TECHNICAL UNIVERSITY OF CRETE

DEPARTMENT OF ELECTRONIC AND COMPUTER ENGINEERING

DIGITAL SIGNAL & IMAGE PROCESSING LAB

Visualization and Comparison of Topological Networks from Multiple Approaches for Cancer Prognosis

---

Diploma Thesis

Tsakaneli Stauroula

Chania, 2015

Thesis Committee

Thesis Supervisor Professor Michael Zervakis,
Professor Euripides Petrakis
Professor Katerina Mania

Abstract

The ultimate goal of the genomic revolution, is understanding the genetic causes, the blueprint that specifies the exact ways that genetic components, like genes and proteins, interact to make a complex living system, behind phenotypic characteristics of organisms. Nowadays, genome-wide gene expression technologies have been available and are of great importance in many scientific areas such as clinical prognosis, diagnosis and treatment. This availability has made at least a part of this goal closer and led, both biologists and computational scientists, to introduce a variety of methodological approaches, well suited for both qualitative and quantitative level modeling and simulation, for the analysis of genetic interactions in terms of predicting the genetic and proteomic associations as well as modeling the relationships among the studied genetic components. These approaches have the potential to elucidate the effect of the nature and topology of interactions on the systemic properties of organisms.

In this thesis, we model and process, by implementing two different methodological approaches, the relationships between genes and proteins, in order to examine relationships as well as novel genomic signatures, fundamental and of great significance in the creation of breast cancer and cancer metastasis. These approaches are two different algorithms, HotNet2 and Activity Vector, which create gene interaction subnetworks after processing gene expression data, which have been selected from a larger dataset, and protein-protein interaction networks. Finally, we evaluate the results, for their biological significance and their statistical prediction in an independent dataset.

Περίληψη

Απώτερος στόχος της γονιδιωματικής επανάστασης, είναι η κατανόηση των γενετικών αιτίων, το ‘’αποτύπωμα’’ που καθορίζει τους ακριβείς τρόπους που οι γενετικές μόρια, όπως γονίδια και πρωτεΐνες, αλληλεπιδρούν για να κάνουν ένα πολύπλοκο σύστημα διαβίωσης, πίσω από φαινοτυπικά χαρακτηριστικά των οργανισμών. Στισ μέρες μας, είναι διαθέσιμες τεχνολογίες γονιδιακής έκφρασης μεγάλης σημασίας σε πολλούς επιστημονικούς τομείς όπως η κλινική πρόγνωση, διάγνωση και θεραπεία. Η προσφορά αυτή έχει κάνει τουλάχιστον ένα μέρος αυτού του στόχου εφικτό και οδήγησε, βιολόγους σε συνεργασία με ειδικευμένους επιστήμονες/προγραμματιστές, να εισαγάγουν μια ποικιλία μεθοδολογικών προσεγγίσεων, κατάλληλη τόσο σε ποιοτικό όσο και ποσοτικό επίπεδο, για την μοντελοποίηση και την προσομοίωση, της ανάλυση των γενετικών αλληλεπιδράσεων όσον αφορά την πρόβλεψη των γενετικών και πρωτεομικών ενώσεων, καθώς και για την μοντελοποίηση των σχέσεων μεταξύ των προαναφερθέντων στοιχείων. Οι προσεγγίσεις αυτές έχουν τη δυνατότητα να αποσαφηνιστεί η επίδραση της φύσης και της τοπολογίας των αλληλεπιδράσεων αυτών, στις συστημικές ιδιότητες των οργανισμών.
Σε αυτή την διπλωματική, μοντελοποιούμε και επεξεργαζόμαστε, με την εφαρμογή δύο διαφορετικών μεθοδολογικών προσεγγίσεων, τις σχέσεις μεταξύ γονιδίων και πρωτεϊνών, προκειμένου να εξεταστούν οι μεταξύ τους σχέσεις και αλληλεπιδράσεις, με σκοπό και την εξαγωγή νέων γονιδιωματικών υπογραφών, οι οποίες έχουν καθοριστική και μεγάλη σημασία στη δημιουργία καρκίνου του μαστού και στη μετάσταση του καρκίνου. Αυτές οι προσεγγίσεις είναι δύο διαφορετικοί αλγόριθμοι, ο HotNet2 και ο Activity Vector,με την εφαρμογή των οποίων δημιουργούνται γονιδιακά υποδίκτυα αλληλεπίδρασης μετά την επεξεργασία των δεδομένων γονιδιακής έκφρασης, τα οποία έχουν επιλεγεί από ένα μεγαλύτερο σύνολο δεδομένων, και σε συνδιασμό με πρωτεινικά δίκτυα αλληλεπίδρασης. Τέλος, τα αποτελέσματα μας αξιολογούνται βάσει της βιολογικής σημασίας τους καθώς και την στατιστική τους ικανότητα για την πρόβλεψη του καρκίνου σε ένα ανεξάρτητο δείγμα.

Acknowledgements

I would like to thank:

My thesis supervisor, Professor Michalis Zervakis for his guidance, patience, support and for giving me the chance to expand my knowledge in the exciting field of bioinformatics. Dr. Katerina Bei for her support, devoted time and biological insight that she offered me as well as Stelios Sfakianakis for providing the dataset and sharing his knowledge.

Moreover, Professor Euripides Petrakis and Professor Katerina Mania for their contribution as members of the thesis committee.

Last but not least, I would like to give my warmest thanks, to my dear friends, Vasia, Sofia Marietta, and Xara, my family and Michael and also my friend Vaggelis Koukourakis for their love, support and encouragement. I would not be here today without them.

Table of Contents

List of Figures *7*
List of Tables 8
1 Introduction *10*
1.1 Breast Cancer 11
1.2 Breast cancer and Bioinformatics 12
1.3 Genomic Analysis 12
1.4 Network Analysis 13
1.5 Related Work 14
1.6 Thesis Outline and Innovation 16
2 Theoretical Background *17*
A.BIOLOGICAL BACKGROUND *18*
2.1 The Human Genome 18
B.BIOINFORMATICS BACKGROUND 23
2.2 Machine Learning and Pattern Recognition 23
2.2.1 Dataset 23
2.2.2 Patterns –Classes – Features 24
2.2.3 Implementation of pattern recognition 25
2.4 Feature Subset Selection (FSS) 27
2.4.1 Filter methods 29
2.4.2 Wrapper methods 31
2.4.3 Embedded methods 31
2.5 Classification 32
2.5.1 Classification analysis 33
2.5.2 Classifiers 33
2.5.2.1 Linear and Non Linear Classifiers 33
2.6 Classification Methods 34
2.6.1 Support Vector Machines(SVM) 34
2.6.2 Relevance Vector Machines(RVM) 38
2.7 Evaluation Methods 38

2.7.1 Holdout Validation 38
2.7.2 K-Fold Cross Validation (K-Fold CV) 39
2.7.3 Leave One Out Cross Validation (LOOCV) 40
2.7.4 Repeated Random Sub-Sampling Validation 41
2.7.5 Bootstrap Resampling Validation 42
2.8 Networks 42
3 Methodology *46*
3.1 Algorithms implementing Biological Networks 47
3.1.1 Kernel-based algorithms 47
3.1.2 Diffusion kernels 48
3.1.2.1Random walk 49
3.1.2.2 Random walk with restart 50
3.2 Hotnet2 51
3.2.1 Null Hypothesis 54
3.2.2 Parameters β and δ 55
3.3 Activity Vector 56
3.4 Evaluation of the Results 59
3.4.1 Statistical Evaluation-Generalization 59
3.4.2 Biological Evaluation 59
3.5 Significant Analysis of Microarrays (SAM) 61
4 Results *66*
4.1 Dataset 66
4.2 HotNet2 Algorithm results 67
4.3 Activity Vector Algorithm results 69
4.4 Generalization Ability of Genomic Signature 73
4.5 Biological Evaluation 75
5 Conclusion *82*
6 Implementation Aspects *85*
References 86

List of Figures

Figure1.1 Anatomy of the female breast. 11

Figure2.1 Illustration of a gene, part of a cell, with the double-stranded DNA and achromosome.Available online:http://www.mayoclinic.org/testsprocedures/genetictesting/multimedia/genetic-disorders/sls-20076216. 20

Figure2.2 Pattern recognition: (i) class Membership-Description Space Ω, (ii) Realized pattern /space P (iii) Measurement space F/genes 24

Figure2.3 Pattern Classifier 25

Figure2.4 Pattern recognition process 26

Figure2.5 Filter process 30

Figure2.6 Wrapper process 31

Figure2.7 Embedded method process 32

Figure2.8 Feature Subset Selection Methods 32

Figure2.9 Linear (i) and non-linear (ii) problems Available online:http://sebastianraschka.com/Articles/2014_naive_bayes_1.html 34

Figure2.10 The SVM learns a hyperplane which best separates two classes. Red dots have a label $y_i = +1$ while blue dots have a label $y_i = -1$. 37

Figure2.11 Holdout validation method 39

Figure2.12 K-Fold Cross Validation method 40

Figure2.13 Leave One Out validation method 40

Figure2.14 Repeated random sub-sampling validation method 41

Figure2.15 Bootstrap resampling validations 42

Figure2.16 Gene Network Inference Available online: http://www.genome.jp/tools/genies/help.html. 45

Figure3.1 Heat diffusion process 52

Figure3.2 Generalization of Hotnet2 for clinical data 53

Figure3.3 Generalized process of the Activity vector algorithm Available online: Chuang, Han-Yu, et al. "Network-based classification of breast cancer metastasis." Molecular systems biology 3.1 (2007): 140. 58

Figure3.4 Proposed methodology 60
Figure3.5 Assign experiments to two groups (1, 2) 61
Figure3.6 Highlighting and invoking SAM 62
Figure3.7 SAM Dialog Box 63
Figure3.8 (a) original grouping, (b) randomized grouping 63
Figure3.9 The SAM Plot Controller on the front side, The SAM Plot sheet on the second side 64
Figure3.10 Processing data set with SAM 65
Figure4.1: Dataset structure 66
Figure4.2 (A) Subnetwork of k minimum size of components 9 (B) Subnetworks of k minimum size of components 10. Color depicts gene scores. 69
Figure4.3 Two representative Activity Vector subnetworks 70
Figure4.4 Process for resulting genomic signature 72
Figure4.5 Standard deviation of 6 multiple genes in the new dataset 73

List of Tables

Table4.1 p-values for the significance of clusters of a given size, based on expected numbers of clusters from permutation tests of gene scores.....69
Table4.2 Mutual genes between HotNet2 and Activity vector resulting subnetworks 70
Table4.3 Activity vector highest scoring modules involving common genes to HotNet2. 71
Table4.4 Generalization Ability of Genomic Signature Results. 74
Table4.5 List of genes that participate in Activity vector subnetworks. The 89 genes of Activity vector network are mapped to the corresponding Entrez Gene IDs and described according to their encoded gene products. Breast cancer-associated genes are highlighted in red. Brown highlighted Entrez Gene IDs are the overlapping genes within the Activity vector subnetworks. The starting node of each subnetwork is highlighted in a green background, while the 10 common genes of both algorithms are highlighted in a blue background 75

T**able 4.5 (continue)** List of genes that participate in Activity vector subnetworks. The 89 genes of Activity vector network are mapped to the corresponding Entrez Gene IDs and described according to their encoded gene products. Breast cancer-associated genes are highlighted in red. Brown highlighted Entrez Gene IDs are the overlapping genes within the Activity vector subnetworks. The starting node of each subnetwork is highlighted in a green background, while the 10 common genes of both algorithms are highlighted in a blue background..... 76

Table 4.6 List of genes that participate in HotNet2 subnetworks. The 51 genes of HotNet2 network are mapped to the corresponding Entrez Gene IDs and described according to their encoded gene products. Breast cancer-associated genes are highlighted in red. The 10 common genes of both algorithms are highlighted in a blue background..... 77

Table 4.7 Enriched pathways in HotNet2 subnetworks. Breast cancer-associated genes are highlighted in red. Common genes of both algorithms are in italic and bold. The common pathways of both WebGestalt and G2SBC enrichment analyses are highlighted in a green background..... 78

Table 4.8 Enriched pathways in Activity Vector subnetworks. Breast cancer-associated genes are highlighted in red. Starting nodes are underlined, while common genes of both algorithms are in italic and bold. The common pathways of both WebGestalt and G2SBC enrichment analyses are highlighted in a green background..... 79

Table 4.9 Comparison of both Activity Vector and HotNet2 resulting networks and subnetworks according to their enrichment analysis..... 80

1

Introduction

Cancer refers to any one of a large number of diseases, characterized by abnormal changes in cells, which cause mutations in genes, or by uncontrollable division of cells which have the ability to infiltrate and destroy normal body tissue, causing metastasis. These mutations are responsible for deregulating the physiological growth of cells and prevent them from maintaining healthy. The genes are in each cell's nucleus, which acts as the brain that controls each cell. Normally, when old cells wear out, healthy new cells take their place through an orderly process of cell growth. But over time, mutations occur and can change the normal flow that genes work in our body. As a result, they "turn on" certain genes and "turn off" others in a cell. That deregulated cell starts dividing without control or order, producing more cells same as it concluding in the formation of a tumor. A tumor may be benign or malignant. As benign is characterized when it is not dangerous to health and malignant when it is potentially dangerous. Benign tumors are not considered cancerous, their cells are close to normal in appearance, they grow slowly, and they do not invade nearby tissues or spread to other parts of the body. Malignant tumors are cancerous. If left unchecked, malignant cells eventually can spread beyond the original tumor to other parts of the body and cause metastasis. Several studies in the recent years have been proposed in order to identify pathways that give rise to metastasis. The study of network-based pathway identification and classification supports the notion that cancer is a 'disease of pathways'.

As we mentioned above everything starts from the "brain" that controls each cell. Genes have a fundamental role in the heredity process and their expression level stands as the measure to judge their effect on every biological process. In this work, given a dataset of preprocessed gene expression data from control and cancer patients, we visualize and compare the network topology of the pathways which involve the genetic/proteomic relationships that are considered significant in breast cancer development, from two methodological approaches. Our data consist of 4.174 differentially expressed genes. In the next chapter we explain the biological and mathematical knowledge in the area of bioinformatics, needed for our study. In order to capture and model the relationships between those genes and the proteins they encode, in chapter 3 we present the two different methodologies, two different algorithms, HotNet and Activity Vector .Both approaches integrate, differently, the gene expression and network data sets which are given as input and conclude to subnetworks discriminative of metastasis. The point of

HotNet is to find “groups” of genes, network-connected, that are statistically significant even though they’re not necessarily individually significant.

Activity Vector traces markers of metastasis within gene expression profiles. These markers, unlike the approaches proposed in other studies; do not consist of individual genes or proteins, but from the relationships and interactions between these molecules. Under the form of proteins subnetworks, derived from a larger protein-protein interaction network, these markers can be used to identify genetic alterations and to predict the likelihood of metastasis in unknown samples.
After generating the subnetworks from each methodology, we examine how the resulting pathways behave and relate. Last but not least we evaluate them, in a new independent dataset, after applying a classification algorithm, SVM, and also taking into consideration their biological significance. Finally, our results are presented in chapter 4.

1.1 Breast Cancer

Breast cancer is an uncontrolled growth of breast cells. It is a malignant tumor that has developed from cells in the breast. Most of the time breast cancer either begins in the cells of the lobules, which are the milk-producing glands, or the ducts, the passages that drain milk from the lobules to the nipple. Less commonly, breast cancer can begin in the stromal tissues, which include the fatty and fibrous connective tissues of the breast.

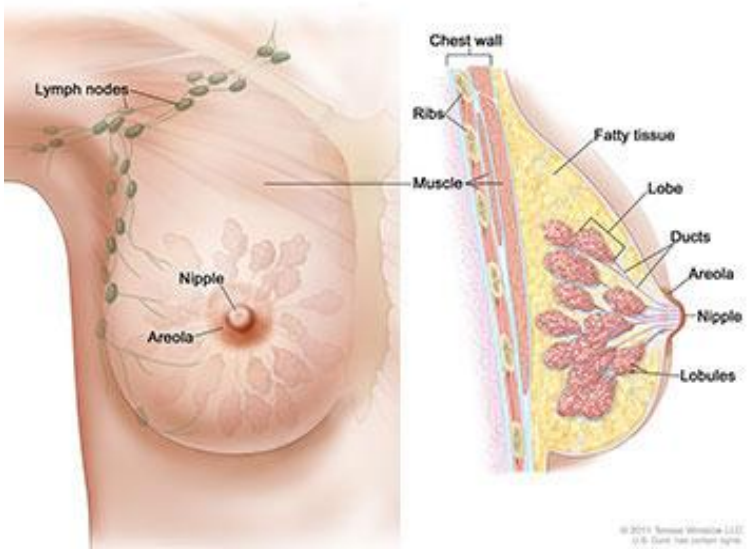


Figure1.1 Anatomy of the female breast. The nipple, areola, lymph nodes, lobes, lobules, ducts, and other parts are shown.[1]

Over time, cancer cells can invade nearby healthy breast tissue and make their way into the underarm lymph nodes, (Figure1.1) small organs that filter out foreign substances in the body. If cancer cells get into the lymph nodes, they then have a pathway into other parts of the body. Breast cancer is always caused by a genetic abnormality, a wrong change in the genetic material. However, only 5-10% of cancers are due to an abnormality inherited from the parents. Instead, 85-90% of breast cancers are due to genetic abnormalities that happen as a result of the aging process. [1, 2]
Cancer is one of the prevalent diseases that bring about death worldwide. Given that scientists have sequenced the human genome, it is now time to use these genomic data, and the high-throughput technology developed to generate them, to confront major health problems such as cancer. [3] The medical and scientific community has therefore come in the conclusion, that it is very important to develop individualized treatment. To achieve the above objective new techniques ought to be developed, in different domains such as bioengineering and bioinformatics.

1.2 Breast cancer and Bioinformatics

Breast cancer occurs in both men and women. Although a cure for each stage of breast cancer has not yet been found, identifying the genetic mutations that cause the disease can play an important role.
Bioinformatics is an integrative area combining biological, statistical and computational sciences. Bioinformatics enables cancer researchers not only to manage, analyze and understand the currently accumulated, valuable, high-throughput data, but also to integrate these in their current research programs. The need for bioinformatics will become even more important as new technologies increase the already exponential rate at which cancer data are generated.
The main aim of bioinformatics is the application of statistics and computer science in the field of molecular biology. We have therefore to do with the development and the advancement of databases, algorithms, computational and statistical techniques and theory to solve formal and practical problems arising from the management and analysis of biological data.

1.3 Genomic Analysis

In genetics the term *Genomics* refers to the field that combines recombinant DNA, DNA sequencing methods, and bioinformatics to sequence, assemble, and analyze the function and structure of genomes. Advances in genomics have triggered a revolution in discovery-based research to understand even the most complex biological systems.

Gene expression profiling is being applied in many areas of research in order to identify new targets for treatment, resistance mechanisms and to improve the current tools of prognosis and treatment. However the vast scale of data generated, in combination with the different protocols, platforms and analysis methods make the studies difficult for the clinicians to understand. In addition, computational scientists and statisticians that participate in the process of data analysis are often not well informed of the sample collection processes or the impact of genetics .Therefore a pressing need has occurred for better understanding of the challenges and limitations of microarray approaches, both in experimental design and data analysis.
The investigation of the roles and functions of single genes is a primary focus of molecular biology or genetics and is a common topic of modern medical and biological research. Research of single genes does not fall into the definition of genomics unless the aim of this genetic, pathway, and functional information analysis is to elucidate its effect on, place in, and response to the entire genome's networks. [4]

1.4 Network Analysis

In the field of Bioinformatics the main goal of several studies has been revealing the pathways that give rise to cancer as well as identifying genetic alterations that determine clinical phenotypes. The relationships between fundamental molecules such as DNA, RNA, proteins and metabolites and the interactions between them, can be described as networks. Networks can be modeled and simulated using various methodological approaches [5, 6]. Once the model has been chosen, the parameters need to be fit to the data. Even the simplest network models are complex systems involving many parameters, and fitting them is a non-trivial process, known as network inference, network identification, or reverse engineering. Genetic networks are often described statistically using graphical models. The interpretation of the network structure constitutes a serious challenge in microarray analysis due to the fact that the sample size is small compared to the number of considered genes. As a result many standard algorithms for graphical models are considered inapplicable. In order to better understand genetic networks we have to look at graph theory and models.
Graph theoretical models (GTMs) are used mainly to describe the topology, or architecture, of a network. These models feature relationships between genes and possibly their nature, but not dynamics: the time component is not modeled at all and simulations cannot be performed. GTMs are particularly useful for knowledge representation, as most of the current knowledge about gene networks is presented and stored in databases in a graph format. In GTMs gene networks are represented by a graph structure, G (V, E), where V = {1, 2,….., n} represent the gene regulatory elements, e.g. genes, proteins, etc., and E = {(I, j) |I, j ∈ V } the interactions between them, e.g. activation, inhibition, causality, binding

specificity, etc. Most often G is a simple graph and the edges represent relationships between pairs of nodes, although hyper edges, connecting three or more nodes at once, are sometimes appropriate. Edges can be directed, indicating that one (or more) nodes are precursors to other nodes. They can also be weighted, the weights indicating the strengths of the relationships. Either the nodes, or the edges, or both are sometimes labeled with the function, or nature of the relationship, i.e. activator, activation, inhibitor, inhibition, etc. The edges imply relationships which can be interpreted as temporal (e.g. causal relationship) or interactional.

Many biologically pertinent questions about gene regulation and networks have direct counterparts in graph theory and can be answered using well established methods and algorithms on graphs. There is also great interest in network medicine for modeling biological systems. Methods using high-throughput data for inference of regulatory networks rely on searching for patterns of partial correlation or conditional probabilities that indicate causal influence. Such patterns of partial correlations found in the high-throughput data, possibly combined with other supplemental data on the genes or proteins in the proposed networks, or combined with other information on the organism, form the basis upon which such algorithms work. Such algorithms can be of use in inferring the topology of any network where the change in state of one node can affect the state of other nodes.

1.5 Related Work

In the field of Bioinformatics, in order to comprehend cancer mechanisms and improve the methods of prognosis, many studies focus on the analysis of gene expression profiles to identify genomic relationships linked to metastasis as well as pathways and associations between mutations and phenotype. The incorporation of Protein-protein interaction (PPI) networks, co-expression networks or pathways from databases such as KEGG, has been proposed to overcome variability of prognostic signatures and to increase prognostic performance.

Studies have been made that focus on the interaction or association between single gene and clinical outcomes. Pauling et al.in [7] has proposed a mutual-information based integrative network analysis for the identification of gene pairs associated with clinical outcome. The produced networks are analyzed over multiple genomic profiles. This study has led to the development of a tool named MINA that integrates the proposed analysis.

Friedman et al., 2000 in [8] proposes an approach based on the well-studied statistical tool of Bayesian networks. This study focuses on revealing interaction between genes by taking into account their expression levels. Furthermore, this approach aims to uncover patterns by examining statistical properties of dependence and conditional independence in

genomic data. Several approaches have been proposed to score known pathways by the coherency of expression changes among their member genes [9-15].

A number of approaches [17-18] have been demonstrated for extracting relevant subnetworks based on coherent expression patterns of their genes or on conservation of subnetworks across multiple species ,Sharan et al, 2005 in[19],using protein–protein interaction networks which derive from literature, the yeast two-hybrid system, or mass spectrometry [16]. More recently, methods to discover mutated subnetworks have been introduced.[20-21].

This work is an implementation of two studies Chuang et al.,2007 in [22] and Vandin et al. in [23-24]. Chuang et al.,2007 study proposes a protein network-based approach that identifies markers not as individual genes but as subnetworks extracted from protein interaction databases. Whereas, Vandin et al., proposed mutated sub-networks which are associated with clinical outcome by developing and applying HotNet [23] and HotNet2[24] algorithm. Pearson's correlation based approaches [25, 26], clustering and classification algorithms [27, 28, 29,30] have been successfully used to elucidate the functional relationship between genes and pathways, but they are unlikely to directly output the specific gene networks in response to abnormal physiological conditions such as diseases, due to experimental errors and the genetic complexity [31, 32]. Their main drawback is their limited performance when the experimental data is insufficient, especially when the number of the features under examination exceeds the number of samples. This makes the estimation of a network structure a challenging problem due to the uncertainty of calculation of the correlation matrix. The information contained in the expression data is limited by the tissue quality, the experimental design, noise, and measurement errors. These factors negatively affect the estimation of causal relationships in network structure and the derivations of dependencies enclosed between neighbored genes and/or proteins [33].

In this context, our goal in this thesis is to implement the two different methodologies into our genomic dataset and locate the structural differences within the network between the two populations (cancer-control).

The gene expression profile of each gene differentiates along the samples and according to the group that each sample belongs; the value of each gene alters significantly. Therefore we aim in finding the genes that most differ between the two groups and are more likely to dominate in our networks.

The resulting subnetworks will give us the information we need in order to determine how the genes behave and probably going to behave ,as well as how they influence each other so as to have a better knowledge in predicting "cancer triggering" relations/pathways.

1.6 Thesis Outline and Innovation

The necessary theoretical background for the development of this thesis is covered in chapter 2. This chapter is divided in two parts. The first, concerning the human genome and biological concepts regarding DNA microarrays form the Biological background. Gene networks and methodologies concerning the analysis of DNA microarray data as well as the construction of gene networks under multiple methodological approaches compose the second part, the mathematical background involving the knowledge in the field of bioinformatics and its applications. In chapter 3, we introduce the proposed methodology concerning this study. Moreover, we analyze in detail the steps chosen for the elaboration of the two selected methodologies Activity Vector and Hotnet2 algorithm, for the gene network construction as well as an introduction to the evaluation method implemented for the generalization ability of the observed results .The integration of the breast cancer gene expression dataset and the two methodologies is presented in section 4, as well as the generation of networks from our data along with their organization in subnetworks. Our results were evaluated after applying the SVM (Support Vector Machine), classifier for statistical prediction and after the examination of their biological significance. Finally, the bioinformating tools for the biological assessment of our results as well as the computer/software requirements needed are mentioned.

