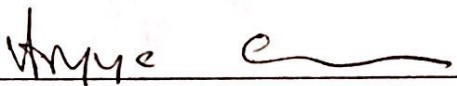


APPROVAL SHEET

Title of Dissertation: Network Analytics towards Drug Repositioning using
Phosphorylated Proteins

Name of Candidate: Iyanuoluwa E Odebode
Doctor of Philosophy, 2019

Dissertation and Abstract Approved: _____


Dr Aryya Gangopadhyay
Professor
Department of Information Systems

Date Approved: 4/29/19

ABSTRACT

Title of dissertation: NETWORK ANALYTICS TOWARDS DRUG
REPOSITIONING: USING
PHOSPHORYLATED PROTEINS

Iyanuoluwa E Odebode, Doctor of Philosophy, 2019

Dissertation directed by: Dr. Aryya Gangopadhyay
Department of Information Systems

Drug Repositioning is an approach to discovering a new use of old drugs. However, current successes in drug repositioning have primarily been a result of serendipity or clinical observations, [1] such as the observed use of sildenafil citrate (Viagra) mostly for the treatment of erectile dysfunction, but now repositioned for the treatment of pulmonary arterial hypertension, [2], leprosy [3], and erectile dysfunction induced depression.

Besides, thalidomide used for inducing sedation is now known to be important for the treatment of multiple myeloma [1]. To transform this process systematically, many computational approaches have emerged with recent advances in computational technology, to automatically identify possible drug repositioning candidates by accessing an overwhelmingly volume of biomedical data [1]. In this study, we described an effort made on computational drug repositioning by applying sequence encoding, sequence analysis, and network analytics against the phosphorylated proteins.

We propose a novel framework for computational drug repositioning with multiple components, 1) Sequence analysis and sequence prioritization 2) biological interaction network construction by integrating heterogeneous interactions among gene, disease, protein, SNP, etc.; 2) Phosphorylated Network creation 3) high-influence node detection/prioritization by applying network analysis and perturbation; 4) Multimodal Network creation and clustering 5) drug candidate identification and evaluation by using sequence analysis and cheminformatics techniques. We used the PIM Substrates and rhabdomyosarcoma as a case study.

NETWORK ANALYTICS TOWARDS DRUG REPOSITIONING USING PHOSPHORYLATED PROTEINS

by

Iyanuoluwa E Odebode

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
2019

Advisory Committee:

Dr. Aryya Gangopadhyay

Dr. Charles Bieberich

Dr. Vandana Janeja

Dr. Nirmalya Roy

Dr. Jiaqi Gong

© Copyright by
Iyanuoluwa E Odebode
2019

Dedication

I dedicate my dissertation to my precious wife. Thank you for staying through with me and supporting me through the process of writing my dissertation. Thanks for all your encouragement, care, and dedication in making sure I did the best that I possibly can.

Acknowledgments

I am very grateful to Dr. Aryya Gangopadhyay, my mentor and committee chair for his words of advice and encouragement during the process of my dissertation. Thank you so much for your time, your dedication to ensuring that I put the best of my effort into completing well. I am also very grateful to the members of my committee, Dr. Vandana Janeja, Dr. Nirmalya Roy, Dr. Charles Bieberich, and Dr. Jiaqi Gong. I appreciate your time and effort in making sure the process of my dissertation went well. Thank you for your constant advice and inputs towards ensuring a successful completion. Thanks also to Xiang Li from the Bieberich Lab, who was always available to help direct me when it came to the biological portion of my experiment. To my colleagues at the data analytics lab thanks for making a positive impact on my life as a graduate student. The Bieberich Lab also have been such a great support to me as a graduate student. I want to say a special thank you to my family Gabi Odebode, Joy Odebode, Sam Odebode, and John Odebode for their support, for enduring with me during the process of the completion of my dissertation. Thanks for your words of encouragement. I am very grateful. Thank you, Gabi, for being there all the time and supporting me and standing in for me in places I could not go because I was busy working on my dissertation.

I am also very grateful for my extended family members, my parents, brothers, and in-laws for their constant encouragement to push through when things seemed difficult.

Contents

List of Tables	x
List of Figures	xi
1 Introduction	1
1.1 Challenges in Drug Repositioning	1
1.2 Challenges in Cancer Research	4
1.3 PIM Kinase in Cancer Research	10
1.4 Motivation	14
1.5 Summary of our Contributions	15
1.6 Organization of the Dissertation	15
2 Literature Review	17
2.1 Drug Repositioning	17
2.1.1 The traditional drug repositioning approach	18
2.1.1.1 Target-based screening	19
2.1.1.2 Phenotypic drug screening	19
2.1.1.3 Limitation of the Approaches	20
2.1.1.4 Other challenges faced include both legal and scientific battles	21
2.1.1.5 Successful drug repositioning: Stories	22
2.1.2 Computational Drug Repositioning	24
2.1.2.1 Computational Strategies to Drug Repositioning	26
2.1.2.2 Phenome	26
2.1.2.3 Genome	27
2.1.2.4 Drug chemical structures/ Drug Combinations	28
2.1.3 Computational Approaches to Drug Repositioning	29
2.1.3.1 Sequence based methods	30
2.1.3.2 Previous work using Sequence-based method for identification of Phosphorylation sites	30
2.1.3.3 Phenotypic and side effect based approaches:	36
2.1.3.4 Network based Approach	37
2.1.3.5 Network based Approach: Disease	38
2.1.3.6 Machine Learning Approach	39
2.1.3.7 Text-mining Approach: EHR Data for Analysis	40
2.1.3.8 Text-mining Approach: Social Media Data	41

2.1.3.9	Chemical based Approach	42
2.1.3.10	Protein structure approach	43
2.1.3.11	Computational Approaches: Success Stories	44
2.1.3.12	Limitations of existing computational approaches.	45
3	Methodology	47
3.1	Research Contributions	47
3.1.1	Introduction to our approach	49
3.1.2	Overall system architecture	50
3.2	Datasets	51
3.2.1	Rhabdomyosarcoma Datasets	51
3.2.2	Diabetes Datasets	52
3.2.3	PIM Datasets	52
3.2.4	Experimental Setup	52
3.3	Identification of Relevant Protein Sequence(PhoSc)	53
3.3.0.1	Sequence Data Preprocessing and Sequence slicing.	54
3.3.0.2	Background in Bioinformatics	56
3.3.0.3	Encoding Scheme:	58
3.3.0.4	Relevance of Sequences using Consensus Sequence	60
3.3.0.5	Relevance of Sequences using Position Scoring Matrix	62
3.3.0.6	Term-Document Matrix and Adjacency Matrix	62
3.3.0.7	Sequence Distance Calculation	64
3.3.0.8	Sequence Distance Calculations using Categorical values or Encoding Schemes	65
3.4	Clustering our Encoding schemes	67
3.4.0.1	Apply k-means clustering on PIM Substrates	68
3.4.0.2	Apply Hierarchical clustering on PIM Substrate	68
3.4.0.3	Application of Singular Value Decomposition on the Sequences	69
3.4.0.4	Applications of n-grams for sequence clustering	70
3.4.0.5	Applications of LSTM for sequence clustering	70
3.4.1	Sum Squared Error	71
3.5	Data Extraction - PhoScExtractor	72
3.5.0.1	SIDER	72
3.5.0.2	DisGeNet	73
3.5.0.3	CTD Database	73
3.5.0.4	dSNP	73
3.5.0.5	PheGenl	74
3.5.0.6	UNIPROT	74
3.5.0.7	DrugBank	74
3.5.0.8	PharmGKB	75
3.5.0.9	BIOGRID	75
3.5.0.10	canSAR	75
3.5.0.11	PGXnet	76
3.5.0.12	DgIDB	76

3.5.0.13	LODD-Linked Open Drug Data	76
3.5.1	Data Integration Approach for Phosphorylated Elements(DIApe)	76
3.6	Phosphorylated Network Databases-Data Processing for Network Analysis and Multimode Clustering	77
3.6.0.1	Example: Creating PIM Database Network	79
3.6.0.2	PIM Database : Querying	79
3.7	Network Analytics Approach to Drug Repositioning -NetAnaPhoS	81
3.7.0.1	Topological Structures	83
3.7.0.2	Erdos Renyi Model - Random Network	83
3.7.0.3	Small World Network	83
3.7.0.4	Scale Free Network	84
3.7.0.5	Brief Overview of Network Properties	84
3.7.0.6	Degree Centrality	85
3.7.0.7	Betweenness Centrality	86
3.7.0.8	Closeness Centrality	87
3.7.0.9	Indegree and Outdegree	87
3.7.0.10	Modularity	87
3.7.0.11	PageRank	88
3.7.0.12	Hub/Authority Scores	88
3.7.0.13	Connected Components	88
3.8	Node Impact Analysis using Single and Sequential Perturbation	89
3.8.0.1	Single Perturbation Analysis	89
3.8.0.2	Sequential Perturbation Analysis	89
3.8.0.3	Summary of this Approach	90
3.9	Multimodal Clustering Approach (DReiM)	91
3.9.0.1	N-cliques, N-clans and N-plexes	91
3.9.0.2	N-cliques	91
3.9.0.3	N-clans	91
3.9.0.4	N-plexes	92
3.9.0.5	Extraction of Datasets for the purpose of Multimodal Clustering	92
3.9.0.6	Bi-mode and Multimode Network	92
3.9.0.7	Creating the bi-mode	95
3.9.0.8	Creating the 3-mode	96
3.9.0.9	Creating the 4-mode	97
3.9.0.10	Creating the 5-mode	98
3.9.0.11	Our approach using eigenvalue decomposition((DReiM + eigenvalue decomposition)	98
3.9.0.12	Our approach using Spectral Clustering Approach(DReiM + Spectral Clustering)	99
3.9.0.13	Time Complexity of the system	100
3.10	Validation of our experiment	102
3.10.0.1	Pubmed	102
3.10.0.2	Clinical Trials	102
3.10.0.3	Precision	103

3.10.0.4	Recall	103
3.10.0.5	Accuracy	104
3.11	Future Work: Gene expression data analysis	104
3.11.0.1	Gene expression level analysis on k562 cells(GEiM)	104
3.11.0.2	Introduction	104
3.11.0.3	Extraction of GSE12056	106
3.11.0.4	GSEA Analysis	106
3.11.0.5	Pathway Analysis	108
3.11.0.6	Identification of Drug Candidate	108
4	Results	109
4.1	Experimental Data - Sequences	110
4.1.1	Experimental Data for PIM Substrates	110
4.1.2	Experimental Data for Rhabdomyosarcoma	111
4.2	Sequence Processing	112
4.2.1	Sequence Extraction	112
4.3	Identification of the Consensus Sequence of PIM Substrates	113
4.3.1	Consensus Sequence Analysis of PIM Substrates	113
4.4	Sequence Encoding to identify Relevant Proteins	116
4.5	Hamming Distance Calculations without applying encoding schemes using standard alphabets	117
4.5.1	Using the Hamming Distance to cluster sequences of length 10	118
4.5.1.1	Using the Hamming Distance to cluster sequences of length 15	119
4.5.2	Using the Hamming Distance to cluster sequences of length 20	120
4.6	Using Singular Value Decomposition with Phosphorylated Alphabet Encoding scheme - Example for clustering	122
4.6.1	Finding the Ideal-k using encoding schemes with ten sequences	122
4.6.2	Clustering using k-means	123
4.6.2.1	Sum Squared Error	125
4.6.3	Result examples using NetAnaPhoS on PIM Substrates	128
4.6.3.1	Network Analytics on PIM Substrate - Using a small example(NetAnaPhoS)	128
4.6.3.2	Top 15 Drug Repositioning Candidate	131
4.6.3.3	Identification of High Influence using PIM Substrates - Example 2	132
4.6.4	Network Characteristics using Top PIM Substrate	133
4.6.5	Our Results: Sequence Clustering	139
4.6.5.1	PhoS and PhoS-con clustering	140
4.6.5.2	Identification of PhoS: using fake sequence + real rhabdomyosarcoma sequences	142
4.6.5.3	Standard Encoding SVD analysis	143
4.6.5.4	Phosphorylated Categorization	145
4.6.5.5	Scaled Phosphorylation Encoding Scheme	148
4.6.6	Result using our Method for Sequence clustering (PhoS)	148

4.6.7	Sequence analysis - Method Comparisons	150
4.6.8	Comparison of PhoSc and PhoSc-con on PIM Substrate	150
4.7	NetAnaPhoS- pim substrates combined analysis of curated databases	151
4.7.0.1	Combined analysis of using Degree	154
4.7.0.2	Combined analysis of using Authority Scores	154
4.7.0.3	Combined analysis of using Betweenness Scores	156
4.7.0.4	Combined analysis of using PageRank Scores	156
4.7.0.5	Some results using Authority Scores	157
4.7.0.6	Singular Value Decomposition for Drug Reposition- ing: Pathway to Drugs using the combined PIM sub- strates(phosphorylated protein)	158
4.7.0.7	Node Impact using Perturbation Analysis	159
4.8	NetAnaPhoS: Rhabdomyosarcoma for network analysis using gene- drug-pathway	160
4.8.0.1	Protein-Protein Interaction Analysis for Rhabdomyosar- coma	161
4.8.0.2	Rhabdomyosarcoma Phosphorylated Protein Extrac- tion.	165
4.8.0.3	Rhabdomyosarcoma Network generation.	166
4.8.0.4	Identification of high influence nodes in protein phos- phorylated in Rhabdomyosarcoma.	166
4.8.0.5	Binning the nodes by the types:	171
4.9	NetAnaPhoS-Network Analytics for Diabetes Mellitus Phosphory- lated Protein Network	172
4.9.0.1	Perturbation Results for Diabetes Mellitus Phospho- rylated Network	175
4.9.0.2	Results from Network Analytics for Diabetes Melli- tus Phosphorylated Protein Network	178
4.10	DREiM: Multimode Clustering on Rhabdomyosarcoma	179
4.10.0.1	Ideal-k for rhabdomyosarcoma for multimode clus- tering	181
4.10.0.2	Clustering results for rhabdomyosarcoma - DREiM .	181
4.10.0.3	Drug Candidates for Multimode Clustering(DREiM)	182
4.10.1	DREiM: Multimode Clustering on Diabetes Mellitus Phospho- rylated Protein Network	182
4.11	Future Work: Gene expression using K562 cells	183
4.11.1	GSEA Analysis	183
4.11.2	Using SVD for Pathway Analysis	185
4.12	Validation of our experiment-PIM Substrate	186
4.12.0.1	Precision	188
4.12.0.2	Recall	188
4.12.0.3	Accuracy	189
4.13	Validation of our experiment for Rhabdomyosarcoma	189
4.13.0.1	Degree	189
4.13.0.2	Betweenness	190

4.13.0.3	Page-rank Scores	191
4.13.0.4	Neural Network using Autoencoder	191
4.13.0.5	Tanimoto coefficient	194
4.13.0.6	Our Method(DReiM)	195
5.14	Conclusion and Future Work	197
	Bibliography	199

List of Tables

3.1	PIM Substrates Sample	55
3.2	Processed PIM Substrates	55
3.3	Phosphorylated Alphabets with Encoding	60
3.4	Structural Alphabets	60
3.5	Charged Alphabets	60
3.6	Functional Alphabets	60
3.7	Volume Alphabets	61
3.8	Phosphorylated Alphabets	61
3.9	Demonstrate Consensus Sequence	61
3.10	Sequence Analysis for PSSM	62
3.11	Position Scoring Matrix - Frequency Table	62
3.12	Term-Document Matrix	63
3.13	Sequence Encoding for two sequences	63
3.14	Adjacency Matrix	64
3.15	Sequences for Hamming Distance Calculations	65
3.16	Sequences for Hamming Distance Categorical Tables	66
4.1	PIM Substrate Sequence with length 20	112
4.2	Sequence Length vs Number of Sequences	113
4.3	The consensus sequence: 10,20,30,40,50	114
4.4	Inter-clusters similarity using sequence of length 10	119
4.5	Inter-clusters similarity using sequence of length 15	120
4.6	Inter-clusters similarity using sequence of length 20	121
4.7	Top 10 PIM Substrates	134
4.8	Statistical Results from Using the One Centroid	135
4.9	Network Properties of Phosphorylated Protein in Rhabdomyosarcoma	166
4.10	Top 5 influence nodes for node types(protein, drugbank, pathway & SNP)	171
4.11	Validation of our experiment	188

List of Figures

1.1	Overview of the Architecture.	16
2.1	Traditional Drug Repositioning.	24
2.2	Computational Drug Repositioning Strategies.	25
2.3	Techniques for finding similarities in sequences using phosphorylated proteins	36
2.4	Dissimilar chemical structure.	43
2.5	Similar chemical structure	43
3.1	Overview of the Architecture.	51
3.2	Overview of PhoSc process	53
3.3	Sequence Extraction	55
3.4	Our Data Integration Approach to creating Phosphorylated Network	77
3.5	Data Extraction for PIM Kinase	78
3.6	Tri-mode for Multimode Clustering	79
3.7	PIM substrates schema	80
3.8	PIM Network SQL Database	81
3.9	Query Pain using PIM Network Database	82
3.10	Specific Gene Relationships in the PIM Network Database	82
3.11	Network Analytics toward Drug Repositioning(NetAnaPhoS)	90
3.12	An example of bi-mode using pathway and drug	93
3.13	An example of tri-mode using gene, pathway and drug	94
3.14	An example of quad-mode using snp, gene, pathway and drug	94
3.15	An example of 5 and 6 modes	94
3.16	Bi-mode pathway and drug example	95
3.17	Three modes using letters v,u,t	96
3.18	Multimode Clustering Architecture	100
3.19	Time Complexity for 4 Modes	101
3.20	Time Complexity for 4 Modes	101
3.21	Gene expression Framework for PIM Substrate Analysis	106
4.1	Important aspects of our work	109
4.2	PIM Substrate Sequence Analysis.	110
4.3	5 Consensus Sequence for the PIM Substrate.	114
4.4	10 Consensus Sequence for the PIM Substrate.	115
4.5	15 Consensus Sequence for the PIM Substrate.	116

4.6	25 Consensus Sequence for the PIM Substrate.	117
4.7	Histogram for 10 consensus sequence up and downstream.	118
4.8	Histogram for 20 consensus sequence up and downstream.	119
4.9	Histogram for 30 consensus sequence up and downstream.	120
4.10	Histogram for 40 consensus sequence up and downstream.	121
4.11	Ideal-k for Standard Alphabets	123
4.12	Ideal-k for Structural Alphabets	124
4.13	Ideal-k for Functional Alphabets	125
4.14	Ideal-k for Charged Alphabets	126
4.15	PIM Substrate Scatterplot	127
4.16	2-Clusters: PIM Substrate Scatterplot	128
4.17	3-Clusters: PIM Substrate Scatterplot	129
4.18	4-Clusters: PIM Substrate Scatterplot	130
4.19	2-Clusters:15 sequence	131
4.20	3-Clusters:15 sequence PIM Substrate	132
4.21	4-Clusters:15 sequence PIM Substrate	133
4.22	2-Clusters:20 sequence PIM Substrate	134
4.23	3-Clusters:20 sequence PIM Substrate	135
4.24	4-Clusters:20 sequence PIM Substrate	136
4.25	Degree: PIM Substrate to Degree	136
4.26	PageRank: PIM Substrate to Pathway	137
4.27	Authority: PIM Substrate to Pathway	137
4.28	Undirected Degree: PIM Substrate to Pathway	138
4.29	Undirected Betweenness: PIM Substrate to Pathway	138
4.30	Undirected Hub: PIM Substrate to Pathway	139
4.31	Top 20 Gene-Pathway-Drug Betweenness	140
4.32	Top Drugs: Using Betweenness	141
4.33	Ideal-k for PhoSc on 10 sequence(Cluster = 3)	142
4.34	Positional Linkage using Hierarchical clustering using 10 sequence(Cluster = 3)	143
4.35	SSE scores using cosine similarity	144
4.36	SSE scores using Euclidean Distance	144
4.37	Ideal-k Standard encoding using cosine similarity	145
4.38	Clustering using Cosine similarity Fake vs Real	146
4.39	Clustering using Euclidean distances to cluster the datasets - Standard Encoding	147
4.40	SSE scores using Phosphorylated Categorization	147
4.41	Phosphorylated Alphabets with fake vs real sequence	148
4.42	Structural Alphabet result using real vs fake sequence	149
4.43	Hydroscaled Phosphorylated Alphabets for fake vs real	150
4.44	PhoSc for Clustering Fake vs Real	151
4.45	PhoSc-con for Clustering Fake vs Real	152
4.46	Precision and Recall using Rhabdomyosarcoma-fake vs real data	152
4.47	Method Comparisons	153
4.48	Method Comparisons: PhoSc and SVD	153

4.49	PIM Related Element Analysis	154
4.50	Degree: PIM Related Element Analysis	155
4.51	Authority: PIM Related Element Analysis	155
4.52	Betweenness: PIM Related Element Analysis	156
4.53	PageRank: PIM Related Element Analysis	157
4.54	Chemical Structure of Tigapode	158
4.55	Degree for Rhabdomyosarcoma	161
4.56	Pagerank for Rhabdomyosarcoma	162
4.57	Betweenness Centrality for Rhabdomyosarcoma	163
4.58	Hub for Rhabdomyosarcoma	163
4.59	protein-protein interaction for rhabdomyosarcoma: Source clustering	164
4.60	protein-protein interaction for rhabdomyosarcoma: Target clustering	164
4.61	Degree of multimodal phosphorylated protein in rhabdomyosarcoma.	168
4.62	Betweenness centrality of multimodal phosphorylated protein in rhabdomyosarcoma.	168
4.63	Authority of multimodal phosphorylated protein in rhabdomyosarcoma.	169
4.64	The score above show Pagerank vs Authority scores indicating receptor to effector score.	170
4.65	The score above show Pagerank vs Hub scores indicating receptor to effector score.	170
4.66	Protein binned influence nodes.	172
4.67	Drug binned influence nodes.	172
4.68	Pathway binned influence nodes.	173
4.69	SNP binned influence nodes.	173
4.70	In-degree for diabetes phosphorylated network.	174
4.71	Out-degree for diabetes phosphorylated network.	174
4.72	PageRank for diabetes phosphorylated network.	175
4.73	Single Perturbation using Betweenness Centrality Measures on Diabetes Mellitus Phosphorylated Network	176
4.74	Sequential Perturbation using Betweenness Centrality Measures on Diabetes Mellitus Phosphorylated Network	177
4.75	Single Perturbation using PageRank on Diabetes Mellitus Phosphorylated Network	177
4.76	Sequential Perturbation using Betweenness on Diabetes Mellitus Network	178
4.77	Single Perturbation using Betweenness on Diabetes Mellitus Network .	178
4.78	Ideal-k rhabdomyosarcoma for multimode clustering.	181
4.79	Clustering results rhabdomyosarcoma for multimode clustering.	182
4.80	Pathway Analysis using Gene expression	184
4.81	Ideal-k using GSEA to discover clustered pathways	185
4.82	Ideal-k using GSEA to discover clustered drugs	186
4.83	silhouette using GSEA to discover clustered drugs	186
4.84	Drug Clusters using Gene expression	187
4.85	Pathway Clusters using Gene expression	187
4.86	Accuracy for Rhabdomyosarcoma for Degree	190
4.87	Accuracy for Rhabdomyosarcoma for Betweenness	190

4.88	Accuracy for Rhabdomyosarcoma for Pagerank	191
4.89	Accuracy for Rhabdomyosarcoma for NN	194
4.90	Accuracy for Rhabdomyosarcoma for Tanimoto Coefficient	196
4.91	Accuracy for Rhabdomyosarcoma for Multimode Network	196

Chapter 1: Introduction

1.1 Challenges in Drug Repositioning

In recent times; there have been many challenges facing pharmaceutical companies and industries as well as the general health sector concerning the process of designing and manufacturing new drugs. The problem is principally center around the high cost of designing, vetting and clinical trial of new drugs and the long duration of these processes. The clinical trials are performed to collect data that support the safe use of a particular medication. The information tells us about the safety and efficacy of the new drug and device development. The drugs or devices go through stages of approval in the clinical trials process before a drug or device is sold in the consumer market. The clinical trial of any drug goes through five phases (Phases 0-4) of investigations that often takes months and even years to ensure safety and efficacy in target end-user-patients.

These processes cost a lot of money, and big pharmaceutical industries spend a lot on legal fees and tests before the drugs are made available to the public. The process of clinical trials also involves human personnel who must be paid for their services [4]. For these reasons, researchers have concerned themselves with finding cheaper ways of producing new drugs or repositioning old drugs for the treatment of

new diseases or finding new indications of using old drugs via clinical observations.

The process of identifying such drugs and their new uses would expectedly be serendipitous, time-consuming or may require expensive clinical trials, but with the introduction of computational approach, the time involved could be reduced drastically [3]. The development of computer technology for this purpose has resulted in computational drug repositioning which has proved to be an effective means of finding new uses for existing drugs [1] or using known drugs and compounds for new indications [5]. Many drugs fail especially the phase III trial during development [3] sometimes due to the difficulty in predicting the clinical efficacy at its targets [3] and such failures automatically result into great losses for the manufacturing company and high costs of pharmaceutical research and development. This position further prompts a drive toward finding new uses for existing old drugs. Furthermore, drug repositioning has the advantages over traditional drug development by bypassing [5] several expensive toxicities and other tests (which were already done for the earlier indication) since we know about its safety. Hence, there is a lower failure rate during development. It is less expensive and has a shorter time to bring the drug to the market. [5]

The increase in genomic and phenotypic data is beginning to encourage scientist to develop more ideas in the field of drug repositioning because, with more data, we have more information relevant to a disease, gene and so on. Historically, drug repositioning has come from serendipitous discoveries in late-stage clinical trials or post-approval. [5] During regular use for appropriate indications. A classic example of a repositioned drug is Thalidomide which was first marketed as a sedative

and used for treating nausea in pregnancy. After few years of its widespread use in Europe, Australia, and Japan, about 10,000 children were born with phocomelia (a congenital malformation of the limbs characterized by absent portions or whole arms) and other congenital disabilities.

This led to its ban in most countries in 1961. However, much later the US Food and Drug Administration approved its use for the treatment of leprosy and, subsequently, for multiple myeloma. [5] Another well-known example is Sildenafil which was designed by the drug company Pfizer for something that would relax the heart blood vessels and treat Angina pectoris. However, its trials in people were disappointing; [5] as the volunteers reported lots of excessive penile erections as an unusual side effect. [6] The drug was then repurposed to treat erectile dysfunction (ED) and had since helped millions of men with ED and later proved to be useful for pulmonary hypertension [7] Similarly, Minoxidil was developed as an antihypertensive drug and got FDA approval to treat high blood pressure. In a study of its efficacy, Chidsey et al. noticed unexpected hair growth on the faces and shoulders of some women subjects. A male subject experienced hair growth on the bald part of his head. The growth was so noticeable that his barber said: Boy, you better find out what you are taking for your high blood pressure! [7] Research was later conducted on the best ways to use Minoxidil for hair baldness, and in 1988 the FDA finally approved it as an anti-baldness medication [7].

Drug repositioning or repurposing has opened a new source of revenue and now holds much appeal with a high potential to accelerate the development of old drugs for new disease(s). More and more companies are scanning the existing

pharmacopeia for repositioning candidates, and the number of repositioning success stories is increasing [8].

Other examples of such drugs that have been repositioned in the last few decades are listed in [5].

For this study, we introduced a computational approach that utilizes sequence encoding and sequence network analysis for drug repositioning. The analysis was performed on a database that contains interactions between factors that relate to a specific disease such as phenotype, biological process/course, as well as chemical compounds or proteins involved. This approach is not limited to using some network analysis techniques but also consists of the use of sequence encoding, sequence analysis, network analysis, and perturbation analytics to assist in repositioning drugs. Expectedly, the methods will aid in increasing the lifeline of old medications, reducing the number of funds spent on approval, re-directing the use of drugs not for one purpose but multiple purposes [4] thus engendering the use of old drugs for treating new diseases.

1.2 Challenges in Cancer Research

With all the research, development and innovative drugs and therapeutic approaches that have helped to fight cancers, now people are surviving from this disease. Cancer therapeutics has come as far as the usage of different drug combinations, radiation therapy, surgery for localized cancer, etc. There is still a lot to do with regards to the treatment of cancer. Studies have shown that clinical trials

in cancer research have been known to have the lowest success rate [9]. One would wonder why this is so.

In the past, research has focused specifically on critical factors that lead to tumor growth [9]. In concentrating on tumor progression, researchers developed therapeutic approaches meant to target these factors. They observed that targeting these factors that leads to tumor progression can introduce an inappropriate amount of toxicity in the body because these factors also play a crucial part in homeostasis in the body [9]. For this reason, the scientist had to go back to the drawing board to figure out how to deal with the complex problem of solving cancer. Many discussions among cancer researchers from different parts of the world from both the academic and pharmaceutical community came to address the current challenges that are hindering the progression of drug developments and various therapeutic approaches that can help stop all forms of cancer.

Here we will address some of the challenges that are currently hindering the development of efficient and effective treatment for different types of cancer. One of the challenges, cancer researchers or scientist, need to enhance their understanding of what a clinical trial is. Clinical trials are a process that a drug, treatment or therapeutic approach has to go through for validation that the drug, treatment or therapeutic approach is viable for a specific disease. Before clinical trials, other extensive research work has been performed, through acquiring a holistic knowledge on the disease in question, addressing the cause of the illness and which approaches can address the condition. Experimentation was done using different models both computational or biological also have been performed in the pre-clinical trial stages.

With all these pre-clinical trial extensive knowledge and testing done before clinical trials, one would expect that researchers have done all that is expected of them to do prior thus having a vast knowledge of both pre-clinical and clinical trials. This is because what was done in the models used, and the result that was attained is expected to be the same result that is anticipated during clinical trials; however, clinical trials do not always correlate with experimentation done on models such as mice. Thus it is understandable for researchers to agree that more knowledge is needed for clinical trials.

In reforming clinical trials through attaining a better understanding of what clinical trials are, there has to be uniformity between researchers or scientist and clinicians to understand the many problems that come with treating individuals. The financial aspect of cancer treatment for patients, the challenges in recruiting patients with cancer to be part of a clinical trial and lastly possible toxicity that comes with a dosage of cancer medications are some of many challenges [9]. It was suggested in this article that the connection between clinicians and researchers has to develop as early as having a graduate school where graduate students are placed in a clinic environment while clinicians are also placed in labs [9]. These different group understanding each others perspective can promote an interaction that will help in understanding what the different communities are experiencing; therefore, working well together to develop better clinical trials for patients with diseases.

When a new drug is developed and sent to the clinic, it is vital for clinicians to comprehend the pharmaceutical concepts such as the pharmacodynamics and pharmacokinetics for the drug [9], the effects on tumor shrinkage and side effects.

Clinicians should attain a better understanding of a balance between drug toxicity to the patient and the effectiveness of the diseases. They should understand the highest amount of dose that can be given to the patient without the dosage causing terrible side effects. Thus researchers should work on identifying toxic side effects during the early stages of drug production and also how it influences the entire body and not just the tumor [9]. Researcher assisting clinicians with a detailed understanding of dosage needed for specific cancers. Clinicians also help play their part in the research aspect by giving a detailed report concerning direct dosage effects on cancer patients with certain cancers. Thus researchers should work on identifying toxic side effects during early stages of drug production and also how the influences the entire body and not just the tumor [9].

The second challenge towards cancer research is improving early detection in cancer. The current initial detection of cancer is done using CT, PET and MRI scans. These scans have been used to distinguish between an aggressive tumor and a tumor that is not progressive in growth or developing slowly [10]. Also, some of these images at times can give an increase in false-positive rate. If these images are improved, it helps doctors to understand the tumor and see if it is localized thus assisting the surgeons in removing cancer and any traces of it better. Though imaging and the improvement of imaging will help with initial detection and the eradication of disease, they also exposed individuals to radiation, which at times can also lead to tumor progression. Early detection is difficult because of another issue which finances; insurance companies are assisting in payment for individuals attain screening for certain types of cancer [9]. It has also been observed that for some

cancers, not only is early detection due to financial challenges or difficulties but also researchers have not developed a way to screen for some cancer early enough [9]. For this reason, specific cancer would have elevated to an aggressive level before there is detection making it difficult for the individual's survival.

Adequate access to resources for researchers is also another challenge that cancer research is currently facing [9]. For cancer research to progress, there has to be an increase in collaborative efforts to create better therapy, predictive and diagnostic tools. Public data that is viable and has been appropriately performed and extracted without any errors have to be available for other researcher and clinicians to learn from to help in developing better treatment approaches [9]. Before a new study initiation for cancer, researchers should acquire all the data collected both on animals and or cellular models and also acquire data from epidemiological data from individuals with that have that specific cancer to have comprehensive knowledge regarding the tumor complexity. The reason for this is because tumor development consists of more than a mutation in the DNA. It also includes the biochemical pathway that is connected. For instance, understanding oncogenes such as Ras and AKT [11] has exposed an area in cancer biology that was unknown. Several therapeutic approaches should be tested on several models via mice models, cellular models and also informatics models. The reason for this is the genetic background may have played a role in tumor growth. Curative and side effects expected in therapeutic approaches or drug development for a cancer patient should be observed in models such as mice before initiating clinical trials. For example, if the side effect is death in patients, then this exact side effect should also be observed in the mice models before the

initiation of clinical trials [9]. Additionally, appropriate data interpretation has to be acquired from models systems before next steps in drug development.

Another suggestion stated in this article was that more in-depth knowledge and understanding of the environment where tumor growth is located has to be acquired [9]. For instances, researchers need to understand what influence does the surrounding tissues or cells have on tumor growth. There has to be a better establishment or development of human-like models that allow the researcher to observe cancer pathways in a realistic biochemical or biological context or environment. This more in-depth knowledge can help in understanding the molecular factors or features that cause the origin of cancer or the type of cancer. For instance in pancreatic cancer there is an 83 percent K-Ras mutation; however, in colon cancer and breast cancer, Ras mutations do not lead to cancer development [12]. Thus, the environment of a tumor/cancer effects can help shed light on the development of the different type of cancer and its behavior. This insight into the environmental impact on cancer will help cancer researchers and clinicians. However, researchers know little concerning the microenvironment influence on tumor development. Researchers also lack vast knowledge of the reason why different types of tumors derived by similar genetic factors respond to therapeutics differently.

Lastly, the use of system biology and the study of the genome has been a great asset to understanding cancer [9]. The study of the genome has lead researchers to ask if cancer is truly a disease of the genes or a disease caused by chemical pathways, e.g., protein interactions. The application of computational biology, health informatics, etc. has also helped to highlight factors that play roles in cancer de-

velopment which has helped scientist to match certain drugs to genomic data [13]. Though system biology has helped in improving cancer research and drug development, much improvement is still needed as some computational research still gives some inaccurate results.

In summary, though cancer research has come a long way, there is still a long way to go as there is still much knowledge that needs to be acquired concerning this disease if it is indeed a disease of the genome or signaling pathway, its characteristics and behavior. There also needs to be better access to viable literature, data, and data interpretation, and improved models for early detection for different types of cancer. Extensive work in both pre-clinical and clinical trials to truly understand and develop drug and other therapeutic approaches that can help destroy tumor or cancer cells while removing exposure to high toxicity to patients. Moreover, clinicians, researchers, and pharmaceutical companies have to develop secure communication and interact habit with one another coming together to understand each side to bring about the best treatments for cancer patients.

1.3 PIM Kinase in Cancer Research

Previously Ras and ATK were mentioned to be oncogenes that have helped attain more understanding in areas of cancer biology that was not previously known. In this section, another protein also currently being studied which is also exposing another side of cancer biology and new drug development towards tumorigenesis is PIM Kinase [14]. PIM kinase belongs to the serine/threonine kinase family; they

have been conserved in the cellular organism through evolution [14]. Within this family consist of three members: PIM 1, PIM 2 and PIM 3. These kinases when compared to each other show high similarities; 71 percent similarities on the amino acid level between PIM 1 and PIM 3 and 61 percent between PIM 1 and PIM 2 [15]. High amounts of PIM 1 has been observed in hematopoietic cells, whereas PIM 2 is observed in lymphoid and brain cells. PIM 3, on the other hand, is found in cells localized in the kidney, breast and brain cells [16,17]. PIM kinase only requires protein stability to be able to function; thus they do not require post-translational modification. PIM consists of over 30 possible recognition sequences for various kinases. Currently, researchers do not know the relevance of these recognition sites [18]. PIM also responds to mitogenic stimuli which causes an upregulation in transcription [19–21].

Furthermore, PIM Kinase activation is also caused by transcription factors that activate downstream growth factor signaling pathway, an example being NF-kB [14,22]. Besides, hypoxia found in the solid tumor has been known to cause the expression of PIM. The upregulation of both PIM 1 and PIM 2 has been observed due to NF-kB response towards FLT3/ITB oncogenic mutants. MLL-X, NuPP-x and MLL-PTD are mutations observed in hematological malignances [22]. These mutations seem to cause upregulation of PIM 1 via a transcription factor known as Hox A9 [17].

Research has identified PIM Kinases in different types of cancer. It was seen to change Mesenchymal cells that ended in lymphoma weakly and leukemia [23,24]. The increase of PIM 1 expression alone could not cause adenocarcinoma develop-

ment in the prostate; however; it was a contributing factor in the severity of prostatic neoplasia [20]. This discovery supports data found where there was an overexpression of PIM 1 in the prostate cell line; this finding showed that PIM 1 was not able to alter benign cells into malignant cell however it caused an increase in tumorigenic abilities of tumor cells [14, 22]. This was seen in both in vitro and in vivo models. Also, increased levels of PIM 1 Kinase was initially observed in lymphoma tumor, human myeloid and lymphoid leukemia [19, 25, 26]. Moreover, a transcription analysis performed using 50 percent samples from prostate cancer patients showed high expression of PIM in malignant tumors, whereas the same analysis performed again this time with benign lesion showed low to no expression of PIM 1 [22]. In the prostates, intraepithelial neoplasia high levels of PIM was observed as well, thus showing that PIM Kinase possibly plays a role in early prostate malignancy development [27].

PIM kinase was noticed to express in various cancers; for this reason, there is an interest to target this protein for new drug development. The interest in targeting it for new drug development [14, 22] is also because of its involvement in cancer-specific pathways, e.g., cell survival, cell migration and lastly cell cycle progression. Additionally, it was shown that the reduction in tumorigenesis in pancreatic cancer cells in mouse models was due to the effects of dominant-negative PIM 1 [28]. Though drug discovery studies are ongoing with PIM Kinase, researchers are focusing on PIM 1 because of its association in tumorigenesis [22]. It was discovered in vivo studies has also demonstrated that the absence of PIM 2 and PIM 3 responds similarly as if all three PIM were absent; meaning that the lack of the two kinases

lowers sarcoma growth that is induced by a treatment called 3-methylcholanthrene carcinogenic as if all three Pim kinases were absent [29].

The identification of the vital role that PIM kinase plays in tumorigenesis in humans has led to more interest in creating small molecule inhibitors that aim or focuses on these kinases [14, 22]. Scientists have discovered many different types of PIM inhibitors [30] however only a few were tested in animal models or cell-based assays to show anti-cancer activity. Among these inhibitors, a few have worked against the kinase family, and this is because most were targeted on PIM 1 [22, 31–33]. Models, where mice did not possess all three PIM, showed that mice did not display side effects with the PIM inhibitors, the same experience might be expected for cancer patients. Additionally, over the past years, many publications have been geared towards new PIM inhibitors. These publications address various chemical structures in PIM inhibitors that are genuinely potent and with promising selectivity profiles compared to other protein kinase [14, 22]. Aside from publications, many companies have developed various small molecules that focus on PIM kinase family [14]. They are focused on developing inhibitors that are structurally different and potent, and also combine it with other therapies. For instance, SMI 4a is a PIM 1 inhibitor developed in the University of South Carolina. It has been observed that SMI 4a caused G1 arrest in the prostate (PC3, Du145, CWR22ru 1) [14]. DHCP-9 is another inhibitor developed by CNRS to inhibit all PIM Kinase family. DHCP-9 has been shown to weaken migration and conquest in the PC-3 prostate cancer cell line [14].

FDA, on the other hand, has suggested to companies a few years ago to start

their Phase 1 clinical trial to test the pharmacokinetics, tolerance of PIM inhibitor SC1-1776 and safety in individuals with solid tumors and also those with refractory non-Hodgkin lymphoma and prostate cancer [14]. The Phase 1 trial was canceled after two years due to protracted cardiac toxicity occurrence during the trial. Furthermore, some of these kinase inhibitors have been shown in other in vitro studies to have acceptable activity and a lessened toxicity profile. There is still an ongoing Phase 1 trial with efforts to attain a well-defined toxicity profile of these drugs in individuals and also find targets for tumor [14]. Currently, data recommend that PIM inhibitor is more effective when combined with treatments such as chemotherapy or other types of targeting agents such as P13k inhibitor [14]. Though efforts are made to develop inhibitors towards PIM kinase family, there is still a large amount of research that needs to be performed to test the combination of PIM inhibitors and other therapies [14].

1.4 Motivation

Existing drugs accounted for 20 percent of drug products released into the market in 2013 [1], the FDA (Food and Drug Administration) has been creating public datasets regarding drug repositioning that allows pharmaceutical industries to make more informed decisions on drugs. [1]. Also, drug repositioning is playing a vital role in precision medicine creating more opportunities for effective treatment of patients [1]. The traditional approaches to drug repositioning have mostly focused on chemical structures and side effects, while other methods have focused on screen-

ing approaches or exploring direct relationships between a drug and disease. The need for integrating information from various sources has become very important. This information includes phenotype, genome, chemical and clinical data, and their integration assists in discovering new uses for old drugs. The datasets are difficult to make sense out and very complex to understand the need to apply a multi-layered approach to understand and analyze this problem becomes very important.

1.5 Summary of our Contributions

The contribution of this work is to create a novel approach or framework for finding drug repositioning candidates. We developed a method for the extraction of relevant information using sequence level data and phosphorylated proteins. We postulated a method for encoding the protein sequence in such a way that we can cluster the protein sequences to find the most relevant protein sequences using PhoSc and PhoSc-con. The relevant sequences were used to capture relevant biological association for various databases using PhoSc Extractor. We utilized the extractor in creating a combination of disease-related elements for network analytics (NetAnaPhoS) and multimodal clustering(DReiM).

We present an overview of the system architecture in figure [3.1](#).

1.6 Organization of the Dissertation

The remaining parts of the dissertation contain Chapter 2 that presents the literature review with research that relates to this work, in Chapter 3 we have the

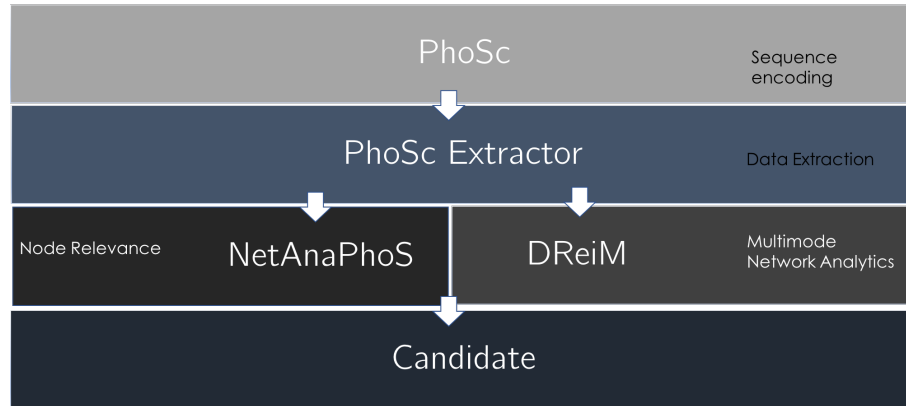


Figure 1.1: Overview of the Architecture.

methodology section that shows the overall architecture of this work and Chapter 4 that contains some results and conclusions and future work in Chapter 5.

Chapter 2: Literature Review

2.1 Drug Repositioning

This chapter discusses two approaches to drug repositioning [1] [2]. This chapter breaks the approaches into two. Both approaches are explained in this chapter, and the most recent literature in these areas discussed to build a foundation for my thesis.

1. The traditional drug repositioning approach
2. The computational drug repositioning approach

Drug Repositioning is the process of developing a new use for old drugs. The concept has been around for many decades [3, 34, 35]. Another way to explain this concept is to define it as the use of newly identified drugs for the treatment of diseases other than those they were intended to treat [35]. The most famous example of drug repositioning is the use of sildenafil for the treatment of erectile dysfunction [2, 36]. Sildenafil was found to treat coronary artery disease by Pfizer in the 1980s [3, 36]. Another good example is thalidomide; the drug was initially developed as a sedative drug especially for treating nausea in pregnant women, and the prescription was eventually found to be effective in treating erythema nodosum leprosum (leprosy)

and much later for multiple myeloma [1,2].

2.1.1 The traditional drug repositioning approach

The traditional drug repositioning approach to drug repositioning is usually referred to as a discovery process [2]. It is generally referred to as a discovery process due to trials and errors involved in the process. The traditional approach involves de novo identification workflow and validation of new molecular entities (NME). It is a time consuming and a very costly process. [4,35] It takes six to nine years to develop a drug if the drug is repositioned it can go directly to the preclinical testing and clinical trial figure 2.1. It will help to reduce the risk and cost further.

The traditional approach is also called an experimental approach to drug repositioning since it involves experimenting and extensive laboratory work [35].

The foundational idea behind traditional drug repositioning has not changed much since prehistoric times; the process is referred to as a discovery process since it consists of a lot of trials and errors [4,35]. The idea is to identify the disease of interest; the disease is selected based on the population of people with this disease and based on clinical needs. The next step will be to gather an array of chemicals to run the test and see the effects and find any relevant chemical of interest. The process is costly, inefficient and very costly. The method is called the screening method. The screening method consists of two methods:

2.1.1.1 Target-based screening

Target-based screens are described as a method used to identify molecules or peptides that affect the expression of a cell [1, 2, 35]. The drugs are screened for a protein or any biomarker that the researcher is interested in or by the evaluation of drug compounds such as the ligands from the collection. After these compounds are identified, an effort is made to identify the biological targets [1, 35]. If a target is found, a biochemical assay screen for the target is conducted, and the result is evaluated on models via a generally expensive process [1–3].

2.1.1.2 Phenotypic drug screening

Phenotypic drug screening is an approach that has been identified in drug discovery research for recognizing molecules and peptides that can distort or alter the phenotype of an organism in the desired way [4, 35]. This approach has a strong history in the drug discovery paradigm. Scientists have screened a bunch of compounds, particularly in a diseased animal to identify compounds that cause an expected change in phenotype. When the mixture is found, the next goal will be to find the biological target for such compounds.

A biological target is defined as a substance within the human organism to which any drug can bind. The reverse could be the case where a biological target can be said to alter disease, thus leading scientist to discover compounds that modulate the activity of this target. The compounds can then be tested on animals to see the outcome [8, 35, 37]. The approach is termed reverse pharmacology. These screening

can be performed in two ways: in-vitro or in-vivo In vitro is the easier of the two screening types and it unlike in vivo utilizes materials such as test tubes, petri dish, etc. to study or test biological units outside the living organism to observe changes in behavior or activities [4,35]. At times the testing or study is performed outside the living organism and placed back inside the organism to observe phenotypic changes. Expressions of several proteins are screened using this screening type.

On the other hand, in-vivo screening represents effects on biological units in living organism to see changes in behavior or activity. This method is usually chosen over in-vitro tests because the experiment is conducted on and the outcome is measured on living organisms. The phenotypic screening techniques do not make any guess on the pathological mechanism and proteins involved. The cell line or model organism of diseases would be used to read the results of the screening [1,2,5].

The target-based screening techniques improve the selection of the chemicals on the ability to bind to a protein. The protein it attaches to is the target protein useful for the pathological process. It binds like the key is to a lock. The better it can lock the more fit the action of the chemical [1,2,5,35].

2.1.1.3 Limitation of the Approaches

It is challenging to find a molecule that will interact with a single protein only. Most proteins have the propensity to bind to multiple drugs thus producing a negative effect. Identifying such proteins can provide significant opportunity drug repositioning [1,2,6,38].

Another limitation is that these drugs are also involved in several biological processes and can achieve multiple functions. Previously, the assumption was that a protein target would perform a single biological role [1,2]. The right protein found to be involved in the disease was a way of recognizing the phenotypic outcomes. The approach ignores the relevance of pathways and networks and this also further produces opportunities for drug repositioning taking these factors into consideration [8]. Considering the approach including these parameters help to create more opportunities for drug repositioning. The better we can understand the involvement of these pathways and networks in the process the more precise our drug repositioning results will be.

Another limitation of the approaches above is that compounds can only be tested based on main indications and it is impossible to check on all possible diseases meaning it possible for a compound to be used for different purposes and yet not identified. These present an opportunity in drug repositioning as well as [2].

To be able to solve this problem, we need available biomedical knowledge and a good understanding of the molecular systems.

2.1.1.4 Other challenges faced include both legal and scientific battles

A biological system is a very complex system and impulsive, and we can see these effects in cancer and other diseases. A disease rarely never keeps their first indications because advancement to the initial evidence is made every year [1,2,6,35,38]. The more information gets available to the public about the disease the

better hints are created to target the disease. It makes it challenging to find new relevant opportunities [35].

The financial challenges also play a significant role in drug repositioning [1, 2, 6, 38]. Before a drug is commercialized, it required not just to show that it is valid based on its features but also essential to make sure that the intellectual property is not expired. If they are expired, there will be no incentive for pharmaceutical companies to continue to pursue the research.

The process of repositioning may also cause a lot of issues since re-positioning is not a part of the regulatory process. It may delay the use of a new indication. During the process depending on the demographics or physiology or the patient, it is tested on an adverse reaction to the drug may occur. The dosage should also be watched for the possibility of an adverse drug reaction [1, 2, 6, 35, 38].

All these factors should be considered to reposition a new drug in the market successfully.

2.1.1.5 Successful drug repositioning: Stories

Sildenafil was initially designed to take care of a condition called angina (chest pain issues that occur due to a restriction to the blood supply to the heart). The theory was that phosphodiesterase-5(PDE5) should increase the blood flow and allow free flow of blood to the heart. The drug was discontinued during the clinical trial stages and during the same time patients started to report a rare side effect which was prolonged erections. Pfizer investigates the drug for the new indication.

Three thousand seven hundred men were studied for this new indication, and the new indication was confirmed leading to the repositioning of this new drug for erectile dysfunction [1,2,6,35,38].

The indication was identified by chance and not logically. The clinical trial was essential to determine the real behavior of the medication, highlighting the challenge from moving from cell-based assays to a real human body.

The more we know about the biological aspects of an organism would help to better predict similar cases like sildenafil.

Thalidomide was removed from the market because of its hazardous nature, but it was introduced later to the market. This story is a fascinating one as the drug was used as a sedative drug or sleep-inducing medicine. It was primarily marketed to treat morning sickness in pregnant women [2] [8]. The drug was assumed to be safe because it was tested on rodents. It was not the case for humans as it caused severe skeletal congenital disabilities in children born from women that took the drug. Over 15,000 newborn suffered this bad effect [2]. The drug was removed from circulation and led to a lot of reforms in the pharmaceutical industry. The story did not end there a practitioner was trying to treat his patient who had erythema nodosum leprosum (inflammation of condition depicted by red nodules under the skin), he decided to use thalidomide. The pain and sores disappeared the next morning [2]. During the clinical trials, thalidomide proofed to treat erythema nodosum leprosum. Thalidomide sales grossed millions of dollars derived from an off-label use for multiple myeloma [2,7]. The lesson from this is that a drug can be harmful in a specific population but important in another. If we can identify an

adverse effect and a drug process, we can target drugs accordingly.

