1-22-2020

Network-based Computational Drug Repurposing and Repositioning for Breast Cancer Disease

Pavithra Ulaganathan
University of Windsor

Network-based Computational Drug Repurposing and Repositioning for Breast Cancer Disease

by

Pavithra Ulaganathan

A Thesis
Submitted to the Faculty of
Graduate Studies through the
School of Computer Science in Partial Fulfillment
of the Requirements for the Degree of Master of
Science at the University of Windsor

Windsor, Ontario, Canada
2020

Network-based Computational Drug Repurposing and Repositioning for Breast Cancer Disease

by

Pavithra Ulaganathan

APPROVED BY:

Dr. M. Hlynka

Department of Mathematics and Statistics

Dr. X. Yuan

School of Computer Science

Dr. A. Ngom, Advisor

School of Computer Science

January 22, 2020

Declaration of Originality

I hereby certify that I am the sole author of this thesis and that no part of this thesis has been published or submitted for publication.

I certify that, to the best of my knowledge, my thesis does not infringe upon anyone's copyright nor violate any proprietary rights and that any ideas, techniques, quotations, or any other material from the work of other people included in my thesis, published or otherwise, are fully acknowledged in accordance with the standard referencing practices. Furthermore, to the extent that I have included copyrighted material that surpasses the bounds of fair dealing within the meaning of the Canada Copyright Act, I certify that I have obtained a written permission from the copyright owner(s) to include such material(s) in my thesis and have included copies of such copyright clearances to my appendix.

I declare that this is a true copy of my thesis, including any final revisions, as approved by my thesis committee and the Graduate Studies office, and that this thesis has not been submitted for a higher degree to any other University or Institution.

Abstract

Pharmaceutical drug development is a complex, time-consuming and expensive process which is also limited to a relatively small number of targets. Drug repositioning is a vital function which involves finding new uses and indications for already approved and existing drugs. It is a cost-effective process in contrast to experimental drug discovery. Previous studies have shown that the network-based method is a versatile platform for drug repositioning as there exists more biological networks which can be used to model interaction between the biological concepts. In this thesis, we are interested in finding the best drugs for one of the most prevailing disease, the Breast Cancer using the existing Protein-protein interaction (PPI) networks. The proposed method is based on the idea that if a perturbation gene expression profile inversely corelates with the disease gene expression profile, the drug may have a curing effect on the disease. Six samples of stroma surrounding invasive breast primary tumours and six matched samples of normal stroma are extracted from the public functional genomics data repository, Gene Expression Omnibus. The perturbation gene expression data corresponding to MCF7 cell line was extracted from the National Institute of Health's (NIH) Library of Integrated Network-Based Cellular Signatures (LINCS) dataset. Machine Learning techniques are used to select the best suited drug for the breast cancer disease. We have used a ranking algorithm to obtain a ranked list of suitable drug repurposing and repositioning candidates.

Dedication

I dedicate this thesis to my beloved parents Mr. Ulaganathan Subramanian and Mrs. Anushya Ulaganathan, and to the rest of my family and friends.

Acknowledgements

I would like to sincerely express my most profound gratitude towards my supervisor Dr. Alioune Ngom, for his support, guidance, encouragement and valuable advice. Thank you for being patient and training me to think better.

I would like to thank my internal reader Dr. Xiaobu Yuan and my external reader Dr. Myron Hlynka for all their valuable inputs and suggestions given to me.

I would also like to thank my friends Pon Vignesh Muthukumarasamy, Chandhini and more each of whom helped me in their own way during my thesis.

Finally, I would like to thank my parents Ulaganathan Subramanian, Anushya Ulaganathan and my grandmother Savithiri Shanmugam, for their immense support from the very beginning.

Collectively, all their support and guidance has enabled me to successfully complete my master's program.

Table of Contents

List of Tables

List of Figures

List of Acronyms

DEG	Differentially Expresssed Gene
DES	Differentially Expressed Sub-Network
LCD	Louvian Community Detection

Chapter 1

Introduction

1.1 Drug Discovery

1.1.1 What is a Drug?

In pharmacology, a drug is a chemical substance, typically of known structure, which, when administered to a living organism, produces a biological effect [1]. A pharmaceutical drug, also called as medication or medicine, is a chemical substance used to treat, cure, prevent, or diagnose a disease or to promote well-being [2].

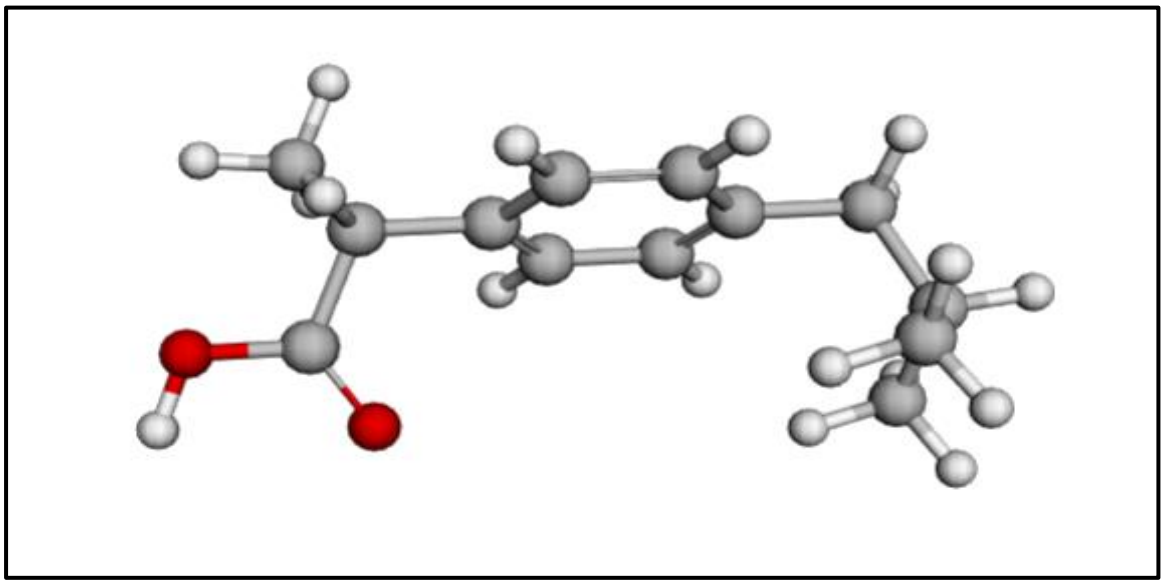


Figure 1.1: 3D molecular structure of Ibuprofen

Figure 1.1 shows the 3D molecular structure of a drug known and sold as Ibuprofen [3], the most common drug in the world. There are some drugs which are not used to specifically treat disease but act as a psychoactive chemical substance influencing a better mood by impacting the central nervous system.

1.1.2 Traditional drug discovery pipeline

Traditional drug discovery and development procedures is highly time consuming and comes at a high development cost. Developing a new drug takes about 10 to 17 years and it costs between $500 billion to $2 million dollar [4]. Approximately 90% of the newly discovered drugs fail in the clinical trials due to their side - effects or adverse effects [5] and only one in 10,000 compounds can make it to market, and less than 20% of drugs entering Phase II clinical trials succeed [6].

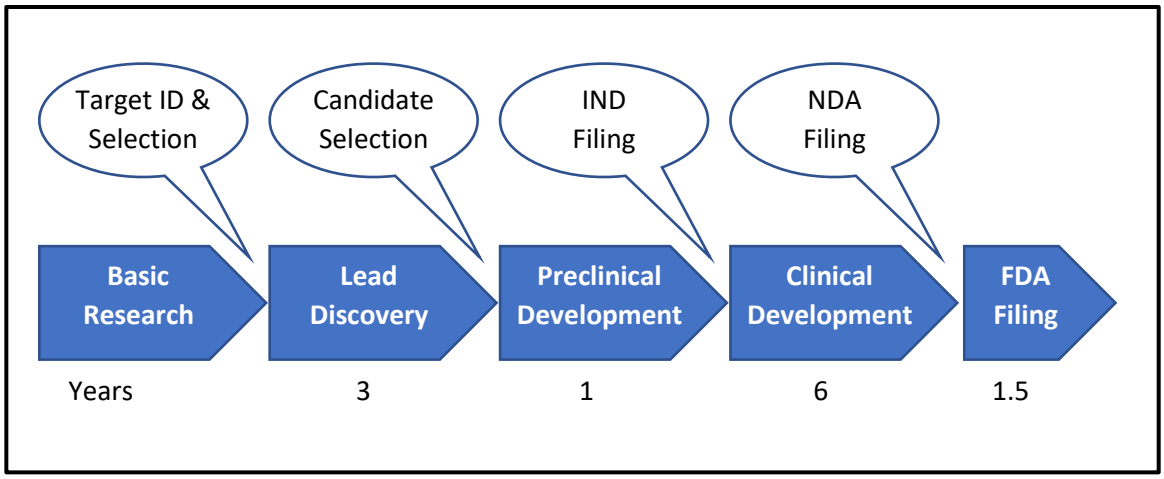


Figure 1.2: Drug Development Cycle

Figure 1.2 shows the steps involved in traditional drug discovery. Step 1 involves basic research and target identification. Target can be defined as the pathogen in which the drug is meant to create an effect on.

Step 2 is the Lead discovery and optimization, which is an initial stage of drug discovery process where the small molecules (drugs) are carefully vetted to observe traces of lead compound, a pharmacological chemical. Step 3 is the Examining the dosage level of drugs and ensuring the safety of the drug. This phase is essential before proceeding to clinical trials. Since a drug cannot be used on humans without having the knowledge of whether it is safe to consume or not, these trials are conducted on other species that have genetics resembling human genetics. Step 4, Clinical trials are where drugs are tested on humans to study

their effect before marketing. The final step involves getting the drug FDA approved, where the FDA review teams thoroughly examine all the submitted data related to the drug or device and decide to approve or not to approve it. To minimize the time and costs associated with traditional drug discovery process computational drug discovery is a preferred alternative.

1.1.3 Computational Drug Discovery

Computational drug discovery is an effective strategy for accelerating and economizing drug discovery and development process [22]. It covers many aspects of drug discovery, including computer programs for designing compounds, tools for systematically assessing potential lead candidates and the development of digital repositories for studying chemical interactions [23]. Because of the availability of biological macromolecule and small molecule information, the applicability of computational drug discovery has been extended to most aspects of the drug discovery and development process [24], from target identification and validation to lead discovery and optimization; the tools can even be applied to preclinical trials, which greatly alters the pipeline for drug discovery and development [25]. Figure 1.3 shows a flowchart for the tasks that computational approaches have been applied to and the computational methods used at each stage.

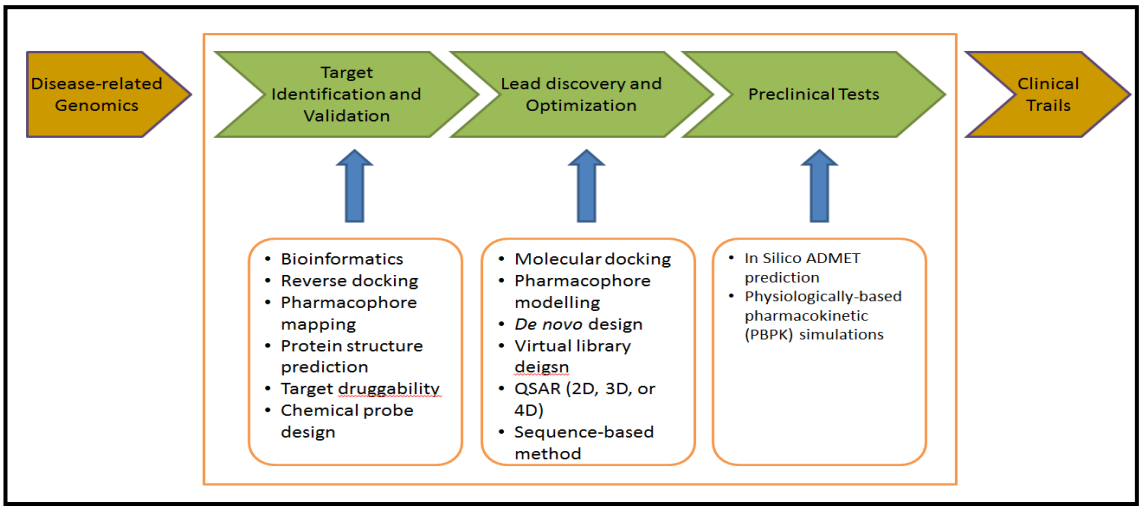


Figure 1.3: Computational Drug Discovery approaches

The use of computational tools could reduce the cost of drug development by up to 50% [26]. Drug Repurposing and Drug Repositioning is one of the remarkable computational drug discovery methodology that is being used to overcome the issues in the traditional drug discovery process.

