AN IMPROVED BICLUSTERING ALGORITHM WITH OVERLAPPING CONTROL FOR IDENTIFICATION OF INFORMATIVE GENES AND PATHWAYS

ROHANI MOHAMMAD KUSAIRI

UNIVERSITI TEKNOLOGI MALAYSIA

AN IMPROVED BICLUSTERING ALGORITHM WITH OVERLAPPING CONTROL FOR IDENTIFICATION OF INFORMATIVE GENES AND PATHWAYS

ROHANI MOHAMMAD KUSAIRI

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

SEPTEMBER 2021

ACKNOWLEDGEMENT

In preparing this thesis, I was in contact with many people, researchers, academicians, and practitioners. They have contributed towards my understanding and thoughts. In particular, I wish to express my sincere appreciation to my main thesis supervisor, TS. Dr Chan Weng Howe, for encouragement, guidance, critics and friendship. I am also very thankful to my co-supervisor TS. Dr Rohayanti Hassan for their guidance, advices and motivation. Without their continued support and interest, this thesis would not have been the same as presented here.

I also would like to express my thanks to my family for their endless care and support that have given me strength in completing this research. Besides, I am grateful to have friends that have been a great assist and always share their knowledge, experiences and suggestions to me. My special thanks to everyone in AIBIG for their helpful comments and suggestion during the weekly seminar.

ABSTRACT

Due to the rise of microarray technology, many tools and methods have been developed to analyse the huge number of gene expression data such as clustering analysis. This clustering analysis is being used for different purposes such as functional annotation, tissue classification and motif identification. Moreover, the clustering methods have made an achievement in the analysis of genetic data by clustering those genes with similar expression patterns into one cluster. Therefore, the genes with similar patterns are obtained and those genes are further analysed to extract the potential biological information. Traditional clustering methods are used to group genes that behave similarly under all conditions but are unable to perform two-dimensional grouping simultaneously. As a result, clusters obtained either contain all rows of data matrix or all columns of data matrix and thus ignoring the local co-expression effects which are present in only a subset of all biological samples. Other than that, clustering methods are unable to assign genes to multiple clusters as they do not correspond to the gene natural behaviour which has more than one function and can participate in multiple pathways. Due to limitations of traditional clustering analysis, a biclustering algorithm as a new method was introduced to identify local patterns in the data by clustering the gene dimension and condition dimension simultaneously. This local correlation information between the subset of genes and conditions is then used to improve the accuracy of clustering results. However, overlapping is another issue in biclustering. As some of the genes may belong to multiple functional categories, overlapping may be considered as one of the bicluster's behaviours but the overlapping among the bicluster need to be controlled to prevent the redundancy of the biclusters formed. This research proposed an improved overlapping control in biclustering algorithms for identification of informative genes from the gene expression data. The overlapping control is crucial in biclusters to hinder the redundancy of the biclusters produced and indirectly the number of the biclusters obtained can be reduced. Experiments were conducted on two microarray data sets (ovarian cancer dataset and glioblastoma cancer dataset). The results obtained were evaluated using 10-fold cross validation and compared with the Qualitative Biclustering Algorithm (Qubic). In addition, the results were further analysed in terms of accuracy, standard deviation, variance and t-test and the proposed method indicated a higher accuracy for Ovarian dataset (96.54%) and glioblastoma dataset (75.68%). This method showed consistent improvement in terms of accuracy of the biclusters when tested using SVM classifier over the Qualitative Biclustering Algorithm (Qubic) method. Biological context verification was then conducted to elucidate the relation of the selected genes such as ERBB2, VCAM1, CD3D and pathways (Endocytosis pathway, Bladder Cancer pathway and Pancreatic Cancer pathway) with the phenotype under study.

ABSTRAK

Berikutan peningkatan teknologi mikrotatasusunan, pelbagai alat dan kaedah telah dibagunkan untuk menganalisis sejumlah besar data ekspresi gen seperti analisis pengelompokan. Analisis ini digunakan untuk pelbagai tujuan berbeza seperti anotasi fungsian, pengklasifikasi tisu dan pengenalan motif. Selain itu, kaedah pengelompokan telah berjaya menganalisis data genetik dengan mengelompokkan gen tersebut dengan corak ekspresi gen yang serupa menjadi satu kelompok. Oleh itu, gen dengan corak yang sama diperoleh dan akan dianalisis dengan lebih lanjut untuk mengekstrak maklumat keupayaan biologi. Kaedah pengelompokan traditional digunakan untuk mengelompokkan kumpulan gen yang mempunyai kondisi yang sama dalam semua keadaan tetapi tidak berupaya untuk melaksanakan pengelompokan dua dimensi secara serentak. Hasil pengelompokan yang diperoleh mempunyai semua baris matriks data atau semua lajur matriks data sehingga mengabaikan korelasi tempatan yang hanya terdapat dalam subset semua sampel biologi. Selain itu, kaedah pengelompokan tidak berupaya untuk mengelaskan gen yang sama kepada beberapa kelompok kerana gen tersebut tidak sepadan dengan tingkah laku semula jadi yang mempunyai lebih dari satu fungsi dan dapat mengambil bahagian dalam beberapa laluan. Disebabkan limitasi kaedah pengelompokan traditional, algorithma bi-kluster diperkenalkan sebagai kaedah baru untuk mengenal pasti corak tempatan dalam data dengan mengelompokkan dimensi gen dan kondisi dimensi secara serentak. Maklumat korelasi tempatan ini antara subset gen dan kondisi kemudian digunakan untuk meningkatkan ketepatan keputusan pengelompokan. Namun, masalah dalam bi-kluster adalah pertindihan. Oleh kerana sesetengah gen mungkin tergolong dalam beberapa kategori fungsian, pertindihan boleh dianggap sebagai salah satu tingkah laku bi-kluster tetapi pertindihan di antara bi-kluster perlu dikawal bagi mengelakkan pertindihan bi-kluster yang terbentuk. Penyelidikan ini mencadangkan peningkatan kawalan pertindihan dalam algoritma bi-kluster untuk mengenal pasti gen yang bermaklumat daripada data ekspresi gen. Kawalan pertindihan sangat penting dalam bi-kluster untuk menghalang pertindihan bi-kluster yang diperoleh dan secara tidak langsung jumlah bi-kluster yang diperoleh dapat dikurangkan. Kajian telah dijalankan ke atas dua data mikrotatasusunan (data set kanser ovari dan data set kanser glioblastoma). Hasil yang diperoleh dinilai menggunakan pengesahan silang 10 kali ganda dan dibandingkan dengan Algoritma Kualitatif Bi-kluster (Qubic). Di samping itu, hasilnya dianalisis lebih lanjut dari segi ketepatan, sisihan piawai, varians dan ujian-t dan kaedah yang dicadangkan menunjukkan ketepatan yang lebih tinggi untuk set data ovari (96.54%) dan set data glioblastoma (75.68%). Kaedah ini menunjukkan peningkatan yang konsisten dari segi ketepatan bi-kluster ketika diuji menggunakan pengkelasan SVM berbanding kaedah Algoritma Kualitatif Bi-kluster (Qubic). Pengesahan konteks biologi telah dilakukan untuk menjelaskan hubungan gen terpilih seperti ERBB2, VCAM1, CD3D dan laluan (laluan Endositosis, laluan Kanser Pundi kencing dan laluan Kanser Pankreas) dengan fenotip yang sedang dikaji.

