Identifying a Novel Disease Genes Association using Tissue Specified Gene-Gene Network (TSG)

Hind Ibrahim Hussein Abdalateef

B.Sc. (Honours) in Statistics\Computer, University of Gezira (2010)

A Dissertation
Submitted to University of Gezira in Partial Fulfilment of the Requirements for the Degree of Master of Science in Computer Science

Department of Computer Science
Faculty of Mathematical and Computer Sciences

September, 2017
Identifying a Novel Disease Genes Association using Tissue Specified Gene-Gene Network (TSG)

Hind Ibrahim Hussein AbdAlateef

Supervision Committee:
Name Position Signature
Dr. Murtada Khalfallah ElbashirMain Supervisor ............... 
Dr. Mohammed Hamad AhmedElhibr Co-supervisor ............... 

Date: September, 2017
Identifying a Novel Disease Genes Association using Tissue Specified Gene-Gene Network (TSG)

Hind Ibrahim Hussein Abdalateef

Examination Committee:
Name  Position  Signature
Dr. Mohammed Hamed Ahmed Elhebir Co-supervisor  ..............
Dr. Abdalla AkodeOsman mohammed  External Examiner  ........
Dr. Mohammed Albarra  Internal Examiner  ............

Date: September, 2017
Identifying a Novel Disease Genes Association using Tissue Specified Gene-Gene Network (TSG)

ABSTRACT

Bioinformatics is an application of computational technology that handles many problems related to molecular biology, which is useful for managing and analysing large datasets, which grows a number of areas, such as gene sequencing and gene expression studies. The prediction of disease genes considered as fundamental task in bioinformatics that can be solved by network based approaches, which is represented in a form of protein-protein interaction network (PPIN) or gene-gene interaction network (GN). In the prediction field there are many studies, one of these studies presents a prediction that uses annotation-based methods, which presents disease association method based on the underlying biological mechanisms to predict disease association. Another presents network based approaches for the prediction represented in a novel method named node removal (NR) to constructing tissue-specific protein-protein interaction networks. The difficulty of identifying the genes causing disease health because of weakness and less efficiency of traditional association approaches in identifying associated genes for diseases. The research aims to discover a new disease gene association for both Alzheimer disease and colorectal cancer, using (DisGeNET) database to determine gene disease association and the Genotype-Tissue Expression (GTEx) to illustrate the most affected tissues by Alzheimer disease and colorectal cancer. A Pearson correlation coefficient (PCC) is calculated for genes to construct weighted tissue specified gene-gene network (TSG), and Support Vector Machine (SVM) is used as a validity predictor for genes-database and for predicted genes. As a result, the accuracy of Alzheimer disease is 70.7143% for all genes and 66.0583% for predicted genes, for
Colorectal cancer the accuracy is 77.6923% for all genes and 67.2063% for predicted genes. Hence SVM indicates superior performance in genes prediction
تحديد الجينات المسببة للأمراض بواسطة شبكة جينية محددة الأنسجة

هند إبراهيم حسين

ملخص البحث

المعلوماتية الحيوية هو تطبيق التكنولوجيا الحاسوبية للتعامل مع معظم المشاكل المتعلقة بالبيولوجيا الجزيئية، وهو مفيد لإدارة وتحليل مجموعات كبيرة من البيانات، كما أنه يمكن تطبيقاته في الكثير من المجالات، مثل تسلسل الجينات ودراسات التعبير الجيني. يعتبر التنبيه بجينات المرض مهمة أساسية في المعلوماتية الحيوية التي يمكن أن تенного على أساس الشبكة، والتمثيل الشبكي يكون في شكل شبكة تفاعل بروتين (PPIN) أو في شكل شبكة تفاعل جين (GGIN). في مجال التنبيه بجينات المرض يوجد العديد من الدراسات المساعدة واحدة من هذه الدراسات استخدم منهجية قائمة على الشروح، حيث اكتشفت طريقة جديدة لارتباط الأمراض اعتماداً على آليات احيائية لتتبع النمو، هناك دراسة أخرى تقرر التنبيه باستخدام منهجية الشبكات وتمثل في بناء طريقة تقوم بناء شبكة تفاعل البروتينات. صعوبة تحديد الجينات المسببة لمرض معين بسبب ضعف وقلة كفاءة النهج التقليدي في تحديد الجينات المرتبطة بالأمراض. الهدف من هذا البحث هو التنبيه بالجينات المسببة لمرض معين على أساس الشبكة الجينية محددة الأنسجة (TSG). قاعدة البيانات (DisGeNET) استخدمت لتحديد ارتباط الجين بمرض معين. الأداة (GTEx) استخدمت لتحديد ارتباط الجين بمرض معين. الدراسة لتحديد شبكات التغيرات لجميع الجينات (PCC) استخدمت وطبقت للتأكد من الأنسجة الأكثر تأثيراً بالمرض. تم حساب قيم فاصلة عصبية (SVM) لبناء شبكة جينية محددة الأنسجة. كما أن خوارزمية آلة التمثيل الداعم استخدمت وطبقت للتأكد من صحة عملية التنبيه. من حساب ومقارنة مقاييس الأداء المثلى (الدقة, الحساسية, الحدودية, منحنى روك, المنطقة تحت منحنى روك) لتقدير أفضل التنبيه في هذه الدراسة تبين أن التنبيه بالجينات المسببة لمرض باستخدام شبكة الجينات مع خوارزمية آلة التمثيل الداعم (SVM) في مرض الزهايمر منح دقة (70.143%) لجميع الجينات و66.0583% للجينات المكتشفة، في سرطان القولون والمستقيم منح دقة (77.6923%) لجميع الجينات و67.2063% لجينات المشتبه بهم. خوارزمية آلة التمثيل الداعم أظهرت أفضل آداء لتنبيه الجينات.
TABLE OF CONTENTS

