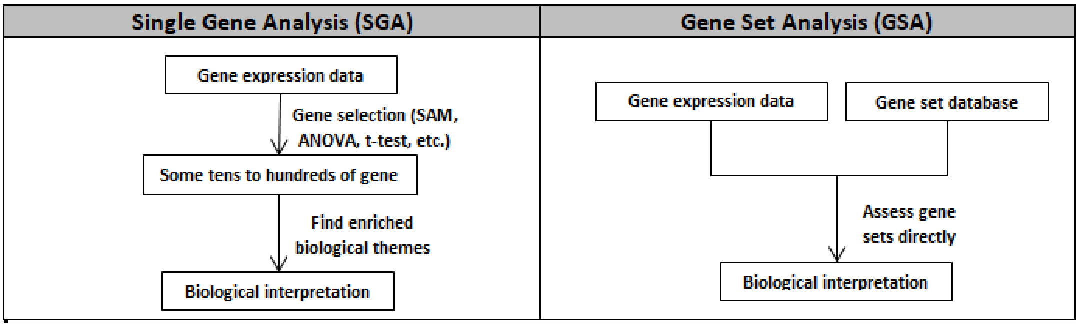

Single Gene Analysis (SGA)
Gene expression data
Gene selection (SAM, ANOVA, t-test, etc.)
Some tens to hundreds of gene
Find enriched biological themes
Biological interpretation
Gene Set Analysis (GSA)
Gene expression data
Gene set database
Assess gene sets directly
Biological interpretation

Figure 1.2 The comparison of single gene analysis (SGA) and the gene-set analysis strategy (GSA)

There are three generations of gene set or pathway analysis that have been described by Khatri *et al.*, (2012). These generations are different from each other based on their step and strategy. First generation is over-representation analysis (ORA) that follows the following strategy. Firstly, from the whole gene expression, it creates the list of input by using specific criteria or thresholds. After that, for every gene set, the inputs of genes that are part of the gene sets are counted. This procedure is repeated for a proper context list of genes. Finally, each gene set is tested for under or over-representation in the list of input genes. Chi-square, hypergeometric and binomial distribution are common tests have used. The second generation is functional class scoring (FCS). This generation has three main steps. Firstly, a gene-level statistic is computed from the differential expression of individual genes by statistic tests such as Kolmogorov-Smirnov statistic (Mootha *et al.*, 2003; Subramanian *et al.*, 2005), ANOVA (Al-Shahrour *et al.*, 2005), FPCA (Ramsay, 2005), Q-statistic (Goeman *et al.*, 2004) and Z-score (Kim and Volsky, 2005). Secondly, the gene-level statistic for all genes in gene sets is accumulated into a single gene set-level statistic. The gene set-level statistic that is commonly used is maxmean statistic (Efron and Tibshirani, 2007) and Wilcoxon rank-sum (Barry *et al.*, 2005). Finally, assessing the significant gene set from gene set-level statistics. The last generation is pathway topology (PT) which has the same step as the FCS method. However, this generation uses additional information such as genes interaction and pathway topology to compute gene-level statistics. In this research, FCS generation and method are used.

Gene set analysis has gained popularity and become the first option to interpret gene expression and protein in recent years because of its advantages. Firstly, gene set analysis has reduced the complexity of analysis by gathering the long list of individual genes into a smaller set of related genes. Secondly, it can have more explanatory power compared to individual gene analysis (Glazko and Emmert-Streib, 2009). Thirdly, it is successful to interpret the gene expression in terms of the molecular pathway, biological function, and genomic function (Zhang *et al.*, 2011). In addition, gene set analysis had emerged widely in microarray analysis due to the large number of open databases which can easily access the high-quality gene set or pathway datasets (Yaari *et al.*, 2013; Zhang *et al.*, 2017). The example of open databases is Molecular signatures database (MSigDB) (Liberzon *et al.*, 2011), Kyoto Encyclopaedia of Genes and Genomes (KEGG) (Kanehisa and Goto, 2000), Reactome (Croft *et al.*, 2010), BioCarta (Nishimura, 2001), and Pathway Interaction Database (Schaefer *et al.*, 2008). These databases are growing exponentially that enabling further opportunities for reveal new functional gene sets (Ideker *et al.*, 2002; Segal *et al.*, 2003; Sharan *et al.*, 2005; Chen and Yuan, 2006).