TABLE OF CONTENTS

	TITLE	PAGE
	DECLARATION	**iii**
	DEDICATION	**iv**
	ACKNOWLEDGEMENT	**v**
	ABSTRACT	**vi**
	ABSTRAK	**vii**
	TABLE OF CONTENTS	**viii**
	LIST OF TABLES	**xii**
	LIST OF FIGURES	**xiii**
	LIST OF ABBREVIATIONS	**xv**
	LIST OF APPENDICES	**xvii**
CHAPTER 1	**INTRODUCTION**	**1**
1.1	Overview	1
1.2	Problem Background	2
1.3	Problem Statement	4
1.4	Research Objectives	6
1.5	Scope of Study	7
1.6	Significance of the Study	8
1.7	Thesis Outline	9
CHAPTER 2	**LITERATURE REVIEW**	**11**
2.1	Introduction	11
2.2	Gene Expression	12
2.3	Microarray Analysis	13
2.4	Pathway-based Analysis	16
2.5	Biclustering Algorithms	21
	2.5.1 Greedy algorithms	25
	2.5.1.1 Cheng and Church's Algorithm (CCA)	25

2.5.1.2 Order-Preserving Submatrix (OPSM) 25
2.5.1.3 Gene Expression Motifs (xMOTIFs) 26
2.5.1.4 Signature Algorithm (ISA) 26
2.5.1.5 Minimum Sum-Squared Residue Coclustering (MSSRCC) 26
2.5.1.6 QUalitative BIClustering (Qubic) 26
2.5.1.7 Combinatorial Algorithm for Expression and Sequence-Based Cluster Extraction (COALESCE) 27
2.5.1.8 Correlated Pattern Biclusters (CPB) 27
2.5.1.9 Large Average Submatrices (LAS) 28
2.5.2 Divide-and-conquer algorithm 28
2.5.3 Exhaustive enumeration algorithm 29
2.5.3.1 Statistical-Algorithmic Method for Bicluster Analysis (SAMBA) 29
2.5.3.2 Bit-Pattern Biclustering Algorithm (BiBit) 29
2.5.3.3 Differentially Expressed Biclusters (DeBi) 30
2.5.4 Distribution parameter identification algorithm 30
2.5.4.1 Plaid 30
2.5.4.2 Spectral 31
2.5.4.3 Bayesian BiClustering (BBC) 31
2.5.4.4 Factor Analysis for Bicluster Acquisition (FABIA) 31
2.5.5 Comparison between Biclustering Algorithms 32
2.6 Software Implementation of Biclustering Algorithms 32
2.7 Overlapping Control in Biclustering Algorithms 35
2.8 Challenges in Biclustering Algorithms 40
2.9 Summary 42

CHAPTER 3 RESEARCH METHODOLOGY 43
3.1 Introduction 43
3.2 Research operational framework 43

3.2.1 Phase 1: Research planning and data preparation 45
3.2.2 Phase 2: Design and implement the improved Biclustering algorithm 45
3.2.3 Phase 3: Result analysis and biological validation 46
3.3 Qubic Algorithm 46
3.4 Datasets 48
3.4.1 Gene Expression Data 48
3.4.2 Pathway Data 51
3.5 Performance Measurement 52
3.5.1 P-Value of Bicluster 52
3.5.2 Classification Accuracy (SVM classifier) 53
3.5.3 Overlap Ratio of Biclusters 54
3.6 Biological Context Verification 55
3.7 Summary 56

CHAPTER 4 AN IMPROVED QUBIC WITH OVERLAPPING CONTROL ELEMENT AND EXPERIMENTAL RESULTS 57
4.1 Introduction 57
4.2 The Proposed Algorithm (Qubic-i) 57
4.3 The Workflow of Algorithm (Qubic-i) 62
4.4 Experimental Results 68
4.4.1 P-Value of Bicluster 68
4.4.2 10-Fold Cross Validation 70
4.4.3 Overlap Ratio of Biclusters 74
4.5 Biological Context Verification 74
4.5.1 Gene Enrichment Percentage 76
4.5.2 Pathway Enrichment Analysis 81
4.6 Discussion 83
4.7 Summary 84

CHAPTER 5 CONCLUSION **85**

5.1 Conclusion 85

5.2 Thesis Contribution 87

5.3 Future Works 88

REFERENCES **89**

LIST OF PUBLICATIONS **105**

LIST OF TABLES

TABLE NO.	TITLE	PAGE
Table 2.1	Comparisons of different approaches in pathway-based microarray analysis	18
Table 2.2	Biclustering Algorithms	23
Table 2.3	Existing Software for Biclustering	34
Table 2.4	Existing methods of overlapping control	37
Table 2.5	Biclustering algorithm categories summary	41
Table 3.1	Summary of the ovarian cancer and glioblastoma datasets.	49
Table 4.1	Results of comparison between Qubic and Qubic-i (ovarian dataset)	69
Table 4.2	Results of comparison between Qubic and Qubic-i (glioblastoma dataset)	69
Table 4.3	SVM classification: 10-Fold cross validation	72
Table 4.4	SVM classification: 10-Fold Cross Validation	73
Table 4.5	Comparisons of average 10-fold CV classification accuracy for top ten bicluster between without gene selection, Qubic and Qubic-i	73
Table 4.6	Biological validation of ovarian cancer dataset (Qubic-i)	79
Table 4.7	Biological validation of glioblastoma cancer dataset (Qubic-i)	80
Table 4.8	Ovarian cancer biclusters with enriched pathways	82
Table 4.9	Glioblastoma cancer biclusters with enriched pathways	82