1.1 Introduction .......................................................................................................... 1
1.2 Research Problem .................................................................................................. 2
1.3 Research Objectives .............................................................................................. 3
1.4 Research organization ......................................................................................... 4
2.1 Background ............................................................................................................ 5
2.2 Literature Review ................................................................................................ 6
3.1 Introduction ........................................................................................................... 7
3.2 Data set features ................................................................................................... 8
3.3 Tissue affected by the disease ............................................................................ 9
3.4 Calculate Pearson correlation coefficient (PCC) .............................................. 10
3.5 Tissue specified gene gene network (TSG) ...................................................... 11
3.6 weighted tissue specified gene gene network (TSG) ........................................ 12
3.7 Support vector machines(SVM) ......................................................................... 13
4.1 Introduction .......................................................................................................... 14
4.2 Results .................................................................................................................. 15
5.1 Conclusions ......................................................................................................... 16
5.2 Recommendations ............................................................................................... 17
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table No.</th>
<th>Page No</th>
</tr>
</thead>
<tbody>
<tr>
<td>Table 4-1: Performance measures of alzheimer's disease</td>
<td>31</td>
</tr>
<tr>
<td>Table 4-2: Performance measures of colorectal cancer</td>
<td>34</td>
</tr>
<tr>
<td>Table 4-3: The Area Under the Roc Curve (AUC)</td>
<td>36</td>
</tr>
</tbody>
</table>
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure No</th>
<th>Page No.</th>
</tr>
</thead>
<tbody>
<tr>
<td>2-1</td>
<td>11</td>
</tr>
<tr>
<td>3-1</td>
<td>18</td>
</tr>
<tr>
<td>3-2</td>
<td>19</td>
</tr>
<tr>
<td>3-3</td>
<td>20</td>
</tr>
<tr>
<td>3-4</td>
<td>23</td>
</tr>
<tr>
<td>3-5</td>
<td>24</td>
</tr>
<tr>
<td>3-6</td>
<td>25</td>
</tr>
<tr>
<td>4-1</td>
<td>29</td>
</tr>
<tr>
<td>4-2</td>
<td>32</td>
</tr>
<tr>
<td>4-3</td>
<td>32</td>
</tr>
<tr>
<td>4-4</td>
<td>35</td>
</tr>
<tr>
<td>4-5</td>
<td>40</td>
</tr>
<tr>
<td>4-6</td>
<td>34</td>
</tr>
<tr>
<td>4-7</td>
<td>35</td>
</tr>
</tbody>
</table>
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GTEx</td>
<td>Genotype-Tissue Expression</td>
</tr>
<tr>
<td>TSG</td>
<td>Tissue-specified gene-gene network</td>
</tr>
<tr>
<td>PCC</td>
<td>Pearson correlation coefficient</td>
</tr>
<tr>
<td>DNA</td>
<td>Deoxyribonucleic acid</td>
</tr>
<tr>
<td>OMIM</td>
<td>Online Mendelian Inheritance in Man</td>
</tr>
<tr>
<td>PPIN</td>
<td>protein-protein interaction network</td>
</tr>
<tr>
<td>NCBI</td>
<td>National Centre for Biotechnology Information</td>
</tr>
<tr>
<td>HPRD</td>
<td>Gene Cards Human protein reference database</td>
</tr>
<tr>
<td>CV</td>
<td>Cross validation</td>
</tr>
<tr>
<td>GEO</td>
<td>Gene expression omnibus</td>
</tr>
<tr>
<td>GNs</td>
<td>Gene networks</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristics</td>
</tr>
<tr>
<td>AUR</td>
<td>Area under roc curve</td>
</tr>
<tr>
<td>TP</td>
<td>True positive</td>
</tr>
<tr>
<td>TN</td>
<td>True negative</td>
</tr>
<tr>
<td>FP</td>
<td>False positive</td>
</tr>
<tr>
<td>FN</td>
<td>False negative</td>
</tr>
</tbody>
</table>
CHAPTER ONE

1.1 INTRODUCTION

Bioinformatics is an interdisciplinary field that develops and applies computational methods to analyse large collections of biological data, such as genetic sequences, cell populations or protein samples, to make new predictions or discover new biology. Bioinformatics is an analysis of biological information by computer programs and statistics. One of the emerging tasks in Bioinformatics is discovering the human genes that cause disease. (Nguyen, Thanh-Phuong and Ho, Tu-Bao, 2012).

Disease Gene Identification is a process which scientists identify the mutant genotypes responsible for an inherited genetic disorder [(Li, Jin Billy and Levanon, Erez Y and Yoon, Jung-Ki and Aach, John and Xie, Bin and LeProust, Emily and Zhang, Kun and Gao, Yuan and Church, George M, 2009]. Disease gene prediction, the task of identifying the most possible candidate disease genes, it is an important issue in computational biology and biomedical research. The predictions of disease-associated genes are usually approached by three main directions: (1) functional annotation based; (2) machine learning based; and (3) network based. (Le, Duc-Hau and Dang, Vu-Tung, 2016). The prediction of disease genes based on the network, this direction based on biological networks, which are constructed based on various kinds of biomedical data, and therefore they are not controlled by the coverage of functional annotation data sources.

Genes network interaction resulting from the impact of a single gene for a particular disease by modifying one gene or another of several other genes; this indicates that the gene interactions are under a biological basis. Representation of the network will be located in the form of mathematical graph including nodes and edges.
Genes represented as nodes where an interaction occurs between two nodes connected by an edge. The path between two nodes is a sequence of adjacent nodes.

In this research, we have a Tissue specified gene-gene network, which is represents the interaction of genes with each other based on the tissues most affected by the genes of a particular disease and based on the Pearson correlation coefficient (PCC) value. Machine learning is a method used to devise complex models and algorithms that lend themselves to prediction [(Alpaydin, 2014)]. Machine learning techniques have been successfully applied to solve various important biomedical problems, such as genome annotation, pattern recognition, classification of microarray data, inference of gene regulatory networks and discovery of gene-gene interaction in disease data.

Support vector machine (SVM) is a classification method that samples hyper planes which separate between two or multiple classes. SVM is an abstract learning machine, which will learn from a training data set and attempt to generalize and make correct predictions about novel data (Jagadeesh, B and Kumar, P Rajesh and Reddy, P Chenna, 2013). Most problems in machine learning are binary classification.

In this research (DisGeNET) database is used to determine gene disease association and then use the Genotype-Tissue Expression (GTEx) software to show the most affected tissues according to a gene that associated with certain disease. Calculate Pearson correlation co-efficient (Jagadeesh, B and Kumar, P Rajesh and Reddy, P Chenna, 2013) That is utilized to construct, tissue specified gene network (TSG), and then network weighted to predict novel disease genes. Support vector machine is applied (SVM) on the total data set, utilized two instances (classes) and used cross validation to divide the data to training and testing data. A variety of performance measures are used in the study accuracy, sensitivity, specificity and ROC analysis.
1.2 RESEARCH PROBLEM
Identifying genes causing disease is a key challenge in human health. Because of weakness and less efficiency of traditional association approaches in identifying associated genes for diseases, there is a need for efficient techniques and strong tools for predicting disease genes to achieve high prediction performance and increase the current knowledge of the disease gene relationship to improve disease diagnoses and treatment.

1.3 RESEARCH OBJECTIVES
Network techniques help in predicting the genetic diseases at the time of identifying genes causing the disease, thus reducing the potential risk of diseases.

The research objectives can be summaries in the following points:

1- Construct tissue-specified gene-gene network (TSG).

2- Predict new disease-gene associations for a particular disease.

3- Measure the effectiveness of tissue-specified gene-gene network (TSG) in terms of accuracy.

1.4 RESEARCH ORGANIZATION
Chapter one: provides the following statements such as introduction, problem statement, research objectives and research organization. Chapter two: this chapter includes the following background and literature review. Chapter three: research methodology, include architecture explain the steps applied in research. Chapter four: discusses the result of the research. Finally, Chapter five: presents conclusions and recommendations.
CHAPTER TWO

BACKGROUND AND LITERATURE REVIEW

2.1. BACKGROUND:

This chapter refers to theory background for many concepts of this research, such as tissues, DNA, Association, biological networks, bioinformatics, machine learning and support vector machine (SVM).