The innovative concept of this thesis involves the process of gene expression data from two different methodological approaches, especially from HotNet2 algorithm were the implementation with input scores from gene expression data has not been explored yet.
Moreover, unlike similar studies, we produce a result which derives and is being evaluated as a group of significant relationships between genes, combining the results from the two methodologies, and not examining each genomic signature separately. Apart from the statistical part, the extraction of subnetworks with two different methodological approaches, we extend our study a step further combining the resulted subnetworks and choosing stronger connected subnetworks as well as their biological significance Finally, we must mention that many approaches have been proposed to compare results of gene selection methods or even information of different type of experiment data. In this thesis we attempt a comparison and combination of significant results of subnetwork operations.

2

Theoretical Background

This chapter is divided in two theoretical parts: first the Biological background is introduced and second is the mathematical background (bioinformatics), needed for the composition of this thesis. In the first section an introduction to the human genome is presented. The domain of the human genome and the significance of DNA microarrays as well as their analysis are covered in section 2.1. Following, in section 2.2, which constitutes the beginning of the second theoretical part, we introduce the scientific field of machine learning and pattern recognition followed by a general interpretation of the data in section 2.3. Moreover, in section 2.4 the process of feature subset selection (FSS), applied in DNA microarray data, which is distinguished in three fundamental algorithms, also presented, wrappers, filters and embedded methods, is interpreted. Continuing on section 2.5 and 2.6, the general process of classification and an introduction of classifiers , including linear and non linear classifiers, along with the classification method (SVM), implemented in this thesis, are covered respectively. Furthermore in section 2.7, different evaluation methods are described such as holdout validation, k-fold cross validation, leave one out cross validation, repeated random sub-sampling validation and bootstrap resampling.
Finally, the relationship of network biology and bioinformatics is introduced in section 2.8 where a part of different biological networks that exist are presented.

A.BIOLOGICAL BACKGROUND

2.1 The Human Genome

Human Genome

The human genome refers to the complete set of human genetic information, the study, analysis and mapping of which, has been the subject of the "Human Genome Project"[34]. All living things are composed of cells, small units of biological activity, surrounded by a semi permeable membrane and have the remarkable ability to of itself in an environment from which other living systems are absent. In the simplest forms of life, such as single cells, a new body is the result of each cell division. Instead, a new person in multicellular organisms created after many cell divisions. The higher organisms such as humans consist of groups of cells that interact with each other thereby ensuring a harmonious cooperation and functioning.

The molecule of DNA (deoxyribonucleic acid) that occurs in 23 pairs of chromosomes of each cell contains the entire genetic signature of living beings. The set of DNA molecules present in a cell are the genetic material called *genome*. Genes, the basic physical and functional unit of heredity, being only a fraction of the total genome, are DNA fragments containing critical information for the synthesis of proteins in a particular cell type. Today, it is estimated that a total of 24,500 genes encode proteins. That number shrinks to 20,500 genes according to recent studies.[35]

The remaining genome is composed of non-coding regions, responsible for regulating the production of proteins, and whose functions may include chromosomal structural integrity.

The discovery that DNA contains the code for life, urged a global effort to understand how the genome sequences of many organisms associated with their health. The study of the human genome led to the genomic revolution since the notification of the first draft sequence of the genome had a huge impact on human cancer research.

DNA

As we have already mentioned, each gene is made of DNA. Deoxyribonucleic acid (DNA) is a molecule that carries most of the genetic instructions used in the development, functioning and reproduction of all known living organisms as well as many viruses.
DNA and RNA are nucleic acids which alongside proteins and carbohydrates, compose the three major macromolecules essential for all known forms of life. Most DNA molecules consist of two biopolymer strands coiled around each other to form a double helix. The two DNA strands called polynucleotide due to the fact that they are composed of simpler units, called nucleotides. Each nucleotide is composed of nitrogen, containing nucleobase, cytosine I, guanine (G), adenine (A), or thymine (T), as well as a monosaccharide sugar, called deoxyribose, and a phosphate group. The nucleotides are joined to one another in a chain by covalent bonds between the sugar of one nucleotide and the phosphate of the next, resulting in an alternating sugar-phosphate backbone. According to base pairing rules (A with T, and C with G), hydrogen bonds bind the nitrogenous bases of the two separate polynucleotide strands to make double-stranded DNA. The total amount of related DNA base pairs on Earth is estimated at 5.0×10^{37}, and weighs 50 billiontonnes.

RNA

As mentioned above RNA is a nucleic acid, a large bimolecul, but contrary to DNA is found not as a double-strand but as a single-strand folded on to itself. RNA genome is the molecule that carries the genetic instructions of many viruses. There are different types of RNA named according to the biological process in which they participate. Messenger RNA (mRNA) carries information from DNA to the ribosome, a large and complex molecul, where the protein synthesis in the cell takes place. The coding sequence of the mRNA determines the amino acid sequence in the protein that is produced. However, many RNAs do not code for protein.

These non-coding RNAs (“ncRNA”) as transfer RNA (tRNA) and ribosomal RNA (rRNA can be encoded by their own genes (RNA genes), but can also derive from an mRNA nucleotide sequence, called introns. Both are involved in the process of translation. There are also non-coding RNAs involved in gene regulation, processing and other roles.

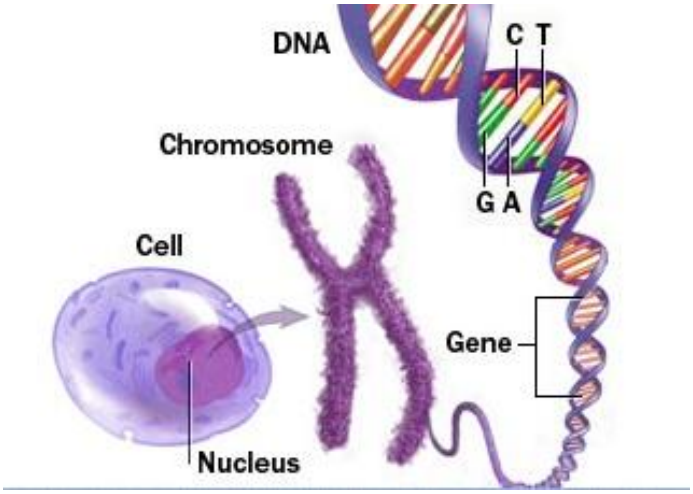


Figure2.1 Illustration of a gene, part of a cell, with the double-stranded DNA and a chromosome.[69]

GENES

A gene is a small piece of the genome. It's the genetic equivalent of the atom: As an atom is the fundamental unit of matter, a gene is the basic physical and functional unit of heredity. Genes, which are made up of DNA, are found on chromosomes and act as instructions to make molecules called proteins. In humans, genes vary in size from a few hundred DNA bases to more than 2 million bases. The Human Genome Project has estimated that humans have between 20,000 and 25,000 genes.

Every person has two copies of each gene, one inherited from each parent. How the two copies interact with each other determines an organism's characteristics. Most genes are the same in all people, but a small number of genes (less than 1 percent of the total) are slightly different between people. This small number of genes is called alleles which are forms of the same gene with small differences in their sequence of DNA bases. These small differences contribute to each person's unique physical features.

Gene Expression

The cell transcribes the genetic sequence into messenger RNA (messenger RNA –mRNA). Then, the mRNA is translated into triplets structuring an amino acid sequence, of the structural components of the protein. Specifically, the codon , as any nucleotide triplet is named, is coding the synthesis of an amino acid according to the standard genetic code which is common to almost all organisms. Essentially, the nucleotide triplet was proposed

by the existence of 20 different amino acids in living organisms. Thus, 3 is the smallest possible number to 43 ≥ 20. In detail, 43 = 64 is the number of all different triads of the four nucleotides obtained by replacement. However, it appears that there is a gap between the number of different triads and the number of different amino acids. In other words, the genetic code is redundant. The first two nucleotides of each codon are fixed and they characterize the encoding of a single amino acid. The third nucleotide of a codon can be altered and yet still encode the synthesis of the same amino acid. The process that determines the level of protein production from a gene is called, gene expression. The gene expression levels (of a gene) indicate the approximate number of RNA copies (of this gene) and correlated to the amount of production of the corresponding proteins (encoding this gene). Thus, the expression of a gene provides a measure of the activity of this gene under specific biochemical conditions. This way we can monitor the effect of a gene in a particular biochemical process by examining the distribution of expression levels.

DNA Microarray and analysis

The DNA microarray technology is a method for gaining information for gene functions as well as probing in parallel, expressions of thousands of genes. Due to the fact that a mutation, or alteration, in a particular gene's DNA may contribute to a certain disease, there was an eager need for the development of a test that can trace these mutations. This was achieved through DNA microarray technology, based on the availability of gene sequences, arrayed on a solid surface [36]. Thousands of DNA probes are arranged in a 2D array, typically on glass slides .Microarrays can be used to study the extent to which certain genes are turned on or off in cells and tissues. In this case, instead of isolating DNA from the samples, RNA is isolated and measured. The total pool of mRNA from experimentally manipulated cells or tissues are used to generate cDNAs, mRNA transcript's sequences, which are labeled using fluorescent nucleotides. A single experiment using this microarray technology can now provide systematic quantitative information on the expression of over 45,000 human transcripts within cells in any given state, enabling the investigator to inquire the whole genome at once. [37] Two types of microarray are in current use; they can be categorized by how the DNA probes are immobilized on the slide: the in situ synthesized Affymetrix GeneChips which utilizes photo-lithography for embedding cDNA probes on silicon chips, and the spotted cDNA (or oligonucleotide) microarrays developed at Stanford University which utilizes robotic spotting of aliquots of purified cDNA clones. Scientists conduct large-scale population studies, to determine how often individuals with a particular mutation actually develop a type of cancer, or to identify the changes in gene sequences that are most often associated with particular diseases. This has become possible because, just as is the case for computer chips, very large numbers of 'features' can be put on microarray chips, representing a very large portion of the human genome.

Microarray can be a valuable tool in order to define transcriptional signatures bound to a pathological condition, to determine whether the DNA from a particular individual contains a mutation in their genes as well as to exclude molecular mechanisms tightly bound to transcription. Microarray analysis frequently does not imply a final answer to a biological problem but allows the discovery of new research paths which let to explore it by a different perspective.

Today, DNA microarrays are used in clinical diagnostic tests for some diseases. Sometimes they are also used to determine which drugs might be best prescribed for particular individuals, because genes determine how our bodies handle the chemistry related to those drugs. With the advent of new DNA sequencing technologies, some of the tests for which microarrays were used in the past now use DNA sequencing instead. But microarray tests still tend to be less expensive than sequencing, so they may be used for very large studies, as well as for some clinical tests.

The principal steps of a microarray analysis are:

- Gene intensity measurements and data normalization.
- Statistical validation of differential expression.
- Functional data mining

B.BIOINFORMATICS BACKGROUND

2.2 Machine Learning and Pattern Recognition

2.2.1 Dataset

Our data is presented as a set of N samples. Each sample contains the expression value of K genes also called predictors. In the dataset, each sample N can be expressed as a vector $\boldsymbol{x_i} \in \boldsymbol{R^K}$ where i = 1,...., N. To each of the samples, a class label y is assigned. The data can also be expressed in array form as $X \in R^{N,K}$ where each row represents a sample containing the expression values of K genes, while the class labels of all samples are expressed as a vector $y \in R^N$.

Pattern recognition [38-40] is classified in the field of machine learning, a scientific area that focuses on the recognition of patterns and regularities in vast amount of data. Machine learning depends to the kind of data that we have at our disposal, thus pattern recognition can be achieved accordingly. The learning method that we can have is supervised learning, unsupervised learning and reinforcement learning.

- **Supervised learning**

Supervised learning entails learning a mapping between a known dataset called the training dataset, a set of input variables X and an output variable Y , and applying this mapping to predict the outputs for unseen data. If the desired output consists of continuous variables, then the task is called regression whereas cases, in which the output falls within discrete values the task is called classification. Supervised learning is the most important methodology in machine learning and it also has a central importance in the processing of class prediction in DNA microarray data analysis.

- **Unsupervised learning**

Unsupervised learning is the type of machine learning that is trying to find hidden structure in data with unlabeled responses. Due to the fact that the data given are unlabeled, this concludes that there is no error or reward signal to evaluate a potential solution. Various unsupervised classification techniques can be employed with DNA microarray data in microarray data analysis that affect statistical analysis, in the part of class discovery.

- **Reinforcement learning**

Reinforcement learning is the type of machine learning where an agent interacts with its environment. The agent senses the environment, and based on this sensory input choosing an action to perform in it. This action changes the environment in some manner and this change is communicated to the agent through a scalar reinforcement signal. Reinforcement learning utilizes a positive or negative reward signal sent to the agent after an action is complete.

2.2.2 Patterns –Classes – Features

DNA microarray analysis falls within supervised learning. In machine learning and pattern recognition the features can be symbolic (e.g. color) or numerically (e.g. height).The combination of some features is the *feature vector*. A *pattern* is a composition of characteristics which are divided into specific decision areas called *classes*. The classes are separated by decision boundaries. The n-dimensional space defined by the feature vector space is called feature space. Feature spaces may overlap each other, allowing patterns of different classes to share same characteristics. Moreover, each pattern can be illustrated in the set of features F. Thus, each feature can be a member not only of different patterns but also different classes. The classification model is a pair of variables {x, ω} where x is a collection of features, feature vector, and ω is the concept of observation, the label. [41-42]

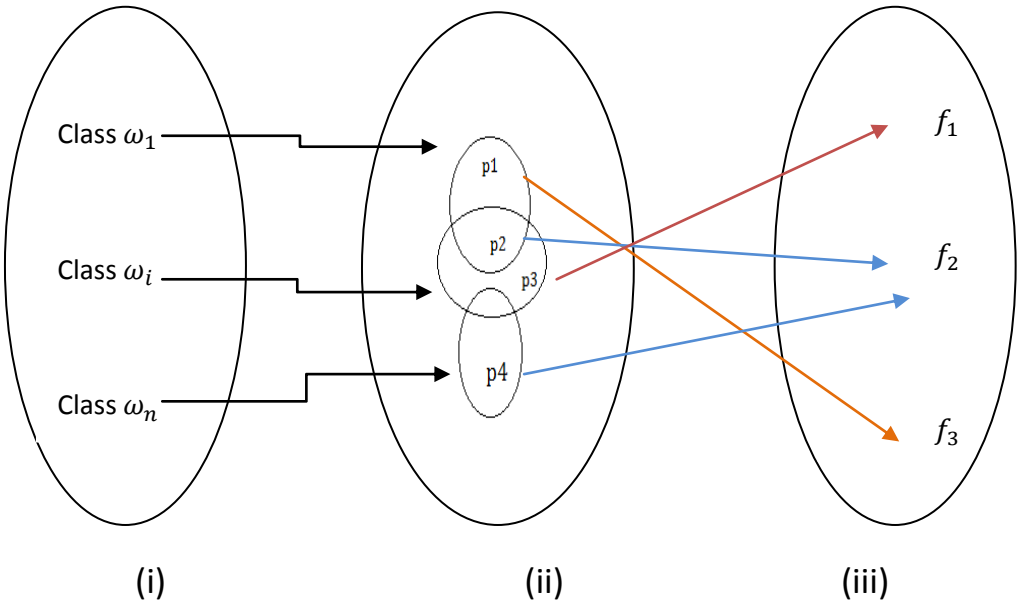


Figure2.2 Pattern recognition: (i) class Membership-Description Space Ω, (ii) Realized pattern space P (iii) Measurement space F/genes

2.2.3 Implementation of pattern recognition

As we mentioned above in machine learning, pattern recognition focuses on making reasonable decisions about the categories of the patterns. Pattern recognition [43] is a two phase process:

a. Training/Learning and
b. Detecting/Classifying

This process begins with the acquisition of data through measurements. These data are subjected to pre-processing in order to isolate patterns of interest. The next step of this process is the selection of features in order to find the best representative subset. In the second phase, the combination of features through supervised, unsupervised or reinforcement learning models, as to assign a pattern to a class, is called Classification. Classification [44] is a representative example of pattern recognition. It implements the process of learning a target function f, that maps each set of features x, in the predefined tagged classes y. Usually the data set is divided into a training set and a control set ,test set. The training set used to build the model and the test set to evaluate it.

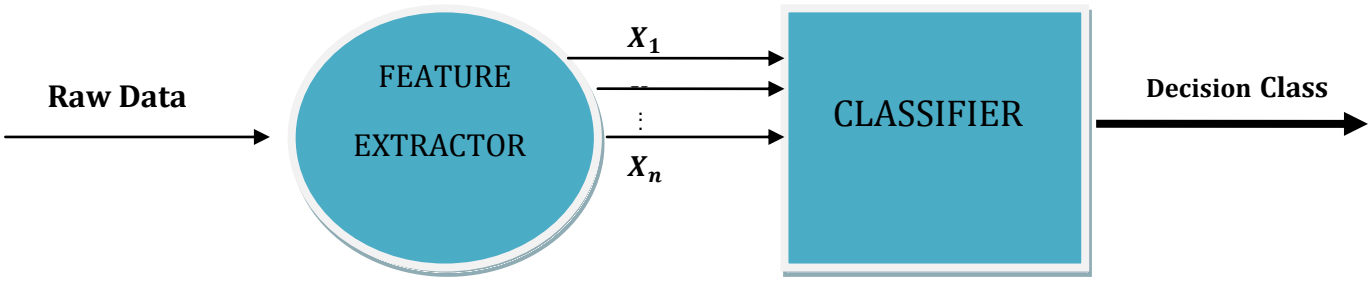


Figure2.3 Pattern Classifier

General process of classification in machine learning is to train classifier to accurately recognize patterns from given training samples and to classify test samples with the trained classifier. Classifier is the algorithm that implements classification and maps input data to class which performs classification. Finally, it is ought to evaluate the decision taken. This involves applying the trained classifier to an independent test set of labeled patterns.
The process is described in detail in sections 2.5-2.7.

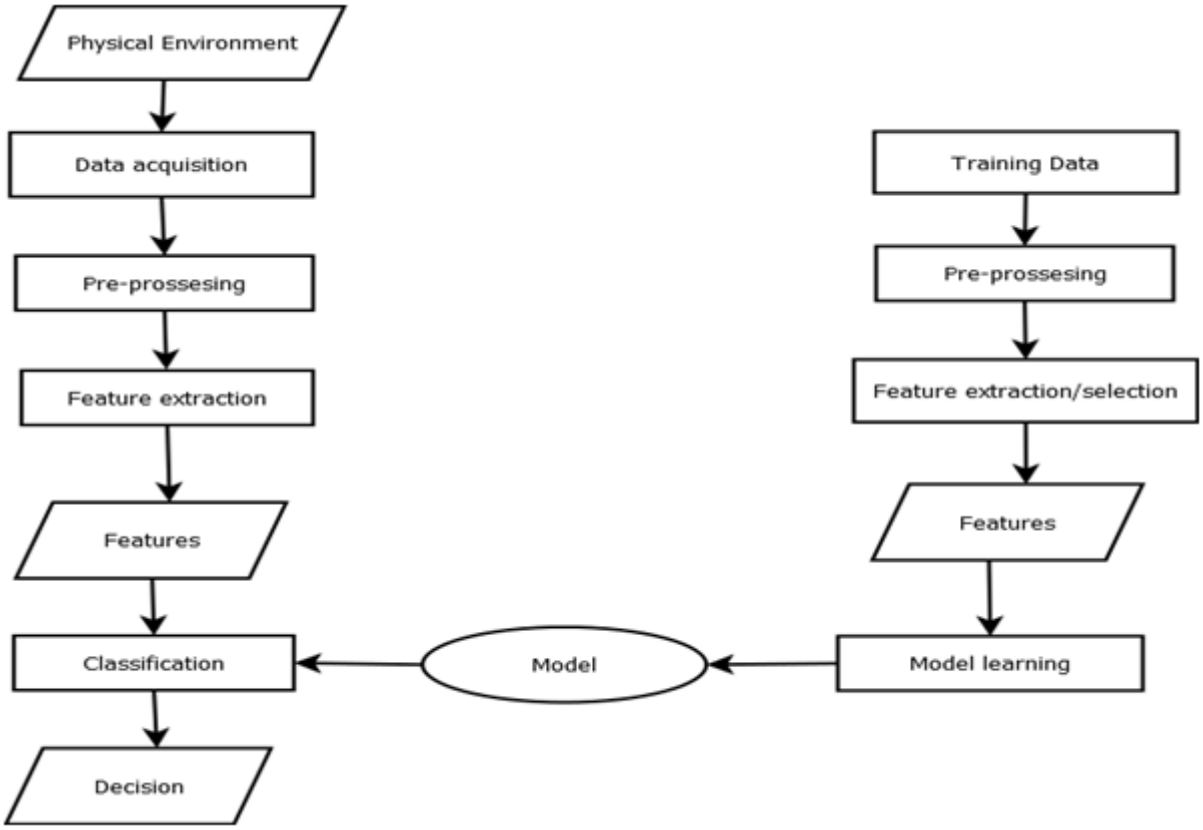


Figure2.4 Pattern recognition process

2.4 Feature Subset Selection (FSS)

After acquiring the gene expression data calculated from the DNA microarray, 2 stages follow: feature selection and pattern classification. The purpose of feature selection can be thought of as the search through the space of feature subsets of genes that might be informative for the prediction by statistical, information theoretical methods, etc. [44, 45, 46, 47].
A typical feature selection process involves two phases:

- Selection of characteristics and
- Fitting the model to evaluate performance.

It consists of three steps:

i. The first step is the creation of a candidate set which contains a subset of the original features through certain research strategies .Three are the main feature selection techniques:

- **Control cases: t-test**
  The basic idea in the t-test is to check if the mean value of the attribute of each class differs significantly from another. T-test is the most popular option when the data follow a normal distribution.

  The aim is to check which of the following two cases applies:
  H1: The feature has a different average value in each class
  H0: The feature has the same average in each class

  If H0 (null hypothesis) is applied then feature is discarded because it is difficult on this basis to distinguish data into categories. On the contrary if H1 (alternative hypothesis) is applicable, the attribute values differ considerably between categories and can be distinguished easily. This feature is selected.

- **The Receiver Operating Characteristic (ROC) curve**
  If when applying the previous method, the respective average values are close, the information may not be sufficient to guarantee good properties classification. The ROC technique gives information on the overlap between categories after quantifying an area defined by two curves.

- **Fisher Discrimination Ratio**
  In order to quantify the resolution of a feature Fisher Discrimination Ratio is used. The ratio is independent of the distribution followed by the class and defined as:

$$\sum\nolimits_w = \sum_{i=1}^{k} \sum_{j=1}^{n_i} \left(x_{ij} - \overline{x_i}\right)\left(x_{ij} - \overline{x_i}\right)^T$$

These criteria do not take into consideration the correlations between features and also do not exploit the cross- correlation coefficient between them. In the scalar selection of characteristics, after choosing a criterion is needed to prioritize features in descending order and calculate the cross-correlation of the first in hierarchy, with all the rest. The cross-correlation process may affect significantly the hierarchy of features.

Additionally, in feature selection a high-dimensional generalization scheme which maximizes the mutual information between the joint distribution and other target variables is found to be useful.

The mutual information (MI) of two discrete random variables X and Y is defined as:

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) log\ \left(\frac{p(x,y)}{p(x)p(y)}\right)$$

, where $p(x,y)$ is the joint probability distribution function of X and Y, and p(x) and p(y) are the marginal probability distribution functions of X and Y respectively. In the case of continuous random variables, the summation is replaced by a definite double integral

$$I(X;Y) = \int_Y \int_X p(x,y) \log\left(\frac{p(x,y)}{p(x)p(y)}\right) dx\, dy$$

, where $p(x,y)$ is now the joint probability density function of X and Y, and and are the marginal probability density functions of X and Y respectively.
Mutual information measures the information that X and Y share. Thus this can be translated as a measurement of the "knowledge" one of these variables gives us, in order to reduce uncertainty about the other. In the case that X and Y are independent, then knowing X does not give any information about Y and vice versa, so their mutual information is zero. On the other hand, if X is a deterministic function of Y and Y is a deterministic function of X then all information conveyed by X is shared with Y:
Knowing X determines the value of Y and vice versa.
As a result, in this case the mutual information is the same as the uncertainty contained in Y (or X) alone, namely the entropy of Y (or X). Moreover, this mutual information is the same as the entropy of X and as the entropy of Y, with a very special case of this is when X and Y are the same random variable.
Mutual information is a measure of the inherent dependence expressed in the joint distribution of X and Y relative to the joint distribution of X and Y under the assumption of independence.

Mutual information therefore measures dependence in the following sense:
$(X; Y) = 0$ if and only if X and Y are independent random variables. this is easy to see in one direction: if X and Y are independent, then $p(x, y) = p(x)\, p(y)$, and therefore:

$$\log\left(\frac{p(x,y)}{p(x)p(y)}\right) = \log 1 = 0$$

Moreover, mutual information is nonnegative I(X; Y) ≥ 0 and symmetric $I(X;Y) = I(Y;X)$.

ii. Continuing on with the steps needed to create a candidate set of features the second step is the evaluation of the candidate set and assess the usefulness of characteristics in the set. Based on the assessment, some features in the candidate set may be rejected or added to selected set of features.

iii. Finally, the last step is to determine whether the current set of selected features is quite good after applying certain switching criteria. If the set meets the prerequisites, a selection algorithm characteristics will return all of the selected features, otherwise, it will be repeated until the stop criterion is satisfied.

In supervised learning, feature selection is often viewed as a search problem in a space of feature subsets. To carry out this search we must specify a starting point, a strategy to traverse the space of subsets, an evaluation function and a stopping criterion. Depending on how and when the utility of selected characteristics is evaluated, different methods may be adopted which are divided into three categories: Filter, Wrapper and embedded models.