LIST OF FIGURES

FIGURE NO.	TITLE	PAGE
Figure 1.1	Illustration of the difference between clustering and biclustering	2
Figure 2.1	The structure of chapter 2	11
Figure 2.2	Brief illustration of the gene structure	12
Figure 2.3	Transcription and translation process	13
Figure 2.4	A systematic illustration of the process flow of microarray technology	14
Figure 2.5	Single sample of raw microarray image	14
Figure 2.6	Illustration of approaches in pathway-based microarray analysis	17
Figure 2.7	Pathway-Topology based scoring methods	19
Figure 3.1	Research methodology framework	44
Figure 3.2	Overview of Qubic algorithm	47
Figure 3.3	Overall flow of Qubic algorithm	48
Figure 3.4	Example of gene expression data	51
Figure 3.5	Example of pathway data	52
Figure 3.6	Confusion matrix for binary classification	53
Figure 4.1	The overlapping of the biclusters (red box)	58
Figure 4.2	Overview of proposed algorithm with additional overlapping control elements (red dashed lines)	60
Figure 4.3	Example output of biclusters without overlapping control element	61
Figure 4.4	Example output of biclusters with overlapping control element	61
Figure 4.5	Discretization of gene expression data	62
Figure 4.6	Graph construction and seed selection	63
Figure 4.7	An initial core is formed	63
Figure 4.8	Core expanding to form pool	64

Figure 4.9	Pseudocode of overlapping control in the proposed algorithm	64
Figure 4.10	Illustration of overlapping of condition 1 (Qubic-i)	66
Figure 4.11	Illustration of overlapping of condition 2 (Qubic-i)	67
Figure 4.12	a) Overlap ratio of Qubic-i and Qubic for ovarian and b) Overlap ratio of Qubic-i and Qubic for glioblastoma dataset	74
Figure 4.13	Flow process of biological context verification	75
Figure 4.14	a) Gene enrichment percentage of biclusters in ovarian cancer dataset and b) Gene enrichment percentage of biclusters in glioblastoma cancer dataset	78

LIST OF ABBREVIATIONS

BC	-	Bicluster
BAPA-IGGFD	-	Bayesian Approach to Pathway Analysis by Intergrating Gene-Gene Functional Directions
BBC	-	Bayesian BiClustering
BPA	-	Bayesian Pathway Analysis
Bimax	-	Binary Inclusive-Maximal Biclustering Algorithm
BiBit	-	Bit-Pattern Biclustering Algorithm
CCA	-	Cheng and Church's Algorithm
CPB	-	Correlated Pattern Bicluster
DE	-	Differential Express
DEGraph	-	Differential Expression Testing for Gene Networks
DeBi	-	Differentially Expressed Bicluster
DEGs	-	Differentially expressed genes
DAG	-	Directed acyclic graph
FABIA	-	Factor Analysis for Bicluster Acquisition
FCS	-	Functional class scoring
GO	-	Gene Ontology
GSA	-	Gene set analysis
GA	-	Genetic Algorithm
GBM	-	Glioblastoma
IGA	-	Individual gene analysis
ISA	-	Iterative Signature Algorithm
KEGG	-	Kyoto Encyclopedia of Genes and Genomes
LAS	-	Large Average Submatrices
mRNA	-	Messenger RNA
mRNA	-	Messenger RNA
MSSRCC	-	Minimum Sum-Squared Residue Co-clustering
NetGSA	-	Network-Based Gene Set Analysis
NP-hard	-	Nondeterministic Polynomial time
OPSMs	-	Order-preserving submatrices

ORA	-	Over-representation analysis
PSO	-	Particle Swarm Optimization
PT	-	Pathway topology
QUBIC	-	Qualitative Biclustering Algorithm
RNA	-	Ribonucleic acid
RNA	-	Ribonucleic acid
COALESCE	-	Sequence-Based Cluster Extraction
SAMBA	-	Statistical-Algorithm Method for Bicluster Analysis
SVM	-	Support Vector Machine
TCGA	-	The Cancer Genome Atlas
TopologyGSA	-	Topology Gene Set Analysis

LIST OF APPENDICES

APPENDIX	TITLE	PAGE
Appendix A	Supplementary materials	99

CHAPTER 1

INTRODUCTION

1.1 Overview

In advancement of microarray technology, the bioinformatics field has grown rapidly with the introduction of a series of high-throughput detection techniques such as microarray experiments that can produce a huge number of valuable gene expression data (Jia et al., 2017). Gene expression data comprises the information about the gene activity, as well as the current state of the cell's activity, whether the cell is normal or cancerous. These data are generally present in the matrix form which consists of rows that represent genes while each column corresponds to sample or experimental conditions and each cell strand of the matrix shows the gene expression level of the genes corresponding to the experimental condition (Li et al., 2017). The gene expression matrix form is important for the extraction of the potential biological information that can be further analysed by biologists. This also aids in the understanding of the mechanism of the gene expression, multiple functions of the genes, and the interaction between genes that could provide the insights and information regarding the disease under study.

The gene expression data can be analysed based on two techniques which are supervised and unsupervised learning. For supervised learning, the documentations or notation of genes or sample is needed to find patterns for the clusters created while unsupervised learning, the gene expression data is analysed to find patterns that can group the genes or samples into clusters without using any type of notations (Mishra and Vipsita 2017). This research focuses on analysing the gene expression data using a clustering analysis method. In the clustering analysis method, the genes are grouped into separate clusters where each cluster consists of genes that show similar expression patterns. These clustering methods analyse the gene expression data matrix in one dimension, the resulting clusters either consist of all rows of data matrix, or all columns

of data matrix (Li et al., 2017). Nonetheless, the relevant genes obtained are not necessarily related to each condition in the column of data matrix as the gene is usually highly expressed only in a subset of conditions and the existence of any local association between genes and conditions in the gene expression data must be taken into account. Due to the drawbacks of traditional clustering methods, the biclustering algorithm is introduced to cluster the genes simultaneously from the gene dimension and condition dimension (Li et al., 2017). As a result, the cluster identifies a subset of genes and conditions and helps to improve the accuracy of the clustering results. The Illustration of clustering and biclustering as shown in Figure 1.1.