Understanding the protein and its process helps to understand better and predict the opportunities here [2].



Figure 2.1: Traditional Drug Repositioning.

2.1.2 Computational Drug Repositioning

Computational drug repositioning is a more promising and useful tool for finding a new use for old drugs [1,2,8]. This field becomes increasingly promising as the data in this field begins to increase. The computational approach is relevant to precision medicine. Advanced analytics techniques are applied to this data sets on a day-to-day basis to find drug repositioning candidates [1,2,8].

Various data sources have been identified to help in drug repositioning. We discussed the drug repositioning data sources in details. Existing drug repositioning computational strategies will also be discussed in this Chapter [1]. Then the most significantly used computational techniques in literature would be elucidated including the approaches used for the validation of the approach [1,2,8,34].

From the traditional approach discussed, it easy to understand the de novo drug discovery approach does not work well. The method is time-consuming and very expensive to manage. The money spent on R and D has increased while the number of new drug approvals has idled [1,2].

[1] group computational strategies for drug repositioning into Phenome, Genome, and Drug chemical structures. Corset, 2014 grouped approaches into chemical structure-based approaches, gene expression and functional genomics-based approaches, protein structure and docking based approaches, phenotypic and side-effect-based approaches, genetic variation-based approaches, disease network-based approaches, and the machine learning and concept combination approaches [2]. A few of these approaches will be discussed in this work include phenotypic and side effect based approaches, disease network-based approaches and the machine learning and concept combination approaches [2].

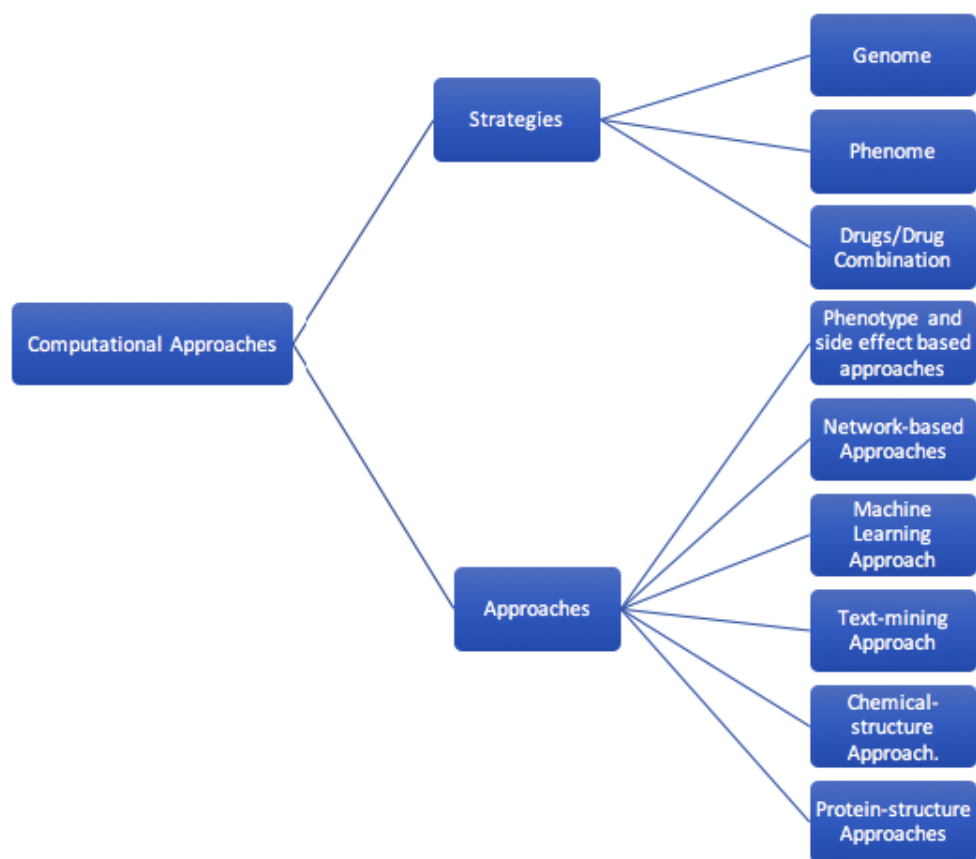


Figure 2.2: Computational Drug Repositioning Strategies.

2.1.2.1 Computational Strategies to Drug Repositioning

We have used [1] to best categorize the computational strategies for drug repositioning.

2.1.2.2 Phenome

The phenome contains a collection of phenotypic information. The data sources that contain this information have emerged to be good sources for drug repositioning. PheWAS, for example, has emerged to be an excellent way to identify associations between genetics and diseases systematically [1, 8].

Denny et al. demonstrated using an application of PheWAS on electronic medical records [39]. His work showed that PheWAS is an excellent tool for enhancing the analysis of genome and for finding associations between gene markers and human diseases [2, 34].

The side effects from clinical information help profile [40] drug-related personal phenotypic information which can be used to develop drug repositioning candidates. Ye et al. showed that similar side effects profile might share similar therapeutic properties [39].

Bisgin et al. also used the latent Dirichlet allocation model for drug repositioned that developed the phenome information from SIDER (Side Effect Resources) [39, 40]. The use of this application will require an understanding of molecular aspects and pathological mechanisms. The phenome data can also be integrated with other types of data for drug repositioning. The work of Hoedndorf et al. in-

volves the integration of genotype-disease associations with drug-gene associations to predict drug-disease associations [41]. A similarity score approach was developed to measure genotype-disease associations. The path leads to the identifying of already existing drug-disease associations and a system like this can be used for drug repositioning.

2.1.2.3 Genome

Genomic data paradigm is growing due to the creation of genomic and transcriptomic data. The available public datasets include information on cell lines, disease samples, normal tissues, and animal models [1]. The combination of the following datasets and other datasets that include phenotypic and clinical data present a good opportunity for drug repositioning. The datasets will provide an understanding of the disease mechanism and explain the mechanism of actions which can help in drug repositioning.

CMap has been widely used to strengthen this effort. The CMap data is an extension of (LINCS), with many gene expression profiles for human cancer cell lines treated with different drug compounds with different conditions [1, 8].

The purpose of CMap is to show a map that contains functional associations between diseases, gene perturbations, and drug actions. The integration of these datasets with NCBI Geodata has led to several studies on drug repositioning. An approach that utilizes this data is called signature reversion. This approach involves finding the inverse of a drug-disease relationship by finding the difference between

drug-gene expression profile and disease gene expression profile.

The approach was used by [42] to compare the gene expression signature of inflammatory bowel syndrome to drug-gene expression signatures consisting of 164 drug compounds from CMap [42]. It has been recently found that miRNA regulate some cell activities. This finding can help to develop drug targets for drug repositioning. An example of this strategy (Liu et al.) found that miRNAs and transcript factors to be enriched in cystic fibrosis associated gene regulations from a public dataset [1, 43]. The feed-forward loop was constructed for cystic fibrosis by building transcriptional factors and miRNA as regulatory elements. Forty-eight existing drugs showed the propensity to affect the expression of miRNA that is part of the feed-forward loop. They were repositioned for drug treatment of cystic fibrosis [1].

2.1.2.4 Drug chemical structures/ Drug Combinations

Drug chemical structures:

The drug chemical can be found in many publicly available datasets, and most of these datasets contain chemical structures and literature based biochemical data [1]. The data can be used for drug repositioning. The concept of using these types of data is that similar chemical structures perform the same chemical function. The structural features may be used to compute their similarities and a topological fingerprint may also be used to calculate their similarities [1, 2].

Swamidass et al. work focus on finding which targets modulate disease relevant phenotype [40, 44]. The work uses chemical structures which target modulate

disease-relevant phenotype [40, 44].

In Wang et al. , drug chemical structure, molecular activity, and side effect were integrated. The three datasets were used to define a kernel function for SVM classifier. The technique was paralleled with other methods and showed high-efficiency [1, 2, 35].

Drug Combinations:

Targeting just a single drug may not be enough to treat a specific disease. The need for a combination of drugs to address a particular disease issue becomes essential. Ceribelle et al. screened 466 drugs that have been approved for cancer therapy [35, 39]. The finding showed that ibrutinib, a kinase inhibitor that hinders B-cells receptor signaling pathway to activate IKK, shows a significant joint effect with JQ1 in killing ABCDLBCL cells both in vitro and in vivo [35, 39, 40]. These indicate that the combine JQ1 with ibrutinib may be a new effective therapy [1, 34, 35]. For information on the drug, combinations use Drug Combination Database.

2.1.3 Computational Approaches to Drug Repositioning

With the advent of computational approaches, we can renew the use of failed drugs for other purposes and shrink down the timeline for identifying drug repositioning candidates by ten times. This process involves designing computer algorithms that would simulate the traditional workflow and allow generation of the potential drug candidate [35]. Below are some of the computation approaches to drug repositioning:

2.1.3.1 Sequence based methods

Sequence analysis is used in bioinformatics to find similarities between DNA sequences, RNA sequences, and protein sequences. The similarity measures help us understand functions, features, and structures [35,45] and how similar they are to each other. One of the methods used to calculate these similarities is sequence alignment [46,47].

In literature, a majority of the sequence-based methods are not directly linked to drug repositioning task or are not typically extended. We have identified a few of those that may be used to connect the dots between sequence clustering and the drug repositioning task [48].

As we mentioned earlier, phosphorylation is very important for regulation of many cellular processes including cell cycle, growth, apoptosis, and signal transduction pathways [48].

Sequence clustering is a useful application in identifying this phosphorylation site. Successful identification of these sites can lead to the discovery of new networks that can lead to drug repositioning [48].

2.1.3.2 Previous work using Sequence-based method for identification of Phosphorylation sites

[48] created a bioinformatics tool for prediction of individual kinase-specific phosphorylation substrates and sites using heterogeneous feature selection method.

PhosphoPredict combines the use of protein sequence and functional features to find the kinase-specific substrates and their phosphorylation sites [48]. The sets of elements that were most relevant to predict the specificity of the substrates for the kinase families were identified [48]. The first step involved removing redundant sequences from the initial datasets [48]. Twelve kinase families were trained using Random Forest Model, and they were taught independently of each other. The tool is efficient for both the prediction of the substrate to kinase family relationship and also to identify the sites of phosphorylation in the kinase families. Phosphopredict involves four significant stages dataset curation, feature extraction, feature selection, and model training and performance evaluation [48]. The feature extraction stage involved sequence features predicted structural features and protein functional features [48]. Then the hypergeometric test was performed to identify those functional features that were under-represented or over-represented, and the mRMR algorithm was applied to select the essential features [48]. The performance of RF-based predictors using 5-fold cross-validation. Though this approach improved the prediction of phosphorylation sites for several kinases, it is limited because the inclusion of additional features improved the prediction accuracy for some kinase/kinase families, it decreased the performance for others [48].

[49–51] used multiple linear regressions to predict sequence values based on their solvent accessibility. Yuan et al. also has done some work in this area using support vector machine and Adamczak et al. using a neural network to predict the real-values of accessible surface area [50, 51]. The method identifies that previously there werent defined accessibility surface area states and because which leads to

random choices, poor comparison, and loss of information [50, 51]. This method applies directly gives the real values of ASA instead of its arbitrarily defined states. This method had a similar performance to when the neural networks and support vector machines were applied to it. The model predicted the accessible surface area with 16.2% means absolute error [50, 51]. The results show the importance of sequence neighbors in the determination of solvent accessibility [50, 51].

[52] is implemented based on profile hidden Markov Model. The pHMM is used to learn the kinase-substrate relationship. The decisive set and the negative set of phosphorylation sites were first to select, and the MDD(Maximal dependence decomposition) was applied to group protein phosphorylation sites into subgroups [50, 52]. Then we applied a certain minimum cluster size on the correct settings using the MDD. If the size of the subgroup is less than a certain size, the subgroup is terminated to be divided until all fall below the minimum cluster size [52]. Based on the profile is hidden Markov model, computational models are learned from the kinase-specific groups of the phosphorylation sites. After evaluating the learned models, the model with the highest accuracy was selected from each kinase-specific group, for use in a web-based prediction tool for identifying protein phosphorylation sites [52].

[52] emphasizes the need to be human-specific to classify protein phosphorylation sites appropriately. The article also points to the use of solvent accessibility as a way to reduce the false positive results for the phosphorylation sites located in the buried region [52].

[53] matched the sites by using a collection of consensus sequence motifs

and their position-specific scoring matrix retrieved from Scansite [54] and the applications of artificial neural networks technique using NetPhosK. The consensus sequence motif is used to identify a family containing closely related kinase [53, 54]. The Kinase are found and designated to families by searching BLASTP for sequences similarity of sequences that represent the set 82 kinases [53, 54]. The hits with at least 50% sequence identity are considered. The probabilistic network of the functional association was used to capture the biological context of the substrates, and they were extracted from the STRING database [55].

They utilized the following biological context on STRING database: genomic context, primary experimental evidence, manually curated pathways, and automatic literature mining [56]. The proximity of the substrates is calculated using the substrates closeness for all the kinases. The Floyd-Warshall algorithm was used to find the most likely path connecting them. This context is used to filter out false-positives. The algorithm was unable to recover the missing sequence motifs.

[57] utilize feature extractions, but the method utilizes a machine learning technique that finds patterns from the raw sequences. The technique will find complex representations of these patterns from the sequence.

The method utilizes a multi-layer convolution neural network. [57] utilizes 33 residues all centered around the potential phosphorylation site. The fragments of the protein then encoded with one at the index of each amino-acid in the protein sequence and 0 at all others. The multi-layer will encode the convolution neural network protein sequence into a fixed two-dimensional hidden state and this done using an attention-based decoder [57]. This allows the model to search for essen-

tial positions to learn soft transformation between input and output. In summary, RNN Encoder was replaced by the multi-layer CNN, and this resulted in a two-dimensional attention mechanism on the sequence dimension and the feature-map dimension that change the RNN decoder into a feedforward neural network to generate a single representation vector [57]. The attention mechanisms help estimate the contributions of each element on both sequence and feature map dimensions and also to arrive at a merged soft-weighted representation of the protein sequence [57]. The problems remain that though Musite worked in identifying phosphorylation, the problem remains with its interpretability and biologically meaningful discoveries [57]. [58] does not only take into account the sequence information but also the functional information regarding substrates which places.

The proposed method takes advantage of not only sequence information but also functional information regarding substrates that are reported to contribute to phosphorylation site prediction [58].

Kinase less than 25 favorable phosphorylation sites were removed, and only 17 kinases were left after. The local sequences of the sites were then extracted seven up and seven downstream. The binary encoding was then applied to the local sequences [58]. Then the protein-protein interaction information was used to incorporate the functional information of the substrate using STRING. The interactions were extracted giving a total of 16,708 proteins from 679 substrates. The two are used to generate the final features [58].

The EasyMKL was used to maximize the distance between the convex hulls of positive and negative samples in the training set. The SVM is used to build a

predictive model and decision function given the combined kernel. The ksrMKL [58] only used functional information; other biological information could be integrated to improve the result.

In [59] this method explores the relationship between protein kinases and disease-related phosphorylation substrates using PSEA [59]. The similarity scores of each peptide against one another were calculated using a substitution matrix from BLOSUM62 [59]. The similarities were then mixed with deriving high and low. Then those were used to find the enrichment score. Then the results were evaluated. [59] didnt need a positive and negative example balance, and we can use this technique to calculate the sequence similarity between peptides directly. It also handles the subset differences hence allowing you to consider the subset difference in kinase subset [59]. The enrichment analysis was calculated using GO terms.

In [60] CMS(composition of monomer strategy) was used to encode the sequences. The monomer spectrum was also used to represent the composition of the amino-acid and the frequency occurrence of each amino-acid window [60]. Then support vector machine was applied to this encoding, and the RBF was used as the kernel function [60].

[55] uses the pattern of sequences alongside the evolutionary information to identify phosphorylation sites. A noise-reducing algorithm was applied to this information to find the ideal phosphorylation sites [55] [56]. The method didnt require sophisticated training algorithm. In this method, similarity scores were calculated using BLOSUM62 and profile-profile alignment that contains the information on evolution [55]. The closer an unknown site is to known location the more phos-

Simple Consensus Pattern based	Sequence similarity-based	Machine Learning
ELM, PROSITE	PostMod, PSEA	Musite

phorylated they are. More features can be added to improve the accuracy of the method [55].

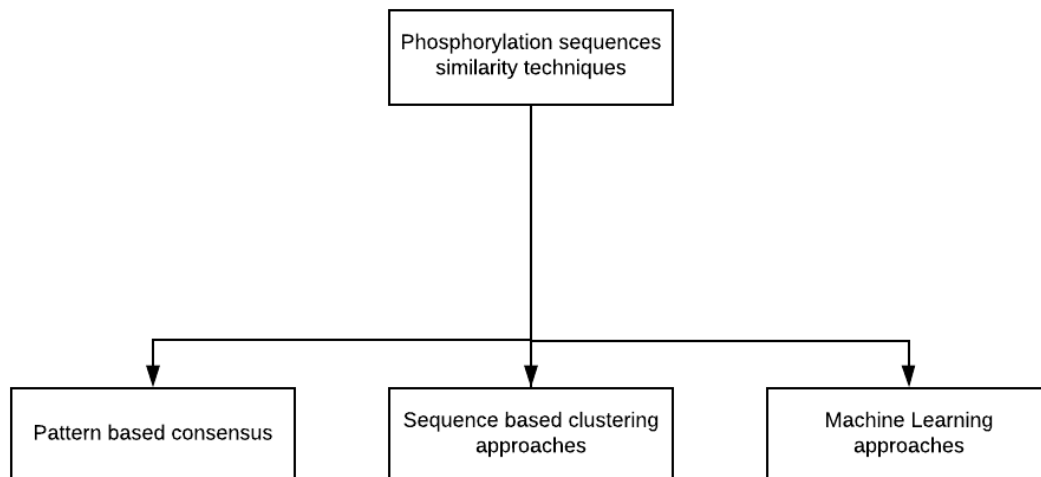


Figure 2.3: Techniques for finding similarities in sequences using phosphorylated proteins

2.1.3.3 Phenotypic and side effect based approaches:

As defined the phenotype previously speaks to the behavior or trait of an organism. It can be physiological, a morphological property [5]. They are traits that constitute an individual or organisms behavior. They are observed by studying that individual genome. When a patient is being observed in the hospital, they are mostly found based on some phenotypic properties. This technique is used for drug repositioning as well.

This approach is useful in drug discovery because it provides direct evidence

over an underlying situation. The side effects are also considered phenotype. It does not matter how much a drug works in an animal model, and we know the real impact after the clinical trial. If we can adequately capture the side effects, we can effectively reposition the drug.

If two drugs have the same binding target, they would have the same side-effect [61]. We can gather a side effect profile for different medications and use the similarities to identify some potential drugs that are off-targets [62]. In Campillo's work, the side effects were mined using some text mining packages and some statistical analysis these texts to find the probability of two drugs having the same targets.

2.1.3.4 Network based Approach

The network-based approach takes into consideration drug-drug network, drug-disease network, protein interaction network, drug-target network, biological pathway network and how the following have been used to achieve drug-repositioning.

[63] established a bi-partite way of defining the network using drug-target construct to achieve drug repositioning. Drug pair similarity used to find the similarity between drug structures and common drug and their interactions. The research was based on previous research [35, 63–66] that used a multilayered approach on gene, disease and drug target to identify new use for old drugs.

Network clustering approaches have also been applied to find new uses for old drugs. [66] used network clustering on heterogeneous data. The idea was to be able

to identify closely connected drugs and diseases that could be used to then mine drug and disease pairs for drug repositioning. The network consisted of two nodes with drugs and disease connection, the genes they shared and other features like the biological process. They were all connected, and the Jaccard score was used to weight the connections.

2.1.3.5 Network based Approach: Disease

The research proposes [35,65] two-step approach to drug repositioning based on the protein to protein interaction network of two diseases. Given that you have two conditions; the protein-protein interaction network of the two diseases was found, and the similarity between the drugs prescribed was used for drug repositioning. The list of the disease that relates to a specific gene was found from meta-database called generator. In the generator, if a disease shares genes with another disease, they were selected. If a drug was found to have the same target in protein to protein network, then the drug was deemed to be repositioned or deem to be a repositioning candidate [35, 65]. The targets for many drugs are still yet to be known so this implied repositioning. The experiment was applied to four different types of diseases that included hypertension, diabetes mellitus, Crohn's disease, and autism. Repositioning candidates were found at both steps [35,65] .

2.1.3.6 Machine Learning Approach

The various data support drug repositioning can be used alongside machine learning-based models to predict associations between drugs and diseases. The machine learning approaches use biomedical concepts to help computers understand how to predict drug repositioning candidates [35,65]. The biomedical concepts help us to train machine-learning algorithms and then predictions from there. One of the recent work in this space involved using SVM (Support Vector Machines). The algorithm was used to predict therapeutic categories, and the misclassified data points were interpreted as drug repositioning assumptions [35,65].

Menden et al. developed a machine learning model that predicts the way cancer cell lines respond to drug treatment [35,62,65]. Genomic level information was collected; they contained cancer cell lines and chemical property information was also obtained. The data collected was used to build a feed-forward perceptron neural network model, and cross-validation and a blind test validated a random forest regression model - the predicted values.

Other methods use machine learning algorithms that use collaborative filtering techniques to predict unknown drug-disease associations.

Zhang et al. designed a unified framework for combining drug similarity and disease similarity [35,65]. A combination of phenome, genome and drug information were combined to arrive at this drug similarity matrix and disease similarity matrix. The report helps to do a drug-disease network analysis further. It was turned into an optimization problem that showed high efficiency in exploring drug repositioning

candidates.

2.1.3.7 Text-mining Approach: EHR Data for Analysis

Work on EHR has advanced over the years, and has grown across the world. EHR contains patient records, but they could be used to help in drug discovery. The EHR includes patients health records; it consists of a progress note, laboratory information about results of individual patients based on the test that was performed in the laboratory, information on the medication used by the patients, the vital signs of individual patients, demographic data, etc. The EHR is rapidly growing in countries like Canada, the US, and the UK. The adoption of EHR in the US has, and extensive research is done in this area.

The introduction of the population into the process of finding relationships between diseases and comorbidities can help understand the complex nature of conditions and help identify treatments. The relationships that exist within this disease network can help identify some pathophysiological mechanisms that can provide some insight into disease etiology and help to identify some new drug targets [35, 65, 66]. If a compound was determined to treat a disease that belongs to a cluster of other conditions, this might help identify a new indication of another illness that belongs to that cluster.

Hidalgo et al., 2009, analyzed 30 million records and disease co-morbidity network was built; where edges and nodes represented the strength of the disease relationships and diseases labeled by ICD9. The results showed that patients de-

velop disease close in proximity to the ones they already have. These edges for this experiment were identified using Pearson correlation and relative risk. They also found that diseases with closer proximities tend to have a worse prediction in comparison to diseases with fewer proximity [61].

2.1.3.8 Text-mining Approach: Social Media Data

Paul and Dredze 2012 worked on exploring recreational drug discussions. They used a multi-dimensional latent text model also known as factorial LDA to capture the corpora using some orthogonal factors and to help researchers create structured output to understand the content of that corpus [35,61,65] better. The structure of factorial LDA was modified to incorporate some existing knowledge, which included informative priors, and some background components. They could learn some factors that corresponded to drug type and delivery method and some aspects like chemistry, culture, and effects. They demonstrated that the improved model keeps a better performance and a more interpretable result [61].

In conclusion, the validation of the computational results needs to be tested experimentally. Usually, the computational approach can generate several results [67]. Ideally, the Identification of 3 best results relevant clinically for the patients is sent for experimental test.

The in vitro and in vivo model can be used to test the computational approach further to ensure the results are useful for repositioning. Kang et al. used the viability of a cell assay to validate the drug combination generated from a search

algorithm. The drug combinations that kills cancer cells were confirmed [68].

2.1.3.9 Chemical based Approach

In chemistry, similar structure implies the same function. The literature contains various approaches to calculate the structural similarity between two chemical compounds such as fingerprinting and clustering algorithms. The algorithms can be used to perform ligand-based similarity calculations where we find a set of already identified active ligands; the ligands are searched against a database to find structurally related drugs that are bioactive as well [35, 69, 70]. The drugs were represented as vectors and used as input to the clustering algorithm [35, 62].

[71] presents a systemic method of discovering new indications for known drugs by finding its relationship with two similar drugs. The study introduces a bipartite approach to finding common drug targets and connections. The study also introduces drug pairwise similarity [71] approach to calculate a score for the similarity between two different drugs. The method in this approach achieved better results than the state of the art approach. [72] The results showed that the combination of both chemical structure and drug target information resulted in higher performance.

The use of chemical approaches to achieve drug-repositioning is a great idea, and they are based on fundamental similarity principles, but a change to a molecular structure can lead a completely different biological outcome.

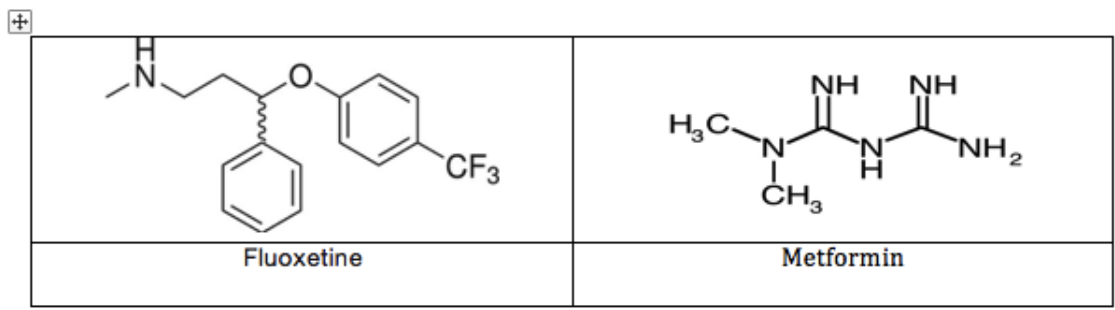


Figure 2.4: Dissimilar chemical structure.

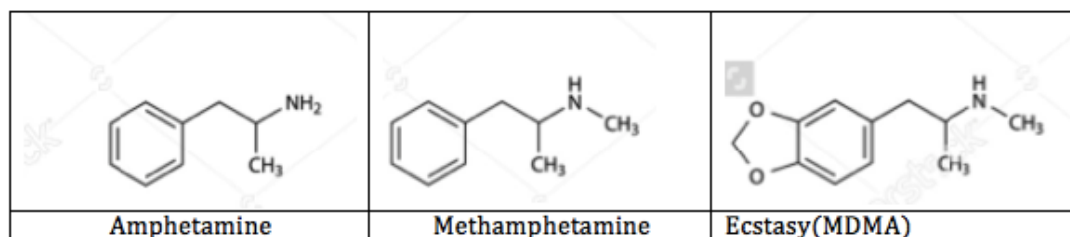


Figure 2.5: Similar chemical structure

2.1.3.10 Protein structure approach

Drugs or biomolecules work by interacting with proteins. This interaction is checked using computational software that can study 3D structures of the target and the drug. The process is known as protein docking. This technique finds the most optimal binding site for both the drug and drug target (protein). These types of studies help to improve the efficacy of the drug. The current dogma is that most drugs are known to interact with more than one protein [35].

[73] The goal for drug repositioning is to identify drugs that do not structurally bind to their proteins and then if the identified drugs are relevant to a disease, then a drug that correctly binds can be repositioned for that disease.

Concerning this idea, lots of research has been done in this area, including [73].

The structural similarity is the closest to the physical reality of the interactions [74]. [74] used 6000 binding site structures to find synapsin I, a protein used for the regulation of neurotransmitter release as a new indication of the drug staurosporine that binds PIM-1 kinase [74]. The finding was experimentally performed and verified in vitro.

2.1.3.11 Computational Approaches: Success Stories

In breast cancer, the repositioning of irinotecan for the treatment of cancer of the breast is one of the success stories. The data sources that were used include Genome (expression, structure, activity) and clinical trials. It was a biomarker-guide repositioning [1, 8].

In cancers, the repositioning of metformin for various types of cancers include breast. Colorectal, endometrial, esophageal, etc. using the clinical trial for cancer treatment with metformin [1, 8, 35]. In cancer for the repositioning of an existing phenothiazine-like antipsychotic drug, called trifluoperazine, as a potential anti-CSC agent using CMap [1].

The results from [72] showed after an evaluation that by finding the drug to diseases pairs on the clinical trials and PubMed [72]. Fluoxetine predicated six indications using a similar drug called Citalopram. 2 of the predicted uses are known to leave for four as new predictions [72]. It was found that alcoholism use was indicated in the clinical trial that was conducted to study fluoxetine in adolescents with alcohol use disorder and significant depression [72] and the other three results

have been investigated with results published.

2.1.3.12 Limitations of existing computational approaches.

1. The existing techniques are focused on drug-drug interactions, or disease-drug, drug-disease interactions. The techniques above do not address multi-modal network issues. Our technique will address multimodal issues using PIM Substrate Network.
2. The data would be large, and a dimension reduction technique or approach needs to be adapted to work with large datasets.
3. The technique utilized for the experiment needs to be generalizable due to the relevance of the technique to rare diseases and orphan diseases. Our technique generalizes the approach for any number of diseases.
4. Our approach will help to predict possible unique pathways that have not been identified before by previous systems as utilize a unique way of filtering the sequences first.
5. Previous studies utilize straightforward approaches to solving drug repositioning problem. Our approach suggests that it's important not to separate the sequence studies from the network studies. We present a multilayered approach to analyzing the data becomes very important. Currently, this approach is focused on a few types of bi-directional relationships. We want to expand on this.

6. The chemical concept of similarity computation is a great concept; however changes to a molecular structure can greatly impact biological outcomes. The focus on the chemical similarity will not deal with the situation because of this reason. Our approach introduces other ways of further evaluating the results in order to get a more authentic drug candidate [35].
7. Some of the techniques above required the drug and disease expression profile must be available to identify the drug candidates. However, the current CMAP database does not currently give us a large set of the drug and disease expression profiles so need to investigate other techniques or using CMAP as a layered approach to arrive at drug candidates become important [35].
8. The limitation for protein structural approach is that the structural data must be available such as PDB contains the records for protein structures. PDB has not covered all spectrum of the proteome. It is challenging to find the binding site based on the protein crystallized structure. The change in one amino acid can cause a dramatic change in the binding site structure. Our approach uses sequence data which are most readily available [35].

Chapter 3: Methodology

3.1 Research Contributions

The contribution of this work is to create a novel approach or framework for finding drug repositioning candidates. The significant contributions of these work are as follows.

1. We postulated the use of phosphorylated encoding (PhoSc) in extracting relevant information using sequence level data. This step is important because it leads us to the analysis of the information using sequence encoding, this encoding scheme helps us to analyze the sequence level data and to connect the data with relevant disease-related elements in order to find the appropriate drugs for repositioning using the phosphorylated proteins. *We utilized* our encoding scheme to *detect* the most pertinent phosphorylated proteins. It was implemented on the PIM substrates to find the most relevant PIM substrates. In the case of the rhabdomyosarcoma, the phosphorylated proteins were collected from phosphosite. *We create a method* to find the most relevant phosphorylated element using these encoding schemes.
2. We also proposed phosphorylated protein database extractor (PhoScExtrac-

tor) that captures the biological networks and other biologically related databases, applying a paradigm that uses a bi-directional path to determine drug repositioning candidates. An example of the bi-directional approach includes disease-gene and gene-disease relationship. In our approach, rather than focusing on disease-gene, gene-drug, gene-gene, or SNP-gene only, we have considered all the elements together putting them into one bucket; thus understanding their complex nature. *We create a database* that contained an integrated network that captures this complex data. Our approach utilizes the phosphorylated database constructed by us and takes into consideration the relationships between protein, drug, disease, environment, bioprocess, gene, chemical compound, and miRNA and utilize the datasets for analyzing and visualizing using network analytics techniques. Thirdly, studies have shown that proteins share functional connections. This implies that any protein in the ribosome and also in the basal transcriptional in the complex. The interactions between these proteins are typically not by chance, but they connect to establish a process. The PPI(protein-protein interaction) implies an interaction between two proteins that dont necessarily have to perpetual or fixed. The current database has shallow coverage of the PPIs (protein-protein interactions). Our research studies the PPIs(protein-protein interaction) in both rhabdomyosarcoma and PIM substrates using singular value decomposition.

3. We have also utilized the database or network to find the highly significant nodes using network analytic techniques for phosphorylated (NetAnaPhoS).

The network that characterizes the significant nodes was used for analysis and towards finding drug repositioning candidates. *We identified vital proteins and biological pathways* in rhabdomyosarcoma and PIM substrates. Pathways are great leads to finding drug candidates. *We utilized influence nodes measures* to identify some of these pathways relevant to finding potential drug candidates.

4. We have developed a novel multilayered approach or multirelational approach to handling a combination of all the networks that were integrated using multiple databases. We applied a novel multimodal network clustering approach to handling numerous bi-mode relationships for rhabdomyosarcoma and PIM substrates (phosphorylated proteins) called DReiM. Our method helps to predict unknown relationships by analyzing the integrated networks. The approach also reveals several patterns in the datasets that contain a complex multimodal system.

3.1.1 Introduction to our approach

The internet has a vast amount of health-related information or data set; the need to make this data comprehensible becomes essential. The data can also play a crucial role in finding significant disease-related elements that could lead to drug repositioning. These data sets are significantly large and even sophisticated. Due to the complex nature of those networks, it is crucial to analyze these data sets in a multidimensional manner.

In this chapter, our novel framework that consists of clustering sequences based

on their function, integration of biological datasets, and applying network approach and multimode approach to analyzing complex biological data towards finding drug candidates was presented.

Previously the identification of drug candidates or other relevant elements that lead to drug repositioning is typically narrow focused. Our contribution integrates both sequence level, gene expression level and network level information towards identifying drug repositioning candidate.

In previous methods, sequences are grouped based on similarities of the alphabets in the sequences. Our approach introduces a sequence encoding technique used to encode protein sequences based on their function. The sequence encoding technique then further helped to group these proteins based on their relevant features. We introduce methods of clustering these sequence.

The method above help to identify relevant protein sequences for data integration and then to further apply network analytics and multimode approaches.

In between these techniques, we developed a gene expression level framework as well alongside to add more drug candidates to our results.

3.1.2 Overall system architecture

The overall architecture of the system shows our approach in a step by step process. The process starts with sequence analysis using encoding schemes and separating the chaff from the wheat. The steps below illustrate the process flow for our technique.

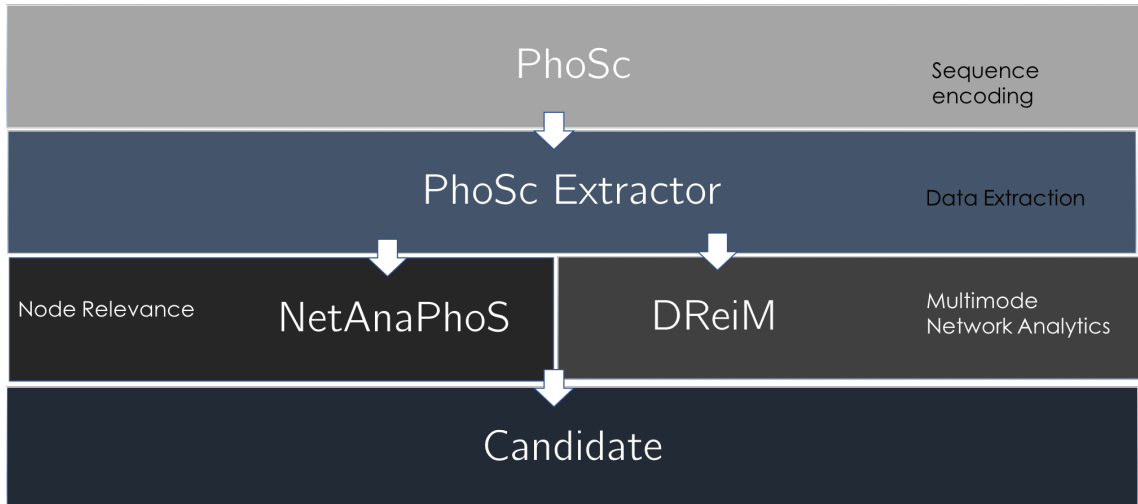


Figure 3.1: Overview of the Architecture.

3.2 Datasets

The datasets for the sequence analytics portion of our experiment were collected from Bieberich Lab. The datasets contain a list of pim kinase or substrates. We also utilized the phosphosite plus website to gather sequence-related information for rhabdomyosarcoma. For the purpose, multimode clustering towards drug repositioning. We have used a wide range of databases such as UniProt, DIS-GeNet, CIDER, and so forth to create the Phosphorylated Network Database. Our databases contain sets of datasets for rhabdomyosarcoma, diabetes and pim substrates.

3.2.1 Rhabdomyosarcoma Datasets

We constructed the Rhabdomyosarcoma Datasets using the Phosphosite plus website to pull the sequence information and then further pull the information for

clustering using multiple sets of databases that will be discussed in the coming section. We created the datasets we used for multimode clustering (RhabPhos).

3.2.2 Diabetes Datasets

We constructed our Diabetes Datasets Phosphorylated Proteins using the CiDER datasets. The CiDER datasets contain groups of subject to object relationships. We drilled downed the datasets to find interactions that take place only when phosphorylation takes place (DiabPhos).

3.2.3 PIM Datasets

We collected the datasets from the Bieberich Lab. The PIM Datasets had three critical columns. The sequence or substrate column, the variable modifications column which is a portion that helps us identify the position of phosphorylation and the protein accession number which represents the protein the sequence belongs. The substrate is just a piece of the large set of sequences [75].

3.2.4 Experimental Setup

For this work, we conducted our experiment using four cores, 32GB RAM, 3401 MHz, and Hard Disk is one terabyte. The experiments were running on a Microsoft Windows Professional version 6.1.7601. The algorithm for developing our database management system was implemented using SQL database. Our algorithms were implemented using python and MATLAB.

We utilized aws sagemaker for implementing the piece of our work that involved large datasets and r studio for chemical similarity computation using our algorithm.

3.3 Identification of Relevant Protein Sequence(PhoSc)

We have formulated a novel process for the identification of relevant pim and rhabdomyosarcoma protein sequences using encoding schemes called PhoSc.

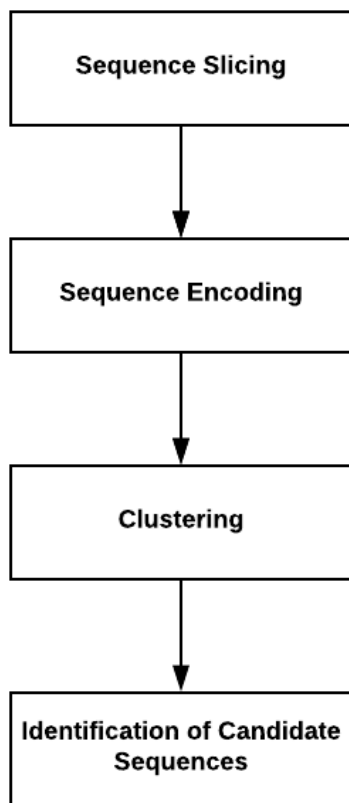


Figure 3.2: Overview of PhoSc process

We utilized the phosphorylated alphabets in encoding schemes section to encode our sequences. We applied this encoding to both pim substrates and rhabdomyosarcoma and clustered our sequences. The first steps are sequence data pro-

cessing and slicing.

3.3.0.1 Sequence Data Preprocessing and Sequence slicing.

This step involves the slicing of the essential component of the sequence from the UniProt database. The PIM Substrates UniProt identification number was used to extract the sequences of each substrate from the UniProt site. The substrates are subsequences from the primary sequence. The substrates are the relevant portions of the sequence because they are the portion of the sequence that phosphorylates. The substrates were extracted alongside their protein accession number, protein description, protein sequence, variable modifications, and variable modification sites.

The variable modification site is used to find the position where phosphorylation took place. The area where the phosphorylation takes place is identified by the number 3. The number 3 position on the sequence is letter S or T.

For example, if 0.0030 is the decimal number. Position 3 as number 3 and therefore represents the location where phosphorylation takes place. 10,20,30,40 and 50 sequences up and down this position is cut for analysis.

The phosphorylation site of the first PIM substrate sequence is below on Table [3.1](#). S is on position 3, and it is on the third position after the decimal. The next step is to find ten sequences up and ten sequences down of t. We would need to extract the whole sequence 1433B_HUMAN from NCBI and search for the substrate across the entire sequence. The position of the S in the substrate was queried in the 1433B_HUMAN sequence on UniProt. The position S was used to identify ten

sequences up and ten sequences down. (Refer to figure 3.3)

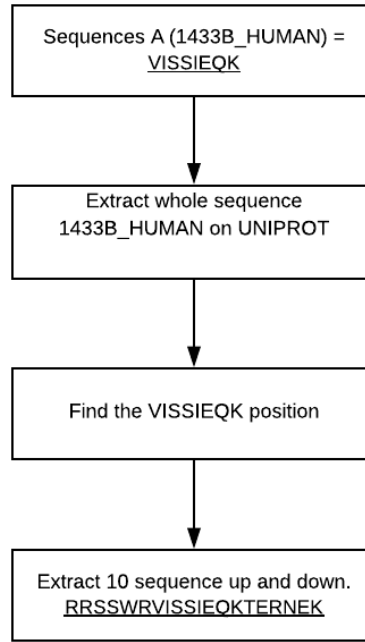


Figure 3.3: Sequence Extraction

From the position with variable modification site, we extracted 10,20,30,40,50 sequence up and down to perform sequence analysis.

Protein Accession ID	Protein Sequence	Variable Modification Sites
1433B_HUMAN	VISSIEQK	0.00300000.0

Table 3.1: PIM Substrates Sample

The sample of results after the processing of the sequence was done is below:

Protein Accession ID	Protein Sequence
1433B_HUMAN	GARRSSWRVISSIEQKTERN

Table 3.2: Processed PIM Substrates

3.3.0.2 Background in Bioinformatics

We will like to explain the three main elements involved in the central dogma of molecular biology. Understanding these terms will help better understand the encoding schemes.

Deoxyribonucleic acid known as DNA, it contains all the genetic information. DNA carries genes that encode the attributes of all living things [76]. It has its name deoxyribonucleic acid because its building blocks which are the nucleotides, Adenine (A), Thymine (T), Guanine (G), and Cytosine (C), the nucleotide contains a nitrogen base, a sugar base known as deoxyribose and a phosphate group [76]. The nucleotides are complementary, A pairs with T, while C pairs with G. They are bound by hydrogen bonds [76]. DNA are two complementary strands that come together to form a helix usually referred to them as a double helix. DNA can fold in various structures. DNA in its condensed form is called chromosomes [76]. When the DNA does not need to replicate or transcribed, it is packed tightly, a shorter form known as the chromosome. However, whenever there is replication or transcription, the DNA is stretched out into a long strand for transcription or replication to take place. DNA can use one of its strands and duplicate another strand using an enzyme known as DNA polymerase [76]. DNA not only replicate itself; however, it can also be transcribed into RNA and then later translated to protein [76].

Ribose nucleic acid is known as RNA. RNA is the transcribed form of the DNA, where one of the DNA strands serves as a template [76]. Transcription of RNA takes place at the nucleus level of the cell where the DNA lives. RNA, unlike

DNA, is one strand [76]. RNA nucleotides consist of Adenine, Uracil, Guanine, and Cytosine. Where Uracil replaces thymine which pairs with A. Once transcribed RNA has coding region exon and non-coding region introns which are spliced out of the strand, this is called post-transcriptional modification [76]. Once introns are spliced out, the RNA is in a messenger RNA form (mRNA) which leaves the nucleus and travels to the ribosome an organelle that translates the mRNA into protein [76]. In the ribosome is rRNA that joins amino acid that is provided by tRNA to form polypeptide chains (protein) [76].

Once the translation has taken place where there is the polypeptide, the post-translation modification takes place [77]. These changes are known as post-translation modifications are chemical modifications to a protein. are called These changes help the protein functions well, as it helps the structure of the protein [77]. The structure of a protein makes the protein function adequately if the protein does not have its proper structure; it will not functions well. Protein folding is one fundamental reason why a protein might not function properly; if the protein does not fold into the right structure [78]. The common post-translational modification includes the cleaving off of precursor proteins, disulfide or covalent bonds formation, acetylation, methylation, phosphorylation, glycosylation, etc. [78]. The typical modification is phosphorylation, acetylation, and methylation. The phosphorylation event adds a phosphate group to the protein. This gives the protein the energy needed to bind to other proteins [78]. Methylation on another hand is the addition of a methyl group. Acetylation, on the other hand, adds an acetyl functional group to the protein [78]. In sum post-translational modification make

the protein to function in its best form.

3.3.0.3 Encoding Scheme:

The encoding scheme was used to encode sequences(PIM Substrates). The encoding schemes help to conserve both biology and chemical properties we used the encoding scheme in our experiment to conserve the phosphorylated alphabets, and we further performed singular value decomposition to reduce the size of the datasets and to filter the best sets of PIM substrates in our dataset. Though the singular value decomposition method didn't give the best results, we came up with a novel way of using a combination of phosphorylated alphabets and a positional way of clustering the sequences.

We utilized other techniques to convert sequences to binary vectors to analyze to identify active substrates relevant for the identification of drug candidates. The standard encoding is the primary encoding scheme. The standard encoding scheme utilizes the 20 amino acid characters (protein related characters). They include A, R, E, N, C, D, Q, L, H, I, G, K, M, F, S, P, Y, W, T, and V. Each of the amino acid character is coded into a n-dimensional vector space, where n represents the number of symbols.

Given a DNA sequence have four letters, DNA sequence = [A,G,C,T] The A represents 0 0 0 1 , G represents 0 0 1 0, C represents 0 1 0 0, and T represents 1 0 0 0. Given a sequence D = AGTCCA (sequence length = 6). The result of D sequence = [0001, 0010, 1000, 0100, 0100, 0001]. The sequence length of D is 24.

The vector dimension will be 24. The encoding generates a sparse matrix. Using the phosphorylated alphabets below the third column shows the encoding of the different categories of the alphabets. Several alphabet choices where the alphabet size is greater than or equal to 4 results in fairly low error rates based on research. The various encoding schemes below were used to encode sequence 10, 20, 30, 40, and 50. The encoding scheme characterizes all the sequences based on their sequence lengths, and the encoded sequences are then used to cluster the proteins. Each of the sequences was represented with their numerical representation. The charged alphabets has three different categories which include the acidic alphabets D and E, the basic alphabets which include R,H, and K, and neutral alphabets represented by A,N,C,Q,G,I,L,M,F,P,S,T,W,Y,V. The numerical representation for the 3 different categories will be $[0\ 0\ 1]$, $[1\ 0\ 0]$, $[0\ 1\ 0]$. $[0\ 0\ 1]$ would be classified as the acidic alphabet's numeric representation, $[1\ 0\ 0]$ would be classified as the basic alphabet's numeric representation, and $[0\ 1\ 0]$ would be classified as the neutral alphabets numeric representation. The final goal will be to use the protein sequences to cluster the sequences to find the clusters of proteins that are relevant in the phosphorylation process. In the next section, we discuss approaches that could lead to the clustering of these protein sequences [37] [79]. This approach includes consensus sequence and PSSM score [79]. We applied our novel approach using phosphorylated alphabets(PhoSc) and phosphorylated alphabets -scaled which takes into consideration the hydropathy index PhoSc-con.

Phosphorylated Categories	Phosphorylated Alphabets	Phosphorylated Encoding
NPR	G,A,V,L,I,M,P	00001
AR	F,W,Y	00010
PUR	S,T,C,N,Q	00100
PR	K,R,H	01000
NR	D,E	10000

Table 3.3: Phosphorylated Alphabets with Encoding

Structural Categories	Structural Alphabets
ambstructure	A,C,G,P,S,T,W,Y
extstructure	R,N,D,Q,E,H,K
intstructure	'T',L,M,F,V

Table 3.4: Structural Alphabets

Charge Categories	Charge Alphabets
acidic	D,E
basic	R,H,K
neutral	A,N,C,Q,G,I,L,M,F,P,S,T,W,Y,V

Table 3.5: Charged Alphabets

Functional Categories	Functional Alphabets
acidic	D,E
basic	R,H,K
hydrophobic	A,'T',L,M,F,P,W',V
polar	N,C,Q,G,S,T,Y

Table 3.6: Functional Alphabets

3.3.0.4 Relevance of Sequences using Consensus Sequence

Using the sequences we can instantly find the consensus sequence. The consensus sequence is a sequence that explains or characterizes a set of sequences. The goal

Volume Categories	Volume Alphabets
small	A,G,S
med	P,C,D,N,T
medlarge	I,L,M,V,E,H,K,Q,R
large	Y,F,W

Table 3.7: Volume Alphabets

Phosphorylated Categories	Phosphorylated Alphabets
NPR	G,A,V,L,I,M,P
AR	F,W,Y
PUR	S,T,C,N,Q
PR	K,R,H
NR	D,E

Table 3.8: Phosphorylated Alphabets

in this project is to be able to find the most probable sequence that characterizes the PIM substrates. It selects the amino acid with the highest chances of occurring. The Table 3.9 below gives us an example. The consensus sequence further helps to determine our close new sequences are to the center of the current sequences. The result section shows the consensus sequence for sequences of length 10,20,30,40 and 50.

	Position 1	Position 2	Position 3	Position 4	Position 5
Sequence 1	F	V	V	Y	E
Sequence 2	G	G	E	Y	F
Sequence 3	T	G	G	E	E
Sequence 4	T	G	G	Y	E
Sequence 5	Y	V	G	Y	V
Consensus	T	G	G	Y	E

Table 3.9: Demonstrate Consensus Sequence

3.3.0.5 Relevance of Sequences using Position Scoring Matrix

The consensus sequence is not enough to judge the performance of multiple alignments because it does not take into consideration the situation where you have more than two symbols in a given position. The position scoring matrix will help us derive the frequency of occurrence of the amino acid at each position. See Table 3.10 and 3.11. The result section shows the PSSM score for sequences of length 10,20,30,40 and 50.

Sequence 1	A	A	G	C	A	A	C
Sequence 2	A	A	G	G	C	A	G
Sequence 3	A	G	G	C	G	A	G
Sequence 4	G	G	C	A	G	C	G

Table 3.10: Sequence Analysis for PSSM

	Position 1	Position 2	Position 3	Position 4	Position 5	Position 6	Position 7
A	3/4	1/2	0	1/4	1/4	3/4	0
G	0	1/2	3/4	1/2	1/2	0	3/4
C	0	0	0	1/2	1/4	1/4	1/4
T	1/4	0	1/4	1/4	0	0	0
PSSM	A	G	G	T	G	A	G

Table 3.11: Position Scoring Matrix - Frequency Table

3.3.0.6 Term-Document Matrix and Adjacency Matrix

Each row of the matrix contains PIM substrate sequence . Each row was encoded using an encoding scheme. Each sequence represent a document. Each amino

acid sequence represent a Term. The representation of Term-Document Matrix and Adjacency Matrix are on Table 3.12 and Table 3.14

The column of the matrix defines the frequency of occurrence of each of the terms that exist in a given document.

	Document 1	Document 2	Document 3	Document 4	Document 5
Term 1	0	1	1	0	0
Term 2	1	0	0	0	1
Term 3	1	0	1	0	1
Term 4	0	1	1	1	0

Table 3.12: Term-Document Matrix

For example, DNA Sequence consists of 4 letters = A,G,C,T. Each of the letters will have their numeric encoding A =0001, G =0010, C=0100, and T=1000. The dimension is 2 x 20.

Sequence 1 = AGTAA Sequence 2 =ATACC

The binary encoding for the sequence above is used to encode each of the sequences. The sequence represents documents and letters (DNA sequences) serve the terms.

