$\mathbb{Z}$

UNIVERSITY OF BONN

MASTER'S THESIS

Towards Mechanism based Biomarkers in Personalized Medicine

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science

in the

Algorithmic Bioinformatics
Bonn-Aachen International Center for Information Technology

September 8, 2017

Declaration of Authorship

I, Sabyasachi PATAJOSHI, declare that this thesis titled, "Towards Mechanism based Biomarkers in Personalized Medicine" and the work presented in it are my own. I confirm that:

I herewith certify that this material is my own work, that I used only those sources and resources referred to in the thesis, and that I have identified citations as such.

Signed:

Date:

"Equipped with his five senses, man explores the universe around him and calls the adventure Science."

Sir Edwin Hubble

University of Bonn

Abstract

Holger Fröhlich

Bonn-Aachen International Center for Information Technology

Master of Science

Towards Mechanism based Biomarkers in Personalized Medicine

by Sabyasachi PATAJOSHI

Mathematical models of biological systems are viewed as the conclusive and comprehensive source of biochemical knowledge. These computational/mathematical models can be represented as reaction graphs. Mathematical analytical methods enable the further decomposition of these models into key modules which can be inferred as sub-networks, these sub-networks (extreme currents in the context of my work) are established to be sufficient to explain the activity of the complete model.

This work explores how far the activity level of these sub-networks or extreme currents is useful in explaining the disease status or survival time in case control and survival data respectively, and the reproducibility and interpretability of the biomarker signatures. Cancers are complex diseases involving the dysregulation of multiple genes and pathways, hence I uniquely approach the analysis of these diseases using an ensemble of biological models covering different aspects of disease biology.

To this end, we borrow the methods from flux balance analysis to compute the extreme currents for a selected set of key models which are affected in most cancers; these extreme currents are mapped to the real gene expression data to generate features for application of machine learning methods. Extreme current features are evaluated from a methodological and application perspective.

Acknowledgements

I am grateful to Prof Dr. Holger Fröhlich for providing me with the great opportunity to work on this exciting idea under his esteemed supervision.

His invaluable inputs during the course of the project has helped me to plan and efficiently implement the advanced machine learning strategies as a part of my exploration. From time to time he has pointed me to several interesting literature which were very important in building my fundamentals in the field of personalised medicine.

Dr. Martin Vogt has provided me immense guidance in terms of technical and writing aspects of this project. His advise on several difficult aspects of the project was extremely beneficial, also his suggestions were fundamental towards improving my writing skills.

I am thankful to Prof Dr. Olga Golubnitschaja and Prof Dr. Martin Hofmann Apitius who gave me the platform to work on this very unique and interesting analysis project. Prof Golubnitschaja's clinical insights were very helpful in understanding the project from a clinical perspective and the impact of the project on people's lives, not to mention the delicious snacks offered by her during our meetings.

Dr. Satya Swarup Samal, a friend and former member of the lab has guided me on several aspects of the project, beginning with mathematical background to the implementation aspects of the project. His wealth of knowledge and lucid presentation of concepts got me interested into the core of the project really fast. I would sincerely thank him for his help.

I would also like to thank Ashar Ahmad whose insights on various technical and machine learning methods were invaluable during stressful situations.

I also acknowledge the help extended by my friend Manish Goel. He ran my programs on his server during the downtime of our server, without his help my thesis would not have been completed in time.

My wife Suman has been a constant source of support motivation behind this thesis, I sincerely thank her for this.

I would finally thank my father, brother and sister-in-law for their support.

Contents

Declaration of Authorship i

Abstract iii

Acknowledgements iv

1 Introduction 1
- 1.0 Thesis Outline . . . 1
- 2.0 Precision Medicine . . . 2
  - 2.1 Need for Precision . . . 2
- 3.0 Cornerstone of Precision Medicine : Biomarkers . . . 4
  - 3.1 Introduction : Biomarkers . . . 4
  - 3.2 Functional classification of Biomarkers . . . 6
    - Prognostic Biomarkers . . . 6
    - Predictive Biomarkers . . . 6
    - Diagnostic Biomarkers . . . 7
  - 3.3 Classification of Biomarkers based on Molecular origin 8
  - 3.4 Biomarker Discovery . . . 10
  - 3.5 Challenges in Traditional Machine learning approaches in Biomarker discovery . . . 12
  - 3.6 Integrating Biological data to Biomarker Discovery . . 13

2 Motivation and Goals 15
- 1.0 Motivation . . . 15
  - 1.1 Lack of Reproducibility . . . 15
  - 1.2 Lack of Interpretability . . . 16
  - 1.3 Poor prediction performance . . . 16
- 2.0 Goals . . . 16
  - 2.1 Extreme Currrents based Approach . . . 16
  - 2.2 Multimodal Approach . . . 17

3 Methods 18

1.0 Extreme Currents based Approach . . . 18
1.1 Selection of Pathway Models . . . 18
1.2 Pathway Models to Extreme Currents . . . 19
1.3 Mapping Extreme Currents to Gene Expression: Feature Matrix . . . 21
1.4 Ensemble of Pathway Models Combined Feature Matrix 22
1.5 Next : Applying Machine learning methods . . . 23
Elastic net (Zou and Hastie, 2005) . . . 23
Sparse Group Lasso . . . 24
Gradient Boosting Machine (Friedman, 2000) . . . 24
Boosting based on Pathway Models . . . 25
Stacking based on Pathway Models . . . 26
2.0 Multimodal Approach . . . 27
2.1 Biomarkers Pathway mapping . . . 27
2.2 Clustering: Non Negative Matrix Factorization . . . 28
3.0 Measuring Prediction Performance and Feature Stability - Common to both approaches . . . 28
3.1 Measuring Predictive ability of a Prognostic Biomarker 28
C-index . . . 28
Cross-validated C-index /AUC . . . 29
3.2 Feature Stability . . . 32

4 Data and Results 34
1.0 Biomodels . . . 34
2.0 Gene Expression Datasets . . . 34
2.1 Breast Cancer data . . . 35
2.2 Glioblastoma data . . . 36
2.3 Prostate Cancer data . . . 36
3.0 Multi modal Data . . . 36
4.0 Prediction Performance in Extreme currents based Methods . 37
4.1 Breast Cancer . . . 37
4.2 Glioblastoma data . . . 37
4.3 Prostate Cancer data . . . 40
5.0 Model Interpretation for Breast Cancer . . . 41
6.0 Biomarker discovery in Multimodal data . . . 43
6.1 Identification of sub clusters . . . 43
6.2 Signature Reproducibility . . . 45
6.3 Cluster Predictability . . . 45

6.4 Cluster Analysis 45
6.5 High Risk Vs Low Risk clusters 46
6.6 High Risk and Low Risk Groups Vs Postmenopausal Benign Patients 47
6.7 High Risk and Low Risk Groups Vs Malign Patients . . 47
6.8 Overall Benign vs Malignant Patients 49

5 Conclusion and Future Directions **51**
1.0 Conclusions 51
1.1 Feature Stability and Sparsity 51
1.2 Predictability 52
1.3 Multimodal Data 52
2.0 Future Directions 53

Bibliography **54**

A Biological Pathway Mappings **59**

B Biomodels **64**

C Marker Importance **68**

List of Figures

1.1 Percentage of patients for which a major drug is active on average for a disease (Cho, Jeon, and Kim, 2012) 3
1.2 Personalised Algorithm 5
1.3 Biomarker Workflow 10

3.1 An example of an extreme current 21
3.2 Overview : Extreme currents based method 22
3.3 Models Stacking 27
3.4 Cross Validation 31
3.5 Overview : Multimodal approach 33

4.1 Breast Cancer Cross-validated AUC, For acronyms refer : 4.1 . 38
4.2 Breast Cancer Cross-validated Feature Frequency of top 30 most frequently selected features, For acronyms refer : 4.1 38
4.3 Breast Cancer - Number of Features vs Feature Frequency , For acronyms refer : 4.1 39
4.4 Glioblastoma Cross-validated AUC, For acronyms refer : 4.1 . 40
4.5 Glioblastoma Cross-validated Feature Frequency, For acronyms refer : 4.1 40
4.6 Glioblastoma - Number of Features vs Feature Frequency, For acronyms refer : 4.1 41
4.7 Prostate cancer Cross-validated AUC, For acronyms refer : 4.1 41
4.8 Prostate cancer cross-validated feature frequency, For acronyms refer : 4.1 42
4.9 Prostate cancer - Number of Features vs Feature Frequency, For acronyms refer : 4.1 42
4.10 Visualization of Extreme Currents 43
4.11 Silhouette Index and Cophenetic Correlation 44
4.12 Consensus Matrix and Silhouette Plot 44
4.13 PCA Plot High vs Low risk clusters 46
4.14 Marker values - High vs Low risk clusters 46
4.15 Marker values - High vs Low risk clusters(Contd.) 47

4.16 PCA of pre-menopausal (blue) and postmenopausal (red) benign patients 48
4.17 Cross validated AUC - A) High risk group Vs Malign B) Low risk group Vs Malign 48
4.18 Cross validated AUC GBM - High vs Low risk clusters 49
4.19 Boxplot of AUC of GBM - Benign Vs Malign, from 10x10 cross validation : A) With menopausal status B) Without menopausal status 50

A.1 Homocysteine 60
A.2 Homocysteine 61
A.3 Actin 62
A.4 Actin 63

List of Tables

1.1 Prognostic Biomarkers 6
1.2 Predictive Biomarkers 7
1.3 Diagnostic Biomarkers 8

4.1 Acronyms 39
4.2 Top Features in Gradient Boosting Machine 42

B.1 Biomodels 65
B.2 Biomodels(Contd.) 66
B.3 Biomodels(Contd.) 67

C.1 Cross validated Marker Frequency in High Vs Low risk cluster 68
C.2 Relative importance of Markers and Trend- Benign vs Malign clusters (With Menopausal status) 69
C.3 Relative importance of Markers and Trend- High Risk vs Malign clusters 70
C.4 Relative importance of Markers and Trend- Benign vs Malign clusters (Without Menopausal status) 71
C.5 Relative importance of Markers and Trend- Low risk vs Malign clusters 72
C.6 Relative importance of Markers and Trend- Low vs High risk clusters 73

Chapter 1

Introduction

1.0 Thesis Outline

This thesis is divided into five chapters, the objective(s) of each chapter is mentioned below :

1. **Introduction** : This chapter introduces the concepts of personalised medicine and bionarkers, towards the end of this chapter I bring up the topic of this thesis - mechanism based biomarkers.
2. **Motivation and Goals** : It contains the challenges faced by current biomarker strategies (Motivation) and how this thesis addresses them (Goals), specifically the goal is to explore two mechanism based approaches- Extreme current based strategy and Multimodal strategy.
3. **Methods** : Methods goes into the details of methods employed in implementing and evaluating the two mechanism based biomarker strategies.
4. **Results** : The results from application of methods for the two strategies is presented.
5. **Conclusion and Future Directions** : The results are discussed and conclusions are made on how far the tested strategies were able to meet the goals. Finally the challenges faced during implementation and possible future work in this direction is discussed.

2.0 Precision Medicine

2.1 Need for Precision

Traditionally in the context of medicine, all patients are treated equally for the same disease. This however has a huge burden in terms of treatment efficacy and cost. Each individual is different and so is its genome which makes them respond differently to drugs and have various levels of vulnerabilities to diseases. Standardization of treatments from multi centre drug trials and personalisation of treatment are both sides of the same coin and a balance must be achieved between the two aspects (Kapalla et al., 2016). Personalisation is the synonym for care as it was around even before the advent of genomics in the form of Prakriti based medicine (Chatterjee and Pancholi, 2011). It has its roots in western medicine as well. "It's far more important to know what person has the disease than what disease the person has"- , as said by Hippocrates -Father of Modern Medicine, very much establishes the tenet of personalised medicine.

Apart from devising novel therapies, companies and government agencies now are also focussed in optimizing treatment strategies and developing approaches towards preventing diseases. Precision, Personalised or Individualised medicine are the terms that usually refer to the goal of treating each person based on its genome and epigenome profile combined with the prior knowledge of the drug response for the person of a particular group.

However the scope of these terms is much beyond choosing drugs based on genetic makeup. Person centered medicine has the patient as the primary focus and the ease and comfort of the patient is prioritised over the treatment of disease. Precision medicine includes predictive, preventive and therapeutic aspects and is based on the genetic, epigenetic and socio cultural attributes of the individual. Precision Medicine has been introduced as a paradigm shifting methodology in health care; however it has been already around in medicine since quite a long time in different forms. Blood transfusion and organ implantation requires extensive testing of the blood and immune response of the patient based on which the donor is selected; this is one of the early on examples of individual medicine.

The predictive and preventive medicine aims to formulate periodic testing that can warn us with early indications of a serious disease enabling us to take appropriate steps to cure it or delay its progression; eventually leading

to a better prognosis. Identifying the key mutations that promote carcinogenesis versus the silent mutations is imperative in devising a preventive strategy. Mutations may include substitutions or single nucleotide polymorphisms or structural variations such as duplications, insertions, inversions, deletions and translocations. An algorithm to rank the mutations according to their virulence is also very important in preventive medicine to enable easier interpretation for the physicians who would otherwise inundate with the wealth of information. Benign tumors must be examined to identify the potential cancer causing mutations so that a more aggressive treatment can be planned for patients having potential mutations even though histopathological examination may not indicate a malignant tumor per se.

Cancer has long been associated with genetic mutations and each cancer has many heterogenenous subtypes. Each tumor is different with a unique set of mutations and so apparently cancer is the disease very much in need of personalised medicine. Figure 1.1 shows the percentage of patients for which a major drug is effective on average. The percentage varies by disease and the lower percentage may be an indicator of high variation in response to a single drug in the particular disease (Cho, Jeon, and Kim, 2012). This again proves that one drug will not work for all.

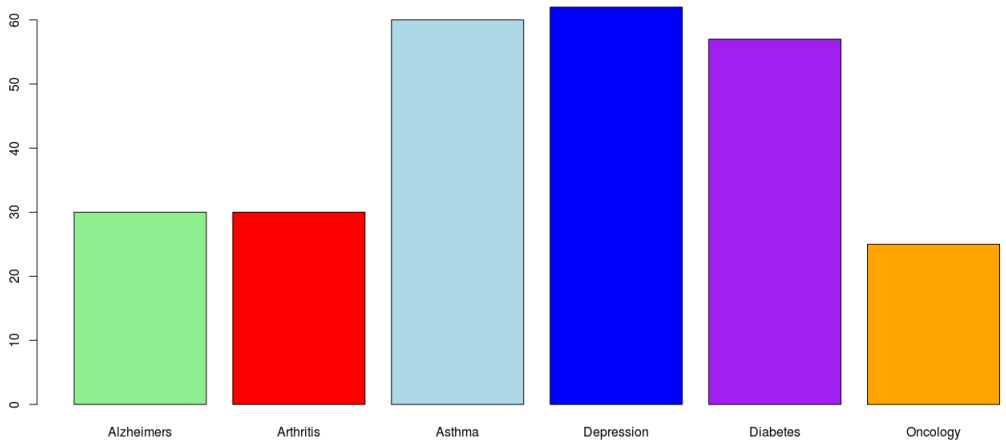

FIGURE 1.1: Percentage of patients for which a major drug is active on average for a disease (Cho, Jeon, and Kim, 2012)

Conventionally cancer is approached by surgical, radiological and chemical modalities. Cancer treatment especially chemotherapy is costly and is accompanied by major side effects, thorough study and planning is done by the oncologist before commencing any treatment. A majority of anti-cancer

drugs work by destroying cancer cells, this particular type of treatment is highly nonspecific and is associated with most side effects. Targeted therapies are new class of drugs that are highly enzyme or receptor specific, i.e. these small molecules or antibodies interact with their substrates and bring on the desired therapeutic effect. These class of drugs are an improvement over cytotoxic drugs as they have much lesser side effects. These targeted drugs target particular mutations and a patient must be identified positive for these mutations for the drug to be effective. Imatinib is approved for the treatment of acute lymphocytic leukemia patients with Philadelphia chromosome positive. Similarly Gefinitib will only work with tumors with mutated and overactive EGFR. These drugs were discovered from the ideas (targeting specific mutations) that form the core of personalised medicine. An example is shown in figure 1.2.

Drug adverse effects is one of the main things along with efficacy that is pivotal while prescribing a drug. Studies have found out that a large number of drug adverse reaction cases have a genetic cause (Daly, 2013). Cases like this one are highly predictable and hence preventable. Personalised medicine will assist in choosing the optimal treatment regimen with maximum efficiency and minimum adverse reactions. Pharmacogenomics is the study of genomic variations that affect the adverse reaction and efficacy of drugs on a individual.

3.0 Cornerstone of Precision Medicine : Biomarkers

3.1 Introduction : Biomarkers

In order to achieve personalization we need to be able to separate an individual from the rest; or find subgroups in the general population. These groups may be different from each other based on the different risks towards developing a disease, different prognosis and different drug responses (Chatterjee and Pancholi, 2011). A reliable biomarker should be able to predict well these differences between different groups, hence can be employed in predictive medicine. Since 2000 there has been a lot of work in the field of biomarker development, there have been 26000 publications in Pubmed under neoplasm

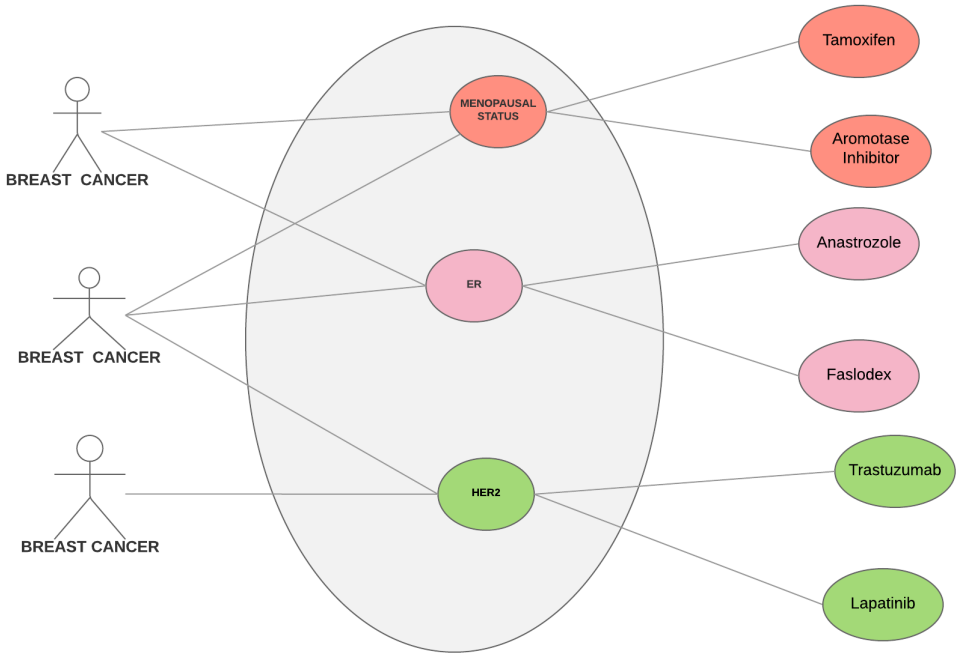

FIGURE 1.2: Personalised Algorithm

and predictive biomarker and 14000 publications under neoplasm and prognostic biomarker (Mehta and Shelling, 2010).

High throughput data such as DNA, RNA, proteins and metabolites are being measured for patient cohorts mixed with control groups; the results are studied to develop statistically significant biomarkers. The high throughput technologies will be discussed in brief in next section of this chapter. Low throughput biomarkers based on histopathology and images are also used. Medical Images are also of paramount importance in order to diagnose, stage and predict the prognosis of a disease. The MRI and PET Images are used to monitor the prognosis or response to treatment and guide the treatment accordingly (Bayouth et al., 2011). However, the medical imaging analysis suffers from the disadvantage of lack of standards, with the image processing algorithms developed by different parties with their own standards. Efforts are underway to standardise the image data acquisition and analysis as a part of Quantitative Imaging Network. Golden et al., (Golden et al., 2013) uses the image derived morphological features and quantitative features based on lesion texture of DCE-MRI images; these features were proven to be very accurate in predicting tumor progression and response to chemotherapy.