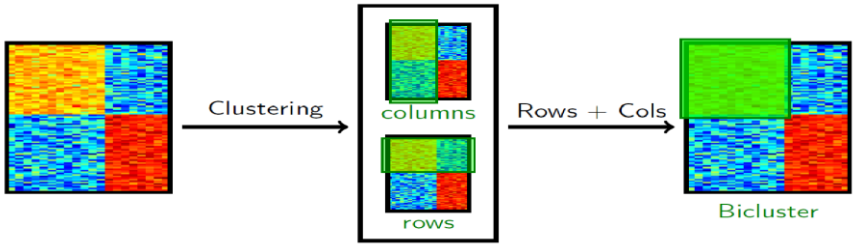

Figure 1.1 Illustration of the difference between clustering and biclustering

1.2 Problem Background

The microarray experiment has generated a massive volume of biological information in the form of gene expression data and this has driven the active developments of various data mining techniques to analyse these data, one of the common analyses is clustering (Nepomuceno et al., 2009). It is crucial to identify groups of genes with similar expression patterns under certain conditions when analyzing large-scale gene expression data (Bhattacharya and cui, 2017). This further drove the active research in using clustering methods for analysis of gene expression data. Genes with identical expression patterns are grouped together into one category using the clustering methods. The clustering is based on the principle to search for the genes with similar patterns, analyse gene function, and investigate gene transcriptional regulation (Li et al., 2017). However, genes are not necessarily related to every condition. The clustering methods can only cluster the genes in one dimension, either

the clustering results contain all rows of data matrix, or all columns of data matrix, but there is usually some local correlation between genes and conditions in gene expression data. The clustering methods group the genes based on similarities over all samples in the dataset to form clusters and this grouping strategy is ineffective at detecting condition-dependent co-expression patterns of the genes (Bhattacharya and cui, 2017). Clustering methods may not be the best suited for some analyses of microarray data for the following two reasons; firstly, there are many genes that encode proteins involved in several functional activities at a time, but the clustering methods cannot identify these genes, because it only allows a gene to belong to one cluster at a time, instead of multiple clusters. Secondly, it is difficult to find the genes that are co-expressed under a few specific conditions but are differently expressed under other conditions because the similarity of the genes in conventional clustering is determined by the entire expression data. Due to the drawbacks in clustering methods, this motivated the development of biclustering algorithms to address this problem.

The biclustering algorithm is widely employed in high-dimensional, large and complex data, especially the gene expression data that has been introduced by Cheng and Church (Zhu et al., 2017). The concept of biclustering algorithm is implemented in analyzing the gene expression data matrix by grouping the genes and condition simultaneously to form a bicluster which overcomes the drawbacks of conventional clustering algorithms. In addition, the correlation between the genes and condition are used repeatedly in biclustering algorithms in order to enhance the accuracy of clustering results (Jia et al., 2017). However, identifying biclusters is difficult due to the fact that the number of possible biclusters is proportional to the number of genes and samples. Besides that, the biclustering algorithm is required to search the sets of biclusters for functional analysis of gene expression dataset. However, extracting complete sets of biclusters from a whole microarray data matrix is a Nondeterministic Polynomial time (NP-hard) problem that requires massive computation. Therefore, in order to avoid computational issues in biclustering, most existing biclustering algorithms use a greedy iterative heuristic approach that locally improves an appropriate searching process which starts from initial seed biclusters. Thus, the greedy heuristic approach of the Qualitative Biclustering Algorithm (Qubic) is the main algorithm used throughout this research. This algorithm is able to identify statistically significant biclusters that include various gene expression data patterns as

well as finding both positively and negatively correlated expression patterns (Ayadi et al., 2012). In addition, the Qubic algorithm is able to identify all the embedded biclusters from large data sets quickly with modest computational setups. The Qubic algorithm shows good performance in analyzing gene expression data by discovering complex relationships among genes and conditions that are difficult to be detected by existing biclustering methods. However, the Qubic algorithm allows overlapping of the biclusters without explicit control of the overlapping. Overlapping control is important where too much overlapping could produce highly similar biclusters which could limit the identification of informative genes (Li et al., 2009).

1.3 Problem Statement

The main purpose of the biclustering algorithm is to group the genes according to similar expression patterns under a subset of experimental conditions. Despite the good performance of the Qubic algorithm in identifying significant biclusters, there are some limitations during the search process where the algorithm does not check on duplicate biclusters and there is no overlapping control mechanism implemented to determine or limit the amount of overlapped elements between biclusters. This could further lead to the redundancy of the biclusters obtained. The overlapping among the biclusters is crucial to measure the ability of an algorithm in ensuring the genes that are not necessarily involved in multiple biological processes are not included in biclusters. The significance of the biclusters can be determined in terms of its overlapping percentage with the previous ones by monitoring the degree of overlapping among biclusters (Pontes et al, 2008).

In general, the overlapping is formed when two biclusters share the same gene (or a group of genes) under the same experimental conditions. Therefore, searching for non-redundant overlapping biclusters is a significant problem in biclustering due to the fact that certain genes belong to different functional groups. So, the biclusters generated from a gene expression matrix can have overlapping within a predetermined threshold or else, the biclusters are considered as redundant (Truong et al., 2013). Thus, the hypothesis for this research is, “Highly overlapping genes among the

biclusters lead to duplicity of the biclusters obtained which restricts in identification of informative genes that provide more meaningful data of the genes expression patterns in the study of cancer disease".

Basically, this research intended to address aforementioned problems based on following research question:

1. How to efficiently extract informative genes from the biclusters that are related to target phenotype of study?
2. How to effectively control the overlap among the identified biclusters to prevent the redundancy of the biclusters produced?
3. How to effectively validate the identified informative biclusters and genes?

1.4 Research Objectives

The goal of this research is to improve the biclustering algorithm by controlling the overlap among biclusters in Qualitative Biclustering Algorithm (Qubic) to efficiently identify the informative genes and pathways. The specific objectives of this research stated in the following points:

(a) To develop a subroutine using Qualitative Biclustering Algorithm (Qubic) for identification of informative genes and pathways.

(b) To improve the Qualitative Biclustering Algorithm (Qubic) by introducing the overlapping control mechanism among the biclusters for more efficient identification of informative genes and pathways.

(c) To verify and validate the results and performance of Qualitative Biclustering Algorithm (Qubic-i) with the previous researchers and biological database.

1.5 Scope of Study

(a) The programming language involved in this research is R in the Windows operating system.

(b) The datasets that will be used are cancer related datasets which are ovarian cancer datasets and glioblastoma (GBM) datasets from The Cancer Genome Atlas (TCGA) project (Jin and Lee, 2014).

(c) There are 168 pathways data from the KEGG database that will be used in this research.

(d) The datasets obtained will be in text file format.

(e) Pre-processing of raw datasets will be done separately.