	Position 1	Position 2	Position 3	Position 4	Position 5
Sequence 1	0 0 0 1	0 0 1 0	1 0 0 0	0 0 0 1	0 0 0 1
Sequence 2	0 0 0 1	1 0 0 0	0 0 0 1	0 1 0 0	0 1 0 0

Table 3.13: Sequence Encoding for two sequences

An adjacency matrix can be represented as a square matrix used to represent a graph. 0 represents no occurrence of a relationship while 1 represents an occurrence of a relationship. The adjacency matrix was used to create the multimodal network.

This will be discussed in subsequent sections.

	Term 1	Term 2	Term 3	Term 4
Term 1	0	1	1	0
Term 2	1	0	0	0
Term 3	1	0	1	0
Term 4	0	1	1	1

Table 3.14: Adjacency Matrix

3.3.0.7 Sequence Distance Calculation

Various distance measures exist to find the distances between sequences. For our experiment, we utilized hamming distance. The hamming distance works best for our method because it takes into consideration both the position and the alphabets. Protein sequences are groups of alphabets that have a specific order. The protein sequence distance can be calculated using hamming distances which is the distance between two binary protein numbers. The distance can be got by counting the number of positions where the values are different. It is the number of a group of bits in the XOR of two binary numbers. The Hamming distance is also great for strings, numbers or binary of equal length all of which are a property of the protein sequence we are studying.

For example, Input: $string1[] = 1011101, string2[] = 1000001$ Output: 3

In the case of PIM Substrate, the PIM Substrate have sequences of equal length, and the Hamming distance is a way of finding the difference between a sequence and another.

	C1	C2	C3	C4	C5
Sequence1	A	C	C	T	G
Sequence2	C	C	T	T	G

Table 3.15: Sequences for Hamming Distance Calculations

Given string 1 and string 2,

$string1[] = ABCDFG$

$string2[] = ABYYFG$

Output: 2 This is because we have the third position and fourth position in string1 different from the third position and fourth position in string2. 2 is the distance of the two sequences. The distance between all PIM Substrate sequence was found for our experiment. The hamming distance was a good tool to find the difference between two sequences. We will further illustrate how the sequence was set up for analysis using the hamming distance. The next section explains another useful application of the hamming distance.

3.3.0.8 Sequence Distance Calculations using Categorical values or Encoding Schemes

For example, suppose that there are five categorical variables, C1 to C5, each with three categories, which we denote by a/g/c/t and that there are two sequences with the following characteristics:

Then the number of matches is 3, and the number of mismatches is 2; hence the distance between the two samples is two divided by 5, the number of variables, that is 0.4. This is called the simple matching coefficient. Sometimes this coefficient is

expressed in terms of similarity, not dissimilarity, in which case it would be equal to 0.6. In this case, we stick to distances, in other words, dissimilarities or mismatches. The coefficient is directly proportional to the squared Euclidean distance calculated between these data in dummy variable form, where each category defines a zero-one variable:

	C1a	C1g	C1c	C1t	C2a	C2g	C2c	C2t	C3a	C3g	C3c	C3t	C4a	C4g	C4c	C4t	C5a	C5g	C5c	C5t
Sequence1	1	0	0	0	0	0	1	0	0	0	1	0	0	0	0	1	0	1	0	0
Sequence2	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	1	0	1	0	0

Table 3.16: Sequences for Hamming Distance Categorical Tables

The squared Euclidean distance sums the squared differences between these two vectors: if there is an agreement (there are two matches in this example) there is zero-sum of squared deviations, but if otherwise, we have two possibilities, +1 or 1, each giving 1 when squared. So the sum of squared differences here is 6, and if this is expressed relative to the maximum discrepancy that can be achieved, namely 10 when there are no matches in the five variables, then this gives the same value 0.6 as before. There are several ways of matching coefficient, and one of them is the chi-square distance for multivariate categorical data, which introduces a weighting of each category inverse to its mean value, as for profile data based on counts.

With sequences of length of 60, with 20 possible categorical values its $60 \times 20 = 1200$. The total occurrences of each category are denoted by (n_1, \dots, n_j) with total $(n = \sum n_j^n)$. Since the totals for each variable equal the sample size, n will be the sample size times the number of variables). Then define c_j as follows: $c_j = n_j/n$ and use $1/c_j$ as weights in a weighted Euclidean distance between the samples coded in dummy variable form. The idea here is, as before, that the rarity of a category

should count more than in the distance than a normal category. Just like the chi-square distance function is at the heart of correspondence analysis of abundance data, so this form of the chi-square for multivariate categorical data is at the heart of multiple correspondence analysis.

With this illustration, we can take an encoding scheme and find similarity between two sequences using the same method as stated above. The result section contained some of the encoding schemes figures.

3.4 Clustering our Encoding schemes

We applied the clustering algorithm to the analysis of the PIM substrates [80] [81] [82]. We first performed sequence encoding, and then applied singular value decomposition to reduce the size of the sparse matrix and to find a small subset of the matrix that characterizes the entire data set [80] [81] [82]. The instances of the PIM substrates have not been pre-classified in any way or form before and did not belong to any class [81] [82] [80]. The PIM substrates are grouped based on similarities. The PIM Substrates are grouped into clusters. The clusters contain proteins that are similar to one another.

We applied intra-connectivity and inter-connectivity to the PIM substrates. The Intra-connectivity measures the density of connections between the PIM substrate instances of a single cluster. If a cluster is highly intra-connected, the PIM substrate instances will be grouped within the same cluster and are highly dependent on each instance. Interconnectivity measures show how distinct individual clusters

are to each other. The results from the SVD calculations were utilized to find the cluster of sequences.

3.4.0.1 Apply k-means clustering on PIM Substrates

The k-means cluster helps to partition the PIM substrates in k number of partitions, each of the points or PIM substrates belonging to a specific cluster. The specific clusters were further studied to find a relationship that exists between two substrates. The closer a substrate is to another the more related they are with each other. We tested kMeans Clustering on the substrates.

3.4.0.2 Apply Hierarchical clustering on PIM Substrate

The hierarchical method of clustering is a method of clustering objects into groups by finding their distances [83] [46]. The first step is finding the distances between the objects. For example in the case of sequences, the distances between all protein sequences were found. Each of these distance was treated at first independently of one another. In other words, each represents a cluster [83] [46]. The algorithm will then find the two clusters that are close to one another and merge two most similar clusters. This will go on until all the clusters are combined.

The distance was calculated by using distance metrics like Euclidean distance. Many other distance metrics can be used. The Euclidean distance is defined as the length of the path between two points [84] [83].

The goal is to use the Euclidean distance to determine hierarchical clusters

between the PIM sequences by first finding the hamming distance between the sequences [46]. It is also vital to have distance metrics have been selected to have a linkage criterion [85]. The linkage criteria help to determine where the distance is being created. The distance is either cluster based on two very similar parts of the cluster which are also known as single-linkage or two least similar which is also referred to as complete linkage or the center of the cluster which is also known as the average-linkage. We have selected the single-linkage in order to group by closely similar clusters. Similar clusters are then typically combined sequentially.

The theory behind utilizing the hierarchical clusters is the proteins within the same cluster are denser than the proteins in different clusters. The clusters also contain network motifs which imply the proteins within that cluster perform the same function [86] [87] [88].

3.4.0.3 Application of Singular Value Decomposition on the Sequences

Since the sequence as a possibility of producing a sparse matrix. The Singular Value Decomposition (SVD) was utilized to reduce the data sizes, the process of singular value decomposition can be considered as a process that removes noisy data and preserves relevant features of the data [89] [90] [91]. When SVD was applied on PIM substrates, it gives a set of singular values [92]. The singular value represents amino acid alphabets contained in an amino acid sequence [93]. It would make searching for the amino acid sequence easier to find [94]. The formula below is used to calculate the singular value decomposition [95] [96].

$$PIMSVD_{n \times m} = U_{n \times n} \Sigma_{n \times m} V^T_{m \times m} \quad (3.1)$$

U is an $n \times n$ orthogonal matrix, Σ is an $m \times n$ orthogonal matrix and V is an $m \times m$ diagonal matrix of singular values [92].

3.4.0.4 Applications of n-grams for sequence clustering

Stuarts et al. started the work of finding the occurring frequency of amino-acid by applying the k-grams method [95] [45]. In k-gram, In the columns of k-grams are possible k-gram sequences and the frequency of their occurrence within each of phosphorylated sequences was computed [79]. Given protein sequences for example, if $k = 2$, at their will exist 200 possible 2-grams [95]. Given 400 sequences, the k-gram matrix will be 400 x 200. The columns will contain a sliding window of k frequency occurrence for k-gram; this will form the data matrix [95]. The result section contains the result for k-gram analytics.

3.4.0.5 Applications of LSTM for sequence clustering

The LSTM is a type of Recurrent Neural Network (RNN) known as Long Short Term Memory. In the RNN are networks, the internal memory allows for information to persist [97]. This means they can remember essential things about the input they received, which enables them to be very precise in predicting whats coming next [97].

The RNN carry the ability to utilize information in the previous timesteps,

giving the past and the present which makes the neural network best suited for sequences or time series data [97].

In our experiment, an LSTM implements an autoencoder using sequence data and since we were working protein sequences. We used the LSTM method to cluster our protein sequences. The LSTM is constructed as an encoder and decoder. The steps involved in the LSTM process is to read an input sequence (ASSNMT), it encodes this sequence, and then it decodes the sequences and recreates it. The models performance is measured by its ability to create again the sequence that was input. The moment the performance required is accomplished the decoder is completely ignored, and the encoder is left to encode input sequences to a specific length [97].

[98] describes how to use lstm to not only reconstruct the sequences but also so predict sequences as well. In this case, [98] used video frames to demonstrate this. This particular help for the unsupervised learning architecture. The input to the model will be sequences of vectors which are encoded and have gone through all the sequences, decodes and outputs the prediction for a sequence we chose to target.

3.4.1 Sum Squared Error

The Sum of Squared Error(SSE) represents the distance to the nearest PIM Substrate cluster [99].

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} (m_i, x) \quad (3.2)$$

The x represents the PIM substrates data point in cluster C_i and m_i is the point that was identified by C_i . The m_i represents the mean of the cluster. The lower the SSE, the more similar PIM substrates that belong to that cluster.

3.5 Data Extraction - PhoScExtractor

In the previous, steps we found the most relevant cluster of proteins involved in the process. The PhoScExtractor was used to pull data from the different sources. We then searched the sequence name using various database below to extract related elements (gene, drug, SNP, disease and so forth). The following databases were used to extract PIM substrates disease related element information. They include CIDER database [66], SIDER database [100], DisGenet database [101], CTD Database [102], mirdSNP database (a condition associated with SNPs) [103], PGXnet database, DGIdb (Drug Gene Interaction Database) [104], canSAR Databases [105], pharmaGKB, UniProt [106], BIOGRID [107], and Drugbank [108]. For our research on rhabdomyosarcoma, we already identified the relevant phosphorylated element. We use the databases to extract relevant disease-related element for rhabdomyosarcoma.

3.5.0.1 SIDER

The SIDER datasets contain small molecules or drugs and their recorded adverse reactions [109]. The datasets were extracted from public information which contains the frequency of the side effects, a drug with side effects, and information on the drug-target relationship [109]. We obtained drug side effect information from

SIDER to find possible drug combinations. [109].

3.5.0.2 DisGeNet

The DisGeNet Database contains several genes and the associated human disease. DisGeNet integrated for a database of expert-curated databases which include scientific literature, animal models and GWAS catalogs [101]. The DisGeNet contains 561,119 gene-disease associations (GDAs) between 17,074 genes and 20,370 diseases, traits or abnormal phenotypes, and 135,588 variant-disease associations (VDAs) between 83,002 SNPs and 9,169 diseases and phenotypes [101]. We extracted disease that related to genes that were associated with a protein on the substrate list [101].

3.5.0.3 CTD Database

The CTD Database also is known as the Comparative Toxicogenomics Database [110]. The database contains manually curated information about the chemical to gene to protein interactions, chemical to disease and gene to disease relationships [110]. They also contain pathway information, and they are currently working on a chemical to phenotype relationships for bio-marker identification [110].

3.5.0.4 dSNP

The Database contains manually curated dSNPs using the 3UTRs of human genes from PubMed [111]. The database contains 786 dSNP-disease associations for 603 unique dSNPs and 204 disease types. We utilize the datasets to extract relevant

dSNP to prostate cancer [111].

3.5.0.5 PheGenl

The database can be used as a phenotype to genotype integrator [111]. The genomic-wide association study contains information from Gene, dbGaP, OMIM, and dbSNP [111]. The database allows us to search for phenotypes that relate to certain genes or SNP [111]. We used the database during our project for this purpose.

3.5.0.6 UNIPROT

The UniProt database contains useful protein sequence resources and the functions of those protein resources [106]. We utilized this resource to find associations between already identified substrates from the Lab [106]. The database contains 555,100 information from SWISS-Prot and 88,032,926 information from TrEMBL [106].

3.5.0.7 DrugBank

The database contains biological and chemical information alongside drug information data and drug target information [108]. The database contains 9591 drug entries with 2037 FDA-approved small molecule drugs. 241 FDA approved biotech (protein/peptide) drugs and 6000 experimental drugs [108]. We extract drug to drug and drug to drug target relationships from this database [108].

3.5.0.8 PharmGKB

The protein to drug datasets was extracted from the pharmaGKB website [112]. The PharmaGKB site contains drug information which includes clinical information about the doses and the labels of the drug [112]. It also contains information about gene-drug associations and genotype-phenotype relationships as well as [112]. The database was also manually curated to find the response of drugs on human genes [112].

3.5.0.9 BIOGRID

The database consists of protein-protein interactions which were utilized during our experiment to find direct interactions between protein extracted from UniProt and other proteins that relate to it [107]. The BIOGRID Database contains 1,493,749 proteins and their genetic interactions, and also some of the chemical associations can be found on the website [107].

3.5.0.10 canSAR

The database contains the gene, disease, and drug data specifically available for canSAR research [105]. The datasets were used to extract relevant gene, drug, disease association with PIM substrates obtained for the Lab [105].

3.5.0.11 PGXnet

The database contains cancer-based drug, gene, and SNP associations [113]. It was extracted from the PharmGKB, GWAS catalog, and PGx Resources. The following interactions were extracted from the dataset [113]. The disease to a gene, drug to disease, SNP to a gene, SNP to a drug, and disease to gene [113].

3.5.0.12 DgIDB

The database contains drug gene interactions [114]. We utilized the database to get drug and gene associations after finding genes from proteins that were extracted from the pim substrates [114].

3.5.0.13 LODD-Linked Open Drug Data

The LODD database consortium took a survey of all available data on drugs on the world wide web [63]. The LODD paper discovered the best ways of organizing and integrating this datasets [63]. The Clinical Trial datasets used for our experiment were pulled from the LODD datasets and used for the evaluation of our results.

3.5.1 Data Integration Approach for Phosphorylated Elements(DIApe)

The datasets were extracted from various data sources. The data extraction section contains a list of data sources that we used to obtain the datasets we used for our experiments. The first step was to acquire the protein sequences and use

their protein accession number to annotate them. The next step is in the selection of databases to use, the databases that contained disease related elements in a bi-directional way was selected. The examples of such interactions are disease-drug, drug-gene, gene-disease and so forth. The next step will be to combine all this bi-directional relationship to form our complex multimodal network for analysis.

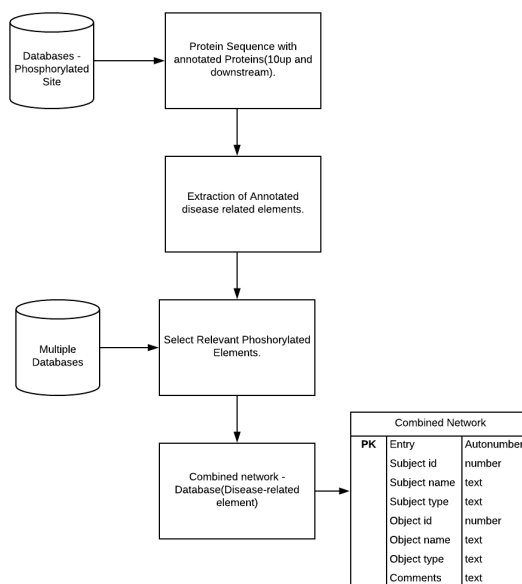


Figure 3.4: Our Data Integration Approach to creating Phosphorylated Network

3.6 Phosphorylated Network Databases-Data Processing for Network Analysis and Multimode Clustering

We acquired the database used for our analyses from various databases. UniProt was selected to find a majority of our datasets [106]. The Universal Protein Resource (UniProt) database is referred to as a tremendous resource for protein sequence and annotated data. Other databases used include ctdbase [102]; this database exposes

us to a full-bodied that helps understand the role our environment plays on our health. DisGeNET [115] incorporates gene and disease relationship curated by experts, [102] also known as Comparative Toxicogenomics Database and drug bank containing the relationship between drug and drug targets. The link we extracted from UniProt, disgenet, and ctdbase include gene-dbSNP, gene-pathway, go-drug bank, protein-protein, protein-gene, pathway-drug. The integrated networks include SNPs, genes, disease, drug bank, go protein and pathway. We combined all the components above into what we called an integrated network for network analysis.

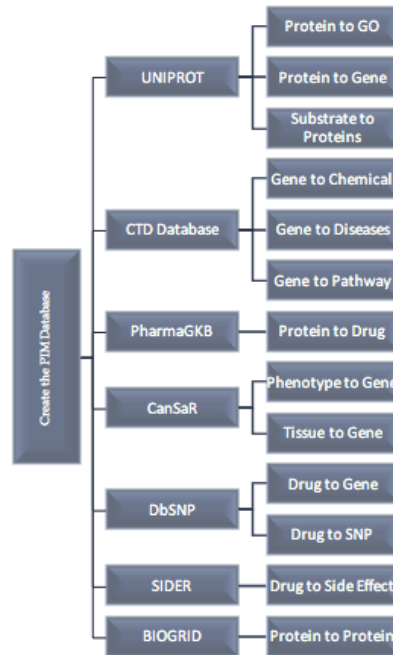


Figure 3.5: Data Extraction for PIM Kinase

In the context of a biological network, we have a direct application of more than one-mode due to the multi-layer nature of the biological network. An example of tri-mode is below:

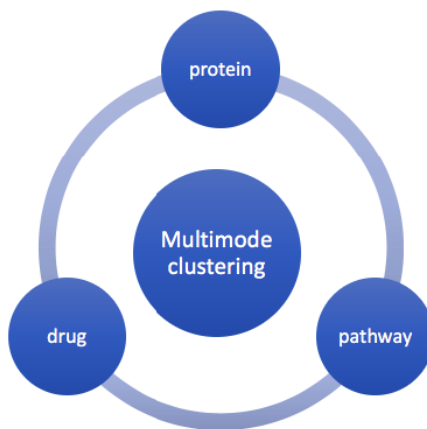


Figure 3.6: Tri-mode for Multimode Clustering

3.6.0.1 Example: Creating PIM Database Network

The following databases were used to extract high influence PIM substrates protein information and to create a database. They include CIDER database, SIDER database, DisGenet database, CTD Database, mirdSNP database (a disease associated with SNPs), PGXnet database, DGIdb (Drug Gene Interaction Database), canSAR Databases, pharmaGKB, UniProt, BIOGRID, and Drugbank.

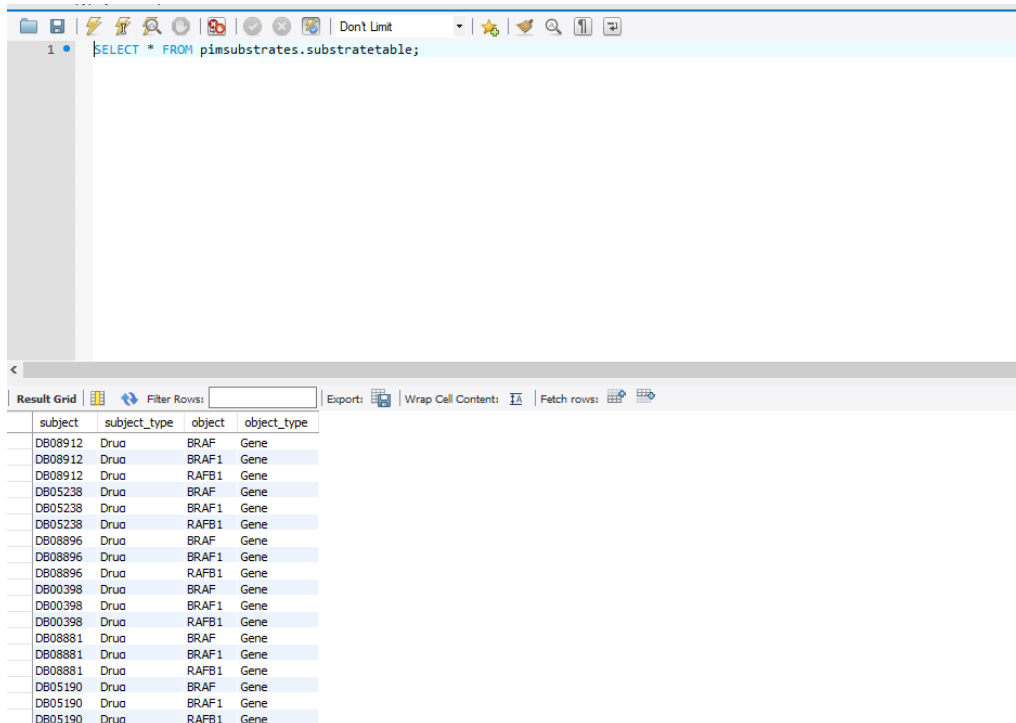
3.6.0.2 PIM Database : Querying

The SQL (Structured Query language) stores the data after cleaning was performed. For each PIM Substrate related interaction; we extracted subjects and objects along with their semantic types. The total number of the semantic types was 9, and this list included drug, gene/protein, SNP, drug. The database could be accessed through the SQL Workbench. Below are a couple of the queries executed on the PIM Database. The figure 3.8 show the entire database network. The figure

Databases	Subject	Object
CTD Databases	Gene	Chemical
CTD Databases	Gene	Diseases
CTD Databases	Gene	Pathway
DgIDB Databases	Drug	Gene
PGXnet	Disease	Gene
PGXnet	Drug	Disease
PGXnet	SNP	Gene
PGXnet	SNP	Drug
PGXnet	Disease	Gene
Pharmagkb	Protein	Drug
Drugbank	Protein	Drug
UNIPROT	Substrate	Proteins
UNIPROT	Proteins	Genes
MIRD	SNP	Gene
CanSar	Phenotype	gene
CanSar	Tissue	gene
DbSNP	Protein	SNP
DisGeNet	Drug	Disease
PharmaGKB	Pathway	Gene
UNIPROT	Protein	GO Ontology
SIDER	Drug	SIDE Effect
PheGenl	Tissue	SNP
BIOGRID	Protein	Protein

Figure 3.7: PIM substrates schema

3.9 shows the drugs with side-effects that relate to pain. The figure 3.10 show the relationship between specific genes and other disease-related elements that relate to it.



The screenshot shows a SQL database interface. At the top, there is a toolbar with various icons and a text input field containing the SQL query: `SELECT * FROM pimsubstrates.substratetable;`. Below the query, there is a "Result Grid" section. It includes a "Filter Rows:" input field, an "Export:" button, a "Wrap Cell Content:" checkbox, and a "Fetch rows:" button. The main area displays a table with the following data:

subject	subject_type	object	object_type
D808912	Drua	BRAF	Gene
D808912	Drua	BRAF1	Gene
D808912	Drua	RAFB1	Gene
D805238	Drua	BRAF	Gene
D805238	Drua	BRAF1	Gene
D805238	Drua	RAFB1	Gene
D808896	Drua	BRAF	Gene
D808896	Drua	BRAF1	Gene
D808896	Drua	RAFB1	Gene
D800398	Drua	BRAF	Gene
D800398	Drua	BRAF1	Gene
D800398	Drua	RAFB1	Gene
D808881	Drua	BRAF	Gene
D808881	Drua	BRAF1	Gene
D808881	Drua	RAFB1	Gene
D805190	Drua	BRAF	Gene
D805190	Drua	BRAF1	Gene
D805190	Drua	RAFB1	Gene

Figure 3.8: PIM Network SQL Database

3.7 Network Analytics Approach to Drug Repositioning -NetAnaPhoS

Network analytics techniques have been applied to a wide range of social media application problems such as Twitter, Facebook and health data in general. For a better understanding of the concepts and contents of this work, some terminologies used to need to be enumerated and explained. Some of these techniques were utilized to identify high influence disease related elements.

The input into our network analytics algorithm is the subject and object from the SQL database.

The screenshot shows a database query interface. At the top, a SQL query is entered in a text area:

```
1 • SELECT * FROM pimsubstrates.substratetable
2 WHERE Object LIKE 'pain%';
```

Below the query area, there is a toolbar with options: "Result Grid", "Filter Rows:" (with an input field), "Export:" (with a download icon), and "Wrap Cell Content:" (with a text icon). Below the toolbar is a table with the following data:

subject	subject_type	object	object_type
Dextropropoxyphene	drug	pain	disease
Diazepam	drug	pain	disease
Morphine	drug	pain	disease
Buprenorphine	drug	pain	disease

Figure 3.9: Query Pain using PIM Network Database

The screenshot shows a database query interface. At the top, a SQL query is entered in a text area:

```
1 • SELECT * FROM pimsubstrates.substratetable
2 WHERE object='BRAF';
```

Below the query area, there is a toolbar with options: "Result Grid", "Filter Rows:" (with an input field), "Export:" (with a download icon), and "Wrap Cell Content:" (with a text icon). Below the toolbar is a table with the following data:

subject	subject_type	object	object_type
D608912	Drug	BRAF	Gene
D605338	Drug	BRAF	Gene
D608896	Drug	BRAF	Gene
D600398	Drug	BRAF	Gene
D608891	Drug	BRAF	Gene
D605190	Drug	BRAF	Gene
P15056	protein	BRAF	Gene

Figure 3.10: Specific Gene Relationships in the PIM Network Database

3.7.0.1 Topological Structures

Our understanding of network topological principles helps to understand the characteristics of the network [116] [117]. In biological science, studying this network model is very important because it gives an idea of the structure of this network. Some of these models include:

3.7.0.2 Erdos Renyi Model - Random Network

In the Erdos-Renyi model, each node pair is connected with a probability Pr ; it is considered a form of random network and starts with a set of nodes. The pair of nodes create above makes up the graph [116] [117]. Each of these pairs of nodes makes up $PrN(N-1)/2$ random links. The path length is proportional to the network \log size. The degree of distribution follows the Poisson distribution. A majority of biological networks have giant connect components, and their average paths lengths are very close to that of Erdos Renyi network [116] [117].

3.7.0.3 Small World Network

A majority of nodes in a small world network are not connected to one another but if you take a look at the node individually, the neighbors of each node on the granular level are likely to be neighbors of one another. We can summarize a small-world network by saying that nodes that have few neighbors and a majority of the nodes can be reached from one another few steps [116] [117]. The connected components can be used for example to control the flow of information, mass or en-

ergy and also implications of a biological network. In biological networks (metabolic networks) they are reachable in a range(3-4) links to one another. It means that if we perturbed the network, it affects the network rapidly [116] [117].

3.7.0.4 Scale Free Network

Scale-free networks are those that have a power law degree distribution. The links originate from a node that displays the power law distribution. In scale-free networks nodes have a wide range of degrees; thus the network is free of any standard scale [116] [117]. The likelihood of linking to a given node i is proportional to the number of existing connections k_i that node has.

$$P \sim \frac{k_i}{\sum_j k_j} \quad (3.3)$$

3.7.0.5 Brief Overview of Network Properties

It is a group of two or more nodes linked together. The nodes in this respect would imply the PIM substrate(gene) and the pathways that are connected to these PIM substrates. The network can be directed, undirected or mixed.

As explained earlier, in this work, nodes can represent the relationship between gene and drug, gene and gene, disease and drug, drug and phenotype, phenotype and gene, etc. A network can be very complex, and thus particular ways to analyze the complex patterns existing within the network is essential. The challenge in studying complex systems is to develop simplified measures that capture some elements of

the complex structure understandably.

The PIM substrates were used to generate a complex network as described above with various types of relationships between disease-related elements.

The PIM Substrate is attached to specific genes and protein, the genes that attached to high influence substrate are selected. The genes selected will be referred to as high profile genes. A high profile gene is sometimes a significant number. The high profile gene is used to extract pathway information using the PIM Substrate Network Database.

Each gene is connected to its corresponding pathway. This forms a gene-pathway network or connection. We can define a network as either directed or undirected. The directed network has a collection of nodes connected by edges and having a direction associated with the connection. The undirected network also is a network that contains nodes that are connected without specific direction, and they are also known as bidirectional.

3.7.0.6 Degree Centrality

The centrality of a network describes the most central nodes in a network. This definition thus varies based on the context in which they are used. It depends on the degree, and also how related they are to the other networks based on closeness, betweenness, and eigenvector.

3.7.0.7 Betweenness Centrality

The betweenness was utilized to find the number of the shortest path from all nodes to all the other nodes that pass through that node. A node with high betweenness centrality has a tremendous influence on the transfer through the network; they follow the shortest paths. We find the shortest way for all nodes in the network to obtain the most influential node for betweenness centrality on the network. The nodes with high betweenness centrality were used to identify the drug repositioning candidates.

3.7.0.8 Closeness Centrality

We also applied closeness centrality to find the length of the average shortest path between a node and all nodes in the graph. We used this to find high influence nodes based on the length of the average shortest path between nodes.

3.7.0.9 Indegree and Outdegree

We calculated nodes that are directly linked to nodes or the number of edges pointing to a particular node. In this experiment, we also calculated the out-degree which indicates the number of nodes that an identified node connects to from itself. The measure was needed in this study to be able to understand how much impact a node has on the other nodes in its network.

3.7.0.10 Modularity

The calculation of the modularity scores contributes to measuring the power of a division of the PIM substrate network and rhabdomyosarcoma into communities. The networks that have high modularity scores have dense connections between the nodes within the community, but rare connections between nodes in different modules. It is used in the detection of community structure. The modularity technique was applied to find PIM Substrate network and rhabdomyosarcoma with high modularity that may imply that a specific cluster is more important to study than the others. We checked the results from the betweenness centrality to find out if high influence nodes also exist in the same community.

3.7.0.11 PageRank

The PageRank score was used to find the number of votes by all phosphorylated network (PIM substrate and rhabdomyosarcoma) disease-related element. The importance score for each of the nodes was generated. The phosphorylated elements are merged in a network form with disease-related elements to form the phosphorylated databases. We used the quality of links connected to a node to determine the score of the nodes itself.

3.7.0.12 Hub/Authority Scores

The high authority nodes or phosphorylated elements are the elements that are pointed to by many hub nodes. The hub nodes, on the other hand, have high

authority nodes that they point to but they are not authority nodes. The phosphorylated elements in the result will be classified a hub if it is one that leads to many high authorities; an authority phosphorylated related element will be classified as an authority if it is one that is pointed to by many high hub pages.

3.7.0.13 Connected Components

We also utilized the connected components to find how significant the nodes are. The connected component of an undirected graph is the maximum set of nodes such that a path connects each pair of nodes. Related components form a partition of the set of graph vertices, meaning that connected components are non-empty, they are pairwise disjoint, and the union of connected components forms the set of all vertices.

3.8 Node Impact Analysis using Single and Sequential Perturbation

Our results from using network analytical techniques were used to perform single and sequential perturbation. We want to know how important a node is compared to the others in this network.

3.8.0.1 Single Perturbation Analysis

The removal of one node in a computational experiment may indicate the alteration of a particular protein, the deletion of a relationship especially looking at expressions that are related between two sets of genes. Our computation allows the

removal of a phosphorylated node and shows how that node affects the entire nodes in the network [118]. In our experiment, we illustrated how important our nodes were by checking how disorganized they become when they are removed.

3.8.0.2 Sequential Perturbation Analysis

It a series of removal nodes one after the other without replacement, and this allows the visualization of the effects of that removal over the network [118]. As they are removed one by one, the network quickly disintegrates [118]. We want to understand how important the nodes were so we used this technique to check.

3.8.0.3 Summary of this Approach

We have used the datasets we retrieved from our database and applied the various centrality measures to our datasets. This experiment aimed to find out the critical nodes and those of them that had a significant impact on the network. We utilized some of the ranks in investigating some of the nodes that our novel multimodal clustering algorithm used to find potential candidates.

3.9 Multimodal Clustering Approach (DReiM)

We introduce a novel approach to clustering phosphorylated network also known as DReiM(Drug Repositioning in Multimodal Networks). We utilized N-cliques, N-clans and N-plexes to decide the multimodal arrangement. We will discuss more what the multimodal arrangement involves in the next sections.

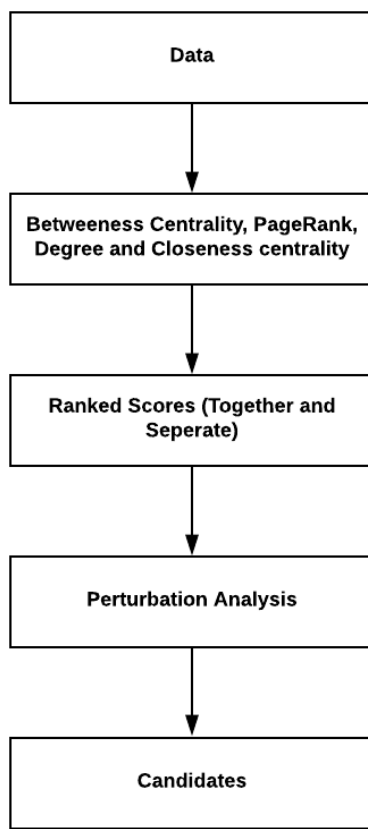


Figure 3.11: Network Analytics toward Drug Repositioning(NetAnaPhoS)

3.9.0.1 N-cliques, N-clans and N-plexes

In our experiment, we used the N-cliques, N-clans and N-plexes to select the best possible groups of disease related elements to determine our multimode setup.

3.9.0.2 N-cliques

It assumes that every member of a subgraph has a connection to one another. The use of the clique is considered a strict definition of the connectedness of vertices but some definitions have help relax this assumption to make the use of cliques

more relevant in general. An object is that called a member of a clique if they are also connected to other objects at a distance greater than 1 or the use of the path distances is applied. e.g., your friends friend.

3.9.0.3 N-clans

The n-clans helps us to add another layer of confirmation of the path distance to how we select our group objects. The n-clan is an extension of the N-cliques.

3.9.0.4 N-plexes

N-plexes help relax the assumption of a maximal complete class. It allows an object that does not have direct ties with the cliques to get added. For example, if the gene has ties with drug and SNP, but not disease; while both drug and SNP have ties with the disease, all four essential disease-related elements will be grouped under that clique using the N-Plex approach. The framework assumes that a vertex belongs to a clique of size t if they have direct connections with $t-k$ members of that clique.

3.9.0.5 Extraction of Datasets for the purpose of Multimodal Clustering

Each mode was extracted in bi-modes. The bi-modes are all diseases related elements with a specific property called or known as phosphorylation. This property serves as a filter that helps to focus on relevant aspects of the disease-related

elements.

3.9.0.6 Bi-mode and Multimode Network

In our approach, we considered our network a phosphorylated network. The phosphorylated network contains genes, diseases, pathways, go ontologies that are connected to a protein that phosphorylates. This phosphorylated network as a resemblance with social networks that include an actor and the relationships that exist between this actor such as the persons they relate with, the companies they work with, and the types of relationships [119]. In the social network, analysis attention is paid more to analysis in single mode [119], whereas this approach makes it difficult to find relationships between two entities. For example, an individual and an event and another example is a user and tag network. This can be easily modeled using bi-mode. Bi-mode creates an opportunity to combine two types of entities together. For example, in the phosphorylated network, the types of entities will be the pathway and drugs.



Figure 3.12: An example of bi-mode using pathway and drug

In the example above, the pathway is considered the source and the drug is regarded as the target. An example of the source to target relationship is an actor and friends relationship. With various other elements that play a role in the bodies function, new entities can be represented, and this will create a multimode phosphorylated network [119]. The multimode network can have 3-modes, 4-modes,

5-modes, and 6-modes. In all the modes we have the primary entities, and the other entities are called the secondary entities. Using this method helps to find relationships that are not typically accomplished by a conventional clustering algorithm. Given the link between a pathway and a drug, a pathway may not be found to relate to a drug in a bi-mode relationship, but they may be found to connect to a drug in a tri-mode relationship. This illustrates that a combination of bi-mode can reveal hidden structural similarity.

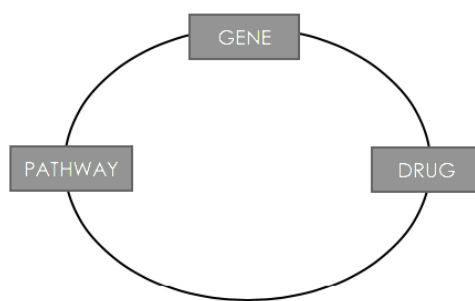


Figure 3.13: An example of tri-mode using gene, pathway and drug

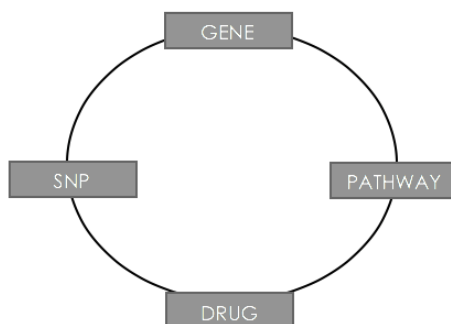


Figure 3.14: An example of quad-mode using snp, gene, pathway and drug

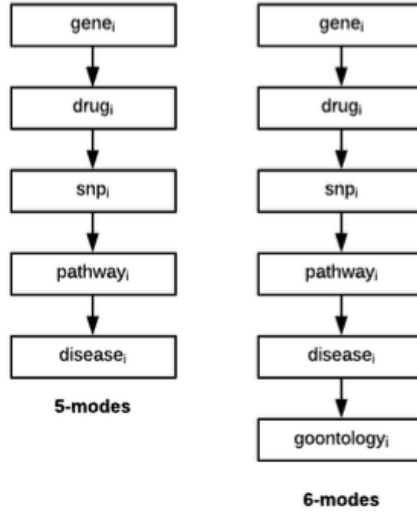


Figure 3.15: An example of 5 and 6 modes

3.9.0.7 Creating the bi-mode

The bi-mode involves two entities. The graph of a 2-mode is regarded as a partite network. All edges will be between pathways and drugs. The figure 3.16 shows an example of the bimode (pathway vs drug) . There are no edges between pathways and drugs.

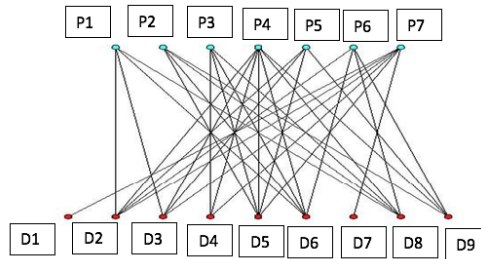


Figure 3.16: Bi-mode pathway and drug example

The adjacency matrix of the bi-mode: The adjacency matrix can be represented in a graph format. The graph $G = [N, E]$, which are a set of nodes

$N = 1, 2, 3, |N|$ and edges $E, i, j = 1, 2, 3, |E|$ that connect the nodes. We utilize this graph to generate our adjacency matrix. The adjacency matrix represents the occurrence of an interaction between two nodes. Each node is an entity. The two entities maybe for example pathway and drugs and there are no links between the entities of different types.

$$A = \begin{cases} 1, & \text{if } (i, j) \in E \\ 0, & \text{otherwise} \end{cases}$$

3.9.0.8 Creating the 3-mode

The multi-mode involves more than two entities. The multimode networks exist with entities in that are in the same or different modes [119]. For the multi-mode, we use the multiple modes to uncover a relationship that exists or associations that exist between the various entities involved. We have used the modes below to demonstrate these types of relationship. To illustrate the layout for 3-mode, we have an example below. We use a ring structure to illustrate this approach.

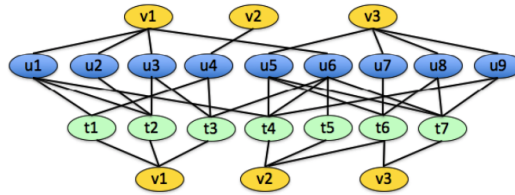


Figure 3.17: Three modes using letters v,u,t

$$A_1 = \begin{bmatrix} u_1 t_1 & u_1 t_2 & \cdots & u_1 t_7 \\ u_2 t_1 & \ddots & \cdots & u_2 t_7 \\ \vdots & \vdots & \ddots & \vdots \\ u_9 t_1 & u_9 t_2 & \cdots & u_9 t_7 \end{bmatrix}$$

$$A_2 = \begin{bmatrix} t_1 v_1 & t_1 v_2 & t_1 v_3 \\ \vdots & \ddots & \vdots \\ t_7 v_1 & t_7 v_2 & t_7 v_3 \end{bmatrix}$$

$$A_3 = \begin{bmatrix} u_1 v_1 & u_1 v_2 & u_1 v_3 \\ \vdots & \ddots & \vdots \\ u_9 v_1 & u_9 v_2 & u_9 v_3 \end{bmatrix}$$

$$L = \begin{bmatrix} D_1^{(9 \times 9)} & A_1^{(9 \times 7)} & A_3^{(9 \times 3)} \\ A_1^{T(7 \times 9)} & D_2^{(7 \times 7)} & A_2^{(7 \times 3)} \\ A_3^{T(3 \times 9)} & A_2^{T(3 \times 7)} & D_3^{(3 \times 3)} \end{bmatrix}$$

3.9.0.9 Creating the 4-mode

The form of the Laplacian Matrix L representation for 4-Modes is shown below. The (u,u) represents the first Diagonal of U which represent D_1 . The (t,t) represents the second diagonal of the matrix represent D_2 . The A_1 represents the (u,t) interactions with u x t type of matrix. The transpose will be the opposite of u x t matrix. This is continued until the complete L matrix is formed.

$$L = \begin{matrix} & \begin{matrix} u & t & v & d \end{matrix} \\ \begin{matrix} u \\ t \\ v \\ d \end{matrix} & \begin{pmatrix} D_1 & A_1 & X_{uv} & A_4 \\ A_1^T & D_2 & A_2 & X_{td} \\ X_{uv}^T & A_2^T & D_3 & A_3 \\ A_4^T & X_{td}^T & A_3^T & D_4 \end{pmatrix} \end{matrix}$$

3.9.0.10 Creating the 5-mode

The form of the Laplacian Matrix L representation for 5-Modes is shown below. The (u,u) represents the first Diagonal of U which represent D_1 . The (t,t) represents the second diagonal of the matrix represent D_2 . The A_1 represents the (u,t) interactions with u x t type of matrix. The transpose will be the opposite of u x t matrix. This is continued until the complete L matrix is formed.

$$L = \begin{matrix} & \begin{matrix} u & t & v & d & z \end{matrix} \\ \begin{matrix} u \\ t \\ v \\ d \\ z \end{matrix} & \begin{pmatrix} D_1 & A_1 & X_{uv} & X_{ud} & A_5 \\ A_1^T & D_2 & A_2 & X_{td} & X_{tz} \\ X_{uv}^T & A_2^T & D_3 & A_3 & X_{vz} \\ X_{ud}^T & X_{td}^T & A_3^T & D_4 & A_4 \\ A_5^T & X_{tz}^T & X_{vz}^T & A_4^T & D_5 \end{pmatrix} \end{matrix}$$

3.9.0.11 Our approach using eigenvalue decomposition((DReiM + eigenvalue decomposition)

Given the Laplacian matrix, the eigenvalue decomposition of this matrix produces eigenvalues and eigenvectors. We select the eigenvector with the second smallest eigenvalue also known as the Fiedler vector. The smallest eigenvalue is always 0 with the corresponding eigenvector having all the one entries.

3.9.0.12 Our approach using Spectral Clustering Approach(DReiM + Spectral Clustering)

The spectral clustering approach explores using Laplacian matrix [120] [119]. The Laplacian matrix requires finding the degree of the matrix D. The D will be on the diagonal. The diagonal is formed using

$$d_i = \sum_{j=1}^n w_{ij}. \text{ The Laplacian matrix } L = D - A.$$

The A is representing the adjacency matrix and D representing the degree matrix.

$$L_{i,j} \begin{cases} d_i, & i = j \\ -w_{ij}, & (i, j) \in E \\ 0, & \text{otherwise} \end{cases}$$

The properties of Laplacian matrix are [120] [119]:

1. They are symmetric positive semi-definite matrix.
2. It has a set of full real and orthogonal eigen vectors.

3. The eigenvalues of the Laplacian matrix are both real and non-negative.
4. 0 is an eigenvalue of L and $e = [1, \dots, 1]^T$ is its eigenvector.
5. The vector x contains the following properties: $x^t L x = \sum_{i=1}^n \sum_{j=1}^n w_{ij} (x_i - x_j)^2$

The formula above is used in graph partitioning to assign nodes to clusters refer to [120] [119] for more.

We applied the kmeans algorithm to find clusters using the multimodal formation. We can repeat the same for 4 Modes, 5 Modes, 6 Modes till 15 Modes.

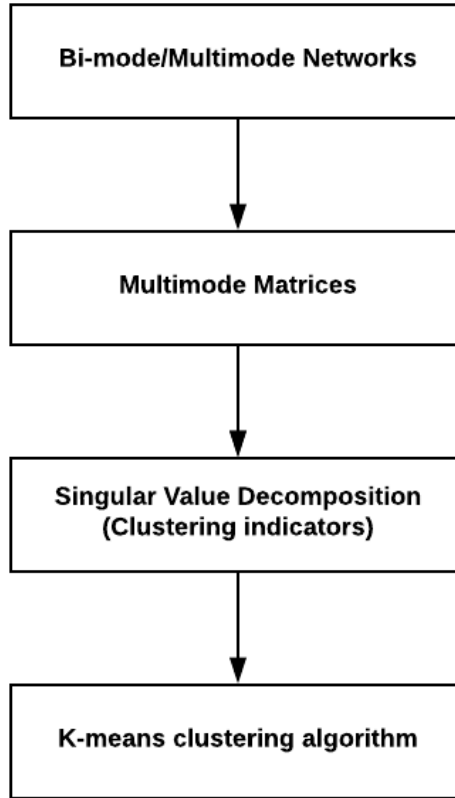


Figure 3.18: Multimode Clustering Architecture

3.9.0.13 Time Complexity of the system

The time complexity of performing multiple k -means clustering to find the ideal- k is shown in Equation 3.5 where, n is the size of the matrix L , t is the number of iterations performed in k -means clustering, i is the number of data points of $U(:, 1 : r)$ (i.e. instances), r is the rank of matrix L and c is the maximum number of clusters.

$$\mathcal{O}(n^2 + 2n^3 + tirc!) \quad (3.4)$$

Input: Maximum number of Clusters, 4 Different two-columned csv files
Output: The graph of wss across k to find Ideal- k

```

Read and Store two-columned csv files into matrices  $ut, tv, vd$  and  $ud$ 
Set  $u$  as the unique elements of  $ut(:,1)$  and  $ud(:,1)$ 
Set  $t$  as the unique elements of  $tv(:,1)$  and  $ut(:,2)$ 
Set  $v$  as the unique elements of  $vd(:,1)$  and  $tv(:,2)$ 
Set  $d$  as the unique elements of  $vd(:,2)$  and  $ud(:,2)$ 
for n=1 to 4 do
    for i = 1 to number of rows of  $I$  do
        Assign 1 to respective row and column of  $A_n(I(i,1), I(i,2))$ 
    end for
end for
for n=1 to 4 do
    for i = 1 to size of  $E$  do
         $D_n(i, i) = \text{sum of } i\text{-th row of } A_n$ 
    end for
end for
Set non-interacting matrices  $X_{uv}$  and  $x_{td}$  to zero matrices
for all matrices  $A_1, A_2, A_3, A_4, X_{uv}$  and  $x_{td}$  do
    Compute its transpose matrices
end for
Concatenate all  $D_1, D_2, D_3, D_4, A_1, A_2, A_3, A_4, X_{uv}$  and  $x_{td}$  and its corresponding transpose matrices to form Matrix  $L$ 
 $U = \text{svd}(L)$ 
 $r = \text{rank}(L)$ 
for k=1 to maximum number of clusters do
     $\text{sumDist} = \text{kmeans}(U(:, 1:r), k)$ 
     $wss(k) = \text{sumsq}(sumDist)$ 
end for
Plot graph of  $wss$  across each value of  $k$ 

```

Figure 3.19: Time Complexity for 4 Modes

Input: Maximum number of Clusters, 4 Different two-columned csv files
Output: The graph of wss across k to find Ideal- k

```

Read and Store two-columned csv files into matrices  $ut, tv, vd, dz$  and  $uz$ 
Set  $u$  as the unique elements of  $ut(:,1)$  and  $uz(:,1)$ 
Set  $t$  as the unique elements of  $tv(:,1)$  and  $ut(:,2)$ 
Set  $v$  as the unique elements of  $vd(:,1)$  and  $tv(:,2)$ 
Set  $d$  as the unique elements of  $dz(:,1)$  and  $vd(:,2)$ 
Set  $z$  as the unique elements of  $dz(:,2)$  and  $uz(:,2)$ 
for n=1 to 5 do                                ▷ Compute Term-Document Matrices  $A_1, A_2, A_3, A_4$  and  $A_5$ 
    for  $i = 1$  to number of rows of  $I$  do        ▷ where  $I$  is  $ut, tv, vd, dz$  or  $uz$  accordingly
        Assign 1 to respective row and column of  $A_n(I(i, 1), I(i, 2))$ 
    end for
end for
for n=1 to 5 do                                ▷ Compute Degree Matrix  $D_1, D_2, D_3, D_4$  and  $D_5$ 
    for  $i = 1$  to size of  $E$  do                  ▷ where  $E$  is  $u, t, v, d$  or  $z$  accordingly
         $D_n(i, i) = \text{sum of } i\text{-th row of } A_n$ 
    end for
end for
Set non-interacting matrices  $X_{uv}, X_{ud}, X_{td}, X_{tz}$  and  $X_{vz}$  to zero matrices
for all matrices  $A_1, A_2, A_3, A_4, A_5, X_{uv}, X_{ud}, X_{td}, X_{tz}$  and  $X_{vz}$  do
    Compute its transpose matrices
end for
Concatenate all  $D_1, D_2, D_3, D_4, D_5, A_1, A_2, A_3, A_4, A_5, X_{uv}, X_{ud}, X_{td}, X_{tz}, X_{vz}$  and its corresponding transpose
matrices to form Matrix  $L$ 
 $U = \text{svd}(L)$                                 ▷ Perform SVD on Matrix  $L$ 
 $r = \text{rank}(L)$                                 ▷ Compute rank of Matrix  $L$ 
for k=1 to maximum number of clusters do
    sumDist = kmeans( $U(:, 1:r), k$ )            ▷ Perform k-means using  $k$ -clusters
    wss(k) = sumsq(sumDist)                    ▷ Compute  $wss$  of each  $k$ -clusters
end for
Plot graph of  $wss$  of across each value of  $k$     ▷ Determine ideal- $K$  using the Elbow method

```

Figure 3.20: Time Complexity for 4 Modes

3.10 Validation of our experiment

In this dissertation, we validate our approach using the current clinical trial datasets and PubMed datasets. We compare the results we found in our experiments to the results in the clinical trial datasets.