The human body is composed of a group of cells, tissues and organs. The cell is the fundamental unit of both structure and function in a living being is the smallest unit capable of carrying out the processes associated with life. Cells of similar structure and specialized function combine to form tissue. (Sherwood, 2015)

2.1.1 Tissues

Tissues are groups of cells of similar specialization. There are four primary types, muscle, nervous, epithelial and connective. Each tissue consists of cells of a single specialized type, along with varying amount of extracellular material. The tissue differs from each other in the following: The size of tissue cells, Forms of tissue cells, Order tissue cells, quantity of interstitial (between cellular) and function tissue. (Sherwood, 2015)

2.1.2 Deoxyribonucleic acid (DNA)

Is a molecule that carries the genetic instructions used in the growth, development, functioning and reproduction of all known living organisms and many viruses (Insan, Nitin Goel and Mane, Vijay and Chaudhary, BL and Danu, Mahesh Singh and Yadav, Amod and Srivastava, Vive, 2013). These instructions are available within each cell as it passes from parents to children. DNA consists of those molecules that
are called nucleotide which in turn is made up of a group of phosphate, a sugar group and nitrogen units. Nitrogen units have 4 types which are arranged according to the instructions of the DNA or genetic code; they are Adenine (A), Thymine (T), Guanine (G) and Cytosine (C). DNA has basic properties:

- **Replication**: Is the presence of molecules and hereditary process can be copied to the life cycle and continuity from one generation to another generation.
- **Generation of form**: structures that can represent an organism which can be described as a form or substance.
- **Mutation**: It is the change that can happen to gene and change it from one form to another and it happens rarely.
- **Gene expression**: is the ability of genes to express themselves and the process by which the DNA sequence of the series gives the structure and function of the cell. The thousands of genes expressed in particular cell determine what that cell can do. [Insan, Nitin Goel and Mane, Vijay and Chaudhary, BL and Danu, Mahesh Singh and Yadav, Amod and Srivastava, Vive, 2013]

2.1.3 Association

Returns to studies that examine the relationship between genetic information and the diseases. Identifying associations of genes with diseases has long been an important goal in computational biology. This association can in practice be derived from many different types of data. (Singh-Blom, U Martin and Natarajan, Nagarajan and Tewari, Ambuj and Woods, John O and Dhillon, Inderjit S and Marcotte, Edward M, 2013)

2.1.3.1 GENE-DISEASE ASSOCIATION

Gene-disease is a major challenge in human health with method to understand disease mechanisms, diagnosis and treatment [(Insan, Nitin Goel and Mane, Vijay and Chaudhary, BL and Danu, Mahesh Singh and Yadav, Amod and Srivastava, Vive, 2013)]. Many similar previous studies are used to conclude genomic periods that associated with disease. One of the most well-known databases that store gene-
disease association is the **Online Mendelian In Man** (OMIM) which provides publication about gene disease relationships. [(Hamosh, Ada and Scott, Alan F and Amberger, Joanna S and Bocchini, Carol A and McKusick, Victor A, 2005)](Hamosh, Ada and Scott, Alan F and Amberger, Joanna S and Bocchini, Carol A and McKusick, Victor A, 2005)\] In computational biology, gene prediction or gene finding refers to the process of identifying the regions of genomic DNA that encode genes. Gene finding is one of the first and most important steps in understanding the genome of a species once it has been sequenced [(Hamosh, Ada and Scott, Alan F and Amberger, Joanna S and Bocchini, Carol A and McKusick, Victor A, 2005)](Hamosh, Ada and Scott, Alan F and Amberger, Joanna S and Bocchini, Carol A and McKusick, Victor A, 2005).

The network based approaches become even more attractive in the research field of disease-causing gene. The most important benefit from the study of genetics of human disease is to predict the risk that the human may have lack to particular disease [(Wray, Naomi R and Goddard, Michael E and Visscher, Peter M, 2007). Knowledge of risk can be used by the clinician for treatment strategies. Also, determining gene-disease association can enhance the development of new techniques for prevention, diagnosis and treatment of disease [((Arzucan and Vu, Thuy and Erkan, G"u"ne and Radev, Dragomir R, 2008)). On the other hand, prediction of new disease genes resulted from the study of human disease might help in assumptions regarding the presence of disease on neighbouring network, which might be caused the same disease genes. This approach will help in the prediction of genetic heterogeneous diseases at the time of identifying genes causing the disease. Identification of particular disease causing genes requires a lot of experience on a large number of candidate genes. Bioinformatics develop computational disease gene prediction method. The various methods for identifying my candidate disease genes in humans cover different concepts, as, functional and literature data, gene-specific characteristics, anatomy-based gene/protein expression data or phenotype comparison analyses.((Van Driel, Marc A and Brunner, Han G, 2006)
2.1.4 Biological networks

Most of the vital processes of the organism, or the so-called processes, reside in the form of networks such as metabolic pathways and protein-protein interaction network (PPIN), which represent how the physical interaction between the proteins and the activity protein with another. Understanding biological networks, analysis, modelling and visualization have become an essential task in the field of sciences life [Franklin, 2013]. As it became the beginning point to understand human disease. Increase topics related to networks in the field of research has increased the importance of dealing with biological networks, although the networks are still incomplete because of the complexities of some datathatrequired network. Prioritizing genes and computational approaches depend on comparing candidate gene to other genes with the same disease. In genetics, gene-gene interaction resulting from the impact of a single gene for a particular disease by modifying one gene or another of several other genes, this indicates that the gene interactions have a biological basis. Gene interaction is a common part of the genetic architecture of complex diseases [Gui, Jiang and Moore, Jason H and Williams, Scott M and Andrews, Peter and Hildege, Hans L and van der Harst, Pim and Navis, Gerjan and Van Gilst, Wiek H and Asselbergs, Folkert W and Gilbert-Diamond, Diane, 2013)]. Simple representation of the network will be in the form of mathematical graph including nodes and edges. Nodes (or vertices) often represent proteins, genes, or metabolites while edges often represent relationships, such as physical interactions or gene expression regulation. (Su, Gang and Morris, John H and Demchak, Barry and Bader, Gary D, 2015)
2.1.5 Bioinformatics

Conceptualizing biology in terms of molecules and applying "informatics techniques" to understand and organize the information associated with these molecules, on a large scale. In short, bioinformatics is a management information system for molecular biology and has many practical applications[(Wheeler, David L and Barrett, Tanya and Benson, Dennis A and Bryant, Stephen H and Canese, Kathi and Chetvernin, Vyacheslav and Church, Deanna M and DiCuccio, Michael and Edgar, Ron and Federhen, Scott and others, 2007)]. It is an analysis of biological information by computer programs and statistical and bioinformatics is trying to use the algorithms and databases to expand biological search. National Centre for Biotechnology Information (NCBI) [(Wheeler, David L and Barrett, Tanya and Benson, Dennis A and Bryant, Stephen H and Canese, Kathi and Chetvernin, Vyacheslav and Church, Deanna M and DiCuccio, Michael and Edgar, Ron and Federhen, Scott and others, 2007)] defines Bioinformatics It's a field of science in which biology and technology of informatics and computer science merged together in a scientific field.

2.1.5.1 GOALS OF BIOINFORMATICS

- The development of efficient algorithms for measuring sequence similarity, Analysis and modelling of different types of data that include amino acids.
- Another goal of Bioinformatics is the extension of experimental data by predictions. A fundamental goal of computational biology is the prediction of protein structure from an amino acid sequence. The use of Bioinformatics in the human genome research broadly within the human Genome Project, which identified genetic series full of human, which consists of about three billion pair.
- Bioinformatics is also used to predict interactions between proteins, given individual structures of the partners. This is known as the “docking problem.”
Computer programs simulate these interactions to predict the optimal spatial relationship between binding partners.