(f) Genecards (www.genecards.com) is used for biological validation of the selected genes in pathways.

(g) Performance measurement based on classification using support vector machine, p-value of biclusters, overlap ratio of biclusters and gene enrichment percentage.

1.6 Significance of the Study

As mentioned in previous section, the clustering methods has two drawbacks as highlighted in this research which are the clustering methods only can cluster in one dimension only, clustering results either contain all rows of data matrix, or all columns of data matrix (Li et al., 2017) and most clustering methods incapable of assigning genes to multiple clusters (Saelens et al., 2018). This has impacted the results of the genes produced in the cluster that might potentially miss local co-expression effects which are present in only a subset of all biological samples. Due to the drawbacks in clustering methods, this motivated the development of biclustering algorithms for solving this problem. The biclustering algorithm clusters gene expression data from the rows (genes) and column (conditions) of data matrix simultaneously, overcoming the drawbacks of conventional clustering approaches (Jia et al., 2017). The aim of this biclustering algorithm is grouping genes presenting similar trends under a subset of experimental conditions. The main significance of this research is to improve the biclustering algorithm by controlling the overlapping among the biclusters to prevent the redundancy of the biclusters produced. Other than that, the investigation has performed for searching the potential improvement of biclustering algorithms with overlapping control mechanisms for identification of informative genes and pathways. Indirectly, this research has given clear insight regarding the identification of informative genes and pathways by using computational approach methods and analysis which provide better understanding of the genes with the targeted phenotype and discover relevant genes that contribute to the development of cancer. Apart from that, the benefits of biclustering over clustering methods in the field of discovery of local expression patterns have been widely studied and documented till now. The finding of this research can potentially be used to aid and support discovery of pertinent biological processes involved in various regulatory mechanisms and provide meaningful data which helps in discovery of many useful drugs or even in treatment design for complex diseases.

1.7 Thesis Outline

This thesis is composed of five chapters. The general description of each chapter is presented as follows:

1. Chapter 1: This chapter presents the introduction of this research including background of the problem, problem statement, goal, objectives, scope and significance of the study.

2. Chapter 2: This chapter presents the concept and recent trends applied by previous researchers related to the research topic. Review the trend of related works regarding the algorithm used in this research.

3. Chapter 3: This chapter presents the research methodology including the research framework adopted in this study, datasets used, proposed algorithms, performance measurements and software requirements to achieve the goal and objectives.

4. Chapter 4: This chapter is composed of the result analysis and the discussion of this research.

5. Chapter 5: This chapter concludes the research study. The contribution, limitations, and future work suggestions for this research are also presented.

REFERENCES

Accolla, R. S., Ramia, E., Tedeschi, A., & Forlani, G. (2019). CIITA-driven MHC class II expressing tumor cells as antigen presenting cell performers: toward the construction of an optimal anti-tumor vaccine. Frontiers in immunology, 10, 1806.

Alavi Majd H, Shahsavari S, Baghestani AR, Tabatabaei SM, Khadem Bashi N, Rezaei Tavirani M, Hamidpour M. Evaluation of Plaid Models in Biclustering of Gene Expression Data. Scientifica. 2016;2016.

Ali, S., Shourideh, M., & Koochekpour, S. (2014). Identification of novel GRM1 mutations and single nucleotide polymorphisms in prostate cancer cell lines and tissues. PloS one, 9(7), e103204.

Ashford, A. L., Dunkley, T. P., Cockerill, M., Rowlinson, R. A., Baak, L. M., Gallo, R., ... & Cook, S. J. (2016). Identification of DYRK1B as a substrate of ERK1/2 and characterisation of the kinase activity of DYRK1B mutants from cancer and metabolic syndrome. Cellular and Molecular Life Sciences, 73(4), 883-900.

Ayadi, W., Elloumi, M., & Hao, J. K. (2009). A biclustering algorithm based on a bicluster enumeration tree: application to dna microarray data. BioData mining, 2(1), 1-16.

Barault, L., Amatu, A., Siravegna, G., Ponzetti, A., Moran, S., Cassingena, A., ... & Di Nicolantonio, F. (2018). Discovery of methylated circulating DNA biomarkers for comprehensive non-invasive monitoring of treatment response in metastatic colorectal cancer. Gut, 67(11), 1995-2005.

Belacel, N., Wang, C., & Cuperlovic-Culf, M. (2010). Clustering: Unsupervised Learning in Large Screening Biological Data.

Ben-Dor A, Chor B, Karp R, Yakhini Z. Discovering local structure in gene expression data: the order-preserving submatrix problem. Journal of computational biology. 2003 Jun 1;10(3-4):373-84.

Bergmann S, Ihmels J, Barkai N. Iterative signature algorithm for the analysis of large-scale gene expression data. Phys Rev E. 2003;67(3): 031902.

Bell D, Berchuck A, Birrer M, Chien J, Cramer D, et al. (2011) Integrated genomic analyses of ovarian carcinoma. Nature 474: 609–615. doi: 10.1038/nature10166

Bhattacharya A, Cui Y. A GPU-accelerated algorithm for biclustering analysis and detection of condition-dependent coexpression network modules. Scientific Reports. 2017 Jun 23;7(1):4162.

Bozda ˘g D, Parvin JD, Catalyurek UV. A biclustering method to discover co-regulated genes using diverse gene expression datasets. In: Bioinformatics and Computational Biology. Berlin: Springer; 2009. p. 151–63.

Chang, H. H., Cheng, Y. C., Tsai, W. C., & Chen, Y. (2020). PSMB8 inhibition decreases tumor angiogenesis in glioblastoma through vascular endothelial growth factor A reduction. Cancer science, 111(11), 4142.

Chaturvedi A, Carroll JD. An alternating combinatorial optimization approach to fitting the indclus and generalized indclus models. J Classif. 1994;11(2):155–70.

Cheng, Y., Church, G.M.: Biclustering of expression data. In: International Conference on Intelligent Systems for Molecular Biology ; Ismb International Conference on Intelligent Systems for Molecular Biology. pp. 590–602 (2000)

Cho H, Dhillon IS, Guan Y, Sra S. Minimum sum-squared residue co-clustering of geneexpression data. In: Sdm. Philadelphia: SIAM; 2004.p.3.

Cho H, Dhillon IS. Coclustering of human cancer microarrays using minimum sum-squared residue coclustering. IEEE/ACM Trans Comput Biol Bioinforma. 2008;5(3):385–400.