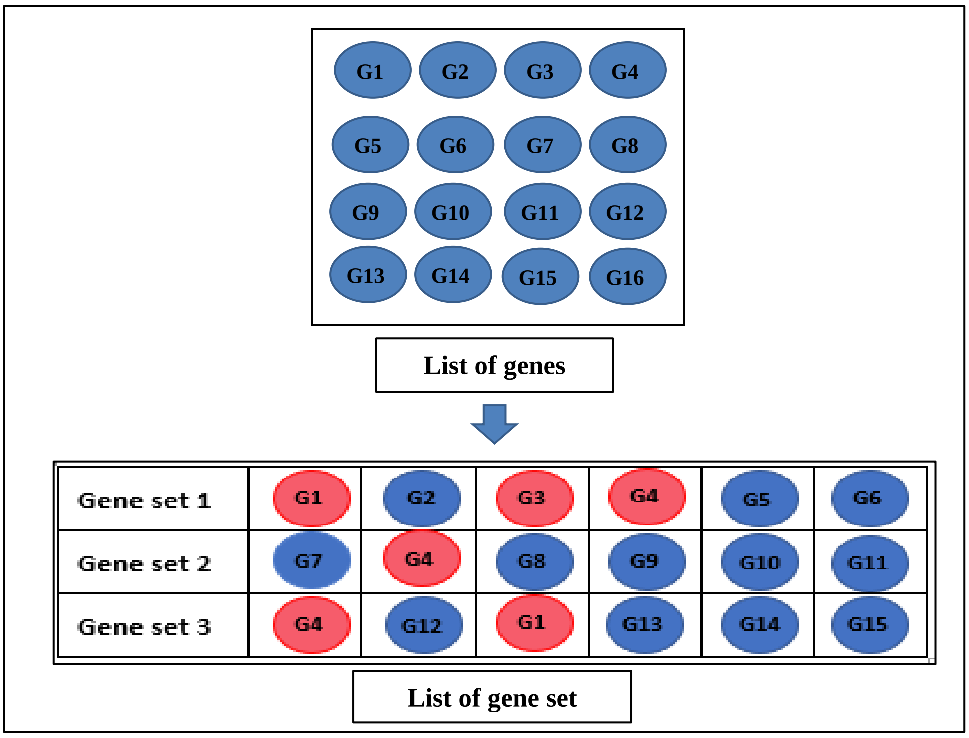
Due to these advantages, gene set analysis has turned into a well-known research area and numerous strategies have been developed to enhance the original Gene set Enrichment analysis (GSEA) procedure by Mootha *et al.*, (2003) and Subramanian *et al.*, (2005) to identify informative gene sets to related biological condition. For example, Parametric Analysis of Gene set Enrichment (PAGE) (Kim and Volsky, 2005), was used the normal distribution in statistical inference that reduces the computation effect compare to using permutation step. Besides, a Generally Applicable Gene-set Enrichment (GAGE) (Luo *et al.*, 2009) was used a two-sample t-test, adjust for the different microarray experiment designs, and separate the experimental gene set and canonical pathway to successfully apply for different sample sizes, profiling techniques, and experimental designs of microarray dataset. Correlation Adjusted Mean Rank gene set test (CAMERA) and its extension (Wu and Smyth, 2012; Yaari *et al.*, 2013) incorporate the adjustment of the inter-gene correlation to increase the false discoveries of numerous differential expression tests and gene-set test considerably. Lastly, Functional Elastic-net regression in Gene Set Enrichment Analysis (FUNNEL-GSEA) (Zhang *et al.*, 2017; Meng *et al.*, 2018) uses for time-course gene expression based on FPCA and use the elastic-net as the weight method or penalized method to decompose the overlapping effect that reduces the statistical power and instability of the inference at the level of the gene set.

The penalized method is the alternative or advance method for gene selection that is crucial for discovering the knowledge with high-dimensional data (Fan and Li, 2006). The penalized method could greatly improve the performance of the fitted model and gene-set analysis method. Thus, many statisticians have attempted to propose several penalization methods and strategies such as LASSO, adaptive LASSO, SCAD, elastic net, and adaptive elastic net to perform model selection and estimation simultaneously. Penalization methods shrink down to zero the coefficient of genes or markers that a have little apparent impact (Ayers and Cordell, 2010) on the phenotype of interest. Through the utilization of the penalization method, it can be discovering the subset of genes that are most associated with the phenotype of interest. Furthermore, penalization methods are able to handle the impacts of the multicollinearity, overfitting issues (Zakariya Yahya Algamal and Lee, 2015), and the effect of the overlapping (Zhang *et al.*, 2017). However, it keeps challenging to choose the better and suitable penalized method to implement in the gene set analysis method to achieve the better result in identify a significant gene set.

1.2 Problem Background

Microarray data analysis has been broadly utilized by researchers to enhance the biological interpretation and understanding of the analysis outcome. The conceptual on the differential expression of single or individual genes shift to sets of biologically related genes and known as gene set analysis (GSA) or pathway analysis. Integration of pathway data and information into the microarray data has enhanced the interpretation and analysis for achievement in microarray analysis.

However, most of the pathway definitions were discovered in the public database are usually curated from numerous studies of cultured cells and domain experts (Adriaens *et al.*, 2008) that obtain under different experimental conditions. Therefore, these gene sets or pathways are not context-specific and there is incredible overlap in these gene sets. The overweight for overlap of the important genes that shares by numerous sets can cause an increase the hypothesis test dependency, encourage type I error (false positive), reduce the power of statistical and instability of inferences at the gene set level (Qiu *et al.*, 2005; Qiu and Yakovlev, 2006, 2007; Gordon *et al.*, 2007; Zhang *et al.*, 2017). Figure 1.3 shows the example of the overlapping gene in the gene set. The red color presents the overlapping gene. One of the examples for overlap gene is G4 that be assigned to the gene set one, two, and three. However, the exact activation for the G4 in the context of influenza viral infection might not be inferred by all the gene sets that can activate that gene.



Gene set 1	G1	G2	G3	G4	G5	G6
Gene set 2	G7	G4	G8	G9	G10	G11
Gene set 3	G4	G12	G1	G13	G14	G15

Figure 1.3 Overlapping gene between the gene sets.

The penalized method is the alternative or advance method for the gene selection to help to reduce the overlapping effect for improving the performance gene set enrichment analysis in identifying the significant gene set. The elastic-net is one of the penalization methods that has been implemented in the FUNNEL-GSEA method. However, the elastic-net penalization method has some limitations. The elastic-net is lacked the oracle property due to the bias estimation same as LASSO even though it outperforms LASSO (Zou and Zhang, 2009; Zeng and Xie, 2014; Zakariya Y Algamal and Lee, 2015; Zakariya Yahya Algamal and Lee, 2015). Consequently, the elastic-net is inconsistent in estimation.

1.3 Problem Statement

Since the gene set analysis dataset consists of overlapping gene causes curated from numerous studies of the expert domain, the penalization method is required to reduce the overlapping effect to improve the performance of gene set enrichment analysis in identifying the significant gene set. The penalization methods are able to discover the subset of marks that are most associated with the disease or phenotype. The previous penalization method regularizes the entire variable coefficient in the gene set equally. As the result, the estimation can be biased for the large coefficient since the heavy shrinkage is imposed on a large coefficient.

1.4 Research aim

The aim of this research is to propose an improved gene set enrichment analysis method to better identifying the significant gene set from the time-course gene expression dataset for further analysis and examination through biological context validation by word search.