- Bioinformatics aims to organizing and arranging the data easy way allow researchers access to the information required easily.[Whisstock, James C and Lesk, Arthur M, 2003]

Bioinformatics study helps in the discovery of solutions to many of the existing problems, which increases individual problem-solving skills. Also provide opportunities for individuals to work with the community and gain experience and how to cooperate. Bioinformatics comprehensive with many databases such as: UCSC Genome browser[(Karolchik, Donna and Baertsch, Robert and Diekhans, Mark and Furey, Terrence S and Hinrichs, Angie and Lu, YT and Roskin, Krishna M and Schwartz, Matthias and Sugnet, Charles W and Thomas, Daryl J and others, 2003)], Ensemble Genome browser[(Kent, W James and Sugnet, Charles W and Furey, Terrence S and Roskin, Krishna M and Pringle, Tom H and Zahler, Alan M and Haussler, David, 2002)], National centre for Biotechnology information (NCBI), Gene Cards Human protein reference database (HPRD)[(Peri, Suraj and Navarro, J Daniel and Kristiansen, Troels Z and Amanchy, Ramars and Surendranath, Vineeth and Muthusamy, Babylakshmi and Gandhi, TKB and Chandrika, KN and Deshpande, Nandan and Suresh, Shubha and others, n.d.]) and online Mendelian Inheritance in Man (OMIM).
2.1.6 Machine learning

Machine learning is a set of methods that can automatically detect patterns in data and then use the uncovered patterns to predict future data, or to perform other kinds of decision making under uncertainty. (Murphy, 2012). Learning machine which is divided into two types as follows:

2.1.6.1 Supervised learning

Learning machine tasks that inferring a function from supervised training data. The training data consists of a set of training examples. In supervised learning, each example is a pair consisting of an input object (a vector) and a desired output value (supervisory signal). (2010, 2012)

2.1.6.2 Unsupervised learning

Is the machine learning task of inferring a function to describe hidden structure from "unlabelled" data. Unsupervised machine learning involves discovery of classes without a prior knowledge or use of sample classification. Unsupervised machine learning is sometimes called cluster analysis. [(Wu, 2014)]

2.1.7 Support vector machine (SVM)

Support Vector Machine is a useful technique for data classification, prediction and regression. A classification task which is usually involves with training and testing data that consist of some data instances, so each instance in the training set contains one “target value” (class labels) and several “attributes” (features). The goal of SVM is to produce a model which predicts the target value of data instances in the testing set which are given only the attributes.[(Nalavade, Kamini and Meshram, BB}, 2012)].
The advantages of support vector machines are:

- High-Dimensionality - The SVM is an effective tool in high-dimensional spaces, which is particularly applicable to document classification and sentiment analysis where the dimensionality can be extremely large.

- Memory Efficiency - Since only a subset of the training points is used in the actual decision process of assigning new members, only these points need to be stored in memory (and calculated upon) when making decisions.

- Versatility - Class separation is often highly non-linear. The ability to apply new kernels that allows substantial flexibility for the decision boundaries, leading to greater classification performance.(Jannsen, 2008)

The disadvantages of support vector machines include:

- If the number of features is much greater than the number of samples, the method is likely to give poor performances.

- SVMs do not directly provide probability estimates, these are calculated using an expensive k-fold cross-validation.(Jannsen, 2008)

2.1.8 The flow of data classification prediction

In general the process of data classification prediction based on SVM includes four steps: determining the training data set and testing data set, data pre-processing, training the SVM model, data classification be generalized to other data later on.

![Figure 2-1: The flow of data classification](image-url)
• **Data pre-processing**: is a data mining technique that involves transforming raw data into an understandable format. Data pre-processing is a proven method of resolving such issues. Data pre-processing prepares raw data for further processing. The following steps which are commonly taken in machine learning for data pre-processing.

  • **Data Cleaning**: Data is cleansed through processes such as filling in missing values, smoothing the noisy data, or resolving the inconsistencies in the data.

  • **Data Integration**: Data with different representations are put together and conflicts within the data are resolved.

  • **Data Transformation**: Data is scaling, normalized, aggregated and generalized.

  • **Data Reduction**: This step aims to present a reduced representation of the data in a data warehouse.

  • **Data Discretization**: Involves the reduction of a number of values of a continuous attribute by dividing the range of attribute intervals.

The training SVM model needs to adjust the relevant parameters, which choose from many ways, such as **cross validation**, which is the process of training learners using one set of data and testing it using a different set. In **v-fold cross-validation**, first divides the training set into subsets of equal size. Sequentially one subset is tested by using the classifier trained on the remaining V-1 subsets. Thus, each instance of the whole training set is predicted once so the cross-validation accuracy is the percentage of data which are correctly classified. The cross-validation procedure can prevent the over fitting problem. And **grid search** which is originally an exhaustive search based on a defined subset of the hyper-parameter space. The hyper-parameters are specified using minimal value (lower bound), maximal value (upper bound). Grid search optimizes the SVM parameters(C, degree.) using a cross validation (CV) technique as a performance metric.
2.2 RELATED WORKS

In 2012, Oded Magger, Yedael Y. Waldman, Eytan Ruppin and Roded Sharan presented the first new large-scale method that aims to enhance the accuracy of the existing network-based gene prioritization algorithms by taking into account tissue-specific information, which achieved by constructing tissue-specific protein–protein interaction (PPI) networks and utilizing its for gene prioritization instead of the standard generic PPI network. Used Node Removal (NR) generate –SPN-by removing genes which are considered unexpressed from the network and Edge Reweight - Reducing the weight of an edge to convert PPIN to tissue specific network.

The result constructed tissue-specific PPI networks, which used the prioritization results of different tissues in order to suggest new disease-tissue associations.

In 2012, Guan Y, Gorenshteyn D, Burmeister M, Wong AK, Schimenti JC proposed a new strategy to solve the problem of tissue specificity of functional genomics data, which reflecting the potentially different roles of proteins and pathways in diverse cell for specific tissue types by generating tissue-specific functional networks, which actively working on the diversity of protein function to represent more accurately of the genes associated with the phenotype. They used these tissue-specific networks to predict genes associated with different phenotypes.

The result demonstrates that tissue-specific functional networks can improve prediction accuracy for phenotype-associated genes compared to a single functional network through computational analyses, and through experimentally confirmed predictions of novel fertility-related genes and visualization of their local networks.

In 2014, Kai Sun, Joana P Goncalves, Chris Larminie and Nataša Pržul uncovered novel disease associations based on the underlying biological mechanisms. Exploring disease-disease associations by using system-level biological data is expected to improve knowledge of disease relationships, which may lead to further improvements
in disease diagnosis, prognosis and treatment. They analysed and compared four publicly available disease-gene association datasets, then applied three disease similarity measures, namely annotation-based measure, function-based measure and topology-based measure, to estimate the similarity scores between diseases. They systematically evaluated disease associations obtained by these measures against a statistical measure of comorbidity. Used GWAS to evaluate our predicted disease associations.

The results showed the correlation between predicted disease associations by similarity measures and from genome-wide association studies are substantially higher than expected at random, Novel disease associations enhance knowledge of disease relationship.

In 2015, Limei Wang, Maozu Guo, Jin Li, Ruijie Zhang, Qiguo Dai, Xiaoyan Liu, Chunyu Wang, Zhixia Teng, Ping Xuan and Mingming Zhang constructed an Expression quantitative trait loci (eQTLs) based gene–gene co-regulation network (GGCRN) to determine genes associated with several diseases with Protein–protein interaction data and EQTL data. They used direct neighbourhood algorithm that works follows, the interaction partners in the network were determined for each known disease gene. The more linkages that exist between a gene and known disease genes for a particular disease, the greater the possibility that it is related to that disease and used Random walk with restart algorithm (RWR) for disease genes mining. Then used the HPRD PPI and Union networks to compare with the algorithms used.