2.4.1 Filter methods

Filter approaches [48, 49] remove irrelevant features according to general characteristics of the data. Filter algorithms provide fast execution, since they do not include repetitions and they are not based on a specific classifier. They have a simple construction, which typically uses a simple search strategy and characteristics evaluation criterion is planned based on a specific criterion, the feature/feature subset relevance. In this method for every possible characteristics combination we choose a criterion (e.g. Bhattacharya distance, Divergence, Scatter Matrices) and select the best combination of features vector. We must note that filter algorithms are relatively robust against overfitting and may fail to select the most “useful” features. The primary advantage of filter methods is their speed and ability to scale, to large datasets.
Filter methods are divided into *multivariate* and *univariate* methods. Multivariate methods are able to find relationships among the features, while univariate methods consider each

feature separately. Uninvariate filter techniques can be divided into two categories: parametric and model-free methods. In parametric methods the data is drawn from a given probability distribution while in model-free methods, or non parametric, the data may not follow a normal distribution. In microarray studies the most widely used techniques are t-test and ANOVA.

Significance Analysis of Microarrays (SAM)

Significance Analysis of Microarrays (SAM) [50, 51], is a filter, univariate, statistical technique for finding significant genes in a set of microarray data. It was proposed by Tusher, Tibshirani and Chu and the software was written by Michael Seo, Balasubramanian Narasimhan and Robert Tibshirani. SAM identifies genes with statistically significant changes in expression by assimilating a set of gene-specific tests. Each gene is assigned a score on the basis of its change in gene expression relative to the standard deviation of repeated measurements for that gene. Genes with scores greater than a threshold are chosen as potentially significant. The percentage of such genes identified by chance is the false discovery rate (FDR). To estimate the FDR, nonsense genes are identified by analyzing permutations of the measurements. The threshold can be adjusted to identify smaller or larger sets of genes, and FDRs are calculated for each set.
The cutoff for significance is determined by a tuning parameter delta, chosen by the user based on the false positive rate. One can also choose a fold change parameter, to ensure that called genes change at least a pre-specified.

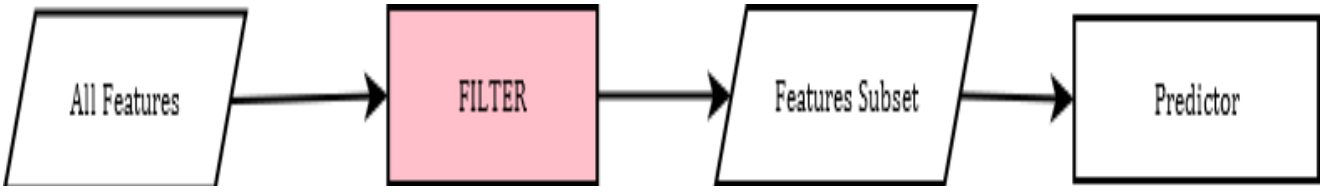


Figure2.5 Filter process

2.4.2 Wrapper methods

Wrapper approaches[48] apply machine learning algorithms to feature subsets and use cross-validation to evaluate the score of feature subsets. Wrapper methodology provides a way to resolve the problem of choice characteristics independent of the learning engine that we have chosen. For each combination of feature vectors to estimate the possibility of false classification is estimated and choose based on the lower smallest error. The execution performance is slow due to repetitions and retraining required and the lack of generality as to the method of identification, however the learning machine can be considered as a black box which makes the method ideal to use anywhere. In this method the criterion that is used is the feature subset "usefulness" measurement. Finally, we must mention that wrapper methods, in principle, result in the most "useful" features, contrary to filter methods which are prone to overfitting. The main disadvantage of wrapper approaches is that during the feature selection process, the classifier must be repeatedly called to evaluate a subset

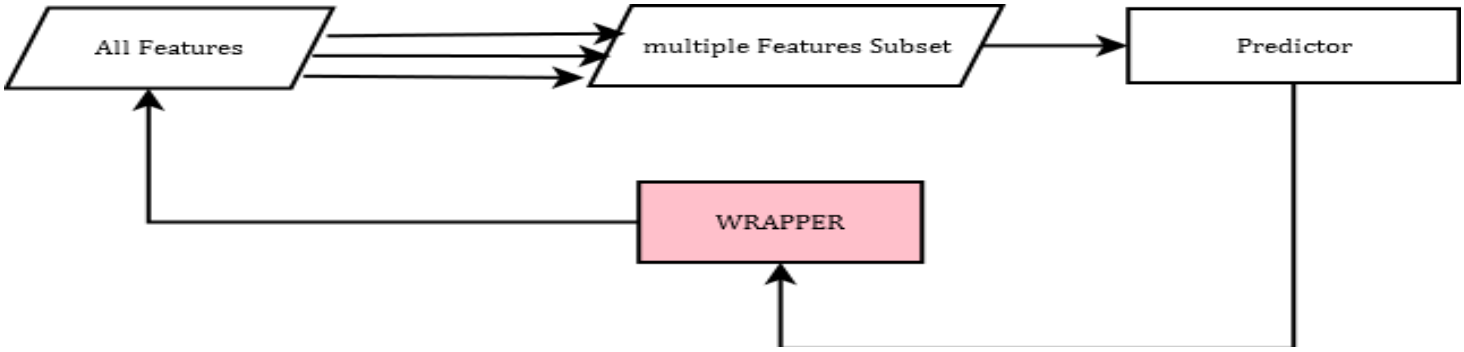


Figure2.6 Wrapper process

2.4.3 Embedded methods

The embedded model algorithms [47, 48] incorporate the feature selection as part of the training/ load process model, and the utility of the characteristics is obtained by optimizing the function of the learning model. This method does not separate the training data in the training dataset and in a set of validation data. Embedded methods are similar to wrappers and they use the same criterion, features subset usefulness. Their advantage is that they are less computationally expensive and less prone to overfitting.

Recursive Feature Elimination (RFE)

Recursive feature elimination is an embedded feature selection approach based on the idea to repeatedly construct a model ,for example an SVM or a regression model, and choose the best or worst performing feature, for example based on coefficients, setting the feature aside and then repeating the process with the rest of the features. This process is applied until all features in the dataset are exhausted. Features are then ranked according to when they were eliminated. As such, it is a greedy optimization for finding the best performing subset of features. The least significant feature is determined through a feature weighting scheme which can be the weight given to each feature by a linear classifier or by non-linear feature weighting methods.

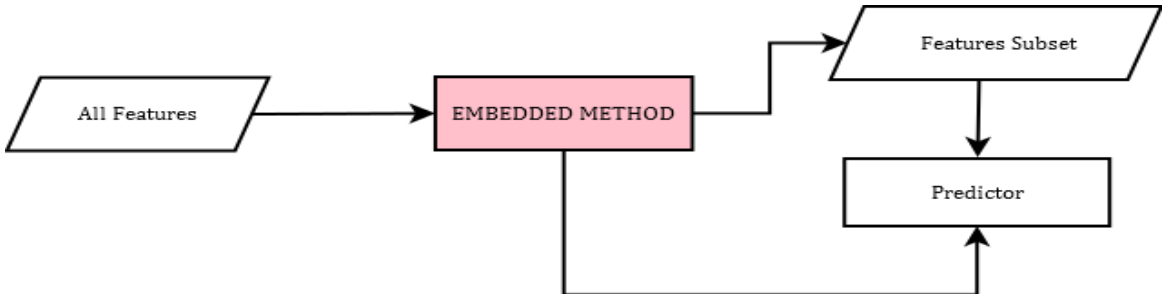


Figure2.7 Embedded method process

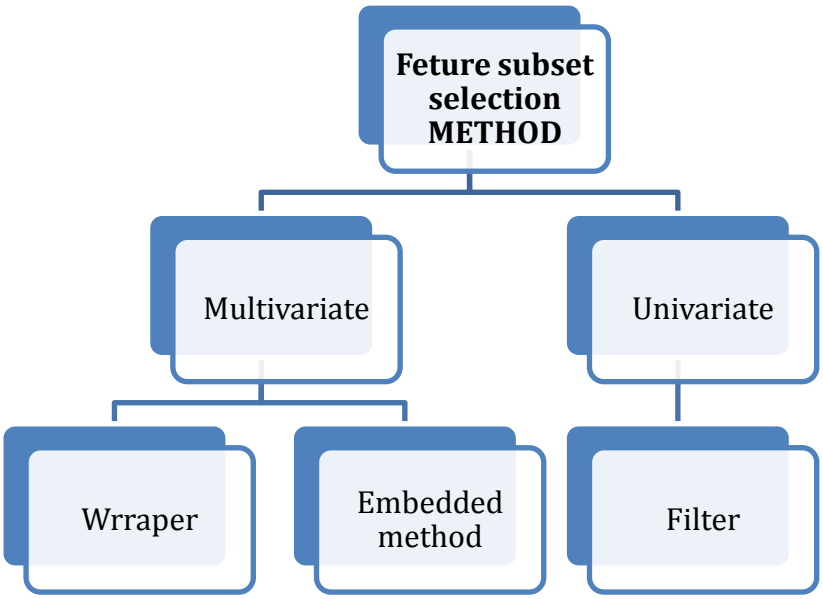


Figure2.8 Feature Subset Selection Methods

2.5 Classification

2.5.1 Classification analysis

As we already mentioned the aim of classification is to find a rule, which, based on external observations, assigns a sample to one of several classes, which implements training a classifier to accurately recognize patterns from given training samples and to classify test samples with the trained classifier. Binary classification is the simplest case where the classifier categorizes the samples of given set into two different classes based on that rule.

2.5.2 Classifiers

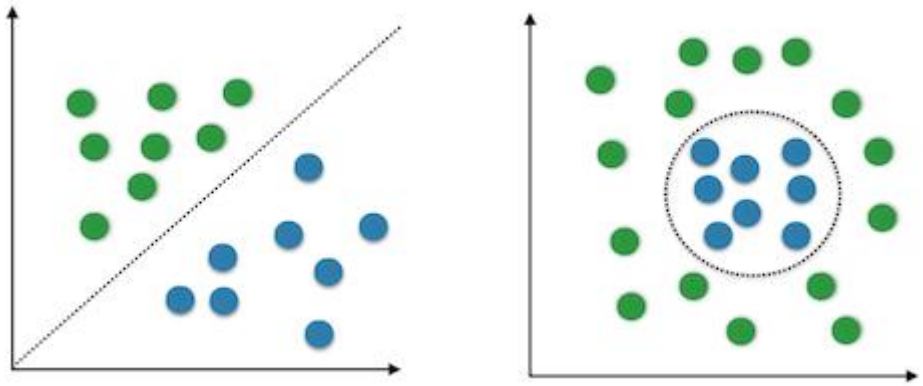
Classifier is the algorithm that implements classification and maps input data to class which performs classification. Classifiers are divided to linear and nonlinear. [49]

2.5.2.1 Linear and non-Linear Classifier

A linear classifier can separate two classes only, when they are linearly separable, i.e. there exists a hyperplane ,in two-dimensional case just a straight line, that separates the data points in both classes. An opposite case is that classes are linearly inseparable. In this case it is still possible that only few data points are in the wrong side of a hyperplane, and thus the error in assuming a linear boundary is small. Depending on the degree of error, linear classifier can still be preferable, because the resulting model is simpler and thus less sensitive for overfitting (poor generalization ability to new data points). However, some classes can be separated only by a non-linear boundary and we need a nonlinear classifier.
More precisely: Let's have numeric attributes $x_{1,\dots\dots\dots}, x_k$ whose values are denoted by $dom(x_i)$. For example if $x_1$can have values between $0 \leq x \leq 1$, then $dom(x_i) = [0,1]$. These compose attribute space: $dom(x_1) \times dom(x_2) \times \dots \times dom(x_k)$.
All data points lie somewhere in this space. If the points fall into two classes, there is some boundary which separates them. If the classes are linearly separable, then in two-dimensional case we can describe the boundary by a line, for 3-dimensional data we need a plane and for higher dimensional data a hyperplane. One way to define this hyperplane is a discriminant function$f(x_1, \dots\dots, x_k)$, which is 0 on the plane, positive, when $(x_1, \dots\dots, x_k)$ belongs to class 1, and negative otherwise. The discriminant function is linear i.e.

$$f = a_1x_1 + a_2x_2 + \dots + a_kx_k + b$$

The simplest example of non-linear boundary is exclusive-or function of two attributes: $XOR(x_1, x_2) = 1$, if x_1 is true or x_2 is true, but not both.
However if we map the datapoints to higher dimensional attribute space, it becomes possible to separate the classes by a hyperplane.
In this study, the linear classifier that is implemented is linear Support Vector Machine (SVM).Other examples of linear classifiers are RLS methods like RR and the LASSO, as well as RVM. An example of a nonlinear classifier is K Nearest Neighbor (K-NN) Classifier which classifies new samples depending on a set of samples closest to them, which are called their "nearest neighbors".



(i) (ii)

Figure2.9 Linear (i) and non-linear (ii) problems. [67]

2.6 Classification Methods

2.6.1 Support Vector Machines (SVM)

Support Vector Machines [52] are supervised learning methods used for classification and regression tasks that originated from statistical theory. SVM is a suitable algorithm to deal with interaction among features and redundant features. The advantage of Support Vector Machines is that they can make use of certain kernels in order to transform the problem, such that we can apply linear classification techniques to non-linear data. Applying the kernel equations arranges the data instances in such a way within the multi-dimensional space, that there is a hyper-plane that separates data instances of one kind from those of another. The kernel equations may be any function that transforms the linearly non-separable data in one domain into another domain where the instances become linearly separable. Kernel equations may be linear, quadratic, Gaussian, or anything else that achieves this particular purpose. Once the data is divided into two distinct categories, our

aim is to get the best hyper-plane to separate the two types of instances. This hyper-plane is important because it decides the target variable value for future predictions. The learnt hyperplane is optimal in the sense that it maximizes the margin while minimizing some measure of loss on the training data. Support vectors are those instances that are either on the separating planes on each side, or a little on the wrong side. SVMs have been shown to work well for high dimensional microarray datasets. One important thing to note is that the data to be separated needs to be binary. Even if the data is not binary, Support Vector Machines handles it as though it is, and completes the analysis through a series of binary assessments on the data.

Linear SVM

In this part of section we further explain the case of the simple linear SVM algorithm [22],[23] in order to be more clearly the concept of support vectors. Linear SVMs are particular linear discriminant classifiers.
Given a training set X of *N* samples of the form:

$$X = \{(x_i, y_i) | \, x_i \in R^m \, , y_i \in \{-1, +1\}\} \, , i = 1, \dots , N$$

where x_i the samples and y_i the class labels, the support vector method approach aims at constructing the maximum - margin hyperplane of dimension $R^{(m-1)}$ that separate the samples having $y_i = +1$ from those having $y_i = -1$. Any hyperplane can be expressed as the set of samples x satisfying:

$$H : w \cdot x - b = 0$$

,where b a real constant and w the normal vector to the hyperplane. The offset of the hyperplane from the origin along the normal vector w can be expressed by the parameter $\frac{b}{\|w\|}$. If the data are linearly separable, there are two hyperlplanes which can be described by the equations :

$$H_1 : w \cdot x - b = 1$$
$$H_2 : w \cdot x - b = -1$$

that fully separate the two classeses without any samples between of them. The region bounded by these hyperplanes is called “the margin” and is equal to $\frac{2}{\|w\|}$. The aim is to maximize the margin, so $\|w\|$ need to be minimized. Given the fact that $\|w\|$ is minimized, samples of either class may fall into the margin, so in order to avoid it, extra constraints need to be applied:

$$w \cdot x_i - b \geq 1 \text{ , for samples of class } y_i = +1$$

$$w \cdot x_i - b \leq -1, \text{for samples of class } y_i = -1$$

The above equations can be expressed in one as:

$$y_i(w \cdot x_i - b) \geq 1, \text{for } i = 1, \dots, N$$

Moreover, the previous constrained equation can be expressed as an optimization problem:

Minimize in w, b

$$\|w\|$$

Subject to

$$y_i(w \cdot x_i - b) \geq 1, \text{for } i = 1, \dots, N$$

This optimization problem is difficult to solve because it is necessary to calculate the norm of w, which involve a square root. Without changing the solution it is possible to substitute $\|w\|$ with $\frac{1}{2}\|w\|^2$. So the optimization problem can be also expressed as:

Minimize in w, b

$$\frac{1}{2}\|w\|^2$$

Subject to

$$y_i(w \cdot x_i - b) \geq 1, \text{for } i = 1, \dots, N$$

By using the Lagrange multipliers $\boldsymbol{\alpha}$, the previous problem can be expressed as a problem of quadratic programming:

$$arg \min_{w,b} \max_{a \geq 0} \left\{ \frac{1}{2} ||w||^2 - \sum_{i=1}^{n} a_i [y_i(w \cdot x_i - b) - 1] \right\}$$

Then, conforming to the stationary Katush – Kuhn – Turkey condition, the solution can be expressed as a linear combination of the training input vectors:

$$w = \sum_{i=1}^{N} a_i \, y_i x_i$$

Only a few of the Lagrange multipliers $\boldsymbol{\alpha}$ will be greater than zero. These corresponding x_i are the support vectors and lie on the margin, satisfying:

$$y_i(w \cdot x_i - b) = 1$$

Solving the above equation for b can derive that the support vectors also satisfy:

$$w \cdot x_i - b = \frac{1}{y_i} \implies b = w \cdot x_i - y_i$$

The b depends on x_i , y_i, so it will vary among the samples. In that manner, a more stable approach for b is to average over all supports vectors:

$$b = \frac{1}{N_{SV}} \sum_{i=1}^{N_{SV}} (w \cdot x_i - y_i)$$

The optimization problem can also be expressed in its dual form, using the fact that $\|w\|^2 = w \cdot w$ and $w = \sum_{i=1}^{N} a_i y_i x_i$. In dual form the classification task takes into account only a function of the supports vectors, which are a small subset of the set of the training samples that lie on the margin. Thus, the problem expressed in dual form is computationally efficient.

Maximize in a_i

$$\check{L}(a) = \sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i,j} a_i a_j y_i y_j \, x_i^T x_j =$$

$$\sum_{i=1}^{N} a_i - \frac{1}{2} \sum_{i,j} a_i a_j y_i y_j \, k(x_i, x_j)$$

, subject to $a_i \geq 0$, $\sum_{i=0}^{N} a_i y_i = 0$ and the kernel function is defined by $K(x_i, x_j) = x_i \cdot x_j$

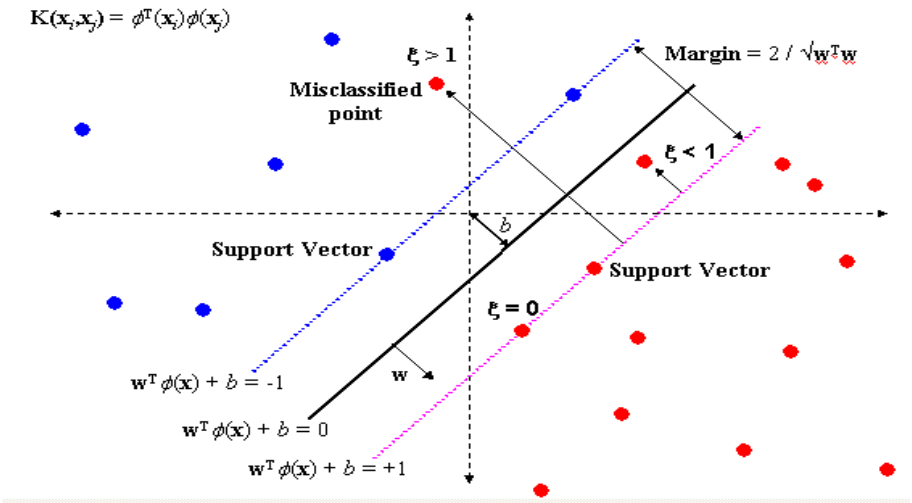


Figure2.10 The SVM learns a hyperplane which best separates two classes. Red dots have a label $y_i = +1$ while blue dots have a label $y_i = -1$

2.6.2 Relevance Vector Machine (RVM)

The relevance vector machine [5],[39] is a sparse kernel technique for both regression and classification. It has an identical functional form to Support Vector Machine (SVM), handles data with less limitations because it consists a special case of Bayesian Logistic Regression that utilizes a specific type of prior probabilities on the feature weights, called Automatic Relevance Determination (ARD) priors that automatically eliminate irrelevant features from the model. RVM is formed as a linear combination of data-centered basis functions, which are called relevance vectors. In comparison to SVMs, the advantages of RVMs are on several aspects including generalization ability and sparseness of the model. In particular, while the SVMs represent decisions, RVMs are based on a Bayesian formulation of a linear model with an applicable prior which is introduced over the weights governed by a set of hyperparameters and bring about a sparse performance. As a consequence, they can generalize well and provide assumptions at low computational cost, since it typically uses dramatically fewer kernel functions.
RVM is a predictive model that directly models the posterior probability of a class C_k , given a sample $p(C_k|\, x)$. The RVM requires class labels of the form$t \in \{0,1\}$, where in the case of binary classification $t_i = 1 \rightarrow x_i \in C_1$, $t_i = 2 \rightarrow x_i \in C_2$. It computes a model which has the form $y(w, x) = \sigma(w^T \cdot \varphi(x))$, where $\varphi(x)$ is a basis function and $\sigma(\cdot)$ the logistic sigmoid function. Thus according to the RVM procedure, each basis function $\varphi(x) = k(x, x_n)$ is given by the kennel and each kernel is associated with one data point. The ARD priors have the form $p(w|a) = \prod_{i=1}^{M} N(w_i|0, a_i^{-1})$. Many of the a_i are led to infinity and the corresponding features are removed from the model, during the ARD process.

2.7 Evaluation Methods

Evaluation methods [36] are techniques for assessing how the results of statistical analysis will generalize to an independent data set. The main idea behind the evaluation methods is to split data, once or several times, for estimating how accurately a predictive model will perform in practice: Part of data, the training set, is used for training each model, and the remaining part, the test set, is used for estimating the accuracy of the model.

2.7.1 Holdout Validation

Holdout Validation is the simplest cross validation method. The dataset is partitioned in two sets, the training set and the testing set. Using the training set only, which consists of the majority of available samples the model, is trained. Then the function is asked to predict the output values for the data in the testing set where the values are unknown. The errors it makes are accumulated to give the mean absolute test set error, which is used to

evaluate the model. The advantage of this method is that it is usually preferable to the residual method and takes no longer to compute. However, the drawback of the method is that its evaluation can have a high variance. The evaluation may depend heavily on which data points end up in the training set and which end up in the test set, and thus the evaluation may be significantly different depending on how the division is made. These limitations of this holdout method can be overcome with other validation methods at the expense of higher computational cost.

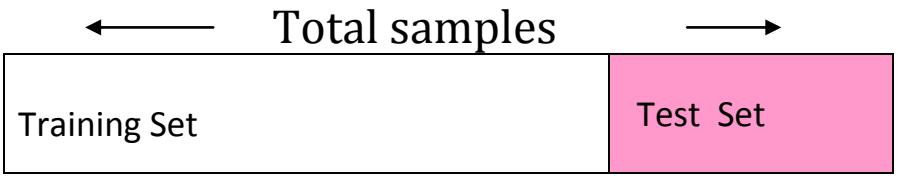


Figure 2.11 Holdout Validation method

2.7.2 K-Fold Cross Validation (K-Fold CV)

As we mention before we can use other cross validation methods to improve over the holdout method. K-fold cross validation is one of them. Here, the data set is divided into *k* subsets, and the holdout method is repeated *k* times. Each time, one of the *k* subsets is used as the test set and the other *k-1* subsets are put together to form a training set. Then the average error across all *k* trials is computed. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set *k-1* times. The variance of the resulting estimate is reduced as *k* is increased. The disadvantage of this method is that the training algorithm has to be rerun from scratch *k* times, which means it takes *k* times as much computation to make an evaluation. A variant of this method is to randomly divide the data into a test and training set *k* different times. The advantage of doing this is that you can independently choose how large each test set is and how many trials you average over.

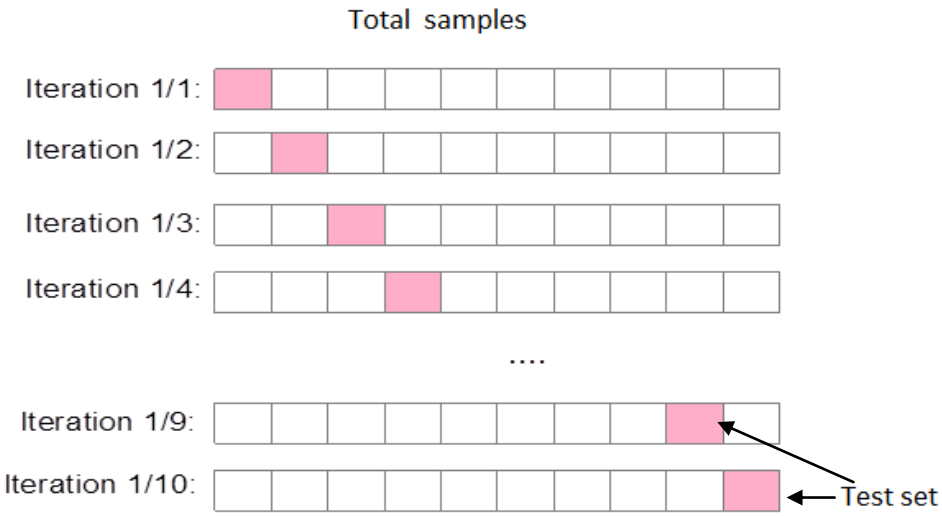


Figure 2.12 K-Fold Cross Validation method

2.7.3 Leave One Out Cross Validation (LOOC)

Leave-one-out cross validation is K-fold cross validation taken to its logical extreme, with K equal to N, the number of data points in the set. This means that for each fold use N-1 samples for training and the remaining sample for testing. As before the average error is computed and used to evaluate the model. The evaluation given by leave-one-out cross validation error is good, but at first it seems very expensive to compute. Fortunately, locally weighted learners can make LOO predictions just as easily as they make regular predictions. That means computing the LOO validation error takes no more time than computing the residual error and it is a much better way to evaluate models.