1.1.4 Drug Repurposing and Drug Repositioning?

Drug Repurposing aims at finding new indications for already existing FDA approved drugs for a disease and therefore increases the available therapeutic choices at a fraction of cost of new drug development [7], whereas Drug Repositioning involves finding indications for drugs that have been developed but failed in the clinical trials or drugs that have not been approved by the FDA.
The drugs used in our research fall under one of the three categories listed below.

- Approved
- Experimental
- Investigational

Approved drugs are those that have passed clinical trials. Experimental drugs are those that have shown to bind proteins in mammals or bacteria. Investigational drugs are at one of the phases of drug design process in one jurisdiction or more.

Drug Repurposing/Repositioning involves the same procedure but differs on the type of drug recommended. The unapproved drugs which closely resemble the properties of approved drugs intended to treat other disease of interest shall be selected as suitable candidates for drug repositioning, while the approved drugs intended to treat other disease of interest shall be selected as suitable candidates for drug repurposing.

1.1.5 Drug Repurposing and Drug Repositioning Methods

Figure 1.4 illustrates the different methods of Drug Repurposing. There are different classifications for Drug Repurposing methods, each of which seeks to categorize the existing methods depending on some important metrics.

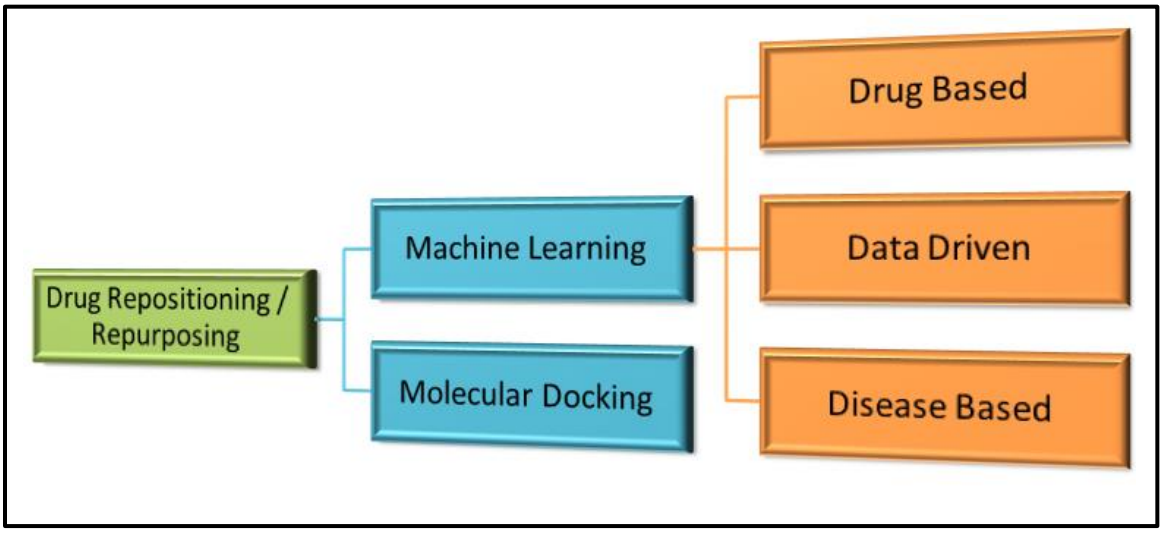


Figure 1.4: Drug Repurposing Methods

Two major Drug Repurposing approaches are docking simulation and machine learning. Molecular docking method try to simulate and model the physical interactions between the drugs and targets and are used in structural molecular biology and computer-assisted drug design [8]. Successful docking methods can efficiently search high-dimensional conformation spaces and accurately rank the candidate dockings using a scoring function [9]. However, there are some limitations in the use of molecular docking in Drug Repurposing. The requirement of known three-dimensional (3D) structure of chemical ligands and protein targets severely limits the application of docking because the structures of many physiologically important proteins are not fully resolved [10]. Moreover, molecular docking methods demand significant computational resources resulting in extended runtimes [11].

Additionally, because of errors in the determined protein structure, and the incomplete modeling of atomic and molecular interactions, the results of molecular docking have high false-positive rates [10].

Machine Learning method treats Drug Repurposing as a supervised learning problem where machine learning algorithms are applied to biological data related to drugs and then link them to treat specific diseases. Machine learning methods appear more favorable than docking simulation, as they can examine a larger number of promising candidates for further experimental screening [11]. Machine learning methods can be further classified as drug-based, disease-based or Data driven methods [10]. Drug based methods try to discover repositioning opportunities by chemical or pharmaceutical perspective investigation, while disease-based methods focus on disease management, symptomatology or pathology.

If more accurate detection of pharmacological properties is needed, drug-based methods which involve pharmacological or chemical information on drugs may be preferred. By contrast, disease-based approaches may be preferred when there is insufficient knowledge of drug pharmacology. Disease-based can be preferred when the focus is on disease or therapeutic category. Each approach presents unique informatics challenges, often requiring elements from both drug- and disease-based methods to be incorporated for a successful process [8]. Data-driven approaches analyze large-scale '-omics' data sets using statistical modeling techniques [12].

The advances in biological sciences, have led the access to a lot of '-omics' molecular data in different levels such as the genome, transcriptome, proteome and metabolome; therefore, using data-driven approaches is an increasingly viable option. Network modelling is one of the most used data-driven approaches.

Networks are simple and versatile data structures on which associations can be inferred through many statistical and computational approaches [13].

A wide variety of concepts in biology are represented in the form of Networks. Figure 1.5 illustrates different types of biological interactions that can be represented by networks.

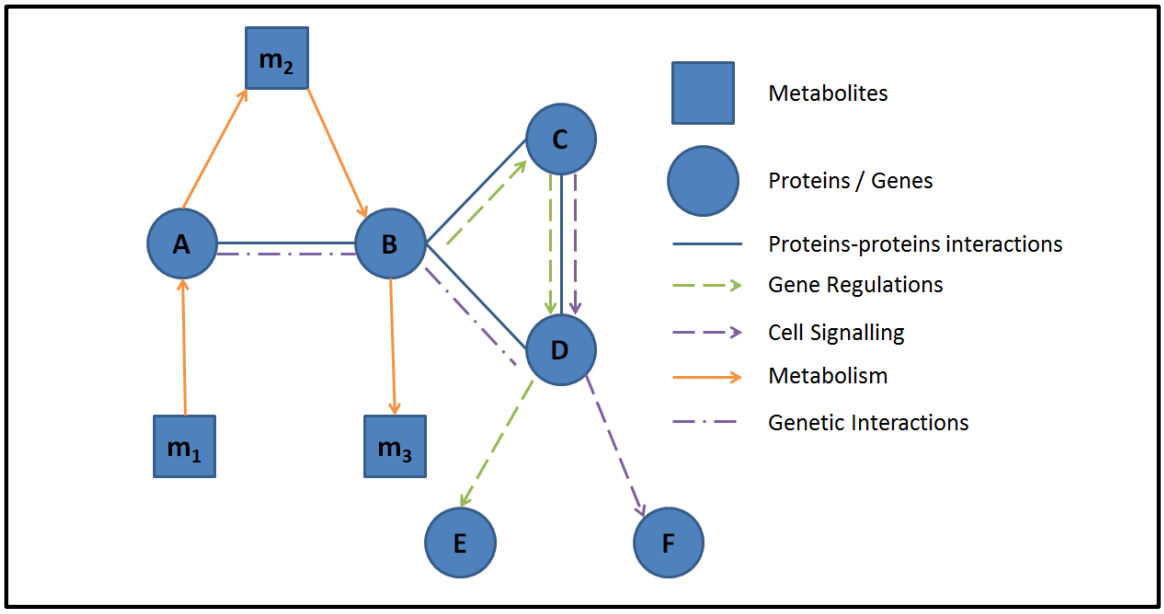


Figure 1.5: Types of Biological Networks

In biological networks, nodes represent various components like genes or proteins and edges represent the relationship between genes, proteins or the functional similarity between genes. To identify a drug target, a network-based strategy first reconstructs a biological network and then simulates its interactions. The resulting interaction relationships between drug targets reveal the potential drug targets [14]. Some of the advantages of the Network-based method is that molecular networks can provide insights into the context in which the drug target works and can, therefore, help understand the drug mechanisms of action [15].

Network algorithms can readily accomplish tasks such as visualizing various existing interactions, adding newly discovered relationships, and superimposing additional properties over primary components and their known interactions [16]. Various kinds of data from different data sets could be represented in one network. Therefore, the topological properties of the network can be used to make predictions when biological data is missing and thus reduces the false-positive rates [17].

1.2 Problem Statement

Given drug perturbation data and gene expression data for breast cancer, we aim to obtain a ranked list of drugs which would make suitable candidates for drug repurposing and drug repositioning for the breast cancer dataset. For this, six samples of stroma surrounding invasive breast primary tumours and six matched samples of normal stroma are extracted from the public functional genomics data repository, Gene Expression Omnibus. The perturbation gene expression data corresponding to MCF7 cell line was extracted from the National Institute of Health's (NIH) Library of Integrated Network-Based Cellular Signatures (LINCS) dataset.

We integrate information from different sources such as molecular interaction networks like Protein-Protein interaction networks (PPI) with the gene expression profiles for a strong Drug Repurposing/Repositioning. We then make use of machine learning method, the Louvain algorithm on the disease network dataset and finally the combinatorial optimization algorithm, the Hungarian Method is used to provide a ranked list of good drug repurposing and repositioning candidates.

1.3 Thesis Motivation

Researching the repurposing of unapproved drugs sharing similarities with approved drugs for the treatment of breast cancer would help speed up the drug design process involving phase of drug discovery and development. As a result, years of time and billions of dollars will have been conserved to help cure breast cancer disease. Most importantly, this thesis does its part in helping us move one step closer to acquiring suitable drugs to tackle breast cancer.

1.4 Thesis Contribution

In this thesis, we have proposed application of existing pre-processing and network clustering methods on the breast cancer dataset to obtain a ranked list of suitable drug repurposing and repositioning candidates. Our novel contribution in this thesis includes the integration of the external biological data with the primary gene expression data in order to increase the quality of drug repurposing or repositioning.

1.5 Thesis Organization

The rest of the thesis/ research work is organized in the following manner.

- In Chapter 2, we discuss literature review in the area of drug repurposing using computational approaches.
- In Chapter 3, we introduce our proposed approach and explain all the techniques used to obtain suitable drug repurposing candidates for the breast cancer dataset.
- In Chapter 4, we present the experimental results and perform an analysis of those results.
- Chapter 5 concludes the research by explaining insights received during the work and setting up the field of opportunities for possible future work.

Chapter 2

Literature Review

This chapter consists of some literature review regarding computational drug repurposing using various disease data. Several computational approaches for drug repurposing have been developed that is worth noting and we discuss some of those works below.

2.1 A Network Approach for Computational Drug Repositioning.

This paper is based on the hypothesis that a drug can be repositioned to another drug's curing area if two drugs share similar molecular and/or chemical properties. The authors of this paper, Jiao Li and Zhiyong Lu [27] constructed a disease-drug-target network based on prior knowledge (i.e., known therapeutic uses of drugs and known drug targets). Different from the other similarity-based methods, in drug pairwise similarity calculation, the authors have adapted a novel bipartite-graph based method to represent the relationships between drugs and their target proteins as a bipartite graph. Furthermore, they added the drug structure information into the drug pairwise similarity calculation and in this way their method boost the target similarity by making use of their corresponding interaction information. Then, the drug pair with higher similarity score is predicted to be repurposed to each other therapeutic area.

Limitation of this method is that the state of many structures and chemical properties of known drug compounds are untrustworthy and many physiological effects cannot be predicted by considering only structural features.