The result showed the eQTL-based GGCRN provided extra gene–gene interaction information than the HPRD PPI network. Disease gene mining using an RWR approach with an integrated network (HPRD PPI and GGCRN) provided faster convergence and identified some new disease-related genes.

In 2016, jajie peng, Kun Bai, Xuequn Shang, Guohua Wang, liangcheng m Yadong Wang, and jin Chen presented a new algorithm to define edge weights in an integrated
network and used the weighted network to predict new disease gene relationships. Used SLN-SRW (Simplified Laplacian Normalization-Supervised Random walk) to compute and model edge weights of integrated networks.

The result proposed a new network based disease gene prediction method to generate and model the edge weight of a new biomedical network that integrate biomedical data from heterogeneous sources, thus enhancing the disease related gene discovery.

In 2017, KeHu, Jing.BoHu, Ju Xiang, Hui.jiaLi,Yan Zhang and Shi Chen proposed a new disease gene prediction method. They used path based similarity (PSI) to measure the similarity between two genes in protein protein interaction network. Used community based similarity (CS) to improve the performance of network based disease gene prediction. The combined similarity methods was proposed by considering path structure and community structure in network.

The result indicated that genes associated with same disease reside in same community of protein protein interaction network and community structure is greatly helpful for disease gene prediction.

<table>
<thead>
<tr>
<th>No</th>
<th>Author</th>
<th>Research Title</th>
<th>Methodology</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2012, Oded Magger, Yedael Y. Waldman, Eytan Ruppin and Roded Sharan</td>
<td>Enhancing the Prioritization of Disease-Causing Genes through Tissue Specific Protein Interaction Networks</td>
<td>Used Node Removal (NR) generate –SPN-by removing genes which are considered unexpressed from the network and Edge Reweight - Reducing the weight of an edge to convert PPIN to tissue specific network</td>
<td>Constructing tissue-specific protein-protein interaction networks</td>
</tr>
<tr>
<td>2</td>
<td>2012 Guan Y, Gorenshteyn D</td>
<td>Tissue-specific functional networks</td>
<td>Used Bayesian network data integration with mutual predicting genotype-phenotype</td>
<td></td>
</tr>
<tr>
<td>Year</td>
<td>Authors</td>
<td>Methodology</td>
<td>Data Source</td>
<td>Results</td>
</tr>
<tr>
<td>------</td>
<td>---------</td>
<td>-------------</td>
<td>-------------</td>
<td>---------</td>
</tr>
<tr>
<td>2014</td>
<td>Burmeister M, Wong AK, Schimenti JC</td>
<td>for prioritizing phenotype and disease genes</td>
<td>information-based regularization</td>
<td>Uncovered Novel disease associations</td>
</tr>
<tr>
<td>2015</td>
<td>Sun, Joana P Goncalves, Chris Larminie and Nataša Pržul</td>
<td>Predicting disease associations via biological network analysis</td>
<td>analysed gene association datasets, applied three disease similarity measures Used GWAS to evaluate our predicted disease associations</td>
<td></td>
</tr>
<tr>
<td>2015</td>
<td>Limei Wang, Maozu Guo, Jin Li, Ruijie Zhang, Qiguo Dai, Xiaoyan Liu, Chunyu Wang, Zhixia Teng, Ping Xuan and Mingming Zhang</td>
<td>Mining disease genes using integrated protein-protein interaction and gene co-regulation information</td>
<td>They used direct neighbourhood algorithm and used Random walk with restart algorithm (RWR) for disease genes mining</td>
<td></td>
</tr>
<tr>
<td>2016</td>
<td>Jajie Peng, Kun Bai, Xuequn Shang, Guohua Wang, Liangcheng Li, Yadong Wang, and Jin Chen</td>
<td>Predicting disease associations via biological network analysis</td>
<td>Used SLN-SRW (Simplified Laplacian Normalization-Supervised Random walk) compute and model edge weights of integrated network and then predict disease genes</td>
<td>Discovery disease genes relationships based on biomedical networks</td>
</tr>
<tr>
<td>2017</td>
<td>KeHu, Jing BoHu, Ju Xiang, Hui jia Li, Yan Zhang and Shi</td>
<td>Predicting disease related genes by path based similarity and community structure in protein-protein interaction network</td>
<td>Used path based similarity (PSI) community based similarity (CS) and then proposed combined similarity methods by considering path structure and community structure in network</td>
<td></td>
</tr>
</tbody>
</table>


In this study, several biological tools were used to predict Alzheimer and colorectal cancer disease based on genes prediction by constructing tissue specified gene network. Using (DisGeNET) to determine gene disease association, and (GTEx) to illustrate the most affected tissues, then extract the actual genes that cause Alzheimer's disease and colorectal cancer through the constructed tissue specified gene network (TSG).
CHAPTER THREE
RESEARCH METHODOLOGY

3.1 INTRODUCTION

The research methodology follows many steps to access the prediction, where firstly, download genes expression features, then second determine gene disease association, and third, illustrate the tissues more affected by certain disease. At the same time calculate PCC for all gene expression features, then fourth, construct tissue specified gene network TSG, and fifth, weight the TSG network to reduce it and then identify the remaining genes as they are discovered genes, as show in figure 3.1.

Figure 3-1: Research methodology
3.2 DATASET FEATURE: GENE EXPRESSION OMNIBUS (GEO) DATABASE

There are 978 genes for 3 kinds of diseases. The human body index transcriptional profiling of tissue-specific gene expression data set was downloaded from the gene expression omnibus (GEO) database in GSE7307 on Apr 09, 2007[(Edgar, Ron and Domrachev, Michael and Lash, Alex E, 2002)]. For determine gene_disease association uses (DisGeNET) is a discovery platform designed to address a variety of questions concerning the genetic underpinning of human diseases. DisGeNET enables to choose the disease's name and then shows the genes most related to the disease. [(Pinerero, Janet and Queralt-Rosinach, Nuria and Bravo, Alex and Deu-Pons, Jordi and Bauer-Mehren, Anna and Baron, Martin and Sanz, Ferran and Furlong, Laura I, 2015)]

![Top 10 gene associations for this disease](image)

**Figure 3-2:** gene associations in Alzheimer's disease by (DisGeNET)

From the above figure we note that the listed genes in the human body are more affected by the Alzheimer's disease by score of association.
3.3 TISSUE AFFECTED BY THE DISEASE

The Genotype-Tissue Expression (GTEx) project aims to provide to the scientific community a resource with which to study human gene expression and regulation and its relationship to genetic variation. This software will collect and analyse multiple human tissues from donors who are also densely genotyped, to assess genetic variation within their genomes. By analysing global RNA expression within individual tissues and treating the expression levels of genes as quantitative traits, variations in gene expression that are highly correlated with genetic variation can be identified as an expression quantitative trait. This software show the most affected tissues according to the name of the gene that is associated to the same disease.[(Lonsdale, John and Thomas, Jeffrey and Salvatore, Mike and Phillips, Rebecca and Lo, Edmund and Shad, Saboor and Hasz, Richard and Walters, Gary and Garcia, Fernando and Young, Nancy and others, 2013)].