1.5 Objectives

The objectives of this research are specified as follows:

a) To propose an improved FUNNEL-GSEA method with integrating adaptive weight parameter in elastic-net penalization method to reduce overlapping effect for better identification of significant gene sets.

b) To discover the significant gene sets to immune response to human influenza infection for Influenza A H3N2 disease.

c) To propose a new validation approach through biological context validation by word search.

1.6 Research Scopes

The scopes of this research are as follow:

a) This research uses Rstudio software to run the source code and R programming language has been used.

b) The dataset of time-course gene expression going to be used in this research is Human influenza infection by influenza A H3N2 or Wisconsin virus that has been downloaded from Gene Expression Omnibus (GEO) repository website with GSE52428 series number.

c) CP: KEGG biological pathway is used as gene set data that has been downloaded from MSigDB database.

d) The research used the “gene set” or group of genes terms to refer as a pathway.

e) The performance measurements used in this research are F-value and p-value

f) The biological context validations by word search are used to validate the significant gene sets to justify the relationship between the gene set and the immune response.

1.7 Significance of the Research

The significance of this research is that the improved method able to better in identifying and more numbers significant gene sets that related to the immune response. It can help researchers and biologists to further study and analyze the significant gene sets for the production of products such as vaccines. Furthermore, the proposed weight parameter in adaptive elastic-net has the ability to produce consistent estimation in penalizing the coefficient of variable and able to reduce the overlapping effect in gene set data that usually affects the performance of methods. Besides, the usage of the time-course gene expression dataset allows for a better interpretation of temporal information and the dynamic behaviors of the gene. Finally, the improved method can be utilized in other biological areas related to human genomes for better interpretation and analysis.

1.8 Thesis Outline

This thesis is arranged into five chapters as follow:

Chapter 1: This chapter presents a detailed explanation of the research domain. This chapter helps to understand the general biological information that relates to this research. It contains the overview of the research domain, problem background, problem statement, research aim, objectives, research scopes, significance of this research, and thesis outline.

Chapter 2: This chapter reviews the revolution and trend from previous researchers that related to gene set enrichment analysis and penalization method.

Chapter 3: This chapter explains the research methodology in detail. It consists of the research framework and research materials such as time-course gene expression datasets and gene set or pathway data. Additionally, this chapter discusses the fundamental software and hardware requirement as well as the performance measurement for the evaluation process.

Chapter 4: This chapter describes the differences between the original FUNNEL-GSEA method from the previous researcher and the improved method. Furthermore, dataset pre-processing is also included in this chapter. This chapter also presents the design and development of the improved method, an improved ADaptive ELastic-NET in Gene Set Enrichment Analysis (ADELNET-GSEA) to identify more numbers of informative gene sets that related to immune response. Then, the result of the improved method and comparison with other methods is presented and discussed. Lastly, will be performed the biological context validation by word search.

Chapter 5: This chapter concludes by emphasizing the achievement of research and recommendations for the future direction of the present research.

REFERENCES

Adriaens, M. E., Jaillard, M., Waagmeester, A., Coort, S. L. M., Pico, A. R. and Evelo, C. T. A. (2008) 'The public road to high-quality curated biological pathways', *Drug discovery today*, 13, pp. 856–862.

Alharthi, A. M., Lee, M. H. and Algamal, Z. Y. (2021) 'Gene selection and classification of microarray gene expression data based on a new adaptive L1-norm elastic net penalty', *Informatics in Medicine Unlocked*, p.100622

Al-Shahrour, F., Díaz-Uriarte, R. and Dopazo, J. (2005) 'Discovering molecular functions significantly related to phenotypes by combining gene expression data and biological information', *Bioinformatics*, 21, pp. 2988–2993.

Algamal, Zakariya Y and Lee, M. H. (2015) 'High Dimensional Logistic Regression Model using Adjusted Elastic Net Penalty', *Pakistan Journal of Statistics and Operation Research*, 11(4), p. 667.

Algamal, Zakariya Yahya and Lee, M. H. (2015) 'Regularized logistic regression with adjusted adaptive elastic net for gene selection in high dimensional cancer classification', *Computers in Biology and Medicine*. Elsevier, 67, pp. 136–145.

Anbari, M. El and Mkhadri, A. (2014) 'Penalized regression combining the L1 norm and a correlation based penalty', *Sankhya B*, 76(1), pp. 82–102.

Ashburner, M., Ball, Catherine, A., Blake, Judith, A., Botstein, D., Butler, H., Cherry, J. M. and Sherlock, G. (2000) 'Gene Ontology: tool for the unification of biology', *Nature genetics*, 25(1), pp. 25–29.

Ayers, K. L. and Cordell, H. J. (2010) 'SNP Selection in genome-wide and candidate gene studies via penalized logistic regression', *Genetic Epidemiology*. John Wiley & Sons, Ltd, 34(8), pp. 879–891.

Barry, W. T., Nobel, A. B. and Wright, F. A. (2005) 'Significance analysis of functional categories in gene expression studies: a structured permutation approach', *Bioinformatics*, 21, pp. 1943–1949.

Beadling, C. and Smith, K. A. (2002) 'DNA array analysis of interleukin-2-regulated immediate/early genes', *Medical Immunology*, 1, p. 2.

Becker, N., Toedt, G., Lichter, P. and Benner, A. (2011) 'Elastic SCAD as a novel penalization method for SVM classification tasks in high-dimensional data', *BMC Bioinformatics*, 12(i).

Ben-Shaul, Y., Bergman, H. and Soreq, H. (2005) 'Identifying subtle interrelated changes in functional gene categories using continuous measures of gene expression', *Bioinformatics*, 21(7), pp. 1129–1137.

Bolón-Canedo, V., Sánchez-Maroño, N., Alonso-Betanzos, A., Benítez, J. M. and Herrera, F. (2014) 'A review of microarray datasets and applied feature selection methods', *Information Sciences*, 282, pp. 111–135.