3.2 Functional classification of Biomarkers

Prognostic Biomarkers

Biomarkers can be classified on various grounds e.g their molecular nature and functional role. Functionally, biomarkers can be classified as Predictive and Prognostic biomarkers (Mehta and Shelling, 2010). Prognostic biomarkers, predicts the outcome of a disease that is the likelihood of recurrence or death, but they don't predict the response to treatment. The below table lists some of the prognostic biomarkers that I have used in my breast cancer project, along with few other examples.

Marker	**Reference**	**Prognosis**
MMP.9	(Ren et al., 2015)	High - Poor
Actin	(De Jong et al., 2010)	High - Poor
Homocysteine	(Ierardi et al., 2013)	High - Poor
Periaxin	(Schulte, 2011)	High - Good
Profilin.1	(Zoidakis et al., 2012)	High - Good
SOD.2	(Miar et al., 2015)	High - Good
Thioredoxin	(Jikimoto et al., 2002)	High - Poor
Rho.A	(Chang et al., 2016)	High - Poor
Calgranulin.A	(Zhou et al., 2008)	Low level - Poor
Prx.Trx	(Cha, Suh, and Kim, 2009)	High - Poor
Comet Assay Class I -IV Percentage	(All, 2014)	Refer paper for details
ING3	(Wang et al., 2007)	Low - Poor
Beta-tubulin	(Sève et al., 2007)	High - Poor

TABLE 1.1: Prognostic Biomarkers

Predictive Biomarkers

The other major class of biomarkers are called Predictive markers, which predict the response of response of a patient to a particular treatment. Some of

the biomarkers may have the dual purpose of being prognostic and predictive. For example - BRCA 1 and CAIX are both predictive and prognostic biomarkers. As a prognostic biomarker, high levels of BRCA1 predicts worse prognosis in patients and as a predictive biomarker high levels of BRCA1 predicts resistance cisplastin therapy. Similarly in Renal Cell Carcinoma, high levels of CAIX predicts good prognosis in patients and it predicts the sensitivity to Interleukin 2 treatment.

The most popular biomarkers are mentioned in Tables 1.1 ,1.2 and 1.3.

Marker	**Cancer**	**Prognosis**
BRCA1	NSCLC	High level – resistance to cisplastin
BRCA1	Breast	High level -response to chemotherapy
EGFR1	NSCLC	EGFR1 mutations – response to gefinitib
EGFR1	Colorectal	EGFR1 amplification -response to anti EGFR1
Rotterdam Signature	Breast	Predict recurrence with tamoxifen
Roche AmpliChip	Breast	Predicts level of CYP2D6 based on genotype and in turn predicts the – Resistance to tamoxifen
MGMT	Glioblastoma	Methylation of MGMT -predicts sensitivity of glioblastoma to temozolmide

TABLE 1.2: Predictive Biomarkers

Diagnostic Biomarkers

These types are one of the most commonly used biomarkers and the oldest in use. Diagnostic biomarkers are used to diagnose the presence of certain conditions (usually diseases). These tests can vary from laboratory tests, physical to imaging tests A combination of the tests may be used to perform diagnosis. Again some biomarkers are used for both diagnostic and predictive purposes. For example certain mutations are used to diagnose and then

used to predict the efficacy of ivacaftor drug on those patients (Davies et al., 2013).

Marker	**Cancer**	**Prognosis**
Sweat Chloride	Cystic Fibrosis	High - Diagnosis
CFTR Mutations	Cystic Fibrosis	Diagnosis
Galactomannan1	Invasive aspergillosis	Presence - Diagnosis
Glycosylated Haemoglobin	Diabetes	High - Diagnosis
Glomerular filtration rate (GFR)	Breast	Low - Kidney damage

TABLE 1.3: Diagnostic Biomarkers

3.3 Classification of Biomarkers based on Molecular origin

Genomic and Transcriptomic Markers

Diseases like cancer and diabetes have a genomic link and genotyping for identification of SNPs in tumor and nontumor DNA is becoming extremely common in context of personalised medicine. Genome wide association studies have made it possible to identify the cancer causing mutations. The limitation of GWAS is that it identifies only common mutations, advanced and economical technologies must be developed to identify more rare mutations in cancer. The genetic testing detects the SNPs, short sequence repeat variations, structural variations like insertions and deletions, translocations, copy number variations etc. Genetic testing is highly useful in monogenic disorders such as hypercholesterolemia and high blood pressure, but they fail to provide such a clear interpretation in case of non mendelian disorders involving multiple genes and environmental factors (Novelli et al., 2008).

The identification of multiple key genes in complex diseases such as diabetes and cancer and their co-expression pattern is helpful in predicting high risk individuals. FDA approved the multigene biomarker '70 gene signature' which is used to distinguish patients at the higher risk of relapse as

compared to patients at a lower risk. The 70 gene signature which is commercially known as MammaPrint is used in making decisions whether to go for chemotherapy or not, thus saving the low risk patient from adverse effects of chemotherapy. The prognostic signature determines the hazard or risk more accurately using the expression of the 70 significant genes. The clinical utility of the 70 gene signature was successfully validated with and without additional factors (Buyse et al., 2006).

Genes often play a role in drug efficacy and adverse reactions. This topic has been extensively discussed in 2.1.

Epigenomic Markers

Epigenetic changes on genome (e.g methylation or acetylation) are essential cellular processes important for cell survival. These processes are disrupted during the disease development and often patterns can be mined from epigenetic modifications to construct predictive, prognostic and diagnostic biomarkers. One of the well studied examples is septin 9 gene methylation, this gene is hypermethylated much more in colon cancer patients(> 50 percentage) as compared to healthy controls (<10 percentage) (Bock, 2009).

Proteomic Markers

Intuitively protein measurements provide a more accurate picture of the system than gene expression measurements which is only an indirect measurement of protein function. Quantitative proteomics can provide the information on the difference in the protein levels between healthy and disease cells using the two most popular methods - Two dimensional electrophoresis and mass spectrometry. However there are only a few protein biomarkers that are being studied as compared to their genomic counterparts. Wide range of proteomic techniques enable the identification of a vast number of biomarker candidates, but currently there is a high false positive rate among proteomic biomarker candidates due to several reasons covered in (Frantzi, Bhat, and Latosinska, 2014) and (Paulovich et al., 2008). A protein panel instead of single measurements may better serve the purpose of protein biomarkers. HER-2,PR and ER are common breast cancer proteomic biomarkers measured by immunohistochemical techniques. PSA and Free PSA proteomic biomarkers are used in measuring prognosis in prostate cancer (Jones et al., 2014). We have measured some important proteins as shown table 1.1 and constructed a biomarker panel to identify high risk patients who are at a greater risk of developing breast cancer.

Biomarker panels

Most complex diseases having the involvement of multiple genes and pathways cannot be adequately addressed by single gene or protein biomarker and often a panel of biomarkers is required (Mujagic et al., 2016) (Jones et al., 2014). 21 gene expression test (Oncotype DX ™) and 70 gene expression test (MammaPrint ®) are the popular approved biomarker panels.

3.4 Biomarker Discovery

Biomarker candidates come from thorough literature review and have to undergo the rigorous process of verification and validation. Figure 1.3 elucidates the process of biomarker discovery and validation process in four major steps. A majority of biomarker candidates are non-reproducible, lack sensitivity and/or specificity or are not clinically useful due to complexity. Till date only a very small number of biomarkers have actually made it to the clinics.

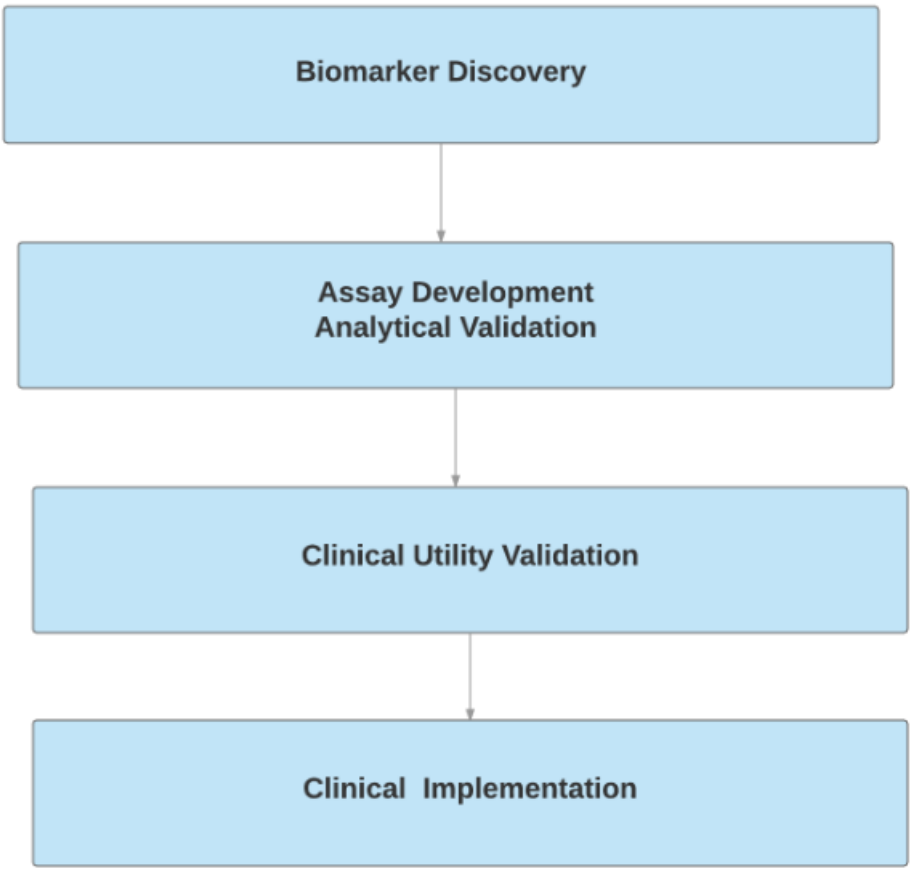

FIGURE 1.3: Biomarker Workflow

Data to Biomarker

In this section I will focus on methods relevant to biomarker discovery, in particular the computational methods involved in this process. Genomically speaking, SNPs are the most precise way to elucidate genomic variation in a population and are partially responsible to a varied phenotype; hence a natural building block to construct biomarkers. Currently 85 million SNPs have been identified in human genome. Technology is able to measure more than a million SNPs in a single SNP array (LaFramboise, 2009).

The magnitude of data with features in millions and samples in thousands entails statistical machine learning and data mining methods in order to discover biomarker signatures. There are two key components of designing any statistical method - Feature selection and Predictive modelling. Transcriptomic or gene expression data from microarray and more recently RNA-seq captures the expression of the genome, this data is in magnitude of thousands as compared to millions in genome. Messenger RNA is finally translated into protein which is the actual working horse of a cell. The complexity of cellular life is achieved by thousands of proteins and their millions of interactions. The proteins and their interactions make up the interactome, another member in the omics family.

Genomic data contains millions of features as SNP and associated phenotype. This is a typical machine learning problem scenario in which we need to find the needle(s) in the haystack, the highly relevant SNPs in this context. The straightforward way to perform univariate testing on each feature to determine its significance towards the phenotype, but the issue with this approach is that it tends to ignore multivariate combination of features as in real world the features are rarely unrelated and they interact. (Drouin et al., 2016)

Let us assume (x,y) are data points in which x is the genome and y is observed phenotype, the data D is in the form

$$\mathcal{S} \stackrel{\text{def}}{=} \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m)\} \sim D^m. \tag{1.1}$$

we consider genome X consists of nucleotides:

$$x \in \{A, T, G, C\} \tag{1.2}$$

Most learning algorithms are designed to learn from a vector representation of the data. Thus, to learn from genomes, we must define a function, that takes a genome as input and maps it to some d dimensional vector space (the feature space).

$$\boldsymbol{\phi} : \{A, T, G, C\}^* \rightarrow \mathbb{R}^d \tag{1.3}$$

Subsequently, a learning algorithm can be applied to the set

$$S' \stackrel{\text{def}}{=} \{(\boldsymbol{\phi}(\mathbf{x}_1), y_1), \ldots, (\boldsymbol{\phi}(\mathbf{x}_m), y_m)\} \tag{1.4}$$

to obtain a model

$$h : \mathbb{R}^d \rightarrow \{0, 1\}. \tag{1.5}$$

The model is a function that, given the feature representation of a genome, estimates the associated phenotype. We strive to obtain a strong learner h that predicts well, i.e., that minimizes the probability, R(h), of making a prediction error for new data, where

$$R(h) \stackrel{\text{def}}{=} \underset{(\mathbf{x},y)\sim D}{Pr} \left[h(\boldsymbol{\phi}(\mathbf{x})) \neq y\right]. \tag{1.6}$$

3.5 Challenges in Traditional Machine learning approaches in Biomarker discovery

There have been attempts to apply the discussed machine learning approaches to discover biomarkers with limited success. The common limitations of biomarker discovery from traditional machine learning approaches :

Low reproducibility and consistency

There are only a handful of biomarkers that have been approved by regulatory authorities for use in clinics. One of the key reasons of low biomarker discovery is low reproducibility i.e, the biomarker panel discovered from one study is not verifiable in another similar study.

So the biomarker is very specific to the study and the associated data, and we need biomarkers that are more generalisable to other experiments.

Sparsity and Interpretability of Models

One of the greatest requirement and challenge is to make biomarkers and models interpretable for the medical community. A biomarker which is more interpretable has a better chance of gaining acceptance than a biomarker which is not so simple to interpret.

Sparsity in a model is one of the most important way to achieve interpretability, as there will be lesser number of features to deal with. Once the model identifies the sparse features, those features can be studied in details. Also sparsity makes the design of biomarker panels cost effective as there will be smaller number of tests required for a panel.

The model based on chosen features is expected to minimise residual error in data and is also expected to generalise well on new data. Usually most models are able to fit the training data very well and don't perform so well in doing prediction on newer data.

3.6 Integrating Biological data to Biomarker Discovery

Integration of biological knowledge with data is one of the possible solutions to address the challenges in traditional machine learning approaches.

Pathway based approaches

Some of the pathway based approaches use the gene set information from the pathways and use that information along with the omics data to develop biomarkers. One of the popular pathway based methods in GSEA (Gene set enrichment analysis). Teschendorff et al., try to incorporate the topology of the pathway by estimating the activities of modules within pathways (Teschendorff et al., 2010) . A nice review of these pathway based methods is provided by Fröhlich et al., (Cun and Fröhlich, 2013) and Bebek et al., (Bebek et al., 2012).

Network based approaches

Protein Protein Interaction networks are another rich source of biological knowledge, which can integrated with omics to develop stable biomarkers. These networks are also much wider in scope as they cover many proteins as compared to pathways which are only limited to few proteins.

The identification of dysregulated subnetworks in the whole PPI network is very important and several proposed approaches exists for the problem. One of the approach is to find the sub networks that explain well the phenotype, the significant sub-networks.

Human PPI network is large, searching for sub-networks with significant dysregulation is a NP-hard problem; greedy ((Chuang et al., 2007),(Chowdhury, 2010)) and branch and bound methods (Dao et al., 2010) are available to find the good approximate solutions to the sub network problem.

The identified sub-networks are used successfully as features for classification and have been shown to outperform regular classifiers.

PPI networks are used along GWAS data to identify the most significant SNPs. PPI networks are also used to find the central proteins/genes and this information is used in feature selection in high dimensional data; the network based feature selection strategy along with regularised models work very well for high dimensional data (Veríssimo et al., 2016).

The details of these methods can be found in (Bebek et al., 2012) and (Samal, 2017).

Mechanism based Biomarker Signature

The objective of my thesis falls is related to this class of biomarkers. I try to re-use the mechanistic knowledge in the ordinary differential equations based detailed quantitative models and build biomarker signatures.

Computational models are available in Biomodels database for signalling and metabolic networks (Chelliah, Laibe, and Le Novere, 2013), and will be explored in more details in next chapters as this thesis is based on computational models from biomodels database.

Metabolic and Signalling pathways are downloadable from KEGG database in interoperable KGML format and can be converted to universal SBML format (Kanehisa et al., 2011).

Chapter 2

Motivation and Goals

Now that background is established, I now focus on the problem areas or motivations followed by the possible solutions i.e goals pursued in this thesis.

1.0 Motivation

As introduced in the previous chapter there are a number of pathway and network based approaches to design biomarkers. In this chapter, the goals and problem areas are briefly described.

1.1 Lack of Reproducibility

"Non-reproducible single occurrences are of no significance to science."

Karl Popper (Casadevall and Fang, 2010)

Reproducibility is one of the greatest challenges in biomarker design and development, primarily the data driven signature based models suffer from this problem heavily for the obvious reasons. The biomarker that has been determined from one study is not necessarily reproducible in another study. Even in the same study, the removal of few samples lead to significant changes in the biomarker signature. These problems led to the integration of prior knowledge from pathways and PPI networks into biomarker signature design.

Reproducibility of biomarkers depend on their stability i.e the robustness of the method being employed to derive the signature.

1.2 Lack of Interpretability

The future of a biomarker is dependent on its interpretability, the level of acceptance of a biomarker by the medical community depends on how far it explains the underlying disease mechanism.

The statistical learning based methods are heavily biased towards the data in context and the multigene complex signatures are difficult to interpret. The next generation network based biomarkers capture PPI network and pathway topology and pathway activity, which improves the interpretability of the biomarker.

Biomarkers can serve as surrogate endpoints in trials involving ethical concerns with clinical end point (e.g invasive procedure), hence the biomarkers need to be interpretable in terms of disease mechanism to serve these kind of purposes.

1.3 Poor prediction performance

Most Biomarkers don't have a very high prediction performance which is a requirement for clinical usage.

2.0 Goals

In order to address the above mentioned concerns, I propose two mechanism based approaches, which are different in implementation but are similar in utilising the mechanism driven features.

2.1 Extreme Currrents based Approach

Knowledge of biological pathways and their mechanism of action is present in literature and in pathway databases such as KEGG and Biomodels ((Kanehisa et al., 2011),(Chelliah, Laibe, and Le Novere, 2013)). Biomodels contains the detailed ordinary differential equations based models with the complete set of equations while KEGG contains pictorial representation of pathways in terms of static diagrams.

The purpose is to transform these mechanistic based models in combination with gene expression data into usable, knowledge rich features which can in turn be used to build classifiers or biomarkers. The mechanism-based biomarkers would not only help to build robust predictive models but can also suggest potential linkage between network sub-activity and phenotype.

2.2 Multimodal Approach

I pursue the novel idea of combining markers from multiple platforms to form a single marker. This idea builds upon the work done by Golubnitschaja et al., 2017 in which two main proteins from tissue remodelling (Rho A and MMP.9) were identified as potential biomarkers; in this project I apply machine learning and data mining methods to devise a multifactorial signature. The markers have been explored in detail in context of their current clinical use as biomarker. The important proteins from several signalling pathways - detoxification pathway, regulation of cytoskeleton pathway and tissue remodelling were selected as markers. Homocysteine and catalase are metabolites that are also evaluated as part of biomarker development. Comet Assay, which uses a combination of genomics and imaging and measure the amount of DNA damage in the cell ((All, 2014)), is also evaluated for their potential as biomarker. This set of markers were measured in patients with benign and malignant tumors and are analysed to find a combined prognostic biomarker.

Chapter 3

Methods

In this chapter, I will discuss the steps that were performed to achieve the goals stated in previous chapter.