Chokeshaiusaha, K., Puthier, D., Nguyen, C., Sudjaidee, P., & Sananmuang, T. (2019). Factor Analysis for Bicluster Acquisition (FABIA) revealed vincristine-sensitive transcript pattern of canine transmissible venereal tumors. Heliyon, 5(5), e01558.

Cormen TH, Leiserson CE, Rivest RL, Stein C. Introduction to Algorithms.Cambridge: MIT Press; 2009.

Demetriou, P., Abu-Shah, E., Valvo, S., McCuaig, S., Mayya, V., Kvalvaag, A., ... & Dustin, M. L. (2020). A dynamic CD2-rich compartment at the outer edge of the immunological synapse boosts and integrates signals. Nature Immunology, 21(10), 1232-1243.

Divina, F., & Aguilar-Ruiz, J. S. (2006). Biclustering of expression data with evolutionary computation. IEEE transactions on knowledge and data engineering, 18(5), 590-602.

Dobosz, P., Stempor, P. A., Roszik, J., Herman, A., Layani, A., Berger, R., ... & Leibowitz-Amit, R. (2020). Checkpoint genes at the cancer side of the immunological synapse in bladder cancer. Translational oncology, 13(2), 193-200.

Feng, S. W., Chen, Y., Tsai, W. C., Chiou, H. Y. C., Wu, S. T., Huang, L. C., ... & Hueng, D. Y. (2016). Overexpression of TELO2 decreases survival in human high-grade gliomas. Oncotarget, 7(29), 46056.

Gostout, B. S., Poland, G. A., Calhoun, E. S., Sohni, Y. R., Giuntoli Ii, R. L., McGovern, R. M., ... & Persing, D. H. (2003). TAP1, TAP2, and HLA-DR2 alleles are predictors of cervical cancer risk☆. Gynecologic oncology, 88(3), 326-332.

Gu J, Liu JS. Bayesian biclustering of gene expression data. BMC Genomics. 2008;9(Suppl 1):4.

Henriques, R., & Madeira, S. C. (2018). BSig: evaluating the statistical significance of biclustering solutions. Data Mining and Knowledge Discovery, 32(1), 124-161.

Huang, Y., Sun, H., Ma, X., Zeng, Y., Pan, Y., Yu, D., ... & Xiang, Y. (2020). HLA-F-AS1/miR-330-3p/PFN1 axis promotes colorectal cancer progression. Life sciences, 254, 117180.

Huttenhower C, Mutungu KT, Indik N, Yang W, Schroeder M, Forman JJ, Troyanskaya OG, Coller HA. Detailing regulatory networks through large scale data integration. Bioinformatics. 2009;25(24):3267–74.

Hochreiter S, Bodenhofer U, Heusel M, Mayr A, Mitterecker A, Kasim A, Khamiakova T, Van Sanden S, Lin D, Talloen W, et al. Fabia: factor analysis for bicluster acquisition. Bioinformatics. 2010;26(12):1520–7.

Ihnatova, I., Popovici, V., & Budinska, E. (2018). A critical comparison of topology-based pathway analysis methods. PloS one, 13(1), e0191154.

Jia Y, Li Y, Liu W, Dong H. An Efficient Weighted Biclustering Algorithm for Gene Expression Data. InParallel and Distributed Computing, Applications and Technologies (PDCAT), 2016 17th International Conference on 2016 Dec 16 (pp. 336-341). IEEE.

Jiang, D., Xie, X., Lu, Z., Liu, L., Qu, Y., Wu, S., ... & Xu, G. (2020). Establishment of a Colorectal Cancer-Related MicroRNA-mRNA Regulatory Network by Microarray and Bioinformatics. Frontiers in Genetics, 11.

Jiang, H., Dong, L., Gong, F., Gu, Y., Zhang, H., Fan, D., & Sun, Z. (2018). Inflammatory genes are novel prognostic biomarkers for colorectal cancer. International journal of molecular medicine, 42(1), 368-380.

Jiang, T., Chen, B., Li, J., & Xu, G. (2019). Indexing and search of order-preserving submatrix for gene expression data. IEEE Access, 7, 184769-184785.

Király, A., Abonyi, J., Laiho, A., & Gyenesei, A. (2012, December). Biclustering of high-throughput gene expression data with bicluster miner. In 2012 IEEE 12th International Conference on Data Mining Workshops (pp. 131-138). IEEE.

Kléma, J., & Malinka, F. (2017). Semantic biclustering for finding local, interpretable and predictive expression patterns. BMC genomics, 18(7), 41-53.

Kluger Y, Basri R, Chang JT, Gerstein M. Spectral biclustering of microarray data: coclustering genes and conditions. Genome Res. 2003;13(4):703–16.

Lazzeroni L, Owen A, et al. Plaid models for gene expression data. Stat Sin. 2002;12(1):61–86.

Lee, C. H., Chen, L. C., Yu, C. C., Lin, W. H., Lin, V. C., Huang, C. Y., ... & Bao, B. Y. (2019). Prognostic value of CD1B in localised prostate cancer. International journal of environmental research and public health, 16(23), 4723.

Lim, S. Y., Yuzhalin, A. E., Gordon-Weeks, A. N., & Muschel, R. J. (2016). Targeting the CCL2-CCR2 signaling axis in cancer metastasis. Oncotarget, 7(19), 28697.

Lin, A., & Yan, W. H. (2018). Heterogeneity of HLA-G expression in cancers: facing the challenges. Frontiers in immunology, 9, 2164.

Lin, C. L., Ying, T. H., Yang, S. F., Wang, S. W., Cheng, S. P., Lee, J. J., & Hsieh, Y. H. (2020). Transcriptional Suppression of miR-7 by MTA2 Induces Sp1-Mediated KLK10 Expression and Metastasis of Cervical Cancer. Molecular Therapy-Nucleic Acids, 20, 699-710.

Li, H., Li, Q., Ma, Z., Zhou, Z., Fan, J., Jin, Y., ... & Liang, P. (2019). AID modulates carcinogenesis network via DNA demethylation in bladder urothelial cell carcinoma. Cell death & disease, 10(4), 1-15.

Li Y, Liu W, Jia Y, Dong H. AWeighted Mutual Information Biclustering Algorithm for Gene Expression Data. Computer Science & Information Systems. 2017 Sep 1;14(3).

Li G, Ma Q, Tang H, Paterson AH, Xu Y. Qubic: a qualitative biclustering algorithm for analyses of gene expression data. Nucleic Acids Res. 2009;37:491.