Branagan, A. R., Duffy, E., Albrecht, R. A., Cooper, D. L., Seropian, S., Parker, T. L., Gan, G., Li, F., Zelterman, D. and Boddupalli, C. S. (2017) 'Clinical and Serologic Responses After a Two-dose Series of High-dose Influenza Vaccine in Plasma Cell Disorders: A Prospective, Single-arm Trial', *Clinical Lymphoma Myeloma and Leukemia*, 17, pp. 296-304. e2.

Breslin, T., Edén, P. and Krogh, M. (2004) 'Comparing functional annotation analyses with Catmap', *BMC Bioinformatics*, 5(1), pp. 1–8.

Bühlmann, P., Rütimann, P., van de Geer, S. and Zhang, C. H. (2013) 'Correlated variables in regression: Clustering and sparse estimation', *Journal of Statistical Planning and Inference*. Elsevier, 143(11), pp. 1835–1858.

Castillo-Davis, C. I. and Hartl, D. L. (2003) 'GeneMerge—post-genomic analysis, data mining, and hypothesis testing', *Bioinformatics*, 19, pp. 891–892.

Chan, W. H. (2016) *Identification of informative genes and pathways using improved penalized support vector machine for cancer classification*. Universiti Teknologi Malaysia.

Chan, W. H., Mohamad, M. S., Deris, S., Corchado, J. M., Omatu, S., Ibrahim, Z. and Kasim, S. (2016) 'An improved gSVM-SCADL2 with firefly algorithm for identification of informative genes and pathways', *International Journal of Bioinformatics Research and Applications*, 12(1), pp. 72–93.

Chen, J. and Yuan, B. (2006) 'Detecting functional modules in the yeast protein–protein interaction network', *Bioinformatics*, 22, pp. 2283–2290.

Chen, X. (2011) 'Adaptive elastic-net sparse principal component analysis for pathway association testing', *Statistical Applications in Genetics and Molecular Biology*, 10(1).

Ciuperca, G. (2018) ‘Adaptive elastic-net and fused estimators in high-dimensional group quantile linear model’, *arXiv*.

Coleman, M. D., Ha, S.-D., Haeryfar, S. M. M., Barr, S. D. and Kim, S. O. (2018) ‘Cathepsin B plays a key role in optimal production of the influenza A virus.’, *Journal of virology & antiviral research*. PMC Canada manuscript submission, 2018, pp. 1–20.

Conesa, A., Nueda, M. J., Ferrer, A. and Talón, M. (2006) ‘maSigPro: a method to identify significantly differential expression profiles in time-course microarray experiments’, *Bioinformatics*, 22, pp. 1096–1102.

Croft, D., O’kelly, G., Wu, G., Haw, R., Gillespie, M., Matthews, L., Caudy, M., Garapati, P., Gopinath, G. and Jassal, B. (2010) ‘Reactome: a database of reactions, pathways and biological processes’, *Nucleic acids research*, 39, pp. D691–D697.

Cypryk, W., Lorey, M., Puustinen, A., Nyman, T. A. and Matikainen, S. (2016) ‘Proteomic and bioinformatic characterization of extracellular vesicles released from human macrophages upon influenza A virus infection’, *Journal of proteome research*, 16, pp. 217–227.

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C. and Lempicki, R. A. (2003) ‘DAVID: database for annotation, visualization, and integrated discovery’, *Genome biology*, 4, p. R60.

Dienz, O., Rud, J. G., Eaton, S. M., Lanthier, P. A., Burg, E., Drew, A., Bunn, J., Suratt, B. T., Haynes, L. and Rincon, M. (2012) ‘Essential role of IL-6 in protection against H1N1 influenza virus by promoting neutrophil survival in the lung’, *Mucosal Immunology*, 5(3), pp. 258–266.

Dinu, I., Potter, J. D., Mueller, T., Liu, Q., Adewale, A. J., Jhangri, G. S., Einecke, G., Famulski, K. S., Halloran, P. and Yasui, Y. (2007) ‘Improving gene set analysis of microarray data by SAM-GS’, *BMC Bioinformatics*, 8(1), pp. 1–13.

Dørum, G., Snipen, L., Solheim, M. and Sæbø, S. (2009) ‘Rotation testing in gene set enrichment analysis for small direct comparison experiments’, *Statistical Applications in Genetics and Molecular Biology*, 8(1).

Edinger, T. O., Pohl, M. O., Yángüez, E. and Stertz, S. (2015) ‘Cathepsin W Is Required for Escape of Influenza A Virus from Late Endosomes.’, *mBio*. American Society for Microbiology (ASM), 6(3), p. e00297.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R. (2004) 'Least angle regression', *The Annals of Statistics*, 32(2), pp. 407–499.

Efron, B. and Tibshirani, R. (2007) 'On testing the significance of sets of genes', *The annals of applied statistics*, 1, pp. 107–129.

Engchuan, W., Meechai, A., Tongsima, S. and Chan, J. H. (2015) *Cross-platform pathway activity transformation and classification of microarray data, In Computational Intelligence in Information Systems*.

Fan, J. and Li, R. (2001) 'Variable Selection via Nonconcave Penalized Likelihood and its Oracle Properties', *Journal of the American Statistical Association*. Taylor & Francis, 96(456), pp. 1348–1360.

Fan, J. and Li, R. (2006) 'Statistical Challenges with High Dimensionality: Feature Selection in Knowledge Discovery'.

Fan, J. and Lv, J. (2010) 'A Selective Overview of Variable Selection in High Dimensional Feature Space', *Statistica Sinica*, 20(1), pp. 101–148.

Fisher, R. A. (1992) 'Statistical Methods for Research Workers.', *In Breakthroughs in statistics. Springer,* pp. 66–70.

Friedman, J., Hastie, T. and Tibshirani, R. (2010) 'Regularization paths for generalized linear models via coordinate descent', *Journal of Statistical Software*, 33(1), pp. 1–22.

