

**Doctoral Dissertation**

**Development of a Human Biomarker  
Database Towards Classification of Diseases  
and Finding Inter-Disease Relations**

**Shaikh Farhad Hossain**

March 17, 2021

Division of Information Science  
Graduate School of Science and Technology  
Nara Institute of Science and Technology  
Japan

A Doctoral Dissertation  
submitted to the Graduate School of Science and Technology,  
Nara Institute of Science and Technology  
in partial fulfillment of the requirements for the degree of  
Doctor of Engineering

Thesis Committee:

Professor Shigehiko Kanaya	(Supervisor)
Professor Kenichi Matsumoto	(Co-supervisor)
Associate Professor Md. Altaf-Ul-Amin	(Co-supervisor)
Associate Professor Naoaki Ono	(Co-supervisor)
Assistant Professor Alex Ming Huang	(Co-supervisor)

# Development of a Human Biomarker Database Towards Classification of Diseases and Finding Inter Disease Relations

Shaikh Farhad Hossain

## Abstract

A biomarker is a measurable indicator of a disease or abnormal state of a body that plays an important role in disease diagnosis, prognosis, and treatment. The biomarker has become a significant topic due to its versatile usage in the medical field and in rapid detection of the presence or severity of some diseases. The volume of biomarker data is rapidly increasing and the identified data is scattered. To provide comprehensive information, the explosively growing data needs to be recorded in a single platform. There is no open-source freely available comprehensive online biomarker database. To fulfill this purpose, we have developed a human biomarker database as part of the KNAPSAcK family of databases which contain a vast quantity of information on the relationships between biomarkers and diseases. We have classified the diseases into 18 disease classes, mostly according to NCBI definitions. Apart from this database development, we also have performed disease classification by separately using protein and metabolite biomarkers based on the network clustering algorithm DPCLUSO and hierarchical clustering. Finally, we reached a conclusion about the relationships among the disease classes. The human biomarker database can be accessed online and the inter-disease relationships may be helpful in understanding the molecular mechanisms of diseases. To our knowledge, this is one of the first approaches to classify diseases based on biomarkers.

**Keywords:** Biomarker, Disease classification, protein, metabolite, Graph

Clustering

---

<sup>1</sup>Doctoral Dissertation, Division of Information Science, Graduate School of Science and Technology, Nara Institute of Science and Technology, NAIST-IS-DD1821052, March 17, 2021.

## Acknowledgements

First and foremost, I thank my creator, the Lord of the Worlds, the Beneficent, and the Most Merciful, for giving me the light and for enabling me to complete this dissertation. There were many obstacles during this journey, but with his help and will, I manage to go through it.

I owe special thanks to my supervisor, Professor Shigehiko Kanaya for giving the opportunity to complete this research work. His continuous support and motivation help me to achieve this immense goal in my study life.

I want to express my gratitude for Professor Kenichi Matsumoto for giving his time to review my thesis with insightful recommendations.

My sincere thanks go to co-supervisor Associate Professor Md. Altaf-Ul-Amin who inspired me throughout my research time with a lot of help and support. During the predicament period of my research work, he always supports me to be patient and strong.

My special thanks go to Associate Professor Naoaki Ono and Assistant Professor Ming Huang for all their support, valuable comments and suggestions. I am also grateful to Mrs. Minako Ohashi for her help in the administrative matter and Mrs. Aki Hirai Morita for helping to develop the KNApSAcK Biomarker Database.

I wish to thank all my fellow lab members in Computational Systems Biology Laboratory for their technical support. Together we had a lot of fun and celebrated different occasion. I surely will miss the time of staying in this lab. I would like to give special thanks, Dr. Mohammad Bozlul Karim. From the beginning of my PhD, he supports me in deciding which directions I wanted to go with my research.

I would express my gratitude to Prof Dr. Liakot Ali from IICT department of Bangladesh University of Engineering and Technology. His inspiration also helped me to achieve this research work.

I would also like to express my gratitude to the authority of the international division of the Nara Institute of Science and Technology for providing me time to time help and support. With their assistance, I feel like spending my entire research time in a friendly environment. I met a lot of different cultural student and people in here which also helped me to learn a lot of knowledge.

My special and highest appreciation to all my family members, especially my parents, Shaikh Golam Mostofa and Mafuja Begum and my beloved wife Tania Sultana and my gentle smiling son Shaikh Fardin Rushan for their immense support and unconditional love. I also apologize to them not giving enough time due to my research work.

I also express my gratitude to my brothers, sister, relatives and friends for their prayer and encouragement.

Last but not least, my special appreciation to google especially for google scholar. I also would like to express my appreciation to all google staff for their information system.

March 17, 2021

Shaikh Farhad Hossain

## List of Abbreviations

A1AD	Alpha-1 antitrypsin deficiency
AD	Alzheimer Disease
APS	Autoimmune poly glandular syndromes
AT	Ataxia telangiectasia
ATM	Ataxia telangiectasia mutated
B-CLL	B-cell chronic lymphocytic leukemia
Bk	Baker's Gamma coefficient
BMD	Bone mineral density
CAH	Congenital adrenal hyperplasia
CF	Cystic fibrosis
CFBD	Cystic fibrosis-related bone disease
CID	Compound Identifier
COPD	Chronic obstructive pulmonary disease.
CTD	Composition Transition and Distribution
DMD	Duchenne muscular dystrophy
GD	Gaucher disease
GI	Gastrointestinal tract
H. pylori	Helicobacter pylori
HDL-c	High-density lipoprotein cholesterol
IBD	Inflammatory bowel disease
MD	Myotonic dystrophy
MD	Menkes disease
MPB	Male pattern baldness
NCBI	The National Center for Biotechnology Information
NIH	National Institutes of Health

OPMD	Oculopharyngeal muscular dystrophy
PCC	Pearson correlation coefficient
PCOS	Polycystic ovary syndrome
PM	Precision medicine
PSSM	Position-Specific Scoring Matrix
RA	Rheumatoid arthritis
RTT	Rett syndrome
TH	Thyroid hormones
VOR	Vestibulo-ocular reflex



# Contents

<b>Chapter 1</b> .....	<b>1</b>
<b>Introduction</b> .....	<b>1</b>
1.1 <i>Application of biomarkers</i> .....	3
1.2 <i>Research objective</i> .....	4
1.3 <i>Outlines</i> .....	6
<b>Chapter 2</b> .....	<b>8</b>
<b>Development of KNApSAcK biomarker database</b> .....	<b>8</b>
2.1 <i>Background</i> .....	8
2.2 <i>KNApSAcK biomarker database</i> .....	9
2.2.1 <i>Features of KNApSAcK biomarker database</i> .....	10
2.3 <i>Human biomarkers database</i> .....	12
2.3.1 <i>Main Menu of Biomarker Database</i> .....	12
2.4 <i>Disease-biomarker classification into 18 disease classes</i> .....	17
<b>Chapter 3</b> .....	<b>22</b>
<b>Materials and methods</b> .....	<b>22</b>
3.1 <i>Classification of disease classes based on biomarkers</i> .....	23
3.2 <i>Classification of disease classes based on protein biomarkers</i> .....	24
3.2.1 <i>Formatting data concerning protein biomarkers</i> .....	24
3.2.2 <i>Recording Accession ID and fasta file download</i> .....	26
3.2.3 <i>Dipeptide composition extraction using ‘protr package’ in R</i> .....	27
3.2.4 <i>Protein biomarker similarity calculation using PCC</i> .....	29

3.2.5 Network visualization by Cytoscape .....	31
3.2.6 Clustering by DPCLUS0 .....	33
3.2.7 Disease classes versus protein clusters Matrix .....	34
<i>3.3 Classification of diseases based on metabolite biomarkers .....</i>	<i>35</i>
3.3.1 Dataset formatting concerning metabolite biomarkers.....	35
3.3.2 Atom pairs fingerprint generation for metabolite biomarkers .....	36
3.3.3 Network of Metabolite biomarkers and clustering .....	39
3.3.4 Visualizing cluster molecule structure.....	40
3.3.5 Disease classes versus metabolite clusters Matrix .....	42
<b>Chapter 4.....</b>	<b>44</b>
<b>Comparison of classification Results and Discussions .....</b>	<b>44</b>
<i>4.1 Hierarchical clustering of 18 disease classes.....</i>	<i>44</i>
<i>4.2 Comparison between Dendrograms.....</i>	<i>46</i>
<i>4.3 Relationship among the 18 disease classes.....</i>	<i>46</i>
<b>Chapter 5.....</b>	<b>60</b>
<b>Conclusions and Future Work.....</b>	<b>60</b>
<b>Reference .....</b>	<b>61</b>
<b>Achievements.....</b>	<b>74</b>
<b>Appendices .....</b>	<b>75</b>

## List of Figures

Figure 1.1	Disease biomarkers; Black, Grey, Red and Blue sphere color C, H, O and N respectively. i. Diabetes biomarker (Glucose - $C_6H_{12}O_6$ ) ii. Encephalopathy biomarker (Glycine - $C_2H_5NO_2$ )..	1
Figure 1.2	Personalized medicine using biomarker diagnostics [11]..	3
Figure 1.3	Drug development chain.....	4
Figure 1.4	Graphical outlines of dissertation..	7
Figure 2.1	Main window of KNApSAcK family databases and arrow indicating biomarker icon.....	9
Figure 2.2	Main window of biomarker database..	11
Figure 2.3	Data display based on partial or exact string matching search.....	11
Figure 2.4	Home page of biomarker database..	13
Figure 2.5	Basic table of biomarker of Human biomarker database..	14
Figure 2.6	Basic table of disease of Human biomarker database.....	14
Figure 2.7	User request form of biomarker database..	15
Figure 2.8	Biomarker disease searching from biomarker database.....	16
Figure 2.9	Searching biomarker or disease by typing string..	16
Figure 2.10	Searching result and import data..	17
Figure 2.11	Disease classes, disease-biomarker relations and biomarker feature connectivity.....	18
Figure 2.12	Pie chart showing the relative frequencies protein biomarkers belonging 18 NCBI disease classes.....	20
Figure 2.13	Pie chart showing the relative frequencies Metabolite biomarkers belonging 18 NCBI disease classes.....	21
Figure 3.1	FASTA file download from NCBI.....	26
Figure 3.2	Threshold PCC Value selection..	31

Figure 3.3	Constructing network based on structural similarity between biomarkers; Node is indicating protein and edge is indicating similarity..	32
Figure 3.4	SDF file downloading from NCBI URL.....	37
Figure 3.5	DPClusO Clusters visualization .	38
Figure 4.1	Disease classification dendrogram based on protein biomarkers.....	45
Figure 4.2	Disease classification dendrogram based on metabolite biomarkers. ....	45
Figure 4.3	Venn diagrams showing common disease classes between protein and metabolite biomarker based clusters; 3 magenta circles are the clusters 1, 2, 3 of figure 4.1 and 4 green circles are the clusters 1, 2, 3, 4 of figure 4.2.....	47

## List of Tables

Table 2.1. 18 disease classes and the number of disease-biomarker relations .....	19
Table 3.1. Protein biomarker, Accession ID, related disease and reference	24
Table 3.2. Protein biomarkers and 18 disease classes mapping.....	25
Table 3.3. Protein biomarkers 400-dimensional descriptors.....	28
Table 3.4. Proteins are considered as nodes and value is considered as edge. .....	29
Table 3.5. Correlation value, Unique Protein ID and none pair. ....	30
Table 3.6. 18 disease classes versus protein cluster data matrix. ....	34
Table 3.7. Metabolite biomarkers and 18 NCBI disease classes mapping. ...	35
Table 3.8. PubChem ID and associated information.....	36
Table 3.9. Checking all 0 columns and deleting from datasets.....	38
Table 3.10. Cluster-based molecular structure visualization (10 out of 43).	41
Table 3.11. 18 disease classes versus metabolite cluster data matrix. ....	43
Table 4.1. Clusters of closely related disease classes.....	48
Table 4.2. Surveyed medical literature to verify our findings for group 1...51	
Table 4.3. Surveyed medical literature to verify our findings for group 2...52	
Table 4.4. Surveyed medical literature to verify our findings for group 3...53	
Table 4.5. Surveyed medical literature to verify our findings for group 4...54	
Table 4.6. Surveyed medical literature to verify our findings for group 5...56	
Table 4.7. Surveyed medical literature to verify our findings for group 6...57	
Table 4.8. Surveyed medical literature to verify our findings for group 7...58	

# Chapter 1

## Introduction

A biomarker (short for biological marker) [1] is defined as a biochemical, cellular, gene related, or molecular alteration that is measurable [2] in biological media, such as blood, body fluids, tissues, or cells by which diseases can be identified.

Figure 1.1 shows two examples of biomarkers.

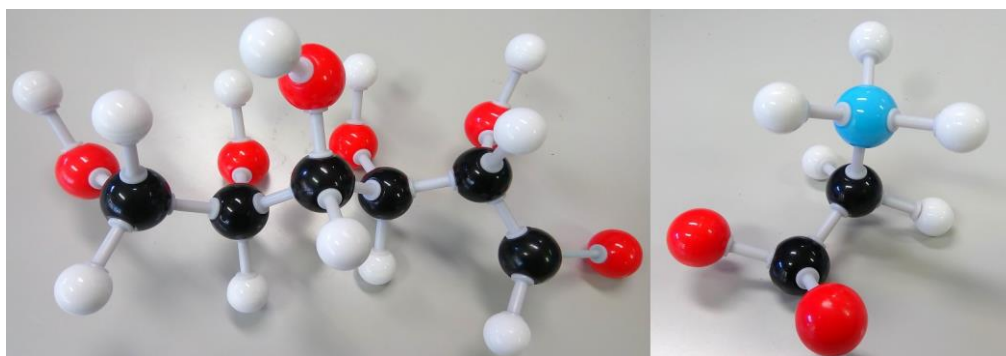


Figure 1.1 Disease biomarkers; Black, Grey, Red and Blue sphere color C, H, O and N respectively. i. Diabetes biomarker (Glucose -  $C_6H_{12}O_6$ ) ii. Encephalopathy biomarker (Glycine -  $C_2H_5NO_2$ ).

A biomarker is an indicator of a disease or disease symptom, which indicates the normal or abnormal condition of a body. Previously, markers of prognosis were considered as biomarkers. Now the concept of biomarker has become widespread. Medical or clinical test results such as from sepsis workups, bone mineral density, x-ray analysis, or EEGs, behavioral/cognitive functioning test results, or growth or other physical measurements or observations such as birth weight, length, fingerprint ridge count, or head circumference also included [3]. For example, human health check includes an assessment of cholesterol, heart rate, blood pressure, triglycerides and fasting glucose levels. Body measurements such as body mass index (BMI), weight and waist-to-hip ratio are routinely used for assessing conditions such as metabolic disorders and obesity. When dealing with exposure assessment, biomarkers are generally classified into three groups:

1. Biomarkers of exposure
2. Biomarkers of effect
3. Biomarkers of susceptibility

### **Biomarkers of Exposure**

Biomarkers of exposure include concentrations of the susceptibility characteristics, exogenous parent chemical, its metabolites, or changes in the body fluids or tissues (e.g., blood lead, etc.) [4]. It reflects only the absence or presence of a substance. It provides the quantification of the amount of substance present in the body and routine surveillance. Biomarkers of exposure are used extensively because they can provide information on the pathway, route and source of exposure.

### **Biomarkers of Effect**

Biomarkers of effect are the quantifiable changes (e.g., biomarkers of early loss of pregnancy, antigen production, benzo-pyrene-DNA adducts, gene suppression, tumor secretions [5] etc.), which shows an exposure to a compound and may show a resulting health effect [6]. Biomarkers measured in tumor tissue are excluded because the disease is diagnosed prior to the bio-measure, and the biomarker is used to ascertain prognosis rather than effect.

### **Biomarkers of Susceptibility**

Biomarkers of susceptibility indicate the detection of a polymorphism such as genetic markers of cancer susceptibility or particular genotype or a natural characteristics of an organism [7]. It may indicate the existence of or the potential for disease or a potential protection against negative health effects. It can assist to define the sensitivity of susceptible as well as critical times when exposures are more harmful. For example, the intensity in asthmatics will indicate how susceptible that

person would be to the respiratory symptom during exposure to brevetoxin, the toxic aerosolize produced during a red tide [8].

## 1.1 Application of biomarkers

The usage of biomarkers is increasing in many health areas such as diagnosing, clinical practice, monitoring disease, ingredient prediction for novel drugs, and precision medicine (PM). In every step of patient care, biomarkers are keeping the vital role. For example, blood sugar may be used to diagnose diabetes whilst HbA1c (glycosylated hemoglobin) monitors blood sugar control [9]. In clinical development, biomarker assays are becoming more important and are used to understand the mechanism of action of a drug as a surrogate marker for monitoring clinical efficacy. These assays can be used to understand the mechanism of action of a drug as a surrogate marker for monitoring clinical efficacy. Prognostic biomarkers are mostly identified from observational data and used to identify patients more likely to have a certain outcome. To identify a predictive biomarker, there should be a comparison system of the treatment to monitor in patients with and without the biomarker.

It has significant importance in PM and is helpful to treat adverse drug reactions [10]. PM is provided to the individual patient based on the disease pattern when no drug affects every patient in the same way (Figure 1.2).

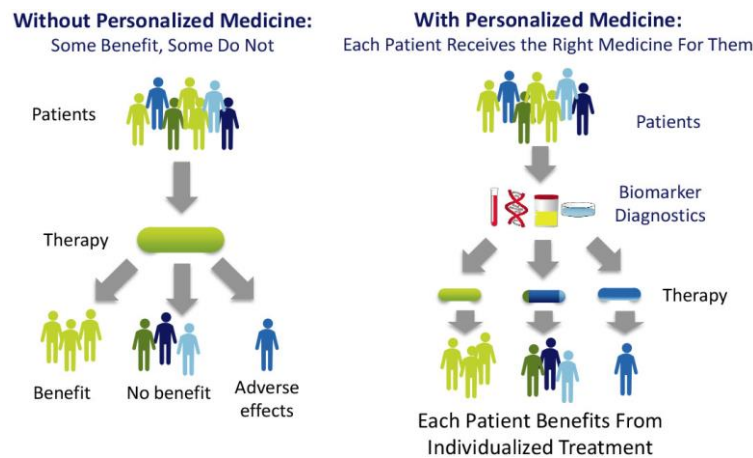


Figure 1.2 Personalized medicine using biomarker diagnostics [11].



According to the PM Coalition, there were 132 personalized medicines on the market in 2016, compared with just five in 2008 and 27% of the new molecular entities approved by the FDA in 2016 can be classified as a PM [12].

Novel drugs are a challenging issue for researcher and companies around the globe. Over the last few years, the pharmaceutical industries have been facing many challenges with increasing time to market, high R & D costs, drug safety, the efficacy of treatment, patient stratification, late attrition and drug resistance [13]. In this situation, biomarkers may be the potential solution for the challenge. In every single step of drug development (Figure 1.3), biomarkers are closely interconnected [14].

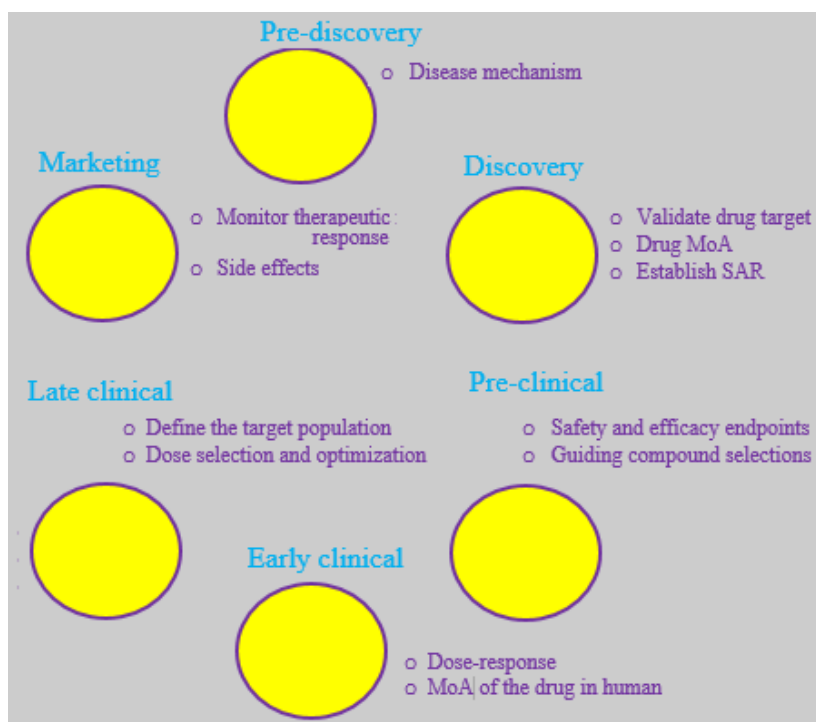


Figure 1.3 Drug development chain.

## 1.2 Research objective

Disease patterns change constantly and identification of accurate biomarkers is also an important challenge. For finding and predicting active medicines, researchers need to read the case studies, mining big data from scattered documents.

It is tough to find the right drug, for the right patient, within the right time period. Recently data management of biomarkers has become a crucial topic because biomarkers are playing significant roles in various disciplines of health research. In this situation, a good quality biomarker database may be a potential solution for the challenge. In medical data-science, biomarker data is going to get much bigger because of the rapid increase of large-scale omics information produced by metabolomics, proteomics, etc. With the growing data volume, the development of a biomarkers database has become a very important issue in the health care field as presently biomarkers are used to detect various human diseases. It is a demand of the time to have an easy access single platform where biomarker data will be stored, and that will provide more accurate and large-scale information as a time-saving tool for drug research. Although some biomarkers databases can be found on the web, they are not comprehensive or free (e.g. GOBIOM [15], BioAgilytix [16], Charles River [17] and upbd [18]). À. Bravo et al. extracted 2803 genes, 2751 diseases and 131,012 disease-biomarker from the literature [19]. The database is a rich resource on biomarker information but all the biomarkers are related to the gene [20]. Srivastava et al. hosted a biomarker database public portal and shared biomarker's title, type, organ(s) [21]. In the database, there is no column of the disease so the database is not so useful for the research. GOBIOM database is the world largest biomarker database that 200000 data sources are included and well-organized database which is developed by Indian 650 scientists research team. The database usage is not free and all data is not for use [15]. BioAgilytix team has developed a biomarker database including 500+ biomarker offerings for a range of disease states. This limited number of the biomarker is not enough for analysis [16]. Institute of Basic Medical Sciences Chinese Academy of Medical Science has published an open-source urinary protein biomarker database contain 1400 protein biomarkers. Blood and urine are important sources of human biomarkers. The data is curated from the literature that has been detected in urine [18].