2.2 A new computational drug repurposing method using established disease-drug pair knowledge.

The paper is based on the method that if a drug-exposure gene expression profile inversely correlates with a disease gene expression profile, the drug may have a therapeutic effect on the disease. The authors of this paper, Draghici et al. [28] at first formed an input matrix by combining the reversed measurements of the genes in the disease profile and the measurements of the same genes in each of the drug profiles. Their workflow consists of transforming the input matrix into a lower dimensionality matrix by incorporating dimensionality reduction technique such as principal component analysis (PCA) or Locally Linear Embedding (LLE). Then the authors have used the known relationship between disease and its FDA approved drugs into a transformed space using distance metric learning algorithm. In this process, the clinically relevant drugs get close to the disease and the Euclidean distance between disease gene expression profile and each of the drug-exposure expression profiles is calculated. Then the drugs are ranked based on the closest to the farthest distance from the disease.

The authors of this paper have worked towards obtaining drug repurposing candidates for three diseases: breast cancer, rheumatoid arthritis and idiopathic pulmonary fibrosis. They have used GEO disease data for breast cancer, CMAP data for rheumatoid arthritis, and LINCS for idiopathic pulmonary fibrosis.

The authors of this paper have made use of only the transcriptional data, so the results are not much reliable. Incorporating transcriptional data with available clinical knowledge such as drug, chemical and disease biomarkers could yield better results.

2.3 A novel computational approach for drug repurposing using systems biology

The authors of this paper, Draghici et al. [29] built a global network (GN) which is the union of all KEGG human signalling pathways. Then a subgraph was extracted from the global network comprising of the shortest paths between the disease related genes and drug targets and termed it drug-disease network (DDN). Then a system level analysis was applied on the gene expression signatures of drug-disease pairs to generate gene perturbation signatures in the drug-disease network. Further, the authors have assigned a repurposing score on the drug disease pair and obtained a ranked drug list with potential therapeutic effects for the given disease based on the repurposing scores. Limitation of this paper is that the gene regulatory network constructed in the proposed method is biased due to the existence of noise in the gene expression data.

2.4 Drug repositioning for cancer therapy based on large-scale drug-induced transcriptional signatures

The authors of this paper, Lee et al. [30] have developed a series of seven classifiers using logistic regression to predict drug repurposing candidates for the treating of glioblastoma, lung cancer, and breast cancer.

They make use of signatures obtained from the chemical structure (S), drug-target relation (T), and gene expression data (E). Suitable drug repurposing candidates were predicted based on similarity of the signatures between the compounds and disease or known its drugs.

Limitation of this method is that the method considers only Differentially Expressed genes (DEG) in the drug dataset. DEG's in the disease dataset are not taken for consideration and structures for most of the drugs are not available.

2.5 Breaking the paradigm: Dr Insight empowers signature free, enhanced drug repurposing

The authors of this paper, Gu et al. [31] have developed a signature free, optimal drug repurposing based on gene expression data, namely Dr. Insight which overcomes the limitations of the existing computational frameworks. The method considers the dysregulation of gene expression from both disease and drug-perturbed data simultaneously, which renders the CEG's as optimal features to investigate the connections among diseases, drugs and genes. The authors have done an extensive comparison on simulated and real cancer datasets and validated the superior performance of Dr Insight over several popular drug-repurposing methods to detect known cancer drugs and drug–target interactions.

Chapter 3

Proposed Methods

In this chapter, we discuss the datasets, pre-processing steps taken, and machine learning techniques used in this thesis.

3.1 Breast Cancer

3.1.1 What is a gene?

A gene is the basic physical and functional unit of heredity. Genes are made up of DNA and every data point generated by a DNA microarray experiment denotes the ratio of expression levels [18]. The results from one experiment with n number of genes on one test subject denotes a series of expression levels. In each of these ratios, the numerator represents expression level of the gene in a varying condition and the denominator denotes the expression level of the gene in a reference condition. Data compiled together to form m such experiments presents a gene expression matrix. The gene expression value will be positive if the production of that gene is increased in that particular test case and will be negative if the generation of that gene is decreased instead [19]. Figure 3.1 shows a sample gene.

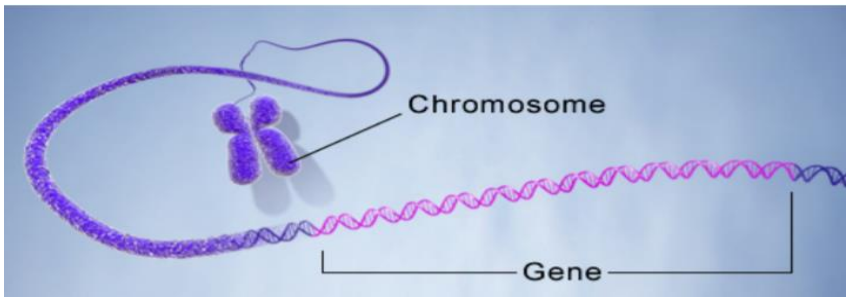


Figure 3.1*: Gene*

3.1.2 What is Breast Cancer?

Cancer occurs as a result of mutations, or abnormal changes, in the genes responsible for regulating the growth of cells and keeping them healthy. A tumour can be benign - not dangerous to health or malignant - has the potential to be dangerous. Breast cancer is a disease in which cells in the breast grow out of control [20]. Usually breast cancer either begins in the cells of the lobules, which are the milk-producing glands, or the ducts.

3.1.3 Why Breast Cancer?

Breast Cancer makes up 25% of all new cancer diagnoses in women across the globe according to the American Cancer Society (ACS) [21].

It is estimated that in Canada in 2019:

- 26,900 women will be diagnosed with breast cancer. This represents 25% of all new cancer cases in women in 2019.
- 5,000 women will die from breast cancer. This represents 13% of all cancer deaths in women in 2019.
- On average, 74 Canadian women will be diagnosed with breast cancer every day and on average, 14 Canadian women will die from breast cancer every day.

3.2 Datasets

3.2.1 Reactive Stroma of Breast and Prostate Cancer

The disease data titled “Reactive Stroma of Breast and Prostate Cancer” was obtained from the National Center for Biotechnology Information’s (NCBI), Gene Expression Omnibus (GEO) portal.

GEO is a public repository that archives and freely distributes comprehensive sets of microarray, next-generation sequencing, and other forms of high-throughput functional genomic data submitted by the scientific community. This dataset consists of gene expression data for a large pool of breast cancer genes. The dataset consists of stroma associated with prostate and breast invasive tumors. It consists of 24 samples which includes, six samples of stroma surrounding invasive breast primary tumours, six samples of stroma surrounding invasive prostate primary tumours and six matched samples of normal stroma for each type of tumour. Out of this we extracted the 6 samples of breast cancer stroma and six matched samples of normal stroma related to the breast cancer. The dataset consisted of 12 columns and 20,322 genes.

3.2.2 LINCS

The drug data was extracted from the pharmacogenomics perturbation data which is the National Institute of Health’s (NIH) Library of Integrated Network-Based Cellular Signatures (LINCS) dataset. This dataset consists of 21,567 drugs in the columns and 12,328 genes in the rows. This dataset is a level 5 LINCS dataset and consists of normalized z-score values. Level 4 LINCS data consists of two sets of data, before administration of drugs and after administration of drugs onto the genes in the dataset. These expression values from both the level 4 datasets are normalized to form the level 5 LINCS dataset. Figure 3.2 shows the LINCS L1000 data processing pipeline.

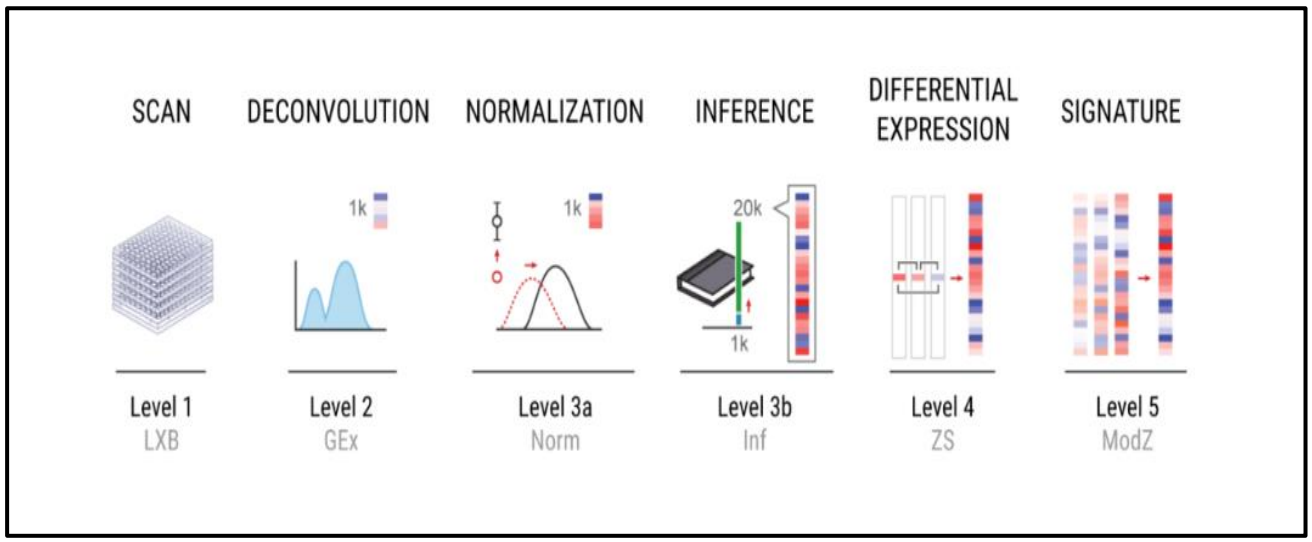


Figure 3.2: LINCS L1000 data processing

L1000 data is provided at five levels of the data processing pipeline [32]:

- Level 1: Raw unprocessed flow cytometry data from Luminex (LXB).
- Level 2: Gene expression values per 1000 genes after deconvolution (GEX).
- Level 3: Quantile-normalized gene expression profiles of landmark genes and imputed transcripts (Q2NORM or INF).
- Level 4: Gene signatures computed using z-scores relative to the plate population as control (ZSPCINF) or relative to the plate vehicle control (ZSVCINF).
- Level 5: Differential gene expression signatures.

3.2.3 Protein-Protein Interaction Networks

Proteins are large biomolecules, or macromolecules, consisting of one or more long chains of amino acid residues. The roles of proteins are many and varied. Protein, DNA, RNA and other biological molecules do not work in isolation; they cooperate with other proteins to perform a biological activity. Two molecules that cooperate to perform a function are said to be interacting. It is the combination of these molecules and their interactions, and not the molecules alone, that characterize the mechanisms of a biological process. Protein–protein

interactions (PPIs) are the physical contacts of high specificity established between two or more protein molecules as a result of biochemical events steered by electrostatic forces including the hydrophobic effect. Protein-protein Interactions (PPI) can be modelled as networks. Each protein is represented as a node, and an edge between any two nodes indicates that these two proteins interact. Figure 3.3 [37] shows an example of Protein-Protein interaction networks, where the proteins are represented by nodes.

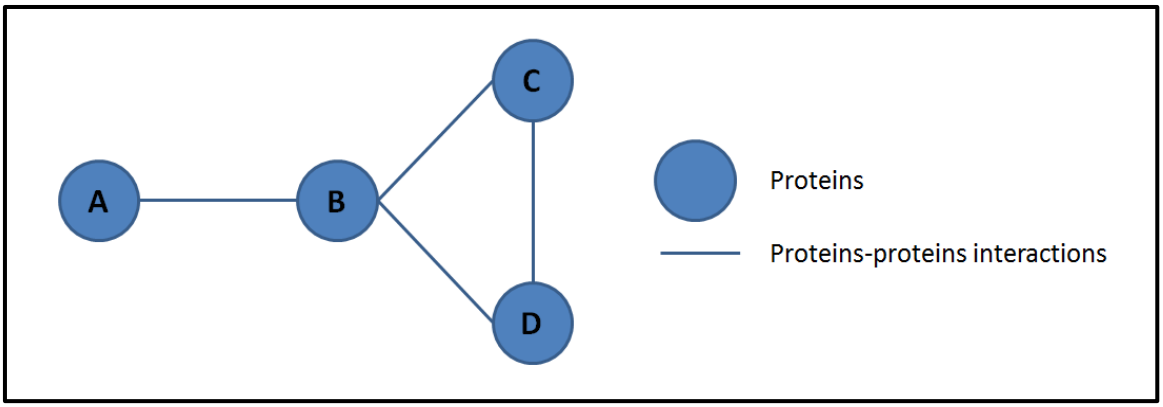


Figure 3.3: Protein-Protein Interaction Network

Pathway Commons is a database of biological pathways and biomolecular interactions aggregated from many source databases. Pathway Commons has biomolecular interaction data from Reactome, NCI Pathways, PhosphoSite, HumanCyc, Transfac, MiRTarBase, Drugbank, Recon X, Comparative Toxicogenomics Database, and KEGG [38].