Li, Z., Chang, C., Kundu, S., & Long, Q. (2020). Bayesian generalized biclustering analysis via adaptive structured shrinkage. Biostatistics, 21(3), 610-624.

Liu, L., Wei, J., & Ruan, J. (2017). Pathway enrichment analysis with networks. Genes, 8(10), 246.

Liu, X., Li, D., Liu, J., Su, Z., & Li, G. (2020). RecBic: a fast and accurate algorithm recognizing trend-preserving biclusters. Bioinformatics, 36(20), 5054-5060.

Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: a survey. IEEE/ACM Trans Comput Biol Bioinforma. 2004;1(1):24–45.

Maleki, F., & Kusalik, A. J. (2020). Gene set overlap: an impediment to achieving high specificity in over-representation analysis. bioRxiv, 319145.

Martínez, V. G., Rubio, C., Martínez-Fernández, M., Segovia, C., López-Calderón, F., Garín, M. I., ... & Dueñas, M. (2017). BMP4 induces M2 macrophage polarization and favors tumor progression in bladder cancer. Clinical Cancer Research, 23(23), 7388-7399.

McLendon R, Friedman A, Bigner D, Van Meir E, Brat D, et al. (2008) Comprehensive genomic characterization defines human glioblastoma genes and core pathways. Nature 455: 1061–1068. doi: 10.1038/nature07385

Millstein, J., Budden, T., Goode, E. L., Anglesio, M. S., Talhouk, A., Intermaggio, M. P., ... & Rogers, P. (2020). Prognostic gene expression signature for high-grade serous ovarian cancer. Annals of Oncology, 31(9), 1240-1250.

Mishra, A., Biswal, B. S., Mohapatra, A., & Vipsita, S. (2016, July). Biclustering of gene expression patterns with an advanced overlapping control strategy. In 2016 IEEE 1st International Conference on Power Electronics, Intelligent Control and Energy Systems (ICPEICES) (pp. 1-5). IEEE.

Mishra, D. (2012). Discovery of overlapping pattern biclusters from gene expression data using hash based pso. Procedia Technology, 4, 390-394.

Mishra S, Vipsita S. Biclustering of gene expression microarray data using dynamic deme parallelized genetic algorithm (DdPGA). InComputational Intelligence in Bioinformatics and Computational Biology (CIBCB), 2017 IEEE Conference on 2017 Aug 23 (pp. 1-8). IEEE.

Miyoshi, N., Ishii, H., Mimori, K., Nishida, N., Tokuoka, M., Akita, H., ... & Mori, M. (2010). Abnormal expression of PFDN4 in colorectal cancer: a novel marker for prognosis. Annals of surgical oncology, 17(11), 3030-3036.

Mudd Jr, T. W., Lu, C., Klement, J. D., & Liu, K. (2021). MS4A1 expression and function in T cells in the colorectal cancer tumor microenvironment. Cellular Immunology, 360, 104260.

Murali T, Kasif S. Extracting conserved gene expression motifs from gene expression data. In: Pacific Symposium on Biocomputing. Stanford: Stanford Medical Informatics; 2003. p. 77–88.

Nepomuceno, J. A., Troncoso, A., & Aguilar-Ruiz, J. S. (2009, November). An Overlapping Control–Biclustering Algorithm from Gene Expression Data. In 2009 Ninth International Conference on Intelligent Systems Design and Applications (pp. 1239-1244). IEEE.

Nguyen, T. M., Shafi, A., Nguyen, T., & Draghici, S. (2019). Identifying significantly impacted pathways: a comprehensive review and assessment. Genome biology, 20(1), 1-15.

Nowak, E. M., Poczęta, M., Bieg, D., & Bednarek, I. (2017). DNA methyltransferase inhibitors influence on the DIRAS3 and STAT3 expression and in vitro migration of ovarian and breast cancer cells. Ginekologia polska, 88(10), 543-551.

Oghabian, A., Kilpinen, S., Hautaniemi, S., & Czeizler, E. (2014). Biclustering methods: biological relevance and application in gene expression analysis. PloS one, 9(3), e90801.

Padilha VA, Campello RJ. A systematic comparative evaluation of biclustering techniques. BMC bioinformatics. 2017 Dec;18(1):55.

Pontes, B., Divina, F., Giráldez, R., & Aguilar-Ruiz, J. S. (2009). Improved biclustering on expression data through overlapping control. International Journal of Intelligent Computing and Cybernetics.

Pontes, B., Giráldez, R., & Aguilar-Ruiz, J. S. (2015). Biclustering on expression data: A review. Journal of biomedical informatics, 57, 163-180.

Preli´c A, Bleuler S, Zimmermann P, Wille A, Bühlmann P, Gruissem W, Hennig L, Thiele L, Zitzler E. A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics. 2006;22(9):1122–9.

Ready, D., Yagiz, K., Amin, P., Yildiz, Y., Funari, V., Bozdag, S., & Cinar, B. (2017). Mapping the STK4/Hippo signaling network in prostate cancer cell. PLoS One, 12(9), e0184590.

Rodriguez-Baena DS, Perez-Pulido AJ, Aguilar JS. A biclustering algorithm for extracting bit-patterns from binary datasets. Bioinformatics. 2011;27(19):2738–45.

Saelens, W., Cannoodt, R., & Saeys, Y. (2018). A comprehensive evaluation of module detection methods for gene expression data. Nature communications, 9(1), 1-12.

Saikia, M., Bhattacharyya, D. K., & Kalita, J. K. (2021). BicGenesis: A Method to Identify ESCC Biomarkers Using the Biclustering Approach. In Proceedings of International Conference on Big Data, Machine Learning and Applications: BigDML 2019 (Vol. 180, p. 1). Springer Nature.

Salem, A., Alotaibi, M., Mroueh, R., Basheer, H. A., & Afarinkia, K. (2020). CCR7 as a therapeutic target in Cancer. Biochimica et Biophysica Acta (BBA)-Reviews on Cancer, 188499.

Savage, S. R., Shi, Z., Liao, Y., & Zhang, B. (2019). Graph algorithms for condensing and consolidating gene set analysis results. Molecular & Cellular Proteomics, 18(8), S141-S152.

Schaaf, H. S., Collins, A., Bekker, A., & Davies, P. D. (2010). Tuberculosis at extremes of age. Respirology, 15(5), 747-763.

Serin A, Vingron M. Debi: Discovering differentially expressed biclusters using a frequent itemset approach. Algorithms for Molecular Biology. 2011 Dec;6(1):18.