To mitigate these problems for the research, I have developed a human biomarker database, which can be accessed online at KNApSAcK family Database site (<http://www.knapsackfamily.com/Biomarker/top.php>). Data were collected from the reliable articles. All of the biomarker information sources are linked to valid references. The database may be useful for the protein, metabolite, disease pattern, diseases similarity, novel drug discovery, drug characteristic research and will play a vital role in personalized medicine (PM). Moreover, a researcher can get biomarker information from a single platform instead of searching multiple sources within a short time without literature survey. Apart from the database development, we have examined disease-disease relations at upper hierarchy i.e. at NCBI disease class level. Disease-disease relations provide the clues to understanding disease mechanisms, drug design etc. because similar diseases share similar pathways and genes. We have adopted two approaches based on protein and metabolite biomarkers to classify the diseases and found remarkable consistency between the results we obtained. Our results are useful to find and explain inter disease interactions, disease pathways and novel drugs.

### **1.3 Outlines**

Chapter 2 describes the development of biomarker database and information on accumulated data is explained. We explain how potential user can utilize this database for systematic studies in biomarkers. We have developed two databases in which one for web version and second for internal usages version. This chapter describes the database menu and data type.

Chapter 3 describes the data formatting, disease-biomarker mapping, classifying disease-biomarker relation into disease class, network clustering based on DPCLUS algorithm. We have made two matrices based on protein biomarker cluster and metabolite biomarker cluster to assess the similarity between adopted methods.

In chapter 4, we apply the hierarchical clustering on two matrices. We have evaluated the similarity between the two hierarchical disease class relations by

using two Baker's Gamma correlation coefficient. We made a group by Venn diagrams based on protein and metabolite relations. The resulted groups were then further accessed by surveyed published medical literature.

Finally, Chapter 5 gives concluding remarks of this dissertation.

Figure 1.4 has been shown at a glance of my full research activities and outcome. It will help to understand easily within a very short time and will help to visualize of the rest chapter.

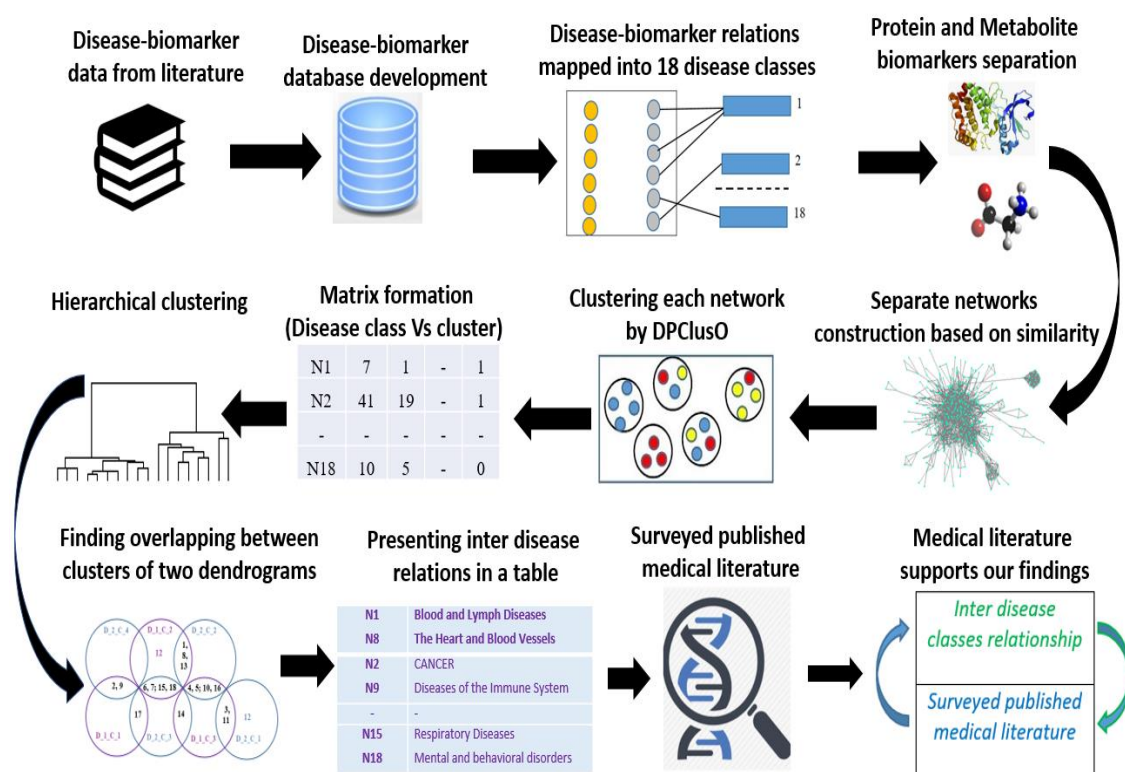


Figure 1.4 Graphical outlines of dissertation.

## Chapter 2

### Development of KNApSAcK biomarker database

Recently biomarker data has become an important topic and it is a demand of the time to have an easy access single platform where biomarker data will be stored, and that will provide more accurate and large-scale information as a time-saving tool for drug research.

#### 2.1 Background

We have accumulated 4539 disease-biomarker associations involving 2181 biomarkers and developed the KNApSAcK human biomarker database. We have developed two database version using same and additional data in which one for web version and second for internal usages version.

- i. KNApSAcK biomarker database
- ii. Human biomarkers database

Web-version has four fields that are biomarker name, disease name, biomarker type(protein/metabolite) and reference. The internal usages version has additional more fields such as FASTA file, InChI Key, Molecular Formula, and Molecular Weight etc. I have developed two versions because web-version has only the main data that will be used according to the research topic. So only, these four fields are sufficient for the user. Addition fields may not be useful for them and may create confusion. Moreover, the additional field may be a burden for the web-service. Internal usages version has additional more field which is useful for our research topic as well as laboratory demand. 'Internal usages version' data may be provided based on a request to the user. Both versions are useful and logical based on our goal.

## 2.2 KNapSack biomarker database

This database is linked with the KNapSack Core (Figure 2.1) and the KNapSack Metabolite Activity Database [22].

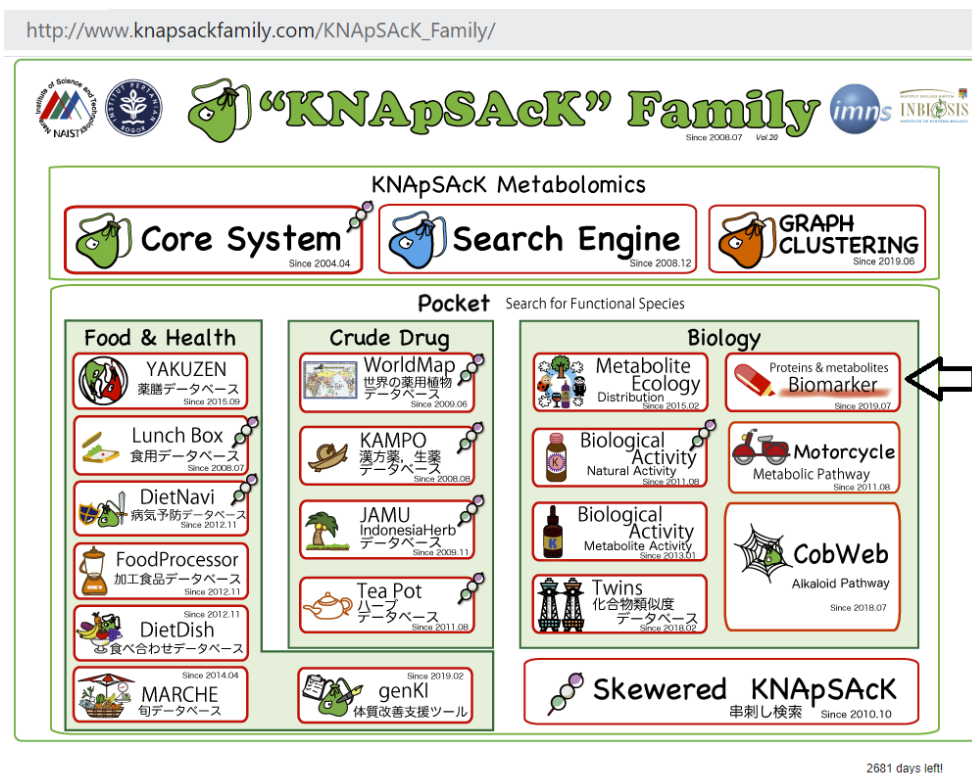


Figure 2.1 Main window of KNapSack family databases and arrow indicating biomarker icon.

Unambiguous and credible biomarkers data was collected from various reliable sources such as NBCI, published patents, proceedings of conferences, approved documents, google scholar, and other recognized documents. The references were hyperlinked to ensure the reliability of the data.

Biomarkers and references were primarily selected according to the following criteria:

1. Biomarker definitions by the National Institutes of Health (NIH) were followed. NIH definition of a biomarker is: “a characteristic that is objectively

measured and evaluated as an indicator of normal biologic processes, pathogenic processes, or pharmacologic responses to a therapeutic intervention” [1].

2. PubMed, Scientific Conferences, and regulatory-approved documents were only considered as the biomarker data source.
3. After the initial selection, the articles were examined by Lab team.
4. Mainly Exposure types of biomarkers were considered


### 2.2.1 Features of KNAPSAcK biomarker database

Our developed human biomarker database is a very rich and up-to-date database where users or researchers can get detailed information about human biomarker data in a single platform. As shown in Figure 2.1, the database can be accessed by browsing to the page of the KNAPSAcK family database ([http://www.knapsackfamily.com/KNAPSAcK\\_Family/](http://www.knapsackfamily.com/KNAPSAcK_Family/)) and then by clicking the ‘biomarker’ icon indicated by an arrow or browsing to the direct database webpage link by <http://www.knapsackfamily.com/Biomarker/top.php> URL. Figure 2.2 shows the main page of the biomarker database. The database has two types of search options called (a) Keyword based data search and (b) All data search. As shown in Figure 2.2, for the Keyword based data search, clicking a radio button can select one of the 4 options (i. All Fields ii. Biomarker iii. Disease and iv. Type). For example, as partially shown in Figure 2.3, after entering the term “pro” and selecting the radio button “All Fields”, if the “List” button is clicked, a Table appears. The database retrieves data based on exact or partial string matching. More detailed descriptions of the search options are available in the instruction manual, which can be downloaded by clicking on the indicated location on the online page of the database (Figure 2.2). Features of our Biomarker Database are as follows:

1. Quick and easy access
2. Online data view without registration
3. Interface with comprehensive search features

4. String searching and intelligent analysis
5. Data sharing with no restriction
6. Dedicated server and routine updates

[knapsackfamily.com/Biomarker/top.php](http://knapsackfamily.com/Biomarker/top.php)



**Information :**

Biomarker project has been performed based on Scientific Literatures. So please check details in references notated when you use information on Biomarker. This web-site is freely available but please cite Biomarker web site (<http://www.knapsackfamily.com/Biomarker/top.php>) when you use publications such as scientific paper and so on.

[Instruction Manual\(English\)](#)

**Keyword Search**

Select by ...

All Fields  Biomarker  Disease  Type

Input data :

**List of all data**

Figure 2.2 Main window of biomarker database.

Number of matched data : 961

Biomarker	Disease	Type	Reference
14-3-3 protein beta/alpha	Renal cell carcinoma	Protein	Shao C, Li M, Li X, Wei L, Zhu L, Yang F, Jia L, Mu Y, Wang J, Guo Z, Zhang D, Yin J, Wang Z, Sun W, Zhang Z, Gao Y. A tool for biomarker discovery in the urinary proteome: a manually curated human and animal urine protein biomarker database. Mol Cell Proteomics. 2011
2-Ethyl-3-hydroxypropionic acid	Schizophrenia	Metabolite	Potential metabolite markers of schizophrenia
3-nitrotyrosine (3-T) (biomarker of peroxynitrite production)	Methamphetamine-induced dopaminergic neurotoxicity	Metabolite	Imam, S.Z., el Yazal, J., Newport, G.D., Itzhak, Y., Cadet, J.L., Slikker, W., Jr., Ali, S.F. Methamphetamine-induced dopaminergic neurotoxicity: role of peroxynitrite and neuroprotective role of antioxidants and peroxynitrite decomposition catalysts. Ann. N.Y. Acad. Sci., vol. 939, 2001
3-oxo-5-alpha-steroid 4-dehydrogenase 2	Prostate disease	Protein	Shao C, Li M, Li X, Wei L, Zhu L, Yang F, Jia L, Mu Y, Wang J, Guo Z, Zhang D, Yin J, Wang Z, Sun W, Zhang Z, Gao Y. A tool for biomarker discovery in the urinary proteome: a manually curated human and animal urine protein biomarker database. Mol Cell Proteomics. 2011
4-vinylcyclohexene (VCH)	reproductive health	Metabolite	Hoyer, P.B., Devine, P.J., Hu, X., Thompson, K.E., Sipes, I.G. Ovarian toxicity of 4-vinylcyclohexene diepoxide: a mechanistic model. Toxicol. Pathol., vol. 29, 2001
5-Oxoproline	Schizophrenia	Metabolite	Potential metabolite markers of schizophrenia

Figure 2.3 Data display based on partial or exact string matching search.



## 2.3 Human biomarkers database

The human biomarker database can be accessed by Local Area Network (LAN) which will provide details information to the users like individual biomarker information, type of biomarkers, source, authors and publication resources and additional biomarker related candidate data. The database is designed to serve three purposes:

1. to search the biomarker details information.
2. to search disease class
3. to show biomarker related additional information

Under the Windows system, all web applications are implemented. NetBeans, Navicat and PHP languages have been used to execute the web interface. The Biomarker database is implemented using the MySQL server. We installed the XAMPP package where Apache as the HTTP service. A database user can import biomarker whole data in Excel formats.

### 2.3.1 Main Menu of Human Biomarker Database

By accessing the 'Human biomarkers database' user can search their expected biomarker data. User will be able to see the default page or home page of our biomarker database. In the home page (Figure 2.4), there is some menus option named Code File, Entry Information, Report. After clicking any individual menu, submenu under every individual menu option will be popped up. Like:

- Code File
  - Biomarker info
  - Disease Info
  - User Info
- Entry Information
  - Biomarker Disease
  - Reference

- Report
  - Search Database
  - Dashboard 1
  - Dashboard 2

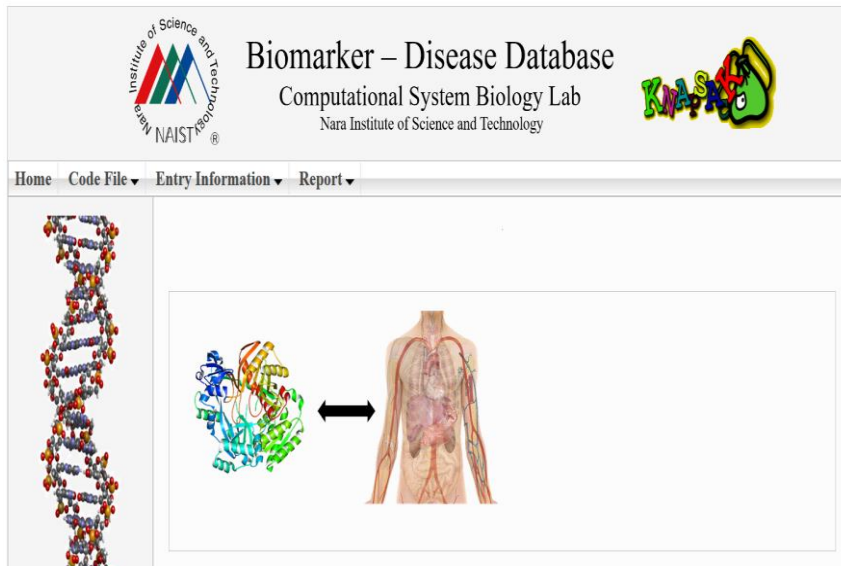


Figure 2.4 Home page of biomarker database.

**Code File:**

Code File menu consists of three tables called biomarker, disease, and user. In Figure 2.5, by clicking the submenu "Biomarker info", a user will see 10 biomarkers per page and clicking "next " button, the user will be able to see more 10 biomarkers and so on. The number of biomarker pages are 307, a user is able to see the whole biomarker of our database. Moreover, by writing full or partial biomarker name or biomarker PubChem ID or both in the specific combo box and clicking the search button a user can specify search terms for any or all of the following fields using advanced search features. By clicking reference URL, a user will get more details information of a specific biomarker.

**Biomarker Information Entry**

Biomarker Name

Biomarker PubchemID

Save Search

FIRST PREV ( Page 34 of 307 ) NEXT LAST

Id	Biomarker Name	PubChem ID
331	2-Aminoadipic acid	-
332	2-Aminobutyrate	6657
333	2-Aminobutyric acid	-
334	2-arachidonoylglycerol	5282280
335	2-chloroacetyl chloride	6377
336	2-Ethyl-3-hydroxypropionic acid	188979
337	2-Hydroxybutyric acid	11266
338	2-Oxoglutarate	51
339	2-S-gluthionyl acetate [C]	11954071
340	21-hydroxylase deficiency	222803

Figure 2.5 Basic table of biomarker of Human biomarker database.

Similarly, in Figure 2.6, by clicking the submenu "Disease Info", the user will see 10 diseases per page and clicking "next" button, the user will be able to see more 10 diseases and so on. The number of disease pages are 137, a user is able to see the whole disease information of our database. Moreover, by writing full or partial disease name or disease alias name or disease description a user can specify search terms for any or all of the following fields using advanced search features. Using 7 combinations of fields, a user can search at the same time.

**Disease Information Entry**

Disease Name

Disease alias

Disease description

Save Search

FIRST PREV ( Page 1 of 137 ) NEXT LAST

Id	Disease Name	Alias	Disease Description
269	Asbestos related cancers	cancer	ref
377	cancer	cancer	-
386	Cancers: liver and lung	cancer	-
862	lung cancer	cancer	ref
863	lung cancers	cancer	ref

Figure 2.6 Basic table of disease of Human biomarker database.

And "User Info" (Figure 2.7) is the third submenu of Code File, a user should fill-up the fields of an individual or organization basic information. There are three types of user 1. admin 2. data entry operator 3.web user. Based on data entry operator and web user request, the admin will provide the different level of access privileges. There are many reasons to fill out a registration form of internal users (e.g. Communications with the user, feedback from the user, support, type of the user). Users are requested to send valuable suggestions and comments as well as any queries about our database and web services. To access this local application, users may use any browser but release after 2015 (recommended) with JavaScript enabled.

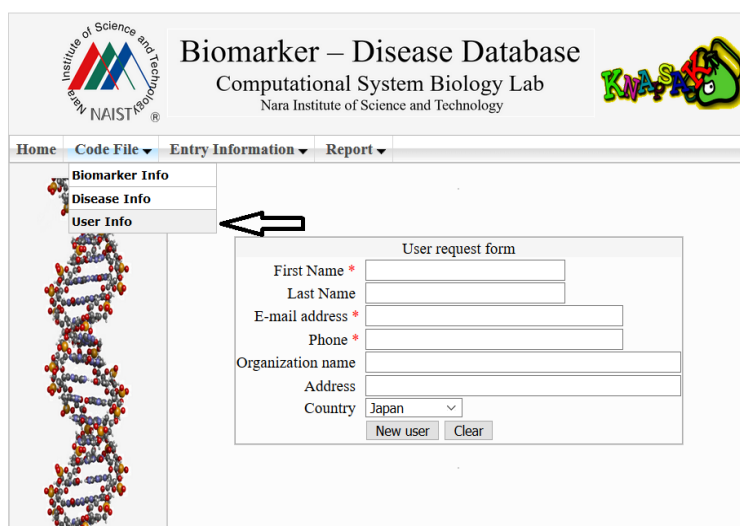


Figure 2.7 User request form of biomarker database.

**Entry Information:**

Entry Information menu consists of two tables called Biomarker Disease and Reference. In Figure 2.8, by clicking the submenu " Biomarker Disease", a user will see 10 biomarkers in the first page and its respective diseases. By clicking "next " button, the user will able to see biomarkers and its respective diseases of next page and so on. The total number of Biomarker Disease page is 464. If a user writes full or partial (minimum 2 characters) disease name or biomarker name or both in the respective left text box, then right combo box will automatically show some options to choose by filtering data using the text box. After selecting the option, clicking the

"search" button will take a little bit time to retrieve result from database. Thus the search engine will show biomarker name with its respective diseases and references.

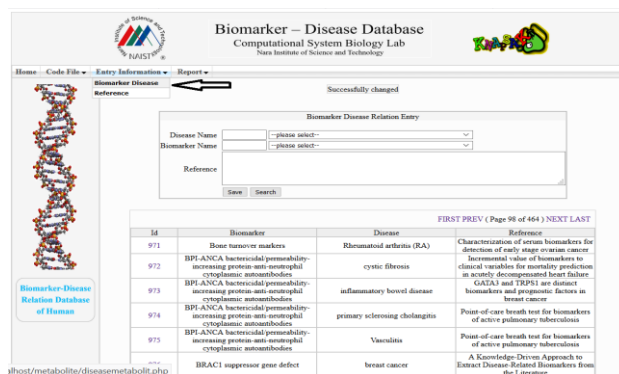


Figure 2.8 Biomarker disease searching from biomarker database.

### Report:

Report is the last and more important menu containing three submenu options named search database, dashboard 1 and dashboard 2 (Figure 2.9). By writing full or partial biomarker name or PubChem ID or disease name or alias name or reference or any combinations (maximum 7) of fields can be searched at the same time.

For example, if a user can remind a partial name of a biomarker i.e. Cy (full name Cyclin) he can easily search on entire data by using the partial mnemonic. If the user writes in Biomarker Name field only "cy" and in Disease Name field "cancer", it is enough for the user to search his expected outcome. Our database will provide specific and details information that string is containing "cy" in biomarker name field and related to cancer disease

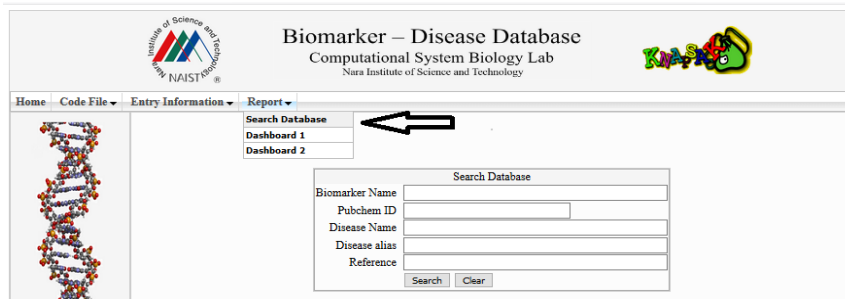


Figure 2.9 Searching biomarker or disease by typing string.