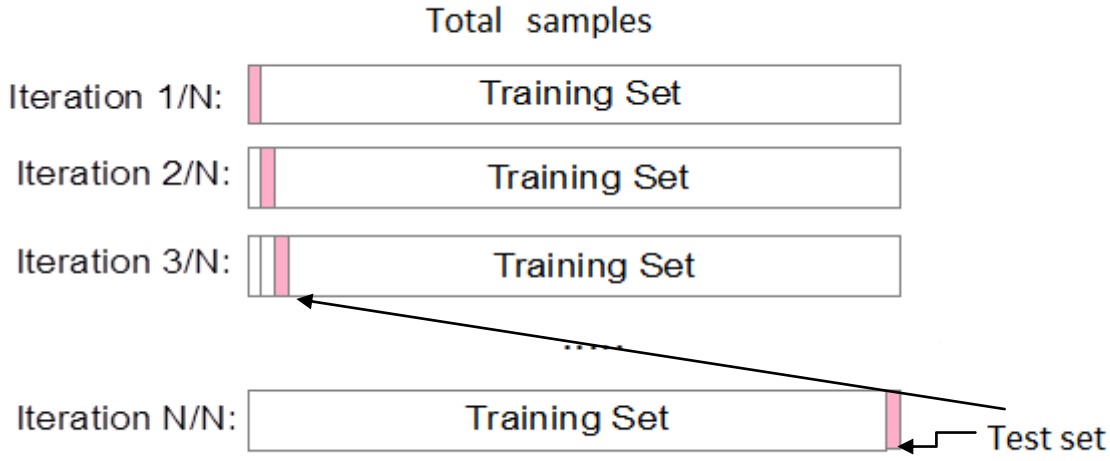


Figure 2.13 Leave One out Validation method

2.7.4 Repeated Random Sub-Sampling Validation

In Repeated random sub-sampling validation [25],[26] the dataset splits K times. Each data split randomly selects a fixed number of samples without replacement. For each such iteration, the model is fit to the training data, and predictive accuracy is assessed using the validation data. The results are then averaged over all iterations. In this method unlike *k*-fold cross validation, the proportion of the training split is not dependent on the number of folds. But the disadvantage using Repeated random sub-sampling is that some observations may never be selected in the validation subsample, whereas others may be selected more than once.

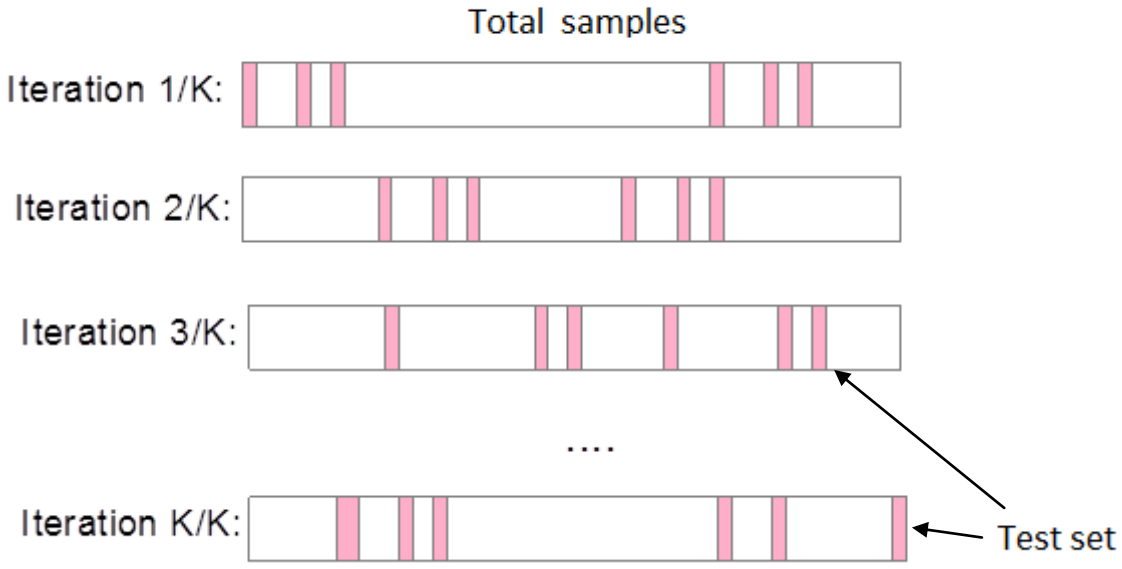


Figure 2.14 Repeated random sub-sampling Validation method

2.7.5 Bootstrap Resampling Validation

The bootstrap resampling validation method [25],[26], also called bootstrapping, is a random sampling technique with replacement. In this method, a number of B bootstrap datasets of fixed size are randomly selected with replacement, from a dataset with N samples, usually the same number of N samples. Then, using the holdout method already mentioned, each bootstrap dataset can be divided into training and test sets. At the end of the procedure in order to get a stable estimation, the statistics are calculated for each bootstrap dataset and are averaged over all bootstrap datasets.

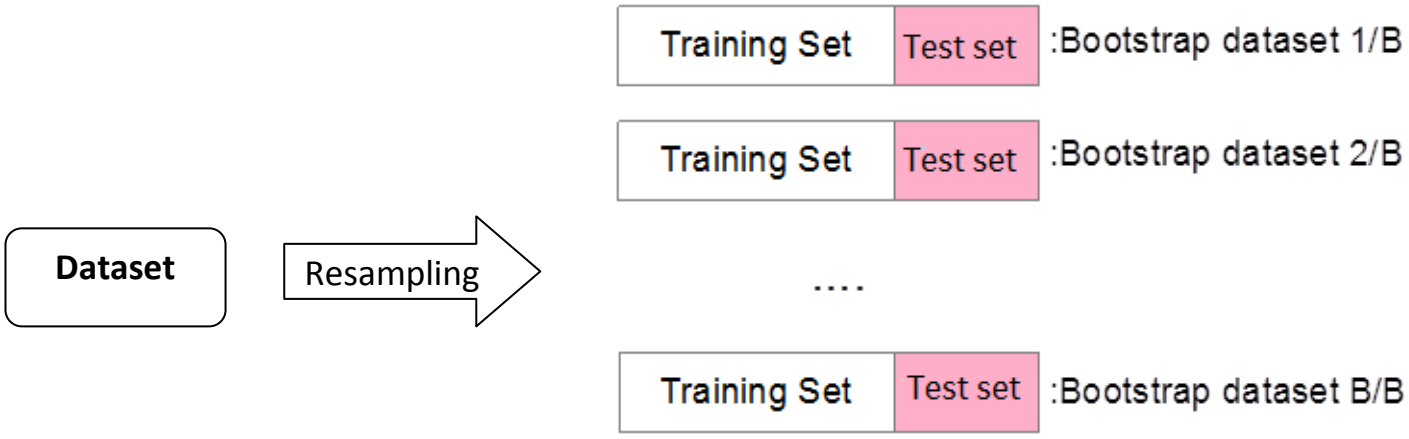


Figure 2.15 Bootstrap Resampling Validation

2.8 Networks

The term Biological Networks is assigned on biological processes which are represented as networks. Biological networks are the interpretation of the interaction between molecules such as DNA, RNA, proteins and metabolites.
There are different types of biological networks. The most fundamental are presented:

- **Metabolic networks**

Metabolic networks refer to the pathways that include the main chemical, mostly enzyme-dependant reactions needed to keep an organism in an internal regulation that maintains a stable, constant condition of a living system, called homeostasis.
Directed edges are drawn between enzymes (proteins that catalyze (accelerate) chemical reactions) and substrates (molecules acted upon by an enzyme) .Thus, enzymes and substrates correspond to nodes, directed edges to metabolic reactions in a metabolic network.

- **Transcriptional regulation networks**

Transcriptional regulation networks are model regulation of gene expression .Gene regulation is a process by which information from genes is turned into gene products (RNA or protein) and enables the cell to control its structure and function. Genes are the nodes and the edges are directed Transcription factor protein X, binds regulatory DNA regions of a gene to regulate (stimulate or repress transcription of a gene) the production rate of protein Y.

- **Protein-protein interaction networks**

Protein-protein interaction networks (PPIs) can be associations of proteins such as functional interactions and their role is highly important for the structure and the function of a cell. These interactions participate in signal induction and play an important role in many diseases (e.g., cancer) .we can encounter stable interactions that form a protein complex (a form of a quaternary protein structure, set of proteins which bind to do a particular function (e.g., ribosome),or transient interactions, which form the dynamic part of PPI networks ,are brief interactions that modify a protein that can further change PPIs –(e.g., protein kineases, add a phosphate group to a target protein). It is estimates state that about 70% of interactions is stable and 30% are dynamic PPI are essential to almost every process in a cell. Thus, understanding PPIs is crucial for understanding life, disease, as well as the development of new drugs .

- **Gene co-expression network (GCN)**

A gene co-expression network (GCN) is an undirected graph, where each node corresponds to a gene, and a pair of nodes is connected with an edge if there is a significant co-expression relationship between them.[53] Having gene expression profiles of a number of genes for several samples or experimental conditions, a gene co-expression network can be constructed by looking for pairs of genes which show a similar expression pattern across samples, since the transcript levels of two co-expressed genes rise and fall together across samples. Gene co-expression networks are of biological interest since co-expressed genes are controlled by the same transcriptional regulatory program, functionally related, or members of the same pathway or protein complex.
The direction and type of co-expression relationships are not determined in gene co-expression networks like in a gene regulatory network (GRN) .Compared to a GRN, a GCN does not attempt to infer the causality relationships between genes and in a

GCN the edges represent only a correlation or dependency relationship among genes. Modules or the highly connected sub graphs in gene co-expression networks correspond to clusters of genes that have a similar function or involve in a common biological process which causes many interactions among themselves.
Gene co-expression networks are usually constructed using datasets generated by high-throughput gene expression profiling technologies such as Microarray or RNA-Sequencing.

- **Bayesian Networks**

Bayesian Networks [54] are a class of graphical probabilistic models that provide a well-ordered representation for the expression of joint probability distributions (JPDs) and inference. Their application is found in many domains such as the of inference of cellular networks, modeling protein signaling pathways, systems biology, data integration, classification and genetic data analysis. They combine two very well developed mathematical areas: probability and graph theory. A Bayesian network consists of an annotated directed acyclic graph $G(X, E)$, where the nodes $x_i \in X$, are random variables representing gene expressions and the edges indicate the dependencies between the nodes. The random variables are drawn from conditional probability distributions$P(x_i|Pa(x_i))$, where $Pa(x_i)$ is the set of parents for each node. A Bayesian network implicitly encodes the Markov Assumption that given its parents; each variable is independent of its non-descendants.
Besides the set of dependencies (children nodes depend on parent nodes) a Bayesian network implies a set of independencies too. This probabilistic framework is very appealing for modeling causal relationships because one can query the joint probability distribution for the probabilities of events (represented by the nodes) given other events.
From the joint distribution one can do inferences, and choose likely causalities.
The complexity of such a distribution is exponential in the general case, but it is polynomial if the number of parents is bounded by a constant for all nodes.

- **Boolean Networks**

Boolean Networks [54] are a class of graphical deterministic models represented as a graph $G(V, E)$, annotated with a set of states $X = \{x_i \mid i = 1, \dots, n\}$, together with a set of Boolean functions $B = \{b_i \mid i = 1, \dots, k\}$,$b_i : \{0,1\}^k \rightarrow \{0,1\}$. Each node v_i has associated to it a function, with inputs the states of the nodes connected to v_i.

The state of node v_i at time t is denoted as $x_i\ (t)$ the state of that node at time t+1 is given by: $x_i\ (t+1) = b_i\ (x_{i1}, x_{i2}, \ldots, x_{ik})$ where x_{ij} are the states of the nodes connected to v_i. This set of functions determines topology connectivity on the set of variables, which then become nodes in a network.

In biological Boolean networks each node represents a gene which takes on two possible values, as mentioned, 0 and 1 and the way these nodes interact with each other is formulated by standard logic functions.

In Boolean networks, genetic interactions and regulations are inextricably linked with the assumption of biological determinism. Though, a gene regulatory network is not a closed system and has interactions with its environment and other genetic networks, and it is also likely that genetic regulations are inherently stochastic; therefore, Boolean networks will have limitations in their modeling power. Probabilistic Boolean networks [55] were introduced to address this issue, such that they are composed of a family of Boolean networks, each of which is considered a context. At any given time, gene regulations are governed by one component Boolean network and network switching is possible such that at a later time instant, genes can interact under a different context. In this sense, probabilistic Boolean networks are more flexible in modeling and interpreting biological data. Interaction networks have proven to be a useful source of information for analyzing genomic data. Using gene expression data we attempt to estimate the network structure using gene and protein information. Boolean Network models belong to the group of qualitative network models, because they do not yield any quantitative predictions of gene expression in the system.

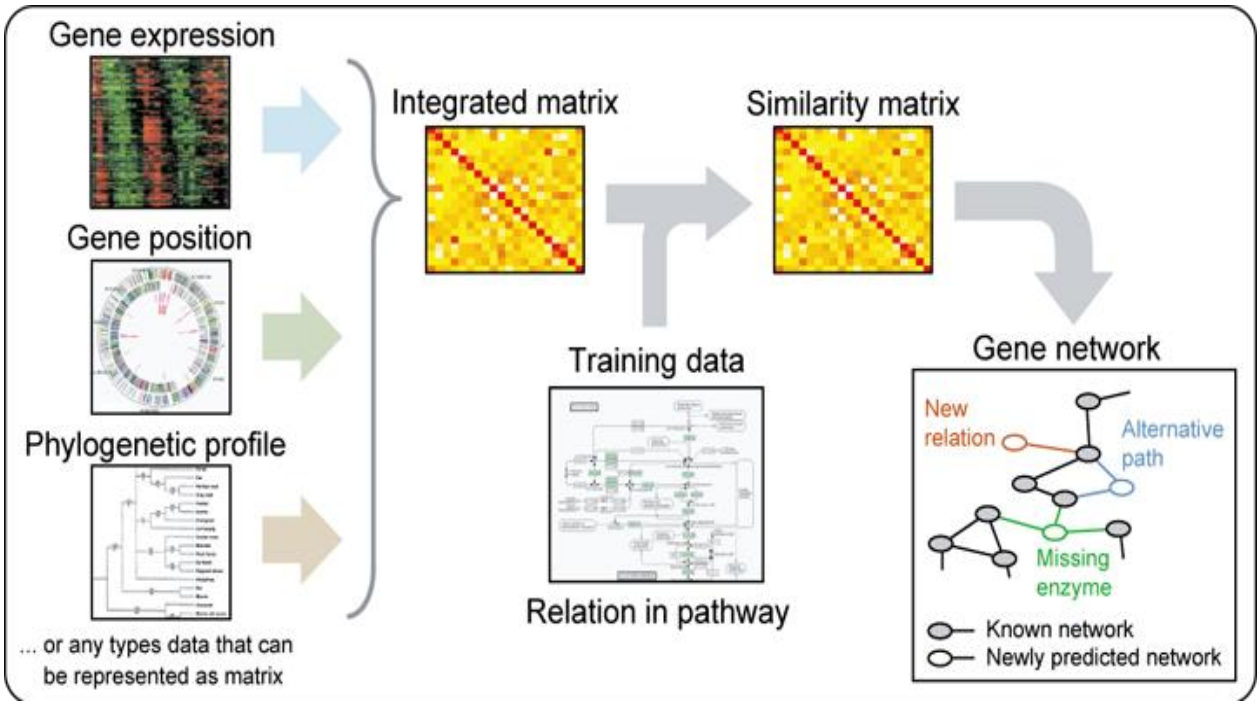


Figure2.16 Gene Network Inference [68]

3

Methodology

Pathway analysis searches for sets of genes differentially expressed in certain phenotypes in order to examine complex diseases, such as cancer. Reasons lying in genetic, physiological, and environmental factors, are responsible for cancer development and thus understanding their interactions enables effective research, diagnosis, and treatment. In this section we extend the network-based algorithm HotNet2 as well as the Activity Vector algorithm, to find gene clusters involved in breast cancer. Breast cancer gene expression data were acquired from two groups of (mRNA) samples (104 control-425 cancer cases). The dataset was pre-processed through a univariate feature subset selection (FSS) method, "Significant analysis of Microarray" (SAM), presented in section 3.4. We have already stated in section 2, part B, the existence of various methods for feature subset selection (FSS) and their significance in DNA Microarray analysis. Section 3.1, refers to the implementation of the two algorithms, respectively. Rather than identifying all genes associated with the disease, both algorithms, seek to identify clusters/markers of genes which are related to each other and to the disease. The results are presented in section 4. Furthermore the most important step in our methodology is the evaluation of our results. In this context, in section 3.2 our goal is to examine how the results of the statistical analysis will generalize to an independent data set. Our results have undergone statistical prediction analysis through classification. A linear classification technique was implemented with the Support Vector Machine (SVM) classifier [57].

3.1 Algorithms implementing Biological Networks

The main focus of networking approaches is to build target-independent networks that describe the pair-wise relations between molecules. The number of personal genomes sequenced has grown rapidly over the last few years and is likely to grow further. Within the last few years, several advanced approaches to address the construction of biological networks from gene-expression data have been proposed, many of them mention in section 1.5. In this study we focus on the work of Chuang et al., 2007 in [22] that proposes a protein network-based approach that identifies markers not as individual genes but as sub networks extracted from protein interaction databases and we present the Activity Vector algorithm while applying it in our gene expression dataset. Moreover we present the Hotnet2 algorithm which extends the work of Vandin et al., 2012 in [23] where HotNet algorithm is proposed for mutated sub-networks associated with clinical outcome.

3.1.1 Kernel-based algorithms

Kernel-based algorithms, such as Gaussian processes [56] or support vector machines [57] are enjoying great popularity in the statistical learning community. The common idea behind these methods is to express prior beliefs about the correlations, the similarities, between pairs of points in data space X in terms of a kernel function K : $X \times X \rightarrow R$, and thereby to implicitly construct a mapping $\varphi : X \rightarrow H\kappa$ to a Hilbert space Hκ, in which the kernel appears as the inner product,

$$K(x, x') = (\varphi(x), (x')) \qquad (1)$$

[57]. With respect to a basis of $H\kappa$, each data point then splits into (a possibly infinite number of) independent features, a property which can be exploited to great effect. Graph-like structures occur in data in many guises, and in order to apply machine learning techniques to such discrete data it is desirable to use a kernel to capture the long-range relationships between data points induced by the local structure of the graph.

3.1.2 Diffusion kernels

Diffusion kernels is a special class of exponential kernels, [58] based on the heat equation [59], and show that these can be regarded as the discretisation of the familiar Gaussian kernel of Euclidean space.

An undirected, unweighted $\{u_{1,}u_2\}$ graph Γ is defined by a vertex set $V$and an edge set E, concluding to a set of unordered pairs $\{u_{1,}u_2\}$ where $\{u_{1,}u_2\} \in \mathrm{V}$ whenever the vertices $u_1\ and\ u_2$ are joined by an edge ($u_1 \sim u_2$).
Equation

$$\frac{d}{d\beta}K_\beta = HK_\beta$$

, where β is real parameter, suggests using an exponential kernel with generator

$$H_{ij}=\begin{cases}1 & for\ i\sim j\\ -d_i & for\ i=j\\ 0 & otherwise\end{cases} \qquad (1)$$

, where d_i is the degree of vertex i, the number of edges emanating from vertex i.
The negative aspect of this matrix is called the Laplacian of Γ , and it plays a central role in spectral graph theory. For any vector $w \in R^{|V|}$, complies that

$$w^T Hw = -\sum_{\{i,j\}\in E}\left(w_i - w_j\right)^2$$

, showing that H is, negative semi-definite. The following functions,

$$\{f: V \to R\}\ by\ (Hf)(x) = \sum x^{'}H_{x,x^{'}}\,f(x')$$

, implement that H can also be considered an operator. In fact, it is easy to show that on a square grid in m-dimensional Euclidean space with grid spacing h, H=h2 is just the finite difference approximation to the familiar continuous Laplacian

$$\Delta = \frac{\partial^2}{\partial x_1^2} + \frac{\partial^2}{\partial x_2^2} + \cdots + \frac{\partial^2}{\partial x_m^2},$$

and that in the limit h →0 this approximation becomes exact.
In analogy with classical physics, where equations of the form

$$\frac{\partial}{\partial t}\psi = \mu\Delta_{\psi}$$

are used to describe the diffusion of heat and other substances through continuous media, the equation

$$\frac{d}{d\beta}K_{\beta} = HK_{\beta}$$

, with H as defined in (1), is called the heat equation on Γ, and the resulting kernels are called diffusion kernels or heat kernels.
We remark that diffusion kernels are not restricted to simple unweighted graphs. For multigraphs or weighted symmetric graphs, all that is needed is to set H_{ij} to be the total weight of all edges between i and j and reweight the diagonal terms accordingly.

3.1.2.1 Random Walks

It is well known that diffusion is closely related to random walks. A random walk on an unweighted graph Γ is a stochastic process generating sequences $z_{0,}z_{1,}z_{2}$ …where $z_1 \in V$ in such a way that $p(z_{1+1} = j|z_1 = i) = \frac{1}{d_i} if\ i{\sim}j$ and zero otherwise. A lazy random walk on Γ with parameter $\beta \leq \frac{1}{max_i d_i}$ is very similar, except that when at vertex i, the process will take each of the edges emanating from i with fixed probability

$$\beta, p(z_{1+1} = j|z_1 = i) = \beta\ for\ i{\sim}j\ .$$

, and will remain in place with probability $1 - d_i\beta$. Considering the distribution p ($z_n|z_0$) in the limit $\Delta_t \to 0$ with N=1/(Δ_t) $and\ \beta = \beta_0\Delta_t$, concludes that diffusion kernels are the continuous time limit of lazy random walks. This analogy also shows that $K(i,j)$can be regarded as a sum over paths from i=j, namely the sum of the probabilities that the lazy walk takes each path.

3.1.2.2 Random walk with restart

Given a gene/protein interaction network, random walk starts from a protein g and at each time step moves to one of the neighbors with the probability $1-\beta$ $(0 \leq \beta \leq 1)$. The walk can also restart from g with probability β.

This process is defined by a transition matrix W.

$$W_{ij}=\begin{cases}\frac{1}{\deg(j)} & if \;\; i\; interacts\; with\; node\; j \\ 0 & otherwise\end{cases}$$

, deg (j) is the number of neighbors or the degree of protein g_i, in the interaction network. The variable β represent the probability with which the walk starting at g_i is forced to restart from g_i.

The random walk will reach a stationary distribution described by the vector $s_i^{\rightarrow}$, the i^{th} column of F.

$$s_i^{\rightarrow\prime}=(1-\beta)Ws_i^{\rightarrow} \;\; +\beta s_i^{\rightarrow}$$

When $s_i^{\rightarrow\prime}=s_i^{\rightarrow}$ we can get

$$s_i^{\rightarrow} = \beta(I-(1-\beta)W)-1e_i^{\rightarrow}$$

$e_i^{\rightarrow}$ is the vector with 1 in position i and 0 is in the remaining positions.

The part $\beta(I-(1-\beta)W)-1$ is called diffusion matrix F.

$$F = \beta(I-(1-\beta)W)-1$$

3.2.2 Hotnet2

HotNet2 (diffusion-oriented subnetworks) is the updated version of the original HotNet algorithm proposed in [7]. It is an algorithm for finding significantly altered subnetworks in a large gene interaction network and it was developed for identifying significantly mutated groups of interacting genes from large cancer sequencing studies. It uses an "insulated" heat diffusion process, a diffusion kernel analogous to random walk with restart to model diffusion of heat, or gene scores, to better capture the local topology of the interaction network and filters the graph by an automatic threshold selection.

HotNet2 rather than selecting connected components in an undirected influence graph identifies strongly connected components in a directed influence graph. It is a general algorithm for choosing high weight subnetworks in a vertex-weighted network. The diffusion process applied encodes the source of heat within the network allowing HotNet2 to uncover significantly "hot" subnetworks with wide ranges of heat scores.HotNet2 also uses an asymmetric influence score, different permutation as a testing method as well as parameter selection procedures.

HotNet2 algorithm performs the following steps:

i. **Heat Diffusion**. HotNet2 employs an insulated heat diffusion process [61,62] that captures the local topology of the interaction network surrounding a protein. At each time step, nodes in the graph pass to and receive heat from their neighbors, but also retain a fraction β of their heat, governed by an insulating parameter β, defined by the Multinet interaction network. In this work β=0.50. The process is run until equilibrium; the amount of heat on each node at equilibrium thus depends on its initial heat, the local topology of the network around the node, and the value β. If a unit heat source is placed at node j (e.g. a gene in g_j in one sample) then the amount of heat on node i is given by the (i, j) entry of the diffusion matrix F defined by,

$$F = \beta(I - (1-\beta)W) - 1$$

where

$$W_{ij} = \begin{cases} \dfrac{1}{\deg(j)} & if \;\; i \; interacts \; with \; node \; j \\ 0 & otherwise \end{cases}$$

, as we have already mentioned.

W is a normalized adjacency matrix of the graph G. We interpret $F(i,j)$ as the influence that a heat source placed on $g_j$has on g_i. The insulated diffusion process is generally asymmetric, i.e. $F(i,j) \neq F(j,i)$.
The diffusion matrix F depends only on the graph G, and not the heat vector $\vec{h}$ which represents the input. Therefore the influence for a given β needs to be computed only once for a given interaction network.