Shabalin AA, Weigman VJ, Perou CM, Nobel AB. Finding large average submatrices in high dimensional data. Annals Appl Stat. 2009;3: 985–1012.

Shi, W., Chen, Z., Li, L., Liu, H., Zhang, R., Cheng, Q., ... & Wu, L. (2019). Unravel the molecular mechanism of XBP1 in regulating the biology of cancer cells. Journal of Cancer, 10(9), 2035.

Simões, I. T., Aranda, F., Casadó-Llombart, S., Velasco-de Andrés, M., Català, C., Álvarez, P., ... & Lozano, F. (2020). Multifaceted effects of soluble human CD6 in experimental cancer models. Journal for immunotherapy of cancer, 8(1).

Singh, M., & Mehrotra, M. (2018). Impact of biclustering on the performance of biclustering based collaborative filtering. Expert Systems With Applications, 113, 443-456.

Sutheeworapong, S., Ota, M., Ohta, H., & Kinoshita, K. (2012). A novel biclustering approach with iterative optimization to analyze gene expression data. Advances and applications in bioinformatics and chemistry: AABC, 5, 23.

Tanay A, Sharan R, Shamir R. Discovering statistically significant biclusters in gene expression data. Bioinformatics. 2002;18(suppl 1):136–44.

Truong, D. T., Battiti, R., & Brunato, M. (2013, December). Discovering non-redundant overlapping biclusters on gene expression data. In 2013 IEEE 13th International Conference on Data Mining (pp. 747-756). IEEE.

Wu, A., Zhang, S., Liu, J., Huang, Y., Deng, W., Shu, G., & Yin, G. (2020). Integrated Analysis of Prognostic and Immune Associated Integrin Family in Ovarian Cancer. Frontiers in genetics, 11, 705.

Xie, J., Ma, A., Zhang, Y., Liu, B., Wan, C., Cao, S., ... & Ma, Q. (2018). QUBIC2: A novel biclustering algorithm for large-scale bulk RNA-sequencing and single-cell RNA-sequencing data analysis. bioRxiv, 409961.

Xie, J., Ma, A., Fennell, A., Ma, Q., & Zhao, J. (2019). It is time to apply biclustering: a comprehensive review of biclustering applications in biological and biomedical data. Briefings in bioinformatics, 20(4), 1450-1465.

Yang, Yi, Yiwen Zang, Chuang Zheng, Zhenyang Li, Xiaodong Gu, Minwei Zhou, Zihao Wang, Jianbin Xiang, Zongyou Chen, and Yiming Zhou. "CD3D is associated with immune checkpoints and predicts favorable clinical outcome in colon cancer." Immunotherapy 12, no. 1 (2020): 25-35.

Yang, Q., Wang, S., Dai, E., Zhou, S., Liu, D., Liu, H., ... & Jiang, W. (2019). Pathway enrichment analysis approach based on topological structure and updated annotation of pathway. Briefings in bioinformatics, 20(1), 168-177.

Ye, W., Zhou, Y., Xu, B., Zhu, D., Rui, X., Xu, M., ... & Jiang, J. (2019). CD247 expression is associated with differentiation and classification in ovarian cancer. Medicine, 98(51).

Yun, T., & Yi, G. S. (2013). Biclustering for the comprehensive search of correlated gene expression patterns using clustered seed expansion. BMC genomics, 14(1), 1-15.

Zhan, S., Li, J., Wang, T., & Ge, W. (2018). Quantitative proteomics analysis of sporadic medullary thyroid cancer reveals FN1 as a potential novel candidate prognostic biomarker. The oncologist, 23(12), 1415.

Zhang, C., Berndt-Paetz, M., & Neuhaus, J. (2020). Identification of key biomarkers in bladder cancer: Evidence from a bioinformatics analysis. Diagnostics, 10(2), 66.

Zhang, L., Chen, S., Zhu, J., Meng, J., & Liu, H. (2020). REW-ISA: unveiling local functional blocks in epi-transcriptome profiling data via an RNA expression-weighted iterative signature algorithm. BMC bioinformatics, 21(1), 1-22.

Zhang, R., Qi, F., Zhao, F., Li, G., Shao, S., Zhang, X., ... & Feng, Y. (2019). Cancer-associated fibroblasts enhance tumor-associated macrophages enrichment and suppress NK cells function in colorectal cancer. Cell death & disease, 10(4), 1-14.

Zhang, W., Ou, J., Lei, F., Hou, T., Wu, S., Niu, C., ... & Zhang, Y. (2016). C14ORF166 overexpression is associated with pelvic lymph node metastasis and poor prognosis in uterine cervical cancer. Tumor Biology, 37(1), 369-379.

Zhang, H., Shao, Y., Chen, W., & Chen, X. (2021). Identifying Mitochondrial-Related Genes NDUFA10 and NDUFV2 as Prognostic Markers for Prostate Cancer through Biclustering. BioMed research international, 2021.

Zhao, H., Wee-Chung Liew, A., Z Wang, D., & Yan, H. (2012). Biclustering analysis for pattern discovery: current techniques, comparative studies and applications. Current Bioinformatics, 7(1), 43-55.

Zhao, Y., Piekos, S., Hoang, T. H., & Shin, D. G. (2020). A framework using topological pathways for deeper analysis of transcriptome data. BMC genomics, 21(1), 1-11.

Zhao, Z., Zhang, G., & Li, W. (2017). Elevated expression of ERCC6 confers resistance to 5-fluorouracil and is associated with poor patient survival in colorectal cancer. DNA and cell biology, 36(9), 781-786.

Zhu X, Qiu J, Xie M, Wang J. A multi-objective biclustering algorithm based on fuzzy mathematics. Neurocomputing. 2017 Aug 30;253:177-82.

Zhu, C. C., Chen, C., Xu, Z. Q., Zhao, J. K., Ou, B. C., Sun, J., ... & Lu, A. G. (2018). CCR6 promotes tumor angiogenesis via the AKT/NF-κB/VEGF pathway in colorectal cancer. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 1864(2), 387-397.

LIST OF PUBLICATIONS

Kusairi, R. M., Moorthy, K., Haron, H., Mohamad, M. S., Napis, S., & Kasim, S. (2017). An Improved Parallelized mRMR for Gene Subset Selection in Cancer Classification. *International Journal on Advanced Science, Engineering and Information Technology*, *7*(4-2), 1595-1600.

Kusairi, R. M., & Chan, W. H. A (2018). Review on Computational Approaches of Biclustering Algorithms for Biological Data Analysis. Postgraduate Annual Research Symposium (PARS), 2018.