3.10.0.1 Pubmed

The database contains over 28 million citations of biomedical literature from journals and books. They also include contents from PubMed central and other publishers. We found drug and disease relationship.

3.10.0.2 Clinical Trials

The clinical trials datasets contain research that has been used on people with the goal of measuring a medical or behavioral intervention. The datasets help researchers locate other new treatments like drugs for the safety of the patients. The clinical trial itself is in four phases. The first phase is the first safety trial of the drugs. It is in this stage that the dose range is established and usually tested on critically ill patients. The second phase checks how effective the medication is across a selected population of patients with the disease being treated. The third phase is an extension or continuation of the second phase with further investigation of safety with a relatively large population. The fourth phase is a post-marketing surveillance phase what happens after the product is marketed. We used this dataset to evaluate our results. We considered drugs in Phase I, Phase II and Phase III.

We considered drugs currently in clinical trials phase I, II, III a candidate since they have been selected by top pharmaceutical to go on the trial. We used clinical trial data sets to determine the validity of our results.

The real positives (TP) are cases in which we predicted yes (these drugs have been found to treat the disease using the clinical trials database), and these drugs treat cancer(prostate cancer or rhabdomyosarcoma) or are on clinical trials to treat prostate cancer or rhabdomyosarcoma. The real negatives (TN): We predicted no, and they don't or can't be used to treat both cancers. The false positives (FP) are drugs that we predicted yes, but can't be used to treat prostate cancer. The false negatives (FN) are drugs that we predicted no (they can't be used to treat any of

the tumors), but they can be used for prostate cancer or rhabdomyosarcoma.

3.10.0.3 Precision

The precision is widely used in classification. The precision measures the exactness of our results. It checks how often the method predicts the correct answer.

$$precision = \frac{tp}{tp + fp} \quad (3.5)$$

3.10.0.4 Recall

The recall checks for the completeness of the results. The number of positive results was labeled positive. The recall represents the rate of true positive or sensitivity score.

$$recall = \frac{tp}{tp + fn} \quad (3.6)$$

3.10.0.5 Accuracy

The accuracy shows how well our results classified correctly in the identification and otherwise.

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (3.7)$$

The F-measure is used to combine both precision and recall.

$$accuracy = \frac{2 * precision * recall}{precision + recall} \quad (3.8)$$

3.11 Future Work: Gene expression data analysis

We introduce the method to some of the work we intend to do in the future. We intend to use gene expression level datasets to find possible drug candidates and drug combination candidates.

3.11.0.1 Gene expression level analysis on k562 cells(GEiM)

3.11.0.2 Introduction

The cell lines used to find out the PIM Substrates is known as K562 cell lines. Lozzio and Lozzio established the K562 cell lines from the pleural effusion of a 53-year-old female who had chronic myelogenous leukemia.

We extracted cancer cell line from cancer cell line encyclopedia (CCLE) and gene expression data from GEO omnibus. The CCLE provides public access analysis and visualization of DNA copy number, mRNA expression, mutation data and more, for 1000 cancer cell lines. K562 cell lines were retrieved from the CCLE to find gene expressions levels and mutations from the K562 cell lines. Xena Python was used to extract gene expression information from cancer cell line encyclopedia from K562 cell lines. Other information contained within the cancer cell encyclopedia includes copy number, gene level, probe level, and phenotype.

Furthermore, CREB regulates a bunch of cellular functions and this includes

cell proliferation, survival and apoptosis [121]. CREB is also identified as a proto-oncogene involving in transformation by promoting abnormal proliferation and survival of myeloid cells [121]. CREB is activated through phosphorylation at serine 133 in response to a variety of cellular and mitogen stress signals. These include peptide hormones, neurotransmitters, calcium influx, and growth factors. CREB is a novel target for prostate cancer [122].

The purpose of this study is to understand the mechanism of CREB in CML to find possible drug candidates. CREB is relevant for regulating cellular function which includes cell survival, cell apoptosis, metabolism and so forth. When they are altered can become an oncogene. This study aims at identifying the mechanism of CREB in chronic myeloid leukemia. CREB is a member of the family of activation transcription factor 1 that binds to cAMP which is an octanucleotide cAMP response element consensus sequence in promoters of target genes. We then use our understanding of the mechanism of CREB in CML to arrive at drug candidates that treat a disease like prostate cancer.

Our framework to arrive drug candidate is below:

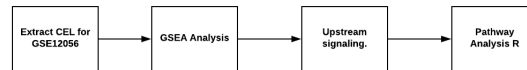


Figure 3.21: Gene expression Framework for PIM Substrate Analysis

3.11.0.3 Extraction of GSE12056

The CEL files were extracted from NCBI and parsed into the gene pattern tool to create the .gct files. The .gct files were further utilized and alongside with cls file (contains the class files for the data in .gct files). The result was used to perform the GSEA Analysis. The method checks a priori defined set of genes shows statistically significant, consistent differences between two biological states (e.g., phenotypes).

3.11.0.4 GSEA Analysis

The difficulty of finding out changes in related gene expression is what leads to GSEA Analysis. Gene expressions are complicated because we have a large amount of them, high erraticism with the samples, and a sometimes inadequate amount of samples. GSEA Analysis helps accomplish the task. GSEA also helps us to find these changes using gene sets related to one another. The process involved calculating an enrichment score. The genes are initially ranked based on the difference in expression levels. The sum is calculated across the ranked genes. This score sum is also referred to as the cumulative sum. The sum grows when the gene is contained in the set and decreases. The degree to which it increases will depend on how related the genes are with the phenotype. The most substantial deviation from zero gives the enrichment score [123] [124].

The phenotype labels were permuted 1000 times and the enrichment score calculated for each of the permutations [123] [124]. The enrichment scores from the actual were compared to those of the permuted data (See [123]). The final step is

that adjustments are made to accommodate multiple hypothesis testing [123].

GSEA helps to find out if the gene sets are distributed randomly or found on the top or bottom of L. Those that are related to the phenotype show top or bottom.

The .cel files containing the gene expression data was uploaded into GSEA, and the class file containing the control and the non-control version was uploaded. The gene set database (MSigDB reference) was selected, and we utilized the transcription factor gene sets so that we can identify the transcription factors that play a significant role and further use this transcription factors for pathway analysis. We permuted 1000 times to get a more accurate p-value. The chip platform utilized was Affymetrix. [123] shed more light on how the gene expression is used to find the up-regulated and down-regulated genes, which will be used to do a pathway analysis.

3.11.0.5 Pathway Analysis

The results from the gene sets were further extracted to find specific genes that belong to this genesets. The genes were then used to perform pathway analysis. The Reactome package in R was used to perform a pathway analysis using the Reactome pathway database. This implements the gene set enrichment analysis, hypergeometric test and enrichment analysis. Our results from the pathway analysis were used to extract drugs that target these pathways. The enrichment analysis

3.11.0.6 Identification of Drug Candidate

The results from Pathway analysis will further help us to find a drug that targets these pathways. We utilized the results from the pathway analysis to find drugs that are relevant to the paths that were significant. See more in the result section. We identified pathways greater than 0.05 p-values and used our bi-mode method to find relevant drug clusters and pathway clusters. We also applied SVD in finding latent structure in a single mode and bi-mode.

Chapter 4: Results

In this chapter, we present the experimental results. The section demonstrates the effectiveness of our approach on the phosphorylated network or proteins (pim substrate and rhabdomyosarcoma). The datasets sources were presented in this section alongside the result using our method section was performed.

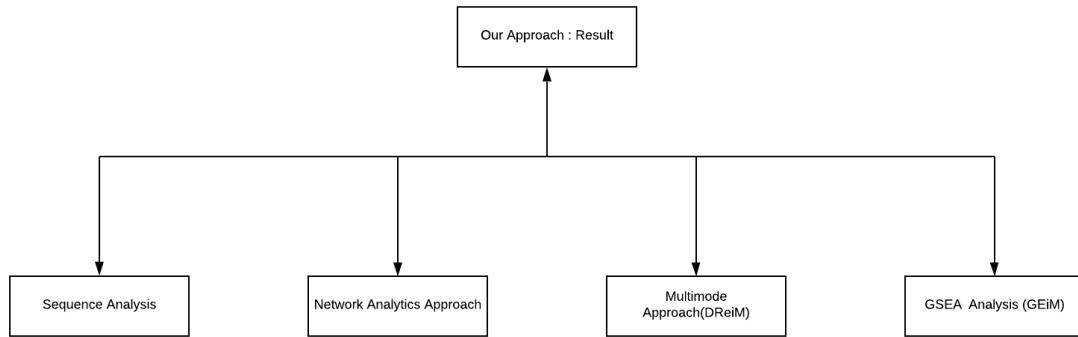


Figure 4.1: Important aspects of our work

The figure 4.1 describes the steps implemented in our framework. This section will start illustrating results starting with sequence analysis, network analytics approach, multimode approach and address future work like GEiM analysis approach.

4.1 Experimental Data - Sequences

We have specifically highlighted how our datasets are pulled in the Methodology section. We explain in this section the datasets used for clustering sequences.

4.1.1 Experimental Data for PIM Substrates

The data sets contained 1269 PIM substrates that have been collected based on the experiment that was performed in the Lab. We utilized the PIM substrates in our research to identify relevant phosphorylated proteins.

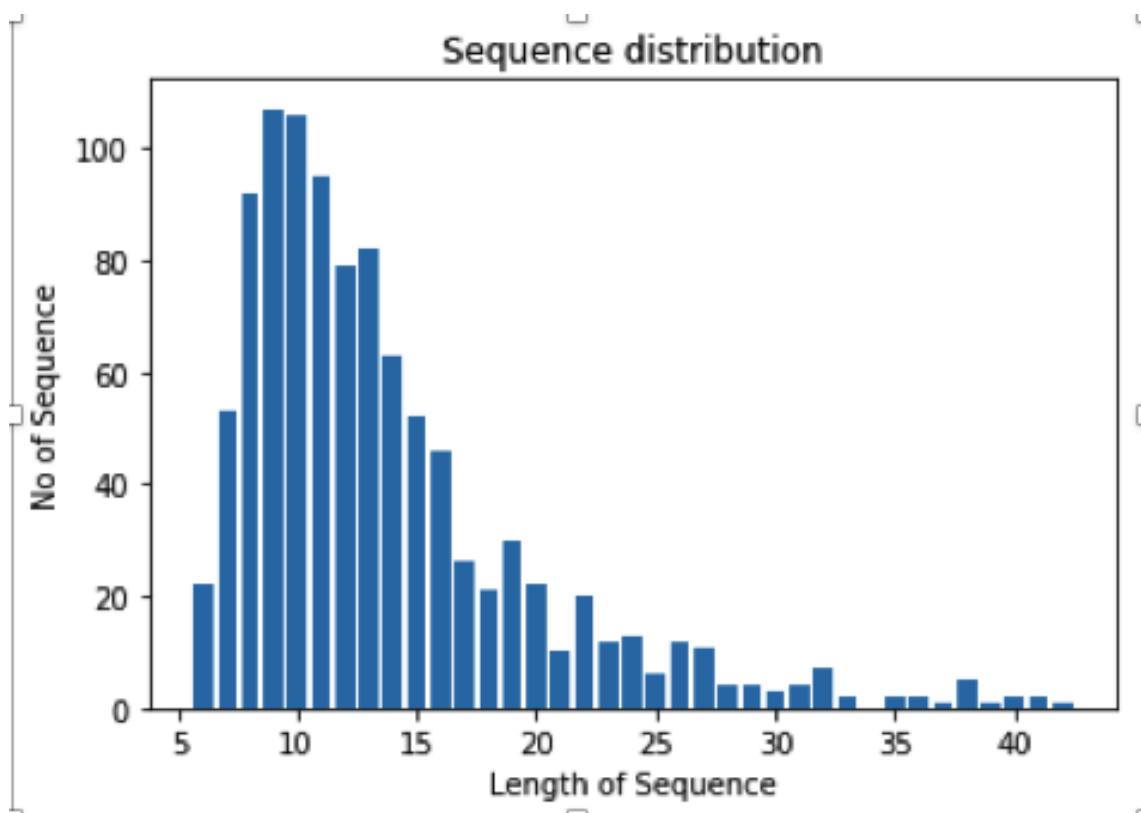


Figure 4.2: PIM Substrate Sequence Analysis.

The distributions help us to analyze the sequence for each PIM Substrate better (see figure 4.2). The length of the sequence is on the Y and number of the

sequences is on the X-axis. 100 sequence contained the length of 10. The database consists of sequences between the range of 6-40 sequences. The sequences were cleaned to remove duplicate sequences and to reduce redundancy. The sequence was processed for analysis. A majority of the sequences fall between 6 and 15. A higher sequence length the lower or, the less number of sequences exist. We used all of the sequences, but they had to undergo a slicing process. The next step explains the sequence processing step results.

4.1.2 Experimental Data for Rhabdomyosarcoma

We directly extrapolated the sequences we used to perform analysis on rhabdomyosarcoma from the phosphositeplus database. The database contained the protein accession number, the diseases this specific accession number was connected to and the raw sequences with the phosphorylation site. We extracted the entire sequence of each of the protein using the protein accession number. We used this sequence to find the ten up and downstream sequence for rhabdomyosarcoma.

4.2 Sequence Processing

The variable modification site indicates the site or location in the sequence where the phosphorylation takes place. At this site (phosphorylation site), we count ten sequences up and ten sequences down. The total length of sequences extracted at the end of the processing is 21. The 1269 sequences have 21 sequences preserved. We removed every sequence with less than 21 sequences and removed every duplicate sequences. The table 4.1 shows the view of the sequences after removing the duplicates and cleaning out the sequences. We followed the same process for sequence length = 15, 20, 30, and 40.

1433BHUMAN	ARRSSWRVISSIEQKTERNE
1433BHUMAN	MKGDYFRYLSEVASGDNKQT
1433EHUMAN	ARRASWRIISSIEQKEENKG
1433FHUMAN	ARRSSWRVISSIEQKTMADG
1433GHUMAN	ARRSSWRVISSIEQKTSADG

Table 4.1: PIM Substrate Sequence with length 20

4.2.1 Sequence Extraction

The table below shows the numbers of protein substrates left after the extraction process was applied to each of the sequences. The observation is the number of substrates reduce as the processing takes because the sequence was cut and the irrelevant proteins were removed. The sequence of length 10 with the largest amount of substrates.

Sequence length	Sequence number
10seq	791
15seq	741
20seq	710
30seq	655
40seq	599
50seq	559

Table 4.2: Sequence Length vs Number of Sequences

4.3 Identification of the Consensus Sequence of PIM Substrates

The consensus sequence results produce the most frequently occurring amino acids based on all the PIM substrate sequences. The figure below shows the consensus sequence for 10, 20, 30, 40, 50 sequences up and downstream [45]. The results in Table 4.3 and figure 4.5 shows that amino sequence character R occurs also very often in the PIM datasets. The S and T represent the phosphorylation site or phosphorylated position in the sequence. The ten up and downstream sequence contains eight counts of R, nine counts of S, two counts of G, and one count of E.

4.3.1 Consensus Sequence Analysis of PIM Substrates

The table below shows the sequences for 10, 15, 20, 30, 40, consensus PIM substrates after trimming was performed on the data. S is the most frequent alphabet because that was the position we selected to perform the trimming exercise.

The figures 4.7 , 4.8, 4.9, and 4.10 below show the graphical representation of the sequences in the 10, 20, 30, 40, 50 up and downstream sequences. The figure 4.7

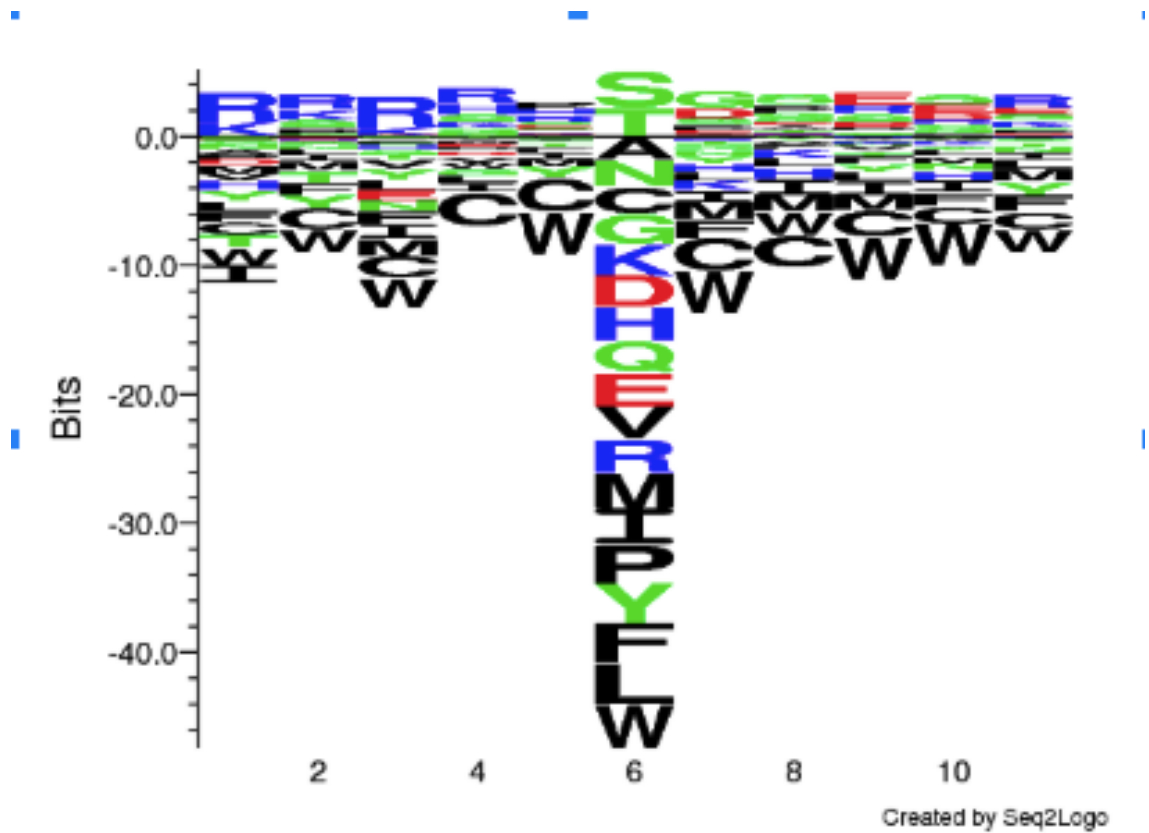


Figure 4.3: 5 Consensus Sequence for the PIM Substrate.

10 up and downstream	RRRRRSRSSSGSESRSRSGS
15 up and downstream	KSKKSKRRKR SRKRS GSESGSKSGKKARKR
20 up and downstream	GSKKAKSKKGKRRKR SRKRS GSESGSKSG KKARKREKKSS
30 up and downstream	AASSKRGSKKGKKKKKSKKGKRRRRSRKR SGSESSSKSGKKASKRKGSSSKRSSKRKSSK
40 up and downstream	EKSKSAAAKASRSSSRKGKSSRSKSSSKSS KRRRRRSRSSSGSSSSSSSGSKASKRELSKKS RSSSRSSSKSKKKSRESSE

Table 4.3: The consensus sequence: 10,20,30,40,50

contains the result for 10 up and downstream showing 9 S sequences 8 R sequences.

The most frequent motif in the sequences was SGSES.

The figure 4.8 shows the result for 20 up and downstream showing 10 S se-

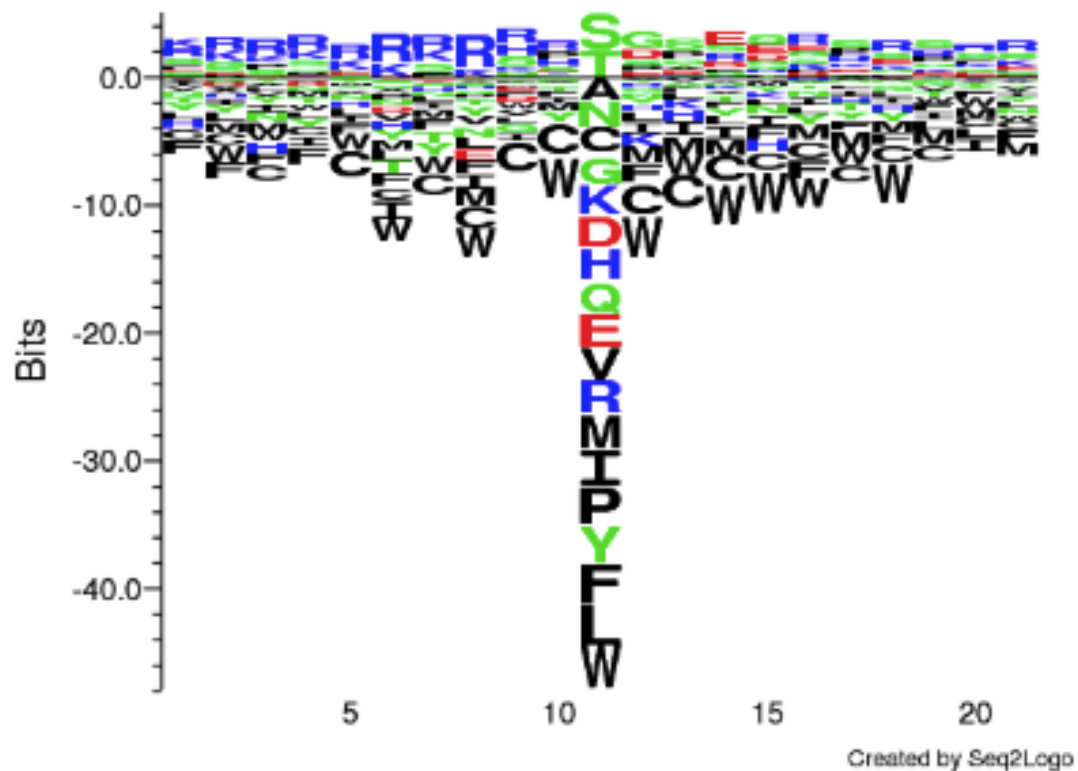


Figure 4.4: 10 Consensus Sequence for the PIM Substrate.

quences, 7 R sequences, A 2 sequences, 2 E sequences, 5 G sequences, and 14 K sequences.

The figure 4.9 shows the result for 30 up and downstream showing 39 S sequences, 13 R sequences, A 5 sequences, 5 E sequences, 3 G sequences, 16 K sequences, and 1 L sequence.

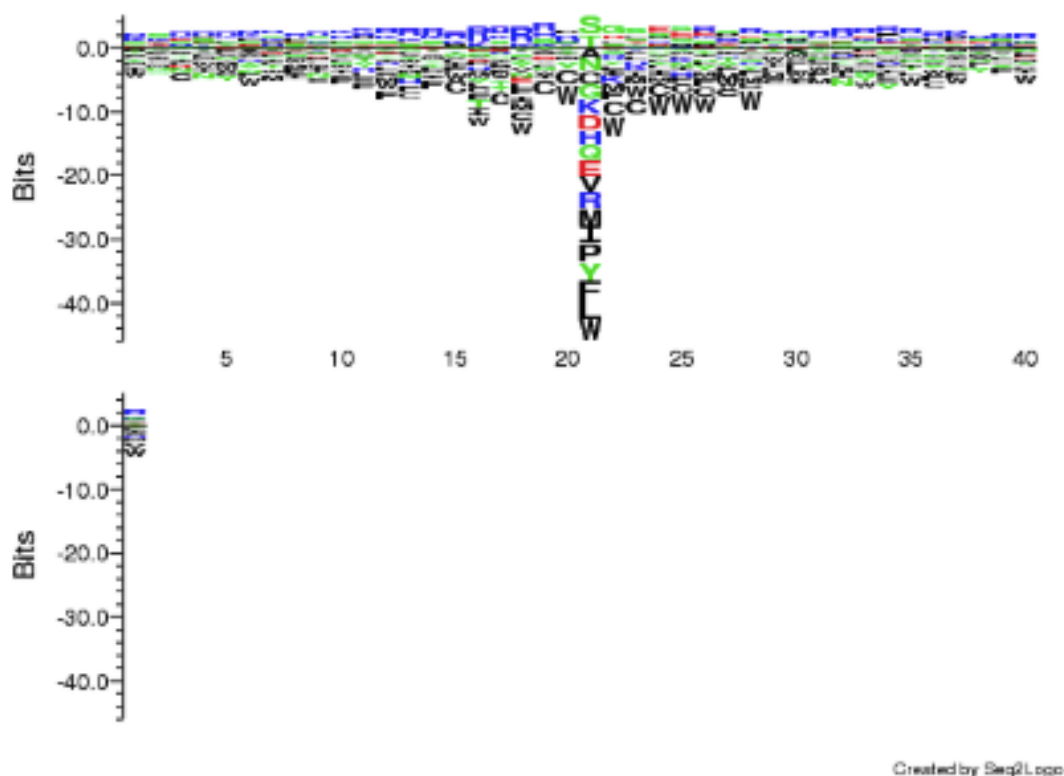


Figure 4.5: 15 Consensus Sequence for the PIM Substrate.

4.4 Sequence Encoding to identify Relevant Proteins

We encoded each of the amino acid sequence using the pim substrate using the derived encoding scheme 3.3. The phosphorylated alphabets were split by the alphabets that characterized the particular category they belonged. The result of encoding these sequence was a sparse matrix. We applied hierarchical clustering to the encoded sequence of 10, 15, 20, 30, 40 PIM Substrates.

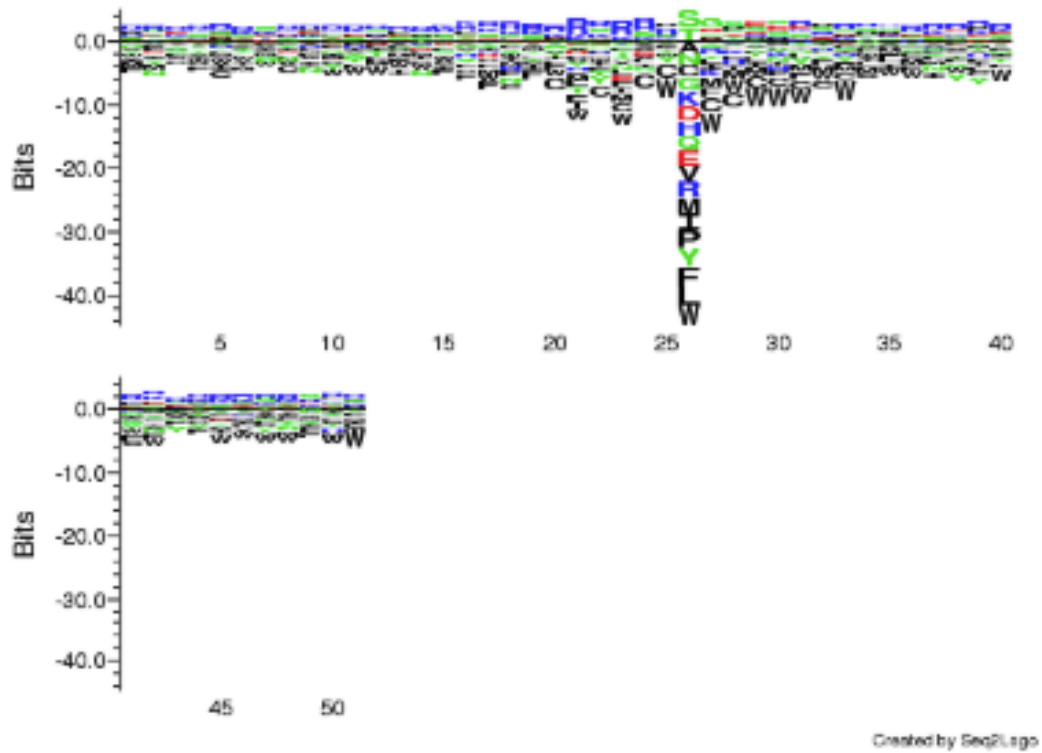


Figure 4.6: 25 Consensus Sequence for the PIM Substrate.

4.5 Hamming Distance Calculations without applying encoding schemes using standard alphabets

The Hamming Distance was used to calculate the distances between the sequences in a square matrix form. The distances were used to find hierarchical clusters. We utilized intercluster and intraccluster differences. The results from sequences of length 10,15,20 are presented in this section.

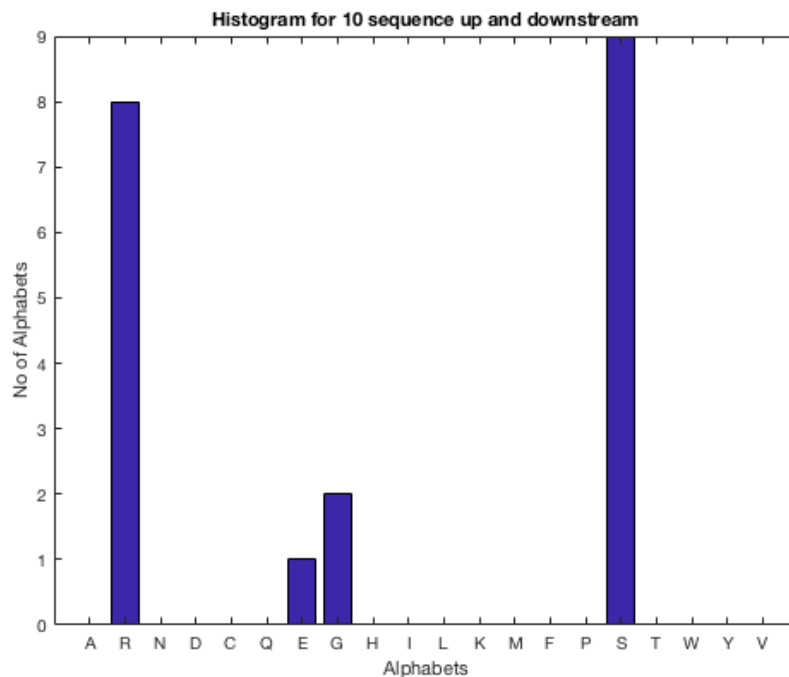


Figure 4.7: Histogram for 10 consensus sequence up and downstream.

4.5.1 Using the Hamming Distance to cluster sequences of length 10

We first of all cluster without using any encoding scheme. The distance between the sequences was calculated to find hierarchical clustering using the hamming distance. The hamming distance help finds a similarity between two sequences by taking the position and length into consideration. The complete linkage method was utilized to detect the clusters. We have applied hamming distance to sequences of length 10,20,30,40, and 50. For the sequence of length 10, $k=3$ clusters and cluster 3 contains 601 protein sequences, and cluster 2 contains 128 protein sequences, and cluster 1 contains 62 protein sequences. The Intra-cluster distance for Cluster 1 is 14.91363163371488, Intra-cluster distance for Cluster 2 is 18.373291015625, and

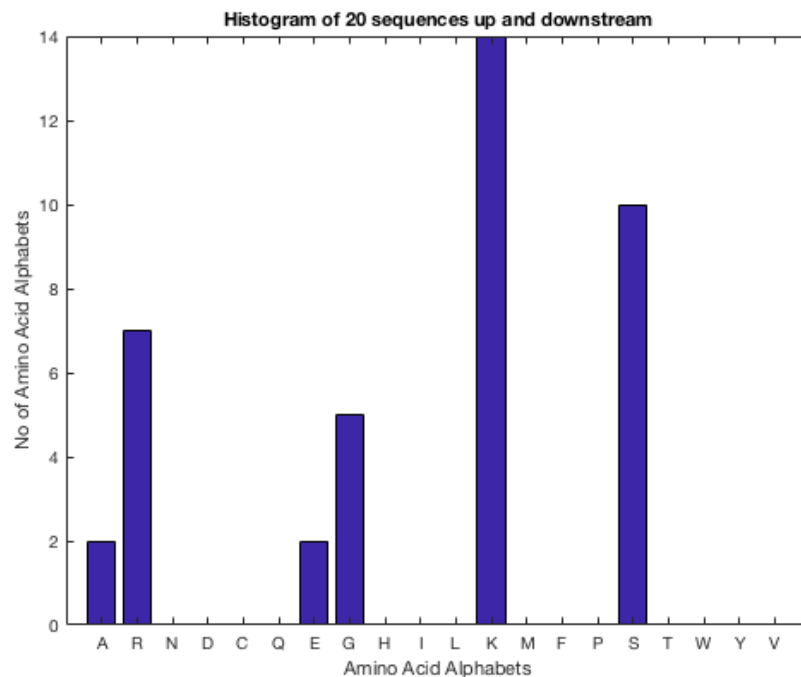


Figure 4.8: Histogram for 20 consensus sequence up and downstream.

Intra-cluster distance for Cluster 3 is 18.6053859208584.

Intercluster/Intracluster similarities	0	1	2
0	14.9136316	16.6434613	16.7595088
1	16.6434613	18.373291	18.4893385
2	16.7595088	18.4893385	18.6053859

Table 4.4: Inter-clusters similarity using sequence of length 10

4.5.1.1 Using the Hamming Distance to cluster sequences of length 15

Using $k=3$, the solution below shows cluster 3 as 459 protein sequences, cluster 2 as 194 protein sequences and cluster 1 as 88 protein sequences. The intra-cluster

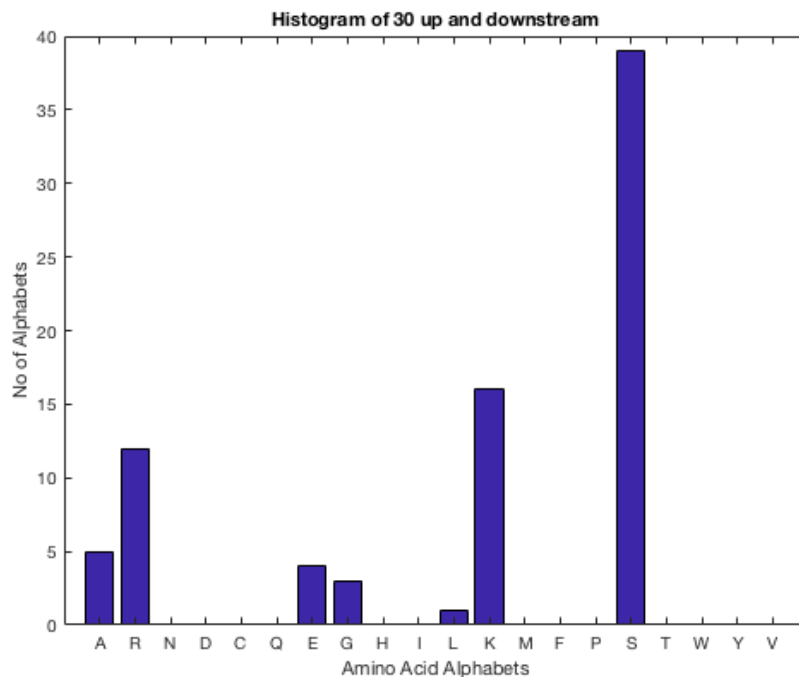


Figure 4.9: Histogram for 30 consensus sequence up and downstream.

distance for Cluster 1 is 24.669, and Cluster 2 is 27.724 and Cluster 3 is 27.749. The inter-cluster distances are on Table 4.5. The intra-cluster distances implied that the sequences in the each of the clusters remained n very closely

Intercluster similarities	0	1	2
0	24.6686467	26.1965829	26.2091341
1	26.1965829	27.7245191	27.7370703
2	26.2091341	27.7370703	27.7496215

Table 4.5: Inter-clusters similarity using sequence of length 15

4.5.2 Using the Hamming Distance to cluster sequences of length 20

Using $k=3$, the solution below shows cluster 3 as 644 protein sequences, cluster 2 as ten protein sequences and cluster 1 as 56 protein sequences. The intracluster

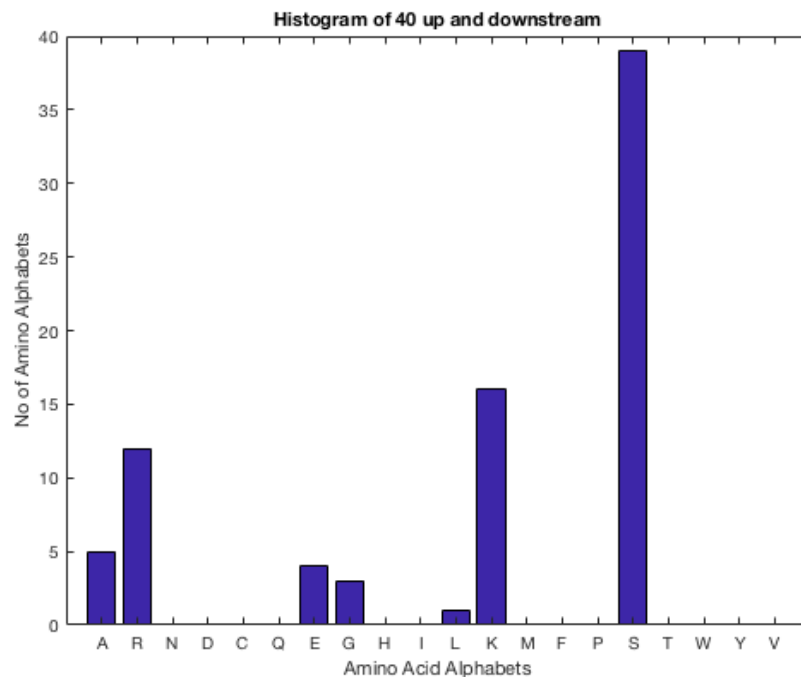


Figure 4.10: Histogram for 40 consensus sequence up and downstream.

distance for Cluster 1 is 32.03, and Cluster 2 is 20.98 and Cluster 3 is 37.535. The inter-cluster distances are on Table 4.6.

Intercluster similarities	0	1	2
0	32.0325255	26.5062628	34.7837004
1	26.5062628	20.98	29.2574376
2	34.7837004	29.2574376	37.5348752

Table 4.6: Inter-clusters similarity using sequence of length 20

We observe very close scores with regard to our well separated and how close the substrates are from each other. We proceed to use phosphorylated alphabets for sequence analysis using the SVD.

4.6 Using Singular Value Decomposition with Phosphorylated Alphabet Encoding scheme - Example for clustering

The singular value decomposition didn't produce excellent performance, but we just wanted to show and compare this performance to the others. For example, the total number of sequences left after cleaning ten up and downstream substrates was 791 sequence. The sequences were encoded using the encoding schemes discussed in the method section 3.3. The singular value decomposition was applied to the encoded sequence, and we selected the best U singular vectors. The 15, 20, 30, 40 sequences were also cleaned and Table 3.3 was applied on the sequences after cleaning. We applied the Singular Value Decomposition technique on this matrix to find Singular values. The SVD was applied in other to find clusters in our PIM Substrate data. The k-means clustering was applied to the datasets to find clusters. The performance using SVD wasn't the best. We will learn more about the results as we moved forward to with other methods.

4.6.1 Finding the Ideal-k using encoding schemes with ten sequences

The substrate sequences of length ten were used to identify the relevant protein sequences. The ideal-k was first identified for using functional alphabets, charged alphabets, standard alphabets, and structural alphabet figure. The ideal-k uses within cluster sums of squares or average silhouette score to find the most optimal k. The most optimal-k was 3. We utilized k=3 during our clustering process to

determine the classes of the pim substrates.

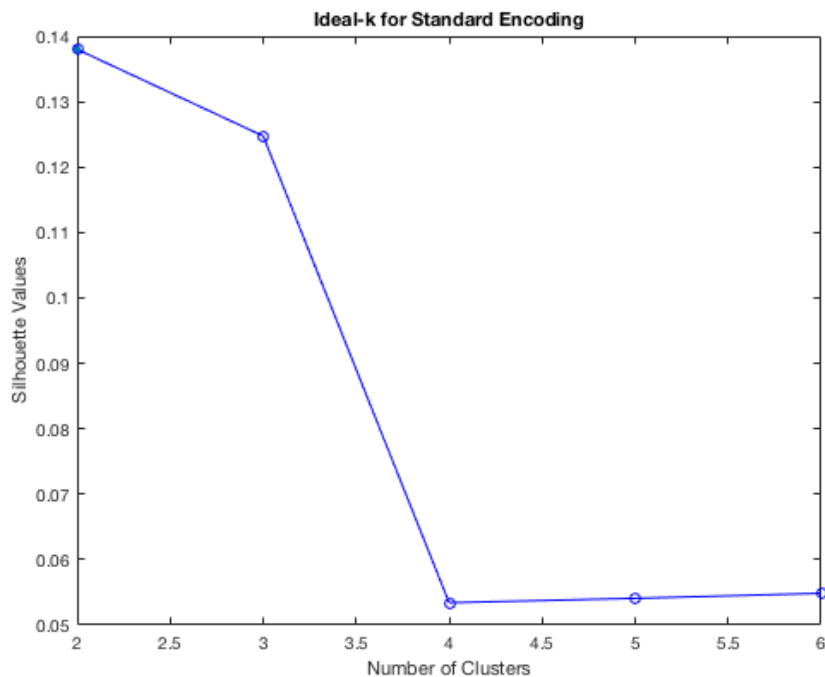


Figure 4.11: Ideal-k for Standard Alphabets

For a majority of the ideal-k's the most optimal-k =3. We chose three and moved into the next step which was clustering with kMeans.

4.6.2 Clustering using k-means

K-means algorithm is used to group the singular vectors in distinct clusters. The clusters are then used to verify the significance of the substrate.

The k-means algorithm in [125] begins with a random set of k center-points (μ). For each updated step, all points y are assigned to their nearest center-point (see equation 4.1).

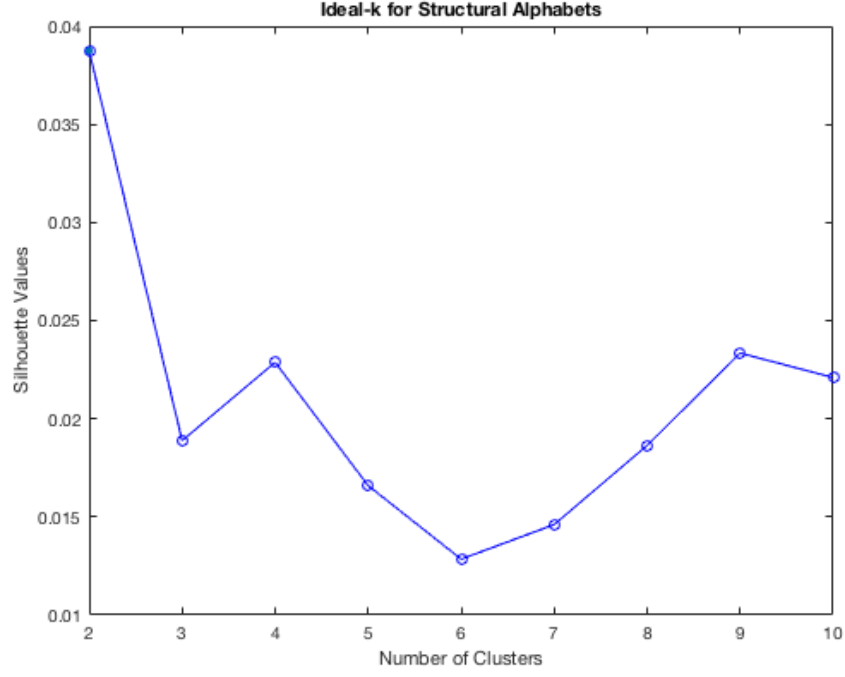


Figure 4.12: Ideal-k for Structural Alphabets

$$X_i^{(t)} = \{y_p : \|y_p - \mu_i^{(t)}\|^2 \leq \|y_p - \mu_j^{(t)}\|^2 \forall j, 1 \leq j \leq k\} \quad (4.1)$$

The center-points of the PIM Substrate get changed by calculating the mean of the points connected to a specific center-point. (see 4.2).

$$\mu_i^{(t+1)} = \frac{1}{|X_i^{(t)}|} \sum_{y_j \in X_i^{(t)}} y_j \quad (4.2)$$

The process continues until all points observed continue to exist at the center-points assigned to them, then the algorithm stops.

The k-means was applied to the PIM Substrates and below shows our clustering results.

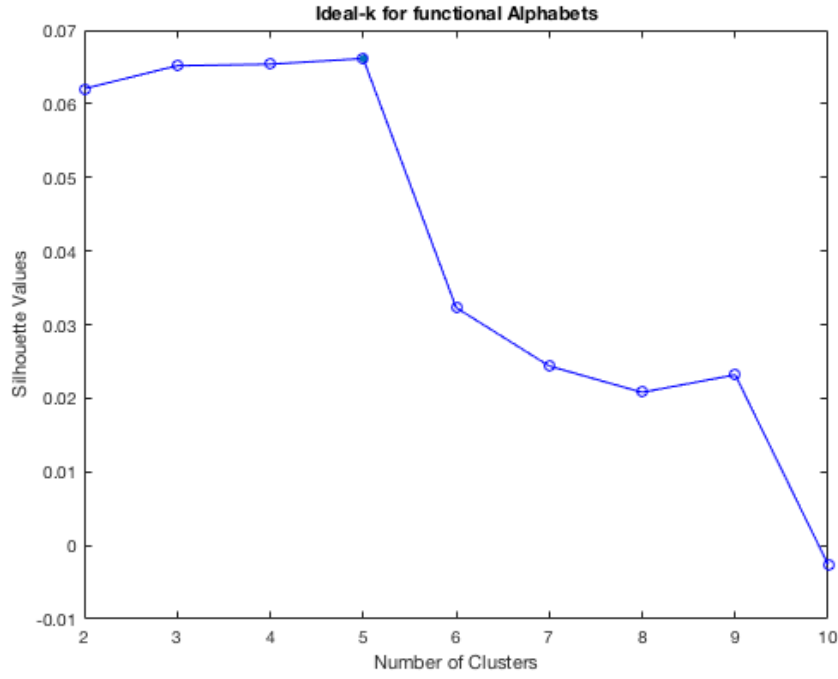


Figure 4.13: Ideal-k for Functional Alphabets

The figure 4.15 scatterplot shows the first two singular vectors. We did not cluster this datasets. The figure 4.16 and the figure 4.17 represents the PIM Substrates with two clusters and three clusters.

4.6.2.1 Sum Squared Error

The Sum of Squared Error(SSE) represents the distance to the nearest PIM Substrate cluster.

$$SSE = \sum_{i=1}^K \sum_{x \in C_i} (m_i, x) \quad (4.3)$$

The x represents the PIM substrates data point in cluster C_i and m_i is the point that was identified by C_i . The m_i represents the mean of the cluster. The

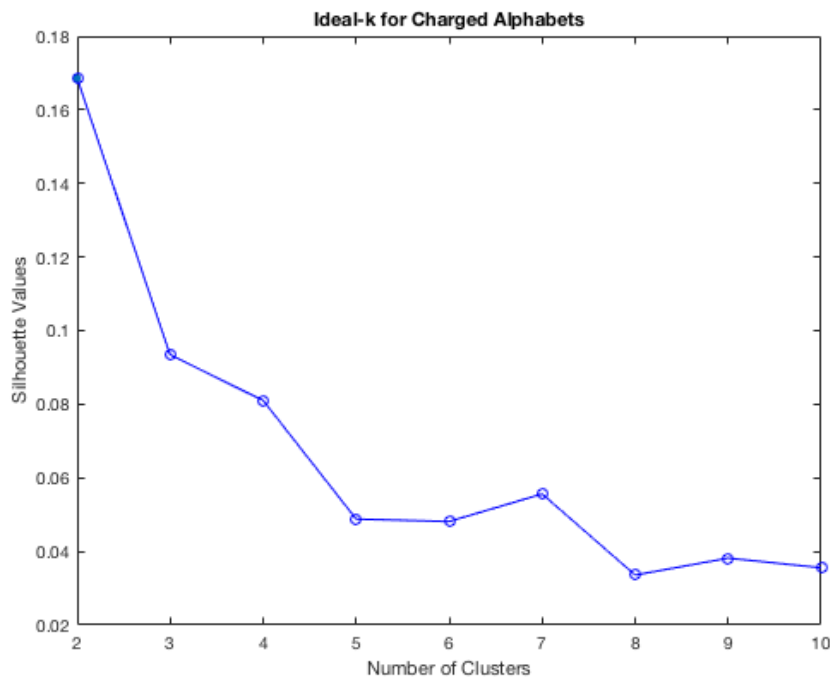


Figure 4.14: Ideal-k for Charged Alphabets

lower the SSE, then the more similar PIM Substrates that belong to that cluster. The SSE was used to select the best clusters and to further evaluate our results figure 4.15, the figure 4.16, the figure 4.17. The result below shows the clusters. The figure 4.17 cluster 2 as the lowest SSE.

The SSE was used to select the best clusters for the 15 up and downstream and to further evaluate our results figure 4.19, figure 4.20 and figure 4.21. The result below shows the clusters. The figure 4.21 cluster 3 as the lowest SSE and we can further study the cluster results to see how well the pim substrates are classified correctly using cluster $k=3$.

The SSE was used to select the best clusters for the 20 up and downstream and to further evaluate our results figure 4.22, figure 4.23 and figure 4.24. The result

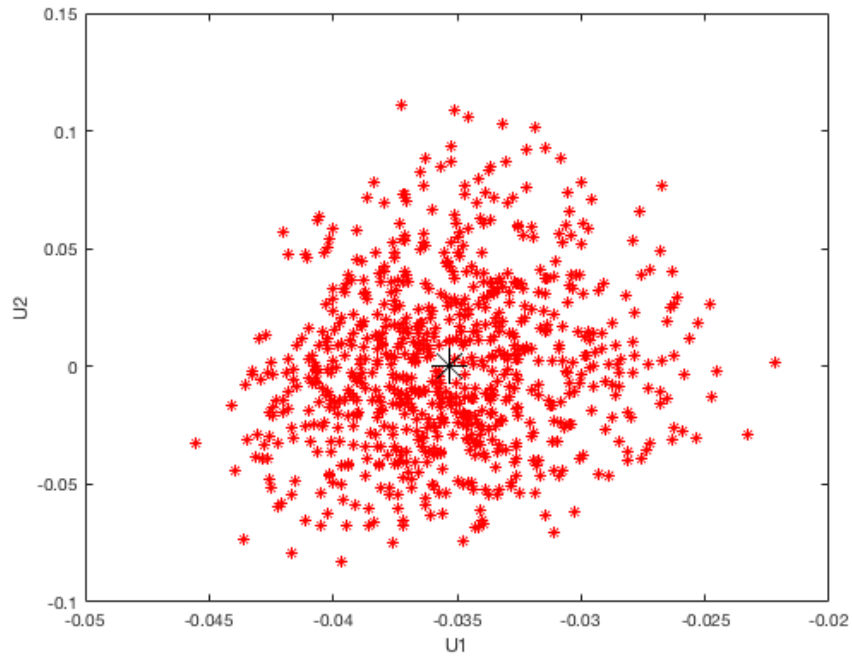


Figure 4.15: PIM Substrate Scatterplot

below shows the clusters. The figure 4.24 cluster 3 as the lowest SSE. We need to classify the pim substrate using the clusters to determine how well this specific method clustered the sequences.

We believe a better approach will be to use PhoSC and PhoSc-con and cluster using hierarchical clustering to find the best clusters.