Gatti, D. M., Barry, W. T., Nobel, A. B., Rusyn, I. and Wright, F. A. (2010) 'Heading Down the Wrong Pathway: On the Influence of Correlation within Gene Sets', *BMC Genomics*, 11(1), pp. 1–10.

Ghosh, S. (2007) 'Adaptive Elastic Net : An Improvement of Elastic Net to achieve Oracle Properties', *Most*.

Ghosh, S. (2011) 'On the grouped selection and model complexity of the adaptive elastic net', *Statistics and Computing*, 21(3), pp. 451–462.

Glazko, G. V and Emmert-Streib, F. (2009) 'Unite and conquer: univariate and multivariate approaches for finding differentially expressed gene sets', *Bioinformatics*, 25, pp. 2348–2354.

Goeman, J. J., Van De Geer, S. A., De Kort, F. and Van Houwelingen, H. C. (2004) 'A global test for groups of genes: testing association with a clinical outcome', *Bioinformatics*, 20, pp. 93–99.

Goli, S., Mahjub, H., Faradmal, J., Mashayekhi, H. and Soltanian, A. R. (2016) 'Survival Prediction and Feature Selection in Patients with Breast Cancer

Using Support Vector Regression', *Computational and Mathematical Methods in Medicine,* 2016.

Gordon, A., Glazko, G., Qiu, X. and Yakovlev, A. (2007) 'Control of the mean number of false discoveries, Bonferroni and stability of multiple testing', *The Annals of Applied Statistics*. Institute of Mathematical Statistics, 1(1), pp. 179–190.

Green, M. L. and Karp, P. D. (2006) 'The outcomes of pathway database computations depend on pathway ontology', *Nucleic Acids Research,* 34, pp. 3687–3697.

Haggag, M. M. M. (2018) 'Adjusting the Penalized Term for the Regularized Regression Models', *Afrika statistika,* 13(2), pp. 1609–1630.

Hagiwara, K. (2018) 'On an improvement of LASSO by scaling'.

Hejblum, B. P., Skinner, J. and Thiébaut, R. (2015) 'Time-Course Gene Set Analysis for Longitudinal Gene Expression Data', *PLoS Computational Biology,* 11(6), pp. 1–21.

Hoerl, A. E. and Kennard, R. W. (1970) 'Ridge Regression: Biased Estimation for Nonorthogonal Problems', *Technometrics,* 12(1), pp. 55–67.

Hu, J., Huang, J. and Qiu, F. (2018) 'A group adaptive elastic-net approach for variable selection in high-dimensional linear regression', *Science China Mathematics,* 61(1), pp. 173–188.

Huang, Y., Zaas, A. K., Rao, A., Dobigeon, N., Woolf, P. J., Veldman, T., Øien, N. C., McClain, M. T., Varkey, J. B., Nicholson, B., Carin, L., Kingsmore, S., Woods, C. W., Ginsburg, G. S. and Hero, A. O. (2011) 'Temporal Dynamics of Host Molecular Responses Differentiate Symptomatic and Asymptomatic Influenza A Infection', *PLoS Genetics*. Edited by N. J. Schork. Public Library of Science, 7(8), p. e1002234.

Hundt, C., Hildebrandt, A. and Schmidt, B. (2016) 'rapidGSEA: Speeding up gene set enrichment analysis on multi-core CPUs and CUDA-enabled GPUs', *BMC Bioinformatics*. BMC Bioinformatics, 17(1), pp. 1–11.

Ideker, T., Ozier, O., Schwikowski, B. and Siegel, A. F. (2002) 'Discovering regulatory and signalling circuits in molecular interaction networks', *Bioinformatics,* 18, pp. S233–S240.

Jha, U. K., Bajorski, P., Fokoue, E., Heuvel, J. Vanden, Aardt, J. van and Anderson, G. (2017) 'Dimensionality Reduction of High-Dimensional Highly Correlated Multivariate Grapevine Dataset', *Open Journal of Statistics,* 07(04), pp. 702–717.

Jia, P., Kao, C. F., Kuo, P. H. and Zhao, Z. (2011) 'A comprehensive network and pathway analysis of candidate genes in major depressive disorder', *BMC Systems Biology,* 5(3), pp. 1–13.

Kanehisa, M. and Goto, S. (2000) 'KEGG: kyoto encyclopedia of genes and genomes', *Nucleic acids research,* 28(1), pp. 27–30.

Kharoubi, R., Oualkacha, K. and Mkhadri, A. (2019) 'The cluster correlation-network support vector machine for high-dimensional binary classification', *Journal of Statistical Computation and Simulation*. Taylor & Francis, 89(6), pp. 1020–1043.

Khatri, P., Sirota, M. and Butte, A. J. (2012) 'Ten years of pathway analysis: current approaches and outstanding challenges', *PLoS computational biology,* 8, p. e1002375.

Khodayari Moez, E., Hajihosseini, M., Andrews, J. L. and Dinu, I. (2019) 'Longitudinal linear combination test for gene set analysis', *BMC Bioinformatics*. BMC Bioinformatics, 20(1), pp. 1–19.

Kim, S.-Y. and Volsky, D. J. (2005) 'PAGE: parametric analysis of gene set enrichment', *BMC bioinformatics,* 6, p. 144.

Kim, Y., Choi, H. and Oh, H. S. (2008) 'Smoothly clipped absolute deviation on high dimensions', *Journal of the American Statistical Association,* 103(484), pp. 1665–1673.

Kumar, S., Ingle, H., Mishra, S., Mahla, R.S., Kumar, A., Kawai, T., Akira, S., Takaoka, A., Raut, A.A. and Kumar, H. (2015) 'IPS-1 differentially induces TRAIL, BCL2, BIRC3 and PRKCE in type I interferons-dependent and-independent anticancer activity', *Cell death & disease*. 6(5), pp.e1758-e1758.