In Figure 2.10, selecting the designed query result, the online user can import biomarker data in Excel formats. For downloading or importing data, the user needs to register in our local intranet service. In short, register users are allowed to download or import data.

The screenshot shows the 'Biomarker - Disease Database' interface. At the top, there are logos for NIAIST and the Computational System Biology Lab. Below the navigation menu, a search form is visible with the following fields: Biomarker Name (containing 'cy'), Pubchem ID, Disease Name (containing 'cancer'), Disease alias, and Reference. A 'Search' button is located below the form. To the left of the search form is a vertical DNA double helix graphic. Below the search form, a table displays search results. The table has five columns: Biomarker, Pubchem ID, Disease, Disease alias, and Reference. The results show two entries for Cytochrome P-450, one associated with 'cancer' and another with 'lung cancer'. Navigation links 'FIRST', 'PREV', 'NEXT', and 'LAST' are visible above the table.

Biomarker	Pubchem ID	Disease	Disease alias	Reference
Cytochrome P-450	121225712	cancer	cancer	Cytochrome P-450 pharmacogenetic and cancer
Cytochrome P-450	121225712	lung cancer	cancer	Cytochrome P-450 pharmacogenetic and cancer

Figure 2.10 Searching result and import data.

## 2.4 Disease-biomarker classification into 18 disease classes

After completion of the database, we have used the disease-biomarker relations for the purpose of disease classification and organized the data for clustering to find disease-disease relations.

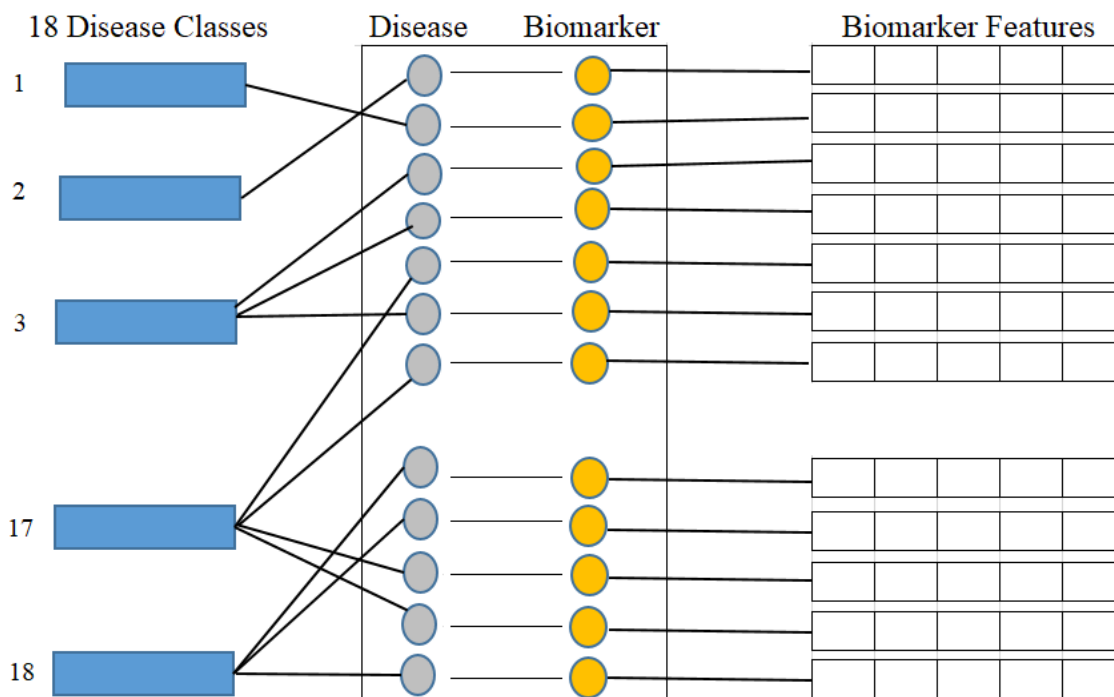


Figure 2.11 Disease classes, disease-biomarker relations and biomarker feature connectivity.

The National Center for Biotechnology Information (NCBI) is a branch of the National Institutes of Health of the United States. NCBI defines and classifies diseases into 16 main classes according to symptoms and disease pattern [23]. In this study, as shown in Table 2.1, we considered 18 disease classes in total, where disease classes N1 to N16 are adopted from the NCBI, and N17 and N18 are determined according to a reference paper [24] and represented by the asterisks symbol in the 'Ref.' column. As shown in Figure 2.11, each biomarker and disease relation is studied and mapped into these 18 disease classes as 'one to many' relations. As illustrated in Figure 2.11, biomarkers are also represented by their structural features. Table 2.1 shows the number, name, and collected disease-biomarker relations in the context of the 18 disease classes. The data contains mainly two types of biomarkers as follows: (1) Protein biomarkers; (2) Chemical or Metabolite biomarkers.

Table 2.1. 18 disease classes and the number of disease-biomarker relations

ID	Ref	Name of Disease Class	Disease-biomarker relations		Total relations
			Protein	Metabolite	
N1	NCBI	Blood and Lymph Diseases	151	73	224
N2	NCBI	CANCER	517	338	855
N3	NCBI	The Digestive System	82	20	102
N4	NCBI	Ear, Nose, and Throat	90	14	104
N5	NCBI	Diseases of the Eye	4	10	14
N6	NCBI	Female-Specific Diseases	172	68	240
N7	NCBI	Glands and Hormones	206	77	283
N8	NCBI	The Heart and Blood Vessels	80	59	139
N9	NCBI	Diseases of the Immune System	262	198	460
N10	NCBI	Male-Specific Diseases	8	7	15
N11	NCBI	Muscle and Bone	40	35	75
N12	NCBI	Neonatal Diseases	88	31	89
N13	NCBI	The Nervous System	71	40	111
N14	NCBI	Nutritional and Metabolic Diseases	86	68	154
N15	NCBI	Respiratory Diseases	268	171	439
N16	NCBI	Skin and Connective Tissue	57	61	118
N17	*	The Urinary System	571	165	736
N18	*	Mental and behavioral disorders	154	197	351

(asterisks refer to a reference paper)



After classifying and counting, belonging protein biomarkers and disease relational percentage in each NCBI disease classes are shown in Figure 2.12.

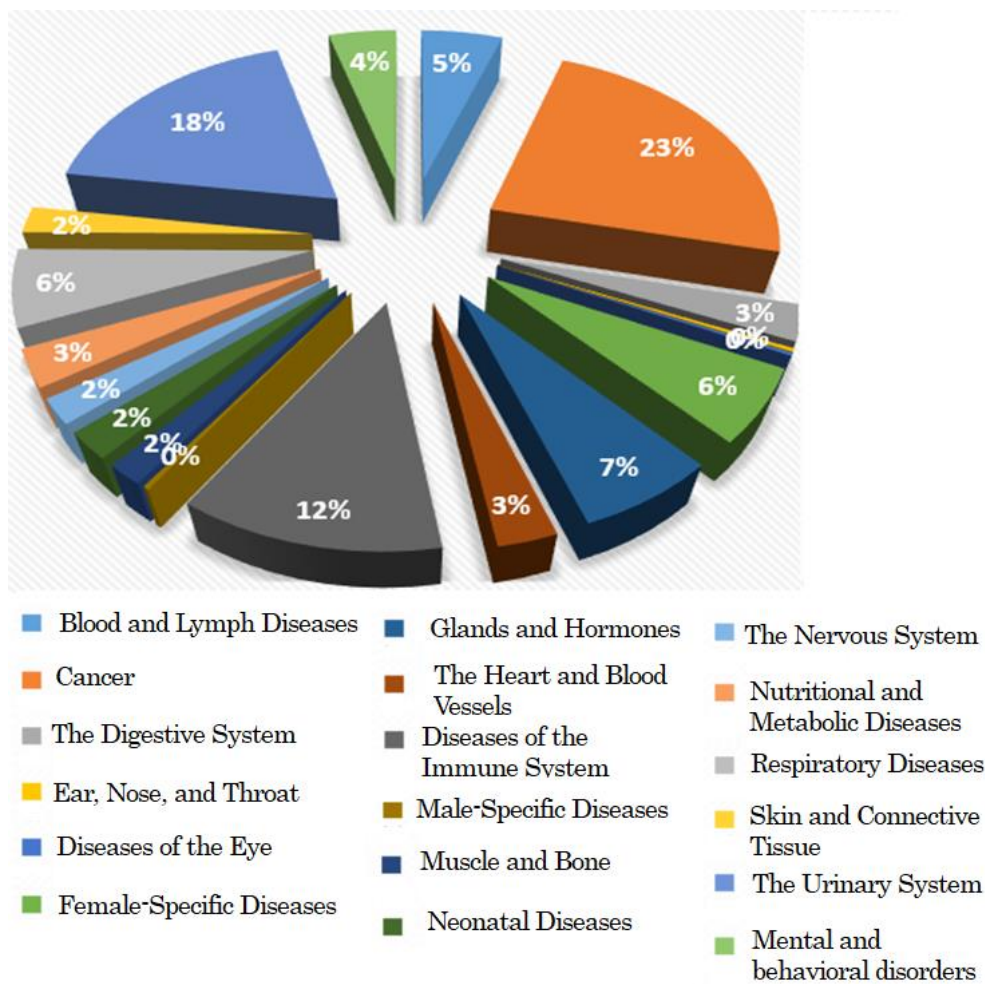


Figure 2.12 Pie chart showing the relative frequencies protein biomarkers belonging 18 NCBI disease classes.

And, the belonging metabolite biomarkers percentage in each NCBI disease classes are shown in Figure 2.13.

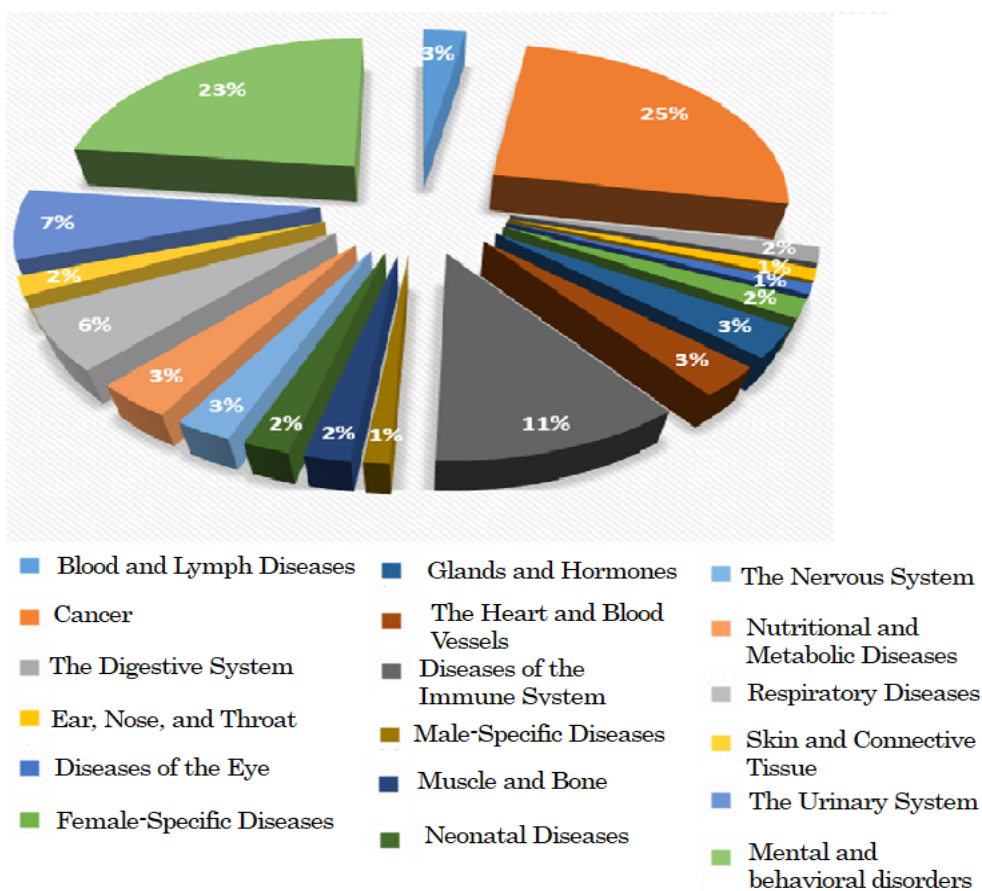


Figure 2.13 Pie chart showing the relative frequencies Metabolite biomarkers belonging 18 NCBI disease classes.

Next, biomarker format files were downloaded from NCBI and similarities between biomarkers were calculated based on the biomarker features. A network was constructed by taking similar biomarker pairs and using a graph clustering algorithm to determine the clusters in the network. Subsequently, we utilized the clusters as characteristic features for disease classes and applied hierarchical clustering to disease classes considering protein and metabolite biomarkers separately (discussed in section 3). We then compared the dendrograms using Baker’s Gamma correlation which is discussed in detail in section 4. Finally, we found significant inter-disease relations among the disease classes that are discussed in section 4.

## Chapter 3

### Materials and methods

Classification enables us to partition a vast expanse of entities that is otherwise disordered into meaningful groups [25]. Disease classification can lead to understanding disease mechanisms, developing drugs, choosing medicines, and guiding medical practice. Disease classification is an old framework which has continued from the seventeenth century until now based on different disease criteria and technology advances [26-27]. Taxonomy is fundamental in biology and originated in the seventeenth century, which uses classification by similar characteristics of individual descriptions among the animal world [28,29,30]. Sydenham notes 1,685 disease symptoms and established a hierarchy in which diseases, symptoms, and the related botany of the treating herbs are linked [31]. In the eighteenth century, de Sauvages clustered diseases by emphasizing a patient symptom-centric structure [32]. In the nineteenth century, laboratory information, clinical signs, radiographs, and electrocardiography were added to recognize disease type and classification. Bertillon recorded the “cause of death,” especially for infection-related deaths, and diseases were classified based on the organ system [33]. In more recent times, the use of high computation facilities and big data involving mRNAs, genes, and metabolites are the basis for classifying diseases, for understanding interactions among diseases, and for prediction of drug ingredients [34, 35, 36].

Marinka et al. show disease-disease associations using molecular data by fusing at systems-level. By using Disease Ontology (DO), they find 107 smaller classes each consisting of closely-related diseases [37]. Lee et al. proposed bipartite human disease association network topology for disease comorbidity where diseases are linked if mutated enzymes associated with them catalyze adjacent metabolic reactions [38]. Kwang et al. developed a human disease network based on gene data and disease phenomena. This work successfully classified 22 types of disease classes

[39]. Emmert et al. construct a human disease and disease gene network for classification, diagnosis, and prediction of disorders and disease genes [40]. Loscalzo et al. proposed complex systems to classify human pathobiology based on the observational correlation between pathological analysis and clinical syndromes. He provides a logical basis for a new approach to classifying human using conventional reductionism of systems biomedicine [41]. Gulbahce et al showed that mutations of susceptibility genes and associated viral infections of human diseases have a direct or indirect association [42]. Arda et al built a multiplex network of 779 human diseases based on genotype-based layer and a phenotype-based layer where a flexible classification of diseases and their molecular underpinnings alongside are manifested [43].

### **3.1 Classification of disease classes based on biomarkers**

In the present work, we are classifying diseases by an upper hierarchy, i.e., based on 18 disease classes. This upper level classification is good for less noisy interpretations of disease relations and avoiding overfitting. Also, Table 2.1 implies that different disease classes are associated with different numbers of biomarkers, i.e., some are linked to many biomarkers whereas others are linked to a small number of biomarkers. Furthermore, the biomarker data we collected is not comprehensive and many new biomarkers will be found in future. Therefore, to compensate for the incompleteness and imbalance of the data, we determined structurally similar clusters of biomarkers and utilized those clusters as features of the disease classes.

The 18 disease classes were classified twice, once based on protein biomarkers, and then again based on metabolite biomarkers. Proteins are biopolymers of amino acids (polypeptides), joined by peptide bonds. Metabolites are amino acids, nucleosides, and the enzyme or coenzyme. Metabolites have various functions, including fuel, structure, signaling, stimulatory and inhibitory effects on enzymes, defense, and interactions with other organisms. Proteins hold the potential to serve as a metabolic fuel source. Proteins are not stored for later use, so excess proteins

are converted into glucose or triglycerides, and used to supply energy or build energy reserves and we know that glucose is a metabolite. Metabolite-protein interactions control a variety of cellular processes, thereby playing a major role in maintaining cellular homeostasis. In the present work, we considered all enzymes as protein biomarkers.

We have adopted two similar procedures separately which are explained in the following sections, and finally the results are compared based on Baker's Gamma correlation.

## **3.2 Classification of disease classes based on protein biomarkers**

A sequence similarity in proteins indicates a functional similarity to a certain extent [44]. A similarity in sequences increases the likelihood of proteins being involved in similar or related signaling and metabolic pathways [45]. Therefore, classification of diseases based on protein biomarkers will obviously be helpful to provide insight into disease mechanisms at the molecular level. When mechanisms are known, that leads to narrowing down potential drug candidates for a disease.

To find disease classifications and inter-disease relations, the protein biomarker data is formatted, mapped to disease classes, protein descriptors are extracted, clustered, and a disease versus clusters matrix is formed. The adopted steps are discussed below.

### **3.2.1 Formatting data concerning protein biomarkers**

As indicated in Table 3.1, the protein biomarkers, respective diseases, and references are arranged in a tabular format. In our data, the number of unique protein biomarkers is 1686 and the protein-disease associations are 3693, because one protein may have associations with multiple diseases.

Table 3.1. Protein biomarker, Accession ID, related disease and reference

Serial No.	Biomarker (Protein)	Accession ID	Disease	Reference
1	Alpha 1-fetoprotein (AFP)	P02773.1	Hepatic cancer	Tatekawa,Y., Asonuma,K., Uemoto,S., Inomata,Y., Tanaka,K.Liver transplantation for biliary atresia associated with malignant hepatic tumors,J.Pediatr.Surg.,vol.36,2001
2	Alpha-2 haptoglobin	AAA88080.1	Schizophrenia	Rohlf,C.Proteomics in neuropsychiatric disorders,Int.J.Neuropsychopharmacol.,vol.4, 2001
3	C-reactive protein	NP_001315986.1	Coronary heart disease	Benzaquen LR;Yu H;Rifai N;High sensitivity C-reactive protein: an emerging role in cardiovascular risk assessment,Crit Rev Clin Lab Sci,vol.39,2002
-	-	-	-	-
3693	Caspase-3	ACR25272.1	Gastric cancer	Chen H, Yang X, Feng Z, Tang R, Ren F, Wei K, Chen G. Prognostic value of Caspase-3 expression in cancers of digestive tract: a meta-analysis and systematic review. Int J Clin Exp Med. 2015;8:10225–10234

The protein biomarkers and disease relations are then classified into the 18 disease classes (mentioned in Table 3.2). As shown in Table 3.2, we created a 1686×18 matrix where rows represent disease-biomarker relations and columns represent 18 disease classes and we put 1 in a cell if the corresponding disease belongs to the corresponding disease class (N1, N2, ....., N18).

Table 3.2. Protein biomarkers and 18 disease classes mapping

<i>Serial No.</i>	<i>Protein Biomarker</i>	<i>Disease</i>	<i>N1</i>	<i>N2</i>	-	<i>N18</i>
1	Alpha 1-fetoprotein (AFP)	Hepatic cancer		1		
2	Alpha-2 haptoglobin	Schizophrenia				1
-	-	-	-	-	-	-

3693	Caspase-3	Gastric cancer		1		
------	-----------	----------------	--	---	--	--

### 3.2.2 Recording Accession ID and fasta file download

The Accession ID is the unique ID of a Fasta file which contains the linear sequence of amino acids within a protein. For our 1686 protein biomarkers, we collected the Accession IDs from the NCBI URL by manual searching [46]. Before recording the Accession IDs, the protein biomarker names in our dataset and the NCBI names were checked carefully for exact matches. Using the Accession IDs, the FASTA files corresponding to protein biomarkers are downloaded (Figure 3.1) from <https://www.ncbi.nlm.nih.gov/protein/> by using NetBeans IDE 8.2 and the JAVA programming language.

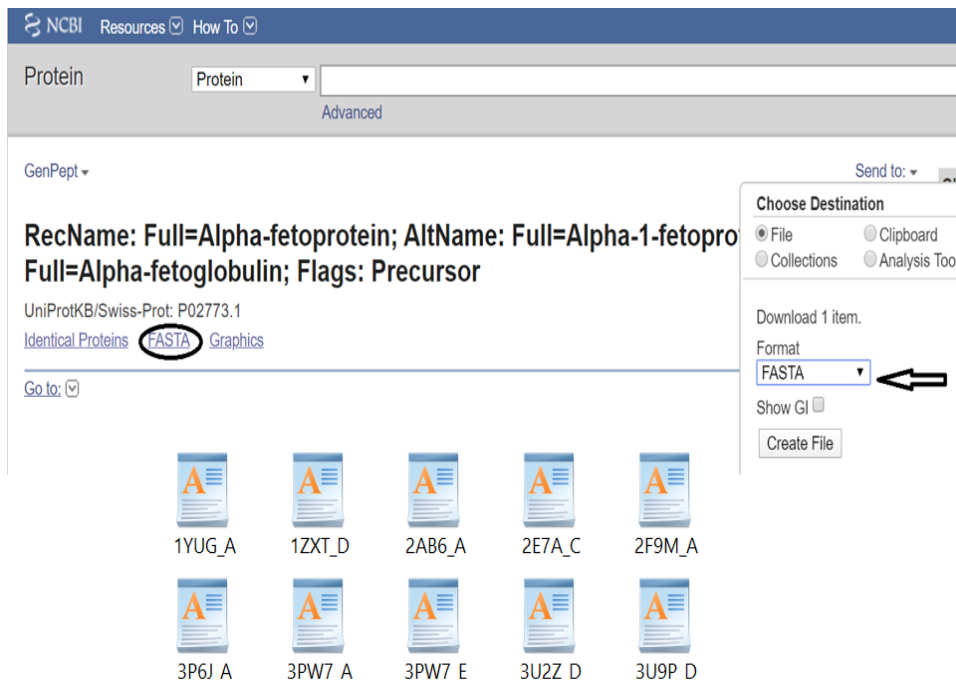


Figure 3.1 FASTA file download from NCBI.