This chapter is followed by results from the application of these methods.

1.0 Extreme Currents based Approach

The extreme currents approach has four steps as explained in details below, an overview can be found in figure 3.5.

1.1 Selection of Pathway Models

My thesis extends the work of Samal et al., (Samal, 2017); they calculated and found the extreme current significantly associated with the phenotype.

Signalling networks are dysregulated in most of the complex diseases like various forms of cancer, diabetes and neurodegenerative diseases; in this work I utilise the repository of signalling networks and apply similar extreme currents analytical methods (Engelhardt et al., 2017).

In most of the multigenic diseases, multiple signalling networks are affected and hence modelling based on a single metabolic or signalling network is incomplete and more pathways should be taken into account, so I intend to develop methods to combine a number of signalling networks to build an aggregate model.

I decided to select the major pathways that are dysregulated in cancer, e.g. EGFR, JAK STAT etc, compute sub-pathways (extreme currents) and their

activity scores for various cancers and eventually test these features for their diagnostic and prognostic abilities.

To this end, I query EBI Biomodels database which contains around 640 curated mathematical models of signalling networks and serves our purpose well.

The list of the models can be found in **Appendix B**.

1.2 Pathway Models to Extreme Currents

In this section I introduce the concept of Extreme currents, their calculation and usage.

Extreme Currents has it's origin in steady state analysis of systems (i.e. under equilibrium); there is no net flux among reactions in a system in steady state.

The law of mass action define the change of rate of metabolites as the product of concentration of reactants. Mathematically,

$$\frac{dC}{dt} = S \cdot r \tag{3.1}$$

where C is the concentration of the metabolites, S is the stoichiometric matrix of dimension mxn where S_{ij} is the stoichiometric coefficient of element i in reaction j; and r is the flux vector of n reactions.

At steady state equilibrium there is no net change in the concentration of metabolites, hence mathematically -

$$S \cdot r = 0 \tag{3.2}$$

The nonnegative constraint is applied on reactions reactions i.e all reactions considered are irreversible (the reversible reaction is split into two reactions)

$$r_{\mathrm{i}} \geq 0 \tag{3.3}$$

There are a number of possible reaction flux vectors (r) that would satisfy the equilibrium condition and can be denoted by the following expression -

$$\{ v \in R^{n} : N \cdot v = 0, v \geq 0 \} \tag{3.4}$$

The set of possible flux vectors is called as flux space and the analysis of this vector space is the key aspect of my thesis. The flux vector space can be considered a convex polyhedron and the vertices are called the extreme currents (Clarke, 1988). In case the reactions of a pathway are represented in a stoichiometric matrix and the extreme currents are calculated, these extreme currents can be interpreted as sub-pathways of the network.

Mathematically, the extreme currents can be thought of basis vectors of the flux space, and the each of the flux vector can be represented as the linear combination of the extreme currents.

$$v = c \cdot E \tag{3.5}$$

where E is the matrix of k extreme currents matrix and c are the linear coefficients called convex parameters.

Elementary flux modes and Extreme Pathways are other similar methods of performing such analysis. Wang et al. (Wang and Albert, 2011) used elementary signalling modes to analyse the essentiality of components in a network and rank them.

Extreme currents, Elementary flux modes and Extreme Pathways can be interpreted as the sub-pathways of a pathway network which can explain all the modes of action of a pathway.

An example of an extreme current identified by Samal et al. (Samal, 2017) as a key feature in differentiating tumor vs normal in Prostate cancer is shown in 3.1

In this thesis work, PoCAB (Samal, Errami, and Weber, 2012) software is used to compute the extreme currents for the mathematical models from Biomodels repository.

I downloaded the SBML files from Biomodels data based and computed the extreme currents using PoCAB software, a number of biomodels resulted in zero extreme currents after removal of spurious cycles (Wagner and Urbanczik, 2005). These extreme currents were then converted into ENTREZ gene lists and stored in text file for mapping into gene expression data, feature generation and later application of machine learning methods.

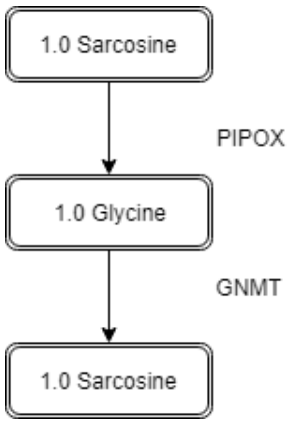

FIGURE 3.1: An example of an extreme current

1.3 Mapping Extreme Currents to Gene Expression: Feature Matrix

Once I generated the extreme currents, the next task was to map then to the gene expression data in order to compute features. The proteins and metabolites in biomodels are annotated and the task for me was to convert them into gene ids so that it can the mapped with gene expression data. The annotations were of various forms, Uniprot /Interpro identifier, Gene ontology, CHEBI , Enzyme nomenclature to name a few.

R platform's inbuilt libraries and Uniprot's webservice APIs were used to perform the conversion from different annotations to entrez identifiers. On the other hand the feature names in gene expression data from different platforms were also converted into entrez identifiers.

Once both gene expression data and extreme currents were converted to the common entrez nomenclature, I mapped the extreme currents into gene expression data. I calculated value of each extreme current as the projection of the expression values of all the genes in extreme current genes on to its first principal component .

Hence finally I had the matrix of all the features for a particular biomodel.

The feature matrix is generated for each of the pathway models, then these feature matrices are combined together as a single matrix while the features are still internally grouped/indexed based on the pathway of origin.

1.4 Ensemble of Pathway Models Combined Feature Matrix

Most of the diseases are complex involving multiple genes and pathways and cannot be explained completely by a single pathway or a model. A list of pathway models are selected based on literature study and are aggregated to build predictive models.

The pathway models can be aggregated to improve prediction by an ensemble technique called model based Gradient boosting. Gradient boosting is explained in details in the next section **Gradient Boosting Machine**.

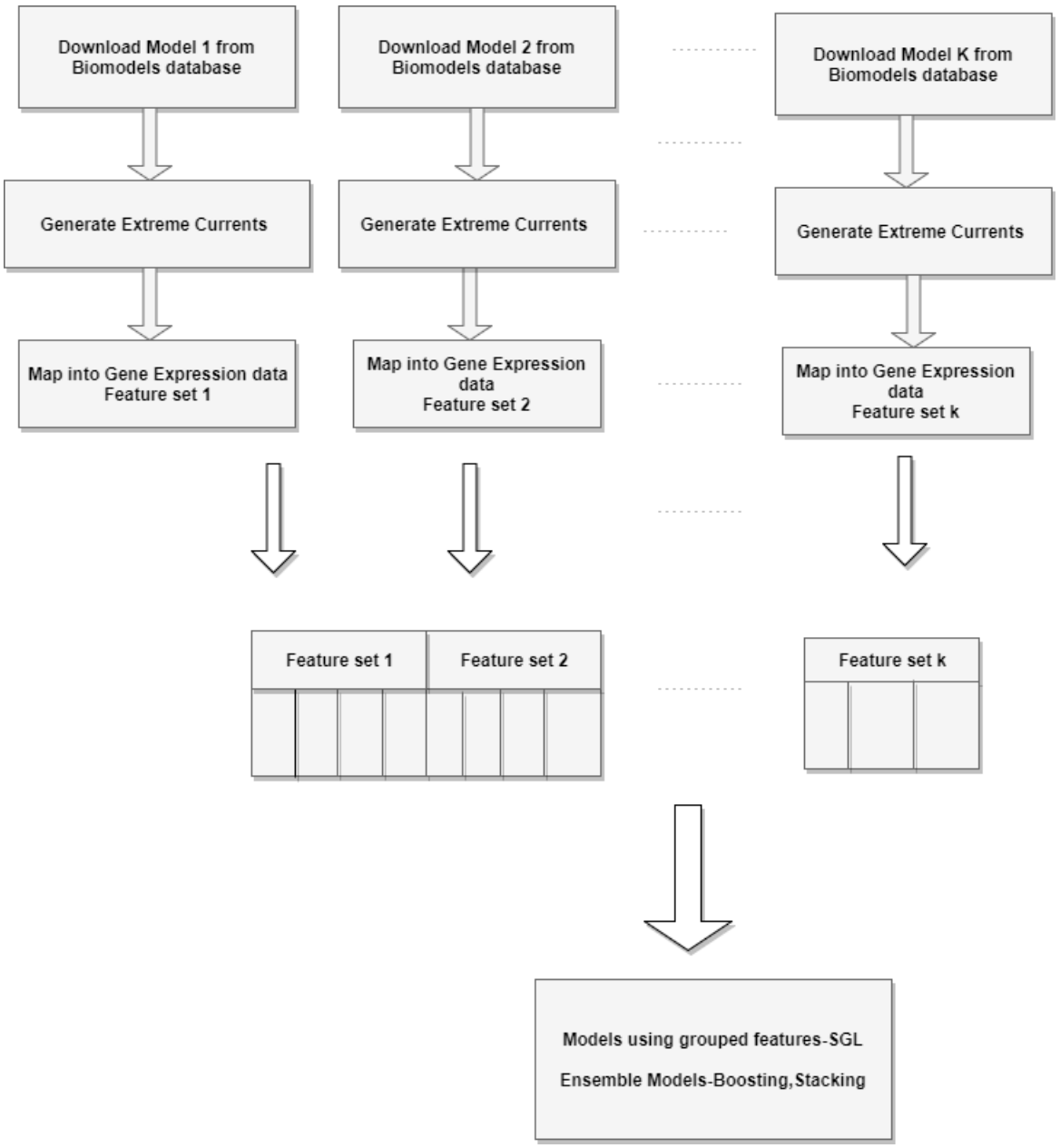

FIGURE 3.2: Overview : Extreme currents based method

1.5 Next : Applying Machine learning methods

Below is the list of machine learning methods used :

Elastic net (Zou and Hastie, 2005)

Regularization is a way to avoid overfitting by penalizing bigger magnitude of parameters against the training error through the use of a regularization term. Generalizable models may be achieved using regularization techniques, there are a number of regularization loss functions suiting different use cases. Extreme currents consists of gene sets containing multiple genes with a tendency to overfit the training data with a poor generalisation to new data, making it suitable to apply regularization techniques.

Elastic net regularisation is the combination of LASSO and ridge penalties, and offers the best of both techniques. The extreme current features are correlated and hence only one would be chosen in Lasso and also the maximum number of features selected in limited to number of samples, hence not suitable in case p > n.

Ridge regression is based on L2 penalty loss and works by shrinking the coefficients of the features. The sparsity of Lasso is desired without its disadvantages, hence elastic net is one of the chosen as it offers the sparsity of lasso minus the demerits. Also, it converges quite fast than the other sophisticated procedures.

$$\hat{\beta} = \underset{\beta}{argmin}(||y - X\beta||^2) + \lambda_2||\beta||^2 + \lambda_1||\beta_1||) \tag{3.6}$$

The expression shows the loss function optimization problem based on L1 and L2 penalties and training error.

Elastic net regularization eliminates the limit on the number of variables selected and encourages grouping; all the variables of a group are selected if. These strategies are useful since number of extreme currents can be large and exceed the number of samples and the extreme currents are similar and occur in groups.

These coefficients are the measure of importance of the extreme current, the larger magnitude meaning more important than the one with smaller magnitude.

Sparse Group Lasso

Sparsity and as well as the grouping of extreme current features are taken care of in Elastic net regularization. Since the features originate from different models, an explicit grouping should be enforced on the model and that is achieved through Sparse Group Lasso.

The features are grouped and hence there are multiple groups (as many as pathways) and sparsity is desirable both a group level and within groups i.e less number of models and less number of extreme currents within them. Sparse group lasso has the unique loss function to achieve group level and within group sparsity (Simon et al., 2013).

$$\min_{\beta} 1/2n \quad ||y - \sum_{l=1}^{m} X^l \beta^l||_2^2 + (1-\alpha)\lambda \sum_{l=1}^{m} \sqrt{p_l}||\beta^l||_2 + \alpha\lambda||\beta||_1 \tag{3.7}$$

X^l are group features and β^l are feature coefficients in the group, β are all coefficients. where $\alpha \quad in \quad [0,1]$ - a optimal combination of the lasso and group lasso penalties ($\alpha = 0$ will become group lasso and $\alpha = 1$ will become lasso)

Gradient Boosting Machine (Friedman, 2000)

SGL and Elastic net are regularised linear models which result in a final single model from the extreme current features, the usage of regularization factor ensures low variance in bias-variance composition.

An ideal model results from both low bias and variance which in turn means better predictions on new data. The extreme currents come from different pathway models and the features from different pathways have a low correlation.

Gradient boosting is one of the popular ensemble methods; instead of building a one monolithic model in a single go, a model is built first with all the features and the misclassified samples in the data are computed; then there is another model building process with the misclassified samples given higher weightage and the process goes on for a prespecified number of times; the model generated at each step is added to the final model. Hence the model

building process works on the gradient at each step. Boosting can reduce the variance and bias . (Friedman, 2000)

The algorithm is depicted as below :

- $We\ have\ (x_1, y_1), (x_2, y_2)...(x_m, y_m)\ with\ values\ of\ X\ and\ y$
  $initialise\ F_0 = 0$
- $Calculate\ the\ pseudoresidual\ wrt\ each\ sample$
  $\frac{\partial L(y_i, F(x_i))}{\partial F(x_i)}\ for\ each\ sample\ i = 1..N$
- $Fit\ the\ the\ hypothesis\ h_t\ to\ (x_i, r_i)\ where\ r_i\ is\ the\ pseudo-residual$
- $Update\ the\ final\ learner\ F_m = F_{m-1} + \nu \cdot h_t$
  $where\ \nu\ is\ the\ step\ learning\ rate$
- $Continue\ steps\ 2-4\ till\ m = m_{stop}$

It is the model of choice in designing biomarker that separates the classes of high risk patients from low risk patients. Gradient boosting models are also used to build classifiers that separate high risk patients from malignant patients and low risk patients from malignant patients; the prediction ability (AUC value) is the indication of the extent of difference between the classes or clusters.

Boosting based on Pathway Models

Gradient Boosting involves a stagewise addition of weak learners to final learner, minimizing the gradient at each step. Model based boosting is built on a set of base learners(linear models,splines), in each boosting iteration step a learner is chosen which minimises the gradient and that learner is then added to the final model.

The stepwise algorithm is as follows for K biomodels :

- $F_{(0)}(X) = 0, F_{k(0)}(X^{(k)}) = 0, k = 1, ...K$
- $Calculate\ the\ pseudo-residuals\ wrt\ each\ Biomodel$
  $\frac{\partial L(y_i, F(x_i))}{\partial F_k(x_i^{(k)})}\ for\ each\ sample\ i = 1..N, k = 1..K$
- $Select\ the\ Biomodel\ h_t\ that\ best\ fits\ the\ pseudo-residuals$
- $Update\ the\ final\ learner\ F_m = F_{m-1} + \nu \cdot h_t$
  $where\ \nu\ is\ the\ step\ learning\ rate$
- $Continue\ steps\ 2-4\ till\ m = m_{stop}$

The set of extreme currents based biomodels are the base learners to model based boosting technique, mboost R package is used to achieve the pathway model based boosting implementation. Each base learner is a biomodel with extreme currents as features, hence each decision tree model consists of the biomodel and associated extreme currents. Pathway based boosting has been proposed and shown by Li et al. in analysis of gene expression data (Wei and Li, 2007).

Stacking based on Pathway Models

Originally introduced by David Wolpert in 1991 (Wolpert, 1992), stacking is a naturally appropriate strategy with ensemble of biological models, where in each of the model is an first level expert and their cross validated responses is used as features for second level expert model.

In the first phase I had to build level 0 learners, cross validation is performed to compute the responses for each sample for each of the models. Hence at the end of this step, I am left with a K-response vector for n training samples. On top of these K learners, a second level learner is built which performs the final prediction.

I re-iterate, each of K base learners in first level is a computational/mathematical biomodel.

Elastic net regularization is used as the base model at both levels of prediction. A pictorial depiction is shown in 3.3

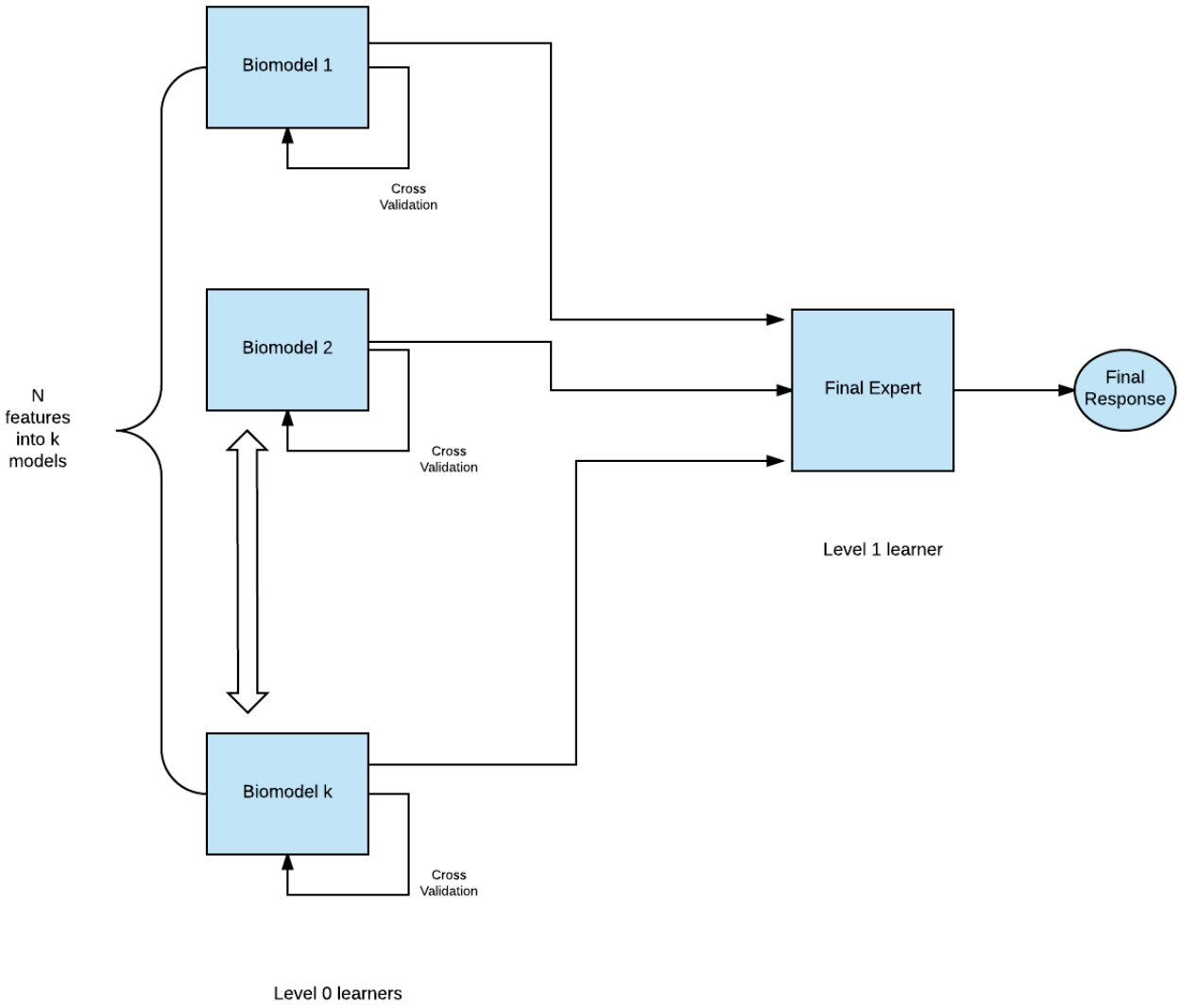

FIGURE 3.3: Models Stacking

2.0 Multimodal Approach

The multimodal approach is explained in details below, an overview can be found in figure 3.5.