Lapuente, D., Storcksdieck Genannt Bonsmann, M., Maaske, A., Stab, V., Heinecke, V., Watzstedt, K., Heß, R., Westendorf, A. M., Bayer, W., Ehrhardt, C. and Tenbusch, M. (2018) 'IL-1β as mucosal vaccine adjuvant: the specific induction of tissue-resident memory T cells improves the heterosubtypic immunity against influenza A viruses.', *Mucosal immunology,* 11(4), pp. 1265–1278.

Lee, H. K., Braynen, W., Keshav, K. and Pavlidis, P. (2005) 'ErmineJ: tool for functional analysis of gene expression data sets', *BMC bioinformatics,* 6, p. 269.

Li, J., Jia, Y. and Zhao, Z. (2013) ‘Partly adaptive elastic net and its application to microarray classification’, *Neural Computing and Applications*, 22(6), pp. 1193–1200.

Li, W., Wang, G., Zhang, H., Zhang, D., Zeng, J., Chen, X., Xu, Y. and Li, K. (2009) ‘Differential suppressive effect of promyelocytic leukemia protein on the replication of different subtypes/strains of influenza A virus’, *Biochemical and Biophysical Research Communications*, 389(1), pp. 84–89.

Li, X., Shen, L., Shang, X. and Liu, W. (2015) ‘Subpathway analysis based on signaling- Pathway impact analysis of signaling pathway’, *PLoS ONE*, 10(7), pp. 1–19.

Liberzon, A., Subramanian, A., Pinchback, R., Thorvaldsdóttir, H., Tamayo, P. and Mesirov, J. P. (2011) ‘Molecular signatures database (MSigDB) 3.0’, *Bioinformatics*, 27(12), pp. 1739–1740.

Liiving, T., Baker, S. M. and Junker, B. H. (2011) ‘Biochemical Fundamentals’, in. Gatersleben, Germany: Springer, London, pp. 19–36.

Lu, M., Zhou, J., Naylor, C., Kirkpatrick, B. D., Haque, R., Petri, W. A. and Ma, J. Z. (2017) ‘Application of penalized linear regression methods to the selection of environmental enteropathy biomarkers’, *Biomarker Research*. Biomarker Research, 5(1), pp. 1–10.

Luo, W., Friedman, M. S., Shedden, K., Hankenson, K. D. and Woolf, P. J. (2009) ‘GAGE: generally applicable gene set enrichment for pathway analysis’, *BMC bioinformatics*, 10, p. 161.

Luo, Z., Li, Z., Chen, K., Liu, R., Li, X., Cao, H. and Zheng, S. J. (2012) ‘Engagement of heterogeneous nuclear ribonucleoprotein M with listeriolysin O induces type I interferon expression and restricts Listeria monocytogenes growth in host cells’, *Immunobiology*. Urban & Fischer, 217(10), pp. 972–981.

Ma, S. and Huang, J. (2008) ‘Penalized feature selection and classification in bioinformatics’, *Briefings in Bioinformatics*, 9(5), pp. 392–403.

Mathur, R., Rotroff, D., Ma, J., Shojaie, A. and Motsinger-Reif, A. (2018) ‘Gene set analysis methods: A systematic comparison’, *BioData Mining*. BioData Mining, 11(1), pp. 1–19.

Matsunaga, T., Ishida, T., Takekawa, M., Nishimura, S., Adachi, M. and Imai, K. (2002) ‘Analysis of Gene Expression During Maturation of Immature Dendritic Cells Derived from Peripheral Blood Monocytes’, *Scandinavian*

Journal of Immunology. John Wiley & Sons, Ltd (10.1111), 56(6), pp. 593–601.

McClain, M.T., Henao, R., Williams, J., Nicholson, B., Veldman, T., Hudson, L., Tsalik, E.L., Lambkin-Williams, R., Gilbert, A., Mann, A. and Ginsburg, G.S. (2016) 'Differential evolution of peripheral cytokine levels in symptomatic and asymptomatic responses to experimental influenza virus challenge', *Clinical & Experimental Immunology*. 183(3), pp.441-451.

Meng, Y., Cai, X. H. and Wang, L. (2018) 'Potential Genes and Pathways of Neonatal Sepsis Based on Functional Gene Set Enrichment Analyses', *Computational and Mathematical Methods in Medicine*. Hindawi, 2018, pp. 1–10.

Mohammed, A., Biegert, G., Adamec, J. and Helikar, T. (2017) 'Identification of potential tissue-specific cancer biomarkers and development of cancer versus normal genomic classifiers.', *Oncotarget*. Impact Journals, LLC, 8(49), pp. 85692–85715.

Mootha, V. K., Lindgren, C. M., Eriksson, K.-F., Subramanian, A., Sihag, S., Lehar, J., Puigserver, P., Carlsson, E., Ridderstråle, M. and Laurila, E. (2003) 'PGC-1α-responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes', *Nature genetics*, 34, p. 267.

Nam, D. and Kim, S.-Y. (2008) 'Gene-set approach for expression pattern analysis', *Briefings in bioinformatics*, 9, pp. 189–197.

Nan, X., Wang, N., Gong, P., Zhang, C., Chen, Y. and Wilkins, D. (2012) 'Biomarker discovery using 1-norm regularization for multiclass earthworm microarray gene expression data', *Neurocomputing*. Elsevier, 92, pp. 36–43.

Nishimura, D. B. (2001) 'Biotech Software & Internet Report: The Computer Software Journal for Scient 2'.

Oemar, N., Schnücker, A. and Reuvers, H. (2020) 'Model selection for Vector Autoregressive processes using the Multi-Step Elastic Net'. *Erasmus University Rotterdam.*

Offenhäuser, C., Lei, N., Roy, S., Collins, B.M., Stow, J.L. and Murray, R.Z. (2011) 'Syntaxin 11 binds Vti1b and regulates late endosome to lysosome fusion in macrophages' *Traffic*. 12(6), pp.762-773.