FASTA files are stored in a searchable descriptor database as a list object. In biochemistry and bioinformatics, a FASTA file corresponding to a protein is a text-

based format for representing amino acid (protein) sequences, in which amino acids are represented using single-letter codes. The FASTA format is easy to manipulate and parse the sequences using text-processing tools such as the R programming language, Python, Perl and Ruby. The linear sequence of amino acids is called the primary structure of a protein. Proteins are made of versatile sequences of twenty types of natural amino acids. To represent the twenty amino acids named Alanine (A), Arginine (R), Asparagine (N), Aspartic acid (D), Cysteine (C), Glutamic acid (E), Glutamine (Q), Glycine (G), Histidine (H), Isoleucine (I), Leucine (L), Lysine (K), Methionine (M), Phenylalanine (F), Proline (P), Serine (S), Threonine (T), Tryptophan (W), Tyrosine (Y), and Valine (V), three-letter codes or single letter codes are used [47]. FASTA files are sequences of these twenty amino acids' single letter codes. The “protcheck(x)” function in the "protr" package is used to check the authenticity of FASTA files and 30 FASTA files were deleted from the protein biomarker list before generating the descriptors.

### 3.2.3 Dipeptide composition extraction using ‘protr package’ in R

In the R language, the protr package [48] is a unique and comprehensive toolkit which is used for generating various numerical representation schemes of protein sequences. It is extensively utilized in chemogenomics and bioinformatics research. Amino acid composition, conjoint raid, autocorrelation, quasi-sequence order, Composition, Transition and Distribution (CTD), profile-based descriptors derived by Position-Specific Scoring Matrix (PSSM), and pseudo amino acid composition are all included in protr as a common used descriptors list. The protein sequence descriptors function named extractX() is used for Amino Acid Composition Descriptor in the protr package where X stands for a descriptor name. There are three Amino Acid Composition Descriptors in the protr package as follows: (i) Amino acid composition, (ii) Dipeptide composition, (iii) Tripeptide composition. We examined all three types of compositions for the protein biomarkers dataset to choose the best one for this study. We have found that Dipeptide composition is a better descriptor than the other



two descriptors. Also, many other studies previously utilized dipeptide compositions to measure the structural similarity between proteins [49,50,51]. Amino acid composition gives the percentages of individual amino acids within the protein that does not contain any information related to sequence pattern and Tripeptide composition descriptor results in zero for most attributes. Finally, Dipeptide composition descriptors for the 1656 protein biomarkers as a 400-dimensional matrix (Table 3.3) was calculated by using the function named extractDC(). This is defined as:

$$f(r,s) = \frac{N_{rs}}{N-1} \quad r,s = 1,2,3, \dots, 20$$

where  $N_{rs}$  is the number of dipeptides represented by amino acid type 'r' and type 's' and N is the length of the sequence.

Table 3.3. Protein biomarkers 400-dimensional descriptors

Protein Accession ID	AA	RA	NA	DA	CA	-	VV
1AT3_B.fasta	0.02032 5	0.01219 5	0	0.00406 5	0.00406 5	-	0.00406 5
1BVK_D.fast a	0	0.00934 6	0	0	0	-	0
1C8P_A.fasta	0	0	0.00990 1	0	0	-	0
1CCD_A.fast a	0.01315 8	0	0.01315 8	0	0	-	0
1DE0_B.fast a	0.01041 7	0.00694 4			0.00694 4	-	0.01041 7
-	-	-	-	-	-	-	-
1E6O_L.fasta	0.01421 8	0.00473 9	0	0.00947 9	0	-	0.00473 9

### 3.2.4 Protein biomarker similarity calculation using PCC

For calculating the structure-based similarity [52, 53] between protein biomarkers, we utilized the Pearson correlation coefficient (PCC) based on 400 dimensional descriptors. PCC was calculated using the following equation:

$$corr(X, Y) = \frac{\sum_{i=1}^l (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^l (X_i - \bar{X})^2 \sum_{i=1}^l (Y_i - \bar{Y})^2}}$$

where  $X$  and  $Y$  are protein accession IDs and  $X_i$ ,  $Y_i$  are the weights of the  $i^{th}$  descriptor;  $\bar{X}$ ,  $\bar{Y}$  are the corresponding means;  $l$  is the descriptor size. The PCC similarity ranges between +1 to -1, where 1 is positive linear correlation, 0 is no linear correlation, and -1 is negative linear correlation. The number of the protein biomarker  $P = 1656$ , so the total number of similarity pairs are  $(P(P-1)/2) = (1656(1656 - 1)/2) = 1370340$ . We are interested in highly positive correlations and therefore, to reduce the computation time protein pairs with correlation values above 0.4 (67865 pairs) are saved in a file using the R programming language. Table 3.4 shows the top 13 rows where column 1<sup>st</sup> and 2<sup>nd</sup> are protein and 3<sup>rd</sup> column is their similarity value.

Table 3.4. Proteins are considered as nodes and value is considered as edge

Protein Node 1	Protein Node 2	Similarity Value
P02452.5	NP_000079.2	0.999961512
P25054.2	AAA03586.1	0.999917487
P05362.2	AAP36234.1	0.999604779
P22626.2	NP_001098083.1	0.998974485
NP_000578.2	BAJ20698.1	0.998853598
P01920.2	AAB41231.1	0.998466191
Q9BUR4.1	AKI70503.1	0.998392605
P06727.3	NP_000473.2	0.998267179
NP_001295327.1	AAA51963.1	0.998206585

P06731.3	NP_001295327.1	0.998199544
NP_776433.1	BAA21517.1	0.997961345
P06731.3	AAA51963.1	0.997944752
P35222.1	BAC87743.1	0.997786537

Protein biomarker pairs are sorted in descending order and next, counts of biomarkers and protein pairs are plotted with respect to PCC to find the optimum PCC value for this study (Table 3.5).

Table 3.5. Correlation value, Unique Protein ID and none pair

PCC	Unique Protein ID	Pair
0.4	1426	67865
0.45	1221	27874
0.5	1089	14507
0.55	850	5223
0.6	702	2565
0.65	523	981
0.7	424	614
0.75	319	371
0.8	261	263
0.85	185	166
0.9	138	111
0.95	79	60

In Figure 3.2, the number of unique protein biomarkers (1426 to 79) and the number of pairs (67865 to 60) are plotted against the PCC values (0.40 - 0.95) where the numbers along the left vertical axis indicate protein biomarkers and those along the right vertical axis indicate the number of pairs. We observe that at 0.6 the slope

of the curve showing protein pairs is very low. Moreover, based on other studies, the PCC value 0.6 can be considered as a reasonably good correlation similarity [24] [54]. Therefore, empirically, we selected 0.6 as the PCC threshold in this study. The number of protein pairs having PCC more than 0.6 is 2565 which contains 702 unique protein biomarkers (42% of total).

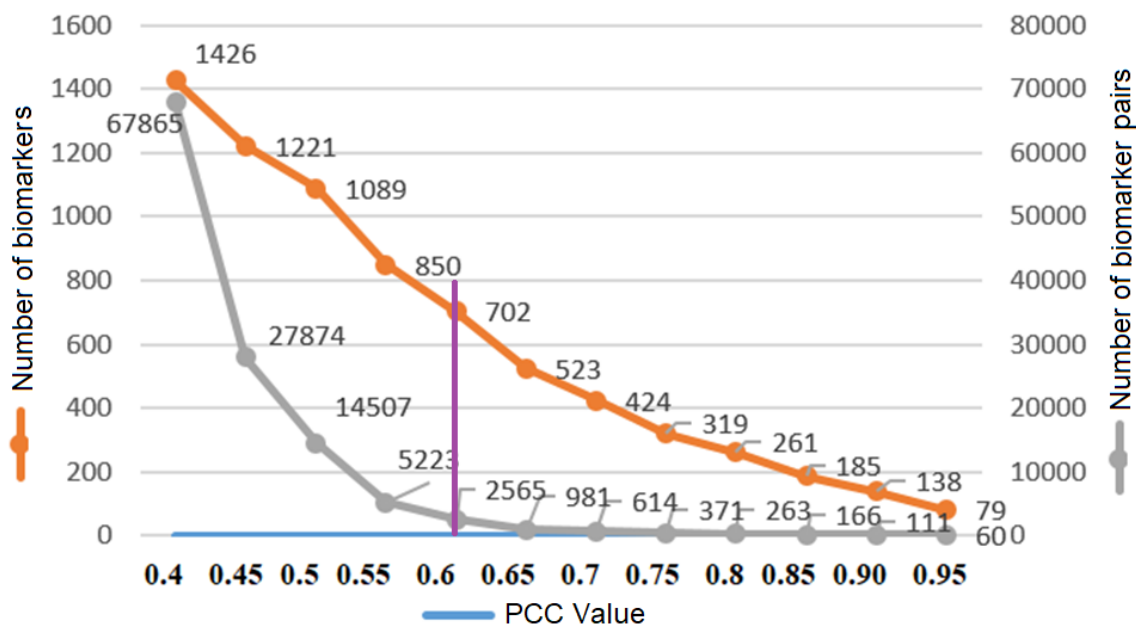


Figure 3.2 Threshold PCC Value selection.

### 3.2.5 Network visualization by Cytoscape

Network analysis is a comprehensive term used in different fields of studies such as biology, economics, sociology, geography, social psychology, business study, etc. Nowadays almost every research where a significant amount of relational data is needed to analyze uses network analysis to find out the kernel information. Conventional database structure like a primary-foreign key relationship or even self-reference relationship is not adequate to represent this heterogeneous relationship of a network. Graphical representation of network data using edge set, vertex set is a very useful way to analyze and understand this relationship. Scientists are trying

to find many important motifs from the network by applying different simple clustering methods.

We visualized the structural similarity based network of protein biomarkers using Cytoscape [55]. Figure 3.3 shows the network, consisting of 2565 protein biomarker pairs having PCC more than 0.6. In the network shown in Figure 3.3, a node is a protein biomarker and an edge represents structural similarity in terms of PCC.

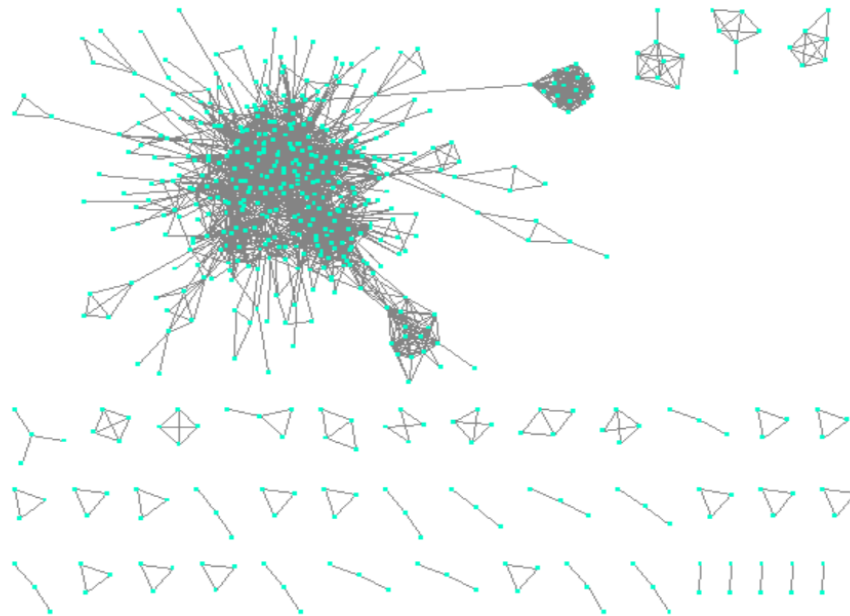


Figure 3.3 Constructing network based on structural similarity between biomarkers; Node is indicating protein and edge is indicating similarity.

### 3.2.6 Clustering by DPCLUSO

Clustering is an unsupervised learning method, which is the task of grouping a set of objects into the same group (cluster) based on similarity or distance measures. This technique is important for knowledge discovery and has been applied in many applications such as machine learning, pattern recognition, image analysis, and

bioinformatics [56,57,58]. In this study, we utilized hierarchical clustering and graph clustering methods for clustering biomarkers.

The constructed protein network was clustered by using the graph clustering algorithm DPclusO. DPclusO is a graph clustering algorithm that is used for extracting densely connected nodes as a cluster from a network [59,60]. The DPclusO algorithm was developed for the detection of protein complexes in large interaction networks. DPclusO can be applied to an undirected simple graph  $G = (N, E)$  that has a finite set of nodes  $N$  and a finite set of edges  $E$ . Density and cluster property are two important parameters in this algorithm. Density  $d$  is a real number ranging from 0 to 1 and cluster  $k$  consisting of  $N_k$  nodes then the density  $d_k$  of the cluster is expressed by the following equation:

$$d_k = \frac{E_k}{E_{k_{max}}} = \frac{2E_k N_k}{N_k(N_k - 1)}$$

Here  $E_{k_{max}}$  is maximum possible edges involving  $N_k$  nodes.

The cluster property is defined by

$$cp_{nk} = \frac{|E_{nk}|}{d_k \times |N_k|}$$

Here  $|E_{nk}|$  indicate the number of edges between the cluster  $k$  and a neighboring node  $n$ . The higher value of  $cp_{nk}$  implies that the node  $n$  has higher priority to be the part of cluster  $k$ .

The overlapping coefficient is defined by

$$OV = \frac{|i|^2}{|a| * |b|}$$

We applied DPclusO with the following settings: cluster property CP = 0.5, density  $d = 0.5$  and Overlapping Coefficient OV=0.05. DPclusO generated 242 protein clusters.

### 3.2.7 Disease classes versus protein clusters matrix

For the purpose of classifying diseases, we utilize the structurally similar clusters of protein biomarkers as features. One biomarker may belong to multiple clusters because we have applied the DPclusO algorithm which generates overlapping clusters and one biomarker may be associated with multiple diseases. We have made a matrix where rows represent the 18 disease classes and columns represent clusters (Table 3.6). An element of the matrix is the number of common protein biomarkers associated with the corresponding disease class and the corresponding cluster. The dimensions of this matrix are 18×242. This matrix is used to classify disease classes and the classification dendrogram is shown and discussed in section 4.

Table 3.6. 18 disease classes versus protein cluster data matrix

	P_Cluster1	P_Cluster2	P_Cluster3	-	P_Cluster241	P_Cluster242
NCBI 1	7	1	4	-	0	1
NCBI 2	41	19	24	-	0	1
NCBI 3	3	4	2	-	0	1
NCBI 4	1	0	1	-	0	0
NCBI 5	0	1	0	-	0	0
NCBI 6	7	7	9	-	0	0
NCBI 7	8	6	9	-	0	0
NCBI 8	8	0	0	-	0	0
NCBI 9	28	6	9	-	0	0
NCBI10	1	2	3	-	0	0
NCBI 11	0	1	0	-	0	0
NCBI 12	5	2	2	-	0	0
NCBI 13	7	2	2	-	0	0

NCBI 14	2	6	3	-	0	1
NCBI 15	12	2	6	-	0	0
NCBI 16	3	2	4	-	0	0
NCBI 17	34	15	10	-	1	0
NCBI 18	10	5	3	-	0	0

### 3.3 Classification of diseases based on metabolite biomarkers

Structural similarity in metabolites often results in activity similarity [61,62]. Structurally similar metabolites might be involved in the same or related metabolic pathways. Structurally similar metabolites might be produced by diseases caused by disruptions in similar pathways. Therefore, it is worthwhile to classify diseases based on metabolite biomarkers for revealing molecular level mechanisms and causes behind diseases.

#### 3.3.1 Dataset formatting concerning metabolite biomarkers

Metabolite biomarkers, associated diseases and references are arranged in a tabular format (Table 3.7). In our dataset, the number of unique metabolite biomarkers is 495 and disease-biomarker associations are 846 because one metabolite may be associated with multiple diseases. Disease and biomarker relations are classified into 18 disease classes. The metabolite biomarker dataset is made into an 846×18 Table where rows are the disease-biomarkers relations and columns are the 18 disease classes. We have put 1 in the cell to indicate an association between the corresponding biomarker and the disease class (Table 3.7).



Table 3.7. Metabolite biomarkers and 18 NCBI disease classes mapping

<i>Serial No.</i>	<i>Biomarker Name</i>	<i>Disease</i>	<i>NCBI 1</i>	<i>NCBI 2</i>	-	<i>NCBI 18</i>
1	Alpha-Tocopherol	Skin cancer (melanoma)		1		
2	Benzene	Leukemias	1			
-	-	-	-	-	-	-
846	Serotonin	Schizophrenia				1

### 3.3.2 Atom pairs fingerprint generation for metabolite biomarkers

PubChem IDs (a public repository for information on chemical substances and their biological activities) of metabolite biomarkers are recorded in the dataset and downloaded from <https://pubchem.NCBI.nlm.nih.gov/> URL by using NetBeans IDE 8.2 and the JAVA programming language. InChI Key, Molecular Formula, and Molecular Weight (Table 3.8) of metabolite biomarkers are also recorded as additional data in the dataset from the NCBI URL by manual searching [63].

Table 3.8. PubChem ID and associated information

<i>Serial No.</i>	<i>Biomarker Name</i>	<i>PubChem ID</i>	<i>Molecular Formula</i>	<i>Weight</i>	<i>InChI Key</i>
1	Alpha-Tocopherol	14985	C <sub>29</sub> H <sub>50</sub> O <sub>2</sub>	430.717 g/mol	GVJHHUAWPYX KBD- IEOSBIPESA-N
2	Benzene	241	C <sub>6</sub> H <sub>6</sub>	78.114 g/mol	UHOVQNZJYSO RNB- UHFFFAOYSA-N
-	-	-	-	-	-

495	Serotonin	5202	C <sub>10</sub> H <sub>12</sub> N <sub>2</sub> O	176.219 g/mol	QZAYGJVTNCV MB- UHFFFAOYSA-N
-----	-----------	------	--	---------------	------------------------------------

Before recording the PubChem ID and associated metadata, datasheets biomarker names and NCBI biomarker names are checked carefully for exact matches. SDF files are stored in a searchable descriptor database as a list object (Figure 3.4). SDF provides 2-dimensional (2D) coordinates for each unique compound structure.

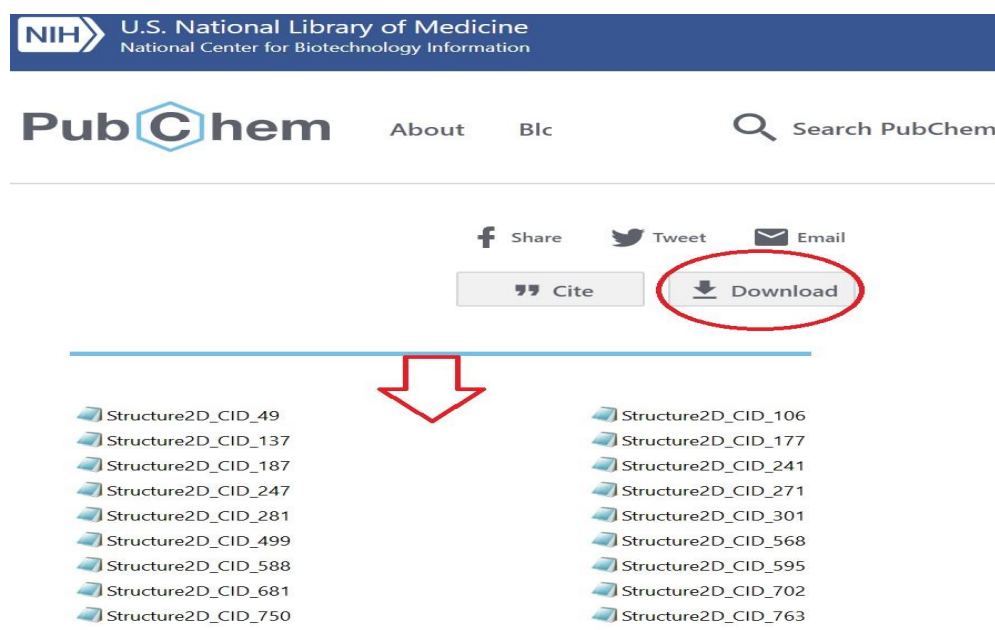


Figure 3.4 SDF file download from NCBI URL.

We have used the ChemmineR (v2.26.0) package [64] to generate atom pair fingerprints from molecular structure description files for the 495 metabolite biomarkers. An atom pair fingerprint is defined by the shortest paths among the non-hydrogen atoms in a molecule. Each path is described by the length of their shortest bond path, the types of atoms in a pair, the non-hydrogen atoms bonded to them, and the number of their pi electrons. There are many molecular fingerprints that are used to represent chemical compounds. Commonly used molecular fingerprints are:

- Atom pairs (AP, 1024 bits)
- PubChem (PubChem, 881 bits)
- CDK (CDK, 1024 bits)
- Extended CDK (Extended, 1024bits),
- Klekota-Roth (KR, 4860 bits),
- MACCS (MACCS, 166 bits),
- Estate (Estate, 79 bits) and
- Substructure (Sub, 307 bits)

In this study, we have used atom pairs fingerprints. By calling the PubChem Compound Identifier (CID) a list and using the functions “sdf2ap” and “desc2fp” with default parameters of ChemmineR, downloaded SDF files are used to generate 1024 bits’ atom pair fingerprints (AP, 1024 bits). Atom pairs fingerprints are binary vectors composed of '0' and '1'.

There are some biomarkers in our list that are not actually compounds. These biomarkers are mainly atoms or ions. These biomarkers show all “0” fingerprints because of no bonding with other atoms. The "Sum" function in Excel is used to check all 0 cell fingerprints and 63 biomarkers are deleted for the subsequent analysis (Table 3.9).

Table 3.9. Checking all 0 columns and deleting from datasets

<i>No.</i>	<i>PubChem</i>	<i>Bit 1</i>	<i>Bit 2</i>	<i>Bit 3</i>	-	<i>Bit 1024</i>
1	14985	0	1	0	-	0
2	241	0	0	1	-	0
<del>3</del>	<del>177</del>	<del>0</del>	<del>0</del>	<del>0</del>	<del>0</del>	<del>0</del>
-	-	-	-	-	-	-
495	5202	0	1	1	0	1

### 3.3.3 Network of Metabolite biomarkers and clustering

The Tanimoto coefficient is utilized for calculating the structure-based similarity [65] between metabolite biomarkers based on 1024-bit atom pair fingerprints. The Tanimoto similarity coefficient ranges between the interval 0 to +1. The number of metabolite biomarkers  $M = 432$ , so the total number of pairs are  $(M(M-1)/2) = (432(432-1)/2) = 93096$ . The Tanimoto similarity between two compounds is calculated by the following equation:

$$Tanimoto_{A,B} = \frac{AB}{A + B - AB}$$

A and B are the number of features that are related to individual compounds and AB is the number of features (or on-bits in the binary fingerprint) common in both compounds. For this study, we have selected the threshold Tanimoto coefficient as 0.85 because metabolite compounds having a Tanimoto coefficient larger than 0.85 represent high similarity. Willett (2014) concluded that the Tanimoto coefficient is standard for similarity searching of 2D fingerprints for different molecular structural similarity measurements and also reported that a Tanimoto coefficient above 0.85 is a good threshold to represent similar structure [66]. We selected 257 metabolite pairs having a Tanimoto similarity more than or equal to 0.85 which contain 30% of the metabolite biomarkers.