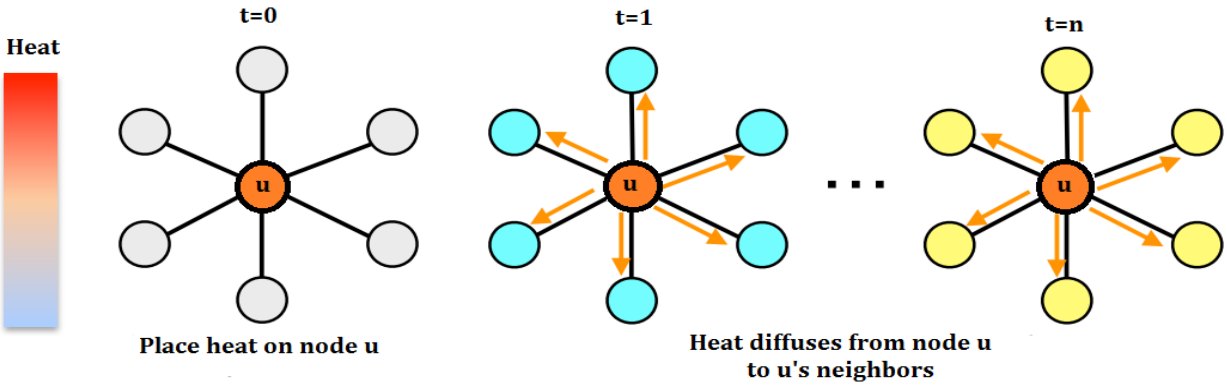


Figure3.1 Heat diffusion process

ii. **Exchanged heat matrix**. The insulated heat diffusion process described above encodes the local topology of the network, assuming unit heat is placed on nodes. To jointly analyze network topology and gene scores given by the initial heat vector $\vec{h}$, we define the exchanged heat matrix E:

$E = FD_{\vec{h}}$ where $D_{\vec{h}}$ is the diagonal matrix with entries $\vec{h}$.

$E(i,j) = F(i,j)\vec{h}(j)$ is the amount of heat that diffuses from node g_j to node $g_i$on the network $when\ \vec{h}(j)$ heat is placed on g_j , which we interpret as the similarity $E(i,j)$,also not symmetric.

iii. **Identification of hot subnetworks**. A weighted directed graph H is formed whose nodes are all measured genes. If $E(i,j) > \delta$, then there is a directed edge, from node j to node i, of weight $E(i,j)$. HotNet2 identifies strongly connected components in H.
A strongly connected component C in a directed graph is a set of nodes such that for every pair u, v of nodes in C there is a path from $u\ to\ v$.

iv. **Statistical test for subnetworks**. HotNet2 returns a list of subnetworks, each containing at least S genes, and employs a two-stage statistical test [63] to assess the statistical significance of the returned list of subnetworks. The first stage of the test computes a p-value for the number of subnetworks with at least S genes that are returned, for different values of S. We can thus determine an S such that the number of connected components of size ≥ S is significant, with a particular p-value. This p-value measures the significance of the number of subnetworks of a minimum size s, but does not say which, if any, of the individual subnetworks is significant. The second stage of the test estimates the false discovery rate (FDR) ,the expected ratio of false discoveries among the subnetworks assessed as significant by the method [64]. To calculate the diffusion matrix HotNet2 has two parameters β and δ, described in section 3.2.2.2, and selects values for both of these parameters using automated procedures. The variable β is selected from the protein-protein interaction network, independently of any gene scores and it is chosen to balance the amount of heat that diffuses from a protein to its immediate neighbors and to the rest of the network. In our case β=0.50. The value of δ is the edge weight parameter. It's used to make sure the HotNet2 will not find large subnetworks using random data.

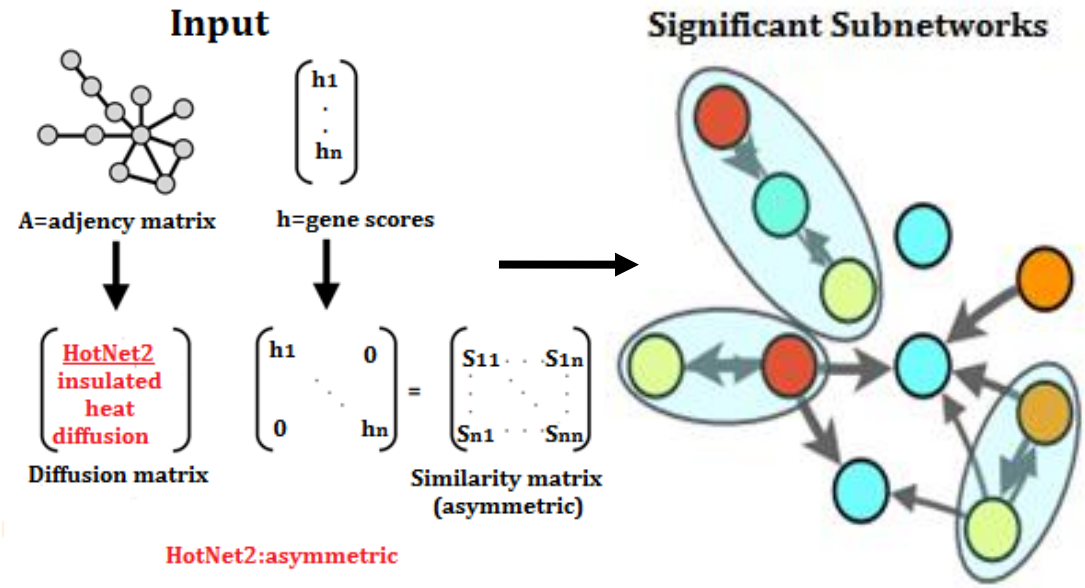


Figure3.2 Generalization of Hotnet2 for clinical data

3.2.2.2 Parameters β and δ

- **Insulated heat-diffusion: β**

  The parameter β is chosen for a given protein-protein interaction network, and remains fixed for different heat datasets. β is chosen in order to balance the amount of heat that diffuses from a protein to its neighbors and to the rest of the network. This was done by computing the amount of heat retained in the neighbors of vertices ("source proteins") with different network centrality. In choosing parameter β the goal is to choose a variable such that all proteins retain most of their heat in their immediate neighbors. The process is described in detailed in [24].

- **Minimum edge weight: δ**

  The edge weight parameter δ is chosen such that to avoid finding large subnetworks using random data. Specifically, for each dataset and interaction network, we generated 100 random networks with the same degree distribution as the original network by performing $Q \times |E|$ connected edge swaps, where E is the set of edges in the interaction network, ensuring that each node retains the same degree as in the original network and that the resulting network is connected, setting Q = 100 [65]. In order to conclude in the δ parameter each permuted network is examined and the minimum δ is identified such that all strongly connected components found have size $\leq$Lmax, for Lmax = 5, 10,15, 20. For each Lmax, the median of these values of δ is calculated across the 100 permuted networks .For each run, the selected value is the smallest δ with the most significant (P < 0, 05) subnetwork sizes k.

3.2.3 Activity Vector

A remaining hurdle to pathway-based analysis is that the majority of human genes have not yet been assigned to a definitive pathway. As we have already mentioned protein–protein interaction networks are used to assign sets of genes to discrete subnetworks. The Activity Vector algorithm is a protein-network-based approach for identifying markers of metastasis within gene expression profiles, which can be used to identify genetic alterations and to predict the likelihood of metastasis in unknown samples. These markers consist of subnetworks of interacting proteins within a larger human protein–protein interaction network and are not individual genes or proteins.

The primary input to Activity vector consists of a gene expression matrix of the 4174 genes and their expression levels, as well as an interaction network. In order to examine both algorithms and their resulting subnetworks under the same construction criteria we used the Multinet interaction network mentioned previously

Activity Vector algorithm performs the following steps:

i. **Creation of subnetwork activity matrix**
The gene expression levels drawn from each type of cancer (i.e., control or cancer) are transformed into a 'subnetwork activity matrix'. For a given subnetwork Mk in the interaction network, the activity α ,represents its vector of activity scores over the samples, is a combined z-score, which is designated the activity α_{kj}, and c represent the corresponding vector of class labels (cancer or control).In order to result in α_{kj}, expression values g_{ij} are normalized to z-transformed scores z_{ij} ,a process that assists in the use of data directly in the calculation of significant changes in gene expression between different samples and conditions. The activity matrix calculation derives from:

$$a_{kj} = \sum_{i} \frac{z_{ij}}{\sqrt{n}}$$

, where n is the number of studied genes. The score z_{ij}, has for each gene i, mean μ=0 and s.d. σ=1, over all samples j and the individual z_{ij} of each member gene in the subnetwork are averaged into a_{kj}.[66].

ii. Searching for significant subnetworks

After overlaying the expression vector of each gene on its corresponding protein in the interaction network, subnetworks with considerably significant activities are found via a greedy search. A subnetwork can be referred as a module of the given protein network. The process for calculating a module starts only with a starting node which can be any node in the protein network and iteratively expands. The module expands under two criteria:

- the new node is adjacent to any node already in the module, and
- it improves the overall score, a value that measures how "good" a module is, of the module.

If no node can be added that meet the two criteria above, the process of building a module stops due to the fact that no addition increases the score over a specified improvement rate r. A module is calculated for each node/protein in the network and all the collected modules are called real modules. Many types of statistic, such as the t or Wilcoxon score tests, could be used to score the relationship between a and c as well as the Mutual Information test which we have already mentioned in chapter 2.4 and we have chosen for this study.
It is considered that the median distance d between any two proteins in the human protein–protein interaction network is five, thus, we set d=2 to provide a sufficient number of neighbors while keeping the search local.
The parameter r is set 0.05 to avoid over-fitting to the expression data used. The majority of searches terminate due to the constraint on r, increasing the value of d has only marginal effect on the results.

iii. Permutation process

Most of the real modules were produced at random and are statistically insignificant. These modules must be removed. The way to filter out insignificant modules is by passing them through statistical tests. First, a gene expression vector and its corresponding protein are randomly associated. After the randomization all modules are recalculated and scored respectively. Finally these modules are collected together and are called random modules.
This randomized process is repeated many times. In this study we have chosen 100 random trials. In general, the more random trials, the better the results, but the computation time becomes longer.

iv. Selection of significant subnetworks

Significant subnetworks are selected based on null distributions estimated from permuted subnetworks, after three significance tests.

- Statistical Test 1: The scores of all random modules are collected together and are placed in a null distribution. If a real module's score is insignificant when compared against the null distribution, it is discarded.

- Statistical Test 2: The random module scores are used to estimate the parameters of a distribution. In this work we have selected mutual information for a scoring method, and the gamma distribution is used. If a real module has an insignificant score compared to the distribution, it is discarded.

- Statistical Test 3: We test whether the mutual information with the disease class is stronger than that obtained with random assignments of classes to patients. The gene expression vectors of a real module are combined into one vector. A score is calculated based on this vector. The order of the vector's columns is then randomized in 20.000 trials. Another score is calculated from this randomized vector. This randomization process is repeated many times, yielding a null distribution of mutual information scores for each trial. Finally, the real score of each subnetwork is indexed on this null distribution and if compared to it is insignificant, the module is discarded. In this study, significant subnetworks are selected hat satisfy all three tests with P_1 <0.05, P_2 <0.05, and P_3 <0.00005.

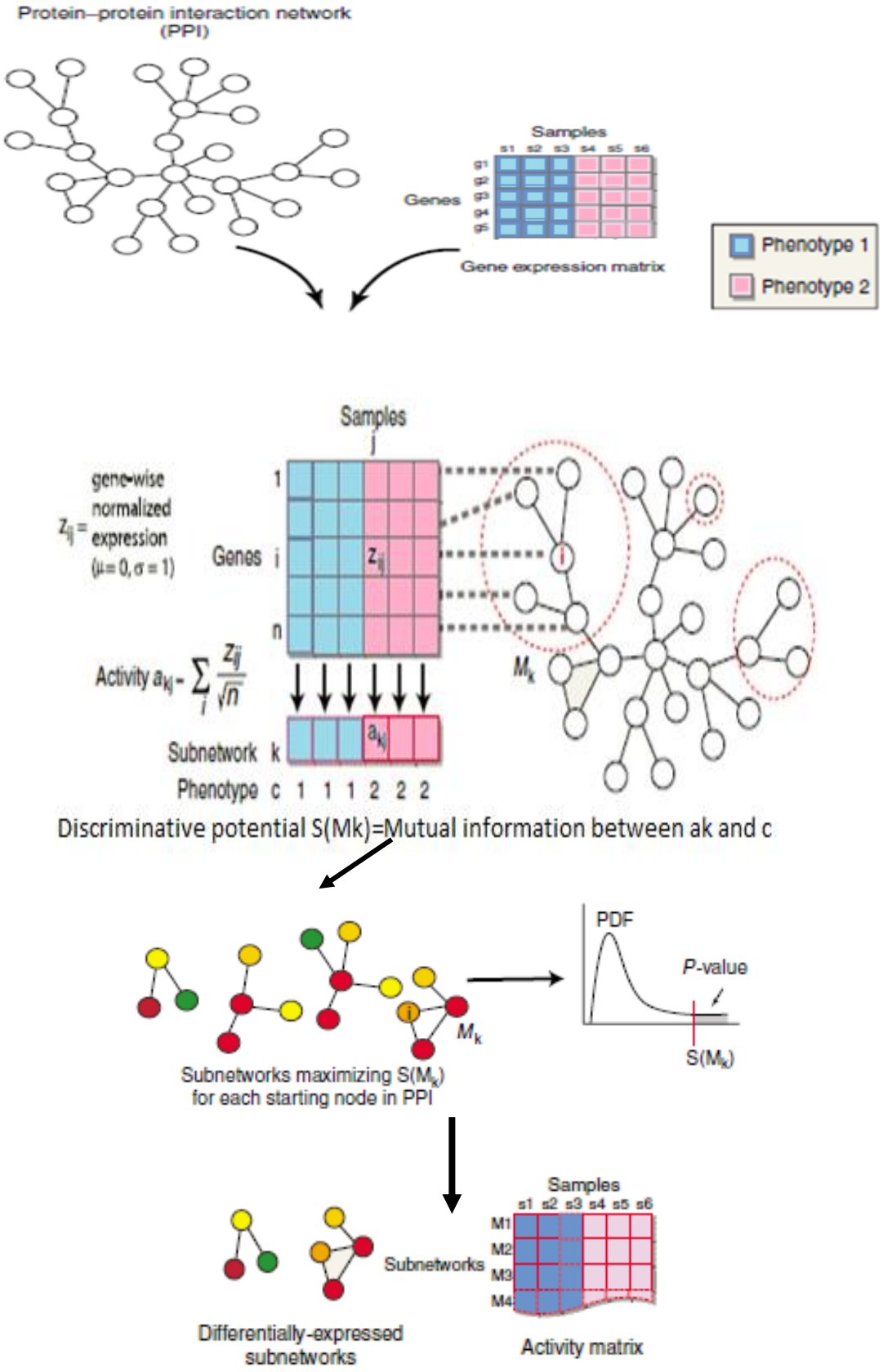


Figure3.3 The generalized process of the Activity vector algorithm [22]

3.3 Evaluation of the Results

3.3.1 Statistical Evaluation-Generalization

In the third part of our methodology the evaluation of our results from the implementation of both algorithms is presented, in section. In microarray and data analysis evaluation methods are used to estimate the generalization ability of genome signature, that is to discover predictive relationships of the results in unknown data .Evaluation methods can be performed in a portion of the existing dataset as well as in an independent/new dataset, called the training set while a test set is used for evaluating whether the discovered relationships are accurate. A test set is a set of data used to assess the strength and utility of a predictive relationship. Cross-validation, explained in section 2.7 is a well known and used strategy because of its simplicity and its universality. The k – fold cross validation approach, implemented in this study, can also be used to assess how the results of a statistical analysis will generalize to an independent data set. In this context, a new independent dataset is used and the procedure of 10 – fold cross validation is repeated.

3.3.2 Biological Evaluation

Apart from the important step of the statistical evaluation of our results and their prediction ability, a fundamental role in the process of evaluation is the biological significance of the resulted genes. In combination, these two methods can help us uncover known as well as new relationships between genes/proteins which if applying either one or the other separately, our conclusion would be incomplete and would lack in terms of statistical as well as biological significance. In section the results from the biological evaluation of our resulted genes is presented.

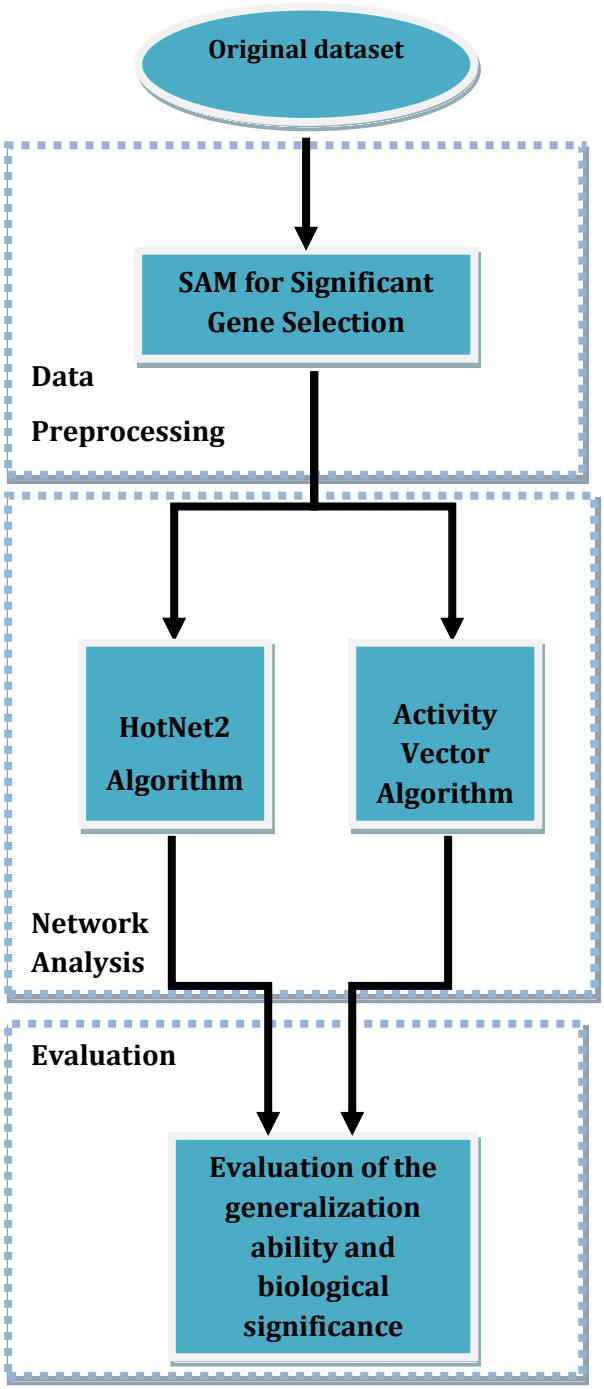


Figure3.4 Proposed methodology

3.4 Significant Analysis of Microarrays (SAM)

As we have already mentioned the feature gene subset selection process is one of the crucial steps in DNA microanalysis. In section 2.4.1.1 we have presented the theoretical knowledge of the filter univariate method called "Significance Analysis of Microarrays" (SAM) [40, 41]. Here we present the general process of SAM in terms of microarray analysis with gene expression values from two classes of pathological subjects. Our dataset is composed of 529 (mRNA) samples which have a large number of gene expression (4174 genes) levels as features. This data set had all ready undergone a filtering process with SAM, in order to reduce the number of genes by keeping the most significant set from a larger dataset. SAM uses a modified t-statistic and permutations of the repeated measurements of the data in order to decide if the gene expression is strongly related to the response.
The SAM procedure follows.

i. **SAM Input**: The input to SAM is gene expression measurements from a set of microarray experiments, as well as a response variable from each experiment. The data should be put in an Excel spreadsheet and have a specific format. The overview of the file must meet some requirements in order to proceed successfully. The first row has the information about the response measurement; all remaining rows have gene expression data, one row per gene. The columns represent the different experimental samples.

ii. **Class definition**: There are many different types of response such as quantitative, one class, two class (unpaired, paired), multiclass, survival data, time course and pattern discovery. In our case, gene expression measurements were separated into two classes (unpaired). These classes are two sets of measurements, in which the experiment units are all different. Particularly, we have two groups: healthy controls and cancer patients, with samples from different patients. Thus the response variable is grouped using numbers 1 (healthy control) – 2 (cancer patient).

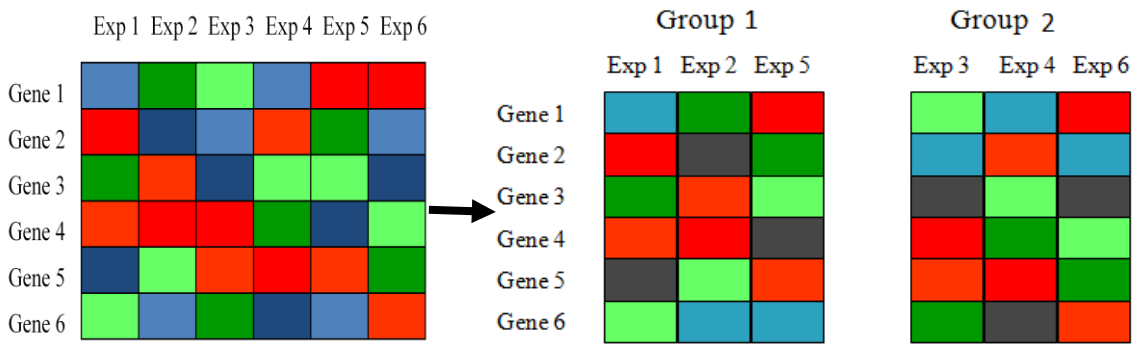


Figure3.5 Assign experiments to two groups (1, 2)

iii. **Processing input:** The area that represents the data should be highlighted. Then the SAM button in the toolbar must be selected and a dialog rises. The dialog box gives the opportunity to the user to select the type of response variable and to change any of values of the default parameters. Moreover the user should specify if the data are from (micro)array or a sequencing experiments and for two class and paired data, one has to specify if the data is in the logged (base 2) scale or not

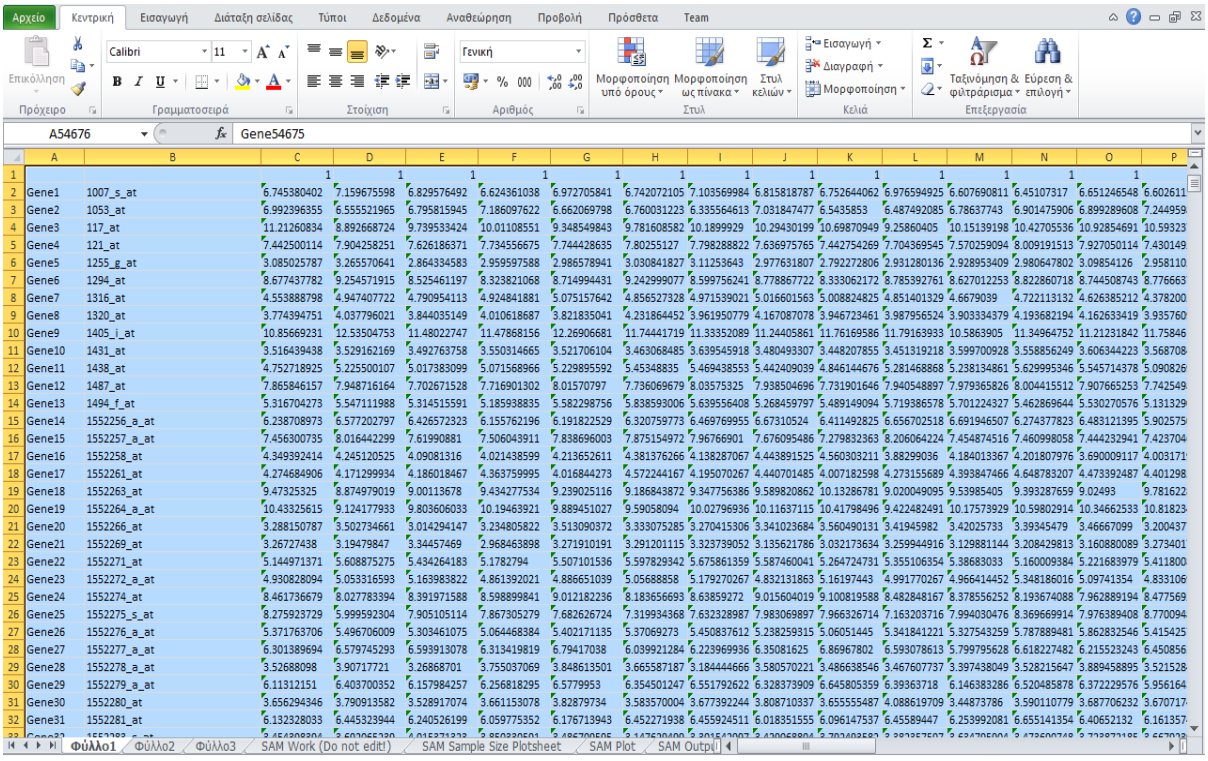


Figure3.6 Highlighting and invoking SAM

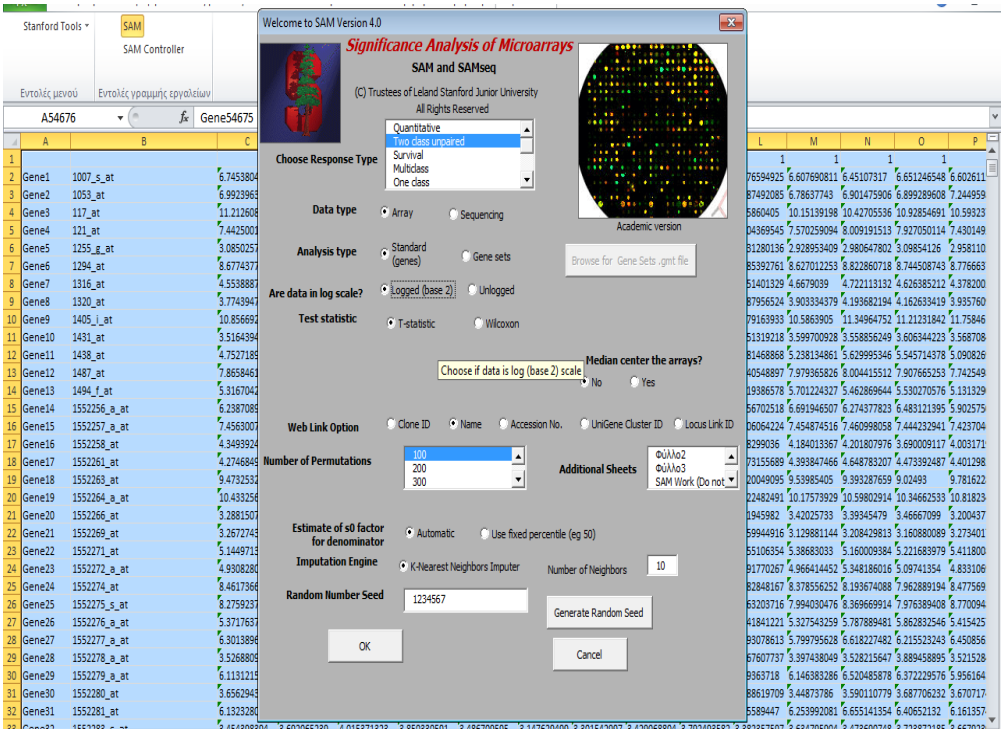


Figure3.7 SAM Dialog Box

In order to determine if the expression of any genes is significantly related to the response the procedure of repeated permutations of the data is as follows:

1. For each gene, compute a statistic d-value, which is the observed d-value for that gene.
2. Order the genes according to their d- values.
3. Randomly shuffle the values of the genes between groups 1 and 2, such that the reshuffled groups 1 and 2 respectively have the same number of elements as the original groups 1 and 2. Compute the d-value for each randomized gene.