Ogutu, J. O., Schulz-Streeck, T. and Piepho, H. P. (2012) 'Genomic selection using regularized linear regression models: ridge regression', *In BMC proceedings. BioMed Central.*, 6(Suppl 2).

Oron, A. P., Jiang, Z. and Gentleman, R. (2008) ‘Gene set enrichment analysis using linear models and diagnostics’, *Bioinformatics*, 24(22), pp. 2586–2591.

Park, T., Yi, S.-G., Lee, S., Lee, S. Y., Yoo, D.-H., Ahn, J.-I. and Lee, Y.-S. (2003) ‘Statistical tests for identifying differentially expressed genes in time-course microarray experiments’, *Bioinformatics*, 19, pp. 694–703.

Piatetsky-Shapiro, G. and Tamayo, P. (2003) ‘Microarray data mining: facing the challenges’, *ACM SIGKDD Explorations Newsletter*, 5, pp. 1–5.

Poli, M. C., Ebstein, F., Nicholas, S. K., de Guzman, M. M., Forbes, L. R., Chinn, I. K., Mace, E. M., Vogel, T. P., Carisey, A. F., Benavides, F., Coban-Akdemir, Z. H., Gibbs, R. A., Jhangiani, S. N., Muzny, D. M., Carvalho, C. M. B., Schady, D. A., Jain, M., Rosenfeld, J. A., Emrick, L., Lewis, R. A., Lee, B., Undiagnosed Diseases Network members, U. D. N., Zieba, B. A., Küry, S., Krüger, E., Lupski, J. R., Bostwick, B. L. and Orange, J. S. (2018) ‘Heterozygous Truncating Variants in POMP Escape Nonsense-Mediated Decay and Cause a Unique Immune Dysregulatory Syndrome.’, *American journal of human genetics*. Elsevier, 102(6), pp. 1126–1142.

Qiu, X., Klebanov, L. and Yakovlev, A. (2005) ‘Correlation Between Gene Expression Levels and Limitations of the Empirical Bayes Methodology for Finding Differentially Expressed Genes’, *Statistical Applications in Genetics and Molecular Biology*. De Gruyter, 4(1).

Qiu, X., Wu, S. and Wu, H. (2015) ‘A new information criterion based on langevin mixture distribution for clustering circular data with application to time course genomic data’, *Statistica Sinica*. Institute of Statistical Science, Academia Sinica, pp. 1459–1476.

Qiu, X. and Yakovlev, A. (2006) ‘Some commnets on instability of false discovery rate stimation’, *Journal of Bioinformatics and Computational Biology*. Imperial College Press, 04(05), pp. 1057–1068.

Qiu, X. and Yakovlev, A. (2007) ‘Comments on probabilistic models behind the concepts of false discovery rate’, *Journal of Bioinformatics and Computational Biology*. Imperial College Press, 05(04), pp. 963–975.

Quackenbush, J. (2002) ‘Microarray data normalization and transformation’, *Nature genetics*, 32, p. 496.

Ramsay, J. (2005) ‘Functional data analysis’, *Encyclopedia of Statistics in Behavioral Science*.

Ramsay, J. . and Silverman, B. . (2008) ‘Functional data analysis’, *Springer Series in Statistics*.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T. and Buetow, K. H. (2008) ‘PID: the pathway interaction database’, *Nucleic acids research*, 37, pp. D674–D679.

Segal, E., Wang, H. and Koller, D. (2003) ‘Discovering molecular pathways from protein interaction and gene expression data’, *Bioinformatics*, 19, pp. i264–i272.

Shannon, J. L., Murphy, M. S., Kantheti, U., Burnett, J. M., Hahn, M. G., Dorrity, T. J., Bacas, C. J., Mattice, E. B., Corpuz, K. D. and Barker, B. R. (2018) ‘Polyglutamine binding protein 1 (PQBP1) inhibits innate immune responses to cytosolic DNA’, *Molecular Immunology*. Pergamon, 99, pp. 182–190.

Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P., Sittler, T., Karp, R. M. and Ideker, T. (2005) ‘Conserved patterns of protein interaction in multiple species’, *Proceedings of the National Academy of Sciences*, 102, pp. 1974–1979.

Storey, J. D., Xiao, W., Leek, J. T., Tompkins, R. G. and Davis, R. W. (2005) ‘Significance analysis of time course microarray experiments’, *Proceedings of the National Academy of Sciences*, 102, pp. 12837–12842.

Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., Paulovich, A., Pomeroy, S. L., Golub, T. R. and Lander, E. S. (2005) ‘Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles’, *Proceedings of the National Academy of Sciences*, 102, pp. 15545–15550.

Takamiya, M., Saigusa, K. and Dewa, K. (2021) ‘DNA microarray analysis of hypothermia-exposed murine lungs for identification of forensic biomarkers’, *Legal Medicine*. Elsevier B.V., 48(July 2019), p. 101789.

Tarca, A. L., Bhatti, G. and Romero, R. (2013) ‘A comparison of gene set analysis methods in terms of sensitivity, prioritization and specificity’, *PLoS ONE*, 8(11).

Tay, J. K., Aghaeepour, N., Hastie, T. and Tibshirani, R. (2020) ‘Feature-weightrd elastic net: using "features of features" for better prediction’, *arXiv preprint arXiv:2006.01395*

Tibshirani, R. (1996) ‘Regression Shrinkage and Selection via the Lasso’, 58(1), pp. 267–288.

Tibshirani, R., Saunders, M., Rosset, S., Zhu, J. and Knight, K. (2005) ‘Sparsity and smoothness via the fused lasso’, *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 67(1), pp. 91–108.

Tsai, C. A. and Chen, J. J. (2009) ‘Multivariate analysis of variance test for gene set analysis’, *Bioinformatics*, 25(7), pp. 897–903.

Tusher, V. G., Tibshirani, R. and Chu, G. (2001) ‘Significance analysis of microarrays applied to the ionizing radiation response’, *Proceedings of the National Academy of Sciences*, 98, pp. 5116–5121.