2.1 Biomarkers Pathway mapping

Most important proteins and metabolites from biological pathways along with DNA damage data from Comet assay are measured for a number of breast cancer and benign tumor patients. Some of these mapping of protein markers onto pathways is shown in A.1,A.2,A.3 and A.4.

The objective is to find high sub groups within benign tumor patients and develop a strategy to find high risk groups.

2.2 Clustering: Non Negative Matrix Factorization

I had the objective to find clusters with a group of patients with a number of features, and I chose NMF to perform this task as it works well in multivariate data.

Non negative matrix factorization(NMF) factorises a data matrix into the product of two smaller matrices; $V \sim WH$, where V is the data matrix of n x m dimensions, W - n x k and H - m x k dimensions (Lee and Seung, 2001).

V contains the measurements of m factors for n patients, W columns are linear combination of factors from the original matrix, also known as meta-markers. H matrix has the probabilities of each sample belonging to a meta-marker or cluster. Using the probabilities from H matrix a clustering structure can be deduced.

NMF is applied to data for a number of times with different initial conditions and the consensus of the results from the runs is taken as the final result.

3.0 Measuring Prediction Performance and Feature Stability - Common to both approaches

3.1 Measuring Predictive ability of a Prognostic Biomarker

C-index

Prognostic biomarkers measure the survival characteristic or risk of a patient under a particular treatment and is used to determine if the treatment is working well, so this situation is not a direct classification problem, hence AUC may not be used as the measure to evaluate the classifier.

Another important aspect of survival data is censoring i.e, we cannot know from the data whether the event (death, metastasis etc) took place or not within the study period for a number of observations.

Harrel's C-index is defined according to Eq. (1.8). That means it reflects (for non-censored patients) the probability of risk scores being ordered in agreement with survival times.

It is the construct to measure the performance of a classifier; probability for a comparable pair, the sample with higher risk prediction will undergo the event sooner than the other.

$$C := P(R_1 > R_2 | T_1 < T_2) \tag{3.8}$$

where R1 and R2 are the risk value predictions for the samples with survival times T1 and T2.

Receiver Operating Characteristic (ROC) Curve - Area Under the curve

ROC curve was originally developed by electrical engineers in World war II used in analysing radio data to differentiate between enemy aircraft and noise, it was later used in medical imaging in evaluating various classification techniques in detecting tuberculosis (Lusted, 1971).

ROC curve is the plot between true positive rate (sensitivity) in y-axis and false positive rate(1-specificity) in x-axis, measured at different decision thresholds. The area under curve is the area under the ROC curve, it varies from 1 (perfect classification) to 0.5 (random classifier) to 0 (the classifier which is wrong all the time).

Cross-validated C-index /AUC

k fold cross validation strategy divides the complete data of n samples into testing (n-k) and k testing samples. Then the model is built on the n-k samples and testing is performed on rest k samples, and AUC or C-index is measured as appropriate.

The k fold cross validation is a significant part in biomarker discovery and has been illustrated in 3.4.

In order to perform cross validation, the entire data is divided into ten folds, nine folds are used to train the model and then predictions by the model for the tenth fold test data is recorded. The model predictions and the actual responses are used to calculate the ROC curve in case of classification and

C-index in case of survival analysis, several R packages are available to calculate these values; this process is performed 10 times in case of a ten fold cross validation as there are ten sets of training and testing data.

The C-index or AUC is calculated 10 times for a 10 fold cross validation, the mean of these ten outputs is considered as the final value. In this project the models are validated by a 10 x 10 cross validation, i.e. 10 fold cross validation is performed 10 times and the final AUC/C-index is the mean of 10 runs of 10 cross fold validations.

Measuring Predictive ability of a Diagnostic Biomarker

A diagnostic biomarker should able to identify patients from healthy with high accuracy. Conceptually it is a statistical classifier to differentiate between healthy and disease samples. Sensitivity and specificity are the known metrics to measure the performance of a classifier.

Specificity It is the proportion of samples with the condition identified as such. It is also called the true positive rate.

P(T+ | D+) = TP / (TP+FN).

Sensitivity It is the proportion of samples without the condition and test negative. It is also called the true negative rate.

P(T- | D-) = TN / (TN + FP)

		True diagnosis		
		Positive	Negative	Total
Screening test	Positive	a	b	$a+b$
	Negative	c	d	$c+d$
	Total	$a+c$	$b+d$	N

Decision Boundary

The classifier computes class based on the side of the boundary the data point fall into. This boundary is drawn by the value of a decision function chosen appropriately e.g sigmoid function in logistic regression. In a two class boundary the points having function value greater than 0.5 are put in one class and rest in another. The sigmoid function is computed as below, the coefficients are fitted from data using newton raphson method.

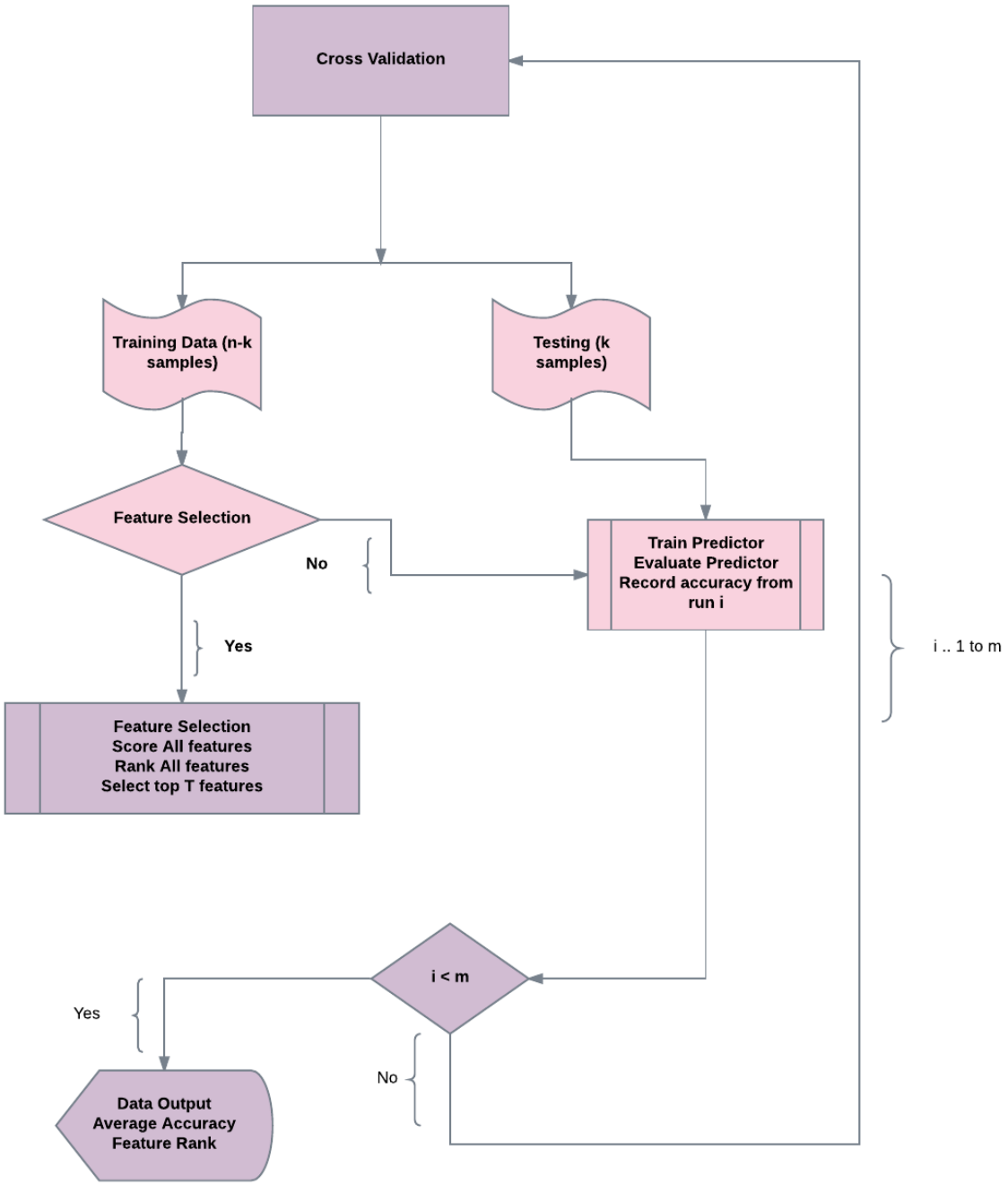

FIGURE 3.4: Cross Validation

$$1/(1 + e^{-(\beta 0 + \beta 1 * x)}) \tag{3.9}$$

3.2 Feature Stability

Feature selection stability over minor change in data is an indicator of suitability of a method for a regression problem, and the identification of frequently chosen features may provide insights into data.

10 x 10 cross validation is performed to determine the prediction performance and feature selection stability is also measured as a part of the same process. The features are extracted from each cross validation fold for 10 x 10 times and the frequencies are noted for the selected features, the top features are plotted for further analysis.

This feature stability is performed for all the extreme current models and multimodal biomarker development.

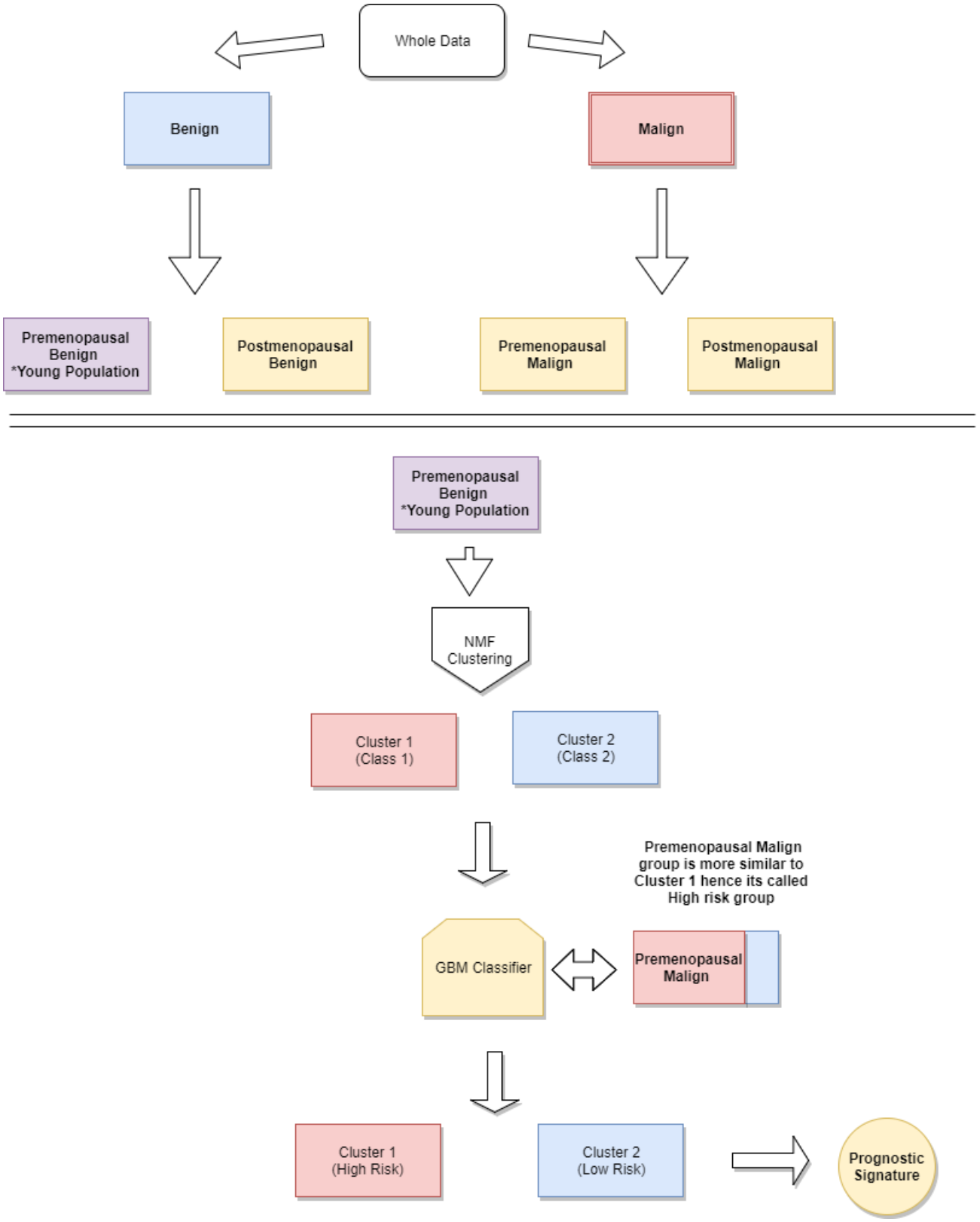

FIGURE 3.5: Overview : Multimodal approach

Chapter 4

Data and Results

This chapter discussed the Results from the application of different methods on three datasets.

In this chapter the Biomodels, Datasets are described first and then the results from applying the methods (from **Methods** chapter) is discussed.

1.0 Biomodels

EBI biomodels database consists of computational and mathematical models of signalling and metabolic networks.

I chose the models based that involved key players and mechanisms in cancer . Models matching the following query terms -TGF Beta, p53, Apoptosis, EGFR, Notch signalling, NFKB, WNT, JAK STAT, MAPK and TGF-beta/SMAD were selected.

The complete list that was finally used in generating features are listed in tables B.1 and B.2.

2.0 Gene Expression Datasets

The extreme current features were generated as explained in Methods chapter and the machine learning procedures from 1.4 Methods chapter is applied on the phenotype and gene-expression data. The following methods were tested :

- **Gradient Boosting**

- **Sparse Group Lasso**
- **Model based Boosting**
- **Elastic net**
- **Stacking**

For each of the set of features, a model is fitted and the accuracy of prediction is performed through a 10 x 10 cross validation.

In addition the performance of these methods were measured against the ground truth. To this end, the four methods that do not use the extreme current strategy or any other apriori information were implemented in addition to extreme current methods. The following non extreme current based methods are tested:

- **Pathway Genes** : The first method simply accounts the expression levels of the genes in the pathway as features.
- **Pathway Activity** : The method computes the value of first principal component of the genes of the pathway and uses it as predictor.
- **Top 100 Genes** : This method uses the top 100 most varying genes as predictors.

For each of the set of features, an elastic net model is fitted and the accuracy of prediction is performed through a 10 x 10 cross validation.

These methods were implemented on the three datasets discussed below.

2.1 Breast Cancer data

The data consists of 295 women with breast cancer with tumor size < 5 cm ((Vijver et al., 2002)). In the patients, 151 were lymph-node-negative and 144 were lymph-node-positive. These patients were followed up for five years. The microarray based gene expression of 24,479 genes of these patients was measured using microarray technology described in (Van't Veer et al., 2002). A R bioconductor conductor package - seventyGeneData is used for the analysis. Time to death is modelled in terms of gene expression covariates.

The mean survival time was 7.863627 years, minimum 0.05 years and maximum 18.34 years. Out of 295 patients, 216 were censored rest 79 non censored.

Out of 24,479 genes, 3,143 were mappable to proteins in biomodels.

2.2 Glioblastoma data

Survival data for glioblastoma was obtained from the TCGA(The cancer genome atlas). After a number of preprocessing steps, the gene expression measure for 342 samples and 10655 features is finalised as the target dataset for our modelling purposes. Again Time to death is chosen as the response variable for our modelling purposes.

93 patients were censored and rest 249 patients were non censored, the mean survival time was 507 days, max was 3880 days and min was 26 days.

Out of 10,655 genes, 2,023 were mappable to proteins in biomodels.

2.3 Prostate Cancer data

The third data set is a case-control data set 48 normal and 47 prostate tumor tissue samples (Börno et al., 2012) and (Brase et al., 2011) and 14960 features. It was downloaded from GEO database.

Out of 14,960 genes, 2,424 were mappable to proteins in biomodels.

3.0 Multi modal Data

The measurements of markers from genomics, proteomics and metabolomics for 180 patients having breast cancer and benign tumors is obtained for further analysis. The data was provided by University Clinic in Bonn.

The multimodal measurements for 180 patients with 37 features has a natural stratification , the data is stratified on the basis of menopausal status. There are 93 malignant tumors and 87 benign tumors. Among those with benign tumors, there are 61 premenopausal and 26 postmenopausal instances. With malignant, there 24 premenopausal and 69 postmenopausal instances.

I investigate the presence of clusters in premenopausal benign patients based on prior knowledge on the subject. The groups with in premenopausal structure would help us in formulating hypothesis on the prognosis of different sub groups in premenopausal benign patients.

4.0 Prediction Performance in Extreme currents based Methods

4.1 Breast Cancer

The results from 10 x 10 cross validation runs are plotted in 4.1. The result show that the extreme currents based methods perform similar to the traditional methods (Using Top 100 genes, all pathway genes and pathway activities). However, the prediction performance has to discussed in light of other criterion like the stability of features and model sparsity.

As already discussed the model stability can be determined by recording the features chosen over 10 x 10 cross validations; a stable model would have the same features selected over and over again leading to a higher feature frequency. Also the lesser number of selected features would make the model more interpretable, so an ideal model would be a sparse model with high frequency of chosen features.

For the purpose of evaluating feature frequency, I plotted the frequencies of top 30 most frequent features for each of the method (4.2). It is observed that GBM and mboost were the most stable models followed by SGL and Stacking.

In order have a complete visualisation, the total number of features were plotted on x-axis and the mean frequency of top 30 most frequent features on y-axis for each of the methods. A complete picture is shown in 4.3, GBM and mboost are stable in terms of features but have quite a high number of features; SGL and stacking on the other case are very sparse as compared to other models but may not be as stable in feature selection as GBM or mboost.