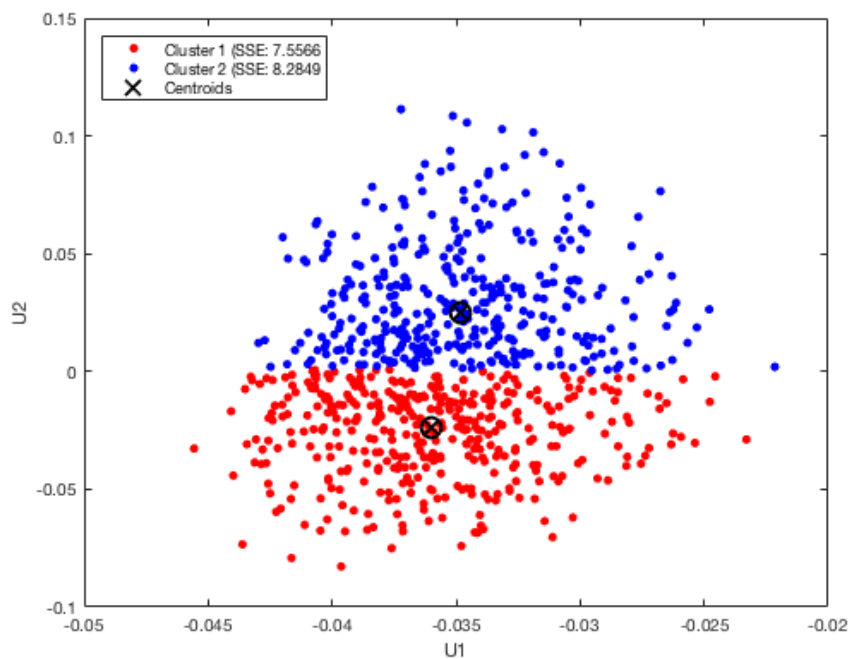


Figure 4.16: 2-Clusters: PIM Substrate Scatterplot

4.6.3 Result examples using NetAnaPhoS on PIM Substrates

4.6.3.1 Network Analytics on PIM Substrate - Using a small example(NetAnaPhoS)

The pathways play a significant role in understanding the mechanisms of the actions of a drug [126]. It also helps scientist understand more about the metabolism of drug [126].

For example, all addictive drugs affect brain pathways. The reward pathway is an example of such.

We extracted a few subsets of our graph(Top 10) or elements of our graph that had highly significant nodes and searched the associations with the PIM Substrate

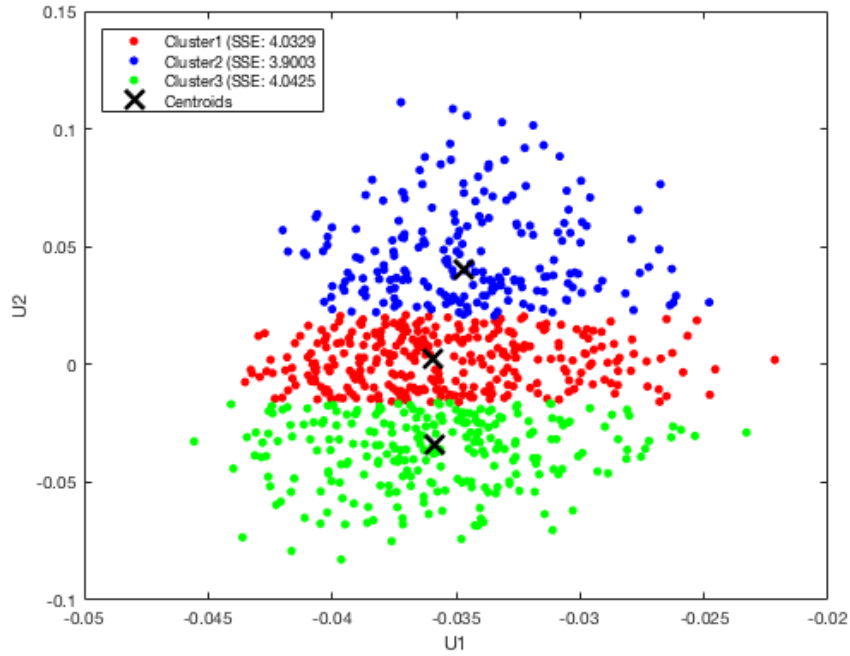


Figure 4.17: 3-Clusters: PIM Substrate Scatterplot

database. Then we applied a network analytic approach to finding high influence nodes using degree, PageRank, authority and hub scores. We applied network analytics using a directed and undirected graph. The Top 10 was used to test this specific case. The nodes = 78 and edges = 65 and average degree = 0.833 and the graph density = 0.011 and connected component = 14.

For the directed case, RS11HUMAN stands out to be a very significant gene based on its high connection to pathways in the network. The gene as the most neighbors and the results is in figure 4.25.

For the directed case, RS11HUMAN and SENP3HUMAN stand out to be a very significant gene using hub scores. RS11 HUMAN comes up again as a vital node.

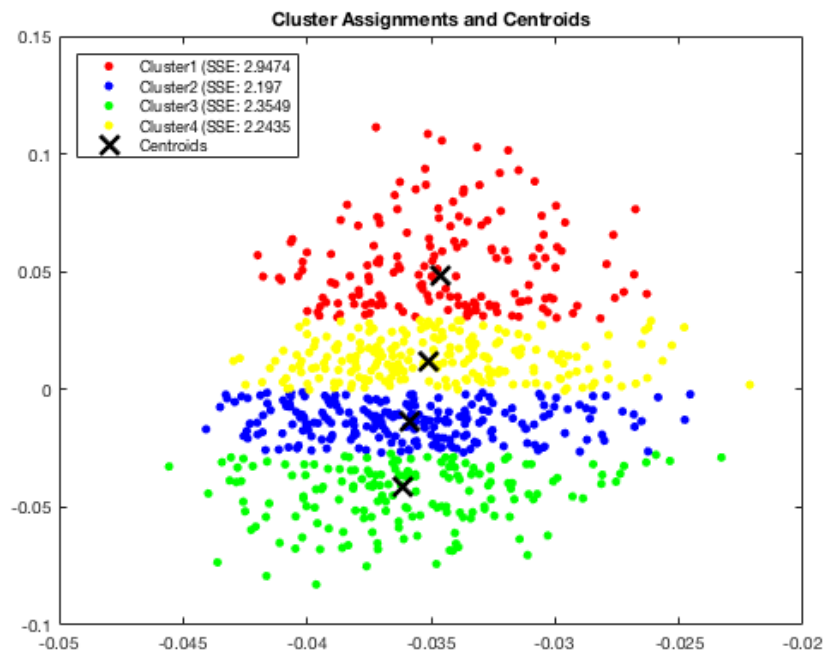


Figure 4.18: 4-Clusters: PIM Substrate Scatterplot

The PageRank scores give us the importance rank of a node. The pathways for this experiment were most important using PageRank. We can investigate these pathways for possible drugs that target these pathways.

For the directed case, multiple pathways stand out to be authorities as well. It will be essential to analyze these pathways to find drugs that would be relevant to treat prostate cancer.

The first case was for the directed network, and we further investigated the results for the undirected network using degree, PageRank and betweenness centrality. The Figure [4.25](#) and figure [4.29](#)

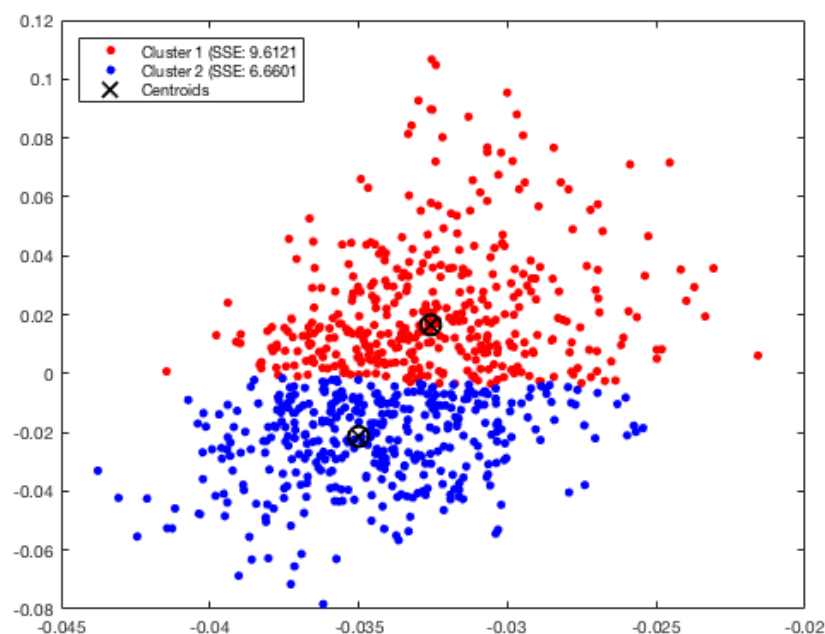


Figure 4.19: 2-Clusters:15 sequence

4.6.3.2 Top 15 Drug Repositioning Candidate

The top 15 authority scores using the directed case was used to find the drug repositioning candidates. The score results show the following drugs/chemicals could be beneficial for the treatment of prostate cancer. E.g., Adenosine-5-Diphosphoribose, Diphthamide, Guanosine-5'-Diphosphate, L-Serine, Pyridoxal Phosphate, 7-Methyl-Gpppa, 7-Methyl-Guanosine-5'-Triphosphate, 7n-Methyl-8-Hydroguanosine-5'-Diphosphate, LY2275796, Tigapotide, Alpha-Hydroxy-Beta-Phenyl-Propionic Acid, Anisomycin, Omacetaxine mepesuccinate, and Puromycin.

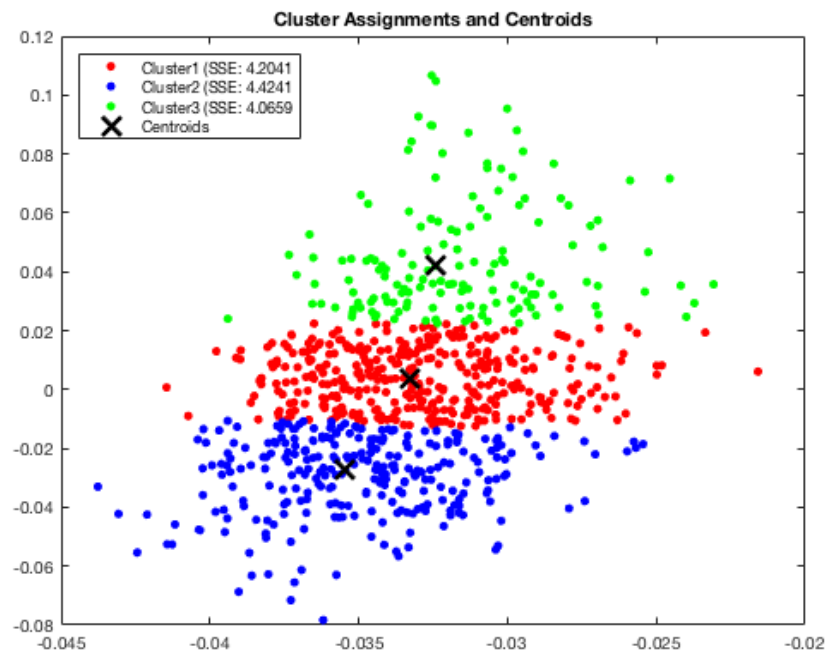


Figure 4.20: 3-Clusters:15 sequence PIM Substrate

4.6.3.3 Identification of High Influence using PIM Substrates - Example 2

In this second example, we identified an approach to clustering sequences if a majority of the sequences belonged to the same cluster. The substrate with high influence was determined by finding the distance to the centroids. The closer a pim substrate is to centroid the more critical the sequences are. For our experiment, we used statistical measures to determine the statistical significance of the result to ensure the best PIM substrates were selected. The result below shows the closeness of the PIM substrates to the centroid and the first ten pim substrates identified using the 4.15 our techniques. The top 25 PIM substrates were used to extract relevant

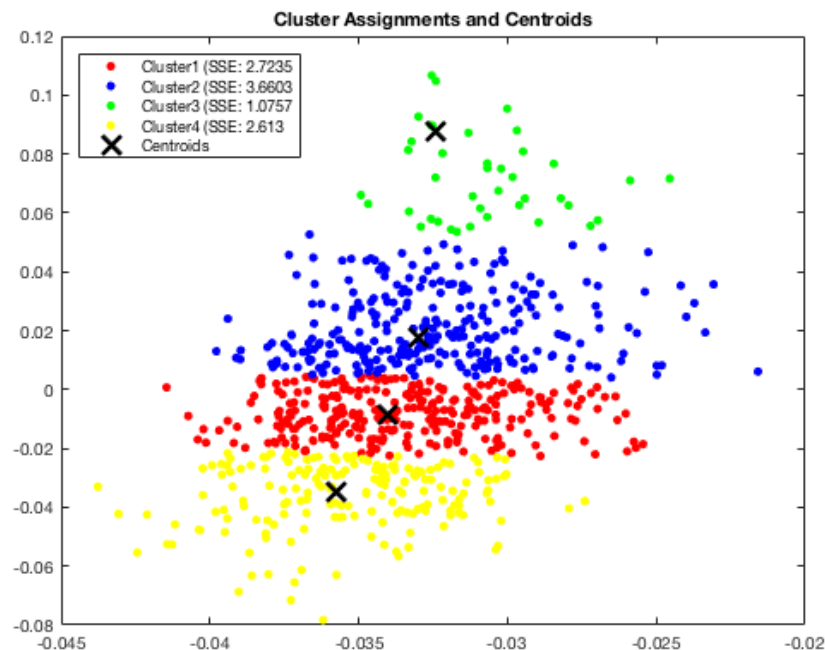


Figure 4.21: 4-Clusters:15 sequence PIM Substrate

pathways. The PIM database was used to select channels relevant or associated with the high 25 PIM substrates. Our results show consistencies with the lab showing a majority of the substrates identified to be connected to the study where determined by our methodology to be relevant for the research. For example, APOEHUMAN allows the binding of lipoprotein particles. It takes the role as a ligand for the LDL (apo B/E) receptor and the specific apo-E receptor of hepatic tissues. [127] justifies the expression in prostate cancer.

4.6.4 Network Characteristics using Top PIM Substrate

The top genes found from the previous step was used to identify Reactome(pathway) associations, and the pathways were further used to determine drugs that are con-

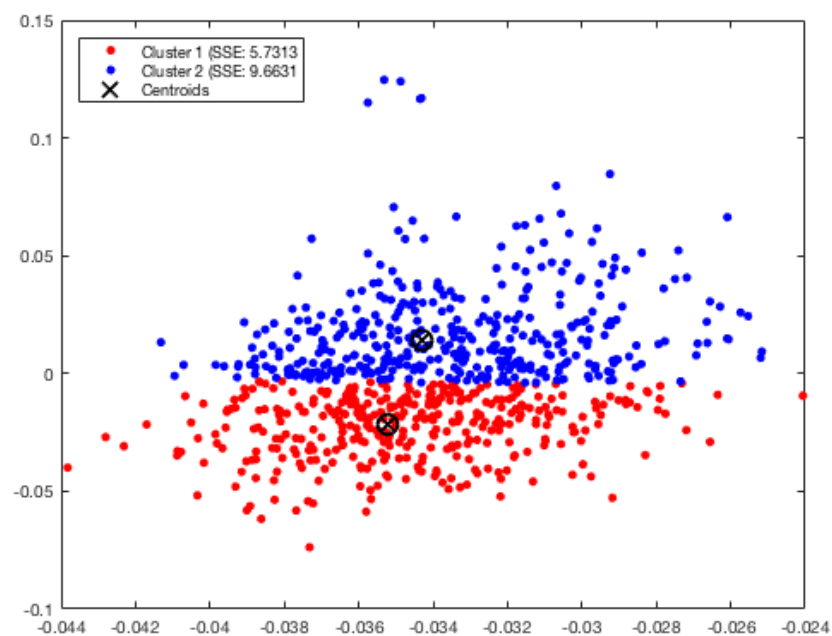


Figure 4.22: 2-Clusters:20 sequence PIM Substrate

EHIL1HUMAN
MDHCHUMAN
ZN609HUMAN
APOEHUMAN
MAP4HUMAN
SFR19HUMAN
TUT4HUMAN
MED1HUMAN
SRRM2HUMAN
CHERPHUMAN

Table 4.7: Top 10 PIM Substrates

nected to these pathways. The first step found all pathways that related to any of the significant genes identified using figure 3.8. The second step was to see all drugs that are connected directly to the pathway that was detected in the previous step.

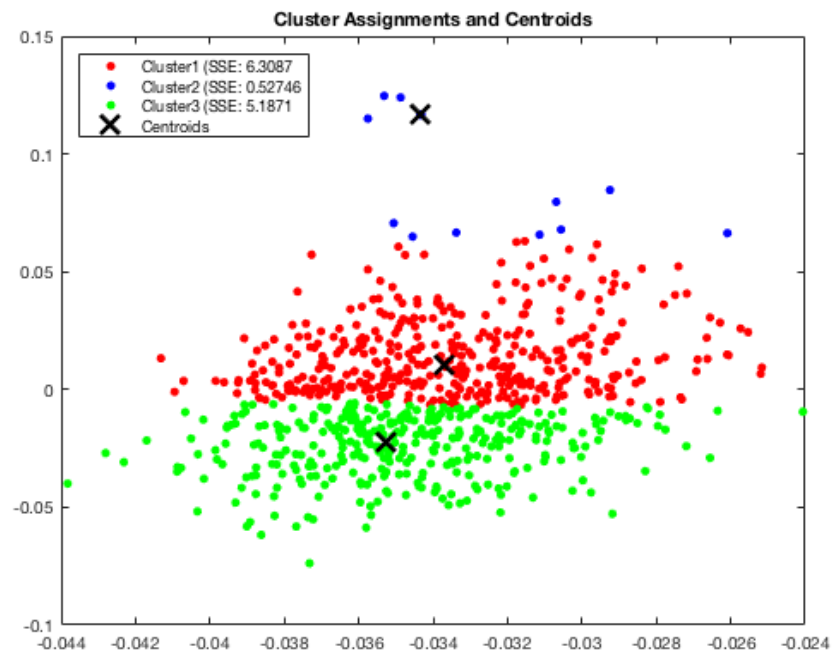


Figure 4.23: 3-Clusters:20 sequence PIM Substrate

Statistics	Score
Mean	0.0012471
Median	0.00012534
Minimum	0.00078323
Q1	0.0011613
Q3	0.00134823
Interquartile Range	0.00018693
Maximum	0.0015905
Range	0.00080727
Lower Fence	-0.0016426
Upper Fence	0.00162861

Table 4.8: Statistical Results from Using the One Centroid

Both the first and second step were combined to perform network analytics. The nodes in this network are 4,537, and the edges are 29,632. The average degree gives

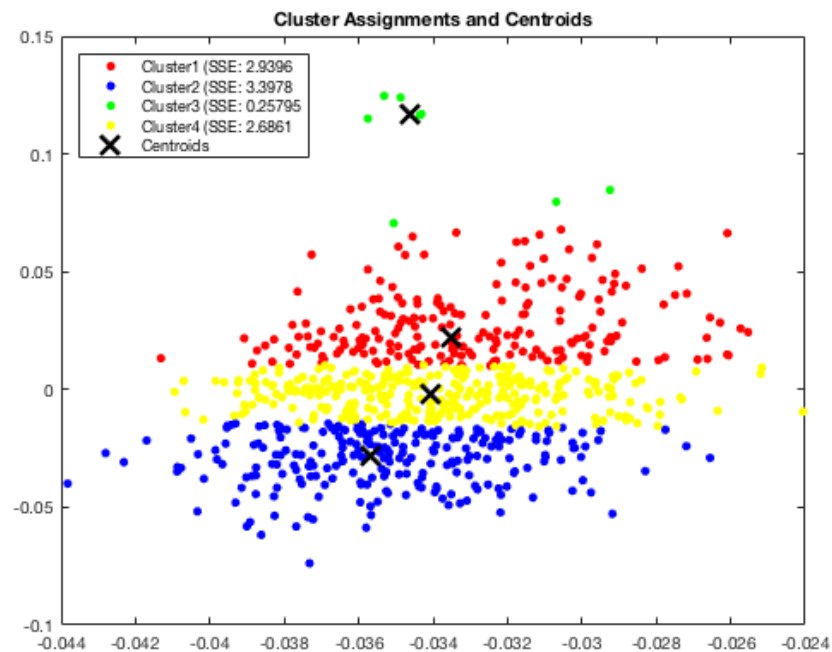


Figure 4.24: 4-Clusters:20 sequence PIM Substrate

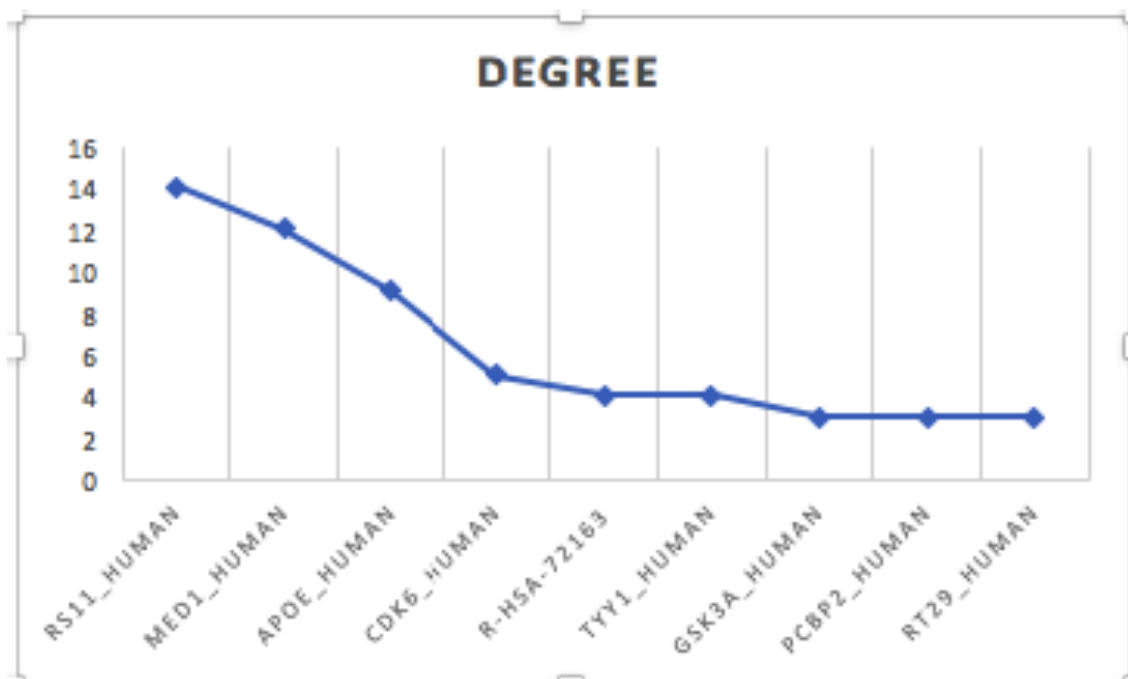


Figure 4.25: Degree: PIM Substrate to Degree

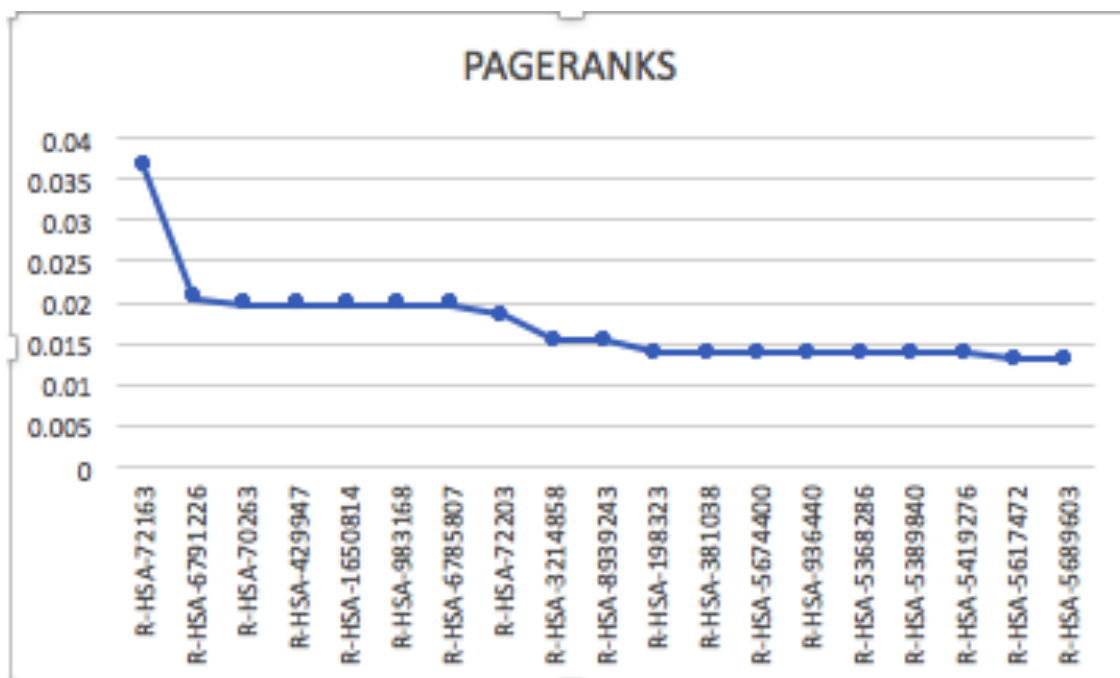


Figure 4.26: PageRank: PIM Substrate to Pathway

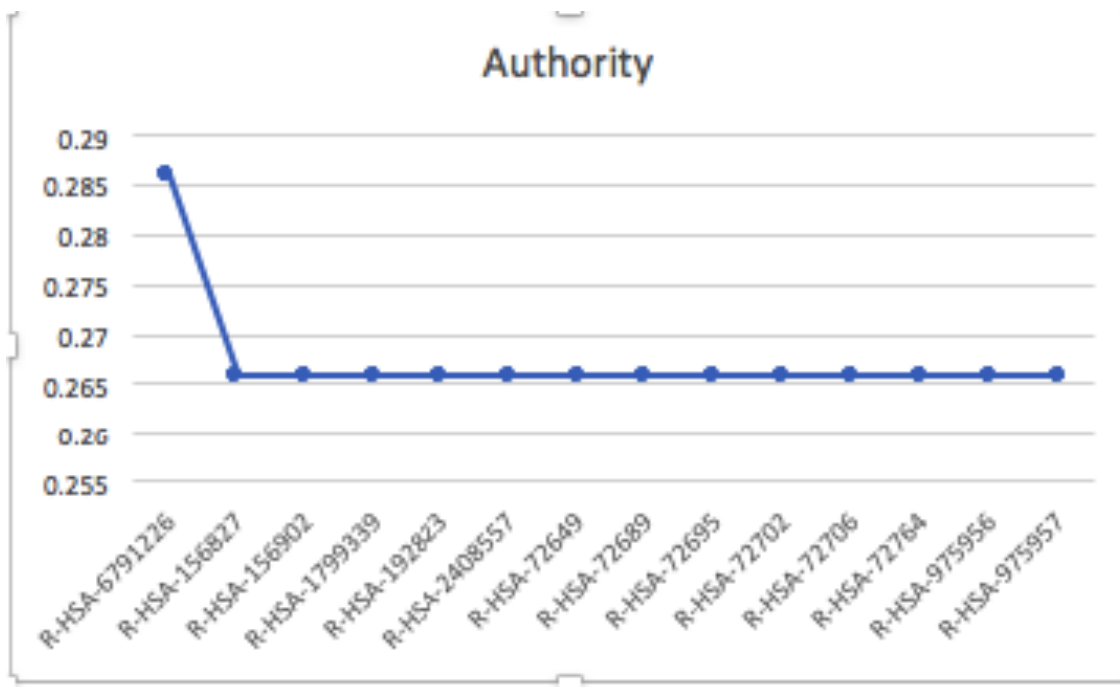


Figure 4.27: Authority: PIM Substrate to Pathway

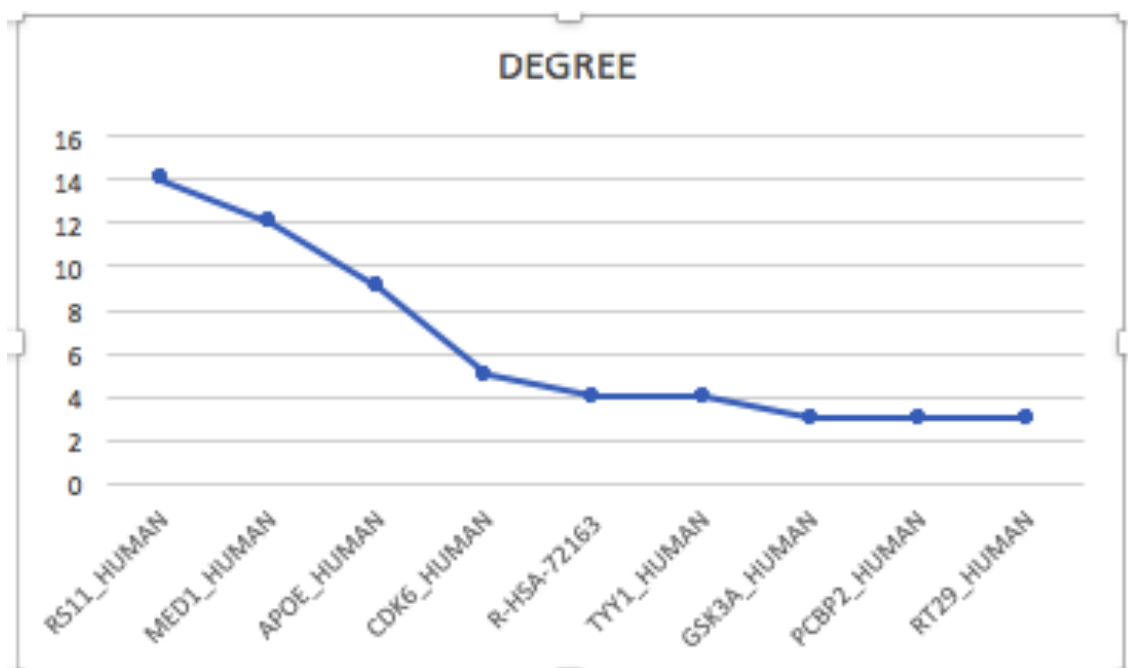


Figure 4.28: Undirected Degree: PIM Substrate to Pathway

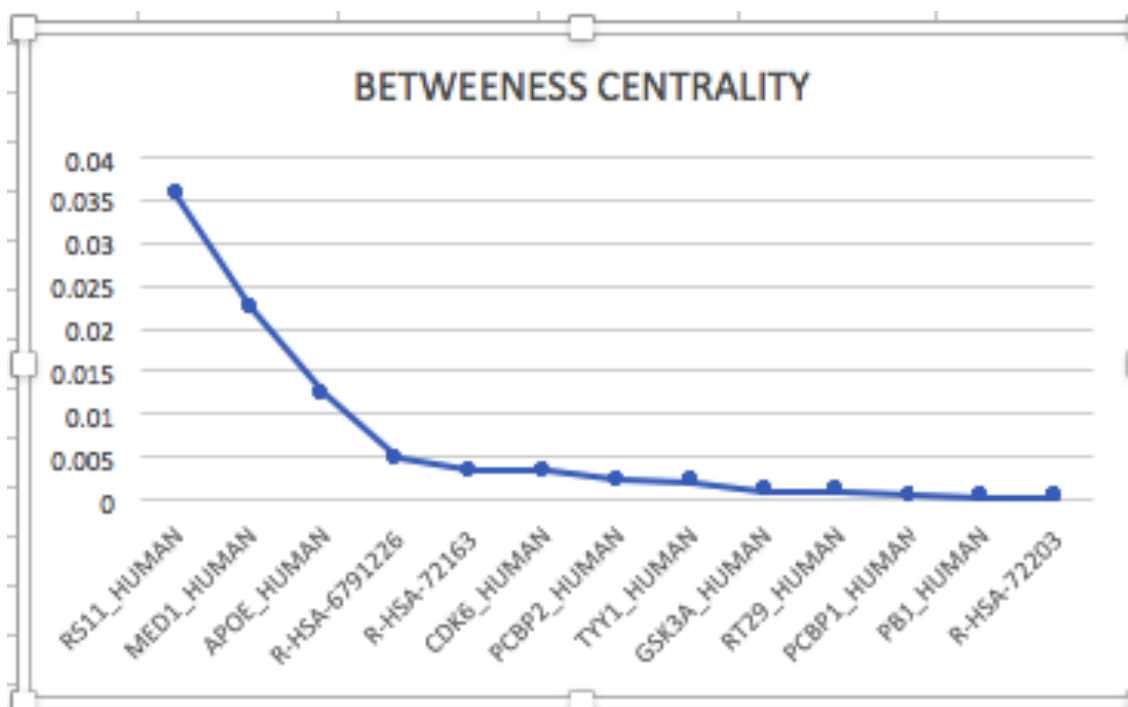


Figure 4.29: Undirected Betweenness: PIM Substrate to Pathway

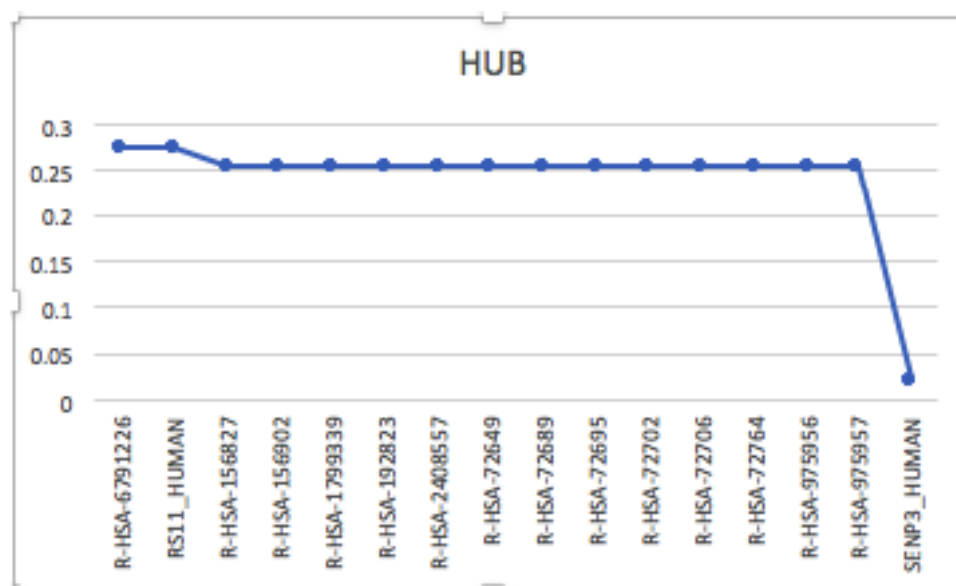


Figure 4.30: Undirected Hub: PIM Substrate to Pathway

13.062. This means each node connects on the average out to 13 other nodes. The network diameter is 10. The Graph density is 0.003. The average path length = 4.247. The figure 4.31 contains a mixture of both genes, pathway, and drugs after we calculated the betweenness centrality. Using the betweenness centrality measure, we select drugs with very high scores. The figure 4.32 shows drugs with very high betweenness centrality.

4.6.5 Our Results: Sequence Clustering

The PhoSc clustering uses the Phosphorylated alphabets, and PhoSc-con uses the hydro scaled Phosphorylated alphabets for clustering phosphorylated sequences. We shed some light on some other encoding approaches, but we will be focused on the encoding scheme with the best results.

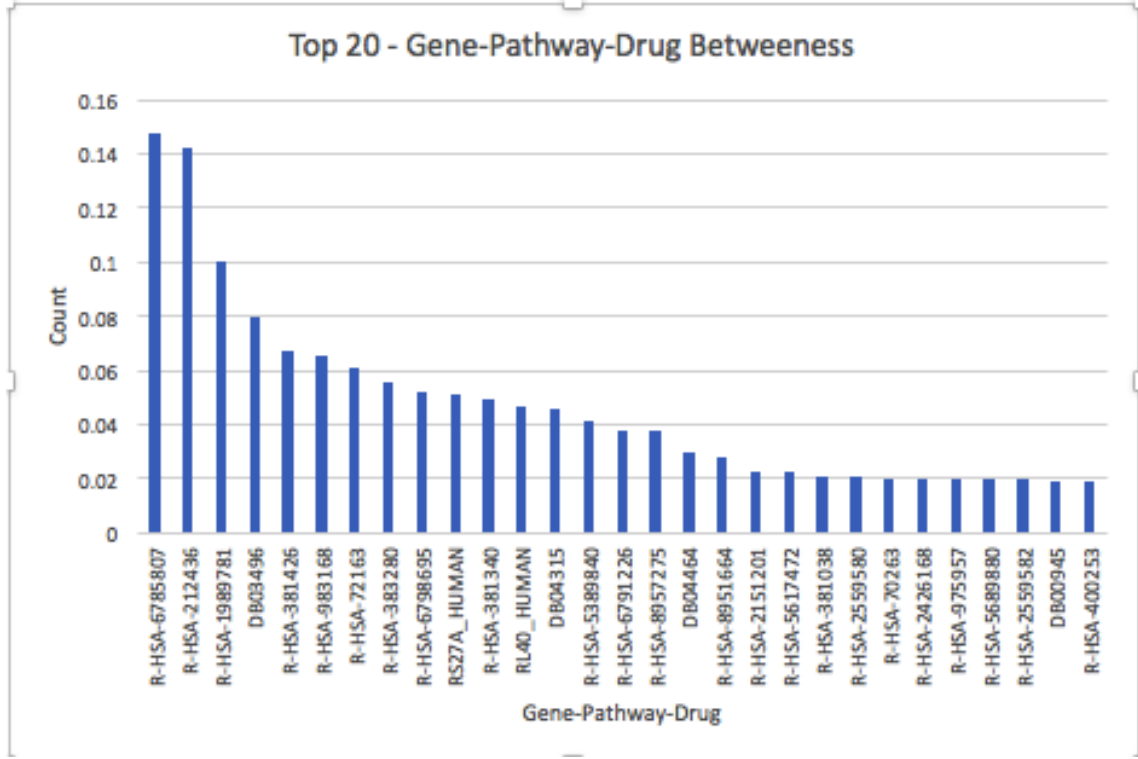


Figure 4.31: Top 20 Gene-Pathway-Drug Betweenness

4.6.5.1 PhoSc and PhoSc-con clustering

We applied phosphorylated alphabets and what we called the phosphorylated alphabets-const to encode the pim substrate sequences. We used the Hierarchical clustering algorithm on the hamming distance (Protein sequence distance) to find clusters of protein sequences that are similar to one another. The hierarchical method of clustering is a method of clustering objects into groups by finding their distances. The first step is finding the distances between the objects. For example in the case of sequences, the distances between all protein sequences were found. Each of these distance was treated at first independently of one another. In other words, each represents a cluster. The algorithm will then find the two clusters that

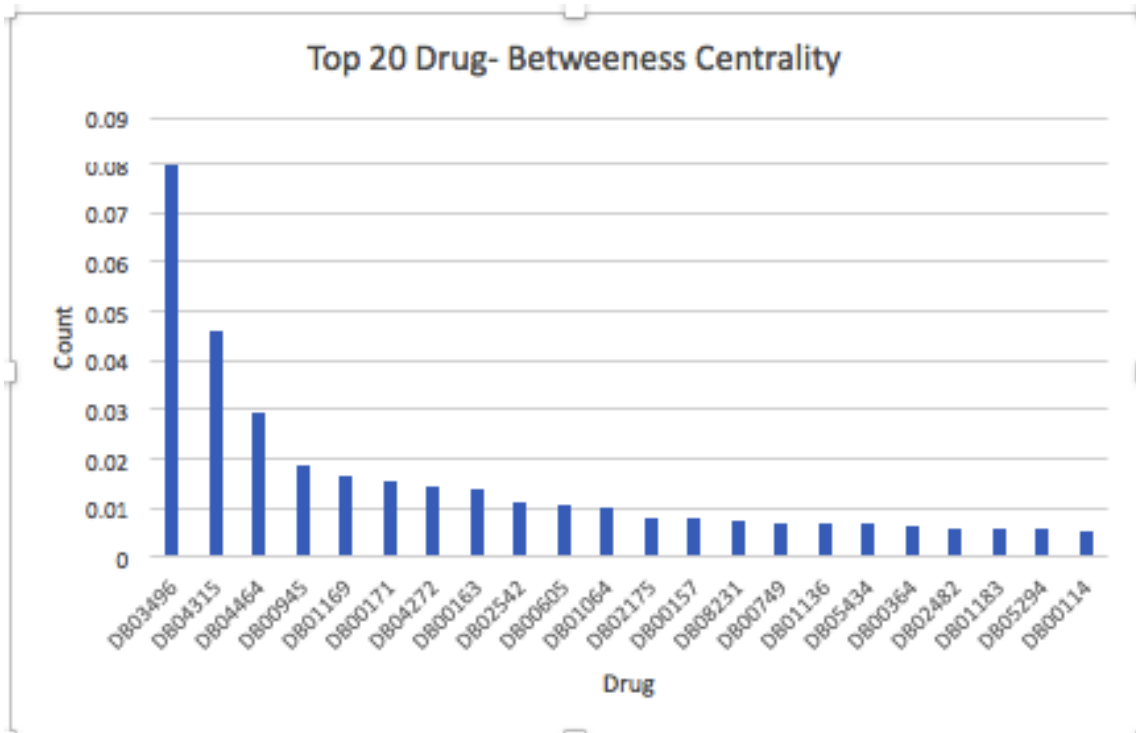


Figure 4.32: Top Drugs: Using Betweenness

are close to one another and merge two most similar clusters. It goes on until all the clusters are combined.

The distance was calculated by using distance metrics like Euclidean distance. Many other distance metrics can be used. The Euclidean distance is defined as the length of a path between two points.

The goal is to use the Euclidean distance to determine hierarchical clusters between the PIM sequences by first finding the hamming distance between the sequences (10 sequences). It is also vital to have distance metrics selected to have a linkage criterion. The linkage criteria help to determine how the distance is created. The distance is either cluster based on two very similar parts of the cluster which are also known as single-linkage or two least similar which is also referred to as complete

linkage or the center of the cluster which is also known as the average-linkage. We have selected the single-linkage in other to group by closely similar clusters. Similar clusters are then typically combined sequentially. Our results show very promising findings.

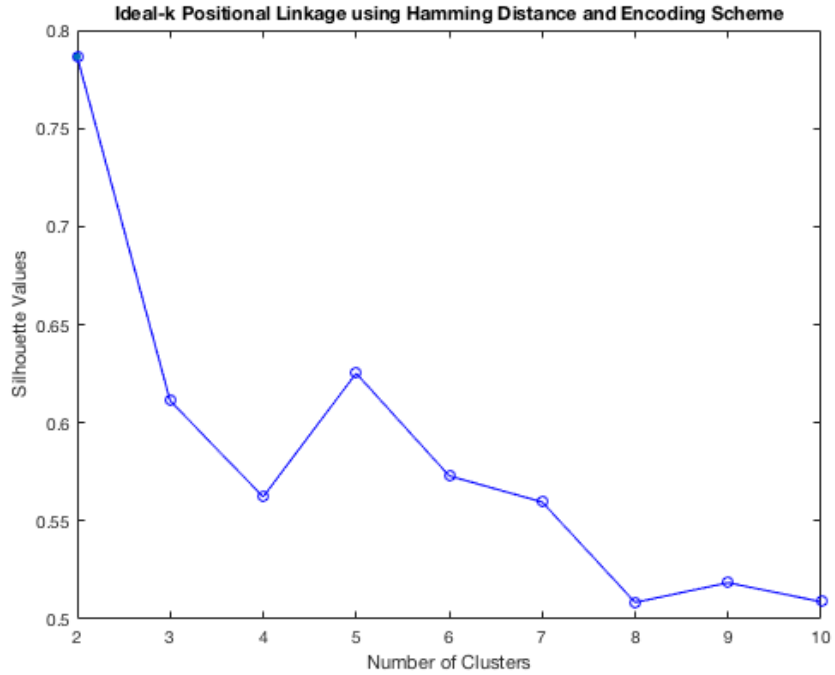


Figure 4.33: Ideal-k for PhoSc on 10 sequence(Cluster = 3)

The diagram below shows the result of clustering the sequences into 3 clusters using positional linkage hierarchical clustering.

4.6.5.2 Identification of PhoSc: using fake sequence + real rhabdomyosarcoma sequences

We used this experiment to demonstrate that our method(PhoSc) can select the rhabdomyosarcoma sequence from the group of sequence that contained both

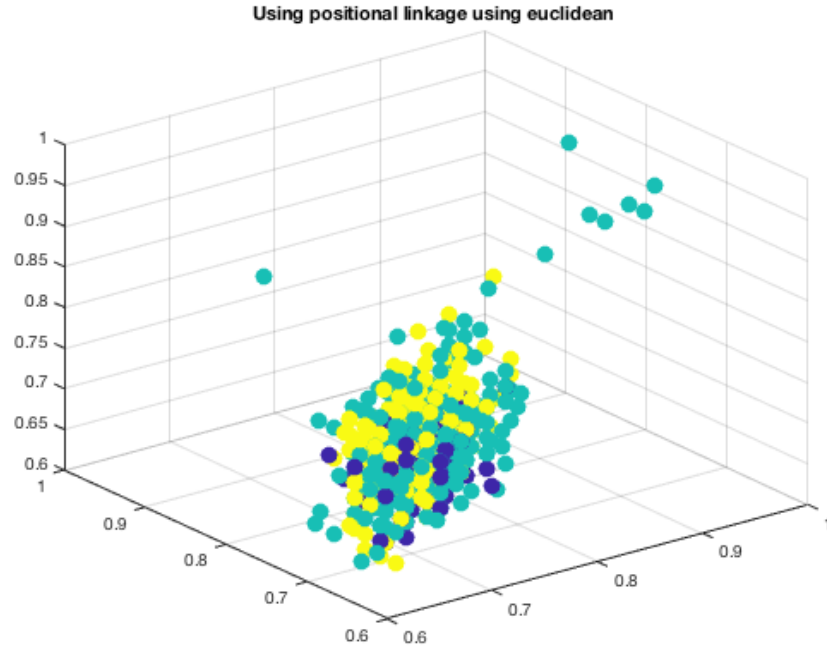


Figure 4.34: Positional Linkage using Hierarchical clustering using 10 sequence(Cluster = 3)

the fake sequences and rhabdomyosarcoma sequences. We had 700 real-sequences for rhabdomyosarcoma and 700 pseudo or fake protein sequences.

4.6.5.3 Standard Encoding SVD analysis

The rank for the matrix was 780. We constructed a matrix similar to the term-document matrix using the binary encoding and apply the SVD after finding the rank of the matrix. The figure below shows the results of the experiment. Since $r = 780$. The figure doesnt display its high dimensional results. However, we can use the inter-cluster and intra-cluster similarities to determine how the cluster are within and how dispersed they are from one another. The cluster 1 contains $SSE = 432.7217$, the SSE scores of the cluster 2 contains $SSE=439.4497$, and the cluster 3

had an SSE score of 457.8834. In this experiment, we used the cosine similarity to measure the distances.

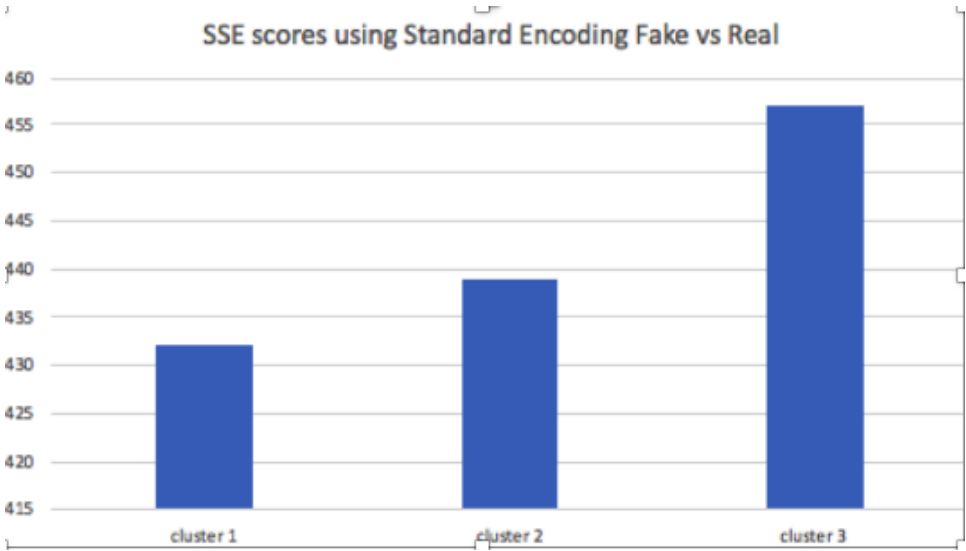


Figure 4.35: SSE scores using cosine similarity

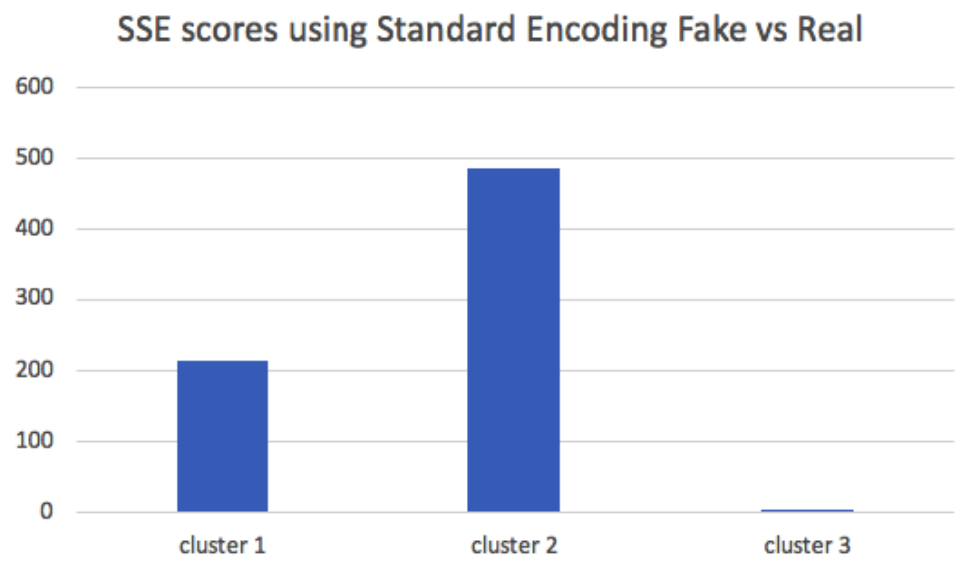


Figure 4.36: SSE scores using Euclidean Distance

We performed the same experiment with the Euclidean distance. The SSE scores are illustrated on the chart. We use the elbow method to determine which

clusters worked the most for our sequences(ideal-k).

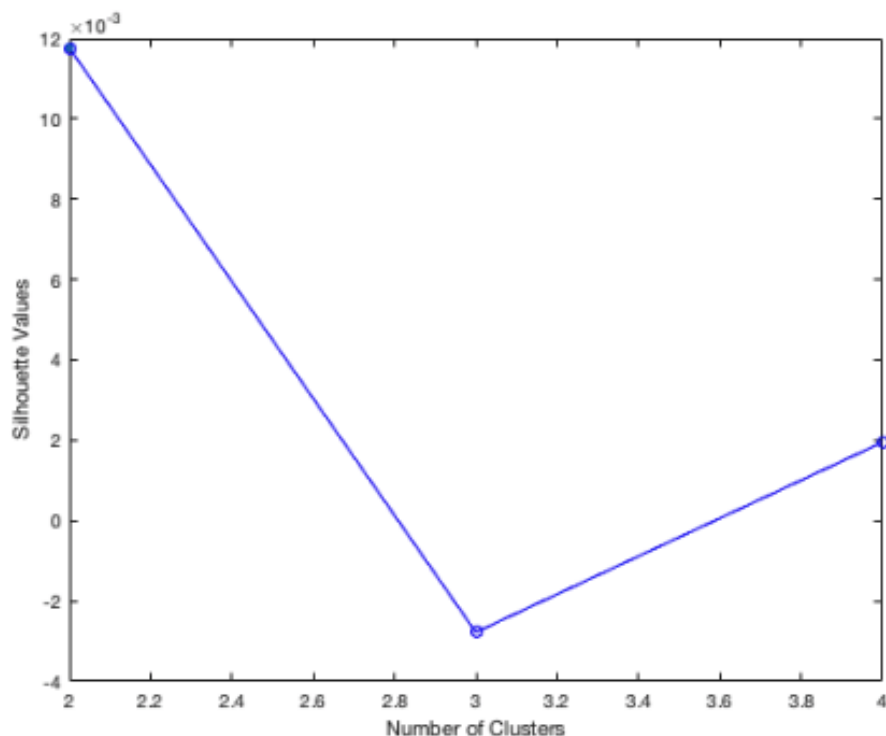


Figure 4.37: Ideal-k Standard encoding using cosine similarity

4.6.5.4 Phosphorylated Categorization

We have applied encoding scheme that contains phosphorylated categorization. This setup gives us a more dispersed chart compared to the standard encoding chart (See Figure). The reason behind this is that we have imposed encoding scheme on the alphabet representation. The SSE scores are below.