In molecular biology, STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) is a biological database and web resource of known and predicted protein–protein interactions [39]. The STRING database contains information from numerous sources, including experimental data, computational prediction methods and public text collections. The resource also serves to highlight functional enrichments in user-provided lists of proteins, using a number of functional classification systems such as GO, Pfam and KEGG

[40]. The STRING database contains information on about 9.6 million proteins from more than 2000 organisms.

3.3 Pre-processing of Datasets

The pre-processing pipeline used on the datasets to be used in the Machine Learning methods in this thesis is explained in the following subsections.

3.3.1 Reactive Stroma of Breast Cancer Dataset

The statistical scores such as p-value, FDR-corrected p-value (or q-value) and Z-scores was calculated using the sick and the healthy samples of the breast cancer disease dataset. Having the possibility of there being a large number of false positives is not statistically good and so we calculate the q-values using the false discovery rate (FDR) approach.

The false discovery rate (FDR) is a method of conceptualizing the rate of type I errors (rejection of a true null hypothesis) in null hypothesis testing when conducting multiple comparisons. The older approaches reduced the number of false positives while also reducing the number of true positives which is not optimal. This newer FDR approach gives us adjusted p-values in every test case. In simpler terms, p-value predicts that there could be 5% false positives in the entire list of DE genes whereas q-value (FDR-adjusted p-value) predicts that there could be 5% false positives in the significant tests.
The false discovery rate formula is [34]:

$$\mathbf{FDR = E\,(V/R \mid R > 0)\,P(R > 0)}$$

Where,

- V = Number of Type I errors (i.e. false positives)
- R = Number of rejected hypotheses

Q-values are the name given to the adjusted p-values found using an optimised FDR approach. The FDR approach is optimised by using characteristics of the p-value distribution to produce a list of q-values.
For the breast cancer dataset, the genes having FDR-corrected p-value(q-values) < 0.05 was the Differentially Expressed genes. Out of 20,322 genes in the breast cancer dataset, 138 genes were identified to be Differentially Expressed.

3.3.2 LINCS Drug Perturbation Data

The LINCS drug dataset consists of drugs related to 7 cell lines. Out of 7 cell lines, we have extracted drugs belonging to the cell line "MCF7". This way we have multiple entries of most drugs, so we have filtered them based on the dosage and time under administration. Within this cell line, we have filtered drugs whose dosage was 1.11 um and whose time under administration was 24 hours.

Out of 21,567 drugs, 1844 drugs passed our criteria. Then we computed p-values based on the z-score by using the normal distribution and FDR corrected p-value(q-value) for each gene per drug profile to select the statistically significant values [35].
Figure 3.4 [5] explains the LINCS pre-processing pipeline. Like the disease dataset, the genes having FDR corrected p-value < 0.05 was considered to be differentially expressed.

We calculated the percentage of differentially expressed genes in each drug profile and eliminated drugs that had fewer than 1% differentially expressed genes. So, in our dataset, out of 12,328 genes, we checked if there are more than 123 DE genes in a drug profile or not and selected only drugs that had more than 123 differentially expressed genes. This step has enabled us to select unique instances of all drugs fitting our criteria. We have extracted a total of 110 drugs based on these filters.

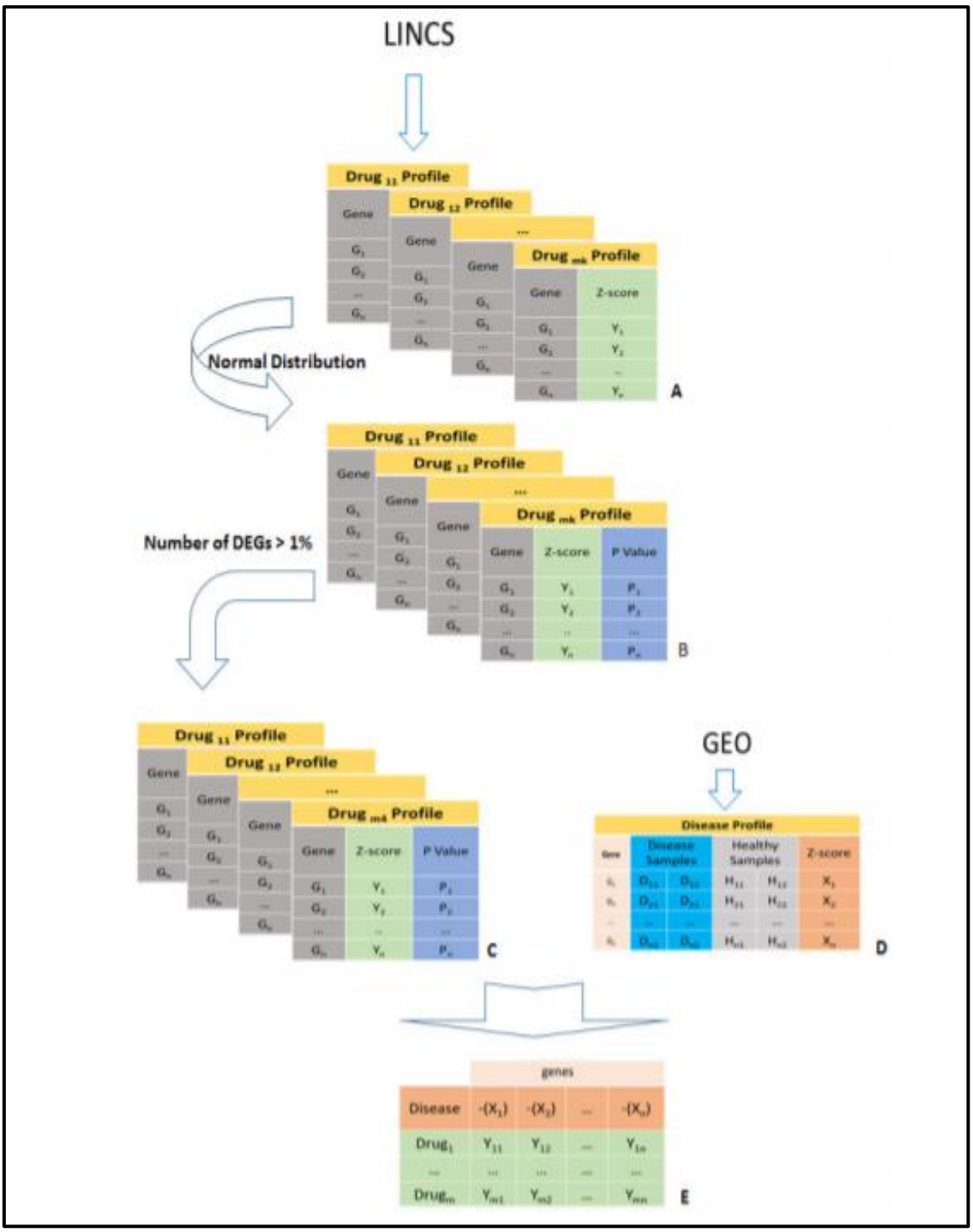


Figure 3.4: LINCS pre-processing pipeline

3.3.3 Protein-Protein Interaction Network Dataset

We obtained the directed protein-protein interaction network by combining protein interactions from the "Pathway Commons Protein-Protein Interactions database" and the "STRING" database. The PPI's in Pathway Commons database are classified as one of the following types in Table 3.1. We extracted the interaction from the Pathway Commons Protein-Protein Interactions database with the interaction types in "controls-expression-of, controls-state-change-of, controls-phosphorylation-of and catalyses-precedes".

Interaction Types	Description
controls-state-change-of	First protein controls a reaction that changes the state of the second protein.
controls-transport-of	First protein controls a reaction that changes the cellular location of the second protein.
controls-phosphorylation-of	First protein controls a reaction that changes the phosphorylation status of the second protein.
controls-expression-of	First protein controls a conversion or a template reaction that changes expression of the second protein.
catalysis-precedes	First protein controls a reaction whose output molecule is input to another reaction controled by the second protein.
in-complex-with	Proteins are members of the same complex.
interacts-with	Proteins are participants of the same Molecular Interaction.
neighbor-of	Proteins are participants or controlers of the same interaction.
consumption-controled-by	The small molecule is consumed by a reaction that is controled by a protein
controls-production-of	The protein controls a reaction of which the small molecule is an output.
controls-transport-of-chemical	The protein controls a reaction that changes cellular location of the small molecule.

Table 3.1: Types of Protein-Protein Interaction

Then, we extracted the directed interactions from the STRING database. The confidence score of an interaction in STRING database is defined as the approximate probability that a predicted link exists between two enzymes in the same metabolic map in the KEGG database. Confidence limits are as follows

- low confidence - 0.25 (or better),
- medium confidence - 0.4,
- high confidence - 0.7,
- highest confidence - 0.9

We removed all the interaction with weak confidence, with score < 250 (i.e., 0.25). The duplicate interactions were removed, and PPI network was formed by combining interactions from the "Pathway Commons Protein-Protein Interactions database" and "STRING" database. PPI network comprises of 904284 unique interactions.

3.4 Methodology

In this thesis, we have proposed two methodologies to find the ranked list of drugs which would make suitable candidates for Drug Repositioning/Repurposing for the disease Breast Cancer. We formed the breast network data from the PPI network by considering only the genes present in the breast disease data and the disease network data consist of 716,426 unique interactions. Then, we constructed the drug network data from the PPI network, for each drug profile by considering only the differentially expressed genes present in each drug profile of the drug data and obtained 110 drug networks. Figure 3.5 represents the pipeline of our overall thesis.