In a previous section, we have discussed DPCLUSO and its default parameter setting. By using the graph clustering algorithm DPCLUSO, 257 metabolite biomarker pairs are converted into a network where a node is a metabolite biomarker and the edge represents the Tanimoto coefficient similarity. By keeping the same parameter settings, the network is clustered and DPCLUSO generates 43 clusters (Figure 3.5).

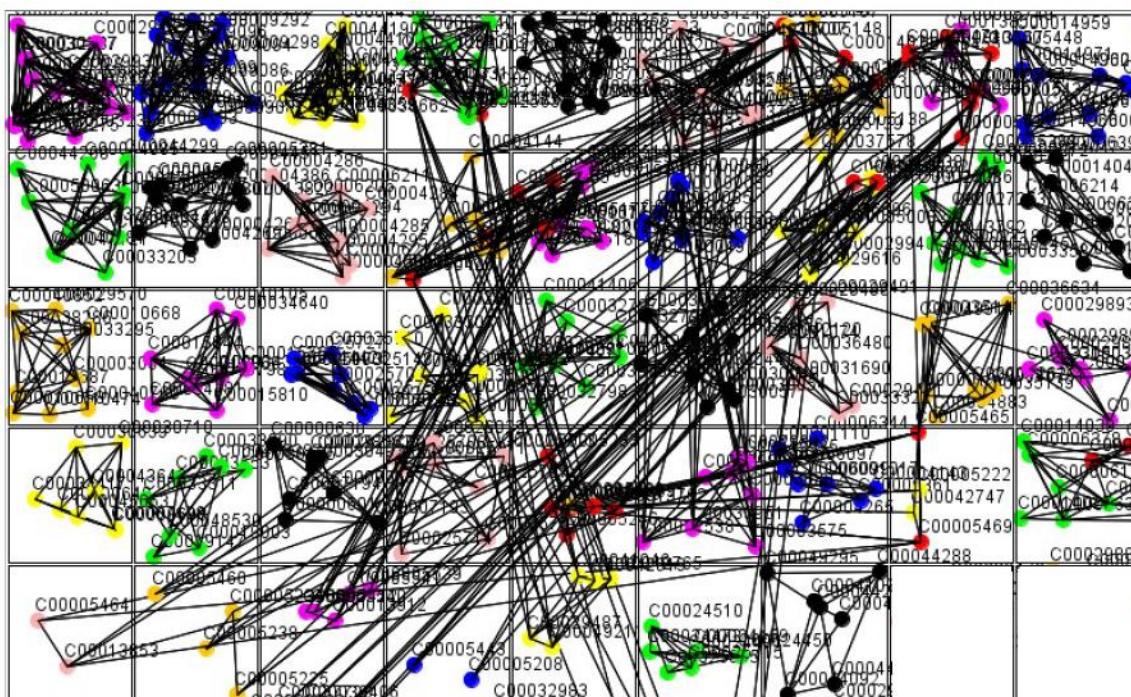


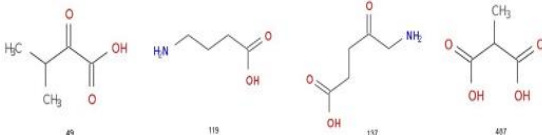
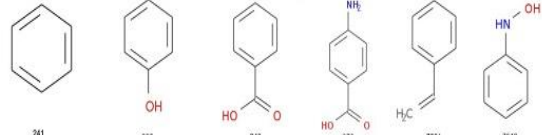
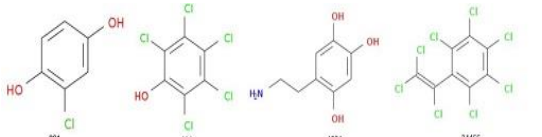
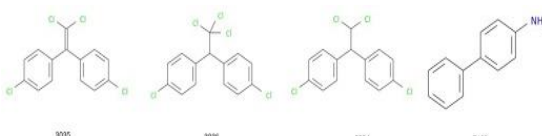
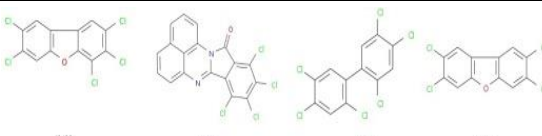
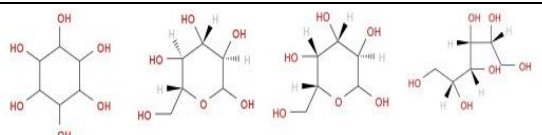
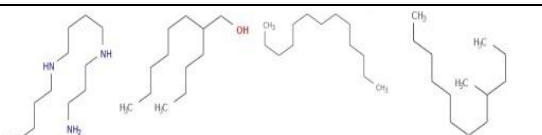
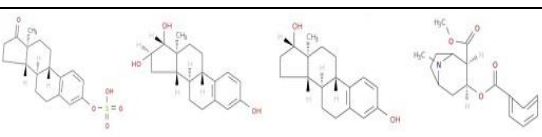
Figure 3.5 DPCLUSO Clusters visualization.

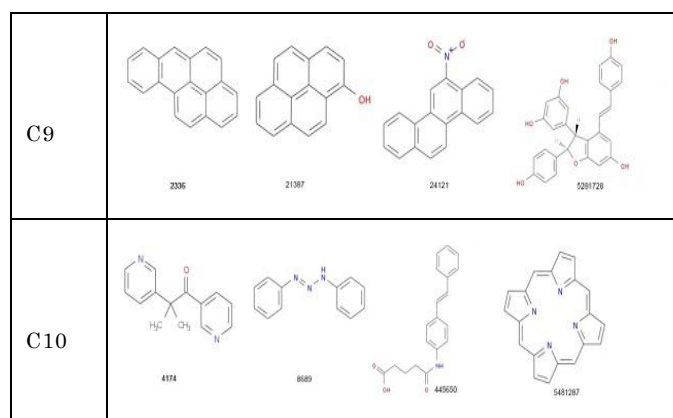
### 3.3.4 Visualizing cluster molecule structure

To verify the similar structure of the metabolite clusters, we have used ChemmineR package within the R language for analyzing and visualizing molecular structures within a cluster. We imported SDF and manually observed the similarity among the compound in the cluster.

"[chemmine.ucr.edu/tools/view\\_job/](http://chemmine.ucr.edu/tools/view_job/)" URL gives a better visualization interface of how generated clusters interact with each other and found that biomarkers belonging to the interacted clusters have similar chemical structure (Table 3.10), which indicates that our clustering software and used fingerprint are perfect selection for this work. Because visualization represent the similar structure of each cluster.

Table 3.10. Cluster-based molecular structure visualization (10 out of 43)

<i>No.</i>	<i>Cluster-based molecular structure visualization</i>
C1	
C2	
C3	
C4	
C5	
C6	
C7	
C8	



### 3.3.5 Disease classes versus metabolite clusters Matrix

We utilize the structurally similar metabolite biomarker clusters as features for classifying the disease classes. In each cluster, related biomarkers of each disease class are counted and recorded. We have made a matrix where rows represent disease classes and columns represent clusters (Table 3.11). An element of the matrix is the number of common metabolite biomarkers associated with the corresponding disease class in the corresponding cluster. This is an  $18 \times 43$  matrix where columns are related to 43 clusters and rows are related to 18 disease classes. The format of the matrix is shown in Table 3.11.

Table 3.11. 18 disease classes versus metabolite cluster data matrix

NCBI	M_Cluster1	M_Cluster2	M_Cluster3	-	M_Cluster43
NCBI 1	3	4	3	-	0
NCBI 2	28	11	8	-	0
NCBI 3	0	0	2	-	0
NCBI 4	0	0	0	-	0
NCBI 5	0	0	0	-	0
NCBI 6	1	1	2	-	0
NCBI 7	1	1	3	-	2
NCBI 8	1	1	0	-	0
NCBI 9	1	1	4	-	0
NCBI 10	1	1	0	-	0
NCBI 11	0	0	1	-	0
NCBI 12	0	0	1	-	0
NCBI 13	0	0	0	-	0
NCBI 14	2	2	1	-	1
NCBI 15	4	2	1	-	0
NCBI 16	2	2	0	-	0
NCBI 17	16	8	6		0
NCBI 18	7	9	7		1



## Chapter 4

### Comparison of classification Results and Discussions

In this chapter, we discuss the hierarchical clustering of disease classes, comparison of dendrograms, and relationships of the 18 disease classes found in our study. We used hierarchical agglomerative clustering method, which starts out by putting each observation into its own separate cluster. It then examines all the distances between all the observations and pairs together the two closest ones to form a new cluster. The process continues until all the observations are included in a single cluster. The result of clustering is usually represented by a dendrogram.

#### 4.1 Hierarchical clustering of 18 disease classes

We applied hierarchical clustering for classifying diseases utilizing the disease classes versus biomarker clusters matrices (Table 3.6 and Table 3.11). We have chosen the hierarchical clustering because it is easy to understand, easy to explain, easy to visualize using dendrograms, and enables distance calculation for better interpretation. For hierarchical clustering, we utilized the Euclidean distance measure given by the following equation:

$$d(i, j) = \sqrt{\sum_{k=1}^n (M_{ik} - M_{jk})^2}$$

Here,  $d(i, j)$ , is the distance between  $i^{th}$  and  $j^{th}$  disease classes and  $M_{ik}$ ,  $M_{jk}$  are the elements of the disease classes versus biomarker clusters matrices. There are several different methods of hierarchical clustering such as Ward's method, single, median, complete, average, and centroid linkage methods depending on how the distance between clusters is measured. We examined all those methods and got almost the same results. Finally, Ward's hierarchical clustering is applied [67] because it is considered a better approach [68,69] and was applied in many other studies. We

applied hierarchical clustering to Table 3.6 and Table 3.11 which are prepared based on protein and metabolite biomarkers respectively. Figure 4.1 and Figure 4.2 shows the disease classification dendrograms respectively based on protein and metabolite biomarkers.

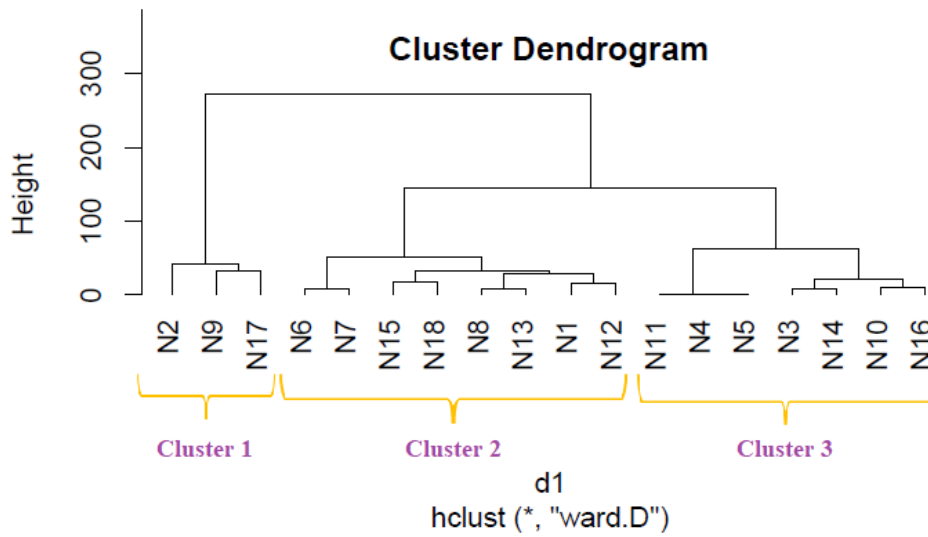


Figure 4.1 Disease classification dendrogram based on protein biomarkers.

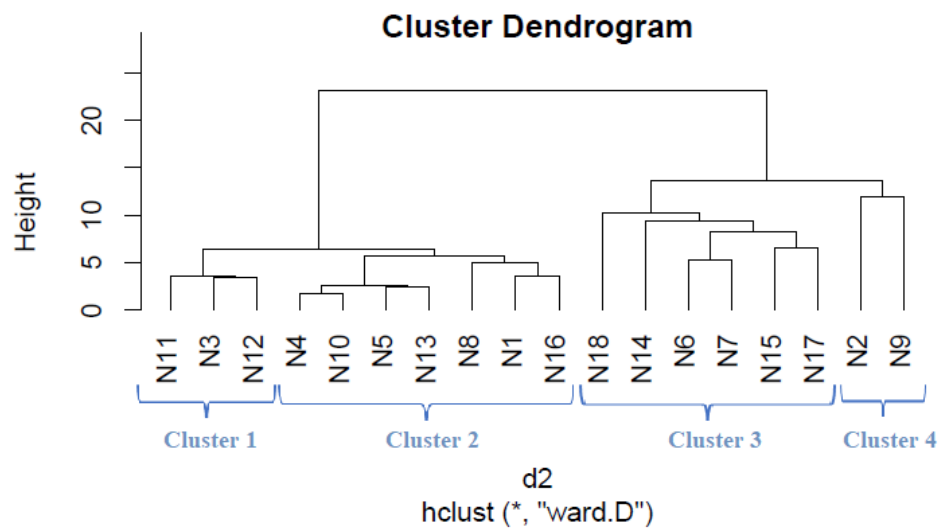


Figure 4.2 Disease classification dendrogram based on metabolite biomarkers.

## 4.2 Comparison between Dendrograms

A dendrogram represents a tree diagram and can display relationships among various objects. We have produced two dendrograms corresponding to two types of biomarkers, i.e., protein and metabolite biomarkers (Figure 4.1 and 4.2). We compared the similarity between the dendrograms using Baker's Gamma correlation coefficient. We observed the highest similarity corresponding to threshold height 3. Baker's Gamma coefficient (Bk) is the calculation of the Mallows-Fowlkes index for a series of  $k$  cuts for global comparison of two dendrogram trees [70,71]. A higher value for the Mallows-Fowlkes index means a greater similarity between the benchmark classifications and the clusters. Baker's Gamma coefficient (Bk) is an external evaluation method to determine the similarity between two hierarchical clustering or a benchmark classification or a clustering.  $k$  is the desired integer number of cluster groups. To compare our produced dendrograms, we used Baker's Gamma correlation coefficient calculated by "dendextend version 1.3.0" package [72] in the R Language. In this work, we obtained the best coefficient for  $k=3$ . "Bk(hc1, hc2, k = 3)" function is executed to measure the similarity between two produced dendrograms and the resulting coefficient is 0.4971546 which indicates a very high similarity between two trees. From this high similarity, it can be concluded that in the context of biomarkers, for most diseases the inter-disease relations are similar both at the protein level and at the metabolite level. This finding is helpful for understanding the molecular mechanisms of the diseases and narrowing down potential drug candidates for a disease.

## 4.3 Relationship among the 18 disease classes

The Baker's Gamma correlation coefficient value 0.4971546 implies that there is a high similarity between the dendrograms of Figure 4.1 and Figure 4.2. We empirically selected 3 and 4 clusters in the dendrograms of Figure 4.1 and Figure 4.2 respectively giving priority to the branching of the dendrogram trees. Figure 4.3 is

drawn based on figures 4.1 and 4.2 showing the common diseases between clusters. In Figure 10, 3 magenta circles are the clusters 1, 2, 3 of Figure 4.1 and 4 green circles are the clusters 1, 2, 3, 4 of Figure 4.2. The disease class IDs that are common between protein and metabolite biomarker based clusters are shown in Figure 4.3. The disease classes included in any cluster of Figure 4.1 can be considered to have a similar mechanism at the protein level and the disease classes included in any cluster of Figure 4.2 can be considered to have a similar mechanism at the metabolite level. Considering the common disease classes between the two sets of clusters (Figure 4.3) and further examining the nearness of the diseases in the dendrograms (Figure 4.1, 4.2) we finally summarize the closely related disease classes as shown in Table 4.1.

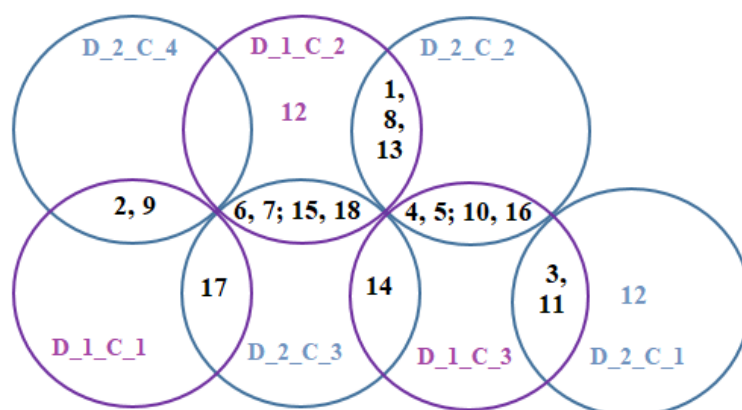


Figure 4.3 Venn diagrams showing common disease classes between protein and metabolite biomarker based clusters; 3 magenta circles are the clusters 1, 2, 3 of Figure 4.1 and 4 green circles are the clusters 1, 2, 3, 4 of Figure 4.2

N12, N14 and N17 are not included in Table 4.1, meaning that they are not similar to any other disease class at both the protein and metabolite level according to our study.

Table 4.1. Groups of closely related disease classes

<i>Group ID</i>	<i>Disease classes ID</i>	<i>Name of Disease Classes</i>
1	N1	Blood and Lymph Diseases
	N8	The Heart and Blood Vessels
	N13	The Nervous System
2	N2	CANCER
	N9	Diseases of the Immune System
3	N3	The Digestive System
	N11	Muscle and Bone
4	N4	Ear, Nose, and Throat
	N5	Diseases of the Eye
5	N6	Female-Specific Diseases
	N7	Glands and Hormones
6	N10	Male-Specific Diseases
	N16	Skin and Connective Tissue
7	N15	Respiratory Diseases
	N18	Mental and behavioral disorders

Therefore, in the context of biomarkers, it can be concluded that a few diseases belong to different groups at the protein level compared to their cohesion at the metabolite level. However, most disease classes that are similar at the protein level are also similar at the metabolite level. Moreover, we studied the PPI network of different literature and inter disease relation based on gene biomarkers and mRNA. That literature working procedure also support our work procedure and support our outcome. Goh et al. collected gene-disease associations data from OMIM database and build a network that is linked with the known disorder-gene. The constructed

human disease network is converted to the cluster and the cluster nodes denoting the disease class based on known disorder–gene and the unknown gene type has a higher likelihood of known type gene [73]. Jimenez et. al classified 1,000 documented disease genes based on function and found striking correlations that which enhanced integration of medicine [74]. Rual et al. provided a human binary PPI network and explained the genotype-phenotype relationships but the interpretation is not completed and insufficient. Liu et al. identified potential biomarkers associated with Basal Cell Carcinoma(BCC). They used the Gene Expression Omnibus (GEO) database of NCBI and built a protein-protein (PPI) network by MOCODE, then analyzed pathway enrichment and ROC. Using KOBAS tool and KEGG database used for gene mapping. Finally, concluded 9 potential genes for BCC from 804 DEGs [75]. Rong et al. detected 20 proteins that are associated with glioblastoma from the PPI network by using TCGA, GeoDiver, GEO2R database and analyzing Cytoscape software, PCC, log2-fold and they mentioned TP53 is the most significant protein [76]. Above procedure also support our adopted procedure.

We have surveyed published medical literature to verify evidence to support our findings which are discussed below in terms of the seven clusters.

**Group 1:** Anemia is often connected to heart disease because the heart must pump more blood to make up oxygen through the body, which can cause an enlarged heart or heart failure, high blood pressure and weakening of the heart muscle, rapid or irregular heartbeat (arrhythmia) [77,78]. B-cell chronic lymphocytic leukemia (B-CLL) forms cancer in blood cells and Ataxia telangiectasia (AT) enlarges blood vessels and affects the brain. Approximately 10–20% of B-CLL occurs by ataxia telangiectasia mutated (ATM) gene damage [79]. A low plasma high-density lipoprotein cholesterol (HDL-c) levels in type I Gaucher disease (GD) creates a deficiency of the lysosomal enzyme acid and affects the blood clotting cells [80]. HDL-c is also an important risk factor of atherosclerotic disease because blood vessels

cannot carry oxygen-rich blood to the heart [81]. Cardiovascular QT syndrome disease drug Donepezil is also used for Alzheimer disease patients [82,83]. To predict Alzheimer Disease (AD) risk,  $\beta$ -Amyloid protein 42 and  $\beta$ -Amyloid protein 40 in the blood are used [84]. Amyotrophic lateral sclerosis (ALS) is a loss of upper and lower motor neurons that affects nerve cells in the brain and spinal cord. Low levels of white blood cell are called CD4 positive T-lymphocytopenia CD4+ T. CD4+ T cells play a neuroprotective role in ALS patients [85,86]. Epilepsy is a disorder which causes seizures due to electrical functioning of the brain. High blood sugar (hyperglycemia) and low blood sugar (hypoglycemia) can affect the nerve cells and low blood glucose can result in a seizure [87,88]. The nervous system is composed of the brain, spinal cord, nerves, and ganglia [89]. The brain cannot work efficiently without sufficient oxygen and the blood is the carrier of oxygen in the brain [90]. Anemia, Gaucher disease (GD), B-cell chronic lymphocytic leukemia (B-CLL) belong to "Blood and Lymph Diseases"; enlarged heart or heart failure, high blood pressure and weakening of the heart muscle, rapid or irregular heartbeat (arrhythmia), atherosclerotic disease, QT syndrome belong to "The Heart and Blood Vessels", and Alzheimer Disease, Amyotrophic lateral sclerosis (ALS), Epilepsy belong to "The Nervous System" disease classes. It is noteworthy that cluster 1 in Table 4.1 includes these three disease classes, "Blood and Lymph Diseases," "The Heart and Blood Vessels," and "The Nervous System." Summary of group 1 discussion is shown in (Table 4.2).