![Figure 3-3: Tissues affected by alzheimer's disease using (GTEx)](image)

The above figure illustrates the tissues may be affected or changed fully or partially in the human body as a result of certain genes affected by Alzheimer's disease. Note that brain tissues are the most damaged when Alzheimer's disease.
3.4 CALCULATE PEARSON CORRELATION COEFFICIENT (PCC)

The Pearson correlation coefficient is a statistical calculation of the strength of two variables’ relationships. Or it’s a measure of how dependent two variables are on one another. In the research, calculate Pearson correlation coefficient (PPC) for gene expression details of query disease were downloaded from (GEO) between any two genes in MATLAB software as a tool for calculating. This means that 978 values are available for correlation between genes.

3.5 TISSUE SPECIFIED GENE GENE (TSG) NETWORK

Gene networks (GNs) is the most important approaches to discover which gene-gene relationships are involved in a specific biological process. [(mez-Vela, Francisco and Díaz-Díaz, Norberto, 2014)] Gene-gene network can be represented as a graph where genes are represented as nodes and their relationships as edge, which in the research was considered a Pearson correlation coefficient (PCC) value. If the value is higher than the threshold (the value of the threshold is taken 0.5 as a default value and as an average of values), it means that the genes more closely together (strong correlation) and remove the lower PCC value genes interaction, as the result of the values of PCC and the tissue affected by query disease construct a network among the genes that interact together as illustrated in the result.
3.6 WEIGHTED TISSUE SPECIFIED GENE GENE (TSG) NETWORK

The weight of the tissue specified gene-gene (TSG) network that constructed previously calculated according to equation (3.1)

\[
S(i, j) = \alpha \left( \sum_{k=1}^{n} \frac{a_{ii} a_{jj}}{N_k} \right) + (1 - \alpha) PCC
\]

---------------------------  Equation(3.1). (Ganegoda, Gamage Upeksha and Wang, JianXin and Wu, Fang-Xiang and Li, Min, 2014)

Notations:

\( S (i,j) \) = similar weight between gene i and gene j where \( i \neq j \)

\( a_{ik} \) = gene i with phenotype k

\( a_{jk} \) = gene j with phenotype k

\( PCC \) = Pearson correlation coefficient

\( \alpha \) = The weighting parameter \( = (0.5) \)

\( [a_{ik} = 1 \) if gene I has phenotype k and \( a_{ik} = 0 \) other wise , \( N_k \) is the number of genes involved in the specific phenotype k, and \( n \) is the total number of phenotypes in query disease. The weighting parameter \( \alpha [0,1] \) with a default value of \( \alpha = (0.5) \), where a phenotype is the composite of an organism's observable characteristic or traits].

To clarify equation (3.1), took the colorectal cancer as an example, where identified all the phenotype that cause colorectal cancer through a database named (Monarch) [driven to truly integrate biological information using semantics, and present it in a novel way, leveraging phenotypes to bridge the knowledge gap. And use of computational reasoning to enable phenotype comparison, both within and across species, with the ultimate goal of improving biomedical research]. - (Zhan, Shuai
These phenotypes are represented in neoplasm of stomach, renal cell carcinoma, hereditary non-polyposis colorectal carcinoma, uterine leiomyosarcoma and transitional cell carcinoma of the bladder. As shown below:

![Figure 3-4: phenotypes of colorectal cancer through Monarch database](image)

Then test each gene individually through the Monarch database and determine if it includes these phenotypes or not. If includes one or more phenotype assign a value of 1, otherwise carries a value of 0.
When applied the equation (3.1), took each two genes together, as in the tissue specified gene-gene network (TSG) and chose the gene values according to their containment for the phenotype $N_k$ values are determined according to the gene's number carrying the phenotype. If two genes have phenotypes, the value of $N_k$ is 2 or 1 if only one contains phenotypes. And then choose the Pearson correlation coefficient value (PCC) for each two genes. For calculating edges weight in all interaction in the tissue specified genes (TSG) network and construct weighted network, must choose the values above (0.5) and remove the lower values. The process of calculating the edge's weight produced what it called discovery or prediction of additional disease genes according to the highest values of the weight.
3.7 SUPPORT VECTOR MACHINES (SVM)

Support vector machines are a set of supervised learning methods which learn from the dataset and used for classification. For the training data we have a set of input vectors, denoted $x_i$, with each input vector having a number of component features. These input vectors are paired with corresponding labels, which we denote $(y_i)$, and there are $M$ such pairs ($i=1,...,m$).

![Diagram of SVM workflow]

**Figure 3-6: Workflow of SVM with cross-validation exercise**

SVM requires that each data instance is represented as a vector of real numbers, so it must convert them into numeric if are not in correct form. The data in this work has already been clarified are split into training and test set. Fitted a prediction model to
the training data [in the training data set each instance set contains one target value (the class labels) and several attribute (the features)], made predictions based on this data and tested the predictions on the test data.

3.7.1 Split the data for 10 Folds Cross Validation

The data were divided into two parts: training data, which represent 90% of total dataset and testing data, which represent the rest 10%. In 10-fold cross-validation (is statistical analysis method used to verify the performance of the classifier) the training set into 10 subsets of equal size. Sequentially one subset is tested using the classifier trained on the remaining 9 subsets. Thus, each instance of the whole training set is predicted once so the cross-validation accuracy is the percentage of data which are correctly classified.

3.7.2 The method involves the following steps

1- Randomly partition the training set into $K$ groups of equal length.
2- For a group of training set, fit the model using all training set except that group.
3- Use that group’s training set to test the model’s predictive performance.
4- Repeat steps 2 and 3 for the other groups.
3.7.3 Performance Measures:

Performance metrics for binary classification are designed to capture trade-offs between four fundamental population quantities: true positives, false positives, truenegatives and false negatives. It comes from the matrix called confusion matrix.

3.7.3.1 Confusion Matrix

<table>
<thead>
<tr>
<th>Total population</th>
<th>Predicted positive</th>
<th>Predicted Negative</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Actual Positive</td>
<td>TP</td>
<td>FN</td>
<td>AP</td>
</tr>
<tr>
<td>Actual Negative</td>
<td>FP</td>
<td>TN</td>
<td>AN</td>
</tr>
<tr>
<td>Total</td>
<td>PP</td>
<td>PN</td>
<td>N</td>
</tr>
</tbody>
</table>

Where

TP: true positives _ FP: false positives
TN: true negatives _ FN: false negatives

3.7.3.2 Accuracy: Is the proportion of the true results, either true positives or true negatives, in a population. (Zhu, Wen and Zeng, Nancy and Wang, Ning and others, 2010)

\[
\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}............................. (3.2)
\]

3.7.3.3 Sensitivity: Is the proportion of true positives that are correctly identified [(Zhu, Wen and Zeng, Nancy and Wang, Ning and others, 2010)]

\[
\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}............................. (3.3)
\]
3.7.3.4 **Specificity**: Is the Proportion of true negatives that are correctly identified (Zhu, Wen and Zeng, Nancy and Wang, Ning and others, 2010)

\[
\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}} \quad \text{------------------------} \quad (3.4)
\]

3.7.3.5 **The receiver operating characteristic (ROC) curve**

Consider as the graphic presentation of the relationship between both sensitivity and specificity (that illustrates sensitivity (TPR) as y-coordinate versus false positive rate (FPR) as x-coordinate) and it helps to decide the optimal model through determining the best threshold for the diagnostic test can be determined from sensitivity and specificity with the presence of prevalence.[32]. In this work a ROC curve was plotted two times for all genes and the other for predicted genes according to the values of sensitivity and specificity.