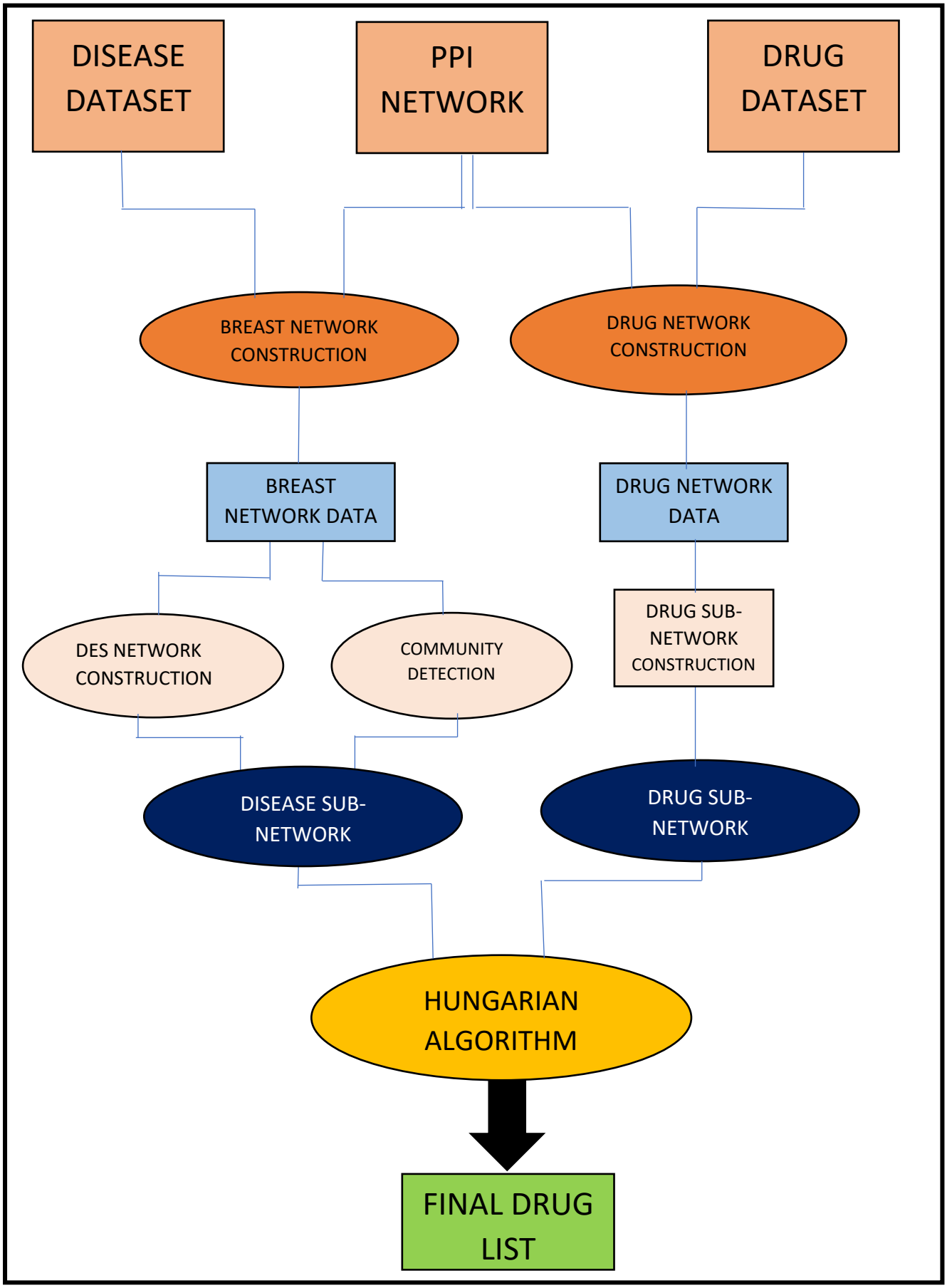


Figure 3.5: Thesis pipeline

In the first method, we aim to find a Differentially Expressed Subnetwork from the disease network data. A gene is declared differentially expressed if an observed difference or change in read counts or expression levels between two experimental conditions is statistically significant [52]. A Differentially Expressed subnetwork is a disease-related subnetwork of differentially expressed interacting genes identified by an appropriate integration of a secondary network data with the primary gene expression data. In our second method, we use community detection algorithms to find the communities, or clusters in the disease and the drug network data. Community detection in networks is one of the most popular topics of modern network science. Communities or clusters are usually groups of vertices having higher probability of being connected to each other than to members of other groups, though other patterns are possible [44]. We will see more about these two methodologies used in this thesis later in this chapter.

3.4.1 Hungarian Algorithm

The standard assignment problem is referred to as the problem to find a one-to-one matching between tasks and agents, in order to optimize the total cost of the assignments. The objective is either to maximize or minimize the total cost. In this thesis we wish to find the optimal assignment of rank of drugs, by maximising the total cost. The classical example of assignment problems is assigning jobs to workers. Hungarian method is the most popular method which solves the assignment problem in polynomial time. It was developed and published by Harold Kuhn in 1955.

The Hungarian Method is based on the principle that if a constant is added to every element of a row and/or a column of cost matrix, the optimum solution of the resulting assignment problem is the same as the original problem and vice versa. The original cost matrix can be reduced to another cost matrix by adding constants to the elements of rows and columns where the total cost or the total completion time of an assignment is zero. Since the optimum solution remains unchanged after this reduction, this assignment is also the optimum solution of the original problem [48].

The Hungarian algorithm consists of the four steps:

Step 1 (Subtract row minima): In the cost-matrix, for each row, the lowest element is subtracted from each element in that row.

Step 2 (Subtract column minima): Similarly, for each column, the lowest element is subtracted from each element in that column.

Step 3 (Cover all zeros with a minimum number of lines): Then all the zeros in the resulting matrix is covered using a minimum number of horizontal and vertical lines. If n lines are required, an optimal assignment exists among the zeros. The algorithm stops. If less than n lines are required, Step 4 is continued.

Step 4 (Create additional zeros): The smallest element (call it k) that is not covered by a line in Step 3 is subtracted from all uncovered elements and k is added to all elements that are covered twice.

A Walk-Through Algorithm:

We consider an example where five salesmen (1, 2, 3, 4,5) need to be assigned to five districts (A, B, C, D, E), one salesman per district. The matrix below shows the cost of assigning a certain worker to a certain district. The objective is to maximize the total cost of the assignment.

Cost Matrix:

		District				
		A	B	C	D	E
Salesman	1	32	38	40	28	40
	2	40	24	28	21	36
	3	41	27	33	30	37
	4	22	38	41	36	36
	5	29	33	40	35	39

Conversion to Minimization Problem:

The given maximization problem is converted into minimization problem by subtracting from the highest sales value (i.e., 41) with all elements of the given table.

		District				
		A	B	C	D	E
Salesman	1	9	3	1	13	1
	2	1	17	13	20	5
	3	0	14	8	11	4
	4	19	3	0	5	5
	5	12	8	1	6	2

Matrix Reduced Row-wise

Reduce the matrix row-wise

		District				
		A	B	C	D	E
Salesman	1	8	2	0	12	0
	2	0	18	12	19	4
	3	0	14	8	11	4
	4	19	3	0	5	5
	5	11	7	0	5	1

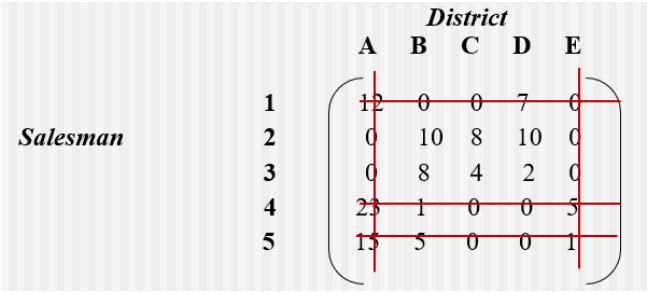
Step 2: Matrix Reduced Column-wise and Zeros Covered

Reduce the matrix column-wise and draw minimum number of lines to cover all the zeros in the matrix, as shown below.

		District				
		A	B	C	D	E
Salesman	1	8	0	0	7	0
	2	0	14	12	14	4
	3	0	12	8	6	4
	4	19	1	0	0	5
	5	11	7	0	0	1

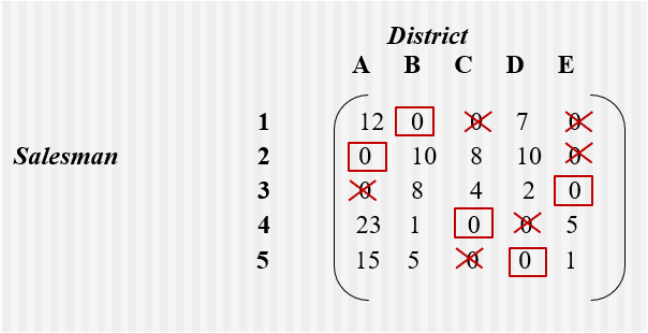
Step 3: Add & Subtract the least Uncovered Element:

Number of lines drawn ≠ Order of matrix. Hence not optimal. Select the least uncovered element, i.e., 4 and subtract it from other uncovered elements, add it to the elements at intersection of line and leave the elements that are covered with single line unchanged.



Step 4: Final Assignments

Now, number of lines drawn = Order of matrix, hence optimality is reached.



Thereby, the salesman 1 is assigned to district B, salesman 2 is assigned A, salesman 3 is assigned to E, salesman 4 is assigned to C and salesman 5 is assigned to district D.

3.4.2 Differentially Expressed Subnetwork Method

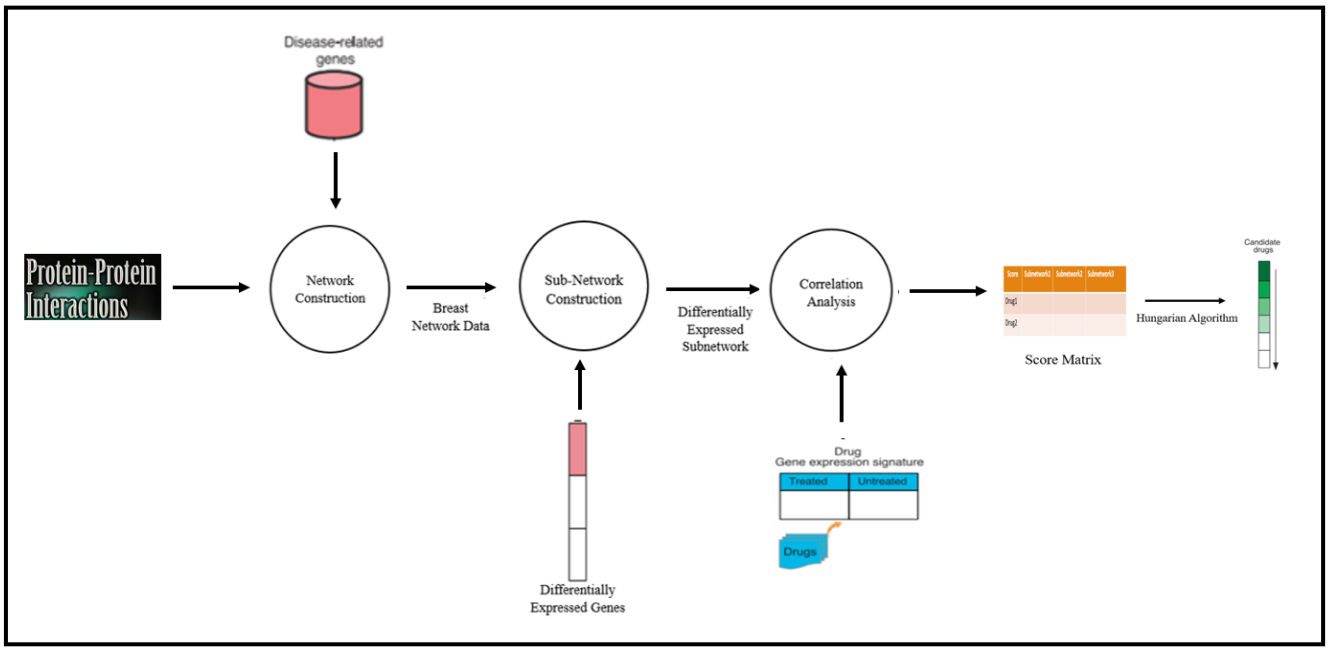


Figure 3.6: Proposed Framework – DES method

In biology, a biomarker is a measurable indicator of the severity or presence of some disease state. More generally a biomarker is anything that can be used as an indicator of a particular disease state or some other physiological state of an organism [41]. In the literature survey, we found that all the papers were using a bioinformatic methods that focuses on identifying biomarkers as small subsets of differentially expressed genes. Differentially expressed genes (DEG's) have limited predictive performance due to the heterogeneity within tumour samples and across patients, moreover insufficient patient sample size and the inherent measurement noise in microarray experiments makes the biomarkers with DEG's unstable [42].

Also, computational methods detecting DEG's do not consider the dependencies or relationships between genes in order to accurately classify the sample data, thus identified biomarker set may contain many DEG's with redundant information yielding decreased prediction performance.

So, to accurately identify effective biomarkers, new bioinformatic methods integrating additional biological information with gene expression data have become necessary [31].

In this thesis, we aim to identify a Differentially Expressed Subnetwork (DES) as an effective biomarker that could help us to find the best candidate drugs for repurposing. Figure 3.5 represents the proposed framework, which involves finding the Differentially Expressed Subnetwork. We have obtained the Differentially Expressed Subnetwork with the help of the Breast Network Data and the differentially expressed genes of the disease data. The Differentially Expressed subnetwork is obtained for each of the DEG's in the disease data, so 138 Differentially Expressed Subnetwork is obtained. Figure 3.6 explains the process of finding the Differentially Expressed Subnetwork.

Starting from a DEG V, the search for the Differentially Expressed Subnetwork proceeds as follows:

- The current aggregate N, initially contains only the differentially Expressed gene V.
- We iteratively aggregate its neighbour nodes U in a greedy manner using Breadth First Search Algorithm.
- A neighbour u is inserted into the current aggregate N if and only if its inclusion increases the correlation between the expression of the genes in the aggregate.
- |"correlation (N + u) – correlation (N)" | > Δ, where Δ is 0.001.
- Then, the same process is repeated on the new aggregate N + u and the process continue till the level 2 neighbours are evaluated.