Table 4.2. Surveyed medical literature to verify our findings for group 1

Group ID	Disease classes ID	Name of Disease Classes	Relation among inter disease
1	N1	Blood and Lymph Diseases	(N1) - Anemia (N8) - Enlarged heart or heart failure (N8) -High blood pressure and weakening of the heart muscle (N8) - Rapid or irregular heartbeat (arrhythmia)
	N8	The Heart and Blood Vessels	(N1) - B-cell chronic lymphocytic leukemia (B-CLL) (N8) - Ataxia telangiectasia (AT)  (N1) - Gaucher disease (GD) (N8) - High-density lipoprotein cholesterol (HDL-c) (N8) - Atherosclerotic disease
	N13	The Nervous System	(N8) - QT syndrome (N13) - Alzheimer Disease (N13) - Amyotrophic lateral sclerosis (ALS) (N1) - Low levels of white blood (T-lymphocytopenia CD4+ T)  (N13) - Epilepsy (N1) - Blood sugar (hyperglycemia) and low blood sugar (hypoglycemia)

**Group 2:** Diabetes weakens the patient's immune system defenses [91]. Patients with diabetes risk developing cancer because insulin is not properly carrying glucose into cells so the pancreas produces more insulin to control blood glucose levels, as a result, the hormone stimulates cell growth [92]. Diabetes interrupts DNA, and makes the genome unstable that can lead to cancer [93]. Diabetes patients have a higher risk of gastric cancer due to higher reinfection rate of *Helicobacter pylori* (*H. pylori*) [94]. Rheumatoid arthritis (RA) is an autoimmune



disease that affects joints. RA patients have an excess risk of lung-cancer because of immune function [95,96]. Autoimmune poly glandular syndromes (APS) is a genetic autoimmune disease that has disorders of several endocrine glands and immune-cell dysfunction. APS is associated with thyroid cancer and multi-centric papillary carcinoma [97]. Cancer, gastric cancer, and thyroid cancer belong to "Cancer" and Rheumatoid arthritis, Autoimmune poly glandular syndromes belong to "Diseases of the Immune System", disease classes. Notice that cluster 2 of Table 4.1 contains these two disease classes, "Cancer" and "Diseases of the Immune System." Summary of group 2 discussion is shown in (Table 4.3).

Table 4.3. Surveyed medical literature to verify our findings for group 2

Group ID	Disease classes ID	Name of Disease Classes	Relation among inter - disease
2	N1	CANCER	(N2) - Cancer (N9) - Diabetes  (N2) - Gastric cancer (N9) - Diabetes (N2) - Helicobacter pylori (H. pylori)
	N9	Diseases of the Immune System	(N9) - Rheumatoid arthritis (N2) - Lung-cancer  (N9) - Autoimmune poly glandular syndromes (N2) - Thyroid cancer

**Group 3:** Inflammatory bowel disease (IBD) is a chronic inflammation of the digestive tract that causes long-lasting ulcers in the intestine [98]. IBD is linked with bone density and alterations in bone geometry which is called metabolic bone disease (MBD) [99]. Intestinal inflammation and autoimmune associated bone disease are closely connected with hyperactivation of autoreactive CD4 T cells [100]. Cystic fibrosis (CF) is a chronic disease that affects the lungs and digestive system. The body produces mucus that obstructs the pancreas. Cystic fibrosis-related bone disease (CFBD) is a common complication of CF patients [101]. CF patients often

have low bone mineral density (BMD) that causes fractures [102]. Vitamin D plays a vital role in both CF and BMD [103]. Duchenne muscular dystrophy (DMD) is a muscle disorder disease. Gastrointestinal tract (GI) consists of a long tube from our mouth to anus. GI motor function is connected with DMD Patients [104]. Myotonic dystrophy (MD) is progressive muscular weakness and affects many other body functions including the GI system, heart and lungs [105]. Inflammatory bowel disease, Cystic fibrosis, and the Gastrointestinal tract belong to “The Digestive System” while metabolic bone disease (MBD), bone mineral density (BMD), Duchenne muscular dystrophy (DMD), and Myotonic dystrophy (MD) belong to the “Muscle and Bone” disease classes. Therefore, these articles support cluster 3 of Table 4.1, including “The Digestive System” and “Muscle and Bone” disease classes. Summary of group 3 discussion is shown in (Table 4.4).

Table 4.4: Surveyed medical literature to verify our findings for group 3

Group ID	Disease classes ID	Name of Disease Classes	Relation among inter disease
3	N3	The Digestive System	(N3) - Inflammatory bowel disease (N11) - Metabolic bone disease (MBD)
	N11	Muscle and Bone	(N3) - Cystic fibrosis(CF) (N11) - Bone mineral density (BMD) Vitamin D (plays a vital role in both CF and BMD)
			(N3) - Gastrointestinal tract (N11) - Duchenne muscular dystrophy (DMD) (N11) - Myotonic dystrophy (MD)

**Group 4:** Cogan’s syndrome is a rheumatic disorder that most commonly affects the eye and the inner ear. Cogan's syndrome can lead to hearing loss, pain in the eyes, decreased vision, inflammation, and vertigo [106]. The vestibular (inner ear) and eye movements that act to stabilize gaze are intimately connected through the Vestibulo-ocular reflex (VOR). Sometimes ear infections with viral or bacterial

conjunctivitis can spread to the eyes [107]. The eye and nose are linked by the Nasolacrimal apparatus and this nasolacrimal apparatus carries tears from the ocular surface to the nose. In many cases, nose disease can affect the eyes and vice versa. For example, allergic rhinitis is an inflammation of the nose which shows watery eyes' sign [108]. Nasal vestibulitis, Sinus and Nasal polyps' diseases, may cause eye pains because of the tissue around the eye. Moreover, the eyes, nose, and cheekbones have the same drains [109]. Oculopharyngeal muscular dystrophy (OPMD) is a muscle disorder that slowly affects the upper eyelids and the throat [110]. Trachoma is a bacterial infection spread via eye, nose, or throat fluids [111]. The mentioned diseases mostly viral, bacterial and drainage pathway-related diseases, are associated with each other based on published medical articles. These diseases belong to "Ear, Nose, and Throat" and "Diseases of the Eye" disease classes. It is worth mentioning that these two disease classes are included in cluster 4 of Table 4.1. Summary of group 4 discussion is shown in (Table 4.5).

Table 4.5. Surveyed medical literature to verify our findings for group 4

Group ID	Disease classes ID	Name of Disease Classes	Relation among inter disease
4	N4	Ear, Nose, and Throat	(N4 + N5) - Cogan syndrome (common) (N5) - Vestibular (inner eye) (N4) - bacterial conjunctivitis(Trachoma)
	N5	Diseases of the Eye	(N4 + N5) - Nasolacrimal apparatus (carries tears, Nose) (N4) - allergic rhinitis (N4) - Nasal vestibulitis (N4) - Sinus and Nasal polyps (N4 + N5) - eyes, nose and cheekbones are the same drains

**Group 5:** Ovarian cancer begins in the ovaries that are obstructed between estrogen and progesterone hormonal balance and create problems in sexual and reproductive development in women [112,113]. Rett syndrome (RTT) is a genetic brain disorder that occurs primarily in girls within 6 to 18 months of age and causes

a disability of language, coordination, and repetitive movements. Children with RTT directly interfere with thyroid hormones (TH) level [114]. Polycystic ovary syndrome (PCOS) is a hormonal disorder which is associated with irregular menstrual cycles, excess facial boils, and acne [115]. Congenital adrenal hyperplasia (CAH) is a common genetic disorder of steroidogenesis that affects fertility due to steroid 21-hydroxylase (21 OH) deficiency. Steroid hormones play a significant role in reproductive function and sexual development [116]. Hyperthyroidism (overactive thyroid) occurs due to excessive production of the hormone thyroxine by the thyroid gland that causes weight loss and irregular or rapid heartbeat. Graves' disease causes hyperthyroidism. Thyroid disease occurs more often in women than in men [117]. Maternal hyperthyroidism increases the risk of miscarriage, premature birth, and a low birth weight baby [118]. Uterine Fibroids are noncancerous growths of the uterus and Endometriosis is cells outside the uterus [119]. Both are a common cause of hormone imbalance [120]. The above discussions imply that "Female-Specific Diseases" are directly or indirectly related to hormones and responsible for hormonal imbalance. Therefore, "Glands and Hormones" related diseases are more common for women compared to men. Moreover, women are emotional than men because of hormone fluctuations [121]. Interestingly, cluster 5 of Table 4.1 reflects such associations between the "Female-Specific" and "Glands and Hormones" disease classes. Summary of group 5 discussion is shown in (Table 4.6).

Table 4.6. Surveyed medical literature to verify our findings for group 5

Group ID	Disease classes ID	Name of Disease Classes	Relation among inter disease
5	N6	Female-Specific Diseases	(N6 + N7) - Ovarian cancer → bstruct between estrogen and progesterone hormonal balance  (N6 + N7) - Rett syndrome (RTT) →occurs primarily in girls within 6 to 18 months (thyroid hormones)  (N6 + N7) - Polycystic ovary syndrome (PCOS)->hormonal disorder
	N7	Glands and Hormones	(N6 + N7) - Congenital adrenal hyperplasia (CAH)→affects fertility (genetic disorder of steroidogenesis)  (N6 + N7) - Steroid hormones → role in reproductive function. Thyroid disease is more often in women than in men  (N6 + N7) - Uterine Fibroids are noncancerous growths of the uterus and Endometriosis  (N6 + N7) - Women are emotional than men because of hormone fluctuations

**Group 6:** Male pattern baldness (MPB) is hair loss on the scalp, which is the most common cause of hair loss in men [122]. Genes and male sex hormones are mostly responsible for MPB. Moreover, dandruff, scalp skin dryness, and skin diseases like Psoriasis, Allergies and Alopecia Areata are causes of hair loss [123]. Peyronie's disease or penis curvature is a disorder caused by fibrous scar tissue inside the penis. It may cause bent penis, erectile dysfunction and can make sex uncomfortable or impossible [124]. Menkes' disease (MD) is an X-linked recessive disorder caused by mutations in the ATP7A gene [125]. Connective tissue and Progressive neurodegeneration are responsible for peculiar 'kinky' hair. Moreover, copper deficiency in the body, failure to gain weight, growth, and nervous system

deterioration are the main characteristic of MD. Patients with MD are the vast majority in males more than in females [126]. The above mentioned diseases Male Pattern Baldness(MPB) and Peyronie's disease belong to “Male-Specific Diseases” while Menkes’ disease (MD) belongs to the “Skin and Connective Tissue” disease class. We found some 'Male' and 'Tissue' related diseases which are linked with female, blood, muscle, nervous system diseases and so on. But more connections are found within Male and tissue-related diseases. The above statements about diseases in the “Male-Specific” and “Skin and Connective Tissue” disease classes are supported by cluster 6 of Table 4.1. Summary of group 6 discussion is shown in (Table 4.7).

Table 4.7. Surveyed medical literature to verify our findings for group 6

Group ID	Disease classes ID	Name of Disease Classes	Relation among inter disease
6	N10	Male-Specific Diseases	(N10 + N16) - Male pattern baldness(MPB)→hair loss on the scalp  (N10 + N16) - Peyronie's disease or penis curvature → fibrous scar tissue inside the penis
	N16	Skin and Connective Tissue	(N10 + N16) - Menkes disease (MD)→Connective tissue  (N10 + N 16) - Copper deficiency → male than female

**Group 7:** Asthma is a chronic disease of the respiratory system in which airways swell or narrow or produce extra mucus that causes breathing difficulties. Bipolar disorder is a mental disorder that includes lows of depression, mania or hypomania (feeling high) and unusual shifts in mood. Severe asthma is associated with bipolar disorder, anxiety disorders, post-traumatic stress, and severe mental disorder. Asthma and Bipolar disorder share a similar pathophysiology and a patient with asthma has 2.12 times higher risk of Bipolar disorder [127]. Alpha-1 antitrypsin deficiency (A1AD) is a genetic disorder that causes lung and liver disease. Anxiety disorders are a group of mental disorders including anxiety, fear, panic, specific

phobias, agoraphobia, worry about future events, and social anxiety disorder. Emotional and Anxiety disorders are common comorbidities in AATD patients [128]. Schizophrenia is a brain disorder that can cause delusions, hallucinations, and extremely disordered thinking and affects how a person feels, thinks, and behaves. Chronic obstructive pulmonary disease (COPD) is a group of lung diseases that causes breathing difficulties and poor airflow. Schizophrenia is connected with weakened lung function and increases the risk of COPD and pneumonia [129]. Marijuana (Cannabis) and tobacco smoke pollute the lungs and, reduce brain activity and reduce the volume of brain regions. Marijuana addicted people are attacked by both Respiratory and Mental disorders [130]. Asthma, Alpha-1 antitrypsin deficiency, and Chronic obstructive pulmonary disease (COPD) belongs to “Respiratory Diseases” while Bipolar disorder, Anxiety disorders, and Schizophrenia belong to the “Mental and Behavioural disorders” disease classes. Diseases in both these disease classes are very close to each other according to medical research, and our study also grouped them in cluster 7 of Table 4.1. Summary of group 7 is shown in (Table 4.8).

Table 4.8. Surveyed medical literature to verify our findings for group 7

<b>Group ID</b>	<b>Disease classes ID</b>	<b>Name of Disease Classes</b>	<b>Relation among inter disease</b>
7	N15	Respiratory Diseases	(N15) - Asthma (N18) - Respiratory Diseases (N18) - Bipolar disorder
	N18	Mental and behavioral disorders	(N15) - Alpha-1 antitrypsin deficiency (N18) - Anxiety disorders  (N18) - Schizophrenia (N15) - Chronic obstructive pulmonary disease (COPD)

**No group (N12, N14, N17: Neonatal Diseases, Nutritional and Metabolic Diseases, The Urinary System)**

Neonatal diseases are not associated with other disease classes because (1) there are a dramatic physiological transition from fetal to neonatal life, (2) organs are still in their developmental stage, and (3) fetuses are affected by the condition of the uterine environments, such as intrauterine inflammation/ infection [131].

Metabolism is the means by which the body derives energy and synthesizes the other molecules it needs from the fats, carbohydrates and proteins we eat as food, by enzymatic reactions helped by minerals and vitamins. Nutritional and Metabolic Disease will only occur if a critical enzyme is disabled, or if a control mechanism for a metabolic pathway is affected. This disease class is not specifically related to others disease classes but it has relation with others disease classes based on enzyme (23).

The urinary tract consists of the kidneys, ureters, bladder and urethra. kidneys filter our blood, creating urine, which travels through the ureters to the bladder, where it is stored. Urinary disorders include any diseases, disorders or conditions that affect your kidneys, ureters, bladder or urethra, or that affect their function. Urinary is itself a system, so it has been said 'The Urinary System' disease class. Within this system, there are many diseases but this disease class is rarely connected with others diseases or others disease classes. We searched medical literature to show the association with other diseases but didn't get satisfactory outcome [132].



## Chapter 5

### Conclusions and Future Work

In the present study, we have developed a human biomarker database, which can be accessed online at the KNApSAcK family Database site (<http://www.knapsackfamily.com/Biomarker/top.php>). I have collected the biomarker data from reliable articles and verified by our Lab team. All of the biomarker information sources are linked to valid references. The database may be useful for research on proteins, metabolites, disease patterns, disease similarities, novel drug discovery, and drug characteristics, and it will play a vital role in personalized medicine (PM). Moreover, within a short time, without doing a literature survey, a researcher can get biomarker information from a single platform, instead of searching multiple sources. In the developed database, there are 1686 protein and 495 metabolite biomarkers involving, respectively, 3693 and 846 diseases-biomarker associations. Apart from the database development, we have examined disease-disease relations in an upper hierarchy, i.e., at the NCBI disease class level. Disease-disease relations provide clues to understanding disease mechanisms, drug design, etc., because similar diseases share similar pathways and genes. We have adopted two approaches based on protein and metabolite biomarkers to classify the diseases and found a remarkable consistency between the results we obtained. A Baker's Gamma correlation value of 0.4971546 was obtained between the dendrograms generated by the two approaches. We have collected FASTA files of protein biomarkers and SDF files of metabolite biomarkers, then extracted descriptors and fingerprints using the programming language R. We have used the Pearson correlation coefficient (PCC) and Tanimoto coefficient to calculate the similarities between protein and metabolite biomarkers respectively. The network clustering algorithm DPCLUSO and hierarchical clustering were applied to extract associations among 18 disease classes. Finally, we determined 7 groups involving 15 of the 18

disease classes based on disease similarities in both protein and metabolite levels. We thoroughly studied medical literature and gathered substantial evidence to support our findings. To our knowledge, this is one of the first approaches to classify diseases based on biomarkers. Our results are useful to find and explain inter-disease interactions, disease pathways, and novel drugs. It will also help to understand the comorbidity. Comorbidity refers to the simultaneous presence of two or more diseases or medical conditions in a patient. Although sometimes discovered after the principal diagnosis, comorbidities often have been present or developing for some time. Examples include diabetes, heart disease, high blood pressure (hypertension), psychiatric disorders, or substance abuse. Our results are also useful to explain comorbidity.

## **Future Work**

We will accumulate more biomarker data including gene and mRNA and perform a comprehensive analysis of the inter-disease relation in the context of different type of human biomarker. We will integrate various analyzing tools with our KNApSAcK biomarkers databases such as BLAST, DPCLUSO, Cytoscape, Rshiny and mathematical analyzer portal. Also, for protein or metabolite biomarker disease class prediction, we will try to develop a deep learning approach to predict biomarker class based on sequence similarity. It is hoped that the KNApSAcK biomarker Database can be as a reference tool for the users to find information on protein and metabolite biomarker with related disease activities for the application in medical, drug and healthcare industry.

## **Reference**

- [1] K. Strimbu and A. T. Jorge, "What are biomarkers?" *Current opinion in HIV and AIDS* vol. 5,6 pp 463-6, 2010.

- [2] L.E. Walker, D. Janigro, U. Heinemann, R. Riikonen, C. Bernard, M. Patel, “WONOEPP appraisal: Molecular and cellular biomarkers for epilepsy”, 57(9), pp 1354-62, 2016.
- [3] L.J. Orchinik, H.G. Taylor, K.A. Espy, “Cognitive outcomes for extremely preterm/extremely low birth weight children in kindergarten”, *Int Neuropsychol Soc.* 17(6):1067-79, 2011.
- [4] W. P. Watson, A. Mutti, “Role of biomarkers in monitoring exposures to chemicals: present position, future prospects”, *Biomarkers*, 9(3), pp 2 11-42, 2004 May-Jun
- [5] H. Hamouchene, V.M. Arlt, I. Giddings, D.H. Phillips, “Influence of cell cycle on responses of MCF-7 cells to benzo[a]pyrene”, *BMC Genomics*, 12:333. Epub 2011 Jun 29.
- [6] G.F. Nordberg, T. Jin, X. Wu, J. Lu, L. Chen, L. Lei, F. Hong, M. Nordberg, “Prevalence of kidney dysfunction in humans - relationship to cadmium dose, metallothionein, immunological and metabolic factors *Biochimie*”, 91(10):1282-5, 2009 Oct
- [7] I. Iavicoli, V. Leso, P.A. Schulte, “Biomarkers of susceptibility: State of the art and implications for occupational exposure to engineered nanomaterials”, *Toxicol Appl Pharmacol*, 299:112-24, 2015
- [8] L.E. Fleming, B. Kirkpatrick, L.C. Backer, “Aerosolized red-tide toxins (brevetoxins) and asthma”, *Chest*, 131(1):187-94, 2007
- [9] Glycated Hemoglobin (HbA1c): Clinical Applications of a Mathematical Concept. *Acta Inform Med*, 24(4):233-238, 2016
- [10] R. La Russa, V. Fineschi, M. Di Sanzo, V. Gatto, A. Santurro, G. Martini, M. Scopetti, P. Frati, “Personalized Medicine and Adverse Drug Reactions: The Experience of an Italian Teaching Hospital”, 18(3):274-281, 2017.
- [11] <https://chartpack.phrma.org/personal-medicines-in-development-chartpack/a-new-treatment-paradigm/a-new-treatment-paradigm> [accessed last on Nov 20, 2020]

- [12] <https://invivo.pharmaintelligence.informa.com/IV005059/Personalized-Medicine-An-Infographic> [accessed last on Nov 20,, 2020].
- [13] N.J.D. Gower, R.J. Barry, M.R. Edmunds, L.C. Titcomb, A. K. Denniston, "Drug discovery in ophthalmology: past success, present challenges, and future opportunities", *BMC Ophthalmol*, pp 16:11, 2016
- [14] A.B. Halim, "Biomarkers in Drug Development: A Useful Tool but Discrepant Results May Have a Major Impact", *intechopen*, 2011
- [15] <https://www.gobiomdbplus.com/about-us> [accessed last on Nov 20, 2020]
- [16] <https://www.bioagilytix.com/biomarker-menu/> [accessed last on Nov 20, 2020]
- [17] <https://wwwapps.criver.com/BiomarkersDB/> [accessed last on Nov 20, 2020]
- [18] <http://upbd.bmicc.cn/biomarker/web/indexdb> [accessed last on Nov 20, 2020]
- [19] À. Bravo, M. Cases, N. Queralt, F. Sanz, L.I. Furlong, "A knowledge-driven approach to extract disease-related biomarkers from the literature", *BioMed Research International*, volume 2014 (2014), Article ID 253128, 11 pages.
- [20] [https://edrn.nci.nih.gov/biomarkers#b\\_start=0](https://edrn.nci.nih.gov/biomarkers#b_start=0)[accessed last on Nov 20, 2020]
- [21] <https://edrn.nci.nih.gov/sites/87-national-cancer-institute/srivastava-sudhir> [accessed last on Nov 20, 2020]
- [22] [http://www.knapsackfamily.com/KNApSAcK\\_Family/](http://www.knapsackfamily.com/KNApSAcK_Family/) [accessed last on Nov 20, 2020]
- [23] <https://www.NCBI.nlm.nih.gov/books/NBK22183/> [accessed last on Nov 20, 2020]
- [24] Wijaya, Sony Hartono, et al. "Supervised clustering based on DPCLUSO: prediction of plant-disease relations using Jamu formulas of KNAPSAcK database." *BioMed Research International* 2014 (2014).
- [25] Jutel, Annemarie. "Classification, disease, and diagnosis." *Perspectives in biology and medicine* 54.2 (2011): 189-205.
- [26] Agustí, Alvar, et al. "Precision medicine in airway diseases: moving to clinical practice." *European Respiratory Journal* 50.4 (2017): 1701655.

- [27] Bell, Mandy J. "A historical overview of preeclampsia - eclampsia." *Journal of Obstetric, Gynecologic & Neonatal Nursing* 39.5 (2010): 510-518.
- [28] Berlin, Brent. *Ethnobiological classification: Principles of categorization of plants and animals in traditional societies*. Princeton University Press, 2014
- [29] Winston, Judith E. *Describing species: practical taxonomic procedure for biologists*. Columbia University Press, 1999.
- [30] Nordenfelt, Lennart. "Identification and classification of diseases: fundamental problems in medical ontology and epistemology." *Studia Philosophica Estonica* (2013): 6-21.
- [31] Sydenham, T. (2011). *Opera Omnis*. Available online at: <https://archive.org/details/b24400750> [accessed last on July 07, 2020].
- [32] Moriyama, Iwao Milton, et al. "History of the statistical classification of diseases and causes of death." (2011).
- [33] American Public Health Association. *The Bertillon classification of causes of death*. R. Smith print. Company, 1890.
- [34] Nicholson, Jeremy K., and Ian D. Wilson. "Understanding 'global' systems biology: metabonomics and the continuum of metabolism." *Nature Reviews Drug Discovery* 2.8 (2003): 668-676.
- [35] Bryant, Penelope A., et al. "Chips with everything: DNA microarrays in infectious diseases." *The Lancet infectious diseases* 4.2 (2004): 100-111.
- [36] Hirsch, Heather A., et al. "A transcriptional signature and common gene networks link cancer with lipid metabolism and diverse human diseases." *Cancer cell* 17.4 (2010): 348-361.
- [37] M. Zitnik ,B Zupan, "Discovering disease-disease associations by fusing systems-level molecular data', *Systems Biology and Applications* ,2013
- [38] D.S Lee, et al, "The implications of human metabolic network topology for disease comorbidity. *Proc. Natl. Acad. Sci. USA* 105, 9880–5 (2008).
- [39] K. Goh, et al., "The human disease network. *Proc. Natl. Acad. Sci. USA* 104, 8685–8690 (2007).