3.7.3.6 **Area under curve (AUC)**

Area under curve is a proportion of the total area that falls below the Roc curve. It is useful and increasingly popular performance measure. The representation of the region is better when it is close to (1).
CHAPTER FOUR
RESULTS AND DISCUSSION

4.1 INTRODUCTION

This chapter shows tissue specified gene-gene network (TSG) before and after calculate the weighting of network, which represent the prediction of disease genes, then applied support vector machine(random samples) in all data set and in part of data, which represent the predicted data (genes). Cross validation will be used to split the data into training and testing data set. And then calculate the performance measures: accuracy, sensitivity, specificity. Used (SPSS) to plot the ROC curve for both the total data and predicted.

4.2 RESULTS

The network in Figure (4.1) represents all the genes for the diseases associated with each other according to the Pearson correlation coefficient (PCC).

4.2.1 Tissue specified gene_gene network

![Figure 4-1: Tissue specified gene_gene network (TSG)]
4.2.2 Weighted tissue specified gene _gene network

4.2.2.1 Alzheimer's disease

Network figure (4.2) represents Genes that have been predicted and cause Alzheimer disease. These genes are APP, APOE, MLMH, APBB2, NOS3, ADAM10 and MAP2K4. The value of the edge between the nodes in the network represents the strength of the correlation between them. Therefore, the correlation between each two genes is higher than 80%.

Figure 4-2: Weighted tissue specified gene _gene network for Alzheimer disease
4.2.3 A performance measure of using SVM

Calculate performance measures of two classes for certain disease twice, once with all genes and once with predicted genes.

4.2.3.1 Alzheimer's disease

Class one shows the Performance measure's result for Alzheimer's disease that presented twice. The first with All genes in the human body, causing many diseases, which include 978 genes and the second with The actual predicted genes that causes Alzheimer 's disease, which include 7 genes. Note that the result of the predicted genes better than all the genes in terms of accuracy, Sensitivity and Specificity.

<table>
<thead>
<tr>
<th>Disease (1)</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>All genes</td>
<td>66.0583%</td>
<td>36.37%</td>
<td>70.69%</td>
</tr>
<tr>
<td>Predicted genes</td>
<td>70.7143%</td>
<td>50.32%</td>
<td>80.00%</td>
</tr>
</tbody>
</table>

Table 4-1: performance measures of Alzheimer's disease

Receiver Operating Characteristics (ROC) Curve for Alzheimer's disease

Figure 4.3 shows ROC curves for Alzheimer's disease with all data (genes). Figure 4.4 shows ROC curves for Alzheimer's disease with the predicted data (genes).

In a Receiver Operating Characteristic curve the true positive rate (Sensitivity) is plotted in function of the false positive rate (1-Specificity) for different cut-off points. ROC curves tend to go from the bottom left corner to the top right corner of the box. The closer the ROC curve get to the top left corner, the better the test is overall. The closer the curve comes to the centre diagonal line, the worse the test. Note in Figure 4.3 that the roc curve tends to be linear or close to the diagonal line while Figure 4.4 shows more curvature of the top left corner.
Figure 4-3: ROC curve for all data

Figure 4-4: ROC curve for predicted data
4.2.2.2 Weighted tissue specified gene _gene network for Colorectal cancer

Network figure (4.5) represents Genes that have been predicted and cause colorectal cancer. These genes are NOS3, BUB1B, AURKA, NRAS, TP53, SRC, PTPN12, CCNB1, BAX, AKT1 and TLR4. The value of the edge between the nodes in the network represents the strength of the correlation between them. Therefore, the correlation between each two genes is higher than 75%.

Figure 4-5: Weighted tissue specified gene _gene network for colorectal cancer

4.2.3.2 A performance measure for Colorectal cancer

Class two shows the Performance measures result for colorectal cancer that presented twice. The first with All genes in the human body, causing many diseases, which include 978 genes and the second with The actual predicted genes that causes colorectal cancer, which include 11 genes. Note that the result of the predicted genes better than all the genes in terms of accuracy, Sensitivity and Specificity.
Table 4-2 Performance measures of colorectal cancer

<table>
<thead>
<tr>
<th>Disease</th>
<th>Accuracy</th>
<th>Sensitivity</th>
<th>Specificity</th>
</tr>
</thead>
<tbody>
<tr>
<td>All genes</td>
<td>67.2063%</td>
<td>64.29%</td>
<td>70.00%</td>
</tr>
<tr>
<td>Predicted genes</td>
<td>77.6923%</td>
<td>70.80%</td>
<td>79.34%</td>
</tr>
</tbody>
</table>

Receiver Operating Characteristics (ROC) Curve for colorectal cancer

The Roc curve for SVM with predicted genes in figure 4.7 seems better than the roc in figure 4.5 with all genes because a roc curve close to upper left corner that represent high values for sensitivity and specificity. While figure 4.6 close to be linear or close to the diagonal line (worst test as mentioned previously).

Figure 4.6 ROC curve for colorectal cancer
Figure 4-7: ROC curve for colorectal cancer
4.2.5 The Area under the Roc Curve (AUC)

AUC values close to 1 indicates a perfect test. AUC values of less than .5 indicates no discriminative values and is represented by a straight, diagonal line extending from the lower left corner to upper right.

Table 4.3 shows the AUCs for classes, and the value of the area under the Roc curve for predicting data is higher than all data.

**Table 4-3: The Area Under the Roc Curve (AUC)**

<table>
<thead>
<tr>
<th>Classes(diseases)</th>
<th>All data</th>
<th>Predicted data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alzheimer disease</td>
<td>0.47%</td>
<td>0.52%</td>
</tr>
<tr>
<td>Colorectal cancer</td>
<td>0.43%</td>
<td>0.62%</td>
</tr>
</tbody>
</table>

The study compared diseases with all genes and predicted genes and found that performed more accurately in predicted genes (area under the curve = 0.53 in Alzheimer disease and area under the curve =0.62 for colorectal cancer ) compared to all genes (area under the curve = 0.47 in Alzheimer disease and area under the curve =0.43 for colorectal cancer).
CHAPTER FIVE

CONCLUSION AND RECOMMENDATIONS

5.1 CONCLUSIONS

The research predicts genes cause certain disease based on constructing a tissue specified gene network (TSG). To validate the prediction applies support vector machine twice. Once with all data and other part of the data (predicted). The result explains the performance of support vector machine with predicted genes is better than the support vector machine with all genes. Performance measures achieved the highest predicted data values for accuracy, sensitivity, specificity and area under Roc (AUR), indicating the validity of predicting disease genes.

5.2 RECOMMENDATIONS

The research recommend to expand the field of prediction of disease genes using other methods such as ProSimas prioritization method and random walks with restart on heterogeneous network (RWRH) for predicting gene disease association.
REFERENCES

(Su, Gang and Morris, John H and Demchak, Barry and Bader, Gary D, 2015).

(Sherwood, 2015).