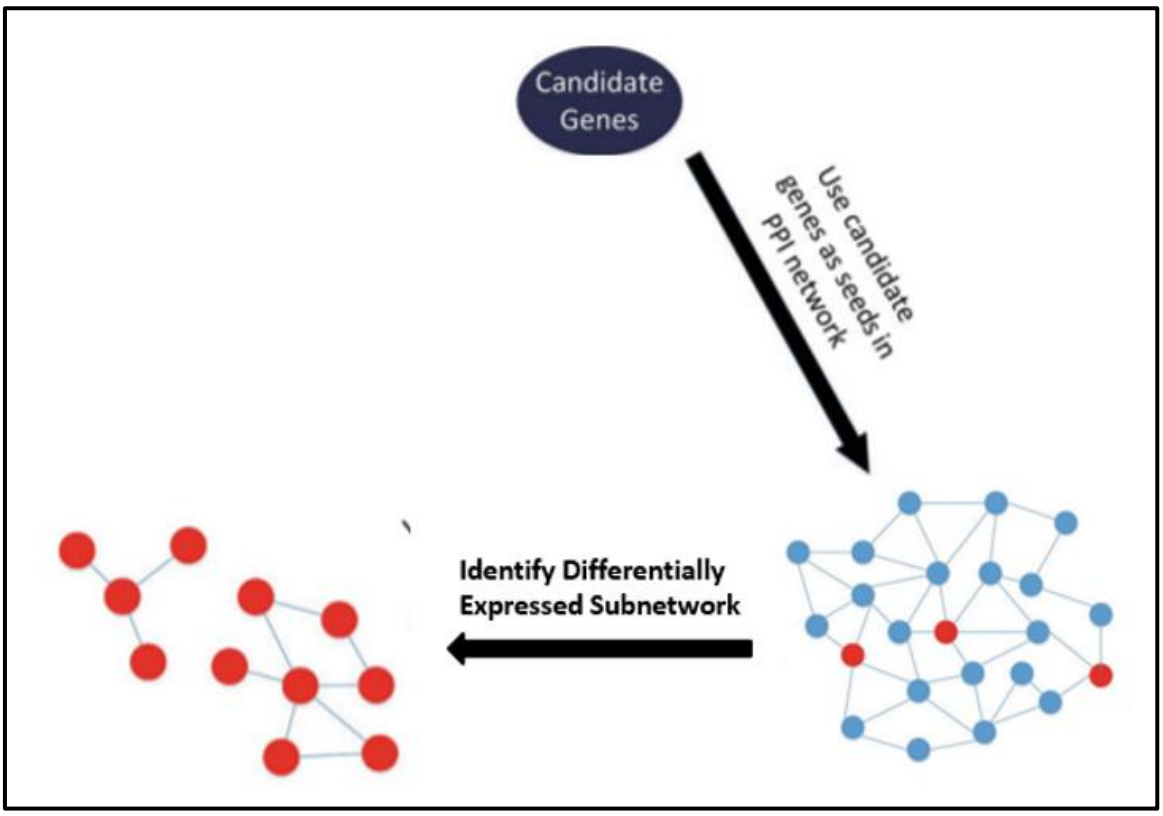


Figure 3.7: Framework for finding Differentially Expressed Subnetwork

After obtaining the Differentially Expressed Subnetwork, the next step was to check for the correlation of genes, between each of the Differentially Expressed Subnetwork and the drug data. For this we have used the ABC model of Network-based method for Drug Repositioning/Repurposing. The ABC model is based on the idea that if a drug perturbed gene expression profile inversely correlates with the disease gene expression profile, the drug may have a curing effect on the disease.

Generally, suppose we know through a data source that a disease C has a certain characteristic B i.e., disease C is caused by a downregulation of gene B and that a compound A has some effect on B i.e., drug A restores the expression of B. Then, we can infer that A will influence C i.e., drug A is a repositioning candidate for disease C [43].

Table 3.2 shows the anti-correlation labels and the correlation which makes the drug, a suitable candidate for repurposing. Down-regulation indicates a decrease in the production of that gene as an effect of the disease. Up-regulation indicates an increase in the production of that particular gene as an effect of the disease.

Disease profile	Drug profile	Correlation label	
Down-regulation	Down-regulation	Direct Correlation	✗
Down-regulation	Up-regulation	Inverse Correlation	✓
Up-regulation	Down-regulation	Inverse Correlation	✓
Up-regulation	Up-regulation	Direct Correlation	✗

Table 3.2: Correlation Table

For each Differentially Expressed Subnetwork(DES), a percentage score is given based on the total number of genes that are inversely correlated between the DES and the drug data i.e. if no genes are inversely correlated between a DES and the drug data, then it is scored 0 and if all the genes are inversely correlated then it is scored 100. Then finally Hungarian Algorithm is applied on the obtained score matrix to get the ranked list of drugs which act as the potential candidate for drug repurposing/repositioning.

3.4.3 Louvain Community Detection Method

In this thesis, in our second method we have used the community detection algorithm on the disease network data followed by correlation analysis to find the repurposing score and the combinatorial optimization algorithm to rank the drugs, based on the repurposing score. We finally have a list of drugs ranked from potentially best suited drug repurposing candidates for the disease breast cancer to potentially less effective drug repurposing candidates.

There are several types of algorithms used for community detection. In this thesis we have used the Louvain community Detection Method (LDM) for detecting communities in networks. Figure 3.7 represents the proposed framework of the second method.

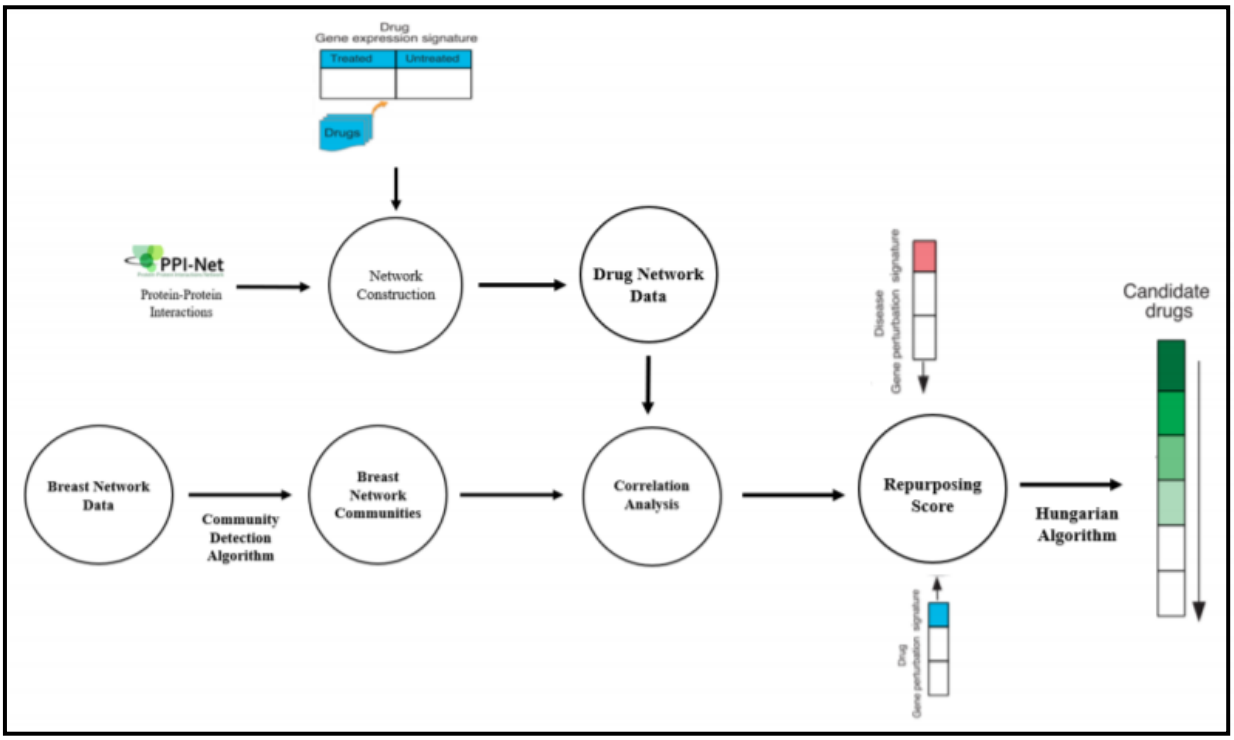


Figure 3.8: Proposed Framework – LCD Method

The Louvain method for community detection is an algorithm for detecting where the modularity quantifies the quality of an assignment of nodes to communities. Modularity is defined as a measure of the structure of networks or graphs. It was designed to measure the strength of division of a network into modules (also called groups, clusters or communities) [45]. The Louvain clustering algorithm is illustrated in figure 3.8 [46].

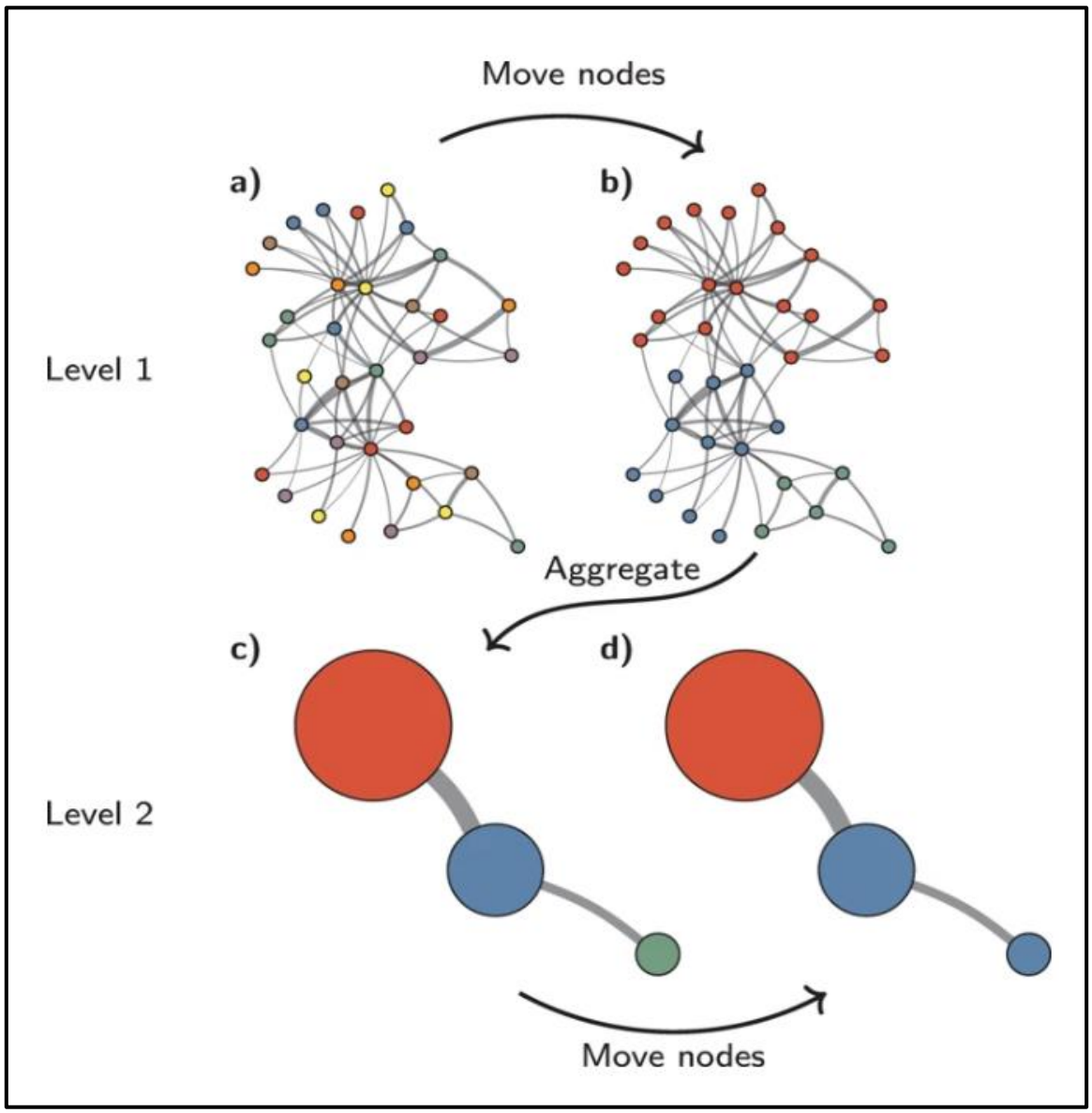


Figure 3.9:Louvain Clustering Algorithm

The algorithm is:

- The Louvain algorithm starts from a singleton partition in which each node is in its own community. The algorithm moves individual nodes from one community to another to find a partition.
- Based on the obtained partition, an aggregate network is created.
- The algorithm then moves individual nodes in the aggregate network.
- These steps are repeated until the quality cannot be increased further.