- [40] F. Emmert-Streib, S. Tripathi, R. Simoes, A. Hawwa, & M. Dehmer, "The human disease network: opportunities for classification, diagnosis and prediction of disorders and disease genes", *Syst. Biomed.* 1, 15–22 (2013).
- [41] J. Loscalzo, I. Kohane, & A.L. Barabási, "Human disease classification in the postgenomic era: a complex systems approach to human pathobiology", *Mol. Syst. Biol.* 3, 124 (2007).
- [42] N. Gulbahce. et al., "Viral perturbations of host networks reflect disease etiology", *PLoS Comput. Biol.* 8, e1002531 (2012).
- [43] A. Halu , M.D. Domenico, "The multiplex network of human diseases", *Systems Biology and Applications* ,Sci Rep. (2019)
- [44] Han, L. Y., et al. "Prediction of functional class of novel viral proteins by a statistical learning method irrespective of sequence similarity." *Virology* 331.1 (2005): 136-143.
- [45] Seeger, Michael, et al. "A novel protein complex involved in signal transduction possessing similarities to 26S proteasome subunits." *The FASEB Journal* 12.6 (1998): 469-478.
- [46] <https://www.ncbi.nlm.nih.gov/protein/> [accessed last on Nov 20, 2020]
- [47] [https://www.ajinomoto.com/aboutus/amino\\_acids/20-amino-acids](https://www.ajinomoto.com/aboutus/amino_acids/20-amino-acids) [accessed last on Nov 20,, 2020]
- [48] <https://cran.r-project.org/web/packages/protr/vignettes/protr.html> [accessed last on Nov 20, 2020]
- [49] Bhasin, Manoj, and Gajendra PS Raghava. "Classification of nuclear receptors based on amino acid composition and dipeptide composition." *Journal of Biological Chemistry* 279.22 (2004): 23262-23266.
- [50] Guruprasad, Kunchur, BV Bhasker Reddy, and Madhusudan W. Pandit. "Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence." *Protein Engineering, Design and Selection* 4.2 (1990): 155-161.

- [51] Khan, Muslim, et al. "Unb-DPC: Identify mycobacterial membrane protein types by incorporating un-biased dipeptide composition into Chou's general PseAAC." *Journal of theoretical biology* 415 (2017): 13-19.
- [52] S. K. Kachigan, *Multivariate Statistical Analysis: A Conceptual Introduction*, Radius Press, New York, NY, USA, 1991
- [53] J. L. Rodgers and W. A. Nicewander, "Thirteen ways to look at the correlations coefficient," *The American Statistician*, vol. 42, pp. 59–66, 1995.
- [54] Chan, Y. H. "Biostatistics 104: correlational analysis." *Singapore Med J* 44.12 (2003): 614-9.
- [55] Min Li, Yu Tang, Dongyan Li, Fangxiang Wu, Jianxin Wang, *CytoCluster: a cytoscape plugin for cluster analysis and visualization of biological networks*.
- [56] R. Shen, A. B. Olshen, and M. Ladanyi, "Integrative clustering of multiple genomic data types using a joint latent variable model with application to breast and lung cancer subtype analysis," *Bioinformatics*, vol. 25, no. 22, pp. 2906–2912, 2009.
- [57] T. Hwang, G. Atluri, M. Xie et al., "Co-clustering phenomegenome for phenotype classification and disease gene discovery," *Nucleic Acids Research*, vol. 40, no. 19, article e146, 2012.
- [58] Y. Liu, Q. Gu, J. P. Hou, J. Han, and J. Ma, "A network-assisted co-clustering algorithm to discover cancer subtypes based on gene expression," *BMC Bioinformatics*, vol. 15, article 37, 2014.
- [59] Altaf-Ul-Amin, Md, et al. "Development and implementation of an algorithm for detection of protein complexes in large interaction networks." *BMC bioinformatics* 7.1 (2006): 207.
- [60] M. B. Karim, N. Wakamatsu, and M. Amin. "DPCLUSOST: A Software Tool for General Purpose Graph Clustering." *Journal of Computer-Aided Chemistry* 18 (2017): 76-93.
- [61] Ohtana, Yuki, et al. "Clustering of 3D - Structure Similarity Based Network of Secondary Metabolites Reveals Their Relationships with Biological Activities." *Molecular Informatics* 33.11 - 12 (2014): 790-801.

- [62] Wakamatsu, Nobutaka, et al. "[Special Issue for Honor Award dedicating to Prof Kimito Funatsu] Prediction of Metabolite Activities by Repetitive Clustering of the Structural Similarity Based Networks." *Journal of Computer Aided Chemistry* 20 (2019): 76-83.
- [63] <https://pubchem.NCBI.nlm.nih.gov/> [accessed last on Nov 20, 2020]
- [64] <https://www.bioconductor.org/packages/release/bioc/vignettes/ChemmineR/inst/doc/ChemmineR.html> [accessed last on Nov 20, 2020]
- [65] D. Bajusz, et al. , "Why is Tanimoto index an appropriate choice for fingerprint-based similarity calculations?", *Journal of Cheminformatics*, 7(1), pp.1–13, 2015.
- [66] J. W. Godden, L. Xue, and J. Bajorath, "Combinatorial preferences affect molecular similarity/diversity calculations using binary fingerprints and tanimoto coefficients," *Journal of Chemical Information and Computer Sciences*, vol. 40, no. 1, pp. 163–166, 2000.
- [67] T. Galili, "dendextend: an R package for visualizing, adjusting and comparing trees of hierarchical clustering," *Bioinformatics*, vol. 31, no. 22, pp. 3718–3720, 2015.
- [68] Szekely, Gabor J., and Maria L. Rizzo. "Hierarchical clustering via joint between-within distances: Extending Ward's minimum variance method." *Journal of classification* 22.2 (2005): 151-184.
- [69] Liu, Kang, et al. "Novel approach to classify plants based on metabolite-content similarity." *BioMed research international* 2017 (2017).
- [70] F. B. Baker, "Stability of two hierarchical grouping techniques Case I: sensitivity to data errors," *Journal of the American Statistical Association*, vol. 69, no. 346, pp. 440–445, 1974.
- [71] Fowlkes, E. B.; Mallows, C. L. (1 September 1983). "A Method for Comparing Two Hierarchical Clusterings". *Journal of the American Statistical Association* 78 (383): 553.
- [72] <https://cran.r-project.org/web/packages/dendextend/vignettes/dendextend.html> [accessed last on Nov 20, 2020]



- [73] Goh, K.-I. et al. The human disease network. *Proc. Natl Acad. Sci. USA* 104, 8685–8690 (2007)
- [74] Jimenez-Sanchez, Gerardo, Barton Childs, and David Valle. "Human disease genes." *Nature* 409.6822 (2001): 853-855..
- [75] Rual J.-F, Venkatesan K, Hao T, Hirozane-Kishikawa T, Dricot A, Li N, Berriz G.F, Gibbons F.D, Dreze M, Ayivi-Guedehoussou N, "Towards a proteome-scale map of the human protein-protein interaction network. *Nature*", 2005; 437: 1173-1178
- [76] Geng, Rong-Xin, et al. "Identification of core biomarkers associated with outcome in glioma: evidence from bioinformatics analysis." *Disease markers* 2018 (2018).
- [77] Ludwig, Heinz, and Kathrin Strasser. "Symptomatology of anemia." *Seminars in oncology*. Vol. 28. WB Saunders, 2001.
- [78] Suzuki, Jun-ichi, et al. "Oxygen therapy prevents ventricular arrhythmias in patients with congestive heart failure and sleep apnea." *Circulation Journal* 70.9 (2006): 1142-1147.
- [79] Stankovic, Tatjana, et al. "Ataxia telangiectasia mutated–deficient B-cell chronic lymphocytic leukemia occurs in pregerminal center cells and results in defective damage response and unrepaired chromosome damage." *Blood, The Journal of the American Society of Hematology* 99.1 (2002): 300-309.
- [80] Alfonso, P., et al. "Effect of enzyme replacement therapy on lipid profile in patients with Gaucher's disease." *Medicina clinica* 120.17 (2003): 641-646.
- [81] Nimkuntod, Porntip, and Pattama Tongdee. "Plasma low-density lipoprotein cholesterol/high-density lipoprotein cholesterol concentration ratio and early marker of carotid artery atherosclerosis." *J Med Assoc Thai* 98.Suppl 4 (2015): S58-63.
- [82] Gurbuz, Ahmet Seyfeddin, et al. "Acquired long QT syndrome and torsades de pointes related to donepezil use in a patient with Alzheimer disease." *The Egyptian Heart Journal* 68.3 (2016): 197-199.

- [83] Howes, Laurence Guy. "Cardiovascular effects of drugs used to treat Alzheimer's disease." *Drug safety* 37.6 (2014): 391-395.
- [84] Murphy, M Paul, and Harry LeVine 3rd. "Alzheimer's disease and the amyloid-beta peptide." *Journal of Alzheimer's disease: JAD* vol. 19,1 (2010): 311-23. doi:10.3233/JAD-2010-1221
- [85] Robelin, Laura, and Jose Luis Gonzalez De Aguilar. "Blood biomarkers for amyotrophic lateral sclerosis: myth or reality?" *BioMed research international* 2014 (2014).
- [86] Chancellor, Andrew M., et al. "The prognosis of adult-onset motor neuron disease: a prospective study based on the Scottish Motor Neuron Disease Register." *Journal of neurology* 240.6 (1993): 339-346.
- [87] Ferguson, Shirley M., et al. "Similarities in mental content of psychotic states, spontaneous seizures, dreams, and responses to electrical brain stimulation in patients with temporal lobe epilepsy." *Psychosomatic Medicine* 31.6 (1969): 479-498.
- [88] Gaby, Alan R. "Natural approaches to epilepsy." *Alternative medicine review* 12.1 (2007): 9.
- [89] Reece, Jane B., and B. Wilbur. "Nervous system." *Campbell Biology*, 9th edn. Pearson-Benjamin Cummings (2011).
- [90] Pardridge, William M. "Transport of nutrients and hormones through the blood-brain barrier." *Diabetologia* 20.1 (1981): 246-254.
- [91] Shenhar-Tsarfaty, Shani, et al. "Weakened cholinergic blockade of inflammation associates with diabetes-related depression." *Molecular Medicine* 22.1 (2016): 156-161.
- [92] American Chemical Society. "How diabetes can increase cancer risk: DNA damaged by high blood sugar." *ScienceDaily*. ScienceDaily, 25 August 2019.
- [93] Herceg, Zdenko, and Zhao-Qi Wang. "Functions of poly (ADP-ribose) polymerase (PARP) in DNA repair, genomic integrity and cell death." *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* 477.1-2 (2001): 97-110.

- [94] Ikeda, Fumie, et al. "Hyperglycemia increases risk of gastric cancer posed by *Helicobacter pylori* infection: a population-based cohort study." *Gastroenterology* 136.4 (2009): 1234-1241.
- [95] Mellekjaer, L., et al. "Rheumatoid arthritis and cancer risk." *European Journal of Cancer* 32.10 (1996): 1753-1757.
- [96] Klinaki, Eleni, et al. "Rheumatoid arthritis and cancer risk results from the Greek European prospective investigation into cancer and nutrition cohort." *European Journal of Cancer Prevention* 27.5 (2018): 502-506.
- [97] Mussa, Alessandro, Patrizia Matarazzo, and Andrea Corrias. "Papillary thyroid cancer and autoimmune polyglandular syndrome." *Journal of Pediatric Endocrinology and Metabolism* 28.7-8 (2015): 793-795.
- [98] Sylvester, Francisco A. "Inflammatory bowel disease: effects on bone and mechanisms." *Understanding the Gut-Bone Signaling Axis*. Springer, Cham, 2017. 133-150.
- [99] Sgambato, Dolores, et al. "Bone alterations in inflammatory bowel diseases." *World journal of clinical cases* 7.15 (2019): 1908.
- [100] [https://www.cell.com/immunity/comments/S1074-7613\(03\)00326-1](https://www.cell.com/immunity/comments/S1074-7613(03)00326-1)  
[accessed last on Nov 20, 2020]
- [101] Parasa, Ramesh Babu, and Nicola Maffulli. "Musculoskeletal involvement in cystic fibrosis." *Bulletin (Hospital for Joint Diseases (New York, NY))* 58.1 (1999): 37-44.
- [102] Aris, R. M., et al. "Abnormal bone turnover in cystic fibrosis adults." *Osteoporosis International* 13.2 (2002): 151-157.
- [103] Tangpricha, Vin. "Vitamin D Nutrition in CF: Guidelines and Beyond." *Pediatric Pulmonology*. Vol. 52. 111 River St, Hoboken 07030-5774, Nj Usa: Wiley, 2017.
- [104] Lo Cascio, Christian M., et al. "Gastrointestinal dysfunction in patients with Duchenne muscular dystrophy." *PLoS One* 11.10 (2016): e0163779.

- [105] Bellini, Massimo, et al. "Gastrointestinal manifestations in myotonic muscular dystrophy." *World journal of gastroenterology: WJG* 12.12 (2006): 1821.
- [106] Mazlumzadeh, Mehrdad, and Eric L. Matteson. "Cogan's syndrome: an audiovestibular, ocular, and systemic autoimmune disease." *Rheumatic Disease Clinics of North America* 33.4 (2007): 855-874.
- [107] El-Mahmood, A. M., O. B. Ogbonna, and M. Raji. "The antibacterial activity of *Azadirachta indica* (neem) seeds extracts against bacterial pathogens associated with eye and ear infections." *Journal of medicinal plants research* 4.14 (2013): 1414-1421.
- [108] Scadding, Glenis K., and Paul K. Keith. "Fluticasone furoate nasal spray consistently and significantly improves both the nasal and ocular symptoms of seasonal allergic rhinitis: a review of the clinical data." *Expert opinion on pharmacotherapy* 9.15 (2008): 2707-2715.
- [109] Sharma, Rajeev. *Mouth-Teeth and Ear-Nose-Throat Disorders*. Diamond Pocket Books Pvt Ltd, 2016.
- [110] Krause-Bachand, Jeanie, and Wilma Koopman. "Living with oculopharyngeal muscular dystrophy: a phenomenological study." *Can J Neurosci Nurs* 30.1 (2008): 35-39.
- [111] Torpy, Janet M., Alison E. Burke, and Richard M. Glass. "Trachoma." *JAMA* 302.9 (2009): 1022-1022.
- [112] Mungenast, Felicitas, and Theresia Thalhammer. "Estrogen biosynthesis and action in ovarian cancer." *Frontiers in endocrinology* 5 (2014): 192.
- [113] Persson, Ingemar. "Estrogens in the causation of breast, endometrial and ovarian cancers—evidence and hypotheses from epidemiological findings." *The Journal of steroid biochemistry and molecular biology* 74.5 (2000): 357-364.
- [114] Stagi, Stefano, et al. "Thyroid function in Rett syndrome." *Hormone Research in Paediatrics* 83.2 (2015): 118-125.

- [115] Somunkiran, Asli, et al. "Anti-Müllerian hormone levels during hormonal contraception in women with polycystic ovary syndrome." *European Journal of Obstetrics & Gynecology and Reproductive Biology* 134.2 (2007): 196-201.
- [116] Reichman, David E., et al. "Fertility in patients with congenital adrenal hyperplasia." *Fertility and sterility* 101.2 (2014): 301-309.
- [117] Srivatsa, Bharath, et al. "Microchimerism of presumed fetal origin in thyroid specimens from women: a case-control study." *The Lancet* 358.9298 (2001): 2034-2038.
- [118] Wang, S., et al. "Effects of maternal subclinical hypothyroidism on obstetrical outcomes during early pregnancy." *Journal of Endocrinological Investigation* 35.3 (2012): 322-325.
- [119] Kim, J. Julie, Takeshi Kurita, and Serdar E. Bulun. "Progesterone action in endometrial cancer, endometriosis, uterine fibroids, and breast cancer." *Endocrine reviews* 34.1 (2013): 130-162.
- [120] Linkov, Faina, et al. "Endometrial hyperplasia, endometrial cancer and prevention: gaps in existing research of modifiable risk factors." *European Journal of Cancer* 44.12 (2008): 1632-1644.
- [121] Collignon, Olivier, et al. "Women process multisensory emotion expressions more efficiently than men." *Neuropsychologia* 48.1 (2010): 220-225.
- [122] Kaufman, Keith D. "Androgens and alopecia." *Molecular and cellular endocrinology* 198.1-2 (2002): 89-95.
- [123] Pingale, Prashant L., et al. "A review on alopecia and its remedies." *Int. J. Pharmacol. Pharmaceut. Sci* 2.3 (2014): 45-52.
- [124] Tunuguntla, Hari Siva Gurunadha Rao. "Management of Peyronie's disease—a review." *World journal of urology* 19.4 (2001): 244-250.
- [125] Kelly, Edward J., and Richard D. Palmiter. "A murine model of Menkes disease reveals a physiological function of metallothionein." *Nature genetics* 13.2 (1996): 219-222.
- [126] Tümer, Zeynep, and Lisbeth B. Møller. "Menkes disease." *European Journal of Human Genetics* 18.5 (2010): 511-518.

- [127] Wu, Ming-Kung, et al. "Significantly higher prevalence rate of asthma and bipolar disorder co-morbidity: a meta-analysis and review under PRISMA guidelines." *Medicine* 95.13 (2016).
- [128] Beiko, Tatsiana, and Charlie Strange. "Anxiety and depression in patients with alpha-1 antitrypsin deficiency: current insights and impact on quality of life." *Therapeutics and clinical risk management* 15 (2019): 959.
- [129] Partti, Krista, et al. "Lung function and respiratory diseases in people with psychosis: population-based study." *The British Journal of Psychiatry* 207.1 (2015): 37-45.
- [130] Gordon, Adam J., James W. Conley, and Joanne M. Gordon. "Medical consequences of marijuana use: a review of current literature." *Current Psychiatry Reports* 15.12 (2013): 419.
- [131] Praud, Jean-Paul, Yuichiro Miura, and Martin G. Frasch. "Animal Models for the Study of Neonatal Disease." *Animal Models for the Study of Human Disease*. Academic Press, 2017. 805-837.
- [132] 132. <https://www.healthgrades.com/right-care/kidneys-and-the-urinary-system/urinary-disorders> [accessed last on Jan 21, 2021]

## **Achievements**

### **Peer review journal paper**

Shaikh Farhad Hossain, Ming Huang, Naoaki Ono, Aki Morita, Shigehiko Kanaya and Md. Altaf-Ul-Amin “Development of a Biomarker Database Towards Performing Disease Classification and Finding Disease Interrelations” Oxford University Press, The Journal of Biological Databases and Curation (2021) Vol. 2021: article ID baab011; doi:10.1093/database/baab011

Shaikh Farhad Hossain, Kazuhisa Hirose, Shigehiko Kanaya and Md. Altaf-Ul-Amin, “Playing Virtual Musical Drums by MEMS 3D Accelerometer Sensor Data and Machine Learning;” Machine Learning and Applications: An International Journal (MLAIJ) ISSN: 2394 – 0840, Volume 10, March, 2019.