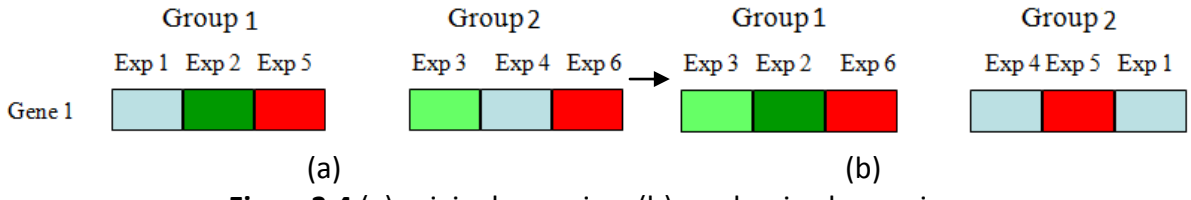


Figure3.4 (a) original grouping, (b) randomized grouping

4. Order the genes according to their permuted d- values.
5. Repeat steps 3 and 4 many times. Thus, each gene has many randomized d-values corresponding to its rank from the observed (unpermitted) d-value (100 or 200 permutations are descent for initial exploratory analysis). Then, take the average of the randomized d-values for each gene which is the expected d-value of that gene.

6. Plot the observed d-values versus the expected d-values
7. For each permutation of the data, compute the number of positive and negative significant genes for a delta parameter, which is the cutoff for significance, chosen by the user based on the false positive rate. The median number of significant genes from these permutations is the median False Discovery Rate (FDR). Thus, any genes designated as significant from the randomized data are being picked up purely by chance. Therefore, the median number picked up over many randomizations is a descent estimate of FDR.

iv. **SAM Output**: While running the SAM, if there is any missing data in your spreadsheet, a new worksheet named SAM Imputed dataset containing the imputed dataset is added to the workbook, unless this worksheet is not added. Therefore, the software adds two more worksheets to the workbook. There is one which is hidden called SAM Plot data which contains the plot of the observed d-values versus the expected d-values and the user can interact with. Particular, a block dialog which is called Sam Plot Controller, shown in figure 3.6, gives the chance to the user to change the delta parameter and examine the effect on the false positive rate. If user wants a more stringent criterion, there is also a fold change parameter that he can select. Positive significant genes are labeled in red on the SAM plot, while negative significant genes are green. The List Delta Table button lists the number of significant genes and the false positive rate for a number of values of delta. The List All Genes prints out all genes in the dataset. After choosing the delta parameter a sheet named SAM Output is showed, including any output.

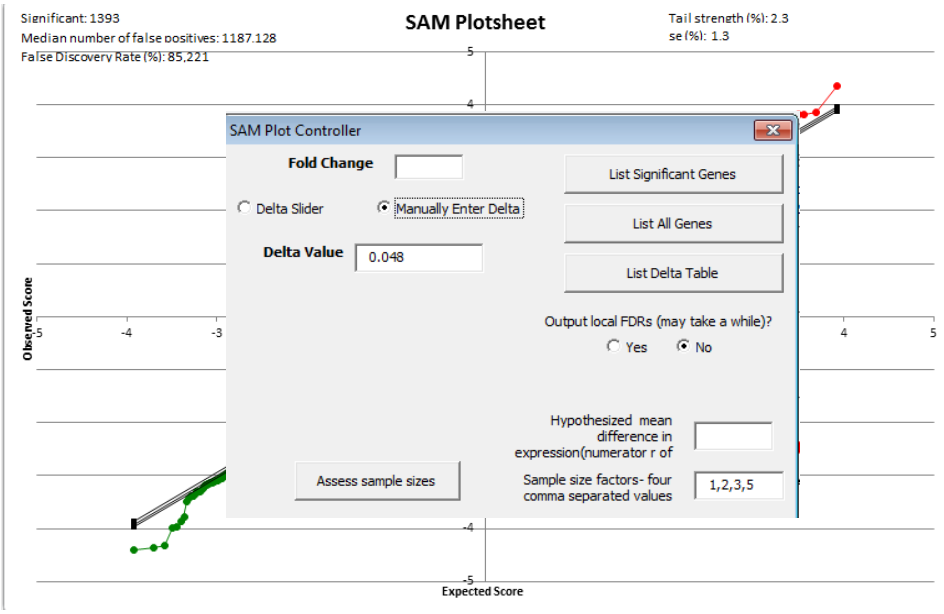


Figure3.8 The SAM Plot Controller on the front side,
The SAM Plot sheet on the second sheet

The output for list of significant genes has a specific format. Particularly, it contains the row number, which is the row in the selected data rectangle, the gene name as well as the gene Id. It also contains the SAM score (d), which is the t-statistic value with the numerator and the denominator(s + s0) of it. Moreover, the q-value, which is the lowest False Discovery Rate at which the gene is called significant as well as the local FDR, which is the false discovery rate for genes with scores d that fall in a window around the score for the given gene are also printed . Finally, in any testing problem, false positive rate (for example FDRs) are calculated, but is also important to consider false negative rates. Thus, a miss rate table is printed which gives the estimated false negative rate for genes that do not make the list of significant genes.

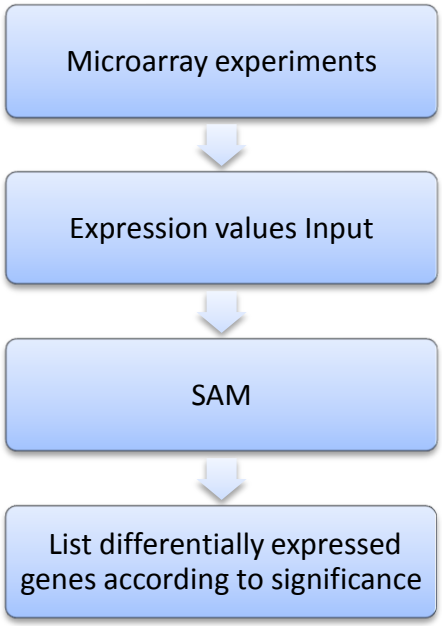


Figure3.9 Processing data set with SAM

In the beginning of this thesis conduction we processed our genes aiming in a further reduction of their number so as to conclude in a smaller subset of significant differentially expressed genes. After using SAM, we reached the number of 84 significant genes. Though, the purpose of the HotNet2 and Activity Vector algorithms is to find significant subnetworks of genes through a large data set of both genes that are significant and genes that are not. Due to the requirements of the two algorithms we have selected to work with the original data set of 4.174 genes.

4

Results

In this chapter we present the results from our proposed methodology. In section 4.1 we introduce the dataset that we have used, followed by section 4.2, 4.3 were the resulting subnetworks from each algorithm implementation is presented. Furthermore, in section 4.4 the statistical significance as well as the resulted genome signature significance of the SVM implementation results is assessed.

4.1 Dataset

The original, preprocessed, dataset consists of 4.174 breast cancer genes and the measurements of their expression values, acquired from two groups of 529 (mRNA) samples, 104 healthy control and 425 cancer cases and it was provided by Stelios Sfakianakis.

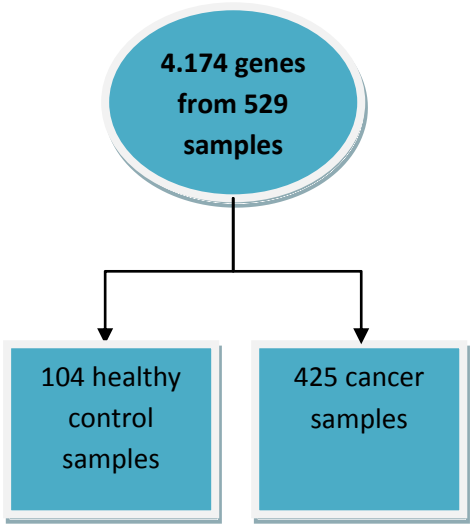


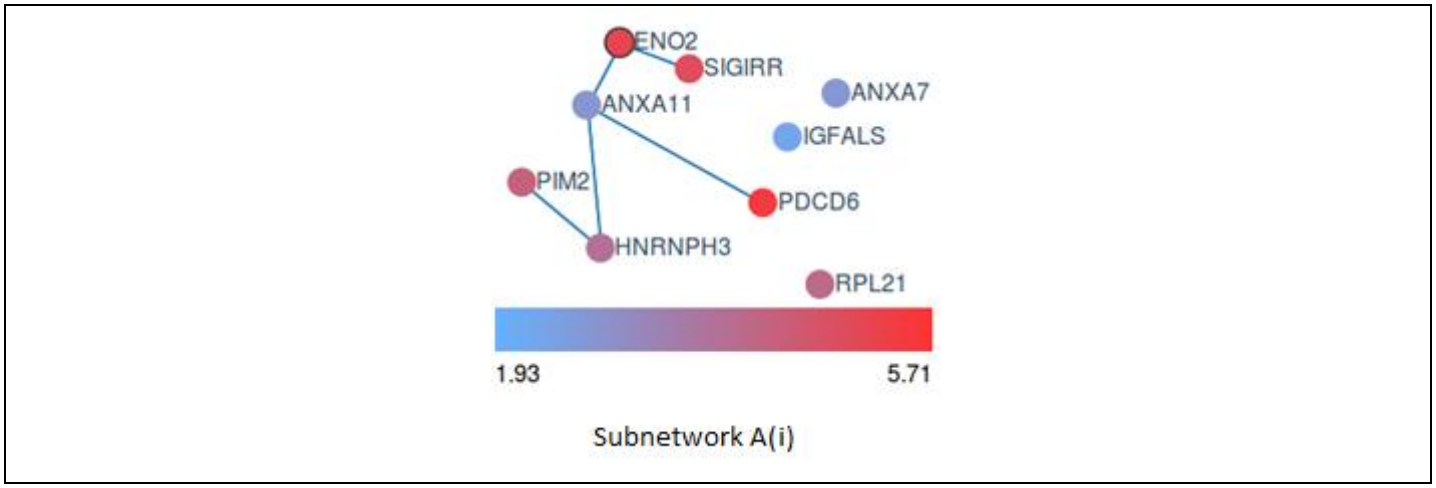
Figure 4.1: Dataset structure

The primary input to HotNet2 consists of a list of genes and gene scores, as well as an interaction network, represented by an influence matrix describing interaction levels between genes. In this work one set of input data was used for network-based analysis and the used network for this set derived from Multinet, an interaction network which includes protein-protein, phosphorylation, metabolic, signaling, genetic and regulatory interactions from multiple databases. The MultiNet network consists of 109,597 interactions among 14,445 proteins. After the removal of self-loop interactions and multiple edges between interactions the resulting MultiNet influence matrix, represented as a sparse square matrix in MATLAB, represented a network of 14,398 genes by defining an influence score for each gene pair based on Multinet's gene interactions and a heat diffusion process. Breast cancer gene expression data was provided by Stelios Sfakianakis. These data consisted of fold changes in expression from healthy control subjects to breast cancer patients for 4.174 preprocessed genes. Corresponding p-values were based on the magnitude of these fold change and gene-specific variability of expression, as well as the adjusted for multiple hypothesis testing p values. Because the testing procedure is carried out for hundreds of gene sets, it is critical to apply a proper adjustment to control for type I error. In our analysis, we use the q value, which is a counterpart of the p value in the context of false discovery rate, to assess the statistical significance of gene associations [60].From these 4.174, 3.385 genes were contained in the Multinet network. HotNet2 assumes larger scores are more important, so in this context, the q-values of the dataset had to be inverted. Thus, the score used for analysis was − log10 (q), which produced an appropriate distribution of scores to differentiate genes without extreme variance.

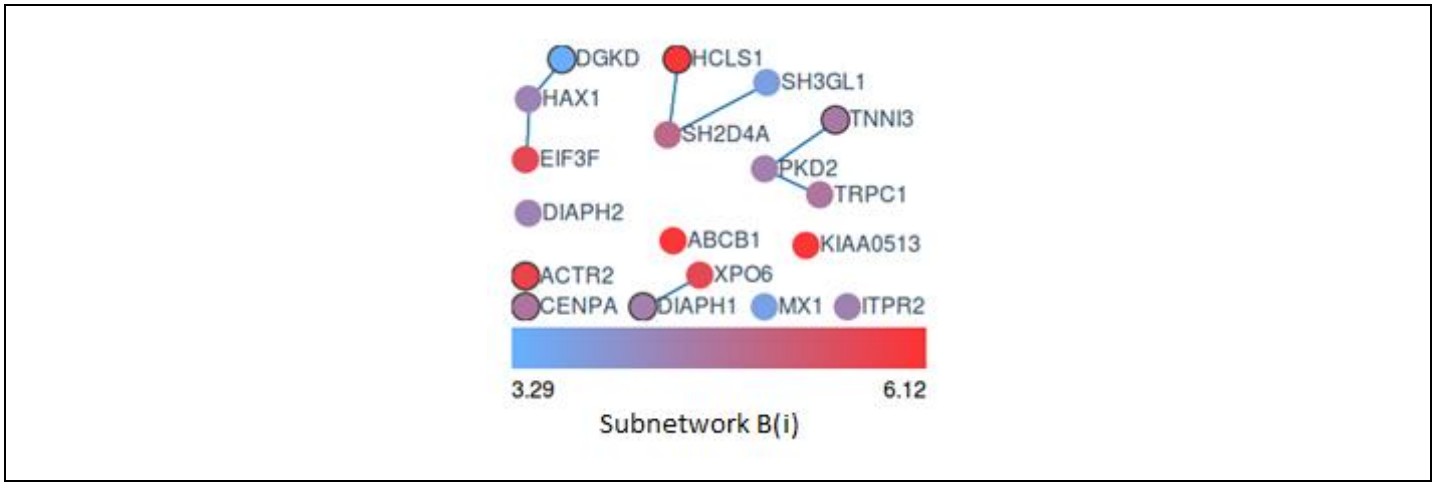
4.2 HotNet2 Algorithm results

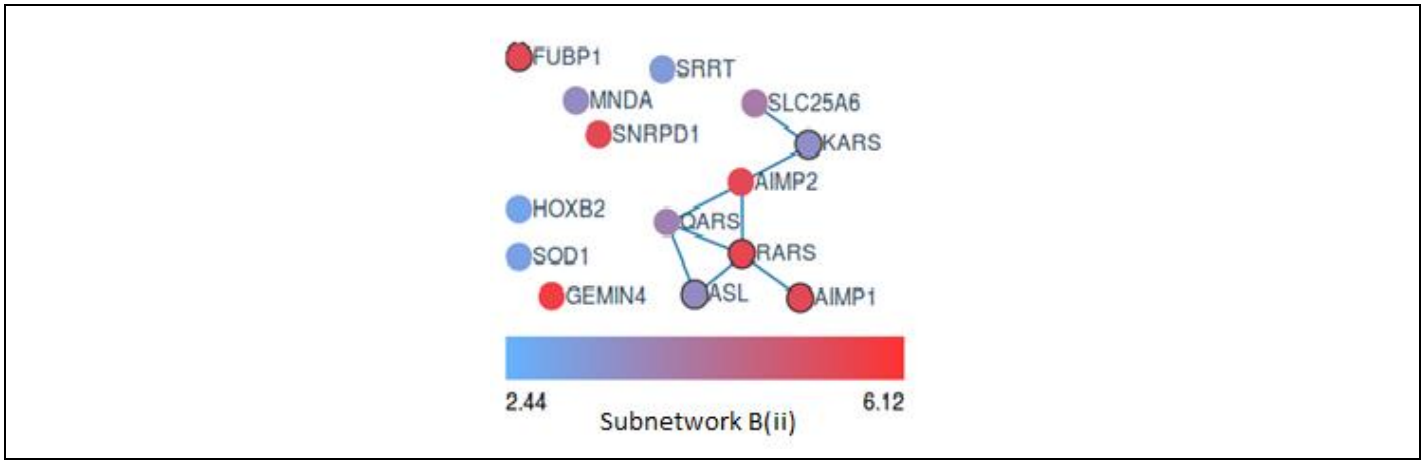
As we have discussed above, the following results were produced from HotNet2 algorithm, after using a gene score conversion of − log10 (q). The gene expression data set was analyzed with the Multinet network.

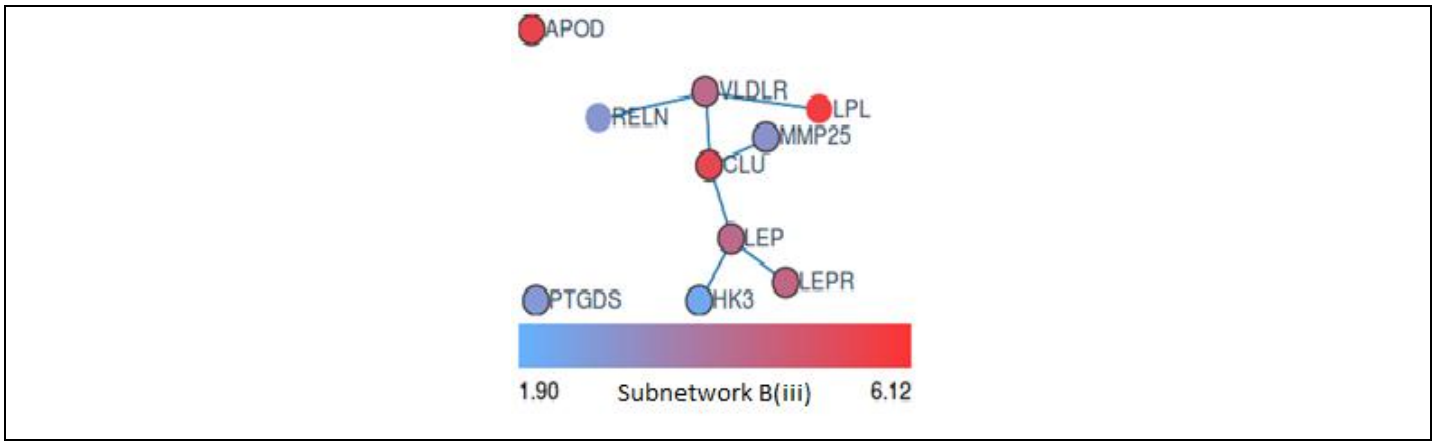
After analyzing the genes scores from the breast cancer data in HotNet2, 4 significant subnetworks were found. In the resulting subnetworks, shown in figure 4.2, where 1 cluster of 9 genes was found and 3 clusters of at least 10 genes. Based on HotNet2 tests permuting the gene scores across the tested genes, the probability of finding 1 clusters of size 9 was 0.07 and 3 clusters of size at least 10 was 0.04 shown in Table4 1. Enrichment p-values are based on a null hyper geometric distribution over genes in the HotNet2 and Bonferroni-corrected for the number of pathways and components tested.



(A)







(B)

Figure 4.2 (A) Subnetwork A(i) of k minimum size of components 9 (B) Subnetworks B(i), B(ii), B(iii) of k minimum size of components 10. Color depicts gene scores.

SIZE k	EXPECTED	ACTUAL	P-VALUE
2	181.6	187	0.3
3	61.24	61	0.61
4	27.7	23	0.95
5	13.83	12	0.81
6	7.71	6	0.84
7	4.56	5	0.46
8	2.84	4	0.3
9	1.61	4	**0.07**
10	1.03	3	**0.04**

Table4.1 p-values for the significance of clusters of a given size, based on expected numbers of clusters from permutation tests of gene scores.

4.3 Activity Vector Algorithm results

In order to analyze the genes scores from the breast cancer data with Activity vector we used the Cytoscape plug-in, PinnacleZ, which is the implementation of the algorithm presented in [22]. As we have already mentioned our goal is to result in significant subnetworks after applying the same criteria in both algorithms. Therefore the input gene expression data set was also analyzed with the Multinet network. After the implementation of the algorithm we resulted in 332 real modules. The common denominator of both methodologies is ten common genes which appear to be common. From the 332 modules we decided to analyze the ones that had these common genes with the HotNet2 resulting subnetworks as well as the highest module score. The 10 common genes are the following:

HCLS1
AIMP2
APOD
SIGIRR
ENO2
PDCD6
ACTR2
HNRNPH3
TRPC1
EIF3F

Table4.2 Mutual genes between HotNet2 and Activity vector resulting subnetworks.

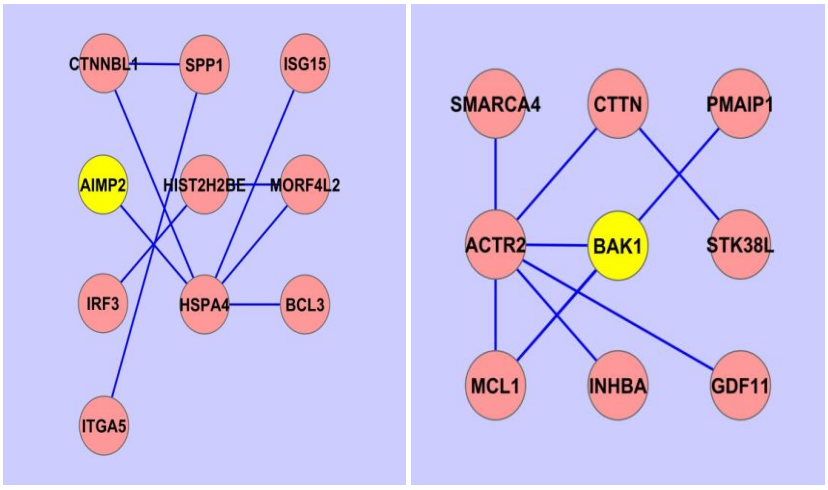


Figure 4.3 Two representative Activity Vector subnetworks

The Activity Vector subnetworks, with the higher module score, involving the 10 common genes, are presented in the next table.

network 1	network 2	network 3	network 4	netwok 5	network 6	network 7	netwok 8	network 9	network 10
BCL3	BCL3	ADRB2	HIST2H2BE	EIF4B	EIF4B	ACTR2	CSNK2A2	GATA2	YY1
BUB3	AIMP2	APOD	CALM2	ATF2	ANKS1A	BAK1	YY1	IRF3	ATF4
CENPE	CTNNBL1	EIF4B	CFL1	CRIM1	CAV1	CTTN	HDAC5	CDC45	BCLAF1
HCLS1	HIST2H2BE	FGFR1	E2F1	CSNK2A2	CSRP1	GDF11	ABCE1	ENO2	ACTG2
HHAT	HSPA4	GRB10	GATA2	IRAK1	EDNRB	INHBA	ATF4	ENSA	CCDC106
MAPK13	IRF3	GRINL1A	H2AFX	PELI2	EPHB3	MCL1	BCLAF1	HIF1A	EEF1G
PITX1	ISG15	HDAC5	LEF1	RBM17	PCBP2	PMAIP1	EEF1A1	MCM3	HNRNPH3
TRAF4	ITGA5	MYO9A	PDCD6	SIGIRR	SCP2	SMARCA4	EIF3F	NFYA	KAT2B
USP7	MORF4L2	NEDD4	THOC4	SOX10	TRIM24	STK38L	RPS7	PTPN1	KAT5
	SPP1	PDCD6IP		YY1	TRPC1			SMARCA5	PTPN4
		PRPF8						SREBF1	TRIM68
		SMURF2						YWHAQ	

Table 4.3 Activity vector highest scoring modules involving common genes to HotNet2.Color red depicts each of the 10 common genes.

In this point we have to mention that network1 and network2 might have small differences after every run of the algorithm, although keeping the same module score. We have decided to work with an instance of the algorithm implementation and concluded in the total number of 89 genes. The genes we have assembled from the combination of both algorithms, consist the most significant genomic signatures from the original 4.174 genes dataset. In section 4.4 that follows we present the statistical and biological evaluation of the results.