4.2 Glioblastoma data

The prediction performance in Glioblastoma data is about as poor as a random classifier with all the extreme current and traditional methods, the AUC is around 0.5 (See: 4.4). Investigating the feature selection trend revealed high feature selection frequency in GBM and mboost techniques, the feature selection frequencies for sparse models dropped further for SGL and Stacking in Glioblastoma dataset as compared with breast cancer dataset. Hence

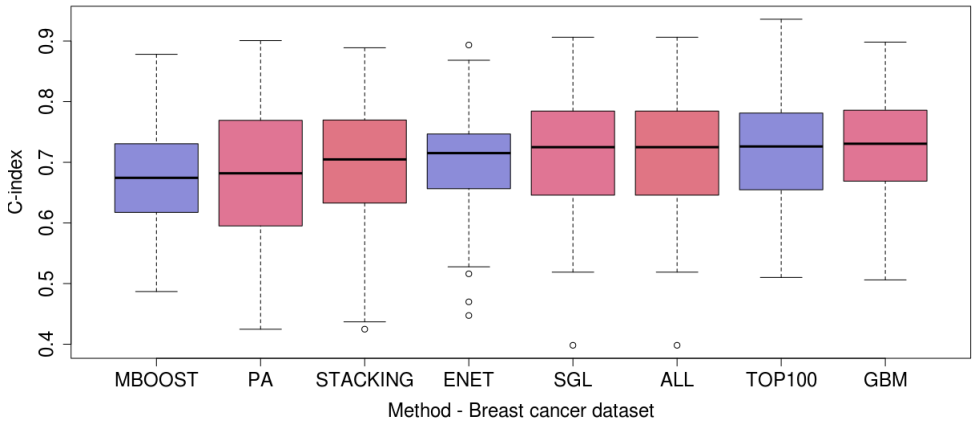

FIGURE 4.1: Breast Cancer Cross-validated AUC, For acronyms refer : 4.1

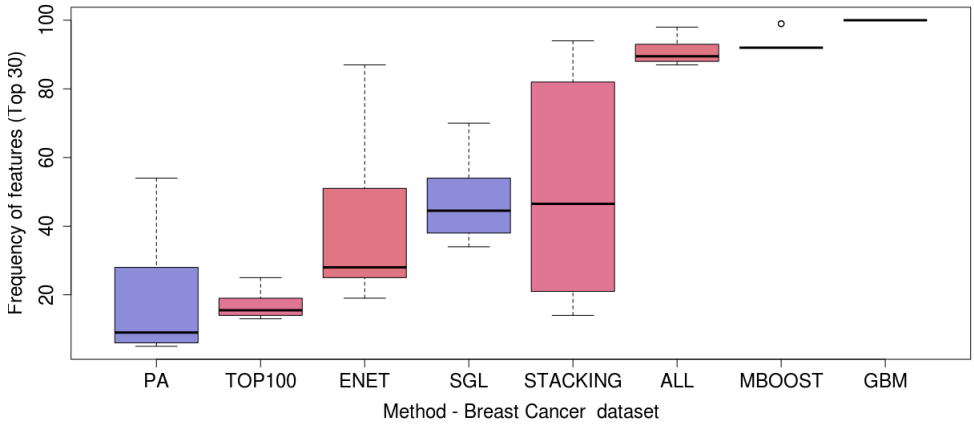

FIGURE 4.2: Breast Cancer Cross-validated Feature Frequency of top 30 most frequently selected features, For acronyms refer : 4.1

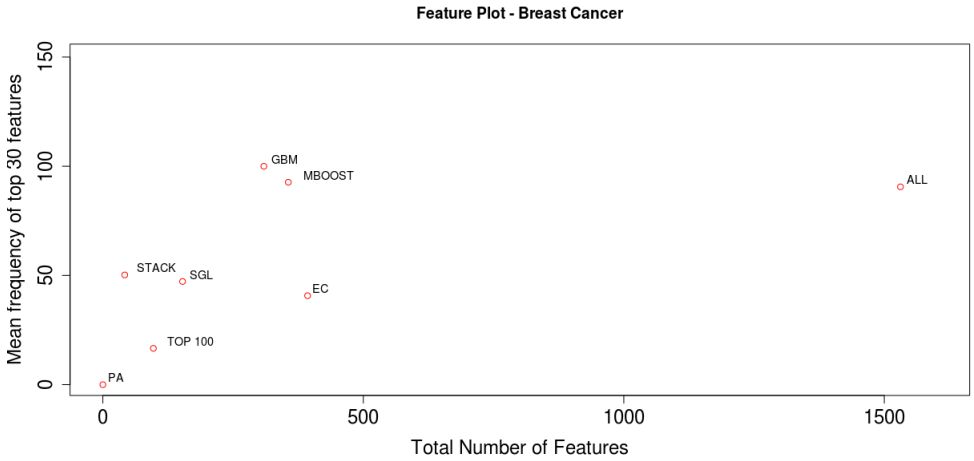

FIGURE 4.3: Breast Cancer - Number of Features vs Feature Frequency , For acronyms refer : 4.1

Short Form	**Expansion**
PA	Pathway Actvity, PCA of all pathway genes
SGL	Sparse Group Lasso
MBOOST	Model Based Boosting
ALL	Elastic net using all pathway mappable genes
ENET	Elastic net using all Extreme currents
TOP100	Elastic net using 100 most varying genes
GBM	Gradient Boosting Machine using all Extreme currents
Stacking	Stacking using models
T	Elastic net using 100 most varying genes
E	Elastic net using all Extreme currents

TABLE 4.1: Acronyms

the same pattern of observation was repeated in glioblastoma. Higher stability and lesser sparsity in GBM and mboost, higher sparsity and lesser stability in SGL and Stacking methods as evident in 4.5 and 4.6.

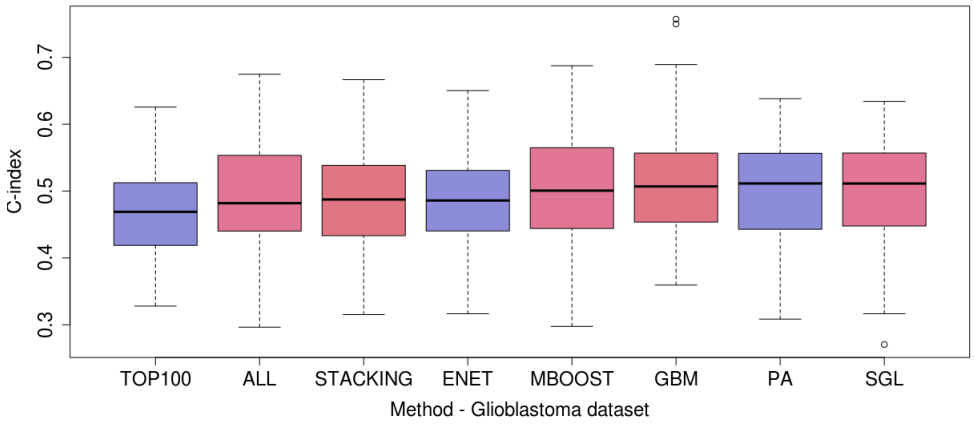

FIGURE 4.4: Glioblastoma Cross-validated AUC, For acronyms refer : 4.1

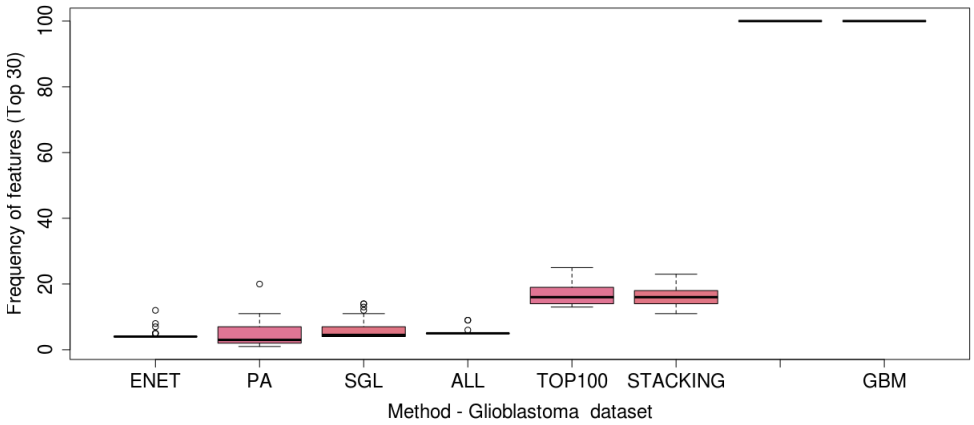

FIGURE 4.5: Glioblastoma Cross-validated Feature Frequency, For acronyms refer : 4.1

4.3 Prostate Cancer data

Glioblastoma and Breast cancer are survival data sets while Prostate cancer is a case control dataset, and the same set of methods were applied to measure the performance of the classifier in differentiating tumor vs healthy.

All the classifiers perform extremely well with AUC approx. 1, so the feature frequencies are compared. The observation pattern continued with GBM and mboost having high feature frequency followed by SGL and Stacking, but GBM and mboost is not as sparse as SGL and Stacking. Please refer figures 4.7, 4.8 and 4.9 for details.

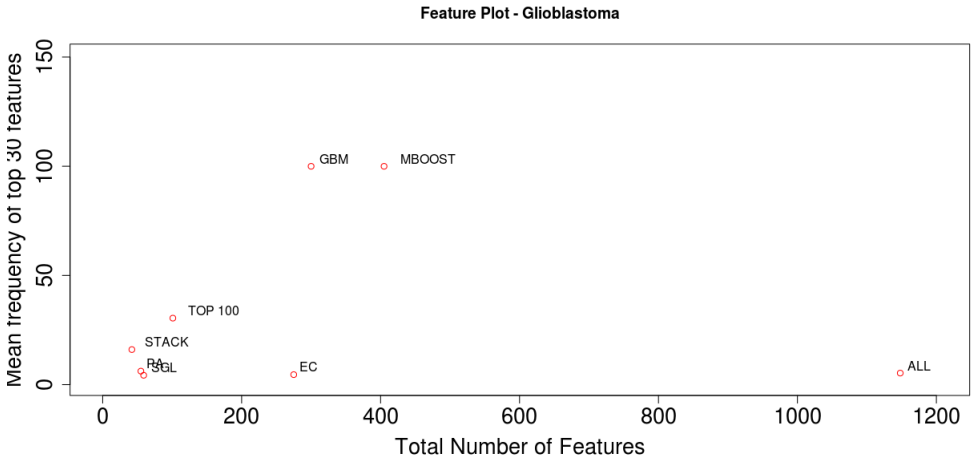

FIGURE 4.6: Glioblastoma - Number of Features vs Feature Frequency, For acronyms refer : 4.1

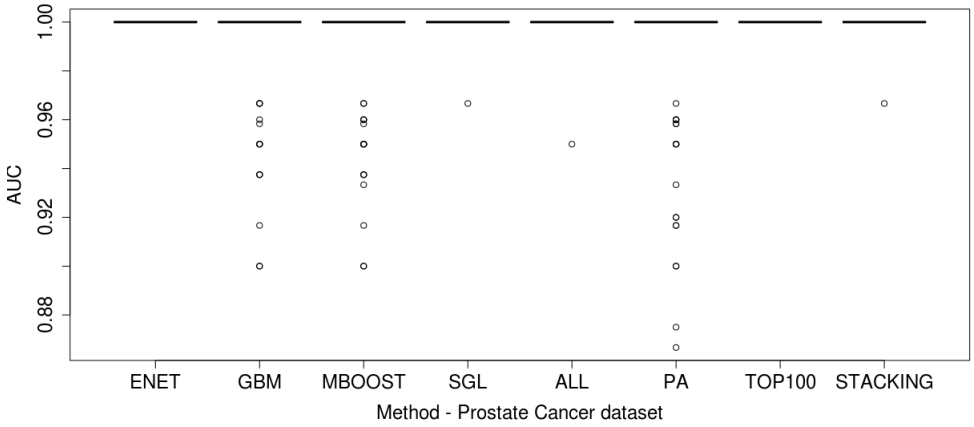

FIGURE 4.7: Prostate cancer Cross-validated AUC, For acronyms refer : 4.1

5.0 Model Interpretation for Breast Cancer

In order to find out which were the most important features in the model, I chose Gradient Boosting model, extracted the coefficients and selected the two most important features for interpretation shown in 4.2.

The Gradient Boosting techniques identified extreme currents from *Stem cell lineage determination Model* (BIOMD0000000209) as most important features. The Model explains the evolution how different lineages arise from the embryonic stem cells (Chickarmane and Peterson, 2008).

The extreme current involves SOX2,OCT4 and GATA6 proteins, the interaction between OCT4 and GATA6 has only been recently discovered from

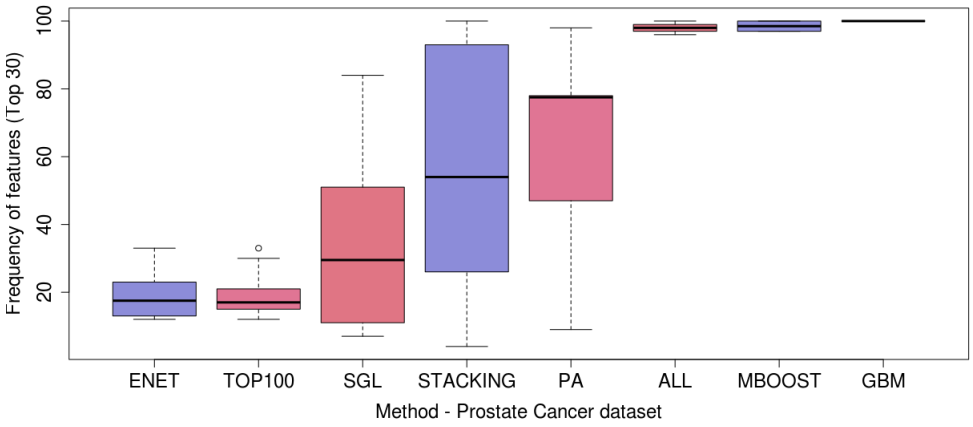

FIGURE 4.8: Prostate cancer cross-validated feature frequency, For acronyms refer : 4.1

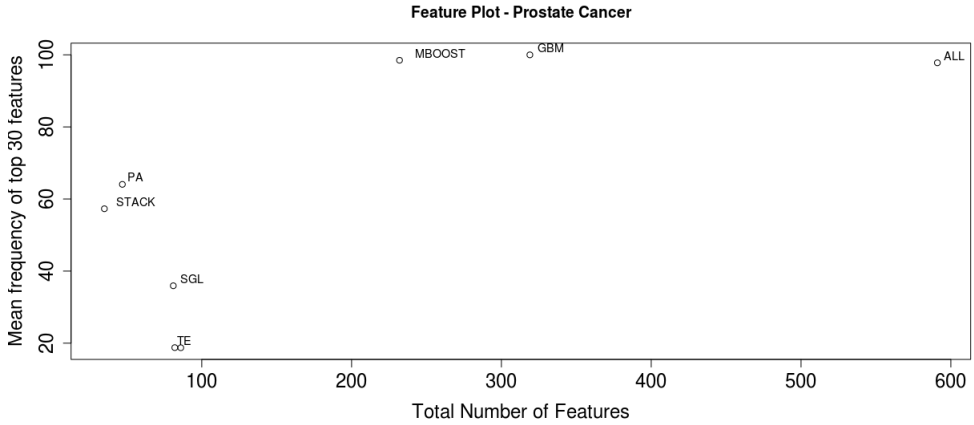

FIGURE 4.9: Prostate cancer - Number of Features vs Feature Frequency, For acronyms refer : 4.1

Marker	**Genes (HGNC Name(s))**	**Pathway**
Extreme Current 1	QSOX2,POU5F1,GATA6	Stem cell lineage determination Model(BIOMD0000000209)
Extreme Current 2	ERBB2	Role of PTEN in Trastuzumab resistance (BIOMD0000000424)

TABLE 4.2: Top Features in Gradient Boosting Machine

ChIP-chip data. It has been phenomenologically observed that strong expression of Oct4 leads to the endoderm lineage, in which Gata-6 is strongly expressed (Chickarmane and Peterson, 2008). So this extreme current measures the extent of endoderm differentiation. The extreme current is visualised in

4.10.

The response variable in the model being time to death (survival time) is strongly correlated to metastasis of breast cancer to other vital organs. There is increasing evidence in the mechanisms involved in differentiation are also involved in tumorigenesis and metastasis (Chou, Provot, and Werb, 2010). Hence this extreme current is significant in determining the metastasis and overall survival and should be investigated as biomarker candidate.

The next extreme current under consideration is from *Role of PTEN in Trastuzumab resistance* Model (BIOMD0000000424) and has only one protein ERBB2 or HER2 which is very well known for its involvement in breast cancer (Gutierrez and Schiff, 2011).

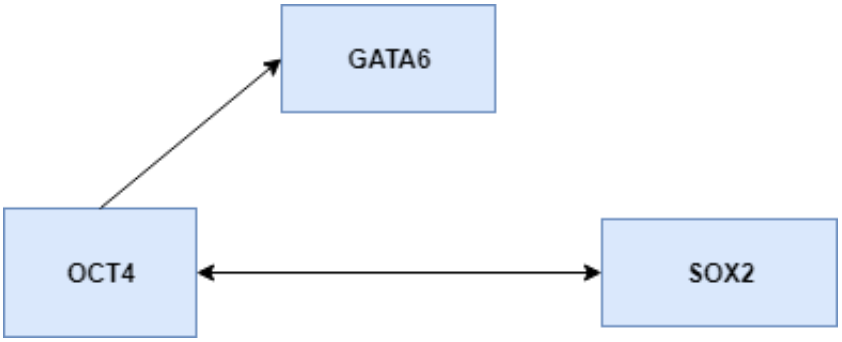

FIGURE 4.10: Visualization of Extreme Currents

6.0 Biomarker discovery in Multimodal data

6.1 Identification of sub clusters

In order to identify high risk groups in young patients, the pre-menopausal patients with benign tumors are selected for the purpose of identification of patients at a higher risk to develop malignant tumor. The NMF clustering method was applied on these premenopausal benign patients (61 of them) with a random start and the clustering was performed 30 times to ensure stability in cluster discovery. During NMF application, the number of clusters was varied from 2 to 10 and silhouette index ((Rousseeuw, 1987)) and cophenetic correlation (Sokal and Rohlf 1962) were noted for clustering results, and it was found that the result with two clusters had the highest silhouette index and cophenetic correlation.

This result is evident in figure 4.11

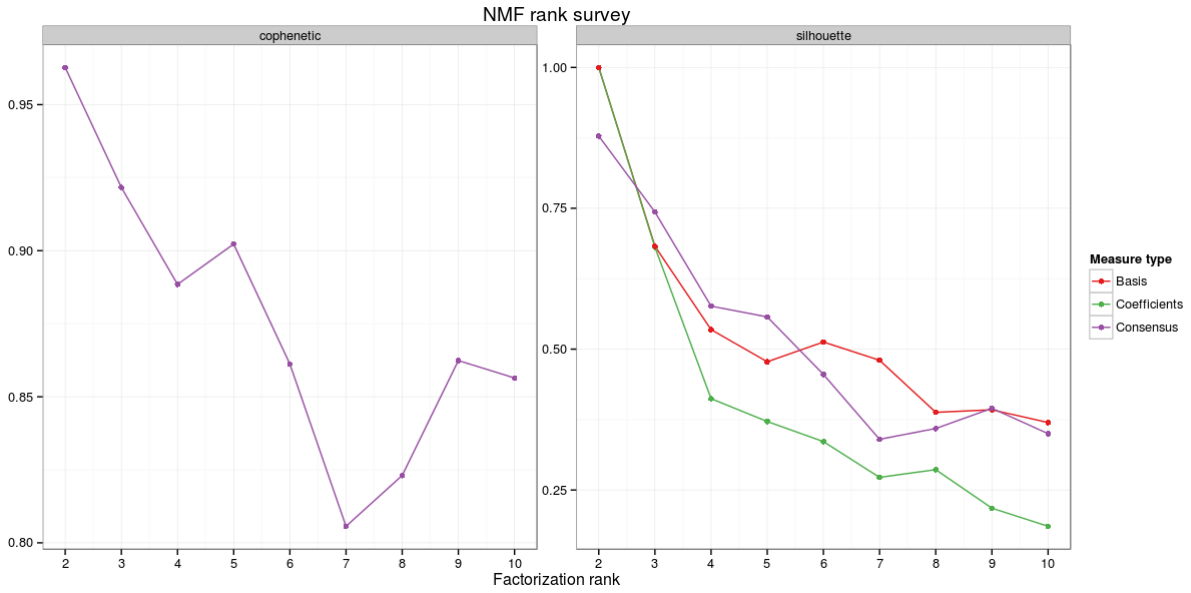

FIGURE 4.11: Silhouette Index and Cophenetic Correlation

The resulting clustering structure is also well supported by the consensus plot which shows the frequencies at which two samples fall in the same group, shown in 4.12. Consensus, cophenetic correlation and silhouette plots also present strong support for the presence of two clusters in pre menopausal benign patients.