Shaikh Farhad Hossain, Mohammad Bozlul Karim, Takematsu Shotaro, Shigehiko Kanaya and Md. Altaf-Ul-Amin, "Development and Mining of a Human Biomarker Database," International Journal of Pharma Medicine and Biological Sciences, Vol. 9, No. 1, pp. 27-32, doi: 10.18178/ijpmbs.9.1.27-32, January 2020

### **Peer review international conference**

Shaikh Farhad Hossain, Ming Huang, Naoaki ONO, Shigehiko Kanaya and Md. Altaf-Ul-Amin, “Inter Disease Relations Based on Human Biomarkers by Network Analysis”, IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE), Athens, Greece, 2019

Shaikh Farhad Hossain, Sony Hartono Wijaya, Ming Huang, Irmanida Batubara, Shigehiko Kanaya and Md. Altaf-Ul-Amin "Prediction of Plant-Disease Relations Based on Unani Formulas by Network Analysis" IEEE 18th International Conference on Bioinformatics and Bioengineering (BIBE), Taichung, Taiwan 2018

Shaikh Farhad Hossain, Kazuhisa Hirose, Ming Huang, Naoaki ONO, Shigehiko Kanaya and Md. Altaf-Ul-Amin, “Multiple Biomarkers of a Single Disease Show Structural Similarity and Minor Atom Variation”, Asian Regional Conference on Systems Biology 2020 (ARCSB2020), Ombak Villa Langkawi, Malaysia, March 1-4, 2020

## Appendix A

### R Code

```
# The source codes below were used to Calculate Pearson Correlation

library(readxl)
Martix <- read_excel("C:/Users/Farhad Hossain/Desktop/Martix1.xlsx")
View(Martix)
p <- cor(x = Martix, method = "pearson")
#library(xlsx)
#write.csv(p, "C:/Users/Farhad Hossain/Desktop/output.csv")

#####

#To arrange the square pearson in two node and it's similarity (Row-wise)

#library(readxl)
#p <- read_excel("C:/Users/Farhad Hossain/Desktop/output1.xlsx")
#out <- data.frame(fn= character(185136), sn= character(185136), cr =
numeric(185136), stringsAsFactors = FALSE)
out <- data.frame(fn= character(1370340), sn= character(1370340), cr =
numeric(1370340), stringsAsFactors = FALSE)
#df <- data.frame(x = rep(NA,185136),y = rep(NA,185136),z = rep(NA,185136))
k <- 1
for(i in 2 : nrow(p))
{
  for(j in 1 : (i-1) )
  {
    # print(i)
    #out <- rbind.data.frame(out, rownames(p)[i] , colnames(p)[j], 1)
    a <- rownames(p)[i]
    b <- colnames(p)[j]
    if (is.numeric(p[i,j]))

    {
      out[k,1] <- a
      out[k,2] <- b
      out[k,3] <- p[i,j]
    }
  }
}
```



```

    k<- k+1
    # print(k)
  }
}

write.csv(out, file = "C:\\Users\\Farhad Hossain\\Desktop\\PhD
Thesis\\yourfile.csv", row.names = FALSE)

```

### # Unique Value check and their frequency after Cluster

```

library("xlsx")
library("openxlsx")
out <- as.data.frame(read.xlsx("C:\\Users\\Farhad Hossain\\Desktop\\New
folder\\7.0_c_UnaniFormula_Cor_Cluster.xlsx", sheet="Sheet1", startRow = 1,
colNames = FALSE, rowNames = FALSE))

gsub("[", "XY", "sd[dasdada]")

p <- data.frame( clsno = numeric(), fn =character(), stringsAsFactors=FALSE)
for(i in 1:nrow(out)) {
  a1<- unlist(strsplit(gsub("\\]", "", gsub("\\[", "", out[i,2])), ", "))
  for(j in 1: length(a1) )
  {
    ch <- trimws(a1[j], "both")
    p = rbind(p, data.frame(clsno=i, fn =ch))
    print(ch)
  }
}

s <- as.vector(p['fn'])
s1 <- table(s)
write.csv(s1, file = "C:\\Users\\Farhad Hossain\\Desktop\\PhD
Thesis\\yourfile.csv", row.names = FALSE)

```

### #####voting node counting of a cluster #####

```

library("openxlsx")
out <- as.data.frame(read.xlsx("C:\\Users\\Farhad Hossain\\Desktop\\PhD
Thesis\\UnaniFormula\\8_(Start here)_Replace Node to Class & Voting\\test1.xlsx",
sheet="cluster", startRow = 1, colNames = TRUE, rowNames = FALSE))

vot <- as.data.frame(read.xlsx("C:\\Users\\Farhad Hossain\\Desktop\\PhD
Thesis\\UnaniFormula\\8_(Start here)_Replace Node to Class & Voting\\test1.xlsx",
sheet="mapping", startRow = 1, colNames = TRUE, rowNames = FALSE))

```

```

dis <- character()
disfre <- character()

#fn =character() ,disease =character(), vot =character(),
outp <- data.frame( clsno = numeric() ,fn =character() ,disease
=character(),vot =character(), stringsAsFactors=FALSE)

for(i in 1:nrow(out)) {
a1<- unlist(strsplit(gsub("\\\\", "", gsub("\\\\[", "", out[i,2])) , ","))
##print(a1)
dis <- NULL
disfre <- NULL
for(j in 1: length(a1)) {
ch <- trimws(a1[j], "both")
fsub <- vot[vot$formula == ch, ]
if ( is.na(fsub[1,2]) == 0) dis <- c(dis, fsub[1,2])
if ( is.na(fsub[1,3]) == 0) dis <- c(dis, fsub[1,3])
if ( is.na(fsub[1,4]) == 0) dis <- c(dis, fsub[1,4])
if ( is.na(fsub[1,5]) == 0) dis <- c(dis, fsub[1,5])
}
dist <- table(dis)
p <-as.data.frame(dist)
p <- p[with(p, order(-p$Freq)), ]
for(k in 1: nrow(p)) {
s<- paste(p[k,1], '-', p[k,2])
disfre <- c(disfre, s)

}
#print(disfre)

outp = rbind(outp, data.frame(clsno=i, fn = out[i,2], disease =
paste(dis,collapse=",") , vot = paste(disfre,collapse=",") ))
}
write.csv(outp, file = "C:\\Users\\Farhad Hossain\\Desktop\\PhD
Thesis\\UnaniFormula\\8_(Start here)_Replace Node to Class & Voting\\out.csv",
row.names = FALSE)

##### Formula replace to Plant number and sorting #####

library("openxlsx")
plant <- as.data.frame(read.xlsx("C:\\Users\\Farhad Hossain\\Desktop\\f and
p\\F & P.xlsx", sheet="data" ,startRow = 1, colNames = TRUE , rowNames = FALSE))

formula <- as.data.frame(read.xlsx("C:\\Users\\Farhad Hossain\\Desktop\\f and
p\\10_Plant to disease relation.xlsx", sheet="out" ,startRow = 1, colNames = TRUE ,
rowNames = FALSE))

```

```

    outp <- data.frame( clsno = numeric() ,cluster =character() ,disease
=character(),plantin =character() ,plantfre =character(), stringsAsFactors=FALSE)

    dis <- character()
    disfre <- character()

    for(i in 1:nrow(formula)) {
      a1<- unlist(strsplit(gsub("\\\\", "", gsub("\\\\[", "", formula[i,2]))
, ""))

      dis <- NULL
      disfre <- NULL

      for(j in 1: length(a1)) {
        ch <- trimws(a1[j], "both")
        pp <- plant[plant$formula == ch, ]

        fsub <- pp[ , colSums(is.na(pp)) == 0]
        p<- as.character(colnames(fsub))
        p<- tail(p,-1)
        dis <- c(dis,p)
      }

      dist <- table(dis)
      p <-as.data.frame(dist)
      p <- p[with(p, order(-p$Freq)), ]

      for(k in 1: nrow(p)) {
        s<- paste(p[k,1], '-',p[k,2])
        disfre <- c(disfre,s)
      }
      #print(dis)
      outp = rbind(outp, data.frame(clsno=i, cluster = formula[i,2], disease =
formula[i,3] , plantin = paste(dis,collapse=","),plantfre =
paste(disfre,collapse=",")) )

    }

    write.csv(outp, file = "C:\\Users\\Farhad Hossain\\Desktop\\f and p\\out.csv",
row.names = FALSE)

```

**##### Frequency distribution #####**

```

library("openxlsx")
displ<-as.data.frame(read.xlsx("C:\\Users\\Farhad
Hossain\\Desktop\\freq\\Disease and Plant.xlsx", sheet="out" ,startRow = 1,
colNames = TRUE , rowNames = FALSE))
pl <- as.character(unique(displ$disease))
ssss <- NULL
for(i in 1: length(pl)) {
  pp <- displ[displ$disease == pl[i], ]

```

```

for(j in 1: nrow(pp)) {
  # j <-1
  ssa <- unlist(strsplit (trimws(pp[j,2]), "both"),',,')
  k <- j
  while (k < nrow(pp) ) {
    k<-k+1
    ssp <- unlist(strsplit (trimws(pp[k,2]), "both"),',,')

    ssnew <- NULL
    for(l in 1: length(ssa))
    {
      # l <-1
      chp <- unlist(strsplit (trimws(ssa[l], "both"),' - ')
      for(m in 1: length(ssp))
      {
        # m <-1
        chm <- unlist(strsplit (trimws(ssp[m], "both"),' - ')
        if (chp[1] == chm[1] )
          {
            fsum      <-      as.character(as.integer(chp[2])
+as.integer(chm[2]))
            ssa[l] <-paste(chp[1],'-',fsum)
            ssp[m] <-paste(chm[1],'-','0')

          }
        }
      }

      pp[j,2] <- paste(ssa,collapse=",")
      pp[k,2] <- paste(ssp,collapse=",")

    }
  }
  ssss <- rbind(ssss,pp)
}

write.csv(ssss, file = "C:\\Users\\Farhad Hossain\\Desktop\\freq\\out.csv",
row.names = FALSE)

```

**#####replace plant name to #####**

```

library("openxlsx")
dis <- as.data.frame(read.xlsx("C:\\Users\\Farhad Hossain\\Desktop\\Plant
Number to Name\\12.0_Predicted plant for disease.xlsx", sheet="Predicted
Class" ,startRow = 1, colNames = TRUE , rowNames = FALSE))
plant <- as.data.frame(read.xlsx("C:\\Users\\Farhad Hossain\\Desktop\\Plant
Number to Name\\Plant Number & Name.xlsx", sheet="Sheet1" ,startRow = 1, colNames
= TRUE , rowNames = FALSE))

```

```

outp <- data.frame( disease=character(), freq = numeric() , disid = numeric(),
plantid =character() ,name =character(), stringsAsFactors=FALSE)

for(i in 1:nrow(dis)) {
s<- trimws(dis[i,4], "both")
r<- plant[plant$pl == s,]
if (nrow(r) > 0 ) t <- r[1,2]
else t<-'-'
# print(t)
outp = rbind(outp, data.frame(disease=dis[i,1],freq=dis[i,2],
disid=dis[i,3],plantid = dis[i,4], name = t ))
}

write.csv(outp, file = "C:\\Users\\Farhad Hossain\\Desktop\\Plant Number to
Name\\out.csv", row.names = FALSE)

##### simpson coefficient Calculation #####

library("openxlsx")
mat <- as.data.frame(read.xlsx("C:\\Users\\Farhad Hossain\\Desktop\\New
folder\\Martix.xlsx", sheet="data" ,startRow = 1, colNames = TRUE , rowNames =
TRUE))

outp <- data.frame( fform =character(185136) ,sform =character(185136), simc
=numeric(185136), a1 =numeric(185136),b1 =numeric(185136),c1 =numeric(185136),
stringsAsFactors=FALSE)
k<-0
for(i in 2:nrow(mat)) {
f1<-as.numeric(mat[i,])
for(j in 1:(i-1)) {
f2 <- as.numeric(mat[j,])
f3<- f1+f2
a <- sum(f3 == 2)
b <- sum(f1 == 1) -a
c <- sum(f2 == 1) -a
if (b < c ) { sc <- a/(a+b)
} else sc <- a/(a+c)

# outp = rbind(outp, data.frame(fform= rownames(mat)[i],sform=
rownames(mat)[1],simc = sc,a1=a, b1=b,c1=c ))
k<- k+1
outp[k,1]=rownames(mat)[i]
outp[k,2]=rownames(mat)[j]
outp[k,3]=sc
outp[k,4]=a
outp[k,5]=b
outp[k,6]=c

}

}

```

```
write.csv(outp, file = "C:\\Users\\Farhad Hossain\\Desktop\\New
folder\\out.csv", row.names = FALSE)
```

**##### dendrogram Construction #####**

```
library("openxlsx")
#chemid <-read.xlsx("D:\\ChemmineData\\hcluster\\matrixmol.xlsx",
sheet="matrixmol" ,startRow = 1, rowNames = TRUE, colNames = TRUE)
chemid <-read.xlsx("D:\\ChemmineData\\k_value_decision\\5_matrixmol.xlsx",
sheet="matrixmol" ,startRow = 1, rowNames = TRUE, colNames = TRUE)
#chemid <-read.xlsx("C:\\Users\\Farhad Hossain\\Desktop\\New
folder\\MMM.xlsx", sheet="Sheet1" ,startRow = 1, rowNames = TRUE, colNames = TRUE)

chemid <- as.matrix(chemid)
#d = as.dist(chemid)
d <- dist(chemid, method = "euclidean")
hc <- hclust(d, method = "ward.D")
plot(hc, hang=-1)
rect.hclust(hc, k=22, border="red") # divide the plot into k-class
mycl <- cutree(hc, k=22) # class of objects

# binding the data and class, and save it as a csv file
temp <- as.matrix(mycl)
result <- cbind(chemid,temp)
#write.csv(result,"D:\\ChemmineData\\k_value_decision\\out.csv",row.names=TR
UE)
write.csv(result,"D:\\ChemmineData\\k_value_decision\\Optimum_5_n_6\\out.csv
",row.names=TRUE)
write.csv(result,"C:\\Users\\Farhad Hossain\\Desktop\\New
folder\\out.csv",row.names=TRUE)
```

**#####dendrogram comparision#####**

```
install.packages("dendextend")
install.packages("dendextendRcpp")
install.packages("microbenchmark")

library("openxlsx")
library("dendextend")

chemid1 <-read.xlsx("C:\\Users\\Farhad
Hossain\\Desktop\\dendextend\\MMM1.xlsx", sheet="Sheet1" ,startRow = 1, rowNames
= TRUE, colNames = TRUE)
chemid2 <-read.xlsx("C:\\Users\\Farhad
Hossain\\Desktop\\dendextend\\MMM2.xlsx", sheet="Sheet1" ,startRow = 1, rowNames
= TRUE, colNames = TRUE)

chemid1 <- as.matrix(chemid1)
chemid2 <- as.matrix(chemid2)

d1 <- dist(chemid1, method = "euclidean")
d2 <- dist(chemid2, method = "euclidean")
```

```

hc1 <- hclust(d1, method = "ward.D")
hc2 <- hclust(d2, method = "ward.D")

plot(hc1, hang=-1)
rect.hclust(hc1, k=3, border="red") # divide the plot into k-class
mycl <- cutree(hc1, k=3) # class of objects

plot(hc2, hang=-1)
rect.hclust(hc2, k=3, border="red") # divide the plot into k-class
mycl <- cutree(hc2, k=3) # class of objects

tree1 <- as.dendrogram(hc1)
tree2 <- as.dendrogram(hc2)

y <- Bk(hc1, hc2, k = 3)
print(y)

#####Mapping call from the datasheet#####

library("openxlsx")

cluster <- as.data.frame(read.xlsx("C:\\Users\\Farhad Hossain\\Desktop\\New
folder\\clustering.xlsx", sheet="Sheet1" ,startRow = 1, colNames = TRUE , rowNames
= FALSE))
mapping <- as.data.frame(read.xlsx("C:\\Users\\Farhad Hossain\\Desktop\\New
folder\\mapping.xlsx", sheet="Sheet1" ,startRow = 1, colNames = TRUE , rowNames =
FALSE))

sink("C:\\Users\\Farhad Hossain\\Desktop\\New folder\\output.txt", append =
T)

result1 <- NULL
result2 <- NULL
result3 <- NULL
result4 <- NULL

for (i in 1:nrow(cluster))

{
  result1 <- cluster[i:i,]
  result4 <- result1
  for (j in 1:nrow(mapping))

  {
    result2 <- mapping[c(j),c(1)]
    if(result1 == result2)
    {
      result3 <- as.data.frame(mapping[c(j),c(2:19)]);
      result4 <- cbind(result4,result3)
    }
  }
}

```

```

}

result4 <- unname(result4)
result4 <- result4[!is.na(result4)]
print(result4)
}
sink()

```

**#####PubChem ID signature extraction#####**

```

if (!requireNamespace("BiocManager", quietly=TRUE))
  install.packages("BiocManager")
BiocManager::install("ChemmineR")

data(sdfsampl)
sdfset <- sdfsampl
install.packages("openxlsx")

#Run code from here
library("ChemmineR")
library("xlsx")
library("openxlsx")

chemid <- read.xlsx("D:\\ChemmineData\\19_matrixmol\\fileid.xlsx",
sheet="Sheet1", startRow = 1, colNames = TRUE)
sdall <- NULL
s1 <- "D:\\ChemmineData\\19_matrixmol\\sdf\\Structure2D_CID_"
s2 <- as.character(chemid[1,1])
s1 <- paste(s1,s2, '.sdf', sep='')
sdall <- read.SDFset(s1)

for(i in 2:nrow(chemid))
{
  s1 <- "D:\\ChemmineData\\19_matrixmol\\sdf\\Structure2D_CID_"
  s2 <- as.character(chemid[i,1])
  s1 <- paste(s1,s2, '.sdf', sep='')
  print(s1)
  sdd <- read.SDFset(s1)
  sdall <- c(sdall,sdd)
}

cid(sdall) <- makeUnique(cid(sdall))
unique_ids <- makeUnique(sdfid(sdall))
sdf.visualize(sdall[1:10])
cid(sdall) <- unique_ids
propma <- data.frame(MF=MF(sdall), MW=MW(sdall), atomcountMA(sdall))
sdallap <- sdf2ap(sdall) # main line of the code
cid(sdallap[1])
which(sapply(as(sdallap, "list"), length)==1)

```



```

foset <- desc2fp(sdallap, descnames = 300, type = "FPset")
fpma <- as.matrix(foset)
write.csv(fpma, file = "D:\\ChemmineData\\19_matrixmol\\out.csv", row.names =
FALSE)

```

**#####PubChem ID structure visualization #####**

```

library("ChemmineR")
library("xlsx")
library("openxlsx")

chemid <-
read.xlsx("D:\\ChemmineData\\hclust_visualize_without_tanimoto\\hclust_22.xlsx",
sheet="Sheet1" ,startRow = 1, colNames = TRUE)
sdall <- NULL
s1 <-
"D:\\ChemmineData\\hclust_visualize_without_tanimoto\\sdf\\Structure2D_CID_"
s2 <-as.character(chemid[1,1])
s1 <- paste(s1,s2, '.sdf', sep='')
sdall <-read.SDFset(s1)
for(i in 2:nrow(chemid))
{
s1 <-
"D:\\ChemmineData\\hclust_visualize_without_tanimoto\\sdf\\Structure2D_CID_"
s2 <-as.character(chemid[i,1])
s1 <- paste(s1,s2, '.sdf', sep='')
print(s1)
sdd <- read.SDFset(s1)
sdall <- c(sdall,sdd)
}
cid(sdall) <- makeUnique(cid(sdall))
unique_ids <- makeUnique(sdfid(sdall))
chemid
sdf.visualize(sdall[1:92])

if (!requireNamespace("BiocManager", quietly=TRUE))
install.packages("BiocManager")
BiocManager::install("ChemmineR")
install.packages("openxlsx")

```

**#####Tanimoto calculation#####**

```

library("xlsx")

```

```

library("ChemmineR") #run package every time
library("openxlsx")
chemid <- read.xlsx("D:\\ChemmineData\\19_matrixmol\\fileid.xlsx",
sheet="Sheet1" ,startRow = 1, colNames = TRUE)
sdall <- NULL

s1 <- "D:\\ChemmineData\\19_matrixmol\\sdf\\Structure2D_CID_"
s2 <- as.character(chemid[1,1])
s1 <- paste(s1,s2, '.sdf', sep='')
sdall <- read.SDFset(s1)
for(i in 2:nrow(chemid))
{
s1 <- "D:\\ChemmineData\\19_matrixmol\\sdf\\Structure2D_CID_"
s2 <- as.character(chemid[i,1])
s1 <- paste(s1,s2, '.sdf', sep='')
print(s1)
sdd <- read.SDFset(s1)
sdall <- c(sdall,sdd)
}

cid(sdall) <- makeUnique(cid(sdall))
unique_ids <- makeUnique(sdfid(sdall))
sdf.visualize(sdall[1:10])
cid(sdall) <- unique_ids
propma <- data.frame(MF=MF(sdall), MW=MW(sdall), atomcountMA(sdall))

sdallap <- sdf2ap(sdall)
cid(sdallap[1])
foset <- desc2fp(sdallap, descnames = 1024, type = "FPset")
out <- data.frame(c1 =character(), c2 = character(), d = numeric())
for(i in 1:nrow(chemid))
{
for(j in i:nrow(chemid))
{
#r<-fpSim(foset[i],foset[j] ,method="Tversky",alpha=1, beta=1)
r <-fpSim(foset[i],foset[j], method="Tanimoto")
# r<-cmp.similarity(sdallap[i],sdallap[j])
if ((cid(sdallap[i]) != cid(sdallap[j])) & (r >= 0.85)) {
out = rbind(out,data.frame(C1=cid(sdallap[i]),C2=cid(sdallap[j]), C3=r))
}
}
}
out
#write.csv(out, file = "C:\\Users\\hira\\workspace\\DPClassGUI1.5\\Combine\\relationall.csv", row.names = FALSE)
write.csv(out, file = "D:\\ChemmineData\\fignerprint\\Out.csv", row.names = FALSE)

#####

```

```

library("openxlsx")
rela <-
as.data.frame(read.xlsx("D:\\ChemmineData\\Threshold_setting\\Threshold_setting.
xlsx", sheet="Sheet1", startRow = 1, colNames = TRUE, rowNames = FALSE))
bio <-
as.data.frame(read.xlsx("D:\\ChemmineData\\Threshold_setting\\Threshold_setting.
xlsx", sheet="Sheet2", startRow = 1, colNames = TRUE, rowNames = FALSE))
outp <- data.frame( bio=character(), s1 = numeric(), s2= numeric(),
stringsAsFactors=FALSE)

for(i in 1:nrow(bio)) {
  p1<- rela[as.integer(bio[i,2]),3]
  outp = rbind(outp, data.frame(bio=bio[i,1], s1=bio[i,2],s2=p1 ))
}
write.csv(outp,file
"D:\\ChemmineData\\Threshold_setting\\out.csv",row.names = FALSE)

```

**#####FASTA File read and descriptor extraction#####**

```

library("protr")
extracell <- readFASTA(system.file(
  "protseq/extracell.fasta",
  package = "protr"
))
extracell <- extracell[(sapply(extracell, protcheck))]

length(extracell)

x1 <- t(sapply(extracell, extractAPAAC));
x <- extractAAC(extracell)
x <- readFASTA(system.file(
  "protseq/FASTA_1686/1BVK_D.fasta",
  package = "protr"
))[[1]]
x <- readFASTA(system.file(
  "protseq/FASTA_1686/0311213A.fasta",
  package = "protr"
))[[1]]
x1<- readFASTA(system.file("protseq/FASTA_1686/0311213A.fasta",package =
"protr"))[[1]]
x <- readFASTA(system.file("protseq/FASTA_1686/0311213A.fasta",package =
"protr"))[[1]]
x <- readFASTA(system.file("protseq/FASTA_1686/1AT3_B.fasta",package =
"protr"))[[1]]
x <- readFASTA(system.file("protseq/FASTA_1686/2109248A.fasta",package =
"protr"))[[1]]
x <- readFASTA(system.file(
  "protseq/FASTA_1686/s2",
  package = "protr"
))[[1]]

```

```

chemid          <-read.xlsx("C:\\Users\\shaikhfarhad-h\\Desktop\\protein
signature\\protein_list.xlsx", sheet="Sheet1" ,startRow = 1, colNames = TRUE)
sdall <- NULL
s1          <-          "C:\\Users\\shaikhfarhad-h\\Documents\\R\\win-
library\\3.5\\protr\\protseq\\FASTA_1686\\"
s2 <-as.character(chemid[1,1])
s1 <- paste(s1,s2,sep='')
print(s1)
sdall <-readFASTA(s1)
x1 <- extractAAC(sdall)
for(i in 2:nrow(chemid))
{
s1          <-          "C:\\Users\\shaikhfarhad-h\\Desktop\\protein
signature\\FASTA_1686\\"
s2 <-as.character(chemid[i,1])
s1 <- paste(s1,s2,sep='')
print(s1)
sdd <- readFASTA(s1)
sdall <- c(sdall,sdd)
}
propma <- data.frame(MF=MF(sdall), MW=MW(sdall), atomcountMA(sdall))
dfs <- list(y1, y2, y3)
x2 <- extractAAC(x1)
fpma <- as.matrix(dfs)
write.csv(fpma, file = "C:\\Users\\shaikhfarhad-h\\Desktop\\protein
signature\\out.csv", row.names = TRUE)
write.csv(fpma, file = "C:\\Users\\shaikhfarhad-h\\Desktop\\protein
signature\\out.csv", startColumn = 2, row.names = FALSE))

```