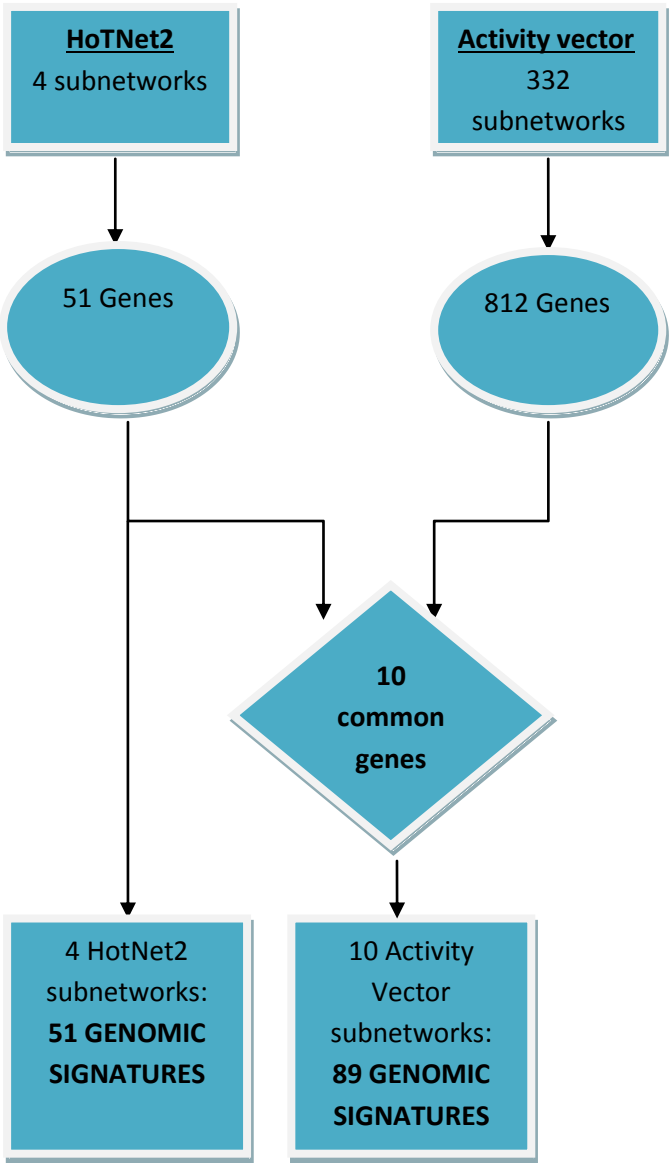


Figure 4.4 Process for resulting genomic signature

4.4 Generalization Ability of Genomic Signature

New Dataset

The new dataset was selected from Gene expression Omnibus (Geo) database involving gene expression profiles of 104 breast cancer and 17 normal breast biopsies. The GEO access number is GSE42568.For each sample, there are measurements of 54.675 genes. Each value is calculated as Log2 GC-RMA signal intensity.

Generalization Ability

As already mentioned in section 3.3.1, the aim of this field is to access the generalization ability of the genomic signature. A good generalization performance is achieved when a genomic signature is able to predict the label of unseen samples correctly. Our final genomic signature consists of 51 genes derived from HotNet2 algorithm as well as 89 genes, after the combination of both methodologies. After the implementation of both algorithms we resulted, as previously mentioned, in 10 common genes. As mentioned above, the 10 genes derived from the application of two different methodologies, that apply different clustering criteria on our data, and seem to stand out for their statistical significance in prediction to a new dataset. Thus, 89 genes, including the 10 common genes, are selected from the new dataset in order to be used for accessing the generalization ability of the model to an independent dataset. However, there are 24 genes that have the same code to the original dataset, while 65 genes of them appear with multiple codes in the new dataset. For this reason, concerning the 65 genes, with more than one code, standard deviation of each gene is extracted in order to decide which code is able to be used as shown in Figure4.5 and the same procedure was applied for every gene. The gene with the higher standard deviation is preferred, since these genes are extended over a wider range of values.

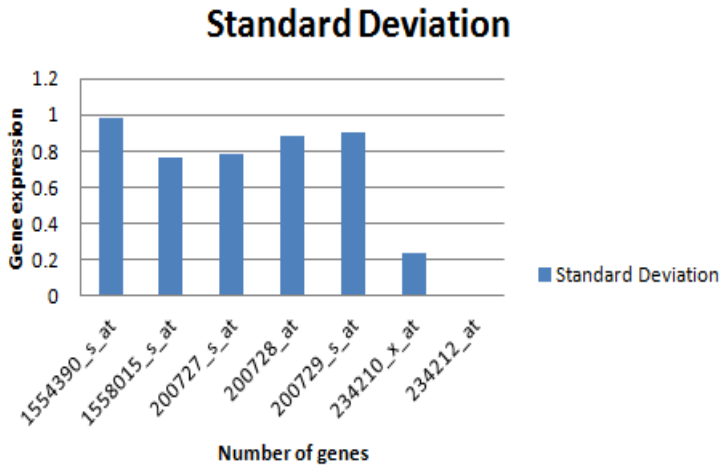


Figure4.5 Standard deviation of 6 multiple genes in the new dataset

Finally, the genomic signature of the new dataset is composed of 89 significant genes and is used to access the generalization ability of the model. As mentioned in section 3.3.1, the 10 fold cross validation approach is applied which generates 9 training datasets and only one test set. This process is repeated 10 rounds. In each round, one of the folds is used for validation, and the other 9 folds for training. Then the SVM as well as the RVM classification method is performed. The process is repeated 200 times and the overall results are averaged. Finally, as we can see in Table4.4 the observed mean classification accuracy is very good, performing very good generalization performance when it comes to the classification of unknown samples. In addition, 11 from 15 genes after applying the SVM algorithm are included in the 21 genes selected from RVM. Additionally, 2 genes (SIGIRR and PDCD6) from the starting dataset of 10 mutual genes are included in the resulting 21 and 15 genes, respectively. The same procedure was implemented to access the classification significance of the original set of the 10 common genes from which the 89 genes derived.

Genomic Signature Size (new dataset)	Mean Classification Accuracy (%)	Mean Genes
89 genes RVM	97.05%	21
89 genes SVM	95%	15
10 genes RVM	96.6%	8
10 genes SVM	93%	10

Table4.4 Generalization Ability of Genomic Signature Results

From the above results it appears that the 10 common genes – from the set of 89 important genes-also provided high classification accuracy in an independent data set. As mentioned above, the 10 genes derived from the application of two different algorithms seem to stand out for their statistical significance in a new dataset. The results enable as to assume that the 10 selected common genes represent significant seeds in order to proceed in an inquiry process, although they are not all part of the resulting 15 and 21 genes selected from the group of 89 genes. However, the two different methodologies considered in this study were selected because of their effectiveness in finding and creating small -potentially significant- subnetworks using a large network of known interactions (MultiNet).So in biological level we examine the set of 89 genes that emerged and we believe are associated with breast cancer. In section 4.5 the biological evaluation of the 89 and 51 final genomic signatures follows.

4.5. Biological Evaluation

In this work, we performed enrichment analyses in order to gain meaningful biological pathways with statistical significance of both Activity Vector and HotNet2 resulting subnetworks that might reflect the phenotype of breast cancer.

Entrez Gene ID	Gene Symbol	Description
SubNetwork 1		**(Module Score: 0.091)**
602	BCL3	B-cell CLL/lymphoma 3
7874	USP7	ubiquitin specific peptidase 7 (herpes virus-associated)
9184	BUB3	budding uninhibited by benzimidazoles 3 homolog (yeast)
55733	HHAT	hedgehog acyltransferase
1062	CENPE	centromere protein E, 312kDa
3059	HCLS1	hematopoietic cell-specific Lyn substrate 1
5603	MAPK13	mitogen-activated protein kinase 13
5307	PITX1	paired-like homeodomain 1
9618	TRAF4	TNF receptor-associated factor 4
SubNetwork 2		**(Module Score: 0.09)**
3661	IRF3	interferon regulatory factor 3
602	BCL3	B-cell CLL/lymphoma 3
7965	AIMP2	aminoacyl tRNA synthetase complex-interacting multifunctional protein 2
9636	ISG15	ISG15 ubiquitin-like modifier
8349	HIST2H2BE	histone cluster 2, H2be
6696	SPP1	secreted phosphoprotein 1
3308	HSPA4	heat shock 70kDa protein 4
9643	MORF4L2	mortality factor 4 like 2
3678	ITGA5	integrin, alpha 5 (fibronectin receptor, alpha polypeptide)
56259	CTNNBL1	catenin, beta like 1
SubNetwork 3		**(Module Score: 0.081)**
10015	PDCD6IP	programmed cell death 6 interacting protein
145781	GCOM1	GRINL1A complex locus
10594	PRPF8	PRP8 pre-mRNA processing factor 8 homolog (S. cerevisiae)
154	ADRB2	adrenergic, beta-2-, receptor, surface
4734	NEDD4	neural precursor cell expressed, developmentally down-regulated 4
347	APOD	apolipoprotein D
10014	HDAC5	histone deacetylase 5
2887	GRB10	growth factor receptor-bound protein 10
4649	MYO9A	myosin IXA
64750	SMURF2	SMAD specific E3 ubiquitin protein ligase 2
1975	EIF4B	eukaryotic translation initiation factor 4B
2260	FGFR1	fibroblast growth factor receptor 1
SubNetwork 4		**(Module Score: 0.1)**
1869	E2F1	E2F transcription factor 1
1072	CFL1	cofilin 1 (non-muscle)
10016	PDCD6	programmed cell death 6
51176	LEF1	lymphoid enhancer-binding factor 1
2624	GATA2	GATA binding protein 2
8349	HIST2H2BE	histone cluster 2, H2be
805	CALM2	calmodulin 2 (phosphorylase kinase, delta)
3014	H2AFX	H2A histone family, member X
SubNetwork 5		**(Module Score: 0.086)**
1386	ATF2	activating transcription factor 2
57161	PELI2	pellino homolog 2 (Drosophila)
59307	SIGIRR	single immunoglobulin and toll-interleukin 1 receptor (TIR) domain
84991	RBM17	RNA binding motif protein 17
51232	CRIM1	cysteine rich transmembrane BMP regulator 1 (chordin-like)
3654	IRAK1	interleukin-1 receptor-associated kinase 1
6663	SOX10	SRY (sex determining region Y)-box 10
7528	YY1	YY1 transcription factor
1459	CSNK2A2	casein kinase 2, alpha prime polypeptide
1975	EIF4B	eukaryotic translation initiation factor 4B

Table 4.5 List of genes that participate in Activity vector subnetworks. The 89 genes of Activity vector network are mapped to the corresponding Entrez Gene IDs and described according to their encoded gene products. Breast cancer-associated genes are highlighted in red. Brown highlighted Entrez Gene IDs are the overlapping genes within the Activity vector subnetworks. The starting node of each subnetwork is highlighted in a green background, while the 10 common genes of both algorithms are highlighted in a blue background

As demonstrated in the methodology and results section, a number of fourteen significant subnetworks have been obtained by using Activity Vector and HotNet2 algorithmic approaches. The genes of the resulting subnetworks were mapped to unique Entrez Gene IDs, and described according to their encoded gene products (Tables 4.4, 4.5), upon which enrichment analysis methods have been applied.

Entrez Gene ID	Gene Symbol	Description
SubNetwork 6		**(Module Score: 0.086)**
5094	PCBP2	poly(rC) binding protein 2
7220	TRPC1	transient receptor potential cation channel, subfamily C, member 1
2049	EPHB3	EPH receptor B3
8805	TRIM24	tripartite motif-containing 24
23294	ANKS1A	ankyrin repeat and sterile alpha motif domain containing 1A
1465	CSRP1	cysteine and glycine-rich protein 1
6342	SCP2	sterol carrier protein 2
1910	EDNRB	endothelin receptor type B
1975	EIF4B	eukaryotic translation initiation factor 4B
857	CAV1	caveolin 1, caveolae protein, 22kDa
SubNetwork 7		**(Module Score: 0.087)**
3624	INHBA	inhibin, beta A
10097	ACTR2	ARP2 actin-related protein 2 homolog (yeast)
23012	STK38L	serine/threonine kinase 38 like
10220	GDF11	growth differentiation factor 11
6597	SMARCA4	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 4
578	BAK1	BCL2-antagonist/killer 1
5366	PMAIP1	phorbol-12-myristate-13-acetate-induced protein 1
4170	MCL1	myeloid cell leukemia sequence 1 (BCL2-related)
2017	CTTN	cortactin
SubNetwork 8		**(Module Score: 0.084)**
9774	BCLAF1	BCL2-associated transcription factor 1
1915	EEF1A1	eukaryotic translation elongation factor 1 alpha 1
6059	ABCE1	ATP-binding cassette, sub-family E (OABP), member 1
10014	HDAC5	histone deacetylase 5
6201	RPS7	ribosomal protein S7
8665	EIF3F	eukaryotic translation initiation factor 3, subunit F
468	ATF4	activating transcription factor 4 (tax-responsive enhancer element B67)
7528	YY1	YY1 transcription factor
1459	CSNK2A2	casein kinase 2, alpha prime polypeptide
SubNetwork 9		**(Module Score: 0.073)**
3661	IRF3	interferon regulatory factor 3
5770	PTPN1	protein tyrosine phosphatase, non-receptor type 1
4800	NFYA	nuclear transcription factor Y, alpha
2624	GATA2	GATA binding protein 2
8467	SMARCA5	SWI/SNF related, matrix associated, actin dependent regulator of chromatin, subfamily a, member 5
2029	ENSA	endosulfine alpha
4172	MCM3	minichromosome maintenance complex component 3
3091	HIF1A	hypoxia inducible factor 1, alpha subunit (basic helix-loop-helix transcription factor)
2026	ENO2	enolase 2 (gamma, neuronal)
10971	YWHAQ	tyrosine 3-monooxygenase/tryptophan 5-monooxygenase activation protein, theta polypeptide
6720	SREBF1	sterol regulatory element binding transcription factor 1
8318	CDC45L	CDC45 cell division cycle 45-like (S. cerevisiae)
SubNetwork 10		**(Module Score: 0.083)**
9774	BCLAF1	BCL2-associated transcription factor 1
72	ACTG2	actin, gamma 2, smooth muscle, enteric
55128	TRIM68	tripartite motif-containing 68
8850	KAT2B	K(lysine) acetyltransferase 2B
29903	CCDC106	coiled-coil domain containing 106
5775	PTPN4	protein tyrosine phosphatase, non-receptor type 4 (megakaryocyte)
3189	HNRNPH3	heterogeneous nuclear ribonucleoprotein H3 (2H9)
10524	KAT5	K(lysine) acetyltransferase 5
1937	EEF1G	eukaryotic translation elongation factor 1 gamma
7528	YY1	YY1 transcription factor
468	ATF4	activating transcription factor 4 (tax-responsive enhancer element B67)

Table 4.5 (continue) List of genes that participate in Activity vector subnetworks. The 89 genes of Activity vector network are mapped to the corresponding Entrez Gene IDs and described according to their encoded gene products. Breast cancer-associated genes are highlighted in red. Brown highlighted Entrez Gene IDs are the overlapping genes within the Activity vector subnetworks. The starting node of each subnetwork is highlighted in a green background, while the 10 common genes of both algorithms are highlighted in a blue background

Entrez Gene ID	Gene Symbol	Description
SubNetwork 1		**(*p*- value = 0.04)**
5243	ABCB1	ATP-binding cassette, sub-family B (MDR/TAP), member 1
10097	ACTR2	ARP2 actin-related protein 2 homolog (yeast)
1058	CENPA	centromere protein A
8527	DGKD	diacylglycerol kinase, delta 130kDa
1729	DIAPH1	diaphanous homolog 1 (Drosophila)
1730	DIAPH2	diaphanous homolog 2 (Drosophila)
8665	EIF3F	eukaryotic translation initiation factor 3, subunit F
10456	HAX1	HCLS1 associated protein X-1
3059	HCLS1	hematopoietic cell-specific Lyn substrate 1
3709	ITPR2	inositol 1,4,5-triphosphate receptor, type 2
9764	KIAA0513	KIAA0513
4599	MX1	(mouse)
5311	PKD2	polycystic kidney disease 2 (autosomal dominant)
63898	SH2D4A	SH2 domain containing 4A
6455	SH3GL1	SH3-domain GRB2-like 1
7137	TNNI3	troponin I type 3 (cardiac)
7220	TRPC1	transient receptor potential cation channel, subfamily C, member 1
23214	XPO6	exportin 6
SubNetwork 2		**(*p*- value = 0.04)**
9255	AIMP1	aminoacyl tRNA synthetase complex-interacting multifunctional protein 1
7965	AIMP2	aminoacyl tRNA synthetase complex-interacting multifunctional protein 2
435	ASL	argininosuccinate lyase
8880	FUBP1	far upstream element (FUSE) binding protein 1
50628	GEMIN4	gem (nuclear organelle) associated protein 4
3212	HOXB2	homeobox B2
3735	KARS	lysyl-tRNA synthetase
4332	MNDA	myeloid cell nuclear differentiation antigen
5859	QARS	glutaminyl-tRNA synthetase
5917	RARS	arginyl-tRNA synthetase
293	SLC25A6	solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 6
6632	SNRPD1	small nuclear ribonucleoprotein D1 polypeptide 16kDa
6647	SOD1	superoxide dismutase 1, soluble
51593	SRRT	serrate RNA effector molecule homolog (Arabidopsis)
SubNetwork 3		**(*p*- value = 0.04)**
347	APOD	apolipoprotein D
1191	CLU	clusterin
3101	HK3	hexokinase 3 (white cell)
3952	LEP	leptin
3953	LEPR	leptin receptor
4023	LPL	lipoprotein lipase
64386	MMP25	matrix metallopeptidase 25
5730	PTGDS	prostaglandin D2 synthase 21kDa (brain)
5649	RELN	reelin
7436	VLDLR	very low density lipoprotein receptor
SubNetwork 4		**(*p-value* = 0.07)**
311	ANXA11	annexin A11
310	ANXA7	annexin A7
2026	ENO2	enolase 2 (gamma, neuronal)
3189	HNRNPH3	heterogeneous nuclear ribonucleoprotein H3 (2H9)
3483	IGFALS	insulin-like growth factor binding protein, acid labile subunit
10016	PDCD6	programmed cell death 6
11040	PIM2	pim-2 oncogene
6144	RPL21	ribosomal protein L21
59307	SIGIRR	single immunoglobulin and toll-interleukin 1 receptor (TIR) domain

Table 4.6 List of genes that participate in HotNet2 subnetworks. The 51 genes of HotNet2 network are mapped to the corresponding Entrez Gene IDs and described according to their encoded gene products. Breast cancer-associated genes are highlighted in red. The 10 common genes of both algorithms are highlighted in a blue background

In particular, we utilized two enrichment analysis systems: (i) The WEB-based GEne SeT AnaLysis Toolkit (WebGestalt) [70], which is a software application that enables the translation of the identified gene sets into a better comprehension of key biological issues, and (ii) The Genes-to-Systems Breast Cancer (G2SBC) Database [71] which is a bioinformatics source that collates and integrates gene, transcript and protein data which are known from the literature to be associated with alterations in breast cancer cells.

Enrichment analysis, defined as a secondary analysis on large gene sets that derived from high-throughput experiments (e.g. microarrays), can provide statistically over-or under-represented terms which are within a genomic signature of interest giving an insight into the underlying biological themes.
Thus, statistical methods such as hypergeometric distribution and cumulative hypergeometric distribution that are used by WebGestalt and G2SBC respectively enable us to identify enriched paths in each genomic subnetwork signature and to consider that these routes have key functions in our study.

Enriched Pathways	
WebGestalt	**G2SBC**
SubNetwork 1 (p-value = 0.04)	
Phosphatidylinositol signaling system (KEGG) (DGKD ITPR2) [adjP=0.0013]	Phosphatidylinositol signaling system (KEGG) (ITPR2 DGKD) [P=3.18E-02]
Pancreatic secretion (KEGG) (***TRPC1*** ITPR2) [adjP=0.0013]	
Calcium signaling pathway (KEGG) (***TRPC1*** ITPR2) [adjP=0.0031]	
Regulation of actin cytoskeleton (KEGG) (DIAPH2 DIAPH1) [adjP=0.0035]	
SubNetwork 2 (p-value = 0.04)	
Aminoacyl-tRNA biosynthesis (KEGG) (KARS RARS QARS) [adjP=3.21e-06]	Aminoacyl-tRNA biosynthesis (KEGG) (KARS QARS RARS) [P=1.94E-04]
	Gene Expression (Reactome) (KARS QARS RARS SNRPD1 ***AIMP2*** AIMP1) [P=3.10E-03]
	Metabolism of non-coding RNA (Reactome) (SNRPD1 GEMIN4) [P=3.82E-03]
SubNetwork 3 (p-value = 0.04)	
Adipocytokine signaling pathway (KEGG) (LEP LEPR) [adjP=0.0005]	Adipocytokine signaling pathway (KEGG) (LEP LEPR) [P=7.94E-03]
Jak-STAT signaling pathway (KEGG) (LEP LEPR) [adjP=0.0015]	Jak-STAT signaling pathway (KEGG) (LEP LEPR) [P=3.89E-02]
Cytokine-cytokine receptor interaction (KEGG) (LEP LEPR) [adjP=0.0021]	
Neuroactive ligand-receptor interaction (KEGG) (LEP LEPR) [adjP=0.0021]	
Metabolic pathways (KEGG) (HK3 PTGDS) [adjP=0.0268]	
AMPK signaling (Wiki) (LEP LEPR) [adjP=0.0003]	
Leptin signaling pathway (Wiki) (LEP LEPR) [adjP=0.0003]	
Adipogenesis (Wiki) (LPL LEP) [adjP=0.0004]	
SubNetwork 4 (p-value = 0.07)	
Lack of statistical significance for the identified biological pathways	Lack of statistical significance for the identified biological pathways

Table 4.7 Enriched pathways in HotNet2 subnetworks. Breast cancer-associated genes are highlighted in red. Common genes of both algorithms are in italic and bold. The common pathways of both WebGestalt and G2SBC enrichment analyses are highlighted in a green background

Tables 4.6 and 4.7 demonstrate the results of enrichment analysis of genomic subnetwork signatures of both algorithmic approaches. Enrichment analysis revealed that a number of genes in each subnetwork of both algorithmic approaches participate in significantly enriched pathways (<0.05), such as EGF-EGFR Signaling Pathway, Apoptosis, Jak-STAT signaling pathway that are well-known cancer-related pathways regarding the “hallmarks of cancer” [72,73].

Enriched Pathways	
WebGestalt	**G2SBC**
SubNetwork 1 (Module Score: 0.091)	
Regulation of toll-like receptor signaling pathway (Wiki) (USP7 MAPK13) [adjP=0.0004]	Lack of statistical significance for the identified biological pathways
SubNetwork 2 (Module Score: 0.09)	
RIG-I-like receptor signaling pathway (KEGG) (IRF3 ISG15) [adjP=0.0003]	RIG-I-like receptor signaling pathway (KEGG) (IRF3 ISG15) [P=8.89E-03]
ECM-receptor interaction (KEGG) (ITGA5 SPP1) [adjP=0.0003]	ECM-receptor interaction (KEGG) (ITGA5 SPP1) [P=1.23E-02]
Toll-like receptor signaling pathway (KEGG/Wiki) (SPP1 IRF3) [adjP=0.0003, adjP=0.0006]	Toll-like receptor signaling pathway (KEGG) (IRF3 SPP1) [P=1.75E-02]
Focal adhesion (KEGG/Wiki) (ITGA5 SPP1) [adjP=0.0009, adjP=0.0008]	Integrin cell surface interactions (Reactome) (ITGA5 SPP1) [P=2.45E-02]
Osteoclast Signaling (Wiki) (SPP1 *AIMP2*) [adjP=3.30e-05]	
Regulation of toll-like receptor signaling pathway (Wiki) (SPP1 IRF3) [adjP=0.0008]	
SubNetwork 3 (Module Score: 0.081)	
Endocytosis (KEGG) (PDCD6IP ADRB2 SMURF2 NEDD4) [adjP=4.40e-07]	Endocytosis (KEGG) (ADRB2 NEDD4 PDCD6IP SMURF2) [P=8.44E-04]
Ubiquitin mediated proteolysis (KEGG) (SMURF2 NEDD4) [adjP=0.0006]	Ubiquitin mediated proteolysis (KEGG) (NEDD4 SMURF2) [P=4.44E-02]
Neural Crest Differentiation (Wiki) (HDAC5 FGFR1) [adjP=0.0005]	
SubNetwork 4 (Module Score: 0.1)	
Melanogenesis (KEGG) (CALM2 LEF1) [adjP=0.0003]	Telomere Maintenance (Reactome) (H2AFX HIST2H2BE) [P=6.92E-03]
Pathways in cancer (KEGG) (LEF1 E2F1) [adjP=0.0020]	Melanogenesis (KEGG) (CALM2 LEF1) [P=1.14E-02]
miRNAs involved in DDR (Wiki) (H2AFX E2F1) [adjP=2.95e-05]	
miRNA regulation of DNA Damage Response (H2AFX E2F1) [adjP=0.0002]	
DNA damage response (Wiki) (H2AFX E2F1) [adjP=0.0002]	
Adipogenesis (Wiki) (GATA2 E2F1) [adjP=0.0004]	
EGF-EGFR Signaling Pathway (Wiki) (CFL1 E2F1) [adjP=0.0006]	
SubNetwork 5 (Module Score: 0.086)	
Regulation of toll-like receptor signaling pathway (Wiki) (***SIGIRR*** IRAK1 PELI2) [adjP=5.26e-06]	Lack of statistical significance for the identified biological pathways
SubNetwork 6 (Module Score: 0.086)	
Calcium signaling pathway (KEGG) (***TRPC1*** EDNRB) [adjP=0.0007]	Calcium signaling pathway (KEGG) (EDNRB ***TRPC1***) [P=5.01E-02]
SubNetwork 7 (Module Score: 0.087)	
Apoptosis (Wiki) (BAK1 MCL1 PMAIP1) [adjP=1.56e-06]	Apoptosis (Reactome) (BAK1 PMAIP1) [P=4.53E-02]
DNA damage response (only ATM dependent) (BAK1 PMAIP1) [adjP=0.0002]	
SubNetwork 8 (Module Score: 0.084)	
RNA transport (KEGG) (***EIF3F*** EEF1A1) [adjP=0.0004]	Metabolism of proteins (Reactome) (EEF1A1 RPS7 ***EIF3F***) [P=1.74E-02]
Translation Factors (Wiki) (***EIF3F*** EEF1A1) [adjP=4.91e-05]	3' -UTR-mediated translational regulation (Reactome) (RPS7 ***EIF3F***) [P=3.18E-02]
SubNetwork 9 (Module Score: 0.073)	
Cell cycle (KEGG/Wiki) (YWHAQ CDC45L MCM3) [adjP=1.00e-05, adjP=1.05e-05]	Cell cycle (KEGG) (MCM3 CDC45L YWHAQ) [P=3.36E-03]
Insulin signaling pathway (KEGG) (PTPN1 SREBF1) [adjP=0.0007]	Insulin signaling pathway (KEGG) (PTPN1 SREBF1) [P=4.38E-02]
SIDS Susceptibility Pathways (Wiki) (YWHAQ NFYA HIF1A GATA2) [adjP=1.70e-06]	DNA Replication (Reactome) (MCM3 CDC45L) [P=4.56E-02]
Adipogenesis (Wiki) (HIF1A GATA2 SREBF1) [adjP=1.15e-05]	
DNA Replication (Wiki) (CDC45L MCM3) [adjP=9.11e-05]	
SREBP signalling (Wiki) (NFYA SREBF1) [adjP=0.0002]	
G1 to S cell cycle control (Wiki) (CDC45L MCM3) [adjP=0.0002]	
SubNetwork 10 (Module Score: 0.083)	
Androgen receptor signaling pathway (Wiki) (10524 8850) [adjP=0.0002]	Lack of statistical significance for the identified biological pathways

Table 4.8 Enriched pathways in Activity Vector subnetworks. Breast cancer-associated genes are highlighted in red. Starting nodes are underlined, while common genes of both algorithms are in italic and bold. The common pathways of both WebGestalt and G2SBC enrichment analyses are highlighted in a green background

Considering the subnetworks of each algorithm, both WebGestalt and G2SBC enrichment analyses yielded significant pathways and some paths are in convergence among them (Tables 4.6, 4.7). This is expected, given that the G2SBC is a specific source with breast cancer annotations. This specificity can also explain the fact that some pathways identified

by G2SBC enrichment analysis did not achieve statistical significance in Activity Vector subnetworks 1, 5 and 10 (Table 4.7).
It is interesting to note that the subnetwork 4 that derived from HotNet2 algorithm lacks of both statistical and biological significance in terms of network structure and biological entity respectively (Table 4.6).