Wang, H., Li, R. and Tsai, C. L. (2007) ‘Tuning parameter selectors for the smoothly clipped absolute deviation method’, *Biometrika*, 94(3), pp. 553–568.

Wang, L., Chen, G. and Li, H. (2007) ‘Group SCAD regression analysis for microarray time course gene expression data’, *Bioinformatics*, 23, pp. 1486–1494.

Wang, Y., Li, J., Yan, W., Chen, Q., Jiang, Z., Zhang, R., Pan, X. and Wang, X. (2018) ‘An active component containing pterodontic acid and pterodondiol isolated from Laggera pterodonta inhibits influenza A virus infection through the TLR7/MyD88/TRAF6/NF-κB signaling pathway.’, *Molecular medicine reports*, 18(1), pp. 523–531.

Winham, S., Wang, C. and Motsinger-Reif, A. A. (2011) ‘A comparison of multifactor dimensionality reduction and L 1-penalized regression to identify gene-gene interactions in genetic association studies’, *Statistical Applications in Genetics and Molecular Biology*, 10(1).

Woods, C. W., McClain, M. T., Chen, M., Zaas, A. K., Nicholson, B. P., Varkey, J., Veldman, T., Kingsmore, S. F., Huang, Y., Lambkin-Williams, R., Gilbert, A. G., Hero, A. O., Ramsburg, E., Glickman, S., Lucas, J. E., Carin, L. and Ginsburg, G. S. (2013) ‘A Host Transcriptional Signature for Presymptomatic Detection of Infection in Humans Exposed to Influenza H1N1 or H3N2’, *PLoS ONE*. Edited by H. Tse. Public Library of Science, 8(1), p. e52198.

Wu, D. and Smyth, G. K. (2012) ‘Camera: a competitive gene set test accounting for inter-gene correlation’, *Nucleic acids research*, 40, pp. e133–e133.

Wu, S., Liu, Z.-P., Qiu, X. and Wu, H. (2014) ‘Modeling Genome-Wide Dynamic Regulatory Network in Mouse Lungs with Influenza Infection Using High-

Dimensional Ordinary Differential Equations', *PLoS ONE*. Edited by A. de la Fuente. Public Library of Science, 9(5), p. e95276.

Wu, S. and Wu, H. (2013) 'More powerful significant testing for time course gene expression data using functional principal component analysis approaches', *BMC bioinformatics*, 14, p. 6.

Xiao, N. and Xu, Q. S. (2015) 'Multi-step adaptive elastic-net: reducing false positives in high-dimensional variable selection', *Journal of Statistical Computation and Simulation*, 85(18), pp. 3755–3765.

Xie, L., Jiang, Y., Ouyang, P., Chen, J., Doan, H., Herndon, B., Sylvester, J. E., Zhang, K., Molteni, A. and Reichle, M. (2007) 'Effects of dietary calorie restriction or exercise on the PI3K and Ras signaling pathways in the skin of mice', *Journal of biological chemistry*, 282, pp. 28025–28035.

Xu, Y., Wu, W., Han, Q., Wang, Y., Li, C., Zhang, P. and Xu, H. (2019) 'Post-translational modification control of RNA-binding protein hnRNPK function.', *Open biology*. The Royal Society, 9(3), p. 180239.

Yaari, G., Bolen, C. R., Thakar, J. and Kleinstein, S. H. (2013) 'Quantitative set analysis for gene expression: a method to quantify gene set differential expression including gene-gene correlations', *Nucleic acids research*, 41, pp. e170–e170.

Yang, L., Ainali, C., Tsoka, S. and Papageorgiou, L. G. (2014) 'Pathway activity inference for multiclass disease classification through a mathematical programming optimisation framework', *BMC Bioinformatics*, 15(1), pp. 1–14.

Yoon, S., Kim, S. Y. and Nam, D. (2016) 'Improving gene-set enrichment analysis of RNA-Seq data with small replicates', *PLoS ONE*, 11(11), pp. 1–16.

Zeng, L. and Xie, J. (2014) 'Group variable selection via SCAD-L2', *Statistics*, 48(1), pp. 49–66.

Zhang, H. H., Ahn, J., Lin, X. and Park, C. (2006) 'Gene selection using support vector machines with non-convex penalty', *Bioinformatics*, 22(1), pp. 88–95.

Zhang, K., Wang, H., Bathke, A. C., Harrar, S. W., Piepho, H.-P. and Deng, Y. (2011) 'Gene set analysis for longitudinal gene expression data', *BMC bioinformatics*, 12, p. 273.

Zhang, Y., Topham, D. J., Thakar, J. and Qiu, X. (2017) 'FUNNEL-GSEA: FUNctioNal ELastic-net regression in time-course gene set enrichment analysis', *Bioinformatics*, 33(13), pp. 1944–1952.

Zhao, P. and Yu, B. (2006) ‘On model selection consistency of Lasso’, *Journal of Machine Learning Research*, (2541–2563), pp. 1–23.

Zou, H. (2006) ‘The adaptive lasso and its oracle properties’, *Journal of the American Statistical Association*, 101(476), pp. 1418–1429.

Zou, H. and Hastie, T. (2005) ‘Regularization and variable selection via the elastic net’, *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, 67(2), pp. 301–320.

Zou, H., Hastie, T. and Tibshirani, R. (2007) ‘On the “degrees of freedom” of the lasso’, *Annals of Statistics*, 35(5), pp. 2173–2192.

Zou, H. and Zhang, H. H. (2009) ‘On the adaptive elastic-net with a diverging number of parameters’, *Annals of statistics*. NIH Public Access, 37(4), pp. 1733–1751.

LIST OF PUBLICATIONS

Hasri, N. M., Wen, N. H., Howe, C. W., Mohamad, M. S., Deris, S., & Kasim, S. (2017). Improved support vector machine using multiple SVM-RFE for cancer classification. International Journal on Advanced Science, Engineering and Information Technology, 7(4-2), 1589-1594.