Lonsdale, John and Thomas, Jeffrey and Salvatore, Mike and Phillips, Rebecca and Lo, Edmund and Shad, Saboor and Hasz, Richard and Walters, Gary and Garcia, Fernando and Young, Nancy and others, 2013. The genotype-tissue expression (GTEx) project. *Nature genetics*, pp. 580--585.


(Le, Duc-Hau and Dang, Vu-Tung, 2016)

(Alpaydin, 2014)

(Jagadeesh, B and Kumar, P Rajesh and Reddy, P Chenna, 2013)

(Vapnik, Vladimir and Chapelle, Olivier, 2000).
(Insan, Nitin Goel and Mane, Vijay and Chaudhary, BL and Danu, Mahesh Singh and Yadav, Amod and Srivastava, Vive, 2013)

(Vanunu, 2010)

(Hamosh, Ada and Scott, Alan F and Amberger, Joanna S and Bocchini, Carol A and McKusick, Victor A, 2005)

(Nguyen, Thanh Phuong and Ho, Tu Bao, 2007)

(Wray, Naomi R and Goddard, Michael E and Visscher, Peter M, 2007)

( Arzucan and Vu, Thuy and Erkan, G{"u}ne{s} and Radev, Dragomir R, 2008)

(Franklin, 2013)


(Luscombe, Nicholas M and Greenbaum, Dov and Gerstein, Mark and others, 2001).

(Wheeler, David L and Barrett, Tanya and Benson, Dennis A and Bryant, Stephen H and Canese, Kathi and Chetvernin, Vyacheslav and Church, Deanna M and DiCuccio, Michael and Edgar, Ron and Federhen, Scott and others, 2007)

({Whisstock, James C and Lesk, Arthur M, 2003}).

(Karolchik, Donna and Baertsch, Robert and Diekhans, Mark and Furey, Terrence S and Hinrichs, Angie and Lu, YT and Roskin, Krishna M and Schwartz, Matthias and Sugnet, Charles W and Thomas, Daryl J and others, 2003).

(Peri, Suraj and Navarro, J Daniel and Kristiansen, Troels Z and Amanchy, Ramars and Surendranath, Vineeth and Muthusamy, Babylakshmi and Gandhi, TKB and Chandrika, KN and Deshpande, Nandan and Suresh, Shubha and others, n.d.)
(Murphy, 2012)
(2010, 2012)
(Wu, 2014)
(Nalavade, Kamini and Meshram, BB, 2012)
(Liu, 2016)
(Keerthi, S Sathiya and Lin, Chih-Jen, 2003)
(Edgar, Ron and Domrachev, Michael and Lash, Alex E, 2002)
(Piñero, Janet and Queralt-Rosinach, Nuria and Bravo, Alex and Deu-Pons, Jordi and Bauer-Mehren, Anna and Baron, Martin and Sanz, Ferran and Furlong, Laura J, 2015)
(Lonsdale, John and Thomas, Jeffrey and Salvatore, Mike and Phillips, Rebecca and Lo, Edmund and Shad, Saboor and Hasz, Richard and Walters, Gary and Garcia, Fernando and Young, Nancy and others, 2013)
(mez-Vela, Francisco and Diaz-Diaz, Norberto, 2014)
(Zhan, Shuai and Reppert, Steven M, 2012)
(Zhu, Wen and Zeng, Nancy and Wang, Ning and others, 2010)
Appendix

Split the data for 10_Folds Cross Validation

clear all;
clc
load testool.mat;
XALL=vertcat(Sheet1);
YALL=vertcat(Sheet3);
X1=XALL;
N= size(X1,1);
Rnd=randperm(size(XALL,1));
X1=XALL(Rnd,:);
Y=YALL(Rnd,:);
foldsize=floor(N/10);
fold1=X1(1:foldsize,:);Yfold1=Y(1:foldsize);fold1(:,size(fold1,2)+1)=1;
fold2=X1(foldsize+1:2*foldsize,:);Yfold2=Y(foldsize+1:2*foldsize);fold2(:,size(fold2,2)+1)=2;
fold3=X1(2*foldsize+1:3*foldsize,:);Yfold3=Y(2*foldsize+1:3*foldsize);fold3(:,size(fold3,2)+1)=3;
fold4=X1(3*foldsize+1:4*foldsize,:);Yfold4=Y(3*foldsize+1:4*foldsize);fold4(:,size(fold4,2)+1)=4;
fold5=X1(4*foldsize+1:5*foldsize,:);Yfold5=Y(4*foldsize+1:5*foldsize);fold5(:,size(fold5,2)+1)=5;
fold6=X1(5*foldsize+1:6*foldsize,:);Yfold6=Y(5*foldsize+1:6*foldsize);fold6(:,size(fold6,2)+1)=6;
fold7=X1(6*foldsize+1:7*foldsize,:);Yfold7=Y(6*foldsize+1:7*foldsize);fold7(:,size(fold7,2)+1)=7;
fold8=X1(7*foldsize+1:8*foldsize,:);Yfold8=Y(7*foldsize+1:8*foldsize);fold8(:,size(fold8,2)+1)=8;
fold9=X1(8*foldsize+1:9*foldsize,:);Yfold9=Y(8*foldsize+1:9*foldsize);fold9(:,size(fold9,2)+1)=9;
fold10=X1(9*foldsize+1:end,:);Yfold10=Y(9*foldsize+1:end);
Allfold=vertcat(fold1,fold2,fold3,fold4,fold5,fold6,fold7,
fold8,fold9,fold10);
Y=vertcat(Yfold1,Yfold2,Yfold3,Yfold4,Yfold5,Yfold6,Yfold7,
Yfold8,Yfold9,Yfold10);
Accuracy=[];
for fold=1:10
Testind=Allfold(:,size(Allfold,2))==fold;
Trainind=Allfold(:,size(Allfold,2))~=fold;
Accuracy=[Accuracy 100*sum(Testind)/length(Allfold)];
end

Xtest=Allfold(Testind,1:size(Allfold,2)-1);Ytest=Y(Testind);
Xtrain=Allfold(Trainind,1:size(Allfold,2)-1);ytrain=Y(Trainind);
    X=Allfold(Trainind,1:size(Allfold,2)-1);y=Y(Trainind);
AllData(fold,:).Testind=Testind;
AllData(fold,:).Trainind=Trainind;
AllData(fold,:).Xtest=Xtest;
AllData(fold,:).Ytest=Ytest;
AllData(fold,:).Xtrain=Xtrain;
AllData(fold,:).ytrain=ytrain;
model = svmtrain(ytrain, Xtrain, '-c 2 -g 0.001 ');
[predict_label, accuracy, prob_estimates] = svmpredict(Ytest,Xtest, model);
TP=sum(Ytest==1) & (predict_label==1);
TN=sum(Ytest==0) & (predict_label==0);
FP=sum(Ytest==0) & (predict_label==1);
FN=sum(Ytest==1) & (predict_label==0);
Qtotal=(TP+TN)/(TP+TN+FP+FN);
Accuracy=[Accuracy accuracy];
Meanaccuracy=sum(Accuracy,2)/10;
end

Equation of weighted tissue specified gene network

Equation of weighteted tissue specified gene netwwork

The confusion matrix to calculate the performance measures

TP=sum((Ytest ==1) & (predict_label ==1));
    TN=sum((Ytest ==0) & (predict_label ==0));
FP=sum((Ytest ==0) & (predict_label ==1));
FN=sum((Ytest ==1) & (predict_label ==0));