Considering the subnetworks of both algorithms, enrichment analyses by WebGestalt and G2SBC revealed that only two pathways are in convergence between Activity Vector and HotNet2 algorithms (Tables 4.6, 4.7). The common enriched pathways are the “Calcium signaling pathway” and “Adipogenesis” paths with implications in breast cancer pathology [74,75]. The intracellular Ca2+ is known as one of the vital signalings in regulation of several cellular functions, and its homeostasis dysregulation is recognized as one of the driving forces in proliferation, migration, invasion, and metastasis [74]. Adipogenesis is “the process during which fibroblast like preadipocytes developed into mature adipocytes”, which in turn are “non-trivial, dynamic partners of breast cancer cells”. It is reported that adipocytes located close to invasive cancer cells are fundamental for breast tumor development and progression [75]. Possibly, the essential roles of “Adipogenesis” and “Calcium signaling pathway” in the development and expression of breast cancer phenotype justify their appearance in both Activity Vector and HotNet2 subnetworks.

Moreover, we found that a large number of the identified pathways of both algorithms such as Focal adhesion, Insulin signaling pathway, Toll-like receptor signaling, Pathways in cancer, RIG-I-like receptor signaling pathway, Adipocytokine signaling pathway, Calcium signaling pathway and Cell cycle, reported as “risk pathways” in a recent breast cancer study [76] where a combined meta-analysis on multi-omics breast cancer data (gene expression, DNA methylation, DNA copy number and somatic mutation) was conducted.

Finally, performing an enrichment analysis on the whole genomic network signature of each algorithm, we found a large number of significant pathways, additionally to the aforementioned pathways. In order to achieve an overview of these results, we annotated specific pathway terms with generic pathway terms, as shown in Table 4.8, which also highlight the biological significance of the discovered pathways in relation to breast cancer [73].

Networks and SubNetworks Comparison	
Activity Vector	**HotNet2**
10 SubNetworks (Module Scores: 0.07-0.1)	4 SubNetworks (P-values: 0.04 & 0.07)
Activity Vector Network includes a list of **89 Genes**	HotNet2 Network includes a list of **51 Genes**
31 genes (34.83%) are associated with breast cancer	20 genes (39.21%) are associated with breast cancer
Common Pathways in Networks: RNA transport, **Calcium signaling pathway**, Focal adhesion, Regulation of actincytoskeleton, **Adipogenesis**, AGE-RAGE pathway **Common Pathways in SubNetworks: Calcium signaling pathway, Adipogenesis**	
Enriched Pathways	
Distinct Pathways in SubNetworks and Networks of both algorithms are involved in same processes: Gene Expression, Signal transduction, Signaling molecules and interaction, Cellular community, Cell motility, Digestive system, Endocrine system	
Distinct Activity Vector Pathways are involved in different processes: Folding, sorting and degradation, Cell growth and death, Development, Immune system, Nervous system, Circulatory system, Cancers	**Distinct HotNet2 Pathways** are involved in different processes: Metabolism, Lipid metabolism, Metabolism of nucleotides

Table 4.9 Comparison of both Activity Vector and HotNet2 resulting networks and subnetworks according to their enrichment analysis

Enrichment Analyses on the SVM and RVM genomic signatures of Activity Vector algorithm gave similar results by discovering biologically significant pathways. SVM was slight superior to RVM (data not shown). The appearance of a minor number (one) of common genes in SVM and RVM signatures explains further the crucial role of the interconnected genes in both Activity Vector and HotNet2 genomic subnetwork signatures and the central role of subnetwork/network structure as prediction model.

5

Conclusion

The aim of this thesis was to conclude in a significant genomic signature, to examine the pathways that are involved in cancer development or give rise to metastasis and to compare the results from both algorithms HotNet2 and Activity Vector, implemented. We presented an approach for identifying significant gene interaction networks and accurate genomic signatures by implementing two different methodological approaches, HotNet2 and Activity Vector algorithm. The proposed methodology is performed on a dataset that is composed of two different populations. Particularly, the original dataset consists of 529 samples related to breast cancer, 104 of which correspond to patients that are healthy/control, while 425 constitute the cancer samples, in combination with a large network of known interactions, MultiNet. The dataset was preprocessed through an univariate filter method Significance Analysis of Microarrays (SAM). First we examined the data by applying the HotNet2 algorithm and resulted in 4 significant subnetworks that implicate 51 genomic signatures. Furthermore, using the same criteria we implemented the Activity Vector algorithm which concluded, after a different approach from HotNet2, in 332 real subnetworks/markers. From the resulted subnetworks we examined the possibility of finding common pathways or genes. In particular, we combined the resulting subnetworks from both methodological approaches and we came across 10 common genes. Taking into account these findings we decided to examine and evaluate the Activity Vector networks that include these 10 genomic signatures and their interaction to other genes, due to the fact that we have a larger amount of subnetworks (332 modules), thus the opportunity to examine many more interactions. This way, we concluded in 89 significant genes. Moreover, we are interested in the generalization ability of the observed results. The ability of how the results of a statistical analysis will generalize to an independent data set was evaluated as well as their biological significance. Finally, a good generalization performance is achieved when a genomic signature is able to predict the label of unseen samples correctly. In that manner, a new independent dataset is used and the procedure of 10 – fold cross validation is repeated. The observed mean classification accuracy was approximately 93% for 10 mean genes selected, performing very good generalization performance when it comes to the classification of unknown samples. We decided to focus on the original dataset of the 10 mutual genes as they constitute a good group of seeds in order to start an inquiry process with the group of 89 genes that derived from these 10 genes, and undergone evaluation through two different classification methods SVM, a

deterministic algorithm as well as RVM a probabilistic method. In addition, we found that single genes might not be biologically important and we consider that subnetwork evolution triggers groups of genes with biological implication.

It is ought to be mentioned that as in any computational approach, our findings are limited by the quality and quantity of input data. The methodology could further be applied after including additional samples, and better interaction networks.

As far as the two methodological approaches implemented are concerned, HotNet2 algorithm is an interactive open source algorithm suitable for other applications, both biological and non-biological. In particular, genome-wide association studies (GWAS) and other studies of genetic diseases face an analogous problem of identification of combinations of genetic variants with a statistically significant association to a phenotype. With an appropriate gene score, the HotNet2 algorithm can be applied to such data as we achieved in this study. On the other hand, Activity Vector algorithm is implemented and accessed through a Cytoscape plugin, PinnacleZ that gives us no access to the code but enables us to select different thresholds to be applied on the statistical tests as well as the ability to choose between t-test and mutual information test, to be performed on our data. Finally, both algorithms, especially HotNet2, demand large computational time.

Overall, after and the biological evaluation our analyses result in the following important issues:

- We identify 14 significant subnetworks that include:
  - well-known cancer-related pathways (Toll-like receptor signaling pathway, EGF-EGFR Signaling Pathway, Calcium signaling pathway) considering the “hallmarks of cancer” as well as
  - pathways with recently or less recognized roles in breast cancer including Leptin signaling pathway, and RIG-I-like receptor respectively.
- Significantly enriched pathways ($p<0.05$) indicate:
  - high internal consistency among subnetwork genes (as groups) of each algorithmic approach and
  - relatively low consistency between the resulted HotNet and Activity Vector subnetworks themselves.
- HotNet and Activity Vector subnetworks converge at a higher information level; when a specific pathway term is annotated with a generic pathway term (Signal transduction, Cell motility).
- The existence of a high number of cancer-associated genes in enriched pathways and in subnetworks highlights their major roles in interconnection of genes within the breast cancer-specific subnetworks, and in turn demonstrates the great utility of the subnetwork structure.

- Our network analysis provides a blueprint to explore new therapeutic opportunities across subnetworks in order to meet the requirements for flexible or advanced pathway identification and diagnostic prediction through classification.
- The extracted signatures composed of “89 significant genes” and “51 significant genes” have both discriminative and predictive properties and can play a classification role in discriminating healthy controls from breast cancer patients.
- The subnetwork structure, as a disease model, aids the study of biological interactions and facilitates clinical interventions.
- Considering many networks organization structures reveals significant common procedures but can also be used in a complementary way as to enrich the pathology aetiology and/or progression.

Our analysis verifies that the search of common genes within multiple signatures might ignore important similarities at a higher (pathway) level. Common pathways instead of genes reveal a wealth of similar procedures indicated by groups of genes in each signature

6

Implementation aspects

Finally we ought to mention the hardware/software used for the conduction of this thesis. In order to be able to run HotNet2 algorithm our system was:

Hardware:

- SUPERCASE PC CHASSIS SKP 378, MIDI TOWER
- SUPERCASE PSU 500W, SERIES FORCE, 12CM FA, power Supply
- MSI MB B85M-G43, SOCKET INTEL LGA1150, CS, motherboard
- CORSAIR RAM DIMM 8GB CMV8GX3M1A1600C11, D, Ram
  We upgraded the Ram memory due to the requirements of the system to 2x8GB
- INTEL CPU CORE i5 4590, 4C/4T, 3.30GHz, C, CPU
- CORSAIR SSD 2.5" 120GB CSSD-F120GBLS, MLC, Disk ssd
- TOSHIBA HDD 3.5"2TB, disk 2tb

Software:

- Linux/Unix
- Python 2.7
- NumPy 1.6.2
- SciPy 0.10.1
- NetworkX 1.7
- h5py 2.4.0
- Fortran or C compiler (optional but recommended for performance)

HotNet2 will likely work with additional versions of Python, h5py, NetworkX, NumPy, and SciPy, but alternative configurations has not been tested.

For the implementation of the Activity Vector algorithm our system was:

Software:

- Cytoscape_v2.6.3
- Cytoscape plugin PinnacleZ

References

[1] Available online:http://www.breastcancer.org/

[2] World Health Organization.*World Cancer Report 2014*. pp. Chapter 5.2. ISBN 92-832-0429-8 (2014)

[3] Strausberg, R. L., Simpson, A. J., Old, L. J., & Riggins, G. J. (2004). Oncogenomics and the development of new cancer therapies. *Nature*,*429*(6990), 469-474.

[4] Barnes B, Dupré J (2008). Genomes and what to make of them*. Chicago: University of Chicago Press.* ISBN 978-0-226-17295-8.

[5] Tipping, M. E. (2001). Sparse Bayesian learning and the relevance vector machine. *The journal of machine learning research*, *1*, 211-244.

[6] Jumali, R., Deris, S., Hashim, S. Z., Misman, M. F., & Mohamad, M. S. (2009, April). A study of network-based approach for cancer classification. In*Information Management and Engineering, 2009. ICIME'09. International Conference on* (pp. 505-509). IEEE.

[7] Pauling, J. K., Christensen, A. G., Batra, R., Alcaraz, N., Barbosa, E., Larsen, M. R., ... & Baumbach, J. (2014). Elucidation of epithelial–mesenchymal transition-related pathways in a triple-negative breast cancer cell line model by multi-omics interactome analysis. *Integrative Biology*, *6*(11), 1058-1068.

[8] Friedman, N., Linial, M., Nachman, I., & Pe'er, D. (2000). Using Bayesian networks to analyze expression data. *Journal of computational biology*, *7*(3-4), 601-620.

[9] Brazhnik, P., de la Fuente, A., & Mendes, P. (2002). Gene networks: how to put the function in genomics. *TRENDS in Biotechnology*, *20*(11), 467-472.

[10] Pavlidis, P., Lewis, D. P., & Noble, W. S. (2002, January). Exploring gene expression data with class scores. In *Pacific Symposium on Biocomputing*(Vol. 7, pp. 474-485).

[11] Pavlidis, P., Qin, J., Arango, V., Mann, J. J., & Sibille, E. (2004). Using the gene ontology for microarray data mining: a comparison of methods and application to age effects in human prefrontal cortex. *Neurochemical research*,*29*(6), 1213-1222.

[12] Doniger, S. W., Salomonis, N., Dahlquist, K. D., Vranizan, K., Lawlor, S. C., & Conklin, B. R. (2003). MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome biol*,*4*(1), R7.

[13] Drăghici, S., Khatri, P., Martins, R. P., Ostermeier, G. C., & Krawetz, S. A. (2003). Global functional profiling of gene expression. *Genomics*, *81*(2), 98-104.

[14] Subramanian, A., Tamayo, P., Mootha, V. K., Mukherjee, S., Ebert, B. L., Gillette, M. A., ... & Mesirov, J. P. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles.*Proceedings of the National Academy of Sciences of the United States of America*, *102*(43), 15545-15550.

[15] Tian, L., Greenberg, S. A., Kong, S. W., Altschuler, J., Kohane, I. S., & Park, P. J. (2005). Discovering statistically significant pathways in expression profiling studies. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(38), 13544-13549.

[16] Mendelsohn, A. R., & Brent, R. (1999). Protein interaction methods—toward an endgame. *Science*, *284*(5422), 1948-1950.

[17] Ideker, T., Ozier, O., Schwikowski, B., & Siegel, A. F. (2002). Discovering regulatory and signalling circuits in molecular interaction networks.*Bioinformatics*, *18*(suppl 1), S233-S240.

[18] Chen, J., & Yuan, B. (2006). Detecting functional modules in the yeast protein–protein interaction network. *Bioinformatics*, *22*(18), 2283-2290.

[19] Sharan, R., Suthram, S., Kelley, R. M., Kuhn, T., McCuine, S., Uetz, P., ... & Ideker, T. (2005). Conserved patterns of protein interaction in multiple species.*Proceedings of the National Academy of Sciences of the United States of America*, *102*(6), 1974-1979.

[20] Vandin, F., Upfal, E., & Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology*,*18*(3), 507-522.

[21] Cerami, E., Demir, E., Schultz, N., Taylor, B. S., & Sander, C. (2010). Automated network analysis identifies core pathways in glioblastoma. *PloS one*, *5*(2), e8918.

[22]Chuang, H. Y., Lee, E., Liu, Y. T., Lee, D., & Ideker, T. (2007). Network-based classification of breast cancer metastasis. *Molecular systems biology*, *3*(1), 140.

[23] Vandin, F., Clay, P., Upfal, E., & Raphael, B. J. (2012). Discovery of mutated subnetworks associated with clinical data in cancer. In *Pac Symp Biocomput*(Vol. 2012, pp. 55-66).

[24] Leiserson, M. D., Vandin, F., Wu, H. T., Dobson, J. R., Eldridge, J. V., Thomas, J. L., ... & Raphael, B. J. . (2015) Pan-cancer network analysis identifies combinations of rare somatic mutations across pathways and protein complexes. *Nature genetics,* 47(2), 106-114

[25]Schäfer, J., & Strimmer, K. (2005). An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics*, *21*(6), 754-764.

[26] Hanisch, D., Zien, A., Zimmer, R., & Lengauer, T. (2002, July). Co-clustering of biological networks and gene expression data. In *ISMB* (pp. 145-154).

[27]Hanisch, D., Zien, A., Zimmer, R., & Lengauer, T. (2002, July). Co-clustering of biological networks and gene expression data. In *ISMB* (pp. 145-154).

[28] Rapaport, F., Zinovyev, A., Dutreix, M., Barillot, E., & Vert, J. P. (2007). Classification of microarray data using gene networks. *BMC bioinformatics*,*8*(1), 35.

[29]Benso, A., Carlo, S. D., & Politano, G. (2011). A cDNA microarray gene expression data classifier for clinical diagnostics based on graph theory. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, *8*(3), 577-591.

[30] Parzen, E. (1962). On estimation of a probability density function and mode.*The annals of mathematical statistics*, 1065-1076.

[31] Huang, S. (1999). Gene expression profiling, genetic networks, and cellular states: an integrating concept for tumorigenesis and drug discovery. *Journal of Molecular Medicine*, *77*(6), 469-480.

[32]Wang, K., Narayanan, M., Zhong, H., Tompa, M., Schadt, E. E., & Zhu, J. (2009). Meta-analysis of inter-species liver co-expression networks elucidates traits associated with common human diseases. *PLoS Comput Biol*, *5*(12), e1000616.

[33] Wong, F., Carter, C. K., & Kohn, R. (2003). Efficient estimation of covariance selection models. *Biometrika*, *90*(4), 809-830.

[34]"National Human Genome Research Institute", [Online]. Available:http://www.genome.gov

[35]"Genetic Home Reference", [Online].
Available:http://ghr.nlm.nih.gov/handbook/basics/gene

[36] Jörnsten, R., Wang, H. Y., Welsh, W. J., & Ouyang, M. (2005). DNA microarray data imputation and significance analysis of differential expression.*Bioinformatics*, *21*(22), 4155-4161.

[37] Asyali, M. H., Colak, D., Demirkaya, O., & Inan, M. S. (2006). Gene expression profile classification: a review. *Current Bioinformatics*, *1*(1), 55-73.

[38] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, *3*, 1157-1182.

[39] Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

[40] Lison, P. *An introduction to machine learning.* (1996)

[41] de Sa Marques, J. P. (2001). Pattern Recognition: Concepts, Methods, and Applications.

[42] Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, *3*, 1157-1182.

[43] Shinde, S. P. (2011). Implementation of Pattern Recognition techniques and Overview of its applications in various areas of artificial intelligence.

[44] "Statistical classification", [Online].
Available: http://en.wikipedia.org/wiki/Classification.

[45] Cho, S. B., & Won, H. H. (2003, January). Machine learning in DNA microarray analysis for cancer classification. In *Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics 2003-Volume 19* (pp. 189-198). Australian Computer Society, Inc..

[46] Kim, Y. (2001). *Feature selection in supervised and unsupervised learning via evolutionary search* (Doctoral dissertation, The University of Iowa).

[47] Dougherty, E. R. (2005). Feature-selection overfitting with small-sample classifier design.

[48] Saeys, Y., Inza, I., & Larrañaga, P. (2007). A review of feature selection techniques in bioinformatics. *bioinformatics*, *23*(19), 2507-2517.

[49]"Linear and non linear classifiers" [Online].
Available:http://cs.brown.edu/courses/cs1955/fall2009/docs/lecture_10-27.pdf

[50] Gil Chu _ Jun Li, Balasubramanian Narasimhan, Robert Tibshirani, Virginia Tusher "Significance Analysis of Microarrays", Users guide and technical document

[51] Tusher, V. G., Tibshirani, R., & Chu, G. (2001). Significance analysis of microarrays applied to the ionizing radiation response. *Proceedings of the National Academy of Sciences*, *98*(9), 5116-5121.

[52] Aizerman, A., Braverman, E. M., & Rozoner, L. I. (1964). Theoretical foundations of the potential function method in pattern recognition learning.*Automation and remote control*, *25*, 821-837.

[53] Weirauch, M. T. (2011). Gene coexpression networks for the analysis of DNA microarray data. *Applied statistics for network biology: methods in systems biology*, 215-250.

[54] Xiao, Y. (2009). A tutorial on analysis and simulation of boolean gene regulatory network models. *Current genomics*, *10*(7), 511.

[55] Albert, I., Thakar, J., Li, S., Zhang, R., & Albert, R. (2008). Boolean network simulations for life scientists. *Source code for biology and medicine*, *3*(1), 1-8.

[56] MacKay, D. J. (1998). Introduction to Gaussian processes. *NATO ASI Series F Computer and Systems Sciences*, *168*, 133-166.

[57] Burges, C. J. (1998). A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, *2*(2), 121-167.

[58]Kondor, R. I., & Lafferty, J. (2002, July). Diffusion kernels on graphs and other discrete structures. In *Proceedings of the 19th international conference on machine learning* (pp. 315-322).

[59]"Heat equation" [Online] Available: en.wikipedia.org/wiki/Heat_equation

[60]Storey, J. D., & Tibshirani, R. (2003). Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences*, *100*(16), 9440-9445

[61]Lebanon, G., & Lafferty, J. D. (2003). Information Diffusion Kernels. In*Advances in Neural Information Processing Systems 15: Proceedings of the 2002 Conference* (Vol. 15, p. 391). MIT Press.

[62] Chung, F. (2007). The heat kernel as the pagerank of a graph. *Proceedings of the National Academy of Sciences*, *104*(50), 19735-19740.

[63]Berkhin, P. (2006). Bookmark-coloring algorithm for personalized pagerank computing. *Internet Mathematics*, *3*(1), 41-62.

[64] Vandin, F., Upfal, E., & Raphael, B. J. (2011). Algorithms for detecting significantly mutated pathways in cancer. *Journal of Computational Biology*,*18*(3), 507-522.

[65] Milo, R., Kashtan, N., Itzkovitz, S., Newman, M. E. J., & Alon, U. (2003). On the uniform generation of random graphs with prescribed degree sequences.*arXiv preprint cond-mat/0312028*.

[66] Cheadle, C., Vawter, M. P., Freed, W. J., & Becker, K. G. (2003). Analysis of microarray data using Z score transformation. *The Journal of molecular diagnostics*, *5*(2), 73-81.

[67] Naive Bayes and Text Classification I [Online]
Available:http://sebastianraschka.com/Articles/2014_naive_bayes_1.html

[68] Gene Network Inference Enginebased on Supervised Analysis [Online]
Available:http://www.genome.jp/tools/genies/help.html

[69] How genetic disorders are inherited [Online] Available: http://www.mayoclinic.org/tests-procedures/genetic-testing/multimedia/genetic-disorders/sls-20076216

[70] J. Wang, D. Duncan, Z. Shi, and B. Zhang, "WEB-based GEne SeT AnaLysis Toolkit (WebGestalt): update 2013," Nucleic Acids Res, vol. 41 (Web Server issue), pp. W77-83, 2013. Available: http://bioinfo.vanderbilt.edu/webgestalt/

[71] E. Mosca, R. Alfieri, I. Merelli, F. Viti, A. Calabria, and L. Milanesi, "A multilevel data integration resource for breast cancer study," BMC Syst Biol, vol. 4, pp. 76, 2010. Available: http://www.itb.cnr.it/breastcancer/

[72] P. Dutta, and W. X. Li, "Role of the JAK-STAT Signalling Pathway in Cancer, eLS, 2013.

[73] D. Hanahan, and R.A. Weinberg, "Hallmarks of cancer: the next generation," Cell, vol. 144, no. 5, pp. 646-674, 2011.

[74] Y.F. Chen, Y.T. Chen, W.T. Chiu, and M.R. Shen, "Remodeling of calcium signaling in tumor progression," J Biomed Sci, vol. 20, pp. 23, 2013.

[75] J Tan, E Buache, M.P. Chenard, N. Dali-Youcef, and M.C, Rio, "Adipocyte is a non-trivial, dynamic partner of breast cancer cells," Int J Dev Biol, vol. 55, no. 7-9, pp. 851-859, 2011..

[76] L. Wang, Y. Xiao, Y. Ping, J. Li, H. Zhao, F. Li, J. Hu, H. Zhang, Y. Deng, J. Tian, and X. Li, "Integrating multi-omics for uncovering the architecture of cross-talking pathways in breast cancer, PLoS One, vol. 9, no. 8, pp. e104282, 2014.

[77] S. Andò, I. Barone, C. Giordano, D. Bonofiglio, and S. Catalano, "The Multifaceted Mechanism of Leptin Signaling within Tumor Microenvironment in Driving Breast Cancer Growth and Progression," Front Oncol, vol. 4, pp. 340, 2014.