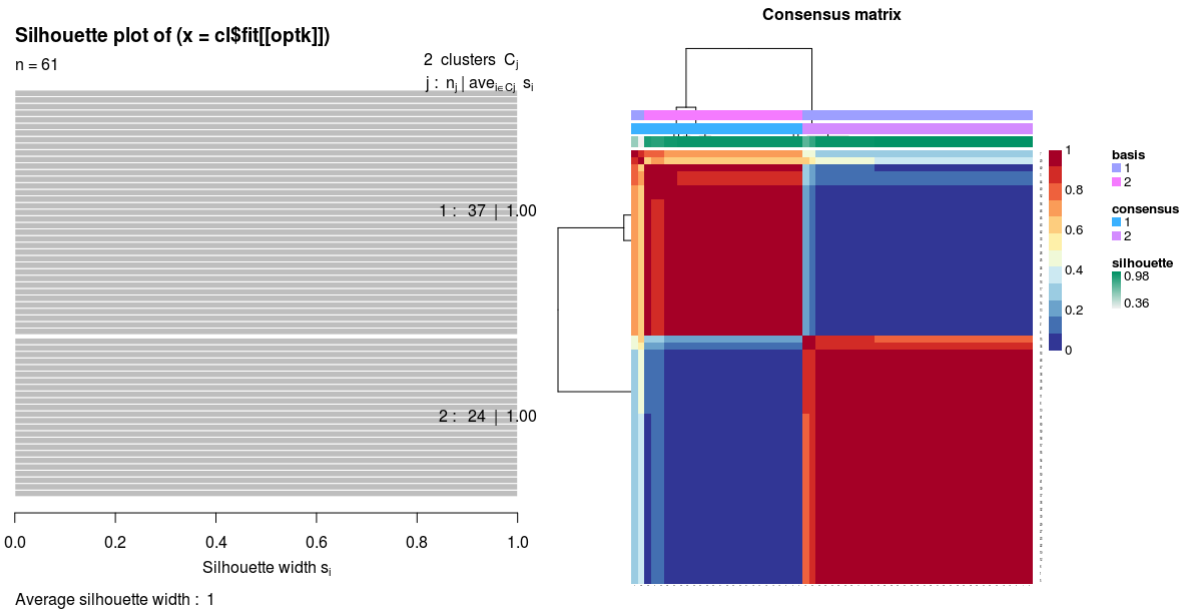

FIGURE 4.12: Consensus Matrix and Silhouette Plot

6.2 Signature Reproducibility

To test the stability of the NMF based marker signature, the NMF based clustering was performed on training data during a 10 fold cross validation performed 10 times, the frequency of features was recorded as in C.1. As observed all of the markers were highly stable except CA Percentage III which is low at 68 percentage.

Hence the signature was highly reproducible.

6.3 Cluster Predictability

In order to test the ability to assign a cluster to a new sample, I design a Gradient Boosting Machine classifier based on two clusters (that were the result of NMF clustering). The classifier is based on gradient boosting procedure as explained in previous chapter, and the optimal number of iterations was chosen via a cross validation technique.

The performance of the GBM classifier was measured through a 10 x 10 fold cross validation. During the cross validation, the data was split into 10 folds and 9 folds of the data was used for training and one fold for testing and this process was done for each fold i.e 10 times; this whole process was repeated 10 times. Each time the AUC was measured and is shown in 4.18.

6.4 Cluster Analysis

The two clusters were obtained as a result of NMF clustering on 61 pre menopausal benign patients. A GBM classifier was constructed on whole 61 patients and their classes, and I tried to classify the 24 premenopausal malign patients through that classifier. It was observed that 14/24 patients (79 percentage) fell in one of the clusters indicating a similarity of cancer patients to this subgroup, hence I termed this particular sub-group as high risk group.

The enrichment of high risk subgroup (green) with cancer patients(red) is shown in the PCA plot (4.13).

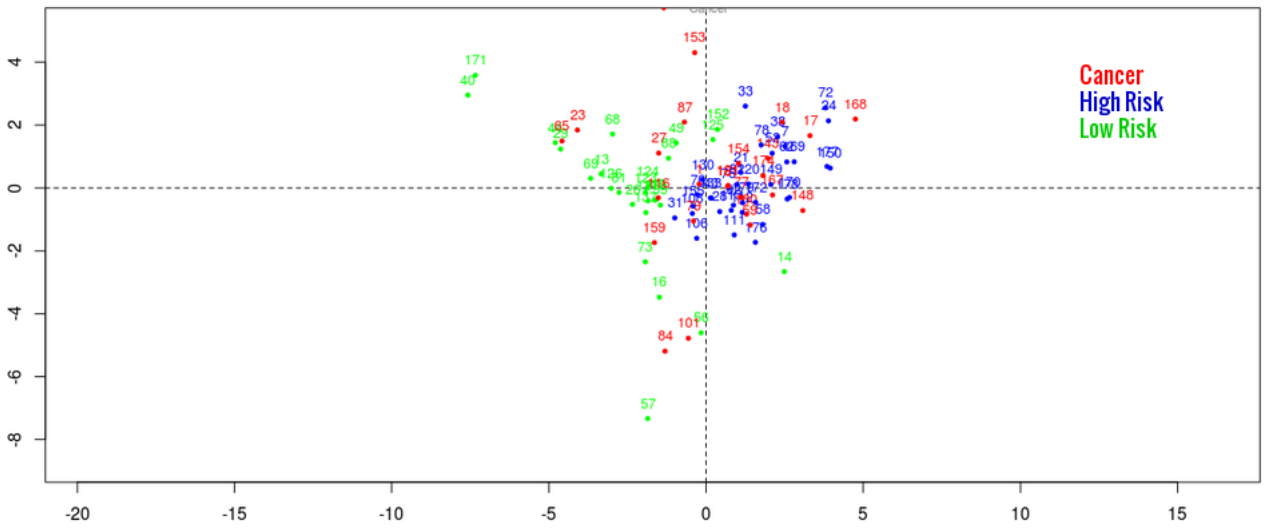

FIGURE 4.13: PCA Plot High vs Low risk clusters

6.5 High Risk Vs Low Risk clusters

I investigated little bit more into the markers separating High risk and low risk clusters by analysing the marker selection frequency over the 10 x 10 cross validation and the relative importance of the marker (C.1, C.6). The absolute values of the significant markers ($p < 0.05$) were also investigated (4.14 and 4.15) which all led to the same biomarker signature.

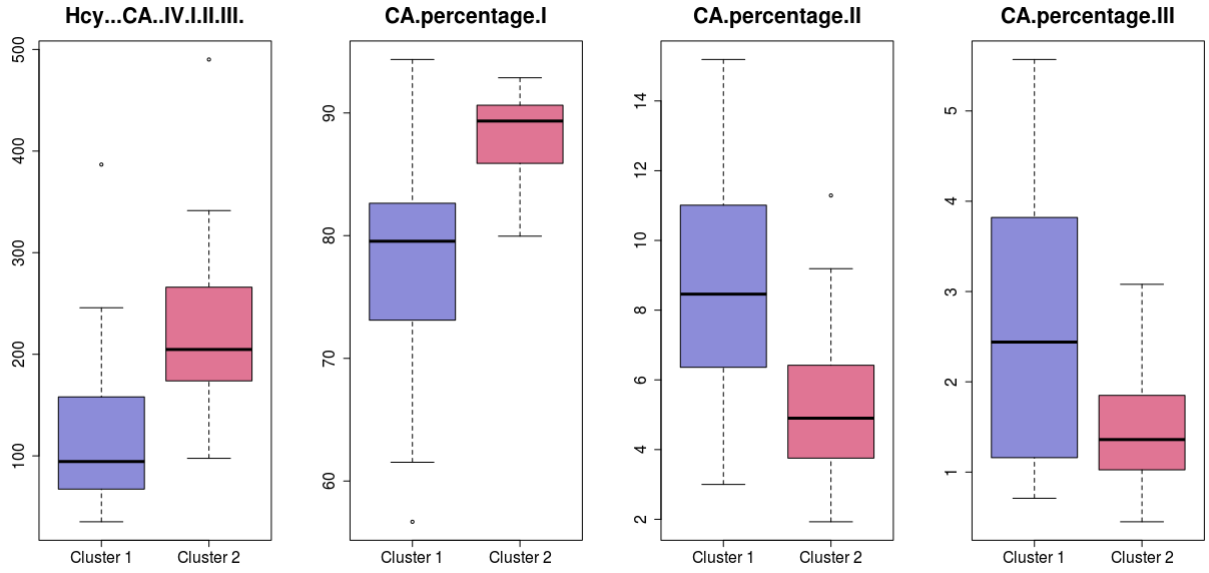

FIGURE 4.14: Marker values - High vs Low risk clusters

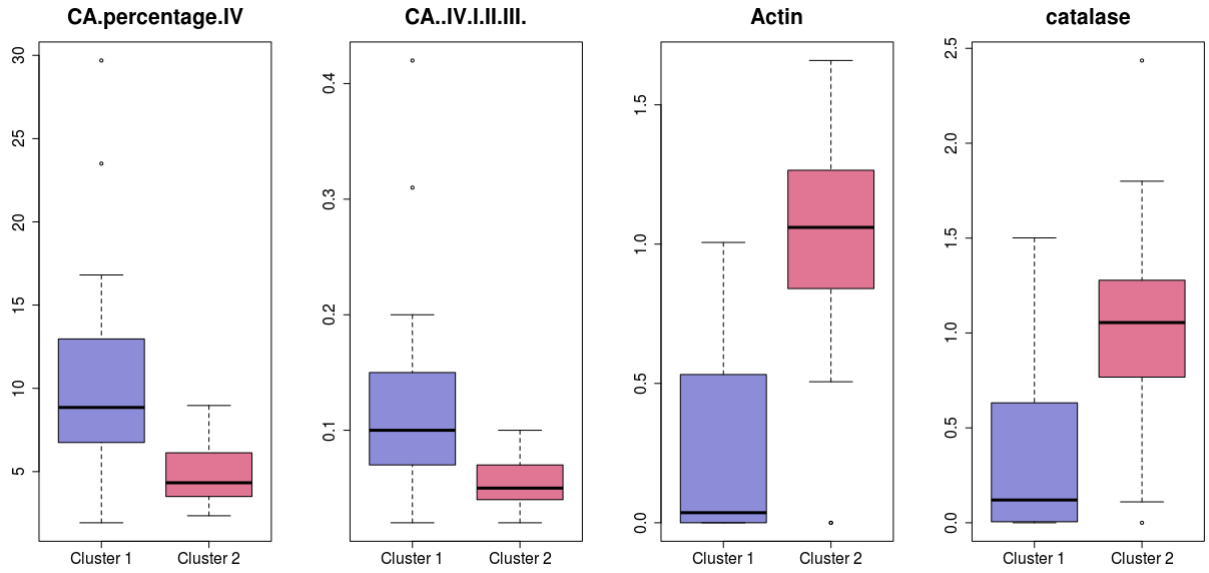

FIGURE 4.15: Marker values - High vs Low risk clusters(Contd.)

6.6 High Risk and Low Risk Groups Vs Postmenopausal Benign Patients

In order to investigate how the high and low risk clusters correspond to postmenopausal benign patients, I plotted a PCA plot for all premenopausal benign and postmenopausal benign patients.

As can be seen in 4.16, the postmenopausal benign patients follow the clustering structure that I established for premenopausal benign patients, this suggests that the clustering pattern might be independent of the menopausal status.

6.7 High Risk and Low Risk Groups Vs Malign Patients

With the goal to compare the low and high risk groups to malign patients, I built GBM classifiers separating high risk from malign patients and low risk from malign patients and evaluated the classifier's potential through a 10 x 10 cross validation.

The GBM classifier was able to separate the low risk group from the malign patients with 77 percentage AUC and the high risk group from the malign patients with 67 percentage AUC (AUC Results :4.17), which validates the dissimilarity of low risk patients with malign patients.

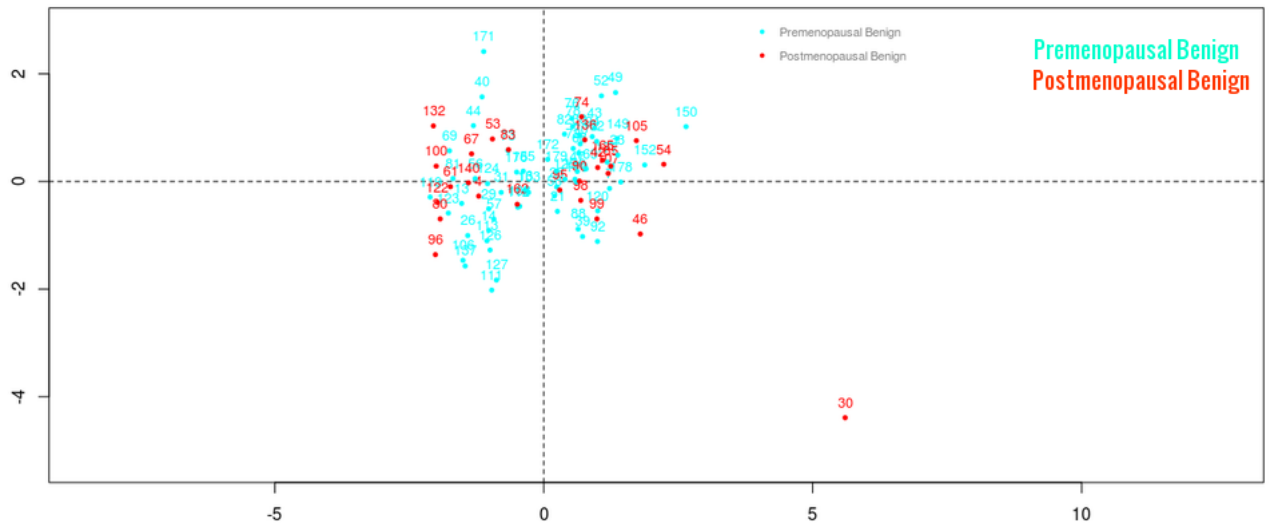

FIGURE 4.16: PCA of pre-menopausal (blue) and postmenopausal (red) benign patients

The relative importance and frequency of markers in High risk Vs Malign is shown in C.3 and the same table for Low risk Vs Malign is C.5.

Malign group has a lower SOD.2 (lower antioxidant activity), lower CA.Percentage I (lower undamaged DNA) and higher CA.Percentage II (higher damaged DNA) than the high risk group. Malign group has a higher Actin level than the low risk group.

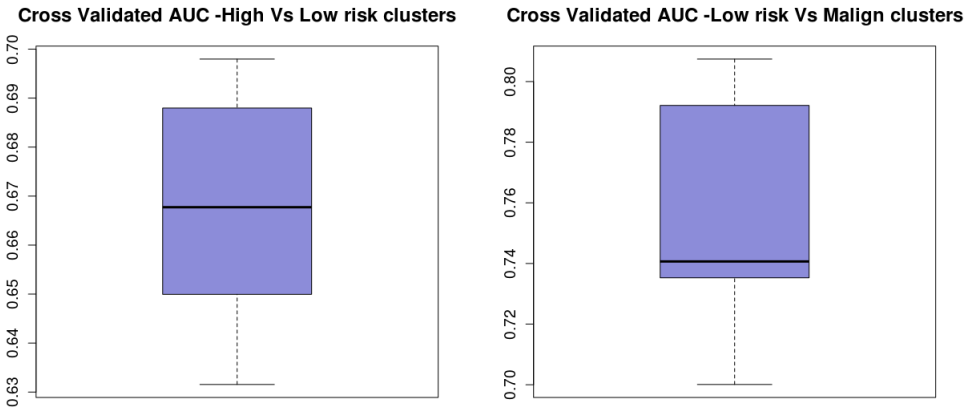

FIGURE 4.17: Cross validated AUC - A) High risk group Vs Malign B) Low risk group Vs Malign

6.8 Overall Benign vs Malignant Patients

Next, I was interested in seeing how different are the benign patients from the malignant patients. I excluded age as a predictor as I wanted to see the classifiers primary on molecular markers, I constructed two GBM classifiers one with menopausal status as predictor and the other one without it.

The GBM classifier for separating Benign Vs Malignant patients have a 60 percentage AUC as compared with 70 percentage AUC of GBM classifiers in differentiating low risk cluster and malignant patients and 80 percentage AUC in differentiating high risk cluster and malignant patients. The classifier weighed heavily on homocysteine and Comet Assay Percentage II (Ref: 4.19),

The relative marker importance for the GBM classifier without menopausal status in differentiating overall benign vs malign is shown in C.4, and the figures considering menopausal status is C.2.

Hcy and CA.Percentage II features were identified to be of high importance between benign and malign patients, Hcy or Homocysteine was found high in malignant patients indicating a poorer health state and lower levels of SOD..2 indicates overall decreased levels of this anti-oxidant in malignant group.

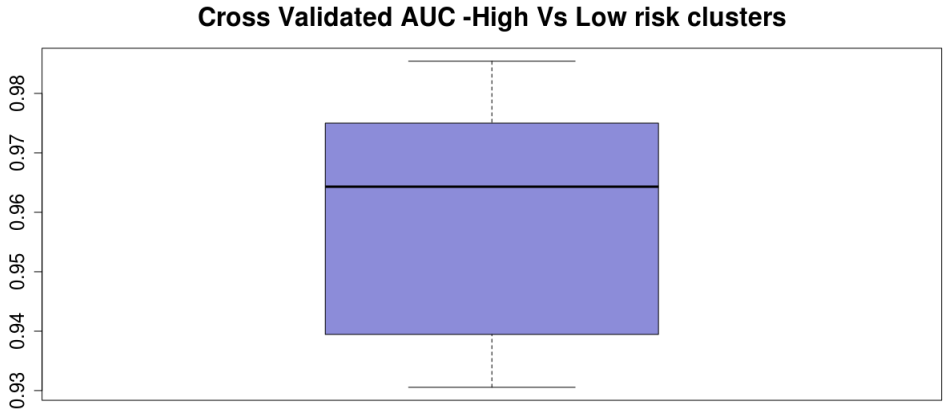

FIGURE 4.18: Cross validated AUC GBM - High vs Low risk clusters

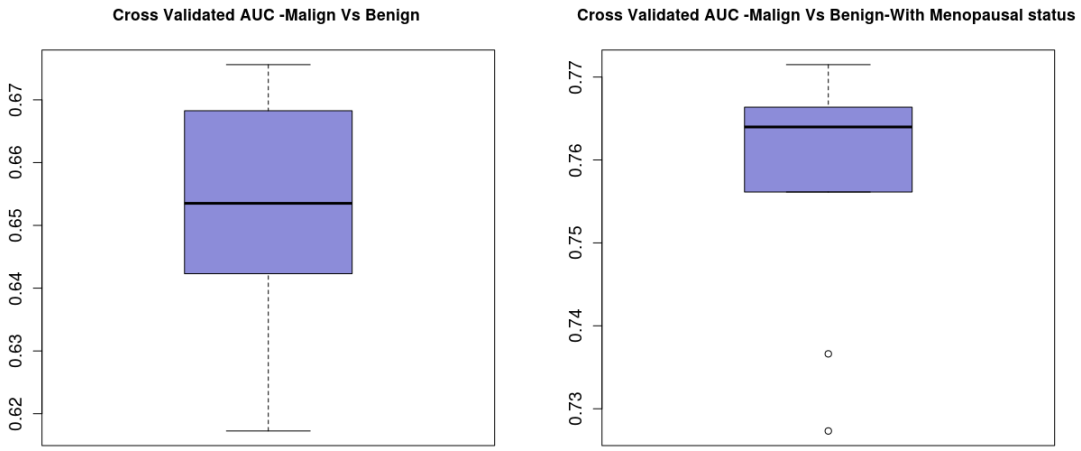

FIGURE 4.19: Boxplot of AUC of GBM - Benign Vs Malign, from 10x10 cross validation : A) With menopausal status B) Without menopausal status

Chapter 5

Conclusion and Future Directions

I will conclude the results and overall thesis project in this chapter, followed by a discussion of possible future directions.

1.0 Conclusions

The main objective was to investigate mechanism based biomarkers from a methodological point of view as well as from an application perspective. In order to acheive this objective, I evaluated the feature stability, sparsity and predictability of extreme currents based methods and compare it against the traditional methods.

1.1 Feature Stability and Sparsity

10 x 10 fold cross validation was applied on all the extreme current and other methods to find out how stable the feature selection is for these methods, i.e we want to find out how much the selected set of features vary between runs. The lesser variance between the set of selected features indicates a stable or reproducible model.