The ideal-k for this specific encoding schemes shows the best $k = 3$. The figure below shows the ideal-k for the phosphorylated encoding schemes. This method was selected using [128] who demonstrate that using these techniques helps to separate

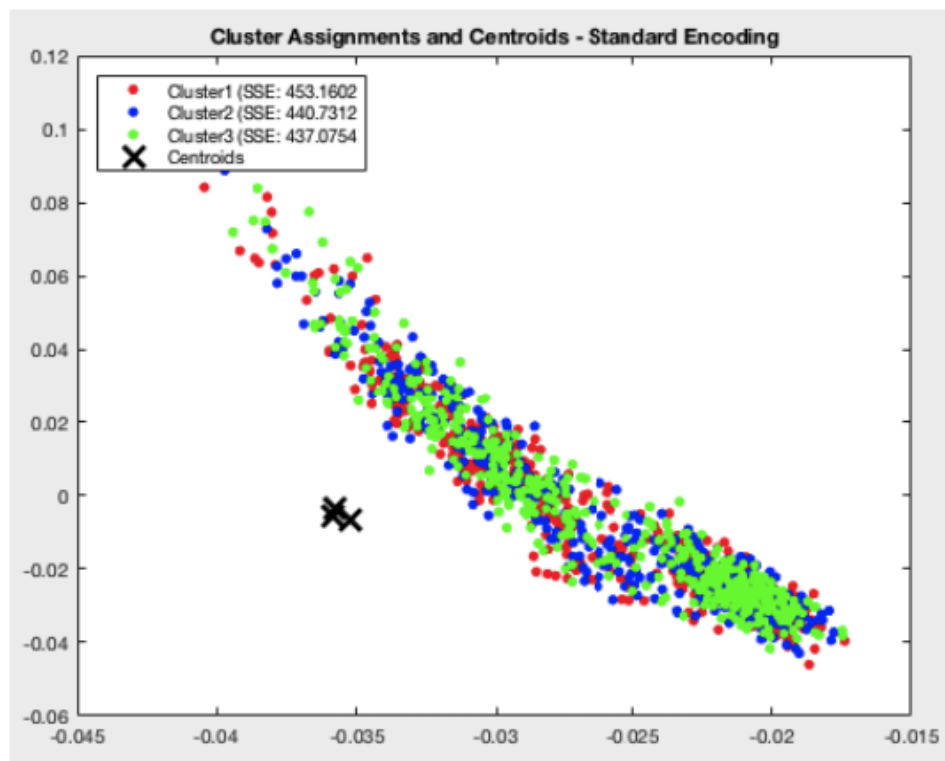


Figure 4.38: Clustering using Cosine similarity Fake vs Real

the phosphorylated elements from the others. The silhouette diagram for the clusters also shows our close a point is to the other points in its cluster when compared to how similar it is in other clusters. A majority of our cluster points have high silhouette scores indicating that $k=3$ is appropriate for clustering the phosphorylated alphabets.

$$S_i = (b_i a_i) / \max(a_i, b_i)$$

The clustering results using the encoding scheme structural alphabets showed that all clusters had very similar SSE scores cluster 1 with 31.29, cluster 2 with 21.16, and cluster 3 with 28.2156. This result shows the level closeness within each cluster. The silhouette score figure shows below with an ideal- k of 4. The silhouette

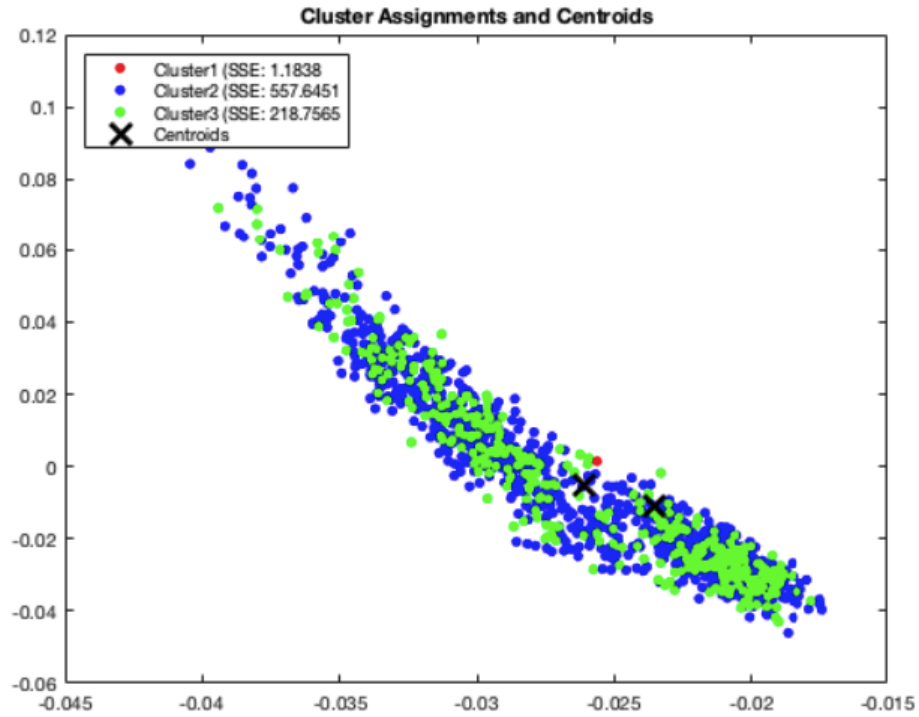


Figure 4.39: Clustering using Euclidean distances to cluster the datasets - Standard Encoding

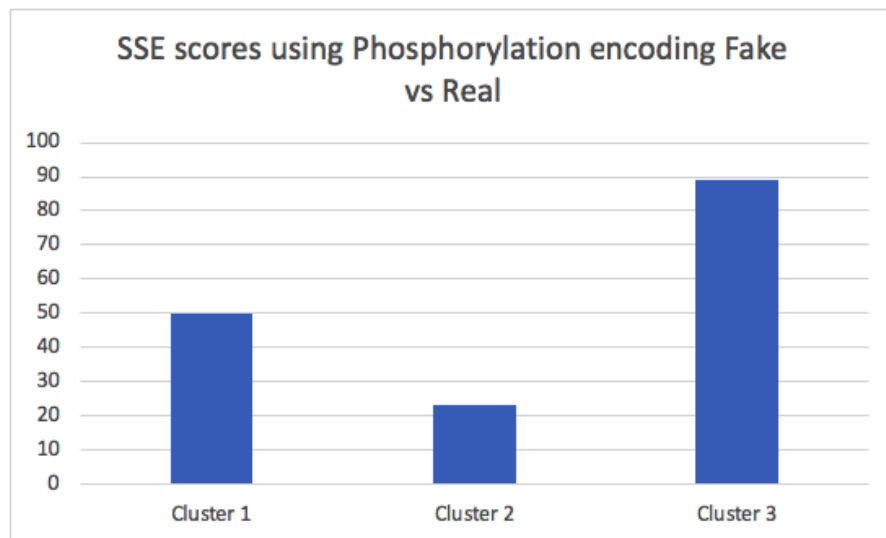


Figure 4.40: SSE scores using Phosphorylated Categorization

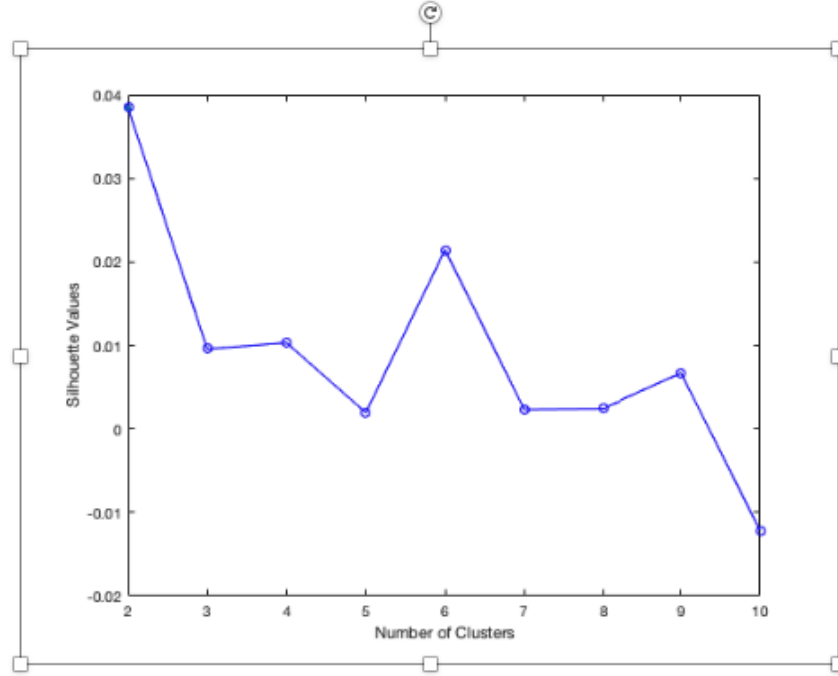


Figure 4.41: Phosphorylated Alphabets with fake vs real sequence

value comparison also shows points that are clustered correctly where they were supposed to belong.

4.6.5.5 Scaled Phosphorylation Encoding Scheme

We introduced a scaled approach to viewing our encoding schemes which improved the results of the clusters for ideal-k. The figure below shows the silhouette when the scaled is introduced.

4.6.6 Result using our Method for Sequence clustering (PhoSc)

The linkage result show cluster stability. We intend to do the same test on the encoding scheme that we impose some constraints on.

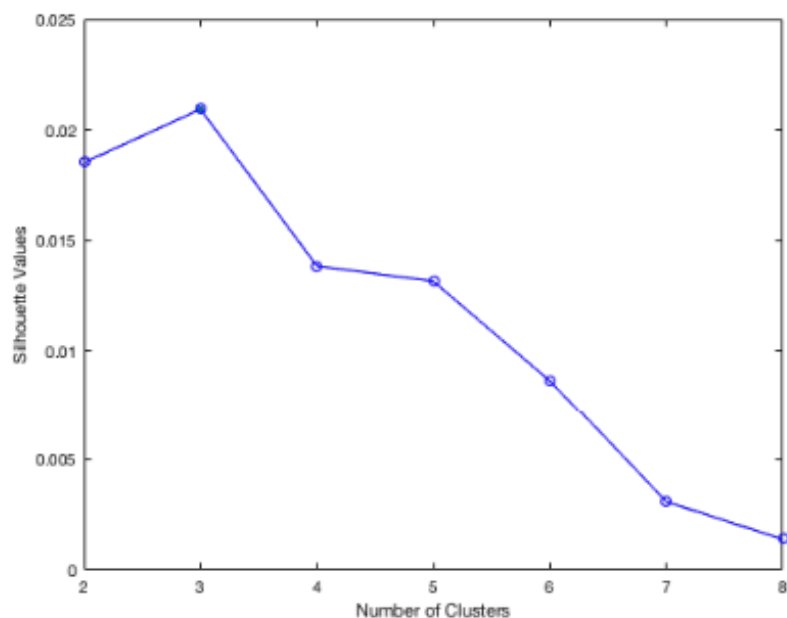


Figure 4.42: Structural Alphabet result using real vs fake sequence

The ideal-k was stable using linkage and phosphorylated scaled encoding schemes. The clustering results using linkage and phosphorylated hydro scaled encoding worked well and classified a majority of the fake proteins in the same cluster. The precision scores for the pseudo proteins using linkage and hydro scaled phosphorylated elements was 90%. The recall scores were 82.7%.

The ideal-k using PhoSc gave us excellent cluster stability as a result below indicates making linkage + Phosphorylated Alphabets a stable encoding scheme to use to select best clusters for network analytics, multimodal clustering.

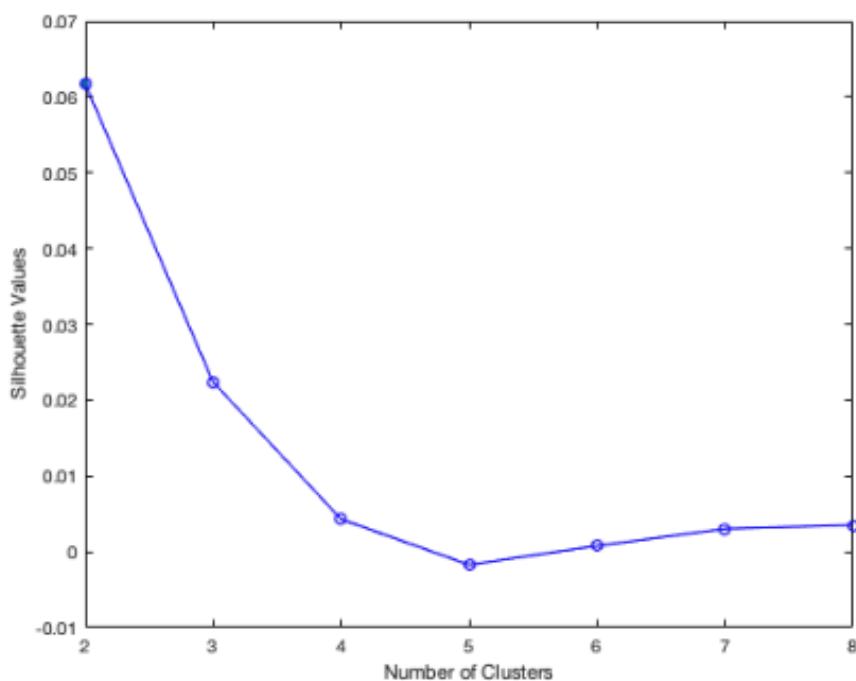


Figure 4.43: Hydroscaled Phosphorylated Alphabets for fake vs real

4.6.7 Sequence analysis - Method Comparisons

To validate our approach, we found the precision and recall. The precision was 92%, and the recall was 85%. Our results compare to others.

4.6.8 Comparison of PhoSc and PhoSc-con on PIM Substrate

With regards to PhoSc, Cluster 1 contained a majority of the lab tested substrate with the precision of 90% and the recall of 71%. Our approach can be used to identify or predict potential phosphorylated elements as well.

With regards to PhoSc-con, Cluster 3 contained a majority of the lab tested

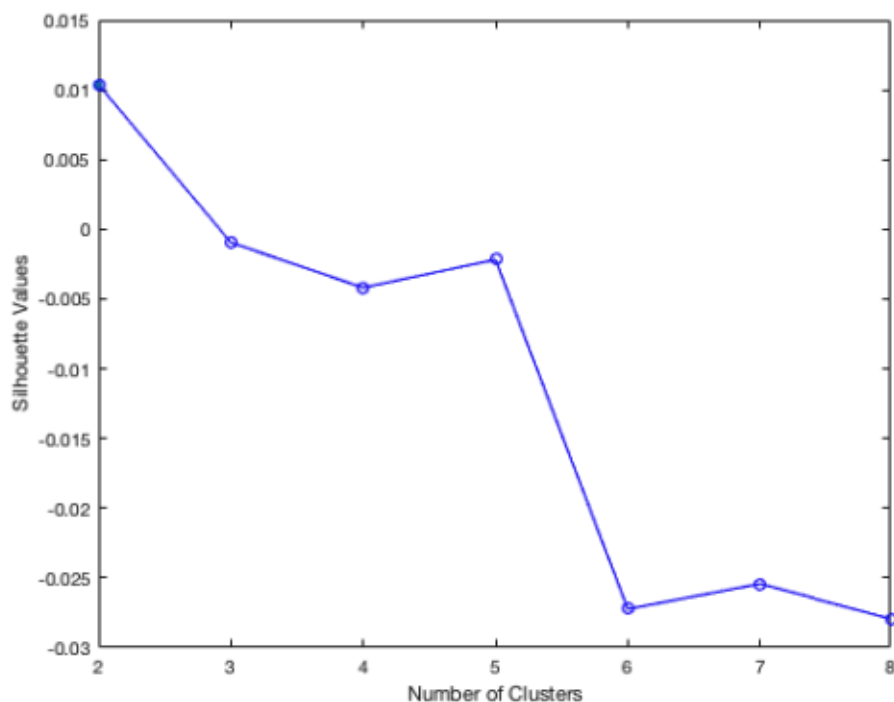


Figure 4.44: PhoSc for Clustering Fake vs Real

substrate with the precision of 100% and the recall of 76%. Our approach can be used to identify or predict potential phosphorylated elements as well.

We compare the results with the SVD. The SVD scores give us 50% precision while the recall was 33%.

4.7 NetAnaPhoS- pim substrates combined analysis of curated databases

We have combined all substrates and generated a network with them. The total number of interactions extracted given the PIM substrates was 153,846 and the number of nodes 39980. The network diameter is 4.426. The modularity is 0.768. The figure 4.49 shows the number of all the disease-related elements connected to

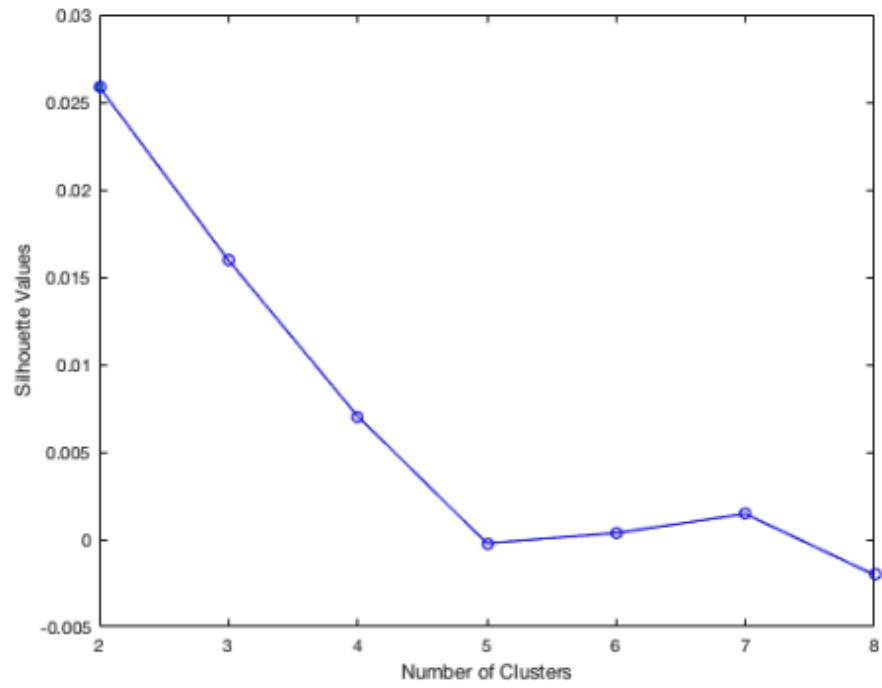


Figure 4.45: PhoSc-con for Clustering Fake vs Real

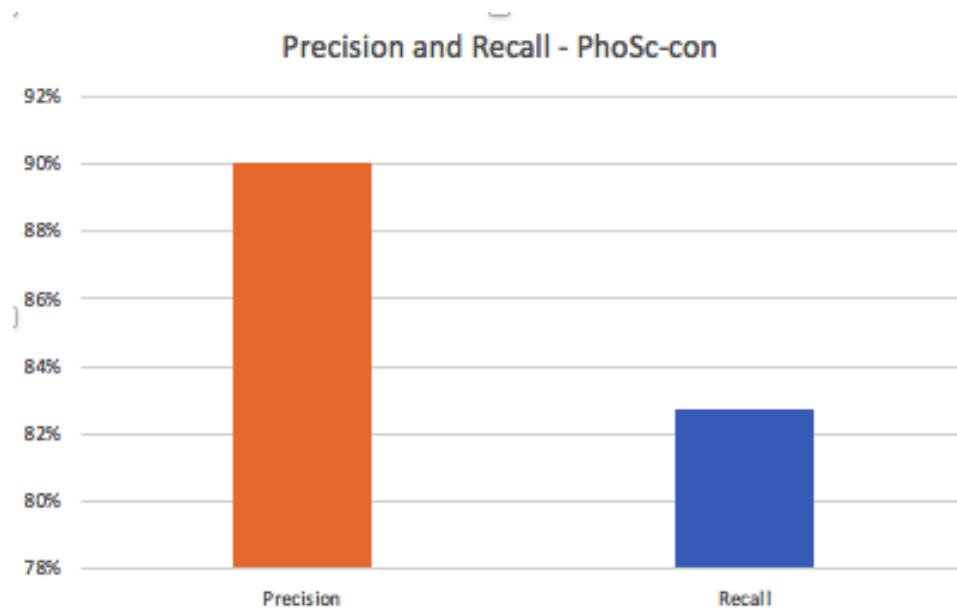


Figure 4.46: Precision and Recall using Rhabdomyosarcoma-fake vs real data

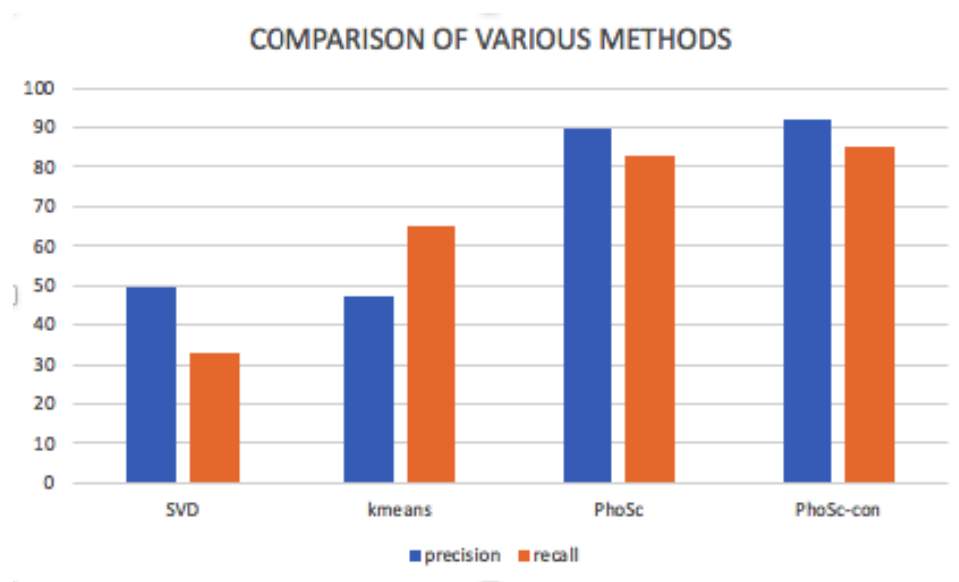


Figure 4.47: Method Comparisons

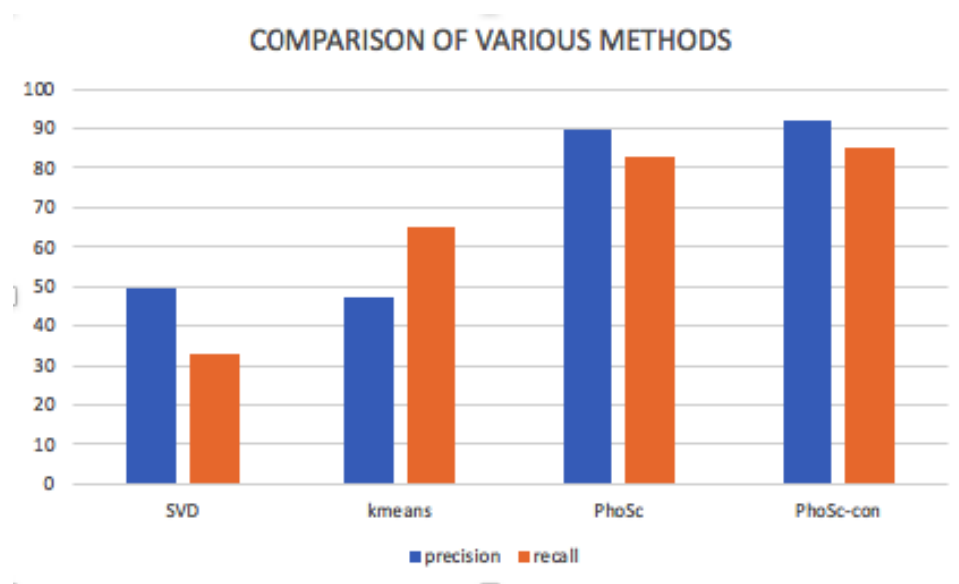


Figure 4.48: Method Comparisons: PhoSc and SVD

the PIM Substrate after sequence cleaning was performed on the substrate. We will be utilizing authority, hub, page rank, and the degree to select the drugs for repositioning candidates.

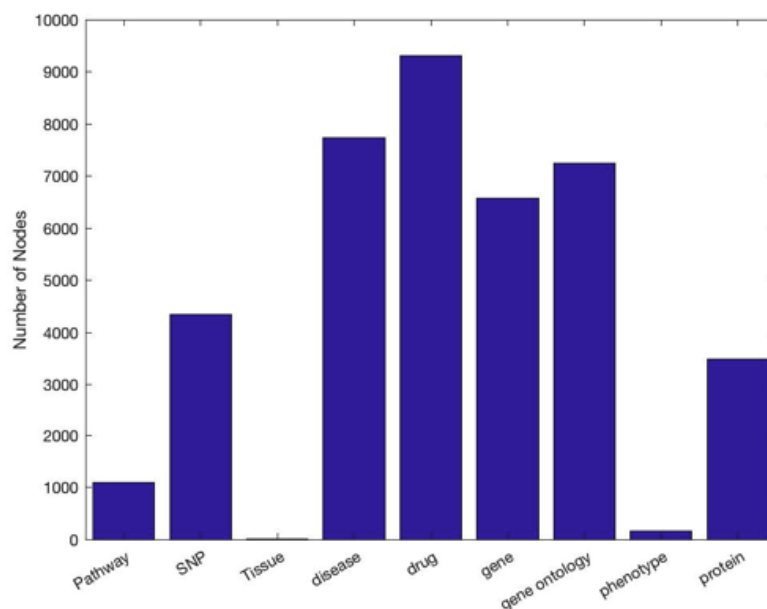


Figure 4.49: PIM Related Element Analysis

4.7.0.1 Combined analysis of using Degree

The degree scores show us the number of neighbors connected to every node. The figure 4.50 shows the nodes with the highest neighbors. The theory is that nodes with the highest neighbors must be critical because other nodes connect to it.

4.7.0.2 Combined analysis of using Authority Scores

The authority nodes are referred to as essential pages. For example, if you would love to search for anything in the world, the entire page or web page will be Google. This means that the Google page is so much influential, it is defined as an authority. We have selected the nodes with top authorities given the combined network. The authority scores highlight a lot of drugs have excellent authority, and

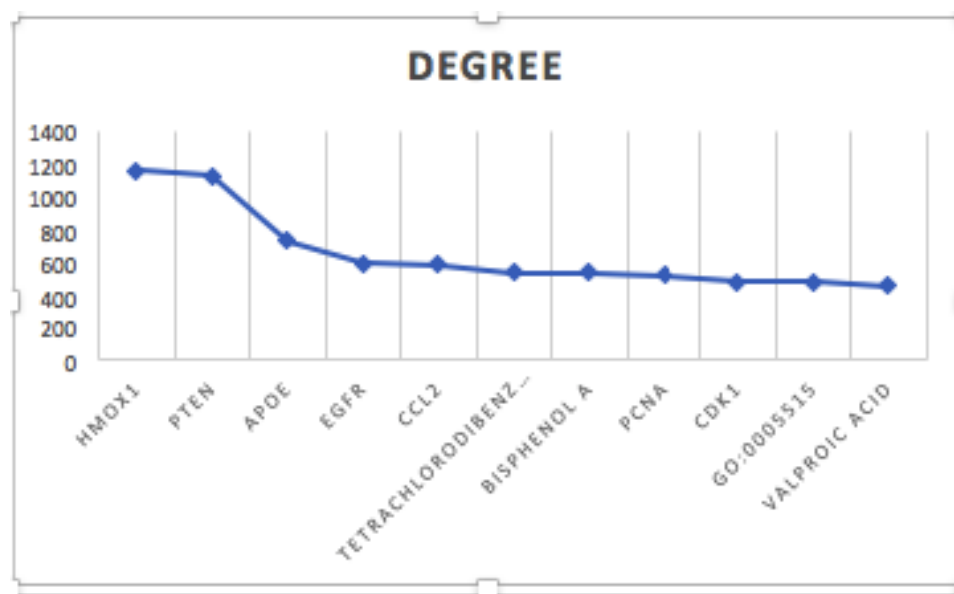


Figure 4.50: Degree: PIM Related Element Analysis

we selected the measure to find some of our initial drug repositioning results.

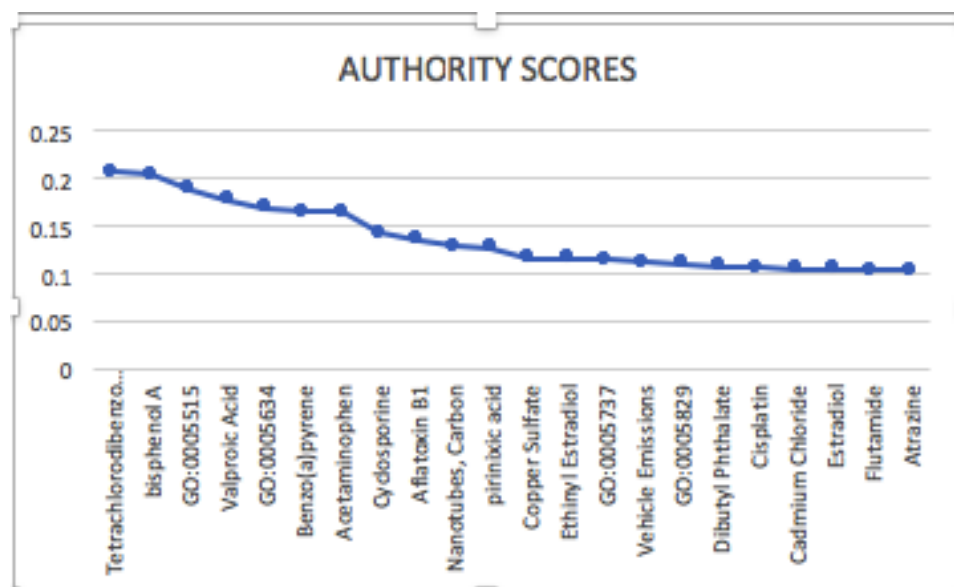


Figure 4.51: Authority: PIM Related Element Analysis

4.7.0.3 Combined analysis of using Betweenness Scores

The betweenness centrality of a node is calculated by the percentage of the number of shortest paths between any two pair of nodes. The figure 4.52 shows results from applying the betweenness centrality measure. A majority of the genes identified by the betweenness scores are highly relevant to prostate cancer.

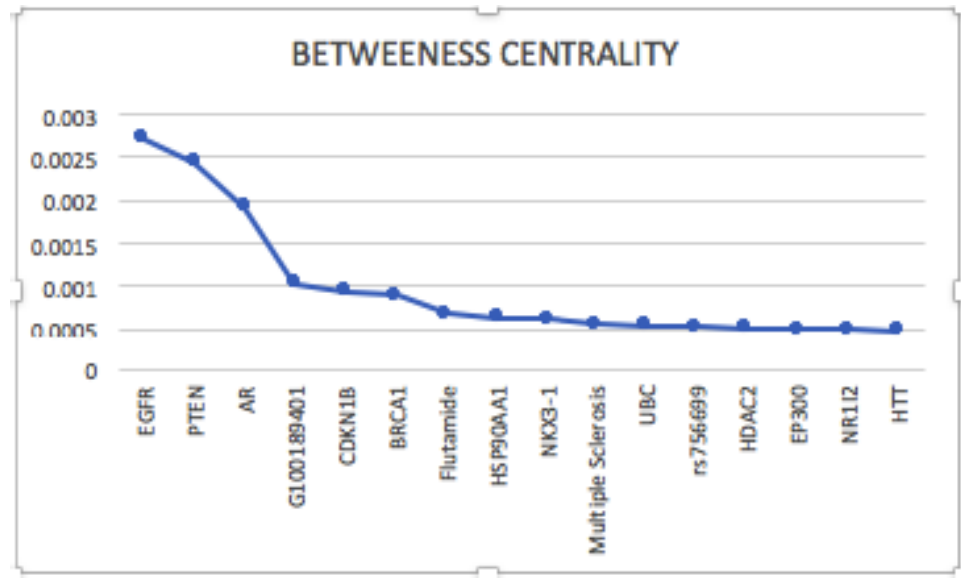


Figure 4.52: Betweenness: PIM Related Element Analysis

4.7.0.4 Combined analysis of using PageRank Scores

The PageRank scores measure the overall importance of each node by voting. The PageRank of a node uses a combination of the number of links to a particular node and the value of those links as criteria for finding the importance of that node. PageRank scores helped to identify some disease and genes using the combination of all disease-related elements. Some of the results need to be studied by the bi-

ological lab. The figure 4.53 shows the Top 15 disease-related elements with high Pagerank scores. For example, the AR gene is expressed throughout prostate cancer progression.

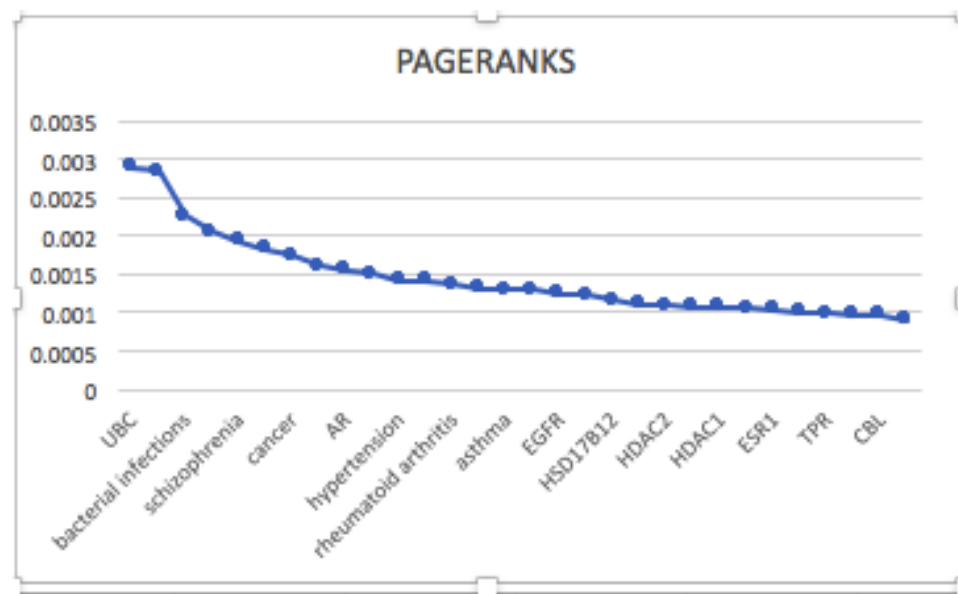


Figure 4.53: PageRank: PIM Related Element Analysis

4.7.0.5 Some results using Authority Scores

Implementing network analytics techniques helps to find indirect connections or indirect relationships that exist in a network or graph. The authority score graph, for example, identifies pathways relevant to the repositioning of drugs. We used these pathways to identify drugs that target this pathway. An example is below is tigapotide which is pending Phase II Clinical Trials. Our results found that Tigapotide is repurposable for prostate cancer.

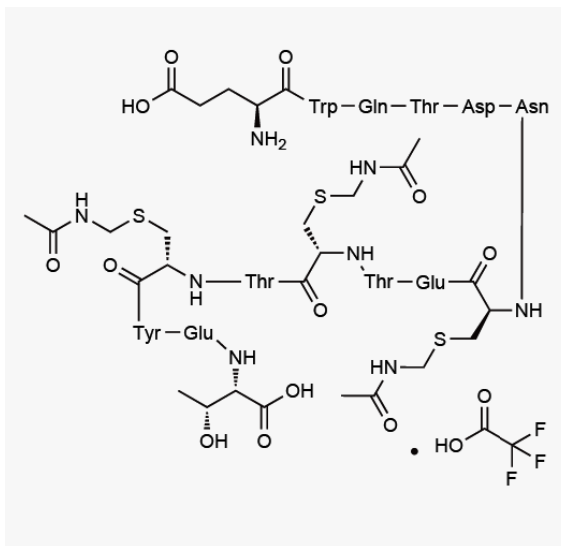


Figure 4.54: Chemical Structure of Tigapode

4.7.0.6 Singular Value Decomposition for Drug Repositioning: Pathway to Drugs using the combined PIM substrates(phosphorylated protein)

The combined network for pim substrates phosphorylated element was used to do a pathway analysis using SVD. Given that we had a pathway and drug interactions, we can also apply the Singular Value Decomposition on the pathway to drug information by creating a term-document matrix. The rows and columns with both pathways and drug information and applying the Singular Value Decomposition on the adjacency matrix and finding the centroid. We can directly apply the Euclidean distance metric to find Reactome to closest to the centroid. Some of the results include R-HSA-6791226: Major pathway of rRNA processing in the nucleolus and cytosol and R-HSA-5607764: CLEC7A (Dectin-1) signaling. The following results

need to be confirmed or verified by a biologist to confirm the actual viability and relationship to the PIM Substrate. The result can then be further associated with relevant drugs that go through these pathways. Twenty-five drugs were identified to go through this pathway. Some drug candidates that target these pathways include DB07374. Anisomycin, DB04865. Omacetaxine mepesuccinate, DB08437. Puromycin, DB04805. Virginiamycin S1, DB06151. Acetylcysteine. DB01169. Arsenic trioxide, DB00995. Auranofin, DB00244. Mesalazine, DB05183. MLN0415, DB00795. Sulfasalazine, DB00795. Sulfasalazine, DB00482. Celecoxib, DB04522. Phosphoserine, DB02010. Staurosporine, DB06616. Bosutinib.

4.7.0.7 Node Impact using Perturbation Analysis

We conducted a perturbation on the PIM network generated from the above step to identify significant biological associations composed of high-influence nodes. A two-step approach inspired by our previous study [37] that was targeting on individualized medicine has been designed [8] identifying high-influence nodes to diabetes based on different centralities, [129] and determining the biological associations based on the perturbation results. To be specific, we first selected the high-influence nodes. The influence of a node can be measured by several centralities including the degree centrality, closeness centrality, as well as betweenness and PageRank centralities. The degree centrality of a node is defined as the total number of links of that node. Nodes with high degree centrality will be expected to have a high influence because such influence is contributed from their immediate

neighbors. The closeness centrality of a node is measured by the average distance of that node from all other nodes in the network. The betweenness centrality of a node is calculated by the percentage of the number of shortest paths between pairs of nodes the nodes with high betweenness centrality measures bridge nodes that connect a significant sub-networks. The PageRank of a node uses a combination of the number of links to a particular node and the value of those links as criteria for finding the importance of that node. Nodes were selected for removal based on the values of their centrality measures. Once the high-influence nodes were chosen, we removed them and measured the effect on the modularity of the network. The node with the highest centrality measure was removed first. The modularity produced by the node removal were examined to determine the high-influence nodes; higher modularity corresponds to nodes with higher influence. In this step, we identify the high-influence nodes and performed perturbation to determine the significant biological associations with the PIM Substrate. We found EGFR, PTEN, and AR to disrupt the network significantly.

4.8 NetAnaPhoS: Rhabdomyosarcoma for network analysis using gene-drug-pathway

The result below shows the scores of the various centrality measures using a tri-mode(gene-drug-pathway). The P63 for example in figure 4.57 as shown to be directly relevant in the studies of rhabdomyosarcoma. This could be a possible target for drug repositioning. Neutrophil degranulation was also found to be a

relevant pathway. We also found some other relevant disease-related element using betweenness centrality like atk1 and DB04313. ABCB1 was also considered an influence node found using the degree centrality measure. The ABCB1 has been found relevant in rhabdomyosarcoma network making it a very relevant target. For more of the centrality results check figure 4.55, figure 4.58, figure 4.57, and figure 4.56.

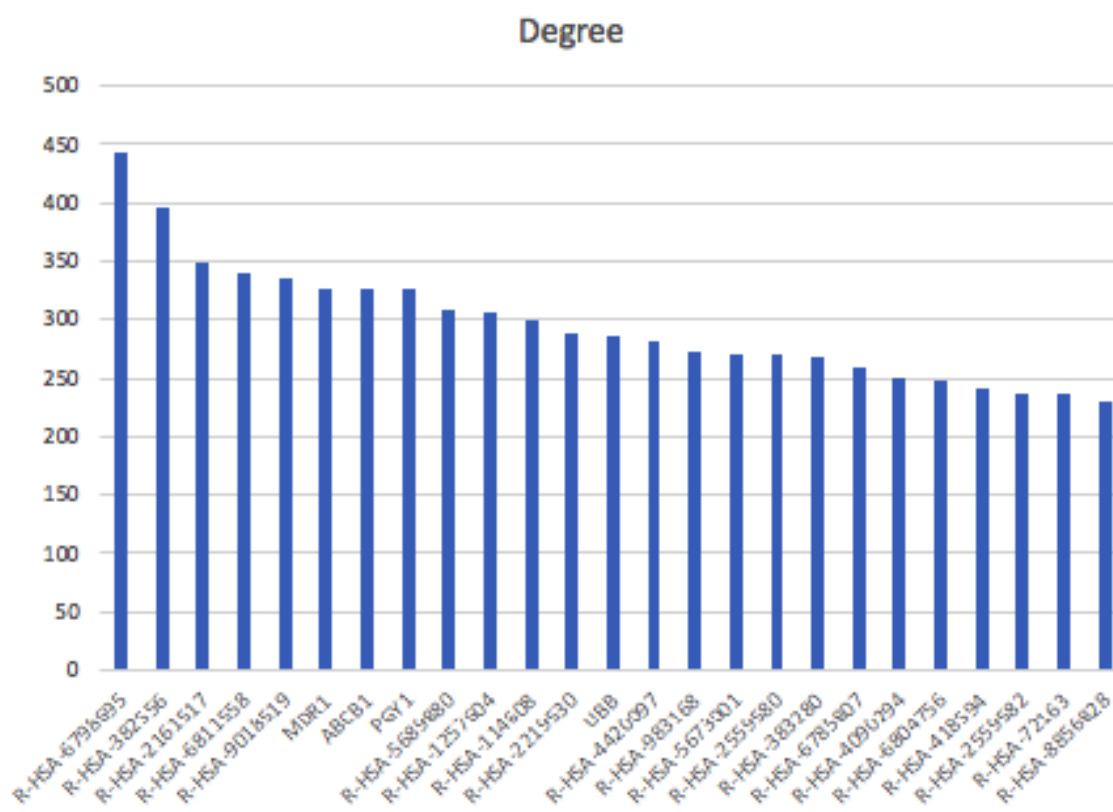


Figure 4.55: Degree for Rhabdomyosarcoma

4.8.0.1 Protein-Protein Interaction Analysis for Rhabdomyosarcoma

Our goal for analyzing protein-protein interactions was to find critical proteins that in rhabdomyosarcoma and PIM substrates and to use these highly essential

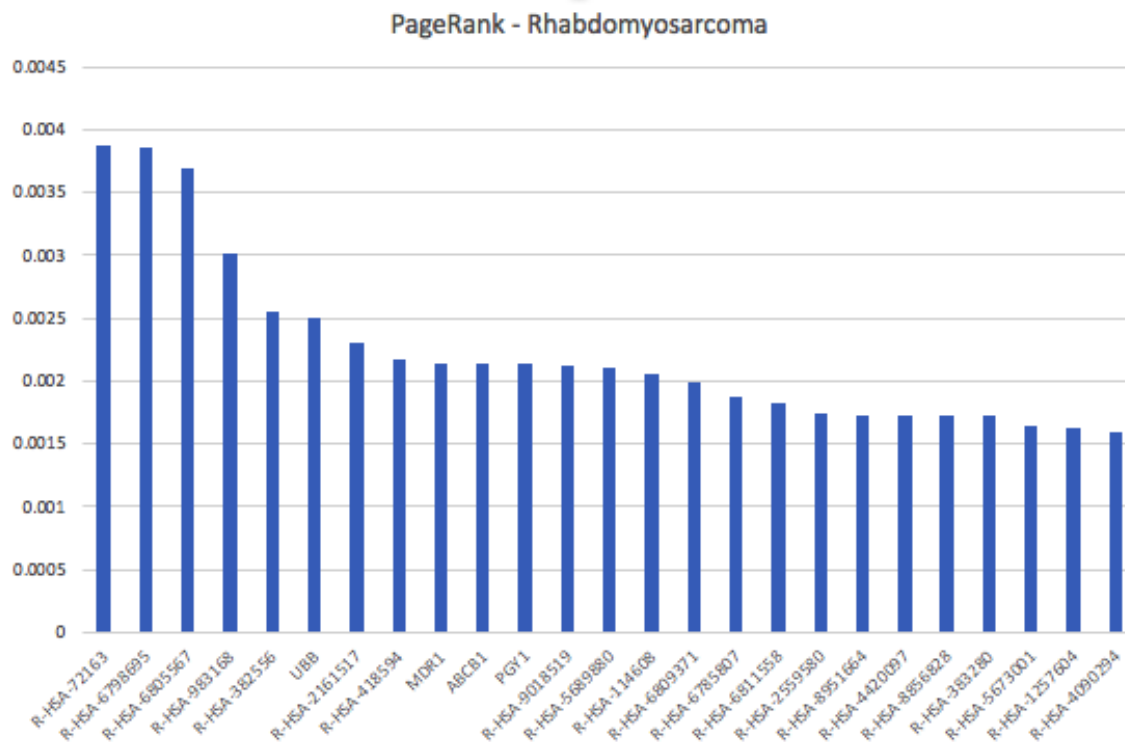


Figure 4.56: Pagerank for Rhabdomyosarcoma

proteins to find biological regulatory pathways.

Rhabdomyosarcoma (protein-protein interactions):

The total number of nodes for the source proteins are 632, and the total number of nodes for the target proteins are 3115, and the total number of interactions is 5785.

The first proteins or the source proteins all fall within the same clusters showing that a majority of these proteins interact with similar proteins.

The second sets of proteins have distinct clusters in blue and two other groups.