The algorithm optimizes a quality function such as modularity in two elementary phases. The first phase involves local moving of nodes and the second phase involves the aggregation of the network. In the local moving phase, individual nodes are moved to the community that yields the largest increase in the quality function. In the aggregation phase, an aggregate network is created based on the partition obtained in the local moving phase. Each community in this partition becomes a node in the aggregate network. The two phases are repeated until the quality function cannot be increased further [47].

We applied the Louvain community detection algorithm on the Breast Network data and obtained 14 communities. Then the correlation analysis was performed on each of the drug network data and the corresponding disease network communities. In this method, we have computed two repurposing scores based on the correlation.

Case 1:

In the first case, the score is computed based on the total number of genes that are inversely correlated between each of Breast network communities and the drug networks. Figure 3.9 explains the computation of the inversely corelated gene score. In fig 3.9, Circle A consist of the disease genes and circle B consist of the drug genes. Genes A, B, D and G are in common between the disease and the drug genes. Among the common genes, the genes which are inversely correlated are highlighted in red colour. These inversely correlated common genes are used to calculate the first score.

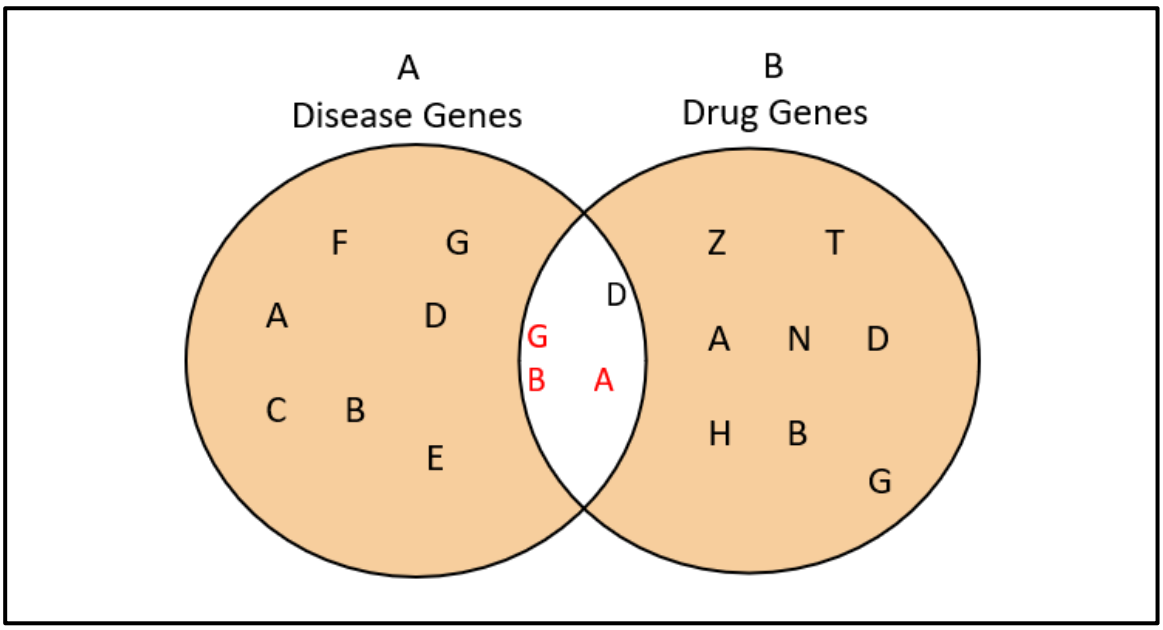


Figure 3.10: Gene Score calculation

Equation 3.1 shows the formula used to calculate the inversely corelated gene score.

$$score_1 = \frac{\#IC_{genes}}{|N|}$$

Where,

N – Total number of genes in common between the disease and the drug data.

IC_{genes} – Total number of inversely correlated genes among the common genes between the disease and the drug data.

Case 2:

In the second case, the score is computed based on the total number of interactions that are inversely correlated between each of Breast network communities and the drug networks. Figure 3.10 explains the computation of the inversely corelated edge score. In fig 3.10, Circle A consist of the interactions in the disease network and circle B consist of the interactions in the drug network. Interactions A B, C → D, X → Z are in common between the disease and the drug network.

Initially, each interaction is scored based on the number of genes that are inversely correlated in an interaction. Genes which are inversely correlated among the common interactions are highlighted in red colour. If both the genes in an interaction are inversely correlated, then the interaction is scored 2. If any one of the genes in an interaction is inversely correlated then the interaction is scored 1 and if no genes in an interaction is inversely correlated, then it is scored 0.

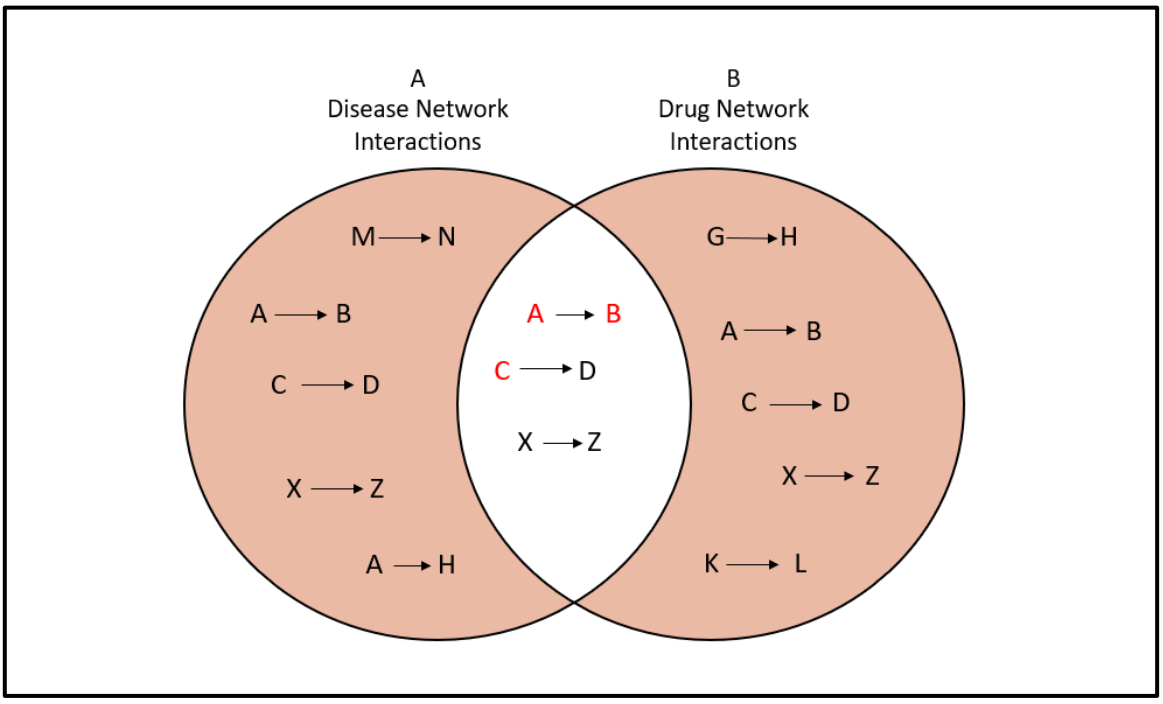


Figure 3.11: Edge Score Calculation

Equation 3.2 shows the formula used to calculate the inversely corelated edge score.

$$score_2 = \frac{1}{\#E}\sum_{1}^{\#E}\frac{s_e}{2}$$

Where,

$\boldsymbol{E}$ = Total number of common interactions between the disease network and the drug network.

$\boldsymbol{s_e}$ = The score of each common interaction.

$\boldsymbol{s_e}$ = 0 if both the genes are not inversely correlated.

$\boldsymbol{s_e}$ = 1 if one of the genes are inversely correlated.

$\boldsymbol{s_e}$ = 2 if both the genes are inversely correlated.

Then the final repurposing score between each drug and disease data is computed by taking average of the Gene score and edge score. Finally, we applied the Hungarian algorithm to the obtained score matrix and found the list of ranked drugs from potentially best suited drug repurposing candidates for the disease breast cancer to potentially less effective drug repurposing candidates.

Chapter 4

Results and Discussion

In this chapter, we shall go through the results of both the proposed methods for the disease Breast Cancer and compare the results obtained. The results showcase several unapproved drugs alongside approved drugs closest to the disease indicating that the unapproved drugs share similarities with the approved drugs which means that they are worth pursuing for repurposing.

4.1 Results

The following tables shows the ranked top 10 drugs for the disease breast cancer obtained using our proposed methods. We selected the top 10% drugs from the drugs lists obtained by applying the proposed methods on the breast cancer datasets. These drugs are ranked according to the repurposing scores computed by the systematic method from the highest to the lowest. We have used online drug databases such as Drug Bank [49] to obtain each drugs' FDA status. Table 4.1 shows the list of ranked drugs obtained by the DES method. The ranked list of drugs obtained from the DES method for the disease breast cancer, comprises of 4 approved drugs and 6 unapproved drugs. Table 4.2 shows the list of top 10 drugs obtained from our LCD method . The ranked list of drugs obtained from our LCD method for the disease breast cancer, comprises of 8 approved drugs and 2 unapproved drugs.

Rank	Drugs	FDA Status
1	Pralatrexate	Approved
2	BIIB 021	Investigational
3	Idasanutlin	Investigational
4	Genz-644282	Approved
5	Inositol	Approved
6	Bardoxolone methyl	Investigational
7	sitagliptin	Approved
8	combretastatin A4	Investigational
9	AS703026	Investigational
10	CYT997	Investigational

Table 4.1: Ranked list of Drugs – DES Method

Rank	Drugs	FDA Status
1	Daunorubicin	Approved
2	Mepivacaine	Approved
3	Mitoxantrone	Approved
4	Ixazomib citrate	Approved
5	AT-7519	Investigational
6	IKK2-inhibitor-V	Approved
7	R-547	Investigational
8	Genz-644282	Approved
9	Sorafenib	Approved
10	L-ergothioneine	Approved

Table 4.2: Ranked list of Drugs – LCD Method

4.2 Discussion

Tamoxifen, Gemcitabine, Epirubicin, Exemestane, Capecitabine, Paclitaxel, Doxorubicin, gemcitabine, Fulvestrant, Exemestane, Neratinib, Docetaxel are some of the FDA approved drugs for the disease breast cancer. These drugs were included in our list of input drugs from the LINCS drug dataset to validate our proposed methods. Table 4.3 shows the validation results for both the proposed methods. FDA approved drugs for the disease breast cancer are highlighted in green colour. Both our proposed methods were able to find 8 out of the 10 FDA approved drugs for the disease breast cancer.

Rank	Method 1	Method 2
1	Gemcitabine	Capecitabine
2	Exemestane	Paclitaxel
3	Paclitaxel	Mepivacaine
4	Triptolide	Doxorubicin
5	Tamoxifen	Tofacitinib
6	Capecitabine	Fulvestrant
7	Doxorubicin	Gemcitabine
8	Neratinib	Exemestane
9	Fulvestrant	Neratinib
10	Lacidipine	Docetaxel

Table 4.3: Validation Results

Some of our proposed drugs are already in the clinical study for the treatment of breast cancer. For example, one of our proposed drugs by method 2, **Sorafenib**, marketed as Nexavar by Bayer, is a drug approved for the treatment of advanced renal cell carcinoma (primary kidney cancer). A recent phase II study in 229 human epidermal growth factor receptor 2 (HER2)-negative metastatic breast cancer patients investigated with the combination of sorafenib and capecitabine inhibited the proliferation of breast cancer [50].

Daunorubicin, one of the proposed drugs by method 2 is an anthracycline used in treatment of leukemia. Clinical studies proved that Human DNA TOP2A is a marker of cell proliferation in breast cancer. Based on this evidence, Daunorubicin which inhibits Human DNA TOP2A may have a potential therapeutic effect on breast cancer. This hypothesis is under phase I clinical study evaluating the effectiveness of Daunorubicin in treating breast cancer patients (ClinicalTrials.gov identifier: NCT00004207). **Mepivacaine**, a local anaesthetic that is chemically related to bupivacaine but pharmacologically related to lidocaine. It is indicated for infiltration, nerve block, and epidural anaesthesia. In a recent study, human breast cancer cell lines, MDA-MB-231 and MCF7, were incubated with mepivacaine and found that the high concentration of mepivacaine, significantly inhibited the breast cancer cell survival [51].

Chapter 5

Conclusion and Future Work

In this thesis, we aimed to find suitable drug repurposing candidates for the disease breast cancer using the Network-based method. We used Reactive Stroma of Breast and Prostate cancer disease datasets, LINCS drug datasets and the Protein-Protein interaction (PPI) networks from the "Pathway Commons Protein-Protein Interactions database" and the "STRING" database. We performed a series of pre-processing steps on these datasets and proposed two different methods for achieving the drug repurposing/repositioning for the disease breast cancer. In our first method, we have discussed methods to identify a Differentially Expressed Subnetwork as an effective biomarker that helped us to find the best candidate drugs for repurposing. We also discussed the ABC model of Network based drug repurposing/repositioning.

In our second method, our proposed frameworks constructs the drug network data and we have used the community detection algorithm on the disease network data followed by correlation analysis on the disease communities and the drug network data to find the repurposing score followed by the usage of combinatorial optimization algorithm to rank the drugs, based on the repurposing score.

5.1 Possible Future Work

This thesis is just a small step towards Drug repurposing and Drug repositioning. There are many directions for future research. Future work that can be conducted includes the following: -

- Our pre-processing steps and methods can be applied on a different cancer dataset such as prostate cancer.
- Using side-effect similarity of unapproved drugs with that of approved drugs, drug repurposing candidates can be obtained.

These ideas can be an open problem that can be explored in the future

Bibliography

[1] H.P., Rang; M.M, Dale; J.M., Ritter; R.J., Flower; G., Henderson (2011). "What is Pharmacology". Rang & Dale's pharmacology (7th ed.). Edinburgh: Churchill Livingstone. p. 1.

[2] "Drug". Dictionary.com Unabridged. v 1.1. Random House. 20 September 2007. Archived from the original on 14 September 2007 – via Dictionary.com.

[3] "Refinement of Ibuprofen at 100K by Single-Crystal Pulsed Neutron Diffraction". Acta Crystallographica Section C 53: 951-954. DOI:10.1107/S0108270197003193.

[4] J P Hughes, S Rees, S B Kalindjian, and Karen L Philpott. Principles of early drug discovery. British journal of pharmacology, 162(6):1239–1249, 2011.

[5] A. P. M. D. A. a. S. D. Nafiseh Saberian, "A new computational drug repurposing method using established disease–drug pair knowledge," Bioinformatics, no. Oxford University Press., pp. 1-7, 2019.

[6] P. Agrawal*, "Advantages and Challenges in Drug 5e-3rofilLng," Journal of Pharmacovigilance, vol. S2, no. e002, p. 2, 2015.

[7] T I Oprea, Julie E Bauman, Cristian G Bologa, Tione Buranda, Alexandre Chigaev, Bruce S Edwards, Jonathan W Jarvik, Hattie D Gresham, Mark K Haynes, Brian Hjelle, et al. Drug repurposing from an academic perspective. Drug Discovery Today: Therapeutic Strategies, 8(3-4):61–69, 2011.

[8] N. G. S. R. M. V. a. J. R. G. Maryam Lotfi Shahreza, “A review of network-based approaches to drug repositioning,” Briefings in Bioinformatics, pp. 1-15, 2017.

[9] Morris GM, Lim-Wilby M. Molecular docking. In: A Kukol (ed).Molecular Modeling of Proteins. Totowa, NJ: Humana Press,2008, 365–82.

[10] Dudley JT, Deshpande T, Butte AJ. Exploiting drug–disease relationships for computational drug repositioning. Brief Bioinform 2011;12:303–11.

[11] Ding H, Takigawa I, Mamitsuka H, et al. Similarity-based machine learning methods for predicting drug-target interactions: a brief review. Brief Bioinform 2014;15:734–47.

[12] Zou J, Zheng MW, Li G, et al. Advanced systems biology methods in drug discovery and translational biomedicine. Biomed Res Int 2013;2013:8

[13] S. &. P. A. Alaimo, “Network-Based Drug Repositioning: Approaches, Resources, and Research Directions.,” Computational Methods for Drug Repurposing, p. 97–113, 2018.

[14] Jiang Z, Zhou Y. Using gene networks to drug target identification. J Integr Bioinform 2005;2:14.

[15] Dai YF, Zhao XM. A survey on the computational approaches to identify drug targets in the postgenomic era. Biomed Res Int 2015;2015:9.

[16] .Arrell DK, Terzic A. Network systems biology for drug discovery. Clin Pharmacol Ther 2010;88:120–5

[17] Azuaje F. Drug interaction networks: an introduction to translational and clinical applications. Cardiovasc Res 2013;97:631–41

[18] National Institutes of Health. What is a gene? - genetics home reference - nih, https://ghr.nlm.nih.gov/primer/basics/gene, 2019. Last accessed March, 2019.

[19] Gene expression data, https://compbio.soe.ucsc.edu/genex/expressdata. html, 2019. Last accessed March 2019.

[20] American Cancer Society. What is breast cancer? — breast cancer definition, https://www.cancer.org/cancer/breast-cancer/about/ what-is-breast-cancer.html, 2019. Last accessed March, 2019.

[21]Breast cancer statistics - canadian cancer society, http://www.cancer.ca/en/cancer-information/cancer-type/breast/statistics/?region=on, 2019.Last accessed March, 2019.

[22]Ou-Yang, Si-Sheng et al. "Computational drug discovery." Acta pharmacologica Sinica vol. 33,9 (2012): 1131-40. doi:10.1038/aps.2012.109

[23]Song CM, Lim SJ, Tong JC. Recent advances in computer-aided drug design. Brief Bioinform. 2009;10:579–91.

[24]Shekhar C. *In silico* pharmacology: computer-aided methods could transform drug development. Chem Biol. 2008;15:413–4.

[25]Jorgensen WL. The many roles of computation in drug discovery. Science. 2004;303:1813–8.

[26] Tan JJ, Cong XJ, Hu LM, Wang CX, Jia L, Liang XJ. Therapeutic strategies underpinning the development of novel techniques for the treatment of HIV infection. Drug Discov Today.

[27] Li, Jiao., & Lu, Zhiyong. A Network Approach for Computational Drug Repositioning. IEEE Second International Conference on Healthcare Informatics, Imaging and Systems Biology, 2012.

[28] Nafiseh Saberian, Azam Peyvandipour, Michele Donato, Sahar Ansari, and Sorin Draghici. A new computational drug repurposing method using established disease-drug pair knowledge. Bioinformatics, 2019

[29] Azam Peyvandipour, Nafiseh Saberian, Adib Shafi, Michele Donato, and Sorin Draghici. A novel computational approach for drug repurposing using systems biology. Bioinformatics, 34(16):2817–2825, 2018.

[30] Haeseung Lee, Seungmin Kang, and Wankyu Kim. Drug repositioning for cancer therapy based on large-scale drug-induced transcriptional signatures. PloS one, 11(3):e0150460, 2016

[31] Jinyan Chan, Xuan Wang, Jacob A Turner, Nicole E Baldwin, and Jinghua Gu. Breaking the paradigm: Dr insight empowers signature-free, enhanced drug repurposing. Bioinformatics, 2019.

[32] Koleti, A., Terryn, R., Stathias, V., Chung, C., Cooper, D. J., Turner, J. P., … Schürer, S. C. (2017). Data Portal for the Library of Integrated Network-based Cellular Signatures (LINCS) program: integrated access to diverse large-scale cellular perturbation response data. Nucleic Acids Research, 46(D1), D558–D566.

[33] Benjamini, Yoav; Hochberg, Yosef (1995). "Controlling the false discovery rate: a practical and powerful approach to multiple testing" (PDF). Journal of the Royal Statistical Society, Series B.

[34] Akey, J. (n.d.). Lecture 10: Multiple Testing. Article posted on the University of Washington website. Retrieved October 29, 2017 from:http://www.gs.washington.edu/academics/courses/akey/56008/lecture/lecture10.pdf.

[35] Anat Reiner, Daniel Yekutieli, and Yoav Benjamini. Identifying differentially expressed genes using false discovery rate controlling procedures. Bioinformatics, 19(3):368–375, 2003.

[36] R. Singh, J. Xu, and B. Berger, “Pairwise global alignment of protein interaction networks by matching neighborhood topology,”in Proc. 11th Annu. Int. Conf. Res. Comput. Molecular Biol., 2007,pp. 16–31.

[37] Kuchaiev, O., Przulj, N. Global Network Alignment. Nat Prec (2010)

[38] Rouillard AD, Gundersen GW, Fernandez NF, Wang Z, Monteiro CD, McDermott MG, Ma'ayan A. *The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins*. Database (Oxford). 2016 Jul 3;2016. pii: baw100.

[39] Szklarczyk, Damian; Gable, Annika L.; Lyon, David; Junge, Alexander; Wyder, Stefan; Huerta-Cepas, Jaime; Simonovic, Milan; Doncheva, Nadezhda T.; Morris, John H.; Bork, Peer; Jensen, Lars J. (8 January 2019). "STRING v11: protein-protein association networks with increased coverage, supporting functional discovery in genome-wide experimental datasets". Nucleic Acids Research. 47 (D1): D607–D613.

[40] Szklarczyk, Damian; Morris, John H.; Cook, Helen; Kuhn, Michael; Wyder, Stefan; Simonovic, Milan; Santos, Alberto; Doncheva, Nadezhda T.; Roth, Alexander; Bork, Peer; Jensen, Lars J. (4 January 2017). "The STRING database in 2017: quality-controlled protein-protein association networks, made broadly accessible". Nucleic Acids Research. 45 (D1): D362–D368.

[41] "Biomarker Technology Platforms for Cancer Diagnoses and Therapies". TriMark Publications, LLC. July 2014.

[42] Firoozbakht, F., Rezaeian, I., D'agnillo, M., Porter, L., Rueda, L., & Ngom, A. (2017). An Integrative Approach for Identifying Network Biomarkers of Breast Cancer Subtypes Using Genomic, Interactomic, and Transcriptomic Data. Journal of Computational Biology, 24(8), 756–766

[43] S. &. P. A. Alaimo, "Network-Based Drug Repositioning: Approaches, Resources, and Research Directions.," *Computational Methods for Drug Repurposing,* p. 97–113, 2018

[44] Fortunato, S., & Hric, D. (2016). Community detection in networks: A user guide. Physics Reports, 659, 1–44.

[45] Newman, M. E. J. (2006). "Modularity and community structure in networks". Proceedings of the National Academy of Sciences of the United States of America. 103 (23): 8577–8696.

[46] Traag, V.A., Waltman, L. & van Eck, N.J. From Louvain to Leiden: guaranteeing well-connected communities. Sci Rep 9, 5233 (2019) doi:10.1038/s41598-019-41695-z.

[47] Blondel, V. D., Guillaume, J.-L., Lambiotte, R. & Lefebvre, E. Fast unfolding of communities in large networks. J. Stat. Mech. Theory Exp. 10008,6, https://doi.org/10.1088/1742-5468/2008/10/P10008 (2008).

[48]"O'Reilly Media, Inc," 2019. [Online]. Available: https://learning.oreilly.com/library/view/quantitative-techniques-theory/9789332512085/xhtml/ch4sec2.xhtml. [Accessed 12 2019].

[49] Drugbank. Drugbank, https://www.drugbank.ca/, 2019. Last accessed March, 2019.

[50] Loibl, S., Rokitta, D., Conrad, B., Harbeck, N., Wüllner, M., Warm, M., … von Minckwitz, G. (2014). Sorafenib in the Treatment of Early Breast Cancer: Results of the Neoadjuvant Phase II Study - SOFIA*. Breast Care, 9(3), 169–174. doi:10.1159/000363430

[51] Li, R., Xiao, C., Liu, H., Huang, Y., Dilger, J. P., & Lin, J. (2018). Effects of local anesthetics on breast cancer cell viability and migration. BMC Cancer, 18(1). doi:10.1186/s12885-018-4576-2

[52] Anjum A, Jaggi S, Varghese E, Lall S, Bhowmik A, Rai A. Identification of Differentially Expressed Genes in RNA-seq Data of Arabidopsis thaliana: A Compound Distribution Approach. J Comput Biol. 2016;23(4):239–247. doi:10.1089/cmb.2015.0205

Vita Auctoris

NAME	Pavithra Ulaganathan
PLACE OF BIRTH	Coimbatore, Tamil Nadu
EDUCATION	Bachelor of Technology in Computer Science, Amrita University, Coimbatore, Tamil Nadu, India, 2017 Master of Science in Computer Science University of Windsor, Windsor, ON, Canada, 2019