In breast cancer data results, all the extreme current methods had a higher feature stability than top 100 and pathway activity traditional methods 4.2.

The traditional method with an Elastic net model based on pathway mappable genes had a higher feature selection stability than Stacking, SGL and Elastic net extreme currents methods but lower than mboost and GBM extreme currents methods. However the total number of features is very high in this particular traditional method which is a big disadvantage.

In prostate cancer data results, stability of extreme currents based method was comparable with the traditional methods. Again the traditional method with all pathway genes (Elastic net model on all pathway mappable genes) had a higher stability in feature selection than Stacking, SGL and Elastic net extreme currents methods but lower than mboost and GBM extreme currents method, but as discussed earlier this method which takes in account all genes in the pathway has a very high number of features.

Finally in glioblastoma data results the feature stability of most of the extreme current and traditional methods had low feature stability, except mboost and GBM methods.

Extreme currents based mboost and GBM methods has a higher stability than elastic net based methods based on top 100 genes, all pathway mappable genes and pathway activities. SGL and Stacking based extreme current methods had a good combination of feature stability and sparsity.

1.2 Predictability

It was found that most of the extreme currents based methods perform about as good as the considered traditional methods.

1.3 Multimodal Data

The analysis of multimodal breast cancer data resulted in the identification of high and low risk clusters among premenopausal benign patients, high risk cluster members are hypothesized to be at a greater risk to develop breast cancer.

Eventually this separation of two groups also lead to the identification of markers involved in the NMF clustering signature, these markers were also validated through a manual inspection of the values of these markers between the groups and a GBM classifier. Consequently a highly predictive prognostic signature was developed which could differentiate between high and low risk samples.

2.0 Future Directions

The extreme currents methods can be extended to other datasets such as colorectal cancer and other TCGA projects.

The extreme currents approach in this thesis is based on reaction graphs that are available for detailed computational models. However most biochemical information usually is not available in such a detailed level, e.g. as static pathway diagrams in KEGG. Hence a tool is needed to transform pathway diagrams to reaction graphs, so that the mathematical analytic methods such as extreme currents can be applied.

Hence the development of such a tool would make the information in many pathway databases accessible and we can apply extreme current methods to them.

A number of reactants,products and modifiers in biomodels were not mappable to genes due to lack of any strategy in place to map those annotations to genes, methods can be developed to resolve this issue so that more information can be integrated into extreme current features. One of the strategies to resolve the annotations to specific proteins is literature search, i.e matching the annotation with the most appropriate protein based on the information from the annotation and interacting proteins in the computational model.

Currently the boosting procedure uses a decision tree as base learner, this can also be extended to different base learners.

Reproducibility is an important objective for biomarkers, hence the mechanism based biomarker strategies explored in this study can be tested for reproducibility on similar studies.

Another main objective of mechanism based biomarkers is to improve the interpretability of the models as the traditional models can be quite difficult to interpret. In future the major and most important extreme currents responsible for the phenotype can be identified and studied in details, these types of analyses may have application in studying drug responses as well apart from identifying prognostic biomarkers.

Finally, the multimodal signature from the breast cancer study can be validated in trials to be of any clinical utility.

Bibliography

All, Yeghiazaryan et (2014). "Innovative strategies for prediction and targeted prevention of glaucoma in healthy vasospastic individuals: context of neurodegenerative pathologies". In: *The EPMA Journal* 5.Suppl 1, A99.

Bayouth, John E. et al. (2011). "Image-based Biomarkers in Clinical Practice". In: URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4270476/.

Bebek, Gurkan et al. (2012). "Network biology methods integrating biological data for translational science". In: *Briefings in bioinformatics* 13.4, pp. 446–459.

Bock, Christoph (2009). "Epigenetic biomarker development". In:

Börno, Stefan T et al. (2012). "Genome-wide DNA methylation events in TMPRSS2–ERG fusion-negative prostate cancers implicate an EZH2-dependent mechanism with miR-26a hypermethylation". In: *Cancer discovery* 2.11, pp. 1024–1035.

Brase, Jan C et al. (2011). "TMPRSS2-ERG-specific transcriptional modulation is associated with prostate cancer biomarkers and TGF-β signaling". In: *BMC cancer* 11.1, p. 507.

Buyse, Marc et al. (2006). "Validation and clinical utility of a 70-gene prognostic signature for women with node-negative breast cancer". In: *Journal of the National Cancer Institute* 98.17, pp. 1183–1192.

Casadevall, Arturo and Ferric C Fang (2010). *Reproducible science*.

Cha, Mee-Kyung, Kyung-Hoon Suh, and Il-Han Kim (2009). "Overexpression of peroxiredoxin I and thioredoxin1 in human breast carcinoma". In: *Journal of Experimental & Clinical Cancer Research* 28.1, p. 93.

Chang, Hae Ryung et al. (2016). "Systematic approach identifies RHOA as a potential biomarker therapeutic target for Asian gastric cancer". In: *Oncotarget* 7.49, pp. 81435–81451.

Chatterjee, Bijoya, Jigisha Pancholi, et al. (2011). "Prakriti-based medicine: A step towards personalized medicine". In: *AYU (An international quarterly journal of research in Ayurveda)* 32.2, p. 141. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3296331/.

Chelliah, Vijayalakshmi, Camille Laibe, and Nicolas Le Novere (2013). "BioModels database: a repository of mathematical models of biological processes". In: *In silico Systems Biology*, pp. 189–199.

Chickarmane, Vijay and Carsten Peterson (2008). "A computational model for understanding stem cell, trophectoderm and endoderm lineage determination". In: *PLoS one* 3.10, e3478.

Cho, Sang-Hoon, Jongsu Jeon, and Seung Il Kim (2012). "Personalized Medicine in Breast Cancer: A Systematic Review". In: URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3468779/.

Chou, Jonathan, Sylvain Provot, and Zena Werb (2010). "GATA3 in development and cancer differentiation: cells GATA have it!" In: *Journal of cellular physiology* 222.1, pp. 42–49.

Chowdhury, Salim Akhter (2010). "Identification of coordinately dysregulated subnetworks in complex phenotypes". PhD thesis. Case Western Reserve University.

Chuang, Han-Yu et al. (2007). "Network-based classification of breast cancer metastasis". In: *Molecular systems biology* 3.1, p. 140.

Clarke, Bruce L (1988). "Stoichiometric network analysis". In: *Cell Biochemistry and Biophysics* 12.1, pp. 237–253.

Cun, Yupeng and Holger Fröhlich (2013). "Network and data integration for biomarker signature discovery via network smoothed t-statistics". In: *PloS one* 8.9, e73074.

Daly, Ann K (2013). "Pharmacogenomics of adverse drug reactions". In: *Genome medicine* 5.1, p. 5.

Dao, Phuong et al. (2010). "Inferring cancer subnetwork markers using density-constrained biclustering". In: *Bioinformatics* 26.18, pp. i625–i631.

Davies, Jane C et al. (2013). "Efficacy and safety of ivacaftor in patients aged 6 to 11 years with cystic fibrosis with a G551D mutation". In: *American journal of respiratory and critical care medicine* 187.11, pp. 1219–1225.

De Jong, Ebbing P et al. (2010). "Quantitative proteomics reveals myosin and actin as promising saliva biomarkers for distinguishing pre-malignant and malignant oral lesions". In: *PloS one* 5.6, e11148.

Drouin, Alexandre et al. (2016). "Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons". In: *BMC genomics* 17.1, p. 754.

Engelhardt, Benjamin et al. (2017). "Modelling and mathematical analysis of the M receptor-dependent joint signalling and secondary messenger network in CHO cells". In: *Mathematical medicine and biology: a journal of the IMA*, dqx003.

Frantzi, Maria, Akshay Bhat, and Agnieszka Latosinska (2014). "Clinical proteomic biomarkers: relevant issues on study design & technical considerations in biomarker development". In: *Clinical and translational medicine* 3.1, p. 7.

Friedman, Jerome H. (2000). "Greedy Function Approximation: A Gradient Boosting Machine". In: *Annals of Statistics* 29, pp. 1189–1232.

Golden, Daniel I. et al. (2013). "Qualitative and quantitative image-based biomarkers of therapeutic response in triple-negative breast cancer." In: URL: https://www.ncbi.nlm.nih.gov/pubmed/24303300.

Gutierrez, Carolina and Rachel Schiff (2011). "HER2: biology, detection, and clinical implications". In: *Archives of pathology & laboratory medicine* 135.1, pp. 55–62.

Ierardi, Daniela Filippini et al. (2013). "Homocysteine as a Biomarker for Predicting Disease-Free Survival in Breast Cancer". In:

Jikimoto, Takumi et al. (2002). "Thioredoxin as a biomarker for oxidative stress in patients with rheumatoid arthritis". In: *Molecular immunology* 38.10, pp. 765–772.

Jones, MP et al. (2014). "A biomarker panel and psychological morbidity differentiates the irritable bowel syndrome from health and provides novel pathophysiological leads". In: *Alimentary pharmacology & therapeutics* 39.4, pp. 426–437.

Kanehisa, Minoru et al. (2011). "KEGG for integration and interpretation of large-scale molecular data sets". In: *Nucleic acids research* 40.D1, pp. D109–D114.

Kapalla, Marko et al. (2016). "Medicine in the early twenty-first century: paradigm and anticipation-EPMA position paper 2016". In:

LaFramboise, Thomas (2009). "Single nucleotide polymorphism arrays: a decade of biological, computational and technological advances". In: *Nucleic acids research*, gkp552.

Lee, Daniel D and H Sebastian Seung (2001). "Algorithms for non-negative matrix factorization". In: *Advances in neural information processing systems*, pp. 556–562.

Lusted, Lee B (1971). "Signal detectability and medical decision-making". In: *Science* 171.3977, pp. 1217–1219.

Mehta, Sunali and Andrew Shelling (2010). "Predictive and prognostic molecular markers for cancer medicine". In: URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3126011/.

Miar, Ana et al. (2015). "Manganese superoxide dismutase (SOD2/MnSOD)/catalase and SOD2/GPx1 ratios as biomarkers for tumor progression and metastasis in prostate, colon, and lung cancer". In: *Free Radical Biology and Medicine* 85, pp. 45–55.

Mujagic, Zlatan et al. (2016). "A novel biomarker panel for irritable bowel syndrome and the application in the general population". In: *Scientific reports* 6.

Novelli, Giuseppe et al. (2008). "Genetic tests and genomic biomarkers: regulation, qualification and validation". In: *Clinical cases in mineral and bone metabolism* 5.2, p. 149.

Paulovich, Amanda G et al. (2008). "The interface between biomarker discovery and clinical validation: the tar pit of the protein biomarker pipeline". In: *Proteomics-Clinical Applications* 2.10-11, pp. 1386–1402.

Ren, Fanghui et al. (2015). "Overexpression of MMP family members functions as prognostic biomarker for breast cancer patients: a systematic review and meta-analysis". In: *PloS one* 10.8, e0135544.

Rousseeuw, Peter J. (1987). "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis". In: *Journal of Computational and Applied Mathematics* 20, pp. 53 –65. ISSN: 0377-0427. DOI: http://dx.doi.org/10.1016/0377-0427(87)90125-7. URL: http://www.sciencedirect.com/science/article/pii/0377042787901257.

Samal, Fröhlich (2017). "Linking Metabolic Network Features to Phenotypes using Sparse Group Lasso". In: *Bioinformatics*.

Samal, Satya, Hassan Errami, and Andreas Weber (2012). "PoCaB: a software infrastructure to explore algebraic methods for bio-chemical reaction networks". In: *Computer Algebra in Scientific Computing*. Springer, pp. 294–307.

Schulte, Janin (2011). "Peroxiredoxin 4: a multifunctional biomarker worthy of further exploration". In: *BMC medicine* 9.1, p. 137.

Sève, Pascal et al. (2007). "Class III β-tubulin expression and benefit from adjuvant cisplatin/vinorelbine chemotherapy in operable non–small cell lung cancer: analysis of NCIC JBR. 10". In: *Clinical Cancer Research* 13.3, pp. 994–999.

Simon, Noah et al. (2013). "A sparse-group lasso". In: *Journal of Computational and Graphical Statistics* 22.2, pp. 231–245.

Teschendorff, Andrew E et al. (2010). "Improved prognostic classification of breast cancer defined by antagonistic activation patterns of immune response pathway modules". In: *BMC cancer* 10.1, p. 604.

Van't Veer, Laura J et al. (2002). "Gene expression profiling predicts clinical outcome of breast cancer". In: *nature* 415.6871, pp. 530–536.

Veríssimo, André et al. (2016). "DegreeCox–a network-based regularization method for survival analysis". In: *BMC bioinformatics* 17.16, p. 449.

Vijver, Marc J. van de et al. (2002). "A Gene-Expression Signature as a Predictor of Survival in Breast Cancer". In: *New England Journal of Medicine* 347.25. PMID: 12490681, pp. 1999–2009. DOI: 10.1056/NEJMoa021967. eprint: http://dx.doi.org/10.1056/NEJMoa021967. URL: http://dx.doi.org/10.1056/NEJMoa021967.

Wagner, Clemens and Robert Urbanczik (2005). "The geometry of the flux cone of a metabolic network". In: *Biophysical journal* 89.6, pp. 3837–3845.

Wang, Rui-Sheng and Réka Albert (2011). "Elementary signaling modes predict the essentiality of signal transduction network components". In: *BMC systems biology* 5.1, p. 44.

Wang, Yemin et al. (2007). "Prognostic significance of nuclear ING3 expression in human cutaneous melanoma". In: *Clinical cancer research* 13.14, pp. 4111–4116.

Wei, Zhi and Hongzhe Li (2007). "Nonparametric pathway-based regression models for analysis of genomic data". In: *Biostatistics* 8.2, p. 265. DOI: 10.1093/biostatistics/kxl007. eprint: /oup/backfile/content_public/journal/biostatistics/8/2/10.1093/biostatistics/kxl007/2/kxl007.pdf. URL: +http://dx.doi.org/10.1093/biostatistics/kxl007.

Wolpert, David H (1992). "Stacked generalization". In: *Neural networks* 5.2, pp. 241–259.

Zhou, Jian-rong et al. (2008). "Identification of Tumor-Associated Proteins in Well Differentiated Laryngeal Squamous Cell Carcinoma by Proteomics". In: *Clinical Proteomics* 3.1, p. 42.

Zoidakis, Jerome et al. (2012). "Profilin 1 is a potential biomarker for bladder cancer aggressiveness". In: *Molecular & Cellular Proteomics* 11.4, pp. M111–009449.

Zou, Hui and Trevor Hastie (2005). "Regularization and variable selection via the elastic net". In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 67.2, pp. 301–320.

Appendix A

Biological Pathway Mappings

Selected pathways highlighting the proteins chosen for multimodal analysis based on it's importance within the pathway.