In the distinct blue cluster, we found IKKB_HUMAN to be based on PROSITE a member of Protein Kinase dom and a member of Protein Kinase ST. We also found that NUA1_HUMAN is a member of Protein_Kinase dom, Protein Kinase ATP,

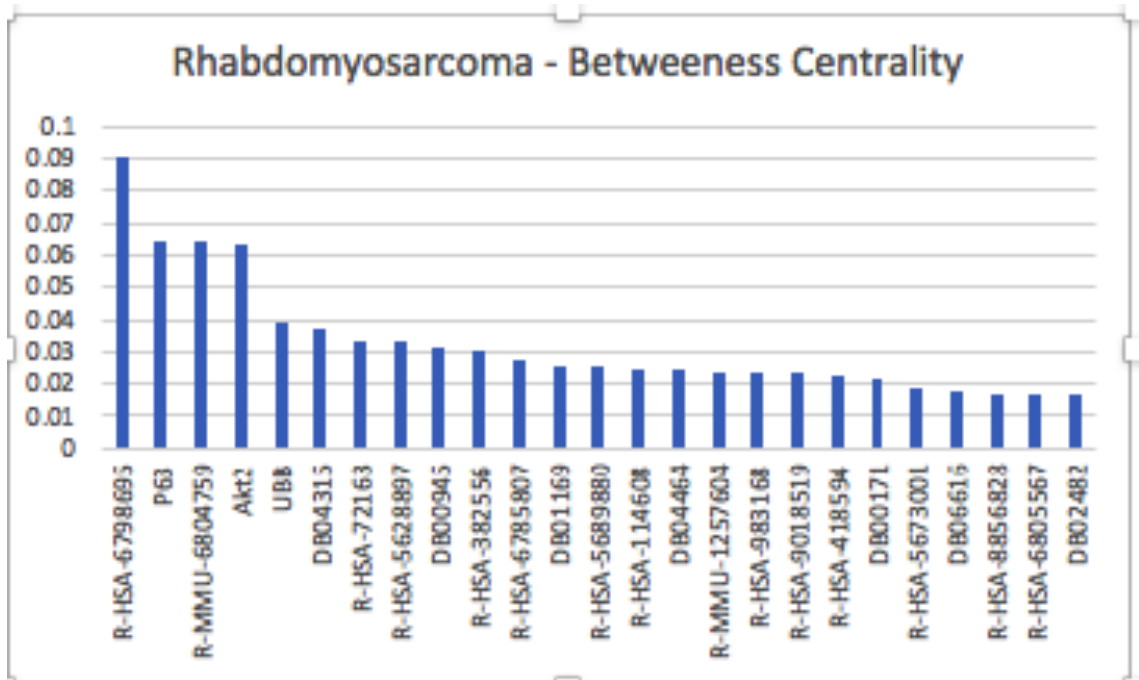


Figure 4.57: Betweenness Centrality for Rhabdomyosarcoma

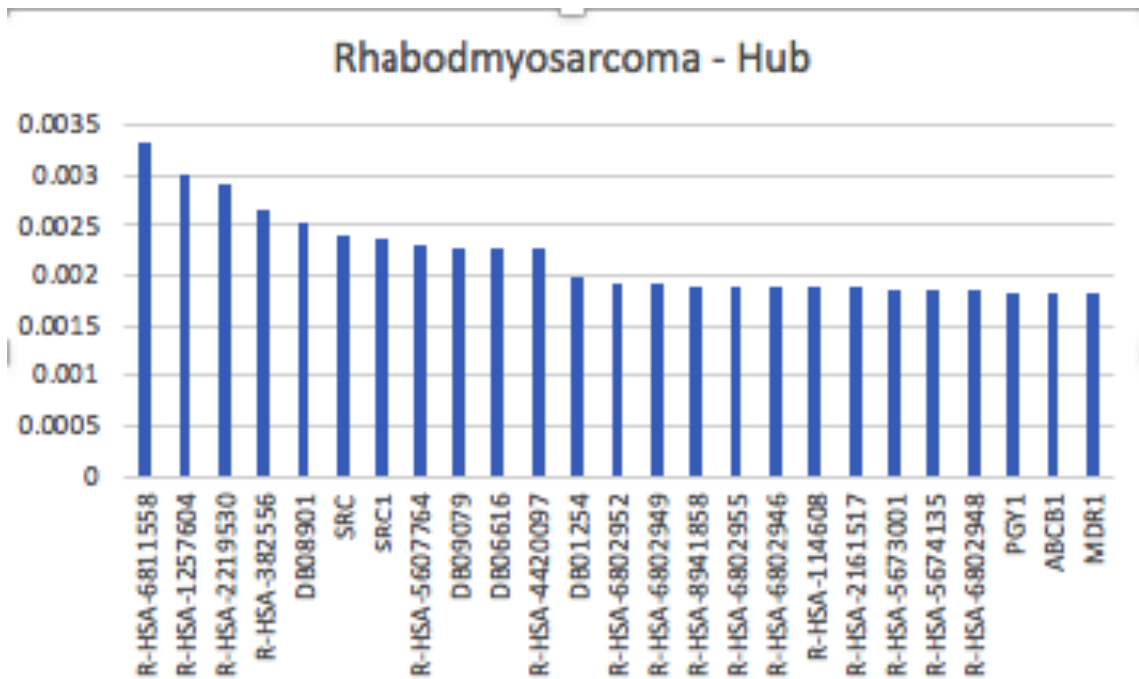


Figure 4.58: Hub for Rhabdomyosarcoma

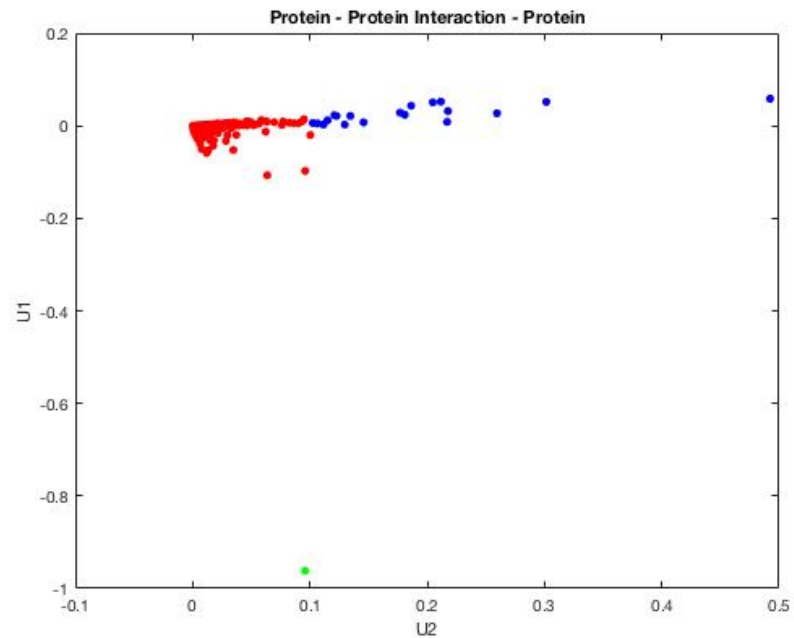


Figure 4.59: protein-protein interaction for rhabdomyosarcoma: Source clustering

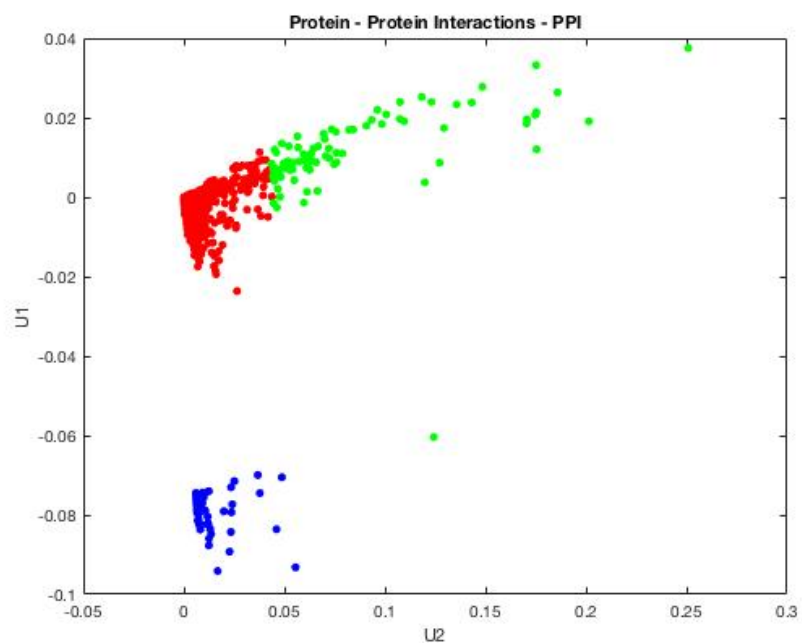


Figure 4.60: protein-protein interaction for rhabdomyosarcoma: Target clustering

and Protein Kinase ST. We also found PLK1_HUMAN is a member of Protein Kinase ATP, Protein Kinase DOM, and Protein Kinase ST. The protein kinase ATP implies that they have an ATP binding region signature, the protein kinase st means that they have serine and threonine protein kinases active site signature and protein kinase dom implies that they belong to the protein kinase domain profile. We had various other new proteins that were identified by our method. The IKKB_HUMAN is linked or connected to sulfasalazine which is a candidate found to treat rhabdomyosarcoma.

4.8.0.2 Rhabdomyosarcoma Phosphorylated Protein Extraction.

PhosphoSitePlus contains 29,082 non-redundant sites on 14, 256 non-redundant proteins and over 90% of the protein sites are from human and mouse. The database also contains phosphorylation sites, acetylation sites, and ubiquitination sites. Our work was focused on the extraction of the rhabdomyosarcoma phosphorylation site from the PhosphoSitePlus database. The database contained gene name, protein, UniProt accession number, the residue position and letter, and the protein sequence. For our research, we were interested in the proteins and genes the sequence were attached. The rhabdomyosarcoma protein database contained 948 proteins and 945 genes. The 948 proteins were used to create our integrated multimodal network.

4.8.0.3 Rhabdomyosarcoma Network generation.

We integrated 108208 biological interactions from different sources which included CTDBASE, DISGENET, UniProt, DRUGBANK and PHOSPHOSITEPLUS to form the rhabdomyosarcoma network. The network contained 21021 nodes and 108208 edges with the average degree 9.686. The network shows a secure connection because we have more edges than the nodes. Table 4.9 shows more of the network properties and illustrates the strength of the connection of the network.

Table 4.9: Network Properties of Phosphorylated Protein in Rhabdomyosarcoma

Network Properties	Network Scores
Nodes	21021
Edges	108208
Average Degree	9.686
Network Diameter	16
Path Length	5.888
Connected Components	201

4.8.0.4 Identification of high influence nodes in protein phosphorylated in Rhabdomyosarcoma.

The next steps will be to utilize different measures in identifying nodes that are highly influential using various centrality measures on our integrated multimodal protein phosphorylated in rhabdomyosarcoma. We measured the level of influence by using: Degree centrality, closeness centrality, betweenness centrality, HITS algo-

rithm, and PageRank. The degree of a specific node is the number of neighbors that the node has. The degree centrality of the graph counts merely how many neighbors a node has. If the network is directed, we will have outdegree and in degree. Figure 2 is the degree of the undirected graph. The closeness centrality is derived by finding the mean distance from one node to the other nodes in a network. The betweenness centrality can be obtained by measuring the degree to that a node lies on the path between another node. The nodes with high betweenness centrality will have a very high influence on the network. They have high control over the information passing through them. The HITS algorithm consists of both authority and hub scores. The nodes with high authority scores are the nodes linked from nodes that are known as a hub. The nodes with high hub scores are the nodes that are connected to nodes that are considered to be authorities. The result below shows the top 10 influence nodes of all the centrality measures. The Pagerank algorithm intuition is that links are synonymous with votes for a node's relevance, the more the votes a node has, the more the important it is. The votes from connected nodes will have more than votes from irrelevant ones. The figures below show the result from centrality measures [130].

The node with the highest degree in figure 2 is called DBO8901. The DB08901 represents ponatinib. In 2016, [131] showed that ponatinib is currently on a clinical trial for treating rhabdomyosarcoma. [131] targeted activated FGFR4 in rhabdomyosarcoma with the inhibitor ponatinib(AP24534). The next drug with the highest degree in figure 4.61 above is DB08896 called Regorafenib. [132] Showed that this drug was considered an inhibitor that can be studied for its anti-tumor ac-

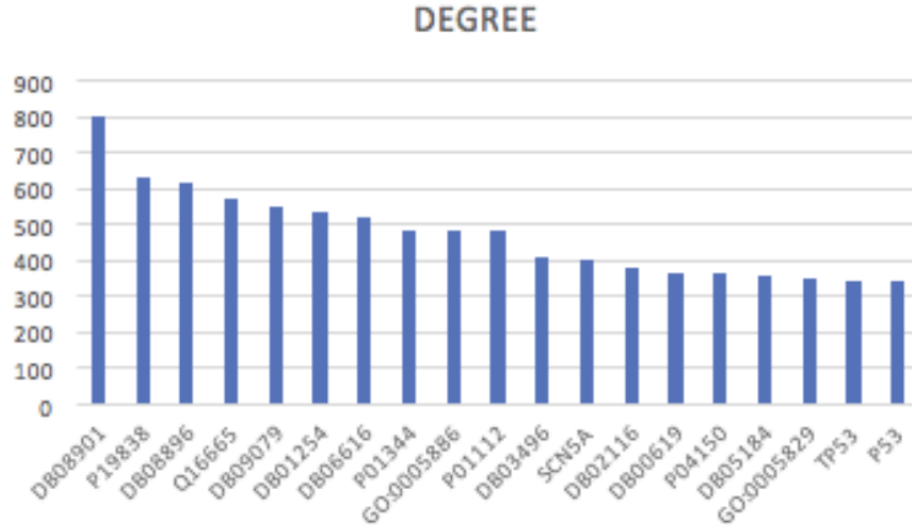


Figure 4.61: Degree of multimodal phosphorylated protein in rhabdomyosarcoma.

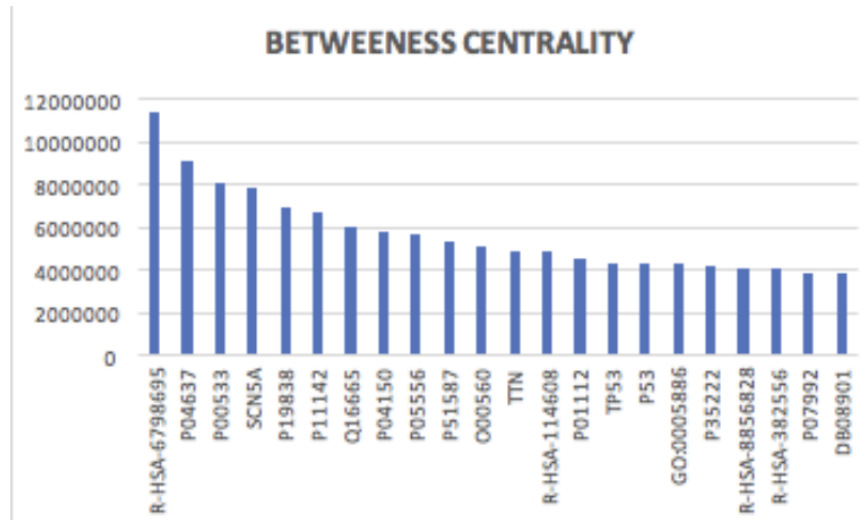


Figure 4.62: Betweenness centrality of multimodal phosphorylated protein in rhabdomyosarcoma.

tivities. These two examples showed using the phosphorylated proteins as the basis to create an integrated network helped to highlight the critical aspects of the study of rhabdomyosarcoma. This helps us pay attention to the relevant aspects of the network. P19838 NF-kappa-B with gene name NFKB1. [133] suggests that abnormal expressions may contribute to myogenesis and rhabdomyosarcoma. DB080901

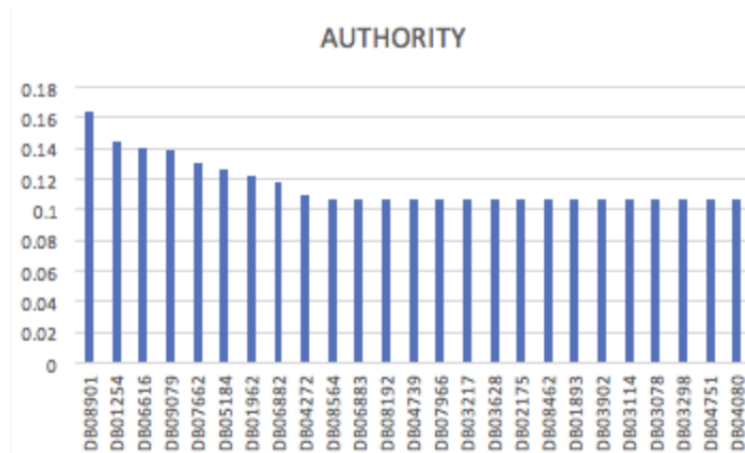


Figure 4.63: Authority of multimodal phosphorylated protein in rhabdomyosarcoma.

appears again to be a high influence node on the betweenness centrality graph.

[134] shows the p53 mutations are frequently detected in young children with rhabdomyosarcoma. Reversing the mutation will reduce the risk of having the disease. The results from the authority scores show that drug nodes have the most authority in our datasets. The same DB080901 was the drug with the highest influence for authority scores as well. TTN was found to be activated in 3 tumor samples [135].

The betweenness centrality results show RHSA-6798695 as the highest betweenness centrality score (see figure 4.62). R-HSA-6798695 is referred to as neutrophil degranulation. [136] indicates that neutrophil degranulation had some anti-tumor functions with regards to rhabdomyosarcoma.

The authority scores also show that DB01254(dasatinib) was the next highest influence node. The result showed that a majority of nodes with high influence in (figure 4.63) were drugs. This implies that a majority of the drugs with high authority are linked from other nodes that are known as hubs.

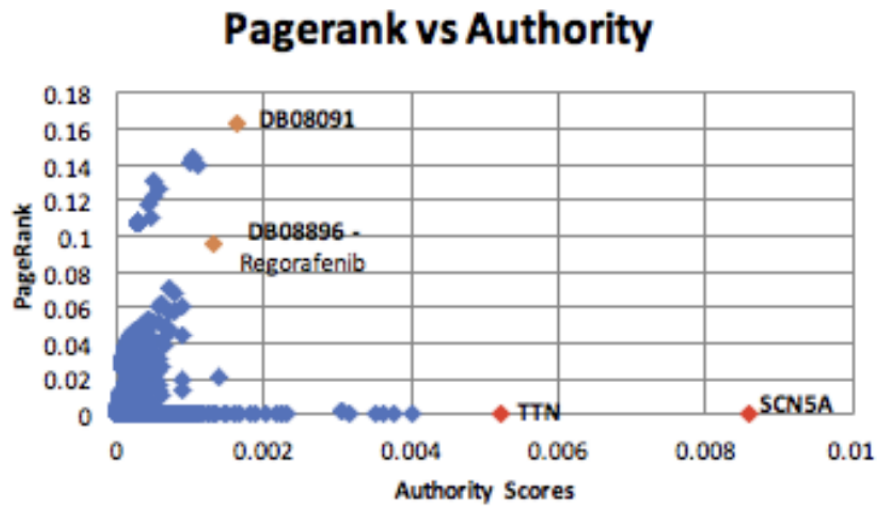


Figure 4.64: The score above show Pagerank vs Authority scores indicating receptor to effector score.

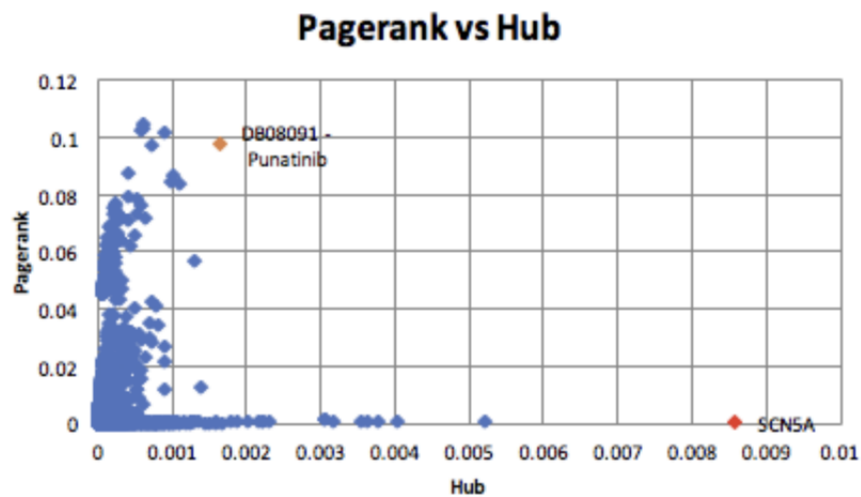


Figure 4.65: The score above show Pagerank vs Hub scores indicating receptor to effector score.

We plotted the PageRank vs. hub score and the page rank vs. authority score. The results show that most nodes are clustered around the origin. The results from the Pagerank vs. Hub shows that some of the sequences were clustered around receptor which means these nodes are nodes that are prone to be influenced by a large number of other nodes. DB08896 in (see figure 4.64) shows that regorafenib

is influenced by and influences other nodes in the network.

4.8.0.5 Binning the nodes by the types:

These sections above observed the results while all the node types are combined to view the node types independently of each of other. We rank the nodes by types below using degree centrality measure to find the topmost influence nodes for each of the influence types. The degree scores then cluster the nodes. The result of binning the drugs still gives us the DB08901(Ponatinib) as the first influence node. The top 5 drugs include Ponatinib, Regorafenib, nintedanib, dasatinib, bosutinib, and alvocidib. All of the drugs are currently being used to treat rhabdomyosarcoma or clinical trial phase.

Table 4.10: Top 5 influence nodes for node types(protein, drugbank, pathway & SNP)

Protein	DrugBank	Pathway	SNP
P19838	DB08901	R-HSA-6798695	rs747622981
P01344	DB08896	R-HSA-6811558	rs7766641
P01112	DB09079	R-HSA-4420097	rs9311651
P04150	DB01254	R-HSA-2559580	rs6778837
P53	DB06616	R-HSA-5673001	rs6445902

Other node types include protein, SNP, drug, and pathway. The top 5 most influential protein nodes: P19838, P01344, P01112, P04150, and P53. p53, for example, is a well-known tumor protein. P01344 is associated with Wilms a tumor, Beckwith-Wiedemann syndrome, rhabdomyosarcoma, and Silver-Russell syndrome. The top 5 influential pathways are R-HSA-6798695, R-HSA-6811558, R-

HSA-4420097, R-HSA-2559580, R-HSA-5673001. The top pathway list can also be used to find more useful candidates that will treat the rhabdomyosarcoma.

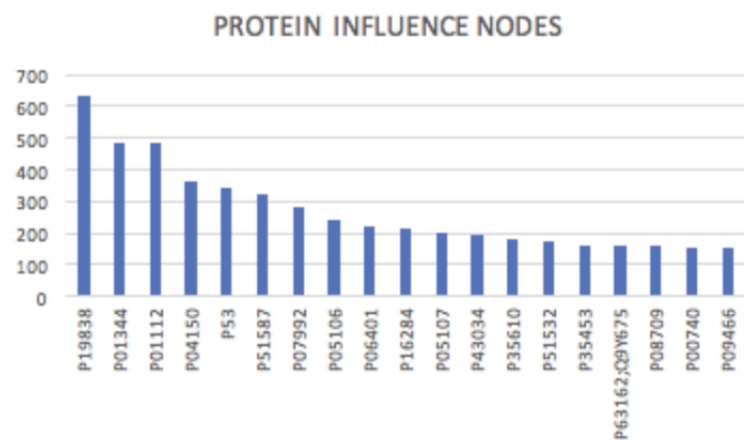


Figure 4.66: Protein binned influence nodes.

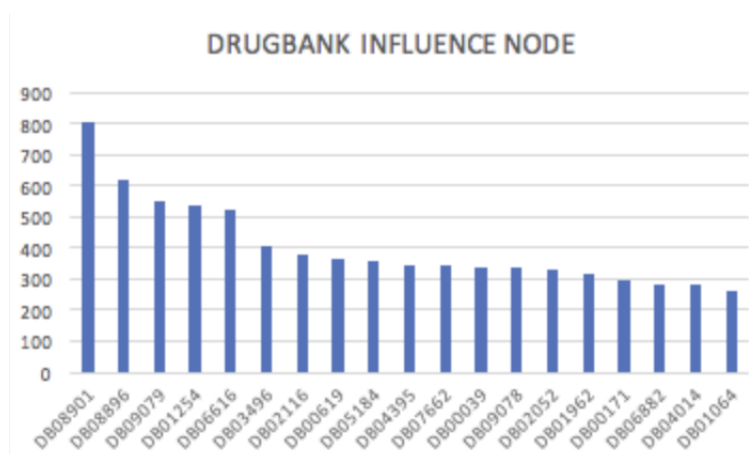


Figure 4.67: Drug binned influence nodes.

4.9 NetAnaPhoS-Network Analytics for Diabetes Mellitus Phosphorylated Protein Network

In this section, we analyzed diabetes phosphorylated protein network using the various centrality measures. The nodes of the network are 277, and the edges

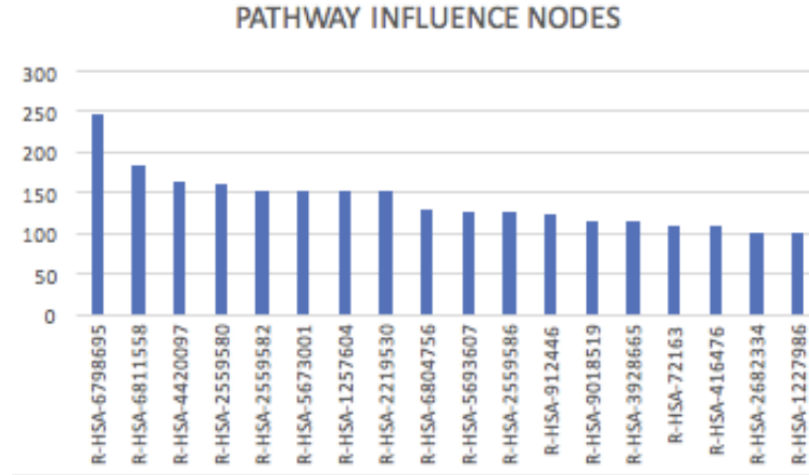


Figure 4.68: Pathway binned influence nodes.

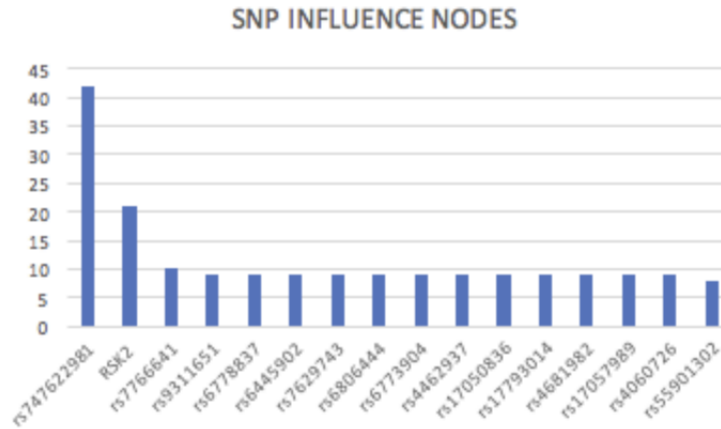


Figure 4.69: SNP binned influence nodes.

are 486. The average degree is 1.755, the network diameter is 3, and the graph density is 0.006.

Indegree: The Indegree shows IRS1, AKT1, and AMPK. We have selected the first 25 nodes that rank the highest using the indegree.

Outdegree: We have selected the first 25 nodes that rank the highest using the outdegree. The high-fat diet is connected to a high number of other disease-related elements. We can take a high-fat diet and find the subgraph of the high-fat

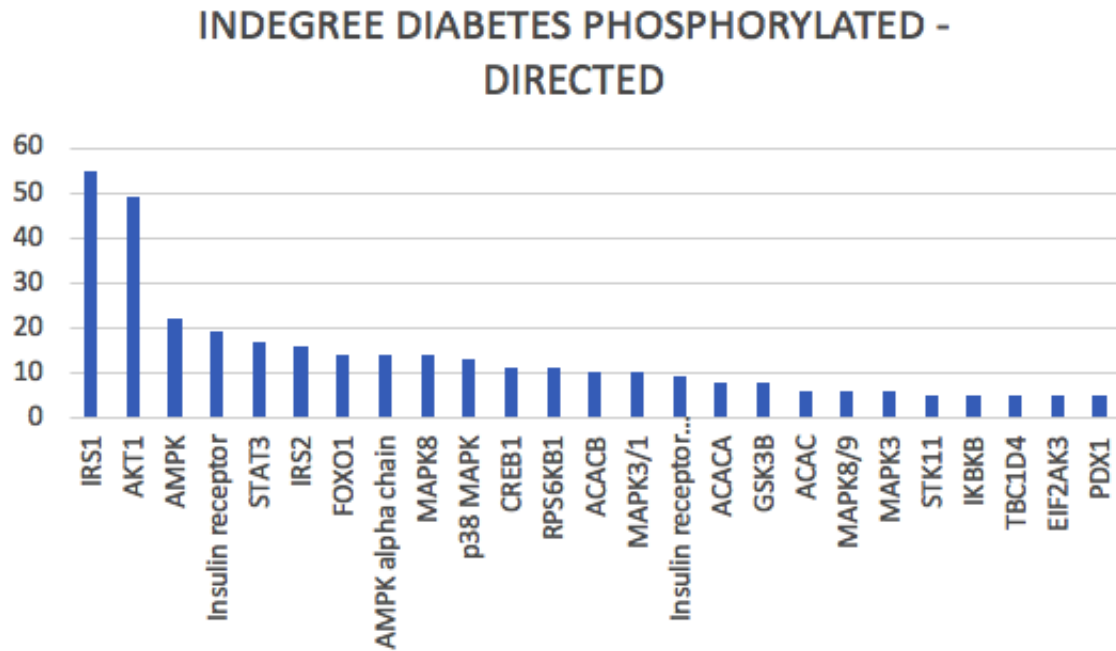


Figure 4.70: In-degree for diabetes phosphorylated network.

diet to see if we can find a possible drug candidate.

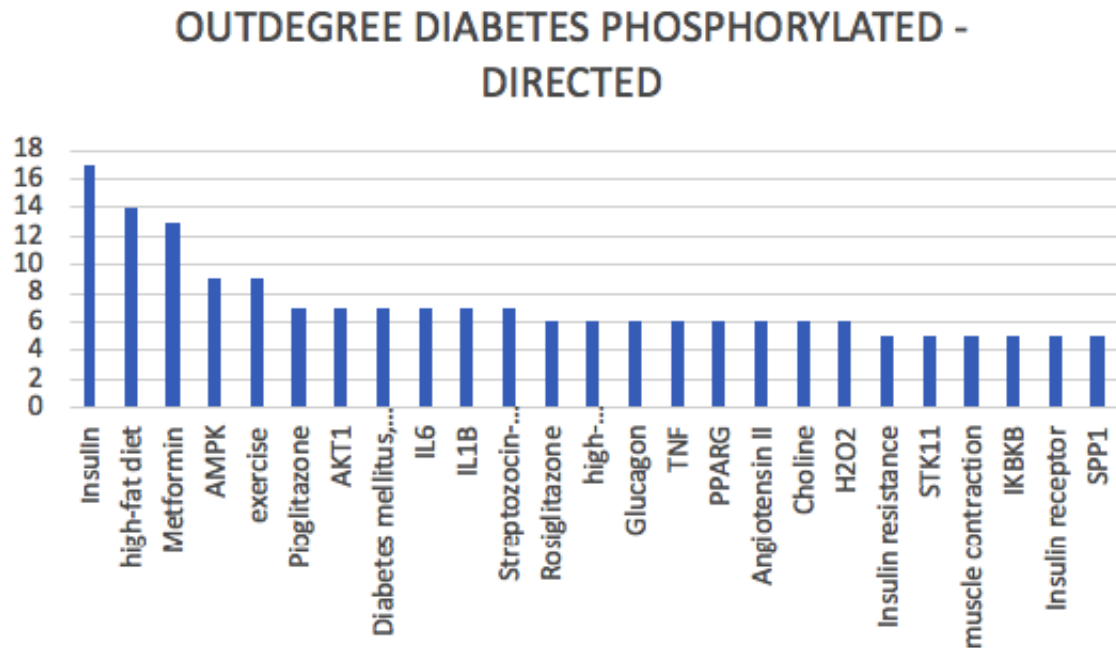


Figure 4.71: Out-degree for diabetes phosphorylated network.

PageRank: The PageRank helps to examine the quality of links to a web

page. The result shows the phosphorylated network nodes scores for PageRank given the entire phosphorylated network.

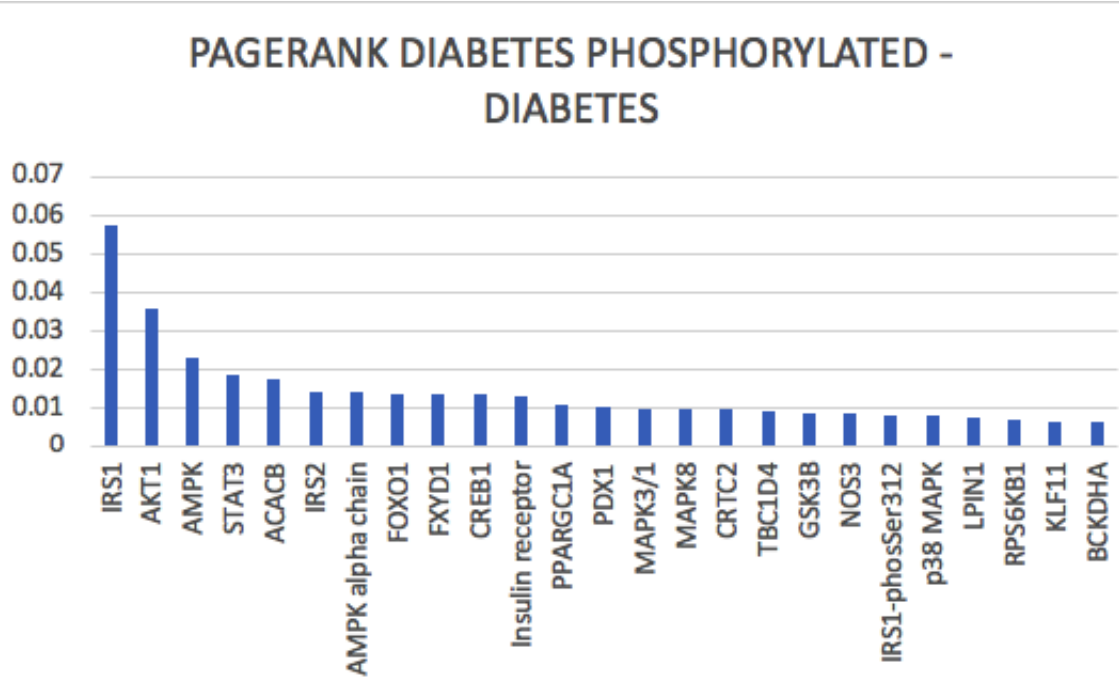


Figure 4.72: PageRank for diabetes phosphorylated network.

4.9.0.1 Perturbation Results for Diabetes Mellitus Phosphorylated Network

We have the results from our experiment using serial and sequential perturbation technique. We have studied the PageRank, degree, authority scores and betweenness centrality for both single and sequential perturbation.

The nodes are 277 and edges are 486. We applied single and sequential perturbation on the datasets. The singular perturbation indicates that you take the node with the most significant score for each of the centrality measures, and you remove them either one by one without replacement or one by one with replacement to find

the measure of how distorted the graph currently is. The more disconnected the graph is, the more the node is. We used the connected components to determine the impact of the node on the graph.

The Betweenness centrality results for serial and sequential perturbation using Diabetes mellitus phosphorylated elements.

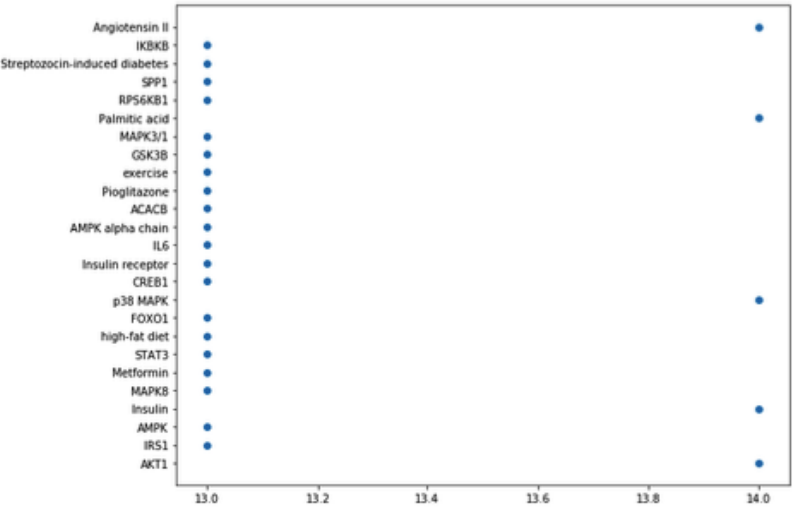


Figure 4.73: Single Perturbation using Betweenness Centrality Measures on Diabetes Mellitus Phosphorylated Network

The next measure we tested the single and sequential perturbation approach on was the PageRank.

Comparing the combined diabetes network to the diabetes mellitus phosphorylated network, we observe a noticeable difference in the elements or nodes of the network. This is due to the separation of the phosphorylated version from the unphosphorylated version. We also observed using sequential perturbation produce more significant gaps with regards to the disconnectedness of the graph. We have specially selected the betweenness centrality and PageRank measure to demonstrate

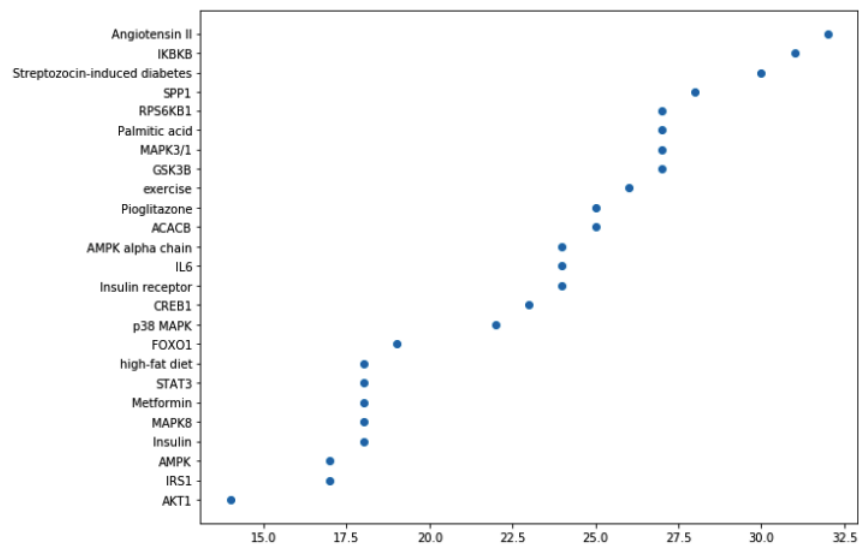


Figure 4.74: Sequential Perturbation using Betweenness Centrality Measures on Diabetes Mellitus Phosphorylated Network

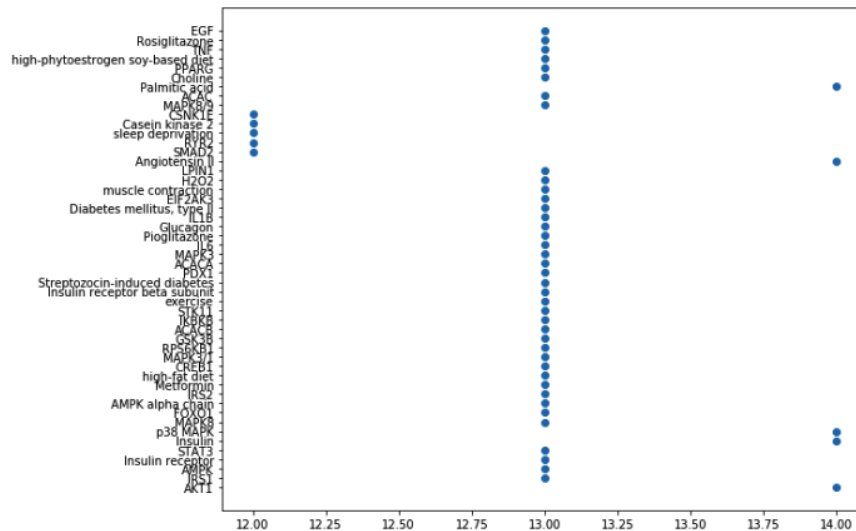


Figure 4.75: Single Perturbation using PageRank on Diabetes Mellitus Phosphorylated Network

this difference.

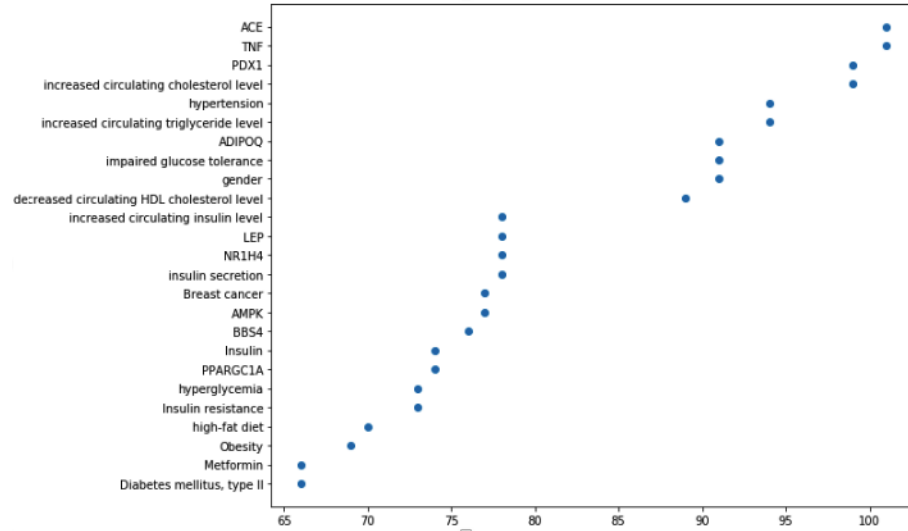


Figure 4.76: Sequential Perturbation using Betweenness on Diabetes Mellitus Network

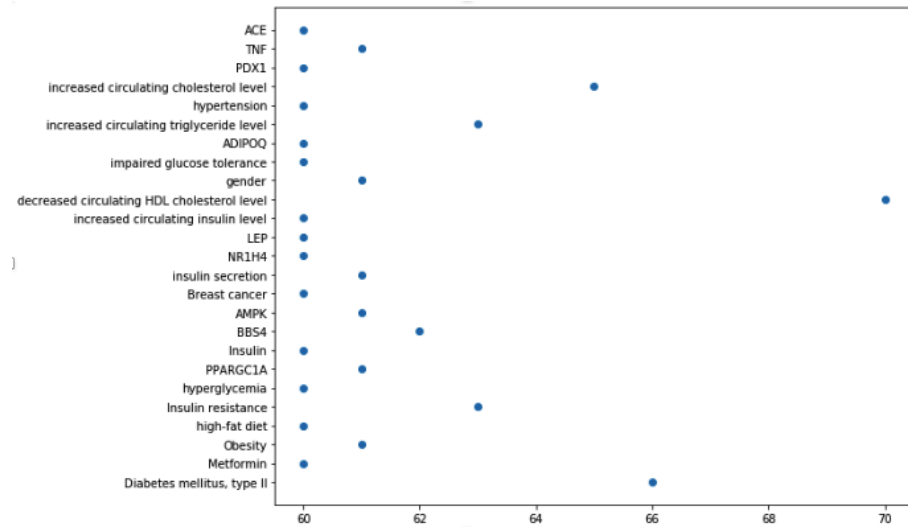


Figure 4.77: Single Perturbation using Betweenness on Diabetes Mellitus Network

4.9.0.2 Results from Network Analytics for Diabetes Mellitus Phosphorylated Protein Network

Our result generated some promising candidates; we confirmed these results by using PubMed and clinical trials to verify their validity. The increase of proin-

inflammatory cytokines such as IL-6 and TNF-alpha has been associated with over-nutrition or obesity and type 2 diabetes [137]. Also for LPIN1, mRNA expression showed low levels of LPIN1 in individuals who are obese and also in individuals with type 2 diabetes [138]. The AMPK is a protein kinase that has been known to regulate glucose homeostasis in the body. Thus having a long association to type 2 diabetes [139]. Studies have also shown a significant link between different genetic variants of PPARGC1A and type 2 diabetes [140]. For AKT1, it plays a role in insulin signaling, and this has been associated with type two diabetes [141]. STK11, A gene variant in STK11 has been associated with high levels of insulin. Thus suggesting that it has a link or a connection to diabetes two [142]. MAPK3 has also been associated with the treatment of Type 2 diabetes [143].

4.10 DREiM: Multimode Clustering on Rhabdomysarcoma

The ring structure described in the methodology section depicts the arrangement of three different connection modes that form a tri-mode. Multimode networks involve multiple types of entities; bi-mode is a more abbreviated version of a multi-mode. An example of a bi-mode will be a relationship between pathway and drug. The bi-mode connections will contain edges between pathway and drugs but no edges between the pathway and no edges between drug.

For the clustering of the tri-mode, we use protein, pathways, and drugs to create tri-mode. We created matrix A_1 , A_2 , and A_3 . The matrix below was used to create the adjacency matrix and to create the Laplacian matrix. The figure below

shows the clustering results for the matrix.

$$A_1 = \begin{bmatrix} u_1 t_1 & u_1 t_2 & \cdots & u_1 t_7 \\ u_2 t_1 & \ddots & \cdots & u_2 t_7 \\ \vdots & \vdots & \ddots & \vdots \\ u_9 t_1 & u_9 t_2 & \cdots & u_9 t_7 \end{bmatrix}$$

The matrix represents protein and pathway. The protein is U while the path represents the t. It represents the first bi-mode.

$$A_2 = \begin{bmatrix} t_1 v_1 & t_1 v_2 & t_1 v_3 \\ \vdots & \ddots & \vdots \\ t_7 v_1 & t_7 v_2 & t_7 v_3 \end{bmatrix}$$

The matrix represents the next bi-mode in the ring structure which is the pathways and the drug. The drug is representing v.

$$A_3 = \begin{bmatrix} u_1 v_1 & u_1 v_2 & u_1 v_3 \\ \vdots & \ddots & \vdots \\ u_9 v_1 & u_9 v_2 & u_9 v_3 \end{bmatrix}$$

The matrix represents the last portion of the ring structure. The equation below shows the construction of the new matrix for decomposition.

$$L = \begin{bmatrix} D_1^{(9 \times 9)} & A_1^{(9 \times 7)} & A_3^{(9 \times 3)} \\ A_1^{T(7 \times 9)} & D_2^{(7 \times 7)} & A_2^{(7 \times 3)} \\ A_3^{T(3 \times 9)} & A_2^{T(3 \times 7)} & D_3^{(3 \times 3)} \end{bmatrix}$$

The next step will be to apply eigenvalue decomposition on the L matrix and utilize the k-means algorithm to find the clusters and to select which of the clusters contain the highly influential nodes.

4.10.0.1 Ideal-k for rhabdomyosarcoma for multimode clustering

We used ideal-k to determine how many clusters to use to classify our rhabdomyosarcoma datasets (see figure 4.78). We have selected ideal-k to be 3 for the multimode network (DReiM) for rhabdomyosarcoma (gene-drug-pathway)

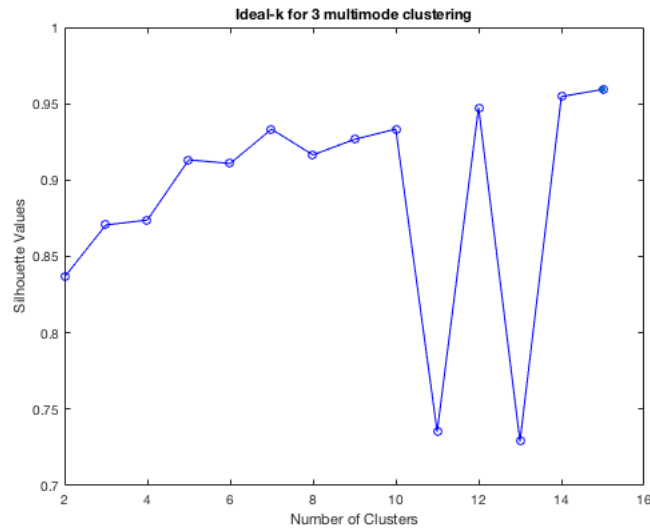


Figure 4.78: Ideal-k rhabdomyosarcoma for multimode clustering.

4.10.0.2 Clustering results for rhabdomyosarcoma - DREiM

The figure below shows the clusters of rhabdomyosarcoma datasets. The results below show the gene-drug-pathway spectral clustering results using DReiM (Multimode Clustering)(see figure 4.79). The figure shows red, blue, and green clusters.

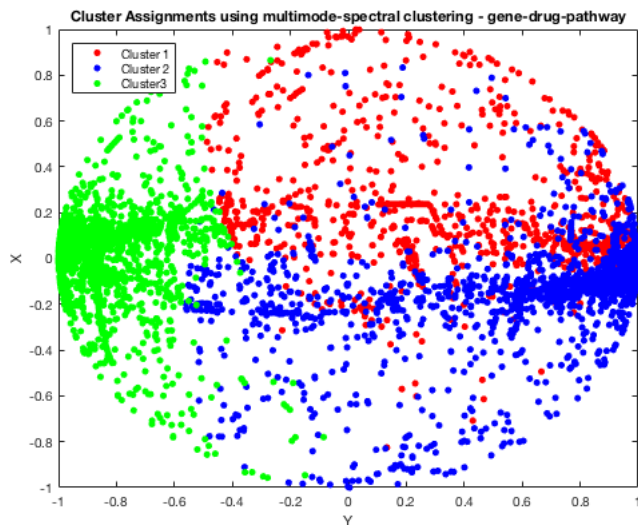


Figure 4.79: Clustering results rhabdomyosarcoma for multimode clustering.

4.10.0.3 Drug Candidates for Multimode Clustering(DReiM)

We check to make sure all the 5 Top Influential Drugs were in the same cluster. The top 5 drugs include Ponatinib, Regorafenib, nintedanib, dasatinib, bosutinib, and alvocidib were all found in the same cluster using DReiM method.

4.10.1 DReiM: Multimode Clustering on Diabetes Mellitus Phosphorylated Protein Network

We have applied DReiM to our Diabetes Mellitus datasets using 4-mode. The results below show the ideal-k for our results, and we have studied the cluster that not only contained the high influence node but also belong to the cluster that contained the 98% precision. Some of the results are discussed in this section.

Nebivolol was shown to improve blood glucose [144]. Besides, ionomycin has been shown to lower glucose in diabetic mouse models [145]. In a clinical trial,

subjects who used valsartan for five years reduced the risk of diabetes by 14% [146]. Metformin is a drug used to treat individuals with type 2 diabetes [147]. Oleic acid helps to prevent type 2 diabetes in people, and it has antidiabetic effects [148].

4.11 Future Work: Gene expression using K562 cells

The CEL files were extracted from NCBI and parsed into the gene pattern tool to create the .gct files. The .gct files were further utilized and alongside with cls file (contains the class files for the files in .gct files). The result was used to perform the GSEA Analysis. The method checks that a priori defined a set of genes shows statistically significant, consistent differences between two biological states (e.g., phenotypes).

4.11.1 GSEA Analysis

The .cel files containing the gene expression data were uploaded into GSEA, and the class file containing the control and the non-control version was uploaded as well. The gene set database was selected, and we utilized the transcription factor gene sets so that we can identify the transcription factors that play a significant role and further use this transcription factors for pathway analysis. We permuted 1000 times to get a more accurate p-value. The chip platform utilized was Affymetrix. 0 represents the k562 creb control, and one represents when k562 when creb cells are knocked out. 86 out of 572 gene sets are upregulated for control for sample phenotype 0. 486/572 gene sets are upregulated in phenotype 1 for the knocked out

version. The gene sets filter 43 out of 615 gene sets, and only 572 gene sets were used for the analysis.

The data sets we used had 20,606 genes. The gene sets that are significantly enriched at nominal p-value $\leq 5\%$ were utilized. The gene sets are considered the positive enrichment gene sets. The first result of the gene set enrichment analysis is the enrichment score (ES), which indicates the extent to which a set of genes is overrepresented at the top or bottom of a ranked list of genes.

The results from the gene sets were further extracted to find specific genes that belong to this genesets. The genes were then used to perform pathway analysis. The Reactome package in R was used to perform a pathway analysis using the Reactome pathway database. It implements the gene set enrichment analysis and enrichment analysis. Our results from the pathway analysis were utilized to extracted drugs that target this pathway.

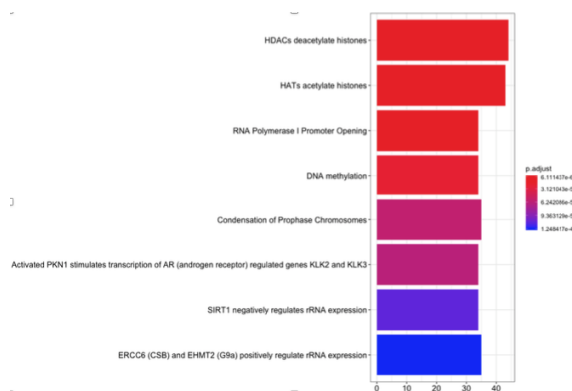


Figure 4.80: Pathway Analysis using Gene expression

4.11.2 Using SVD for Pathway Analysis

We utilized SVD in finding clusters of drugs and pathways that play a role in the treatment of the disease. The clusters show two distinct groups for the pathways and seven distinct groups for the drug. This indicates that clusters with small drugs can be possible targets for drug combination.

We used a term-document matrix set up to create clusters of drug vs. pathway. The goal was to find clusters of drugs and pathways and a combination of drugs and pathway in a cluster.

For Drug vs. Pathway, The ideal-k, we have selected is 3. We had 3-clusters in our result. The figure [4.84](#) shows even more defined clusters.