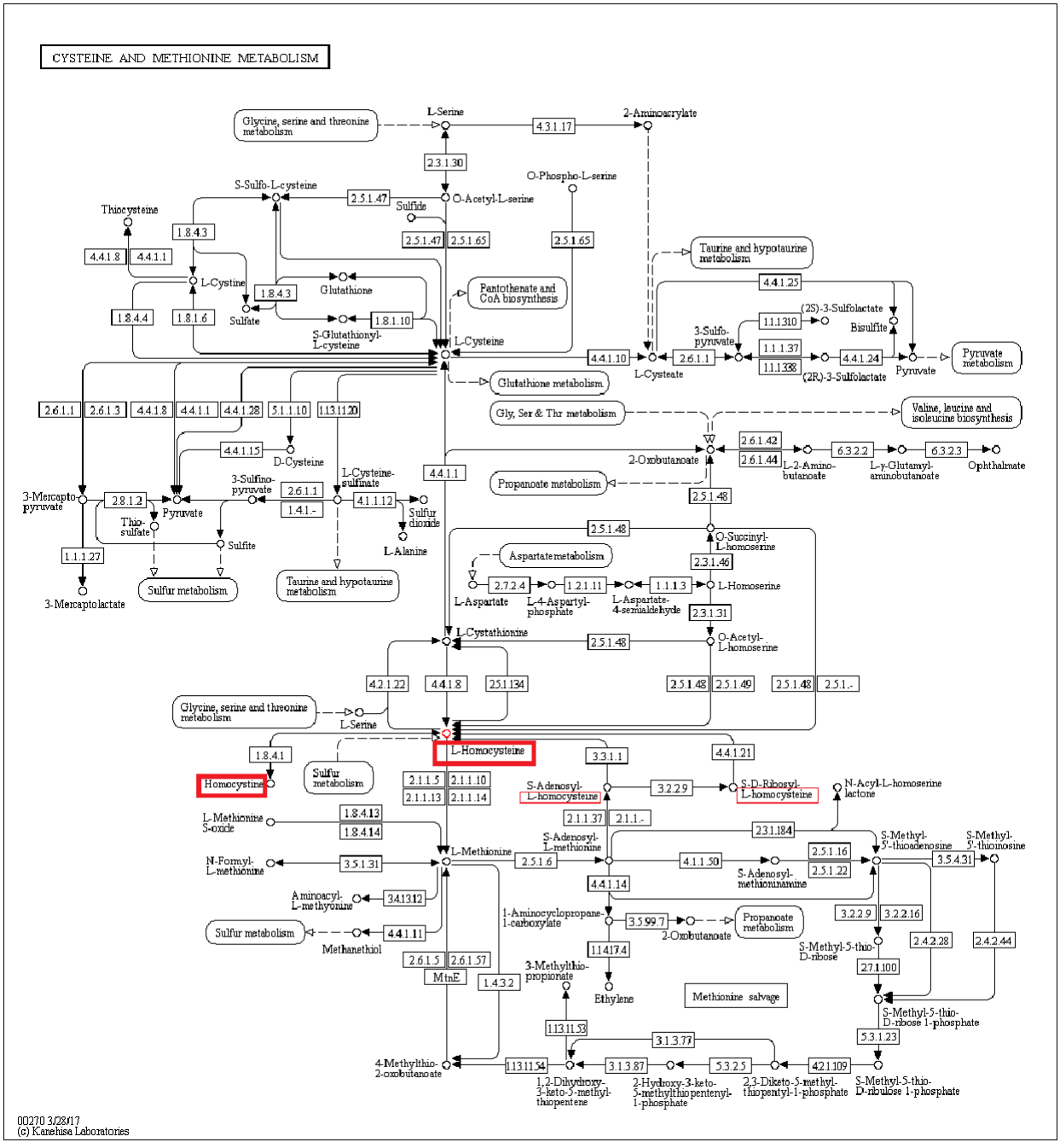

FIGURE A.1: Homocysteine

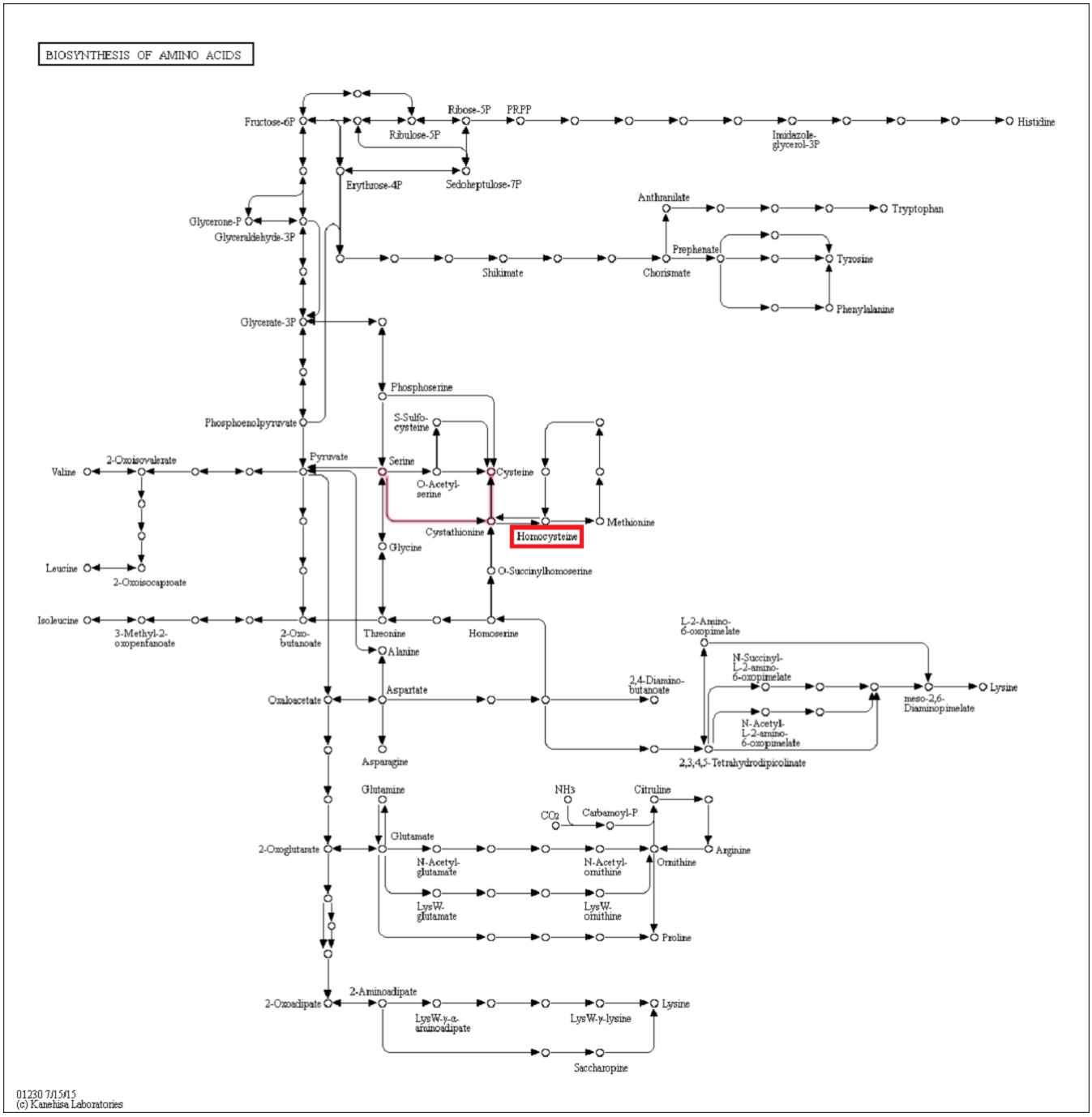

FIGURE A.2: Homocysteine

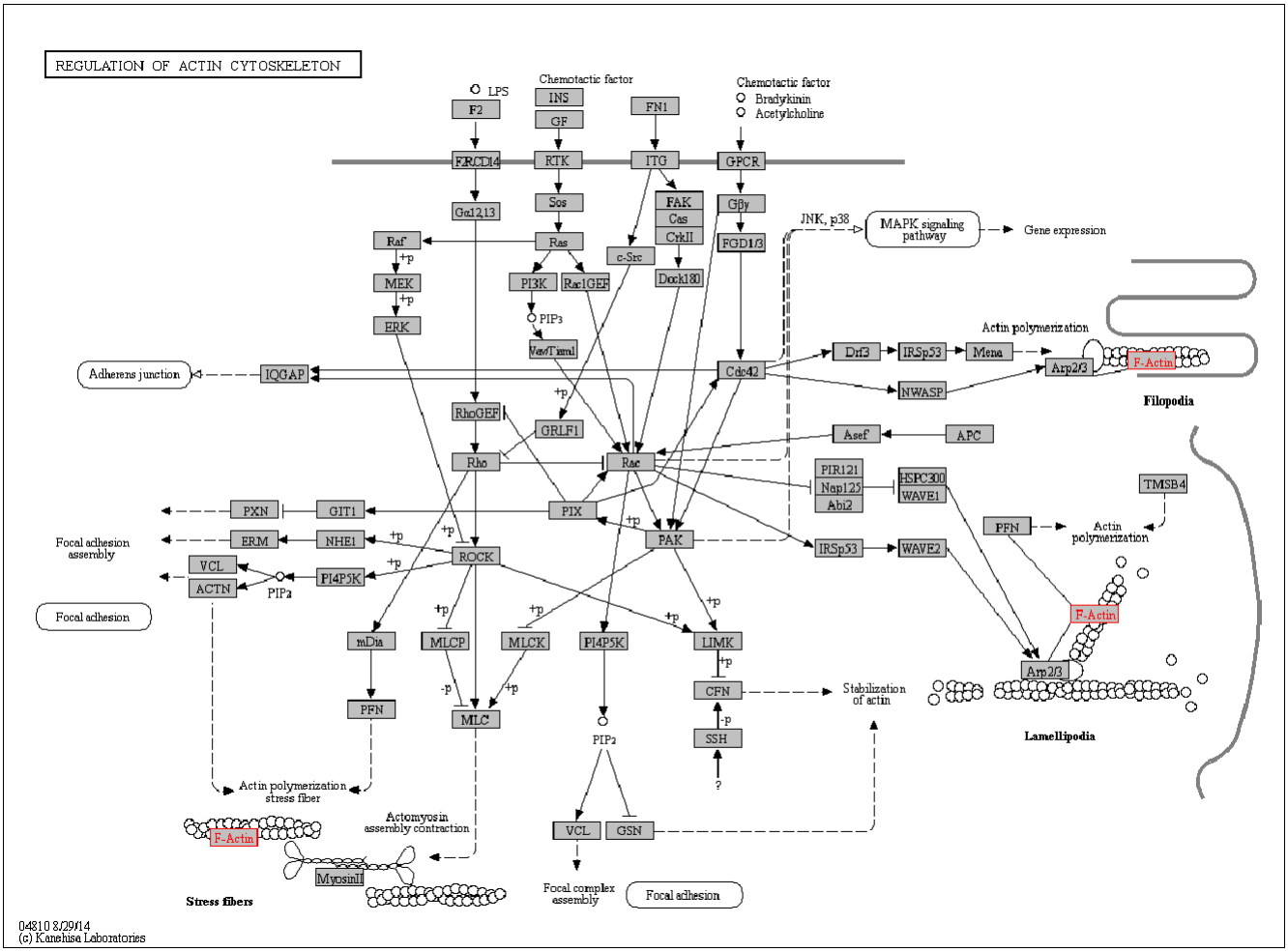

FIGURE A.3: Actin

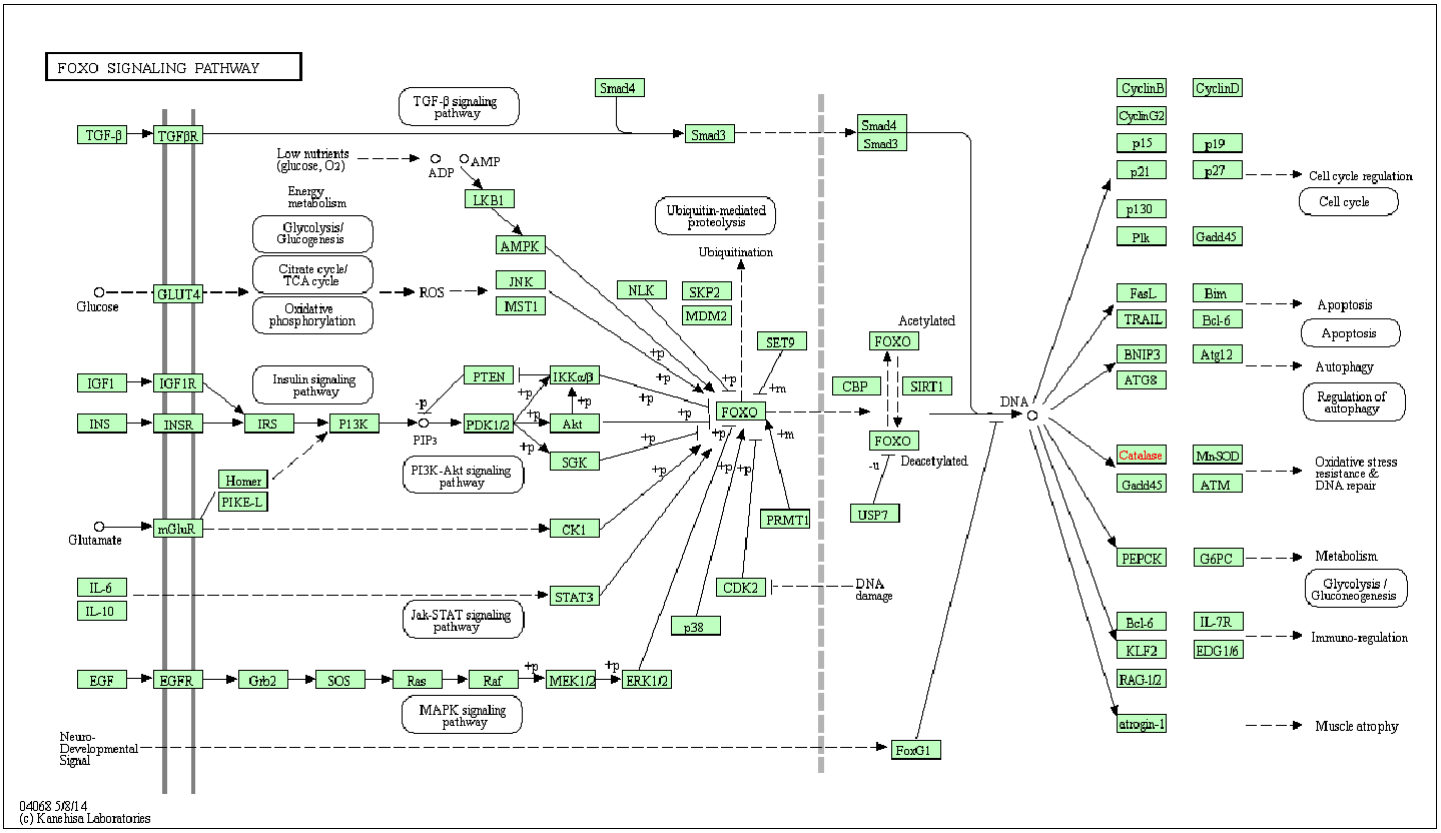

FIGURE A.4: Actin

Appendix B

Biomodels

The following table enlists the Biomodels used in the thesis, the detailed description can be obtained from EBI Biomodels database using the identifier.

For example to access Cdc2 and Cyclin interactions Model (BIOMD0000000005) model, the Url is : https://www.ebi.ac.uk/biomodels-main/BIOMD0000000005.

In general, the Url format for a model is https://www.ebi.ac.uk/biomodels-main/BIOMDXXXXXXXXXX

As discussed in earlier section, these models represent various signalling pathways involved directly or indirectly in cancer.

Sl No.	**Model**
1	BIOMD0000000005
2	BIOMD0000000010
3	BIOMD0000000033
4	BIOMD0000000149
5	BIOMD0000000151
6	BIOMD0000000154
7	BIOMD0000000155
8	BIOMD0000000156
9	BIOMD0000000157
10	BIOMD0000000158
11	BIOMD0000000159
12	BIOMD0000000173
13	BIOMD0000000175
14	BIOMD0000000188
15	BIOMD0000000189
16	BIOMD0000000201
17	BIOMD0000000203
18	BIOMD0000000204
19	BIOMD0000000209
20	BIOMD0000000210

TABLE B.1: Biomodels

Sl No.	**Model**
21	BIOMD0000000223
22	BIOMD0000000246
23	BIOMD0000000286
24	BIOMD0000000287
25	BIOMD0000000344
26	BIOMD0000000407
27	BIOMD0000000424
28	BIOMD0000000427
29	BIOMD0000000430
30	BIOMD0000000431
31	BIOMD0000000432
32	BIOMD0000000433
33	BIOMD0000000440
34	BIOMD0000000441
35	BIOMD0000000442
36	BIOMD0000000443
37	BIOMD0000000444
38	BIOMD0000000452
39	BIOMD0000000453
40	BIOMD0000000462
41	BIOMD0000000465
42	BIOMD0000000468
43	BIOMD0000000477
44	BIOMD0000000488
45	BIOMD0000000494
46	BIOMD0000000504

TABLE B.2: Biomodels(Contd.)

Sl No.	**Model**
47	BIOMD0000000523
48	BIOMD0000000524
49	BIOMD0000000525
50	BIOMD0000000526
51	BIOMD0000000541
52	BIOMD0000000543
53	BIOMD0000000544
54	BIOMD0000000552
55	BIOMD0000000553

TABLE B.3: Biomodels(Contd.)

Appendix C

Marker Importance

Appendix C lists the importance of markers separating two particular clusters.

Marker	**Percentage**
Hcy CA I - IV	94.00
Comet Assay Percentage I	100.00
Comet Assay Percentage II	89.00
Comet Assay Percentage III	68.00
Comet Assay Percentage IV	99.00
Comet Assay Percentage I-IV	99.00
Actin	98.00
Catalase	92.00

TABLE C.1: Cross validated Marker Frequency in High Vs Low risk cluster

Variable Name	**Importance**	**Frequency**	**Trend(In Malignant)**
Menopausalstatus	29.02	100	↑
CA.percentage.II	10.5	100	↓
SOD.2.Actin	7.34	100	↓
Profilin.1..Actin	6.48	100	↑
SOD.2	5.6	100	↓
Hcy	4.35	100	↑
Profilin.1	3.03	100	↑
Calgranulin.A.Actin	2.96	100	↑
catalase.SOD.2	2.69	100	↑
Trx...Actin	2.61	100	NA
CA.percentage.III	2.26	100	↓
Rho.A (Sum of both bands)	1.99	100	↑
PRX...Actin	1.84	100	NA
Trx	1.84	100	↑
Hcy...CA..IV.I.II.III.	1.67	100	↑
CA..IV.III.	1.63	100	↑
CA.percentage.IV	1.46	100	↓
catalase.Actin	1.33	100	↓
Rho.A...Actin	1.32	100	↑
Calgranulin.A	1.29	100	↑
PRX	1.19	100	↓
Rho.A	1.17	100	NA
PRX.Trx	1.09	100	↓
Hcy...CA..IV.III.	0.89	100	↑
catalase	0.83	100	NA
Rho.A.(Upper band)	0.8	100	↑
Actin...Profilin.1	0.77	100	NA
Actin	0.6	100	↑
Rho.A...Actin	0.52	100	↑
CA.percentage.I	0.5	100	↑
MMP.9	0.28	100	↑
CA..IV.I.II.III.	0.17	100	↓

TABLE C.2: Relative importance of Markers and Trend- Benign vs Malign clusters (With Menopausal status)

Variable Name	**Importance**	**Frequency**	**Trend(In Malignant)**
SOD.2	100	8.0407	↓
CA.percentage.I	100	7.8005	↓
CA.percentage.II	100	7.4476	↑
Actin	100	7.2151	↓
Hcy...CA..IV.I.II.III.	100	7.2042	↓
Hcy	100	7.0788	↑
CA.percentage.IV	100	7.0377	↑
Actin...Profilin.1	100	6.5187	↓
catalase.Actin	100	4.7773	↓
catalase	100	4.6558	↓
PRX	100	3.377	NA
Summe.beider.RhoA.Banden	100	3.369	NA
Calgranulin.A.Actin	100	3.2108	NA
PRX...Actin	100	2.733	NA
CA.percentage.III	100	2.4051	↑
CA..IV.III.	100	2.0344	NA
catalase.SOD.2	100	1.9404	↓
Profilin.1	100	1.7984	NA
Rho.A	100	1.7758	NA
Trx	100	1.7288	↓
SOD.2.Actin	100	1.6077	NA
Rho.A...obere.Bande.	100	1.1674	↓
Hcy...CA..IV.III.	100	1.0612	NA
Calgranulin.A	100	0.9182	NA
Trx...Actin	100	0.9138	NA
PRX.Trx	100	0.8426	NA
Profilin.1..Actin	100	0.6865	NA
obere.Bande.Rho.A...Actin	100	0.354	NA
Rho.A...Actin	100	0.2759	NA
CA..IV.I.II.III.	82	0.0237	NA

TABLE C.3: Relative importance of Markers and Trend- High Risk vs Malign clusters

Variable Name	**Importance**	**Frequency**	**Trend (In Malignant)**
Hcy	17	100	↑
CA.percentage.II	14	100	↓
SOD.2	9.2	100	↓
SOD.2.Actin	8.1	100	↓
Profilin.1..Actin	6	100	↑
catalase.SOD.2	3.8	100	↑
Calgranulin.A.Actin	3.8	100	NA
CA..IV.III.	3.2	100	↓
Trx	3	100	↑
Trx...Actin	2.8	100	↑
PRX	2.8	100	↓
catalase.Actin	2.5	100	↓
PRX...Actin	2.3	100	NA
Hcy...CA..IV.I.II.III.	2	100	↑
Summe.beider.RhoA.Banden	1.8	100	↑
Hcy...CA..IV.III.	1.8	100	↑
Profilin.1	1.7	100	↑
CA.percentage.III	1.5	100	↓
PRX.Trx	1.4	100	↓
Rho.A...obere.Bande.	1.3	100	↑
Rho.A	1.3	100	NA
Actin...Profilin.1	1.3	100	↑
Calgranulin.A	1.2	100	NA
CA.percentage.IV	1.1	100	↓
catalase	0.89	100	↓
obere.Bande.Rho.A...Actin	0.8	100	↑
CA.percentage.I	0.77	100	↑
Actin	0.68	100	↓
MMP.9	0.67	100	↑
Rho.A...Actin	0.66	100	↑
CA..IV.I.II.III.	0.23	99	↓

TABLE C.4: Relative importance of Markers and Trend- Benign vs Malign clusters (Without Menopausal status)

Variable Name	**Importance**	**Frequency**	**Trend(In Malignant)**
CA.percentage.II	100	9.7546	↓
Actin...Profilin.1	100	9.0212	↑
CA.percentage.III	100	8.8571	↓
SOD.2.Actin	100	8.5215	NA
Hcy...CA..IV.I.II.III.	100	8.3151	↑
CA.percentage.I	100	7.3897	↑
catalase.SOD.2	100	7.2051	↑
Trx...Actin	100	5.3045	↑
SOD.2	100	4.7144	NA
PRX.Trx	100	3.5398	NA
Trx	100	3.5383	↑
CA.percentage.IV	100	3.4507	↓
Actin	100	2.5805	↑
Calgranulin.A	100	2.4345	↓
catalase	100	2.1919	↑
Hcy	100	2.1139	NA
Hcy...CA..IV.III.	100	1.5725	NA
CA..IV.III.	100	1.5506	NA
Profilin.1	100	1.3613	↑
catalase.Actin	100	1.1443	↑
Calgranulin.A.Actin	100	0.7649	NA
Rho.A...Actin	100	0.7516	↓
CA..IV.I.II.III.	100	0.6642	NA
Profilin.1..Actin	100	0.5882	NA
Summe.beider.RhoA.Banden	100	0.579	NA
PRX...Actin	100	0.5265	NA
Rho.A	100	0.4456	↓
MMP.9	100	0.4124	NA
Rho.A...obere.Bande.	100	0.246	NA
PRX	100	0.2337	NA
obere.Bande.Rho.A...Actin	100	0.2267	NA

TABLE C.5: Relative importance of Markers and Trend- Low risk vs Malign clusters

Variable Name	**Frequency**	**Importance**	**Trend (In High Risk)**
CA.percentage.I	100	39.8	↑
Catalase	100	19.49	↑
Hcy + CA I-IV	100	13.7	↑
Actin	100	12.7	↑
CA IV	100	6.63	↓
CA III	100	3.96	↓
CA II	100	3.61	↓
CA I-IV	100	0.095	↓

TABLE C.6: Relative importance of Markers and Trend- Low vs High risk clusters