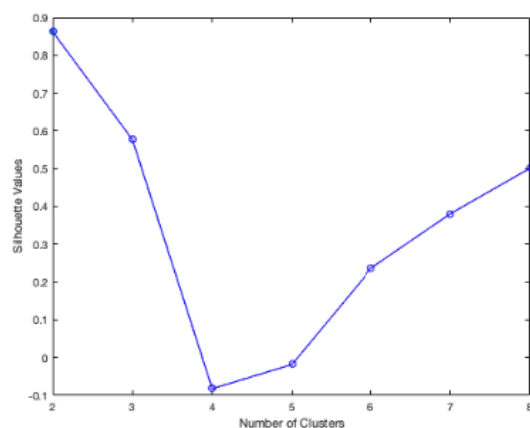


Figure 4.81: Ideal-k using GSEA to discover clustered pathways

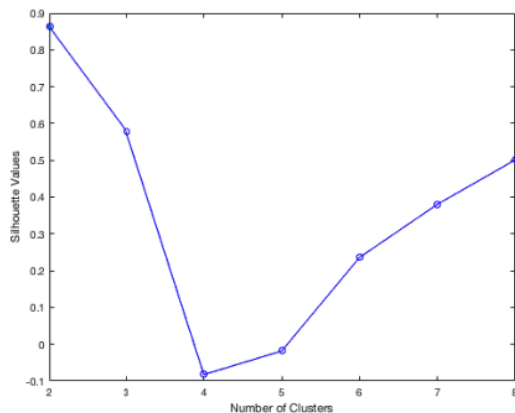


Figure 4.82: Ideal-k using GSEA to discover clustered drugs

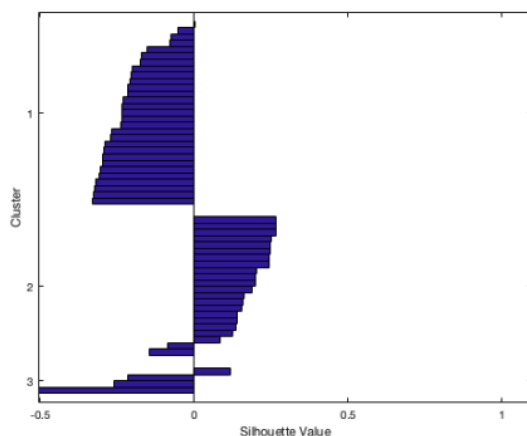


Figure 4.83: silhouette using GSEA to discover clustered drugs

4.12 Validation of our experiment-PIM Substrate

We used degree, authority, and PageRank to select drug candidates. One hundred sixty-eight drug candidates were found by our current algorithm using the Top 50 degree, Top 50 authority, Top 50, and Top 50 PageRank scores. We cleaned the result datasets by removing chemical names that weren't a specific drug name and that left us with 168 drugs. We considered drugs currently in clinical trials

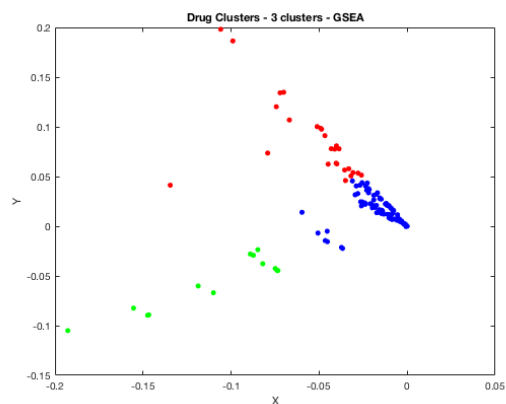


Figure 4.84: Drug Clusters using Gene expression

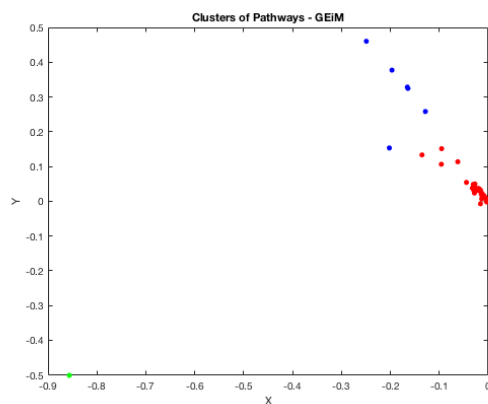


Figure 4.85: Pathway Clusters using Gene expression

phase I, II, III a candidate since they have been selected by top pharmaceutical to go on the trial. We used clinical trial data sets to determine the validity of our results.

The real positives (TP) are cases in which we predicted yes (these drugs treat prostate cancer or are on clinical trials to treat prostate cancer), and these drugs to treat cancer or are on clinical trials to treat prostate cancer. The real negatives (TN): We predicted no, and they don't or can't be repositioned for prostate cancer. The false positives (FP): We predicted yes, but they can't be repositioned for prostate

cancer. (Also known as a "Type I error.") The false negatives (FN): We predicted no, but they can be repositioned for prostate cancer. (Also known as a "Type II error.")

The result is below on the table.

N=168	PREDICTED NO	PREDICTED YES	
ACTUAL NO	TN=27	FP=33	60
ACTUAL YES	FN=6	TP=98	104
	33	131	

Table 4.11: Validation of our experiment

4.12.0.1 Precision

The precision is widely used in classification. The precision measures the exactness of our results. It checks how often the method predicts a correct answer. We used Equation 4.4 to find the precision score and the $score = 0.74$.

$$precision = \frac{tp}{tp + fp} \quad (4.4)$$

4.12.0.2 Recall

The recall checks for the completeness. The number of positive results that were label positive. The recall represents the rate of true positive or sensitivity score. The Equation 4.5 was used to calculate the recall and the score for the $recall = 0.94$.

$$recall = \frac{tp}{tp + fn} \quad (4.5)$$

4.12.0.3 Accuracy

The accuracy shows how well our results classified correctly in the identification and otherwise. The equation for finding accuracy score is below and the *score* = 0.76.

$$accuracy = \frac{tp + tn}{tp + tn + fp + fn} \quad (4.6)$$

The results above show that Recall is 94%, Precision is 74%, and Accuracy is 76%. The precision with 100% that all label classes belonged to the class they were a label. The precision doesn't tell us anything about results that were mislabeled or misclassified. The recall, on the other hand, doesn't tell us anything about the mislabel classes. F-measure is used to combine both precision and recall and the F-score = 82.8%.

$$accuracy = \frac{2 * precision * recall}{precision + recall} \quad (4.7)$$

4.13 Validation of our experiment for Rhabdomyosarcoma

4.13.0.1 Degree

The results from the degree show the calculation of the degrees for each node (gene, SNP, drug, disease, etc.) in the network. The scores from the degree were used to find clusters by applying the k-means clustering on the degree scores. The degree was clustered into three groups using the k-means clustering. The result from

the k-means clustering was compared with other results using other methods. The performance of the degree is below.

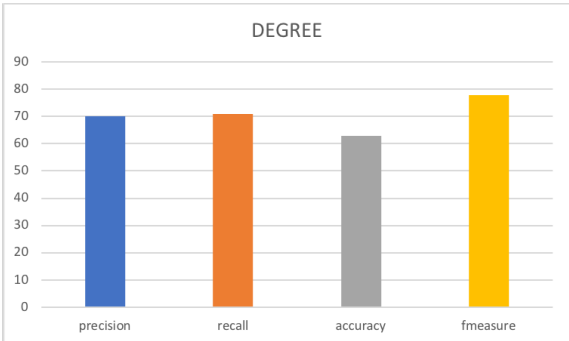


Figure 4.86: Accuracy for Rhabdomyosarcoma for Degree

4.13.0.2 Betweenness

The results from the Betweenness centrality shows the calculation of the Betweenness score for each of the nodes in the network. The scores for each of the nodes was used to find the clusters. The K-means clustering algorithm was used to find groups of nodes. The silhouette score was used to select the best cluster. The Betweenness centrality result performance is below:

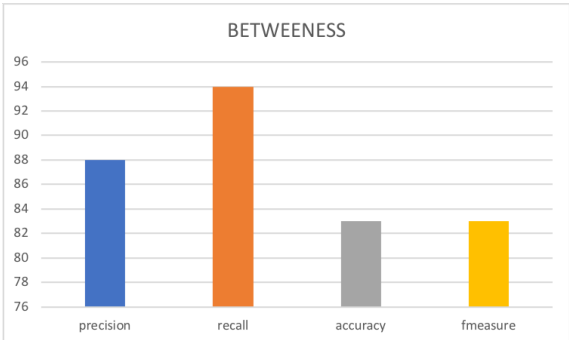


Figure 4.87: Accuracy for Rhabdomyosarcoma for Betweenness

4.13.0.3 Page-rank Scores

The PageRank algorithm was also utilized to find essential nodes. The scores from the PageRank were further used to see groups of clusters using the k-means algorithm, and the results of the algorithm were compared with the algorithm of the multimode network.

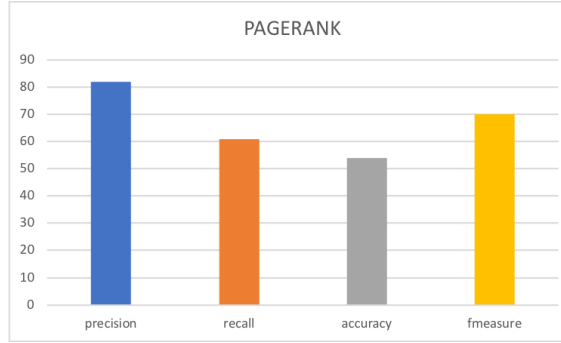


Figure 4.88: Accuracy for Rhabdomyosarcoma for Pagerank

4.13.0.4 Neural Network using Autoencoder

The autoencoder is a neural network that is used for unsupervised learning. The autoencoder applies backpropagation [149] [97] . Therefore making the value that is targeted the same as the value that is inputted. The mathematical representation is $a^{(i)} = b^{(i)}$.

Given that, we have an example of an unlabeled set of training vectors $\{b^{(1)}, b^{(2)}, b^{(3)}, \dots\}$, where $b^{(i)} \in \mathbb{R}^n$.

The autoencoder's ambition is to try to learn a function $h_{W,t}(b) \approx b$. It learns this function to produce a similar output to the original b . It imposes some

constraints on this matrix or network to uncover novel structures about the data. The hidden layers will further help to uncover this structure. The autoencoder does quite similar to the PCA which is to learn in low dimensional spaces. To demonstrate with an example, if you had a 15 x 15 image. The number of pixels for this image will be 225. We can have p_i represent the number of pixels, where $p_i = 225$ pixels. If there are only, $hl = 20$ layers in between where hl represents hidden layers and the output layer is the same as the input layer. It tries to reconstruct the 225 pixels. The input vector is given a function to compress it $ac^{(2)} \in \mathbb{R}^{50}$. The function is also known as the activation function. The activation function formula is written below:

$$\hat{\rho}_j = \frac{1}{m} \sum_{i=1}^m \left[h_{ij}^{(2)}(b^{(i)}) \right] \quad (4.8)$$

The neuron is considered either active, if the values of its output are closest to 1 or not active if its values are closest to 0.

The $h_{ij}^{(2)}$ implies the activation of the hidden unit of j for the autoencoder. $h_{ij}^{(2)}(a)$ gives the activation function when specific input is given in this case a . The activation function across all the hidden unit is average, and this constraint is enforced on the data.

$$\hat{\rho}_j = \rho, \quad (4.9)$$

The expectation is that the average of each of the hidden neurons will be close to 0.05 to meet the requirement of the constraints. To be able to accomplish this a penalty term is applied to optimize that penalizes $\hat{\rho}_j$ from deviating significantly ρ . The penalty choice below was chosen to give a significant result.

$$\sum_{j=1}^{s_2} \rho \log \frac{\rho}{\hat{\rho}_j} + (1 - \rho) \log \frac{1 - \rho}{1 - \hat{\rho}_j}. \quad (4.10)$$

The formula or concept above is based on KL divergence.

$$\sum_{j=1}^{s_2} \text{KL}(\rho || \hat{\rho}_j), \quad (4.11)$$

The overall cost function will result in the formula below:

$$J_{\text{sparse}}(W, b) = J(W, b) + \beta \sum_{j=1}^{s_2} \text{KL}(\rho || \hat{\rho}_j), \quad (4.12)$$

In our experiment, the adjacency matrix was used to build a clustering model using a neural network. The auto-encoder and decoder model was fully built symmetrically. The hyperparameters and loss function was defined, and the model was fit to our datasets. The weights of the trained model were saved and utilized to build clustering layers. The input features are converted into soft labels. The vectors represent the probability of the sample belonging to each cluster. The input is

the above encoder input which is the adjacency matrix, and its output layer is the clustering layer or the clustered layer. The loss function was used to compute loss using the mean squared error. We use KMeans for predicting clusters of the predicted data from Neural Network encoder. The performance of the Neural Network is below:

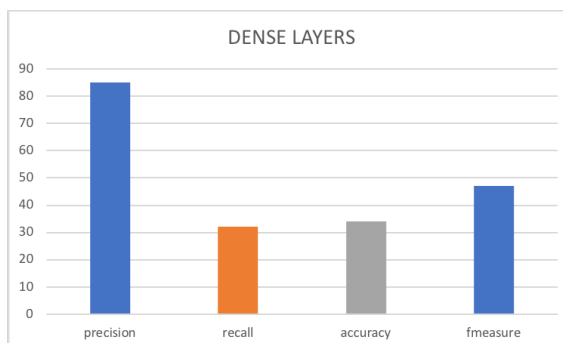


Figure 4.89: Accuracy for Rhabdomyosarcoma for NN

4.13.0.5 Tanimoto coefficient

The drug candidates will be selected by comparing the chemical similarity between FDA approved drugs and the identified drug candidates. This based on the dogma that chemicals with a similar structure might have similar biological activity. In other words, if a drug candidate has an identical chemical structure to one of the FDA approved drugs, then we assume that drug candidate might have similar biological activities to the FDA approved the drug, and it might be useful for treating that disease. To calculate the drug similarity, we do the following:

- 1) The drug name was first to convert to its SMILE representation, the simplified molecular-input line-entry system (SMILES), one type of chemical represen-

tations, by programmatically invoking the PubChem Power User Gateway (PUG) web service; 2) The mostly used PubChem fingerprints based on the SMILES was generated; and 3) The Tanimoto coefficient was calculated for each drug pair based on their fingerprints by using the Chemistry Development Kit (CDK).

The top-ranked drugs with the highest similarity to the FDA approved drugs as potential drug candidates can be further evaluated about their biological activity. When we compared drugs, we identified the current FDA approved drugs.

The figure 2.5 show two drugs that have a similar structure and similar function and the figure 2.4 show two drugs that are dissimilar in structure and function. The goal is to find two drugs that are similar both in structure and function.

Chemical Similarity using the Tanimoto coefficient was applied to our drugs to find the chemical similarity between drug candidates and already existing prostate cancer drugs. A majority of the drugs had their similarities ranging from 0.75 to 1 using the Tanimoto coefficient.

The Tanimoto coefficient was utilized to find the similarity between two drug molecules, and we applied the k-means algorithm to the similarity matrix. The performance of the Tanimoto coefficient is below:

4.13.0.6 Our Method(DReiM)

Our result showed for the multimodal network precision and specificity to be 100%. The general performance of this technique is below. We applied the same method for 4-Modes, 5-Modes, 6-Modes, 7-Modes and up to 15-Modes.

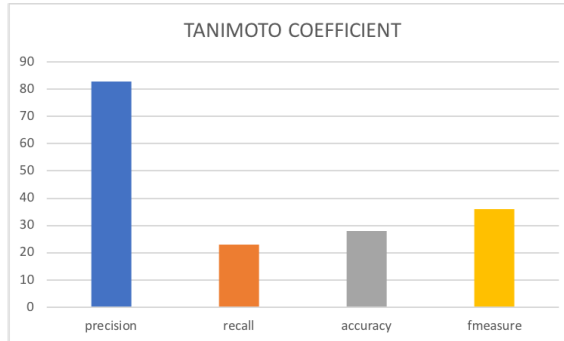


Figure 4.90: Accuracy for Rhabdomyosarcoma for Tanimoto Coefficient

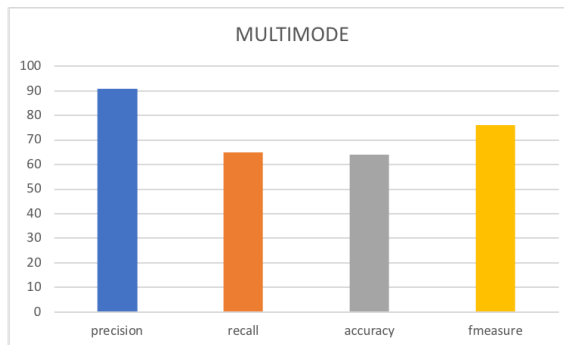


Figure 4.91: Accuracy for Rhabdomyosarcoma for Multimode Network

5.14 Conclusion and Future Work

In this dissertation, the sequence level analytics by the use of different encoding schemes [45] was used to extend the task towards drug repositioning using phosphorylated proteins. Our approach utilized both the network analytical approach and a multimode approach to accomplish this task. Our method can be summarized: i) identification of phosphorylated protein through sequence analysis using the sequence encoding schemes ii) network integration and network analytics and iii) multimodal network approach to find potential drug candidates.

The experimental results using all the following approach was presented using the datasets collected from the Lab for PIM Substrates and rhabdomyosarcoma. Our result shows the need to explore the need for a multilayered approach to analyzing medical data. Our results show that we can find a combination of the disease-related element in a cluster that play major together. Our contribution is a novel approach that helps to incrementally create multimode structures using the various bi-mode relationship that is connected.

We intend to extend our approach to some other research areas that include:

We intend to apply a molecular docking approach to understand proteins that have interactions with one another. Our current method doesn't contain any form of three-dimensional interaction between the proteins. The application of the docking approach will act as a filter instead of the sequence approach we utilized in our current path.

We have applied our approach to PIM substrates and rhabdomyosarcoma. We

intend to implement the method to other conditions as well with the hope to identify critical genes, SNP, and pathways that play essential roles in these conditions.

The work can be extended in a decision support system to predict future interactions between diseases and disease-related elements. We also intend to build a significant data infrastructure that can search PIM Substrate specifically out and predict new substrates for lab tests.

We intend to introduce side effects dataset into our analysis to help identify levels of safety for patients and to enhance our method.

Bibliography

- [1] Elias Zintzaras, Michael Miligkos, Panayiotis Ziakas, Ethan M Balk, Despoina Mademtzoglou, Chrysoula Doxani, Theodoros Mprotsis, Raman Gowri, Paraskevi Xanthopoulou, Ioanna Mpoulimari, et al. Assessment of the relative effectiveness and tolerability of treatments of type 2 diabetes mellitus: a network meta-analysis. *Clinical therapeutics*, 36(10):1443–1453, 2014.
- [2] Ping Zhang, Pankaj Agarwal, and Zoran Obradovic. Computational drug repositioning by ranking and integrating multiple data sources. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 579–594. Springer, 2013.
- [3] Liyang Yu. Sparql: Querying the semantic web. In *A Developers Guide to the Semantic Web*, pages 265–353. Springer, 2014.
- [4] Chao Wu, Ranga C Gudivada, Bruce J Aronow, and Anil G Jegga. Computational drug repositioning through heterogeneous network clustering. *BMC systems biology*, 7(5):S6, 2013.
- [5] Elisabeth Wong, Brittany Baur, Saad Quader, and Chun-Hsi Huang. Biological network motif detection: principles and practice. *Briefings in bioinformatics*, 13(2):202–215, 2011.
- [6] Raimond L Winslow, Natalia Trayanova, Donald Geman, and Michael I Miller. Computational medicine: translating models to clinical care. *Science translational medicine*, 4(158):158rv11–158rv11, 2012.
- [7] Bridget K Wagner and Stuart L Schreiber. The power of sophisticated phenotypic screening and modern mechanism-of-action methods. *Cell chemical biology*, 23(1):3–9, 2016.
- [8] Sebastian Wernicke and Florian Rasche. Fanmod: a tool for fast network motif detection. *Bioinformatics*, 22(9):1152–1153, 2006.
- [9] Ross Cagan and Pablo Meyer. Rethinking cancer: current challenges and opportunities in cancer research, 2017.
- [10] John V Frangioni. New technologies for human cancer imaging. *Journal of clinical oncology*, 26(24):4012–4021, 2008.

- [11] Susumu Hirabayashi, Thomas J Baranski, and Ross L Cagan. Transformed drosophila cells evade diet-mediated insulin resistance through wingless signaling. *Cell*, 154(3):664–675, 2013.
- [12] Bert Vogelstein and Kenneth W Kinzler. The multistep nature of cancer. *Trends in genetics*, 9(4):138–141, 1993.
- [13] Andrea Califano and Mariano J Alvarez. The recurrent architecture of tumour initiation, progression and drug sensitivity. *Nature reviews. Cancer*, 17(2):116, 2017.
- [14] Carmen Blanco-Aparicio and Amancio Carnero. Pim kinases in cancer: diagnostic, prognostic and treatment opportunities. *Biochemical pharmacology*, 85(5):629–643, 2013.
- [15] Martijn C Nawijn, Andrej Alendar, and Anton Berns. For better or for worse: the role of pim oncogenes in tumorigenesis. *Nature reviews. Cancer*, 11(1):23, 2011.
- [16] JD Feldman, L Vician, M Crispino, G Tocco, M Baudry, and Harvey R Herschman. Seizure activity induces pim-1 expression in brain. *Journal of neuroscience research*, 53(4):502–509, 1998.
- [17] Laurent Brault, Christelle Gasser, Franz Bracher, Kilian Huber, Stefan Knapp, and Jürg Schwaller. Pim serine/threonine kinases in the pathogenesis and therapy of hematologic malignancies and solid cancers. *haematologica*, 95(6):1004–1015, 2010.
- [18] CJ Saris, J Domen, and A Berns. The pim-1 oncogene encodes two related protein-serine/threonine kinases by alternative initiation at aug and cug. *The EMBO journal*, 10(3):655, 1991.
- [19] Ze Ping Wang, Nandini Bhattacharya, Matt Weaver, Kate Petersen, Maria Meyer, Leslie Gapter, and Nancy S Magnuson. Pim-1 a serine/threonine kinase with a role in cell survival, proliferation, differentiation and tumorigenesis. *Journal of veterinary science*, 2(3):167–179, 2001.
- [20] Jie Wang, Jongchan Kim, Meejeon Roh, Omar E Franco, Simon W Hayward, Marcia L Wills, and Sarki A Abdulkadir. Pim1 kinase synergizes with c-myc to induce advanced prostate carcinoma. *Oncogene*, 29(17):2477, 2010.
- [21] Eileen White. The pims and outs of survival signaling: role for the pim-2 protein kinase in the suppression of apoptosis by cytokines. *Genes & development*, 17(15):1813–1816, 2003.
- [22] Carmen Blanco-Aparicio, Ana María García Collazo, Julen Oyarzabal, Juan F Leal, María Isabel Albarán, Francisco Ramos Lima, Belén Pequeño, Nuria

- Ajenjo, Mercedes Becerra, Patricia Alfonso, et al. Pim 1 kinase inhibitor etp-45299 suppresses cellular proliferation and synergizes with pi3k inhibition. *Cancer letters*, 300(2):145–153, 2011.
- [23] T Möröy, Antje Grzeschiczek, Sigrid Petzold, and Klaus-Ulrich Hartmann. Expression of a pim-1 transgene accelerates lymphoproliferation and inhibits apoptosis in lpr/lpr mice. *Proceedings of the National Academy of Sciences*, 90(22):10734–10738, 1993.
- [24] T Möröy, S Verbeek, A Ma, P Achacoso, A Berns, and F Alt. E mu n-and e mu l-myc cooperate with e mu pim-1 to generate lymphoid tumors at high frequency in double-transgenic mice. *Oncogene*, 6(11):1941–1948, 1991.
- [25] H Th Cuypers, G Selten, A Berns, and AHM Geurts Van Kessel. Assignment of the human homologue of pim-1, a mouse gene implicated in leukemogenesis, to the pter-q12 region of chromosome 6. *Human genetics*, 72(3):262–265, 1986.
- [26] Gerard Selten, H Theo Cuypers, and Anton Berns. Proviral activation of the putative oncogene pim-1 in mulv induced t-cell lymphomas. *The EMBO journal*, 4(7):1793, 1985.
- [27] HG Van Der Poel, J Zevenhoven, and AM Bergman. Pim1 regulates androgen-dependent survival signaling in prostate cancer cells. *Urologia internationalis*, 84(2):212–220, 2010.
- [28] Jian Chen, Masanobu Kobayashi, Stephanie Darmanin, Yi Qiao, Christopher Gully, Ruiying Zhao, Satoshi Kondo, Hua Wang, Huamin Wang, Sai-Ching Jim Yeung, et al. Hypoxia-mediated up-regulation of pim-1 contributes to solid tumor formation. *The American journal of pathology*, 175(1):400–411, 2009.
- [29] Maja Narlik-Grassow, Carmen Blanco-Aparicio, Yolanda Cecilia, Sandra Peregrina, Beatriz Garcia-Serelde, Sandra Muñoz-Galvan, Marta Cañamero, and Amancio Carnero. The essential role of pim kinases in sarcoma growth and bone invasion. *Carcinogenesis*, 33(8):1479–1486, 2012.
- [30] Tina Morwick. Pim kinase inhibitors: a survey of the patent literature. *Expert opinion on therapeutic patents*, 20(2):193–212, 2010.
- [31] Alex N Bullock, Judit Debreczeni, Ann L Amos, Stefan Knapp, and Benjamin E Turk. Structure and substrate specificity of the pim-1 kinase. *Journal of Biological Chemistry*, 280(50):41675–41682, 2005.
- [32] Shannon M Mumenthaler, Patricia YB Ng, Amanda Hodge, David Bearss, Gregory Berk, Sarath Kanekal, Sanjeev Redkar, Pietro Taverna, David B Agus, and Anjali Jain. Pharmacologic inhibition of pim kinases alters prostate cancer cell growth and resensitizes chemoresistant cells to taxanes. *Molecular cancer therapeutics*, 8(10):2882–2893, 2009.

- [33] Kevin C Qian, Joey Studts, Lian Wang, Kevin Barringer, Anthony Kronkaitis, Charline Peng, Alistair Baptiste, Roger LaFrance, Sheenah Mische, and Bennett Farmer. Expression, purification, crystallization and preliminary crystallographic analysis of human pim-1 kinase. *Acta Crystallographica Section F: Structural Biology and Crystallization Communications*, 61(1):96–99, 2005.
- [34] M Yamaguchi. Adiponectin: production, regulation, and roles in disease. *New York: Nova Science Publishers*, 2012.
- [35] Samuel Croset. Drug repositioning and indication discovery using description logics. 2014.
- [36] Zhiyong Lu, Pankaj Agarwal, ATUL J Butte, et al. Computational drug repositioning. In *Pacific Symposium on Biocomputing*, pages 1–4, 2013.
- [37] Iyanuoluwa Emmanuel Odebode, Aryya Gangopadhyay, and Qian Zhu. Acquisition of diabetes-related biological associations using a motif based network: Preliminary results. In *Bioinformatics and Biomedicine (BIBM), 2015 IEEE International Conference on*, pages 1057–1062. IEEE, 2015.
- [38] Ping Zhang, Fei Wang, and Jianying Hu. Towards drug repositioning: a unified computational framework for integrating multiple aspects of drug similarity and disease similarity. In *AMIA Annual Symposium Proceedings*, volume 2014, page 1258. American Medical Informatics Association, 2014.
- [39] Fabrice Moriaud, Stéphane B Richard, Stewart A Adcock, Laetitia Chanas-Martin, Jean-Sébastien Surgand, Marouane Ben Jelloul, and François Delfaud. Identify drug repurposing candidates by mining the protein data bank. *Briefings in bioinformatics*, 12(4):336–340, 2011.
- [40] Michel Morange. The central dogma of molecular biology. *Resonance*, 14(3):236–247, 2009.
- [41] Ron Milo, Shai Shen-Orr, Shalev Itzkovitz, Nadav Kashtan, Dmitri Chklovskii, and Uri Alon. Network motifs: simple building blocks of complex networks. *Science*, 298(5594):824–827, 2002.
- [42] Jakub M Tomczak and Adam Gonczarek. Decision rules extraction from data stream in the presence of changing context for diabetes treatment. *Knowledge and Information Systems*, pages 1–26, 2013.
- [43] Paolo Paoli, Paolo Cirri, Anna Caselli, Francesco Ranaldi, Giulia Bruschi, Alice Santi, and Guido Camici. The insulin-mimetic effect of morin: a promising molecule in diabetes treatment. *Biochimica et Biophysica Acta (BBA)-General Subjects*, 1830(4):3102–3111, 2013.
- [44] Michel Morange. What history tells us xiii. fifty years of the central dogma. *Journal of biosciences*, 33(2):171–175, 2008.

- [45] Iyanuoluwa Emmanuel Odebode. *The application of information retrieval techniques to the mining of bioinformatics data*. PhD thesis, Morgan State University, 2011.
- [46] Florence Corpet. Multiple sequence alignment with hierarchical clustering. *Nucleic acids research*, 16(22):10881–10890, 1988.
- [47] Robert C Edgar. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, 32(5):1792–1797, 2004.
- [48] Jiangning Song, Huilin Wang, Jiawei Wang, André Leier, Tatiana Marquez-Lago, Bingjiao Yang, Ziding Zhang, Tatsuya Akutsu, Geoffrey I Webb, and Roger J Daly. Phosphopredict: A bioinformatics tool for prediction of human kinase-specific phosphorylation substrates and sites by integrating heterogeneous feature selection. *Scientific Reports*, 7(1):6862, 2017.
- [49] Jung-Ying Wang, Hahn-Ming Lee, and Shandar Ahmad. Prediction and evolutionary information analysis of protein solvent accessibility using multiple linear regression. *Proteins: Structure, Function, and Bioinformatics*, 61(3):481–491, 2005.
- [50] Zheng Yuan, Kevin Burrage, and John S Mattick. Prediction of protein solvent accessibility using support vector machines. *Proteins: Structure, Function, and Bioinformatics*, 48(3):566–570, 2002.
- [51] Rafał Adamczak, Aleksey Porollo, and Jarosław Meller. Combining prediction of secondary structure and solvent accessibility in proteins. *Proteins: Structure, Function, and Bioinformatics*, 59(3):467–475, 2005.
- [52] Hsien-Da Huang, Tzong-Yi Lee, Shih-Wei Tzeng, and Jorng-Tzong Horng. Kinasephos: a web tool for identifying protein kinase-specific phosphorylation sites. *Nucleic acids research*, 33(suppl_2):W226–W229, 2005.
- [53] Rune Linding, Lars Juhl Jensen, Adrian Pasculescu, Marina Olhovsky, Karen Colwill, Peer Bork, Michael B Yaffe, and Tony Pawson. Networkin: a resource for exploring cellular phosphorylation networks. *Nucleic acids research*, 36(suppl_1):D695–D699, 2007.
- [54] John C Obenauer, Lewis C Cantley, and Michael B Yaffe. Scansite 2.0: Proteome-wide prediction of cell signaling interactions using short sequence motifs. *Nucleic acids research*, 31(13):3635–3641, 2003.
- [55] Inkyung Jung, Akihisa Matsuyama, Minoru Yoshida, and Dongsup Kim. Postmod: sequence based prediction of kinase-specific phosphorylation sites with indirect relationship. *BMC bioinformatics*, 11(1):S10, 2010.
- [56] Damian Szklarczyk, Andrea Franceschini, Michael Kuhn, Milan Simonovic, Alexander Roth, Pablo Minguez, Tobias Doerks, Manuel Stark, Jean Muller,

- Peer Bork, et al. The string database in 2011: functional interaction networks of proteins, globally integrated and scored. *Nucleic acids research*, 39(suppl.1):D561–D568, 2010.
- [57] Jianjiong Gao, Jay J Thelen, A Keith Dunker, and Dong Xu. Musite, a tool for global prediction of general and kinase-specific phosphorylation sites. *Molecular & Cellular Proteomics*, 9(12):2586–2600, 2010.
- [58] Minghui Wang, Tao Wang, and Ao Li. ksrmkl: a novel method for identification of kinase–substrate relationships using multiple kernel learning. *PeerJ*, 5:e4182, 2017.
- [59] Sheng-Bao Suo, Jian-Ding Qiu, Shao-Ping Shi, Xiang Chen, and Ru-Ping Liang. Psea: Kinase-specific prediction and analysis of human phosphorylation substrates. *Scientific reports*, 4:4524, 2014.
- [60] Liang Zou, Mang Wang, Yi Shen, Jie Liao, Ao Li, and Minghui Wang. Pkis: computational identification of protein kinases for experimentally discovered protein phosphorylation sites. *BMC bioinformatics*, 14(1):247, 2013.
- [61] Olga Tanaseichuk, A Hadj Khodabakshi, Dimitri Petrov, Jianwei Che, Tao Jiang, Bin Zhou, Andrey Santosyan, and Yingyao Zhou. An efficient hierarchical clustering algorithm for large datasets. *Austin Journal of Proteomics, Bioinformatics and Genomics*, 2(1), 2015.
- [62] James A Shapiro. Revisiting the central dogma in the 21st century. *Annals of the New York Academy of Sciences*, 1178(1):6–28, 2009.
- [63] Matthias Samwald, Anja Jentzsch, Christopher Bouton, Claus Stie Kallesøe, Egon Willighagen, Janos Hajagos, M Scott Marshall, Eric Prud’hommeaux, Oktie Hassanzadeh, Elgar Pichler, et al. Linked open drug data for pharmaceutical research and development. *Journal of cheminformatics*, 3(1):19, 2011.
- [64] Jiao Li, Si Zheng, Bin Chen, Atul J Butte, S Joshua Swamidass, and Zhiyong Lu. A survey of current trends in computational drug repositioning. *Briefings in bioinformatics*, 17(1):2–12, 2015.
- [65] Yu-Chi Lee, Chao-Qiang Lai, Jose M Ordovas, and Laurence D Parnell. A database of gene-environment interactions pertaining to blood lipid traits, cardiovascular disease and type 2 diabetes, 2011.
- [66] Martin Lechner, Veit Höhn, Barbara Brauner, Irmtraud Dunger, Gisela Fobo, Goar Frishman, Corinna Montrone, Gabi Kastenmüller, Brigitte Waegle, and Andreas Ruepp. Cider: multifactorial interaction networks in human diseases. *Genome biology*, 13(7):R62, 2012.
- [67] Uri Alon. Network motifs: theory and experimental approaches. *Nature Reviews Genetics*, 8(6):450–461, 2007.

- [68] Justin Lamb, Emily D Crawford, David Peck, Joshua W Modell, Irene C Blat, Matthew J Wrobel, Jim Lerner, Jean-Philippe Brunet, Aravind Subramanian, Kenneth N Ross, et al. The connectivity map: using gene-expression signatures to connect small molecules, genes, and disease. *science*, 313(5795):1929–1935, 2006.
- [69] V Pogacic, An Bullock, o Fedorov, P Fllippakopoulos, C Gasser, and A Biondi. Structural analysis identifies imidazo[1,2-b]pyridazines as pim kinase inhibitors with in vitro antileukemic activity. *Cancer Research*, 67(14):6916–24, 2007.
- [70] J Makani, SF Ofori-Acquah, O Nnodu, A Wonkam, and K Ohene-Frempong. Sick cell disease: new opportunities and challenges in africa. *The Scientific World Journal*, 2013, 2013.
- [71] Sotiris B Kotsiantis, Ioannis D Zaharakis, and Panayiotis E Pintelas. Machine learning: a review of classification and combining techniques. *Artificial Intelligence Review*, 26(3):159–190, 2006.
- [72] Eugene V Koonin. Does the central dogma still stand? *Biology direct*, 7(1):27, 2012.
- [73] MR Hurle, L Yang, Q Xie, DK Rajpal, P Sanseau, and P Agarwal. Computational drug repositioning: from data to therapeutics. *Clinical Pharmacology & Therapeutics*, 93(4):335–341, 2013.
- [74] L Hood, MA Flores, KR Brogaard, and ND Price. Systems medicine and the emergence of proactive p4 medicine: predictive, preventive, personalized and participatory a2-dekker. *Handbook of Systems Biology*, pages 445–467, 2013.
- [75] Xiang Li, Bin Guan, Minu K Srivastava, Achuth Padmanabhan, Brian S Hampton, and Charles J Bieberich. The reverse in-gel kinase assay to profile physiological kinase substrates. *Nature methods*, 4(11):957–62, Nov 2007.
- [76] Jane B Reece, Lisa A Urry, Michael Lee Cain, Steven Alexander Wasserman, Peter V Minorsky, Robert B Jackson, et al. *Campbell biology*. Number s 1309. Pearson Boston, 2014.
- [77] N Saraswathy and P Ramalingam. 15-phosphoproteomics. *Concepts and techniques in genomics and proteomics*, pages 203–211, 2011.
- [78] Alexander Bürkle. Physiology and pathophysiology of poly (adp-ribosyl) ation. *Bioessays*, 23(9):795–806, 2001.
- [79] Eric Sakk and Iyanuoluwa E Odebode. Vector space information retrieval techniques for bioinformatics data mining. In *Bioinformatics-Trends and Methodologies*. InTech, 2011.
- [80] Rui Xu and Donald C Wunsch. Survey of clustering algorithms. 2005.

- [81] Rui Xu and Don Wunsch. *Clustering*, volume 10. John Wiley & Sons, 2008.
- [82] Tapas Kanungo, David M Mount, Nathan S Netanyahu, Christine D Piatko, Ruth Silverman, and Angela Y Wu. An efficient k-means clustering algorithm: Analysis and implementation. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (7):881–892, 2002.
- [83] William HE Day and Herbert Edelsbrunner. Efficient algorithms for agglomerative hierarchical clustering methods. *Journal of classification*, 1(1):7–24, 1984.
- [84] Michael Steinbach George Karypis, Vipin Kumar, and Michael Steinbach. A comparison of document clustering techniques. In *TextMining Workshop at KDD2000 (May 2000)*, 2000.
- [85] Fionn Murtagh and Pierre Legendre. Wards hierarchical agglomerative clustering method: which algorithms implement wards criterion? *Journal of classification*, 31(3):274–295, 2014.
- [86] Behnam Neyshabur, Ahmadreza Khadem, Somaye Hashemifar, and Seyed Shahriar Arab. Netal: a new graph-based method for global alignment of protein–protein interaction networks. *Bioinformatics*, 29(13):1654–1662, 2013.
- [87] Lucy Skrabanek, Harpreet K Saini, Gary D Bader, and Anton J Enright. Computational prediction of protein–protein interactions. *Molecular biotechnology*, 38(1):1–17, 2008.
- [88] Somaye Hashemifar. Computational prediction and analysis of protein-protein interaction networks. *arXiv e-prints*, page arXiv:1709.01923, Sep 2017.
- [89] Michael W Berry, Susan T Dumais, and Gavin W OBrien. Using linear algebra for intelligent information retrieval. *SIAM review*, 37(4):573–595, 1995.
- [90] Kevin Beyer, Jonathan Goldstein, Raghu Ramakrishnan, and Uri Shaft. When is nearest neighbor meaningful? In *International conference on database theory*, pages 217–235. Springer, 1999.
- [91] Domenico Bordo and Patrick Argos. Suggestions for safe residue substitutions in site-directed mutagenesis. *Journal of molecular biology*, 217(4):721–729, 1991.
- [92] Bogdan Done. *Gene function discovery using latent semantic indexing*. Wayne State University, 2009.
- [93] James U Bowie, Roland Luthy, and David Eisenberg. A method to identify protein sequences that fold into a known three-dimensional structure. *Science*, 253(5016):164–170, 1991.

- [94] Sándor Dominich and Sâandor Dominich. *The modern algebra of information retrieval*. Springer, 2008.
- [95] BRGM Couto, AP Ladeira, and MA Santos. Application of latent semantic indexing to evaluate the similarity of sets of sequences without multiple alignments character-by-character. *Genet Mol Res*, 6(4):983–999, 2007.
- [96] Scott Deerwester, Susan T Dumais, George W Furnas, Thomas K Landauer, and Richard Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.
- [97] Andrew Ng et al. Sparse autoencoder. *CS294A Lecture notes*, 72(2011):1–19, 2011.
- [98] Nitish Srivastava, Elman Mansimov, and pages=843-852 year=2015 Slakhudinov, Ruslan booktitle=International Conference on machine learning. Unsupervised learning of video representations using lstms.
- [99] Andrew Moore. K-means and hierarchical clustering, 2001.
- [100] Hao Ye, Qi Liu, and Jia Wei. Construction of drug network based on side effects and its application for drug repositioning. *PloS one*, 9(2):e87864, 2014.
- [101] Janet Piñero, Àlex Bravo, Núria Queralt-Rosinach, Alba Gutiérrez-Sacristán, Jordi Deu-Pons, Emilio Centeno, Javier García-García, Ferran Sanz, and Laura I Furlong. Disgenet: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic acids research*, 45(D1):D833–D839, 2017.
- [102] Allan Davis. Ctd-comparative toxicogenomics database.
- [103] Andrew E Bruno, Li Li, James L Kalabus, Yuzhuo Pan, Aiming Yu, and Zihua Hu. mirdsnp: a database of disease-associated snps and microRNA target sites on 3’utrs of human genes. *BMC genomics*, 13(1):44, 2012.
- [104] Alex H Wagner, Adam C Coffman, Benjamin J Ainscough, Nicholas C Spies, Zachary L Skidmore, Katie M Campbell, Kilannin Krysiak, Deng Pan, Joshua F McMichael, James M Eldred, et al. Dgidb 2.0: mining clinically relevant drug–gene interactions. *Nucleic acids research*, 44(D1):D1036–D1044, 2015.
- [105] Joseph E Tym, Costas Mitsopoulos, Elizabeth A Coker, Parisa Razaz, Amanda C Schierz, Albert A Antolin, and Bissan Al-Lazikani. cansar: an updated cancer research and drug discovery knowledgebase. *Nucleic acids research*, 44(D1):D938–D943, 2016.
- [106] UniProt Consortium et al. Uniprot: a hub for protein information. *Nucleic acids research*, page gku989, 2014.

- [107] Chris Stark, Bobby-Joe Breitkreutz, Teresa Regul, Lorrie Boucher, Ashton Breitkreutz, and Mike Tyers. Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl_1):D535–D539, 2006.
- [108] David S Wishart, Craig Knox, An Chi Guo, Savita Shrivastava, Murtaza Hassanali, Paul Stothard, Zhan Chang, and Jennifer Woolsey. Drugbank: a comprehensive resource for in silico drug discovery and exploration. *Nucleic acids research*, 34(suppl_1):D668–D672, 2006.
- [109] Michael Kuhn, Monica Campillos, Ivica Letunic, Lars Juhl Jensen, and Peer Bork. A side effect resource to capture phenotypic effects of drugs. *Molecular systems biology*, 6(1):343, 2010.
- [110] Allan Peter Davis, Cynthia J Grondin, Robin J Johnson, Daniela Sciaky, Benjamin L King, Roy McMorran, Jolene Wieggers, Thomas C Wieggers, and Carolyn J Mattingly. The comparative toxicogenomics database: update 2017. *Nucleic acids research*, 45(D1):D972–D978, 2017.
- [111] Stephen T Sherry, M-H Ward, M Kholodov, J Baker, Lon Phan, Elizabeth M Smigielski, and Karl Sirotkin. dbsnp: the ncbi database of genetic variation. *Nucleic acids research*, 29(1):308–311, 2001.
- [112] Teri E Klein, Jeffrey T Chang, Mildred K Cho, Katrina L Easton, Ray Ferguson, Micheal Hewett, Zhen Lin, Y Liu, S Liu, DE Oliver, et al. Integrating genotype and phenotype information: an overview of the pharmgkb project. *The pharmacogenomics journal*, 1(3):167–170, 2001.
- [113] Qian Zhu, Cui Tao, Feichen Shen, and Christopher G Chute. Exploring the pharmacogenomics knowledge base (pharmgkb) for repositioning breast cancer drugs by leveraging web ontology language (owl) and cheminformatics approaches. In *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, page 172. NIH Public Access, 2014.
- [114] Malachi Griffith, Obi L Griffith, Adam C Coffman, James V Weible, Josh F McMichael, Nicholas C Spies, James Koval, Indrani Das, Matthew B Callaway, James M Eldred, et al. Dgidb: mining the druggable genome. *Nature methods*, 10(12):1209–1210, 2013.
- [115] Janet Piñero, Núria Queralt-Rosinach, Àlex Bravo, Jordi Deu-Pons, Anna Bauer-Mehren, Martin Baron, Ferran Sanz, and Laura I Furlong. Disgenet: a discovery platform for the dynamical exploration of human diseases and their genes. *Database*, 2015, 2015.
- [116] Karthik Raman. Construction and analysis of protein–protein interaction networks. *Automated experimentation*, 2(1):2, 2010.
- [117] Albert-László Barabási and Réka Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

- [118] Gitanjali Yadav and Suresh Babu. Nexcade: perturbation analysis for complex networks. *PloS one*, 7(8):e41827, 2012.
- [119] Ahmed Aleroud and Aryya Gangopadhyay. Multimode co-clustering for analyzing terrorist networks. *Information Systems Frontiers*, pages 1–22, 2016.
- [120] Inderjit S Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 269–274. ACM, 2001.
- [121] H Xia, Z Gong, Y Lian, J Zhou, and X Wang. Gene expression profile regulated by creb in k562 cell line. 48(6):2221–2234, 2016.
- [122] Addanki P Kumar, Shylesh Bhaskaran, Manonmani Ganapathy, Katherine Crosby, Michael D Davis, Peter Kochunov, John Schoolfield, I-Tien Yeh, Dean A Troyer, and Rita Ghosh. Akt/creb/cyclin d1 network: a novel target for prostate cancer inhibition in transgenic adenocarcinoma of mouse prostate (tramp) model mediated by nexrutine®, a phellodendron amurense bark extract. *Clinical cancer research: an official journal of the American Association for Cancer Research*, 13(9):2784, 2007.
- [123] Aravind Subramanian, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, Scott L Pomeroy, Todd R Golub, Eric S Lander, et al. Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, 102(43):15545–15550, 2005.
- [124] Vamsi K Mootha, Cecilia M Lindgren, Karl-Fredrik Eriksson, Aravind Subramanian, Smita Sihag, Joseph Lehar, Pere Puigserver, Emma Carlsson, Martin Ridderstråle, Esa Laurila, et al. Pgc-1 α -responsive genes involved in oxidative phosphorylation are coordinately downregulated in human diabetes. *Nature genetics*, 34(3):267, 2003.
- [125] James MacQueen et al. Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability*, volume 1, pages 281–297. Oakland, CA, USA., 1967.
- [126] Hui Zeng, Chengxiang Qiu, and Qinghua Cui. Drug-path: a database for drug-induced pathways. *Database*, 2015, 2015.
- [127] Marco C Venanzoni, Sergio Giunta, Giovan Battista Muraro, Laura Storari, Claudia Crescini, Roberta Mazzucchelli, Rodolfo Montironi, and Arun Seth. Apolipoprotein e expression in localized prostate cancers. *International journal of oncology*, 22(4):779–786, 2003.

- [128] Huiling Chen, Huan-Xiang Zhou, Xiaohua Hu, and Illhoi Yoo. Classification comparison of prediction of solvent accessibility from protein sequences. In *Proceedings of the second conference on Asia-Pacific bioinformatics-Volume 29*, pages 333–338. Australian Computer Society, Inc., 2004.
- [129] Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy, et al. Gephi: an open source software for exploring and manipulating networks. *Icwsm*, 8:361–362, 2009.
- [130] Anam A Gangopadhyay A, Odebode I. A network approach to personalized medicine. *The 2nd International Congress on Personalized Medicine*, 13(11):35, 2013.
- [131] Do-Hee Kim, Yeonui Kwak, Nam Doo Kim, and Taebo Sim. Antitumor effects and molecular mechanisms of ponatinib on endometrial cancer cells harboring activating fgfr2 mutations. *Cancer biology & therapy*, 17(1):65–78, 2016.
- [132] Estelle Daudigeos-Dubus, Ludivine Le Dret, Claudia Lanvers-Kaminsky, Olivia Bawa, Paule Opolon, Albane Vievard, Irène Villa, Mélanie Pagès, Jacques Bosq, Gilles Vassal, et al. Regorafenib: antitumor activity upon mono and combination therapy in preclinical pediatric malignancy models. *PloS one*, 10(11):e0142612, 2015.
- [133] Xiaodong Ma, Lindsey E Becker Buscaglia, Juanita R Barker, and Yong Li. Micrnas in nf- κ b signaling. *Journal of molecular cell biology*, 3(3):159–166, 2011.
- [134] Lisa Diller, Elizabeth Sexsmith, Amy Gottlieb, Frederick P Li, and David Malkin. Germline p53 mutations are frequently detected in young children with rhabdomyosarcoma. *The Journal of clinical investigation*, 95(4):1606–1611, 1995.
- [135] Xiang Chen, Elizabeth Stewart, Anang A Shelat, Chunxu Qu, Armita Bahrami, Mark Hatley, Gang Wu, Cori Bradley, Justina McEvoy, Alberto Pappo, et al. Targeting oxidative stress in embryonal rhabdomyosarcoma. *Cancer cell*, 24(6):710–724, 2013.
- [136] Sandra Vols, Ronit V Sionov, and Zvi Granot. Always look on the bright side: anti-tumor functions of neutrophils. *Current pharmaceutical design*, 23(32):4862–4892, 2017.
- [137] Paresh Dandona, Ahmad Aljada, and Arindam Bandyopadhyay. Inflammation: the link between insulin resistance, obesity and diabetes. *Trends in immunology*, 25(1):4–7, 2004.
- [138] Merce Miranda, Matilde R Chacón, José Gómez, Ana Megía, Victòria Ceperuelo-Mallafre, Sergi Veloso, María Saumoy, Lluís Gallart, Cristóbal

- Richart, Jose Manuel Fernández-Real, et al. Human subcutaneous adipose tissue lpin1 expression in obesity, type 2 diabetes mellitus, and human immunodeficiency virus-associated lipodystrophy syndrome. *Metabolism*, 56(11):1518–1526, 2007.
- [139] Bei B Zhang, Gaochao Zhou, and Cai Li. Ampk: an emerging drug target for diabetes and the metabolic syndrome. *Cell metabolism*, 9(5):407–416, 2009.
- [140] Chao-Qiang Lai, Katherine L Tucker, Laurence D Parnell, Xian Adiconis, Bibiana García-Bailo, John Griffith, Mohsen Meydani, and José M Ordovás. Ppargc1a variation associated with dna damage, diabetes, and cardiovascular diseases: the boston puerto rican health study. *Diabetes*, 57(4):809–816, 2008.
- [141] Jia-Ying Yang, Wu Deng, Yumay Chen, Weiwei Fan, Kenneth M Baldwin, Richard S Jope, Douglas C Wallace, and Ping H Wang. Impaired translocation and activation of mitochondrial akt1 mitigated mitochondrial oxidative phosphorylation complex v activity in diabetic myocardium. *Journal of molecular and cellular cardiology*, 59:167–175, 2013.
- [142] Judit Bassols, Ana Megia, Pilar Soriano-Rodríguez, Marta Díaz, Anna Prats-Puig, Magdalena Gifre, Inmaculada Simón-Muela, Sara Torrent, Anna C Borrell, Joan-Carles Riera-Socasau, et al. A common gene variant in stk11 is associated with metabolic risk markers and diabetes during gestation. *Fertility and sterility*, 100(3):788–792, 2013.
- [143] Satyanarayana Medicherla, Scott Wadsworth, Breda Cullen, Derek Silcock, Jing Y Ma, Ruban Mangadu, Irene Kerr, Sarvajit Chakravarty, Gregory L Luedtke, Sundeep Dugar, et al. p38 mapk inhibition reduces diabetes-induced impairment of wound healing. *Diabetes, metabolic syndrome and obesity: targets and therapy*, 2:91, 2009.
- [144] LM Van Bortel. Efficacy, tolerability and safety of nebivolol in patients with hypertension and diabetes: a post-marketing surveillance study. *Eur Rev Med Pharmacol Sci*, 14(9):749–58, 2010.
- [145] W Zheng, X Feng, L Qiu, Z Pan, R Wang, S Lin, D Hou, L Jin, and Y Li. Identification of the antibiotic ionomycin as an unexpected peroxisome proliferator-activated receptor γ (ppar γ) ligand with a unique binding mode and effective glucose-lowering activity in a mouse model of diabetes. *Diabetologia*, 56(2):401–411, 2013.
- [146] NAVIGATOR Study Group. Effect of valsartan on the incidence of diabetes and cardiovascular events. *New England Journal of Medicine*, 362(16):1477–1490, 2010.
- [147] Lilian Beatriz Aguayo Rojas and Marilia Brito Gomes. Metformin: an old but still the best treatment for type 2 diabetes. *Diabetology & metabolic syndrome*, 5(1):6, 2013.

- [148] Xavier Palomer, Javier Pizarro-Delgado, Emma Barroso, and Manuel Vázquez-Carrera. Palmitic and oleic acid: the yin and yang of fatty acids in type 2 diabetes mellitus. *Trends in Endocrinology & Metabolism*, 29(3):178–190, 2018.
- [149] Henry Leung and Simon Haykin. The complex backpropagation algorithm. *IEEE Transactions on signal processing*, 39(9):2101–2